Speech Prosody 2004
Nara, Japan
March 23-26, 2004

F0 Peaks and Valleys Aligned with Non-Prominent Syllables Can Influence Perceived Prominence in Adjacent Syllables

Stefanie Shattuck-Hufnagel[1], *Laura Dilley*[1], *Nanette Veilleux*[2], *Alejna Brugos*[3], *Rob Speer*[1]

[1]Speech Group, RLE Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Computer Science Department, Simmons College. Boston, MA, USA
[3]Boston University, Boston, MA, USA
stef@speech.mit.edu

Abstract

The occurrence of peaks and valleys of the F0 contour of an utterance on non-prominent syllables in American English (as on the *–ing* or *a-* in *reading again*) raise the question of how to label these inflection points. Analysis of samples from prosodically-labelled corpora of natural speech*(MIT Maptask and BU FM Radio News) show that H* !H* sequences with an f0 peak on a weak syllable between them can occur quite commonly in continuous communicative speech. Informal listening suggests that the alignment of these f0 peaks with specific non-prominent syllables between the two accented syllables can change the perceived relative prominences of the accents. This observation is supported by results of perceptual experiments using synthesized F0 contours: the location of the peak in the weak syllable can shift the perceived strongest prominence from the initial syllable to the final syllable of a word like *lemonade* or *millionaire.* These findings illustrate the pervasiveness of F0 inflection points that are not aligned with syllables perceived as prominent, and suggest that alignment of the inflection point is a critical aspect of the specification of an intonational contour.

1. Introduction

The model of English intonation proposed by Pierrehumbert [1] has become the dominant phonological framework for analyzing intonation [2]. According to this model, termed the autosegmental (AM) model by Ladd [3], one phonetic characteristic which distinguishes pitch accent categories is the temporal alignment of an F0 maximum (peak) or minimum (valley) with respect to syllable boundaries [4][5]. The AM model also formed the basis for the ToBI transcription system for labelling **To**nes (tonal targets) and **B**reak **I**ndices (constituent boundaries) in spoken utterances [6]. One aspect of the system which presents a challenge for labellers is the occurrence of F0 peaks or valleys on non-prominent syllables. For example, in a sequence of two pitch accents H* followed by !H*, there can sometimes be an F0 peak on a nonaccented syllable between the two accented ones, e.g. on the *–ing* or the *a-* of *READing aGAIN.* Examples with early and late nonprominent F0 peaks are shown in figures 1 and 2. Labellers are sometimes confused about whether to label such examples as late peaks associated with the preceding H*, or as the H in an H+!H* complex pitch accent. Silverman and Pierrehumbert [7] have reported that the peak F0 of an H* can be delayed into the onset of the following syllable in this context (in the absence of an intervening word boundary or upcoming stress clash), and so one approach would be to label tokens which have a peak delayed only slightly (i.e. into onset of the syllable directly after the first H* syllable) as H* with delayed peak, but label tokens where the peak occurs further along in the sequence of syllables as H+!H*.

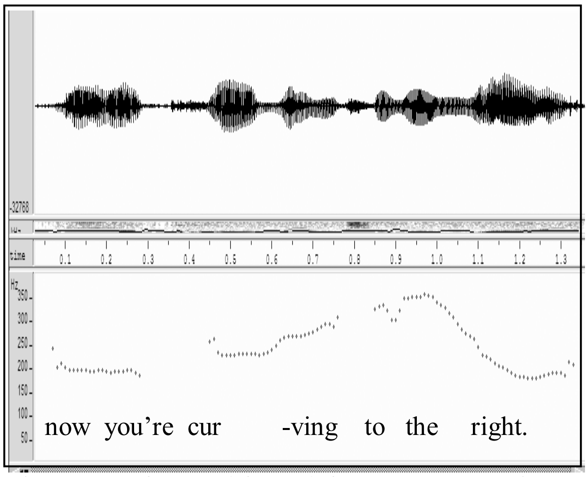

Fig. 1: *wave form and f0 track for* now you're curving to the right *with f0 peak on* the.

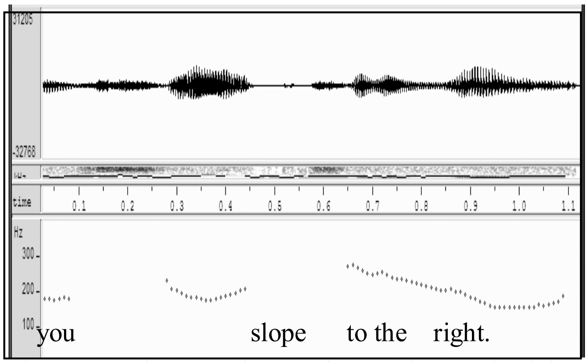

Fig. 2: *wave form and f0 track for* you slope to the right *with f0 peak on* to.

This study explores the appropriateness of such a criterion, by investigating two questions: how often does this configuration (i.e. H* !H* with an F0 peak on an intervening syllable) occur, and what are the perceptual consequences of alignment of the F0 inflection with intervening syllables closer to vs. further from each of the two pitch accents. If this contour occurs only rarely and if the precise location does not distinguish one intonational contour from another, then the

* This work is supported by NIH grants DC0075, DC02978, DC02125. We thank Dr. Olga Gubanova for recording the MIT Maptask corpus and Dr. Lisa Lavoie for transcribing it.

issue may not require much attention. But if the configuration occurs regularly, and the location of the intervening F0 peak distinguishes one accent contour from another, then a criterion needs to be established which will allow labellers to transcribe each category appropriately.

Table 1: Number of analyzable H!H* sequences with and without a non-prominent F0 peak on a weak syllable between the two accents; number of intervening weak syllables ranged from one to more than four.*

Speaker	Peak	No Peak	Total
Maptask	19	27	46
RadioNews 1	6	5	11
RadioNews 2	1	10	11
RadioNews 3	5	6	11

2. Analysis of delayed peaks in speech corpora

To determine whether H* !H* sequences with an F0 peak on an intervening syllable occur rarely or regularly, we sampled two corpora of continuous speech: the BU FM Radio News corpus, and the MIT Maptask corpus. The Maptask sample consisted of 520 words of spontaneous speech (235 seconds) from a single female speaker giving directions for drawing a map; the FM Radio News sample consisted of the same paragraph (73 words) read by three different female professional news readers (164 seconds). The Maptask sample had been labelled by one ToBI labeller and the FM Radio News sample by a different labeller. For both samples, we surveyed all the H* !H* sequences, where H* stands for either H* or L+H*, and !H* stands for either !H* or H+!H*. Tokens for which there was no intervening syllable between the two accents, or for which laughter prevented determination of the f0 contour, were removed from the sample set, and the remaining tokens were categorized according to whether they a) contained an F0 peak on a syllable between the two accented syllables, b) did not contain such a peak, or c) were indeterminate, either because the F0 track was irregular, or the F0 difference was too small to be reliable (2-3 Hz). Examples may be accessed at www.simmons.edu/~veilleux.

Table 1 shows the results for both samples; despite the small sample size, all speakers produced appreciable numbers of H* !H* sequences (as labelled by these ToBI labellers), and three of the four speakers produced an F0 peak on an intervening syllable in 1/2 to 1/3 of the tokens for which this decision could be made. Tokens with intervening peaks include 19 for the Maptask speaker (of 46 candidate H* !H* sequences) and 6, 5 and 1 for the three FM Radio News speakers (of 11 H* !H* sequences each).

Several aspects of this preliminary analysis have potentially significant implications. The first is that H* !H* sequences with an intervening F0 peak are not a rare phenomenon in continuous American English speech, highlighting the importance of developing a criterion for labelling them reliably in ToBI. Second, differences among FM Radio News speakers who read the same paragraph (5 and 6 vs. 1 intervening F0 peak) hint at a possible difference among speakers in how often they produce this configuration.

A third observation of interest concerns the effect of particular peak locations on the perceived relative prominence of the two accented syllables. Informal listening suggested that when the peak occurs on the syllable immediately after the H*, e.g. on the *–ing* in *READing aGAIN*, the initial H* accent is perceived as stronger than the later !H*. In contrast, when the peak occurs on the syllable immediately before the !H*, e.g. on the *a-* of *again*, the following !H* is perceived as the stronger of the two accents. (When only a single unaccented syllable intervenes between the two accented ones, as in *STONy DESert* or *RIGHT aCROSS*, a peak on this syllable also makes the following !H* stronger.) If the location of a peak on one vs. another of a string of unaccented syllables between two accented syllables results in a categorical difference in relative prominence, possibly corresponding to the H*-with-delayed-peak vs. H+!H* distinction, then it will be particularly important to determine where the alignment boundary is. We investigated this question using a new method for studying intonation categories developed in Dilley [8]: judgments of relative prominence. The experiments described here elicited judgments of relative prominence within words, i.e. listeners were asked to report the location of main lexical stress in stimuli for which the location of the F0 peak or valley is moved from an initial strong syllable through an intervening weak syllable in words like *millionaire*, which can take main stress on either the first or last syllable (Webster's Third Dictionary).

3. Experiment and Methods

In the course of developing stimuli for another experiment, Dilley [8] noted that the perceived location of the main stress on a target word seemed to change, depending on the temporal location of an F0 maximum or minimum. In particular, for words like *lemonade* which could have either a S'WS" or S"WS' stress pattern (e.g., *lémonade* vs. *lemonáde*), shifting an F0 peak or valley from the first strong syllable to the following weak syllable induced a perceived shift in the location of the main stress, from the first strong syllable (*'lemonade*) to the last (*lemo 'nade*). This is an expected result if alignment of the inflection point with the weak syllable directly preceding the second strong syllable creates the perception of greater prominence on that strong syllable. The following experiment explores this phenomenon of shiftable perceived prominence quantitatively.

3.1. Stimuli and Subjects

To test the robustness of the shift in perceived prominence as the result of non-prominent f0 peak location, a forced-choice task involving judgment of the relative prominence of two strong syllables was used. The experiment was based on stimulus continua designed by Dilley [8] in which an F0 maximum or minimum was shifted through a SWS or SWWS syllable sequence, corresponding to a target word that could have its main stress on one of the two strong target syllables. Each stimulus series was based on a single natural production of a target phrase, which was recorded in a sound-attenuated room in real time at 16 kHz using custom software designed by Mark Tiede. Praat software [9] was used to shift an F0 maximum or minimum through a SWS or a SWWS syllable sequence in 30 ms increments, as described below.

<u>*millionaire*</u>: The first series (Figure 3a) was based on the word *millionaire* in the phrase *For a millionaire*, produced with an overall falling (statement) intonation. An F0 maximum was shifted through the initial SW sequence of *millionaire* in

30 ms increments to yield 13 stimuli. Lannameraine: The second series (Figure 3b) was based on the nonsense word *Lannameraine* in the phrase *In Lannameraine*, produced with an overall falling (statement) intonation. In this series, an F0 maximum was shifted through the initial SWW sequence of the nonsense word to yield 18 stimuli. lemonade: The third series (Figure 3c) was based on the word *lemonade* in the phrase *Some lemonade*, produced with an overall rising (question) intonation. In this series, an F0 minimum (rather than maximum) was shifted through the initial SW sequence of the target word to yield 10 stimuli. For the *millionaire* and *lemonade* series, it was predicted that shifting the F0 peak or valley from the initial S syllable to the following W syllable would result in a perceived shift in the location of the main stress from the first syllable (*mil–* or *lem–*) to the last syllable (*–naire* or *–nade*). For the SWWS stimulus *Lannameraine*, it was predicted that the shift to judgments that the last syllable was the strongest would take place only for stimuli in which the F0 maximum was aligned with the second weak syllable, but not when it was aligned with the first weak syllable.

There were 19 subjects ranging in age from 18 to 45 years. All were self-reported native speakers of English with normal hearing, and were paid a nominal sum for participation.

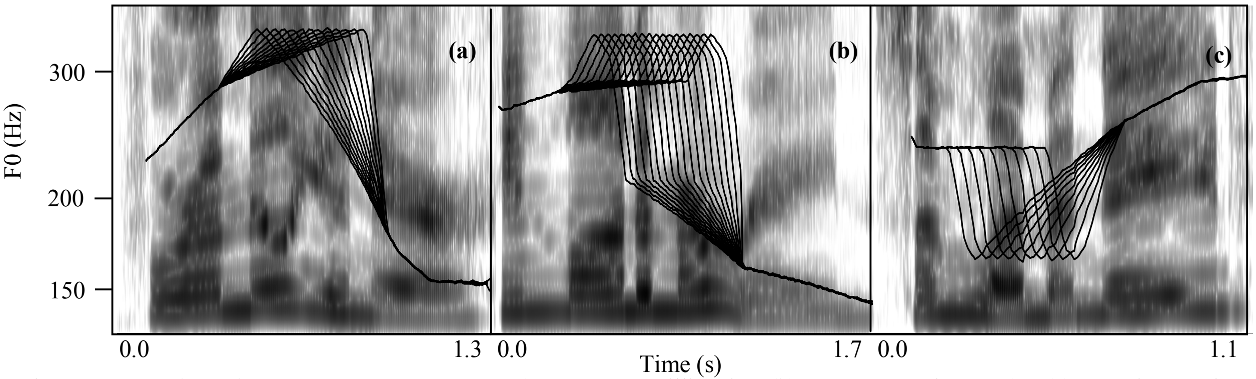

Figure 3: S*timuli used in experiment. Utterance in (a) was* For a millionaire*, (b)* In Lannameraine*, and in (c)* Some lemonade?

3.2. Task

Stimuli within each continuum were presented to listeners over headphones, at a comfortable volume, in blocks. At the beginning of the experiment, subjects were told that they would hear a spoken phrase containing a target word that could have one of two stress patterns. Subjects were instructed to decide which of the two stress patterns the speaker intended, to expect both stress patterns in each block of trials, and not to pay attention to loudness alone.

Stimuli were presented via computer over headphones using LISE software (Listening Interface for Sound Experiments). To hear a stimulus, subjects used a mouse to click a button on the computer screen labeled "Play". Subjects checked the box corresponding to the relative prominence pattern they heard (e.g., MILLionaire or millioNAIRE). The presentation of stimuli was blocked by stimulus continuum. Following practice trials, all stimuli within a continuum were presented in a given random order each of three times in succession. The order of presentation of stimulus blocks was counterbalanced across subjects. This 30 minute procedure resulted in three judgments per stimulus per subject for each of the three stimulus continua not counting skipped trials.

3.3. Analysis

For a given stimulus continuum, data from subjects who responded on fewer than 50% of trials were omitted from consideration. This resulted in two subjects' data being discarded for the *Lannameraine* series only. Moreover, data from subjects who gave responses from only one category were thrown out, since it was reasoned that such a response pattern indicated that the target word could not undergo stress shift in that subject's phonology. This resulted in one subject each being discarded from the *lemonade* and *millionaire* series.

The remaining subjects' responses as a group were somewhat variable. To determine the most consistent pattern of responses for each stimulus continuum, a two-tailed, bivariate correlation analysis was carried out using SPSS software. In this analysis, the response of subjects (e.g., whether the first or last syllable was the strongest) was treated as a binary-valued variable, and the average responses of individual subjects to each stimulus were correlated with one another in pairwise fashion. Subjects whose average responses were uncorrelated with 2/3 or more of the remaining subjects at $\alpha = .05$ were thrown out. There were 14, 15, and 8 subjects who met this latter criterion for the *millionaire*, *Lannameraine*, and *lemonade* series, respectively.

For these subjects, the rate of indicating that the last syllable was the strongest is shown in Figures 4a, b, and c. For the *millionaire* series (Figure 4a), stimuli which have a peak on the initial strong syllable are judged as having main stress on *mil–*, while stimuli with a peak on the intervening weak syllable are judged as having main stress on *–naire*. For the *Lannameraine* series (Figure 4b), stimuli which have a peak on the initial strong syllable are judged as having main stress on *Lan–*, while stimuli with a peak on the second weak syllable are judged as having main stress on *–raine*. Stimuli with a peak on the first weak syllable are judged somewhat more equivocally, but the majority of respondents indicated that *Lan-* was the stronger syllable for stimuli with this alignment. Finally, for the *lemonade* series (Figure 4c), stimuli which have a valley on the initial strong syllable are judged as having main stress on *lem–*, while stimuli with a valley on the following weak syllable are judged as having main stress on –*nade*.

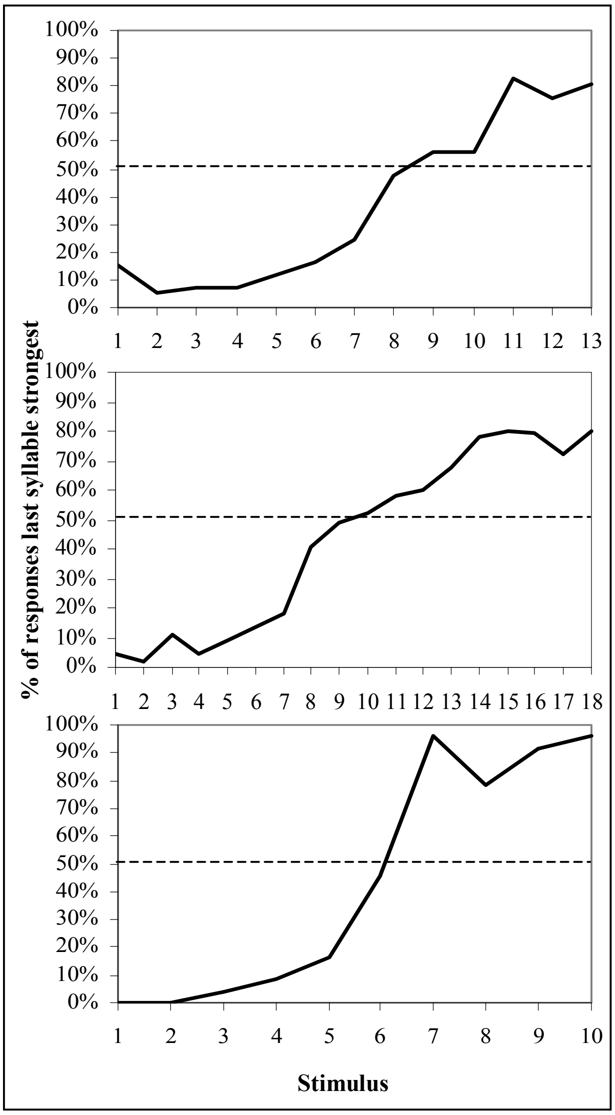

Figure 4: *Percentage of responses for three stimulus continua that the last syllable was the strongest. Top to bottom:* millionaire, Lannameraine, *and* lemonade.

4. Discussion

These results suggest that the timing of an F0 maximum or minimum in a non-prominent syllable can affect the perception of relative prominence among the strong syllables in a word. In a word such as *millionaire,* which may have its main stress in more than one location, shifting an F0 peak or valley to the weak syllable just before the late main stress syllable induces the perception that the upcoming syllable is the strongest syllable in the word. Bolinger [10] and Kohler [11] have reported that, for words with adjacent strong syllables, the perceived location of the main stress may correspond either to the first or to the second strong syllable, depending on the alignment of an F0 peak and other factors. However, we are not aware of any work which suggests that aligning an F0 inflection with the weak syllable preceding a strong syllable induces perception of the strongest prominence on that strong syllable. Critically, in a word such as *Lannameraine* an F0 peak in the first weak syllable induces the perception that the initial syllable is the strongest syllable in the word, while shifting this peak into the second weak syllable induces the perception that the last syllable is the strongest. Thus an F0 peak anywhere in the first weak syllable (not just in its onset) in these experiments is associated with greater prominence on the preceding strong syllable. These findings suggest not only that the alignment of non-prominent inflection points can determine judgments of relative prominence, but also that relative prominence judgment tasks can be used to probe category distinctions in intonation.

The results also suggest the need to revisit some of the claims of earlier studies. For example, Pierrehumbert and Steele [5] showed categorical effects in production using a continuum in which an F0 peak was shifted from the initial strong to a following weak syllable in the word *millionaire*. They interpreted this as evidence for a distinction between L*+H and L+H* pitch accents on *mil-*. However, our results suggest that when an F0 peak is shifted from the strong syllable *mil-* to the weak syllable *lio-* in *millionaire*, listeners perceive a shift in the stress form *mil-* to *-naire*. Further investigation of this phenomenon will show whether accents are perceived on both strong syllables.

5. General Discussion

Both the corpus sample analysis and the perceptual experiments suggest that the alignment of F0 inflection points with non-prominent syllables are important aspects of intonation contours in American English. The patterns in naturally-produced speech indicate that F0 peaks on non-accented syllables between accents are not marginal, but occur relatively frequently. The perceptual results show that the precise location of an F0 peak or valley on a weak syllable between two strong syllables strongly influences the perceived relative prominence of the two strong syllables. This claim is currently being tested in experiments using naturally-produced tokens with differently-located non-prominent f0 peaks and valleys for the relative prominence judgment task. Of particular interest will be strings with more than two weak syllables between the strong ones, the possible de-accenting effect of an early non-prominent peak on the later strong syllable and the role of word boundary location.

6. References

[1] Pierrehumbert, J.B., 1980. The phonology and phonetics of English intonation, Ph.D. thesis, MIT.

[2] Ladd, D.R.; Schepman, A., 2003. "Sagging transitions" between high accent peaks in English: experimental evidence**.** *Journal of Phonetics* 31: 81-112.

[3] Ladd, D.R., 1996. *Intonational Phonology*,Camb.Univ.Pr., 1996.

[4] Beckman, M.E.; Pierrehumbert, J.B.,1986. Intonational structure in Japanese and English, *Phonology Yearbook 3*.

[5] Pierrehumbert, J.B.; Steele, S.A., 1989. Categories of tonal alignment in English, *Phonetica 46.*

[6] Beckman, M.; Ayers-Elam, G.,1997. Guidelines for ToBI labeling, www.ling.ohio-state.edu/research/_phonetics/E_ToBI.

[7] Silverman, K.; Pierrehumbert,J.B.,1990. The timing of prenuclear high accents in English. In Kingston, J.; Beckman, M. (eds), *Laboratory Phonology I.* Cambridge: Cambridge University Press.

[8] Dilley, L.C., forthcoming. The phonetics and phonology of tonal systems. Ph. D. dissertation, MIT.

[9] Boersma, P.,2001. PRAAT, a system for doing phonetics by computer. Glot International 5(9/10): 341-345.

[10] Bolinger D., 1958. A theory of pitch accent in English.*Word 14.*

[11] Kohler ,1987, The linguistic functions of F0 peaks. *Proceedings of the ICPhS* Talinn, Estonia. Vol. 3.