University of Massachusetts Occasional Papers in Linguistics

Volume 22 UMOP 19: *Linguistics in the Laboratory*

Article 2

1996

Salient Pitch Cues in the Perception of Contrastive Focus

Christine Bartels
University of Massachusetts

John Kingston
University of Massachusetts

Follow this and additional works at: <https://scholarworks.umass.edu/umop>

Part of the [Linguistics Commons](#)

Recommended Citation

Bartels, Christine and Kingston, John (1996) "Salient Pitch Cues in the Perception of Contrastive Focus," *University of Massachusetts Occasional Papers in Linguistics*: Vol. 22 , Article 2.
Available at: <https://scholarworks.umass.edu/umop/vol22/iss1/2>

This Article is brought to you for free and open access by the Graduate Linguistics Students Association (GLSA) at ScholarWorks@UMass Amherst. It has been accepted for inclusion in University of Massachusetts Occasional Papers in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Salient Pitch Cues in the Perception of Contrastive Focus

Christine Bartels and John Kingston

University of Massachusetts

1 Background

The concept of contrastiveness can be characterized as "assertion of a proposition accompanied by explicit denial of salient alternatives." The status of this concept in linguistic theory, and in focus theory in particular, has long been a subject of debate. On the one hand, contrastiveness has been claimed to be a basic, bivalent semantic feature independent of focus structure. Thus a focused expression can be either [+contrastive] or [-contrastive]. The strict version of this view (represented by, e.g., Chafe 1976 and Taglicht 1982) has it that to be contrastive, an expression must be set up as one member of a contextually salient pair of opposites; that is, the set of alternatives is limited to two. The laxer version (represented by, e.g., Couper-Kuhlen 1984) requires merely "a closed set of items." In any case, the set of alternatives has to be construable from the context for an assertion to be contrastive. Thus the focus of, for example, the answer to a *wh*-question does not necessarily qualify.

On the other hand, it has been argued that the contrastiveness of a focus is merely a derived, and gradient, property of focus itself. It is said (e.g., by Ladd 1980, Sgall et al. 1986, and Bolinger 1989) that focus always, by definition, picks and asserts one member of a contextually given set and denies the alternative, so (since all assertions are assumed to contain a focus) there is an implied "rather than" in every assertion. For a given utterance, the degree of contrastiveness will correlate with focus scope: the narrower the focus, the smaller the set of alternatives and the more salient the denied members of this set. But even the broadest focus is minimally contrastive, from this second perspective, in singling out one member of the all-encompassing set "of all things." Thus, for example, the answer to a *wh*-question is always contrastive. The point is that, given suitable assumptions about the role of discourse context in providing or evoking the relevant set of alternatives, and about the role of focus as selectively asserting a member of this set, there is no need for an independent semantic feature 'contrastive'.

There have been several attempts in recent years to develop models of intonation that allow for the assignment of a distinct, context-independent interpretational counterpart not only to larger phrasal contours but also to each nuclear component of these contours—in other words, phonological models that allow for a compositional semantics of tunes. Taken to the extreme, such a research program will aim for a direct, one-to-one mapping between the inventory of intonational primitives and the basic semantic and pragmatic aspects they encode. But even a more cautious approach to the relation between tonal and interpretational primitives might consider selecting intonational primitives on nonphonological grounds if the intonation contours actually observed allow for more than one analysis; conversely, the primitives of an empirically successful model of intonation might at times be taken to provide clues to the basic components of utterance interpretation and communicative function.

It is because of this potential interaction that the theoretical status of the concept of contrastiveness is relevant for prosodic research. The literature on intonation contains plenty of discussion of the definition and distribution of “contrastive stress” and “contrastive accent.” In particular, the existence of a distinctive tune in English signaling contrastive vs. mere declarative assertion has often been claimed and equally often been denied, with both sides citing empirical observations regarding the physical characteristics of the pitch contours involved. Now, within the “direct-mapping approach” described above, proponents of the view that contrastiveness is an underived aspect of meaning should in fact expect to find a corresponding distinctive tune or tune feature; whereas those who claim that contrastiveness does not exist as an underlying semantic feature per se should not expect the existence of a characteristic tune signaling contrast independent of intonational cues to focus structure. At the same time, if empirical studies of intonation were to succeed in settling the issue of whether there are distinct acoustic cues which, alone and together, signal contrastiveness, we might allow the outcome of these studies to influence our view of contrastiveness as a semantic or pragmatic concept. This possibility was part of the motivation for the perception study reported here.

The study took off from Janet Pierrehumbert’s widely used model of intonational description (Pierrehumbert 1980, Beckman & Pierrehumbert 1986). According to this model, the underlying phonological representation of a pitch contour consists of a series of tonal targets taken from a limited inventory of accent types. A simple bivalent feature High(H)/Low(L) specifies tone height; importantly, neither absolute pitch level nor relative excursion size or the properties of the pitch movements that connect tonal targets are considered part of the underlying representation. Relevant for our purposes here, the model draws a categorical phonological distinction between the bitonal accent L+H*—a lowpoint/dip in the contour followed by a rapid rise to the peak on the mainstress syllable of the accented word—and the monotonal H*—a less steep rise from midlevel, or interpolation from a preceding accent, if any, to that same peak.

Now, it has been proposed by Pierrehumbert & Hirschberg (1990) that the interpretational difference between L+H* and H* is just the one at issue here: that the bitonal L+H* commonly conveys contrastive assertion, suggesting that the accented item and not some alternative item should be believed by the listener (and is of course believed by the speaker already), whereas the monotonal H* is merely neutrally assertive, indicating new information. What is unclear, though, is which specific acoustic properties of a pitch contour convey to a listener the occurrence of one of these types of accent rather than the other, especially since the dip in pitch accents classified on acoustic grounds as L+H* is not always clearly discernible in the pitch track. (Pitch track transcribers working within the Pierrehumbert framework are reported to often disagree on the nature of a high accent; see, e.g., Pitrelli 1994.) That is the issue we wished to address. We assumed that Pierrehumbert & Hirschberg's distinction between the meanings of the two accent types was descriptively correct and used it as diagnostic in our experiments; that is, we asked listeners to classify intonation contours as effectively contrastive or non-contrastive in interpretation, and assumed that contrastive contours fell into the L+H* category whereas noncontrastive ones were H*'s.

Altogether, three experiments were conducted. Experiments II and III were designed after Experiment I had been carried out, and aimed to avoid its demonstrated and potential problems. They used intonation contours somewhat different in phonetic detail from those employed in Experiment I, and a markedly different experimental design, as will be detailed below. At the same time, these later experiments, taken in combination, addressed the question as to which of the semantic possibilities sketched above might receive more support from intonational facts.

The structure of this paper follows the sequence of experiments. Section 2 describes the experimental method, data analysis, results, and conclusions for Experiment I. Section 3 similarly describes Experiments II and III. We conclude with a general discussion in Section 4.

2 Experiment I

2.1 Experimental Method

Stimuli consisted of the synthetic sentence *Amanda had a banana*, created on a terminal analogue speech synthesizer (Klatt & Klatt 1990). The overall pitch contour involved a high accent on both *Amanda* and *banana* (see Fig. 1). The F0 contour was the same for the first 650 ms of all stimuli. Stimuli differed only in the shape of the second accent, along the following dimensions (see Fig. 2):

Figure 1

1. Peak Height (PH): 8 levels, from 165 to 270 Hz, in 15-Hz steps
2. Dip/No-Dip (DND): 2 levels for rise onset point: 120 Hz (dip) and 150 Hz (nodip); reached via straight interpolation from 150-Hz end point of fixed contour at 650 ms
3. Timing of Rise Onset (RO): 3 points, at 80-ms intervals: 740, 820, and 900 ms ("early," "medial," and "late" relative to medial PT value given below; the 3 points fall into the /b/, first /a/, and first /n/ of *banana*, respectively)
4. Timing of Peak (PA): 3 points, at 80-ms intervals: 960, 1040, and 1120 ms ("early," "medial," and "late" relative to the 210-ms-long stressed /a/ of *banana*; the peak is positioned 50, 130, and 210 ms after vowel onset)

Accent peak height is known to correlate with general emphasis, or "speaker involvement" (in Bolinger's terms). The dip/no-dip dimension of course addresses the most straightforward interpretation of what constitutes the difference between H* and H+L* in Pierrehumbert's framework. The temporal alignment parameters (RO and PA) might contribute in their own right to contrastiveness (e.g.,

relatively late peaks have been claimed to be more emphatic than relatively early peaks), but were also needed to manipulate the steepness of the rise preceding the peak—another potential cue to contrastiveness—independent of dip depth. Thus in these stimuli the slope of the rise toward the peak is a function of dip depth, rise onset, as well as peak alignment settings.

Figure 2

The actual parameter values along each dimension were chosen based on recorded natural productions of the utterance, comparable intonation contours found in Pierrehumbert 1980, and pilot tests. The time intervals between alignment points were kept the same (80 ms) for RO and PA, so that effects due to variations in slope of rise could be separated from effects due to variation in either of the two alignment variables. Aside from slope of rise, the following additional variables can in principle be derived from the above set of basic variables: the interval rise-onset-to-peak (note that the interval between paired early, medial, and late RO and PA points corresponds roughly to the presumably typical bitonal interval of 200-220 ms between L and H* in L+H* accents), slope of fall, interval peak-to-fall-offset, and "area under the peak." Of these, only slope of fall and the peak-to-fall-offset interval, that is, post-peak properties not implicated by the Pierrehumbert categorization of accents, cannot be isolated in the analysis from other basic or derived variables; this limitation had to be accepted in order to contain the number of stimuli, and thus the size of the experiment overall.

Manipulating three of the four basic variables at a time, speech stimuli were synthesized for the following two conditions:

- (1) Rise Onset by Peak Height by Dip/No-Dip:
[RO] $\times$ [PH] $\times$ [DND], with PA fixed at a medial 1040 ms
- (2) Peak Timing by Peak Height by Dip/No-Dip:
[PA] $\times$ [PH] $\times$ [DND], with RO fixed at a medial 820 ms

Each condition involved $3 \times 8 \times 2 = 48$ stimuli; because of overlap, the total number of stimuli for both series combined was 80 rather than 96.

Our subjects were 8 native speakers of American English. They were naive as to the purposes of the experiment. In each of two listening sessions (lasting about 32 min each), they were presented with 4 repetitions of each stimulus (320 trials per session) in random order. Thus each stimulus was presented to them a total of 8 times.

Discourse context: Experiment I was intended to test how variation in semantic contrastiveness of a *fixed-size focus* would bring out the difference between H* and L+H*. Subjects were narratively given the following scenario, after a model due to Bolinger (1989): Amanda is a toddler who spends her days in a day-care center. When her mother picks her up one afternoon, she asks, "So, did Amanda eat anything today?" She addresses this question to one of the two caretakers, the other being within earshot. The caretaker responds by saying "Yes, she ate her apple." Upon which the second caretaker chimes in, "Amanda had a banana." Now, depending on how the second caretaker makes this statement, he might assert either that Amanda ate a banana instead of her apple—the contrastive interpretation—or that she ate a banana in addition to her apple—the noncontrastive, 'expansive' interpretation. (The use of gender here is determined by the

fact that the synthetic voice was male.) Thus one might imagine the first caretaker to react by saying either "No, she did not, she had an apple!" or "Oh really, that too?" respectively. In short, the discourse fragment presented to subjects was as follows:

Speaker A: So, did Amanda eat anything today?

Speaker B: Yes, she ate her apple.

Speaker C: Amanda had **a banana**.

where boldface denotes the stimulus utterance and underlining the presumed focus. The task we posed to our subjects, based on this discourse scenario, was to decide, for each stimulus, "whether the second caretaker implies that Amanda did not have an apple."

2.2 Analysis and Results

Figure 3 below shows the data obtained from one of the subjects; the points connected by curves indicate the number of times a given stimulus was judged 'contrastive' ("Amanda did not have an apple"), out of the total of 8 judgments. The x-axis denotes increasing peak height level in all plots. Each vertical pair of plots presents sets of stimuli that differ in [DND] only (Dip depicted below No-Dip in all cases). Condition (1) is presented by the vertical triplet of such pairs, Condition (2) by the intersecting horizontal triplet.

As can be seen informally from the diagram, which was typical, [PH] was by far the most decisive factor for this subject's judgments: regardless of other parameters varied, all curves rise in more or less similar fashion as a function of [PH]. The analysis of the data did not formally confirm and quantify this observation, but rather took the effect of PH on contrastiveness itself as the affected (dependent) variable and tested for significant effects of other, independent parameters (DND, RO, PA) on this variable.

A series of ANOVAs, pooling data across subjects, was carried out separately for Condition (1) (independent variables DND, RO) and (2) (independent variables DND, PA), using each of the following specific criteria as the dependent measure in turn.

- a. Total number of 'contrastive'-judgments for a given set of DND/RO/PA values, summing over PH levels; i.e., roughly, the area under a given curve in the graphic presentation of Figure 3.
- b. Crossover point; i.e., the PH value at which 'contrastive'-judgments reached 50%.

Figure 3

c. Slope of the curve near the crossover point; this is a measure of the sharpness of the category boundary rather than contrastiveness per se.

Measures (b) and (c) involved a slight idealization of the actual curves, in that only their central portions (the transition left-to-right between the last point denoting 0 'contrastive'-judgments and the first point denoting 8 (100%) 'contrastive'-judgments) were taken into account.

The analysis confirmed the dominant role of PH in determining contrastiveness, in that the mean crossover point across all data, corresponding to a peak height of 208 Hz (or, ignoring outliers, 212 Hz), showed a range across conditions and subjects of only 30 Hz (2 PH levels). This is a small range of variation compared to the difference in peak height found between natural renditions of minimal pairs $H^*/L+H^*$, which is 60-100 Hz.

The analysis of secondary effects due to other variables yielded little, however. For Condition (1), no significant effects of DND or RO on measures (a-c)

could be found. As for Condition (2), an analysis based on measure (a) showed a significant interaction $DND \times PA$ ($P = 0.034$): in the Dip stimuli, but not the No-Dip stimuli, the "late" peak alignment point was significantly less contrastive. The analysis based on (b) supported this result somewhat: the interaction $DND \times PA$ showed a nonsignificant trend ($P = 0.087$) for the latest peak point to have a higher crossover value (by about 12 Hz) than the earlier peak points, but only in Dip stimuli. Otherwise, no effects due to any of the variations in pitch contour, including any of the derived variables mentioned in Section 2.1., could be identified.

2.3 Discussion

Earlier, short pilot studies involving only two dimensions of variation at a time had led us to expect a trading relationship between peak height and dip depth in this experiment. The main result—that contrastiveness was consistently a function of peak height alone—was disappointingly negative in this respect. Neither the presence of a dip preceding the peak—expected to be the most salient non-height cue—nor any other of the parameters varied had a systematic effect on listeners' judgments. Thus, in retrospect, Experiment I must itself be considered a pilot study for Experiments II and III: important mainly for what it failed to show.

The one height-independent effect on contrastiveness, namely reduced contrastiveness in late peaks for Dip contours, is in fact at odds with the widespread notion that later peaks are generally more contrastive or emphatic. Given that the effect was limited to Dip contours, the only hypothesis we have to account for it is that these late peaks were no longer identified as $L+H^*$ accents but had begun to sound like L^*+H accents to the subjects—a category characterized by alignment of a Low tone with the stressed syllable, with different semantic/pragmatic properties altogether. In other words, in our attempt to explore the outer reaches of $(L+)H^*$ territory we may accidentally have strayed into the L^*+H preserve. No-Dip stimuli would not have been affected because they lack the characteristic Low interpreted as L^* here. This hypothesis is based on comparison of peak alignment points relative to vowel length in our stimuli with those studied by Kohler (1991) for German: Kohler found that peaks coinciding with the VC transition (as our late peaks did) were given a categorically different interpretation from those occurring earlier ("lack of speaker commitment" in Pierrehumbert & Steele's (1989) terms). Furthermore, Kohler & Gartenberg (1991) found that accents corresponding to our late peaks were systematically judged less prominent than accents that correspond to our early and medial peaks; so aside from categorical, interpretational differences, there might have been other perceptual mechanisms involved. In subsequent experiments, to be described below, we narrowed the range of PA values to a set of "medial" points, staying clear of the VC transition area.

The weightiest among suspected reasons why Experiment I did not yield any other significant results is potential failure of the Dip contours to be identified as such, i.e. as L+H*. That is to say, the dips may not have been long/deep enough to be sufficiently salient. If listeners perceptually integrate over up to 30 ms around pitch extrema (as 't Hart 1991 might be taken to suggest), the actually perceived pitch difference between rise onsets in Dip vs. No-Dip stimuli may never have exceeded 15-20 Hz (whereas the physical difference was 30 Hz), and that may not have been enough for a clear dip percept. The fact that both early and late dip inversion points happened to coincide with a spectral transition period might also have contributed to reduced perceptual salience (Verhoeven 1994); however, this effect cannot have been decisive because we did not see more of a dip-effect on contrastiveness for the medial dip position (falling into the steady-state portion of a vowel) than for the flanking positions. Most importantly from a conceptual viewpoint, perhaps the fixed sentence-initial part of the synthetic contours extended too far, as Bob Ladd (pers. comm.) suggested; in other words, in natural utterances, the difference between H* and L+H* is not confined to the segment of the contour beginning with the rise onset (which in our stimuli coincided with the instantaneous lowpoint of the dip) but rather extends to the segment before rise onset, for example by requiring a low plateau for L+H* rather than an instantaneous lowpoint preceded by a nondistinctive plateau. All of these possibilities were taken into account in the design of the stimuli for Experiments II and III; these revised stimuli had deeper and more extended dips (see section 3.1 below).

Other potential weaknesses of stimulus and experimental design in Experiment I, each of them individually less likely perhaps to have played a decisive role for the absence of results, include the following:

- (i) A local maximum in amplitude on the stressed syllable had not been shifted concomitantly with peak alignment (see Kohler 1991 for a systematic investigation of the relative importance of amplitude and pitch variation).
- (ii) The multidimensional variation listeners had to cope with in this experiment, compared to the "two-dimensional" pilot studies mentioned above that had yielded a simple trading relationship between peak height and dip depth, had "overwhelmed" them and rendered inoperative the relevant perceptual mechanisms used in natural discourse situations.
- (iii) The dominant role of PH variation, due to the range of values chosen vis-a-vis the range for other variables, had led listeners to adopt a less "integrative" strategy, focusing on PH alone. Note, though, that the range chosen had merely reflected the difference typically found between accent prototypes.
- (iv) The utterance-initial peak on *Amanda* may have been too prominent after all, given the discourse scenario (which clearly made the sentence subject "old infor-

mation”), and had “distracted” listeners from the intonationally variable part of the sentence.

Again, all of these points were addressed by modifications in stimulus properties and experimental design in Experiments II and III.

3 Experiments II and III

3.1 Methods

Stimuli consisted of the same synthetic sentence as in Experiment I. However, aside from slight non-pitch-related improvements in stimulus quality, both the constant and the variable part of the pitch contour were modified compared to the earlier study. In the sentence-initial constant part, which was now limited to the first 460 ms, the peak on *Amanda* was reduced by 20 Hz, as shown in the overview diagram below (Fig. 4).

Figure 4

The variables affecting the shape of the nuclear accent and the preceding segment of the contour leading up to it were now manipulated as follows to create a new set of stimuli (see Fig. 5):

Figure 5

1. Peak Height (PH): 3 levels: 204, 212, and 220 Hz (i.e., the total range of 16 Hz now corresponds to what used to be the difference between any two of the eight levels used in Experiment I)

2. Dip Depth (DD): 6 levels, from 110 Hz (strongest dip, corresponding to utterance-final baseline value) to 150 Hz (no dip); the dip is reached via interpolation from the 150-Hz endpoint of the fixed contour (at 460 ms now) to a point 20 ms before rise onset, itself a variable (cf. below), which created a comparatively lower inter-peak segment vis-a-vis the stimuli of Experiment I, and a brief low plateau rather than an instantaneous lowpoint for the dip itself.
3. Timing of Rise Onset (RO): 6 points, from 780 to 880 ms in 20-ms steps (narrowed overall range compared to Experiment I, in keeping with modified range of peak alignment values as given below; the new values still span thesegments /b/, /a/, /n/ of the initial syllable of *banana*); the time point of rise onset always denotes the end of the 20-ms dip plateau, the beginning of which as well is thus solely a function of RO (cf. (2) above)
4. Timing of Peak (PA): 6 points, from 980 to 1080 ms in 20-ms steps (narrowed overall range compared to Experiment I, eliminating points close to the offset of the 210-ms-long stressed vowel; values now range from 70 to 170 ms relative to vowel onset)

The secondary parameters that can be derived from these basic ones are the same as for Experiment I. Again, parameter values were selected based on available pitch track corpora (especially the ToBI corpus; Beckman & Ayers 1994) and preliminary studies assessing naturalness and comparable perceptual distinctiveness for each dimension of variation, as well as on the considerations laid out in section 2.3 above. The most important change compared to Experiment I was that peak height variation was constrained now to a very narrow window of midrange values (16 Hz) that included the mean crossover point identified in Experiment I; the implication was that whatever effect might be identified for other cues in Experiments II and III would have to be considered secondary to the dominant effect of peak height already established.

Each of the two experiments consisted of nine conditions, each of which involved random presentation to listeners of 8 repetitions of a series of 18 stimuli that varied independently along only two of the dimensions above. The conditions used were the following:

- (1) Peak Height by Dip Depth:
[PH] × [DD], at fixed RO (820 ms) and PA (1020 ms)

- (2a) Peak Height by Peak Timing at early rise:
[PH] × [PA], at fixed DD (134 Hz) and RO (780 ms)
- (2b) Peak Height by Peak Timing at late rise:
[PH] × [PA], at fixed DD (134 Hz) and RO (880 ms)
- (3a) Peak Height by Rise Onset at early peak:
[PH] × [RO], at fixed DD (134 Hz) and PA (980 ms)
- (3b) Peak Height by Rise Onset at late peak:
[PH] × [RO], at fixed DD (134 Hz) and PA (1080 ms)
- (4) Peak Height by whole-contour shift:
[PH] × [RO]&[PA], at fixed DD (134 Hz)
- (5) Peak Timing by Dip Depth:
[PA] × [DD], at fixed RO (820 ms) and PH (212 Hz)
- (6) Rise Onset by Peak Timing:
[RO] × [PA], at fixed DD (134 Hz) and PH (212 Hz)
- (7) Whole-contour shift by Dip Depth:
[RO]&[PA] × [DD], at fixed PH (212 Hz)

Taken together, these conditions in principle allow assessment of the relative contribution to contrastiveness of PH (manipulated in (1-4) but not (5-7)), DD (manipulated in (1,5,7) only), RO (manipulated in (3a,b,4,6,7)), PA (manipulated in (2a,b,4,5,6,7)), and slope of rise (manipulated in all conditions, but constant for a given PH in (4), a given RO/PA interval in (6), and a given DD value in (7), regardless of variation in other dimensions). This paradigm of conditions is not complete, of course, with respect to possible combinations of dimensions, but concentrates on the most likely candidates for contrastiveness cues in order to contain the overall size of the experiment.

Our subjects were 8 native speakers of American English, newly recruited for these revised experiments, who again were naive as to the purpose. During each listening session, they were presented with two conditions chosen in semi-random order (the order varied among subgroups of subjects), consisting of 144 trials each. All subjects completed Experiment II before moving on to Experiment III.

Discourse context: Experiment II was based on the same task as Experiment I; subjects were given the exact same scenario, testing the effect of pitch contour variations on the relative contrastiveness of a *fixed-size focus* (cf. section 2.1 above). Experiment III, by contrast, was intended to assess the effect on a *variably-sized focus*. Subjects were given the following scenario in Experiment III: A

little older now, Amanda is in the hospital, where she has had her appendix removed two days ago. The doctor arrives and asks the nurse on duty to update him on his patients. He might ask the nurse one of two questions: either (i), "So, what's Amanda been up to today?" (where any statement predicating a convalescent activity on Amanda would be a felicitous answer), or (ii) "So, did Amanda have any fruit today?" (where the answer is expected to be of the form 'Amanda had fruit X today', or, of course, "No"). The nurse responds with the stimulus utterance. Thus, there are two alternative discourse fragments to be considered:

(i) Q: So, what's Amanda been up to today?

A: Amanda **had a banana**.

(ii) Q: So, did Amanda have any fruit today?

A: Amanda **had a banana**.

where boldface denotes the stimulus utterance and underlining the presumed focus. The task posed to subjects in this experiment was to decide, for each stimulus, "whether it was more likely to have been the nurse's answer to question (i) or (ii)." Here, (i) represents the broad-focus, less contrastive interpretation and (ii) a narrow-focus, more contrastive interpretation.

3.2 Analysis and Results

Initial inspection of the data showed that although the revised stimuli succeeded in making the effects of PH more comparable to those of other factors, there was once again considerable variation among subjects as to the effects of different conditions on contrastiveness judgments.

For example, Figure 6 depicts the results of one condition (Cond. 4; that is, [PH] × [PA]&[RO]) of Experiment II plotted for three subjects (represented by the three rows of diagrams). The three columns denote the three levels of peak height, increasing from left to right. Each plot gives the number of 'contrastive'-judgments (out of a total of eight for each stimulus) versus the six points of rise onset and peak alignment shifted in tandem for this condition (from early to late, left to right). Clearly, increasing peak height increases contrastiveness for all subjects. However, rise-onset/peak shift tendentially increases contrastiveness for HR, decreases it for KC, and does not produce any clear trend for MG.

In light of this diversity, which was found throughout the data and replicated the absence of trends across subjects found in Experiment I, a different statistical technique was chosen for analysis this time.

For each subject and each condition, correlation coefficients were calculated between the values of each of the two variables manipulated and the fraction

Figure 6

of stimulus tokens judged to be contrastive (option (ii) in case of Experiment III). As it turned out, there was little difference between the results of the two experiments, so data were pooled for them. (we will return to this point below.) Each subject's performance was characterized by counting the instances for each cue (that is, each dimension of variation) across applicable conditions, where correlations exceeded threshold values of either ± 0.4 (medium correlation) or ± 0.6 (strong correlation). In this step instances of opposing sign were subtracted, so a low score may indicate either insignificance or variable interpretation. Cues were then rated as follows:

WEAK cues: Cues for which the fraction of correlations exceeding ± 0.4 constitutes 50% or more of all instances

STRONG cues: Cues for which the fraction of correlations exceeding ± 0.6 constitutes 67% or more of all instances

The results are shown in Figures 7-9. Figure 7 represents the fraction of conditions with correlations exceeding ± 0.4 for each cue and each subject.

Figure 7

The picture turns out to be rather mixed, but one can say that subjects fall into two groups: peak height listeners, who, despite the very narrow range of peak height variation, base their judgment primarily on the height cue (HR, KC, MG, EB); and a second group of listeners, who attend to other cues as well (SD, JW, KH, LW).

The next two plots bring out the different strategies more strongly by suppressing lesser cues. Figure 8 shows all cues that qualified at least as WEAK cues, and Figure 9 shows only STRONG cues. As the comparison demonstrates, the diversity of strategies, peak height aside, remains a fact no matter what threshold is imposed.

However, going back to Figure 7, the diagram that offered the most detailed picture, the following generalizations can be drawn:

1. Dip Depth (DD), the acoustic property presumably corresponding most immediately to the Low-tone of the bitonal accent L+H*, is indeed the most widely used secondary cue (STRONG for one, WEAK for four subjects), though it does not show consistently stronger correlations with the judgment data than other cues.

Figure 8

Figure 9

2. Peak Timing (PA) is another prominent secondary cue. In line with the trend observed in Experiment I, peak timing is negatively correlated with contrastiveness here (for seven of the eight subjects). An early peak contributes more to contrastiveness than a late peak.
3. Rise Onset time (RO) shows a minor positive correlation with contrastiveness (for six subjects), but does not qualify as even a WEAK cue for any. Similarly, whole-contour shift (RO&PA) is only a STRONG cue for subjects who also use PA as a STRONG cue, in the same direction.

The first two of these generalizations, that is, the reliance of listeners on greater dip depth and earlier peak timing as secondary cues, might be taken to point to steepness of slope as the decisive criterion for accent classification. (The primary cue Peak Height would of course collaborate here.) However, a slope account is put in question by the fact that rise onset time did not have as strong an effect as peak arrival time. That is to say, if PA variation served as a cue only indirectly, via manipulation of the slope preceding the peak, any other manipulation effecting this slope in a similar way—specifically, RO manipulation—should show an equally strong correlation with contrastiveness. This was not found to be the case here: although the correlation between RO and contrastiveness goes in the right direction, in the sense that RO values producing steeper slopes are judged more contrastive, the correlation is much weaker than that between PA and contrastiveness.

Alternatively, subjects might be sensitive to differences in excursion size between lowpoint and peak. Obviously both Peak Height and Dip Depth would contribute to this value. An early accent peak might also contribute to perceived excursion size, due to a perceptual mechanism that shifts judgments toward the initial frequency if a transition occurs during the early part of a vowel (Nabelek et al. 1970). Regarding peak timing, it is also conceivable that, independent of strategies involving other cues, an early accent peak might make the subsequent fall more salient; it has been suggested (Couper-Kuhlen 1984) and observed experimentally (Eady & Cooper 1986) that the slope of this fall correlates with contrastiveness. Because the accent was on the penultimate syllable in our stimuli, we could not independently assess the role of such post-peak properties in the present design, which concentrated on properties implicated by Pierrehumbert's accent categorization.

It is worth drawing attention to the fact that our subjects, especially the group of non-peak-height listeners (SD, JW, KH, LW), showed considerable variation in the effect a given cue had on contrastiveness judgments (see Fig. 7 above); note, for instance, the exceptional way in which KH uses alignment cues. Any explanations of these seemingly idiosyncratic aspects of subject behavior must remain speculative for now. (On one view, they may merely constitute evidence for strategic listening subjects would not have employed in natural speech situations;

cf. section 3.3 below.) However, it is known, for example, that the relation between pitch movement delay and prominence is not universal for listeners. If KH, who judged delayed peaks to be more rather than less contrastive, attended more to the rising pitch *movement* in late peaks than to the peak height eventually reached, this might have created a percept of vowel lengthening, a demonstrable effect (see e.g. Lehiste 1976; but cf. van Dommelen 1994 for conflicting experimental results). Perceived lengthening, in turn, would likely have been interpreted as increased prominence.

To summarize the above results, the present data provide only weak support for a categorical contrast between L+H* and H* pitch accents, as a means of encoding contrastive vs. noncontrastive interpretations, that is based on the distinction between the two implied by their labels in Pierrehumbert's model. As already observed in Experiment I, peak height, rather than the cues that would convey the presence of the L tone (dip depth) or bitonality (peak timing combined with rise onset timing), was the most categorical property.

As mentioned, the results of Experiments II and III were pooled in the above analysis because there were no marked differences between them in the use of the phonetic properties manipulated here in conveying differences in contrastiveness for fixed focus versus variable focus. The comparison only showed a slight increase in the number of properties attended to, and a slight decrease in reliance on peak height. Both are likely to be learning effects and evidence of strategic listening (cf. section 3.3 below).

Figure 10

Figure 11

However, analysis of the correlation between contrastiveness judgments in the two experiments revealed considerable heterogeneity in the use of cues both across conditions and across subjects (see Figs. 10 and 11, respectively). The one generalization to make is that conditions involving peak height manipulation produced more consistent performance than those that did not. This observation is consistent with the result that peak height is the dominant cue to contrastiveness.

3.3 Discussion

The heterogeneity of subject behavior indicated by the comparative analysis of Experiments II and III suggests that the use of secondary cues may at least partly involve ad hoc schemes developed by subjects in a given session, rather than stable perceptual strategies. There may be a number of explanations for this. First, it is of course possible that despite comments to the contrary, subjects weren't able to adjust to the experimental setting and the somewhat unnatural, repetitive task, and responded by resorting to ad hoc strategies. More seriously, it is possible that the discourse scenarios used here failed to bring out the crucial interpretational distinction corresponding to the accent categories $L+H^*$ and H^* . For instance, one might argue that in the fixed focus, daycare center scenario, the second caretaker's most natural intonation would need to be contrastive either way: on the banana-not-apple assertion because it is contrastive as defined, signaling denial of a factual alternative, and on the apple-plus-banana scenario as a means to identify the part

of the sentence that “contrasts” with the otherwise identical preceding sentence—a discourse situation normally avoided by varying the wording. (In a sense, there is an alternative being denied in the second scenario as well, namely a speech act.) If contrastiveness is a binary feature independent of focus, we might not have seen a difference between experiments in this case because both experiments failed to set up the needed opposition. Subjects might have gone by cues for a difference in general assertiveness, emphasis, or speaker involvement—that is to say, primarily peak height—and at times have employed ad hoc strategies vis-a-vis other stimulus variations.

At the same time, there is always the possibility that the direct-mapping-hypothesis linking phonological to semantic/pragmatic primitives, which all such considerations are based on, may not hold, at least not for the accents under consideration. Undoubtedly, more experimental work should be done to follow up on the results reported here, which merely constitute negative evidence on this view: perception studies involving a greater number of stimulus utterances and discourse scenarios, and production studies systematically testing what dimensions of acoustic variation are actually employed by subjects producing well-defined semantic contrastiveness differences.

But let us assume for a moment that the design of Experiments II and III was successful in bringing out the perceptual strategies actually employed by listeners to distinguish between $L+H^*$ and H^* on the basis of pitch cues. Specifically, let us assume that inconsistent behavior toward a given cue may be discarded as an attempt by subjects to make use of experimental variation they would not attend to in natural discourse, and that the only pitch cue consistently employed in categorically distinguishing between these accents is peak height, or more precisely, perceived peak height. It has been suggested by Ladd et al. (1994) that if the height of a peak exceeds a certain threshold, it seems to signal to the listener that the prominence of that peak should be judged “on its own terms” rather than relative to the pitch range of the intonational unit it is part of (cf. also Ladd 1994). In fact, one might want to say that a rapid pitch rise to a peak above this threshold systematically serves as an “attention marker” indicating a resetting of the pitch range, as might occur at the beginning of a new intonational unit (Intermediate or Intonational Phrase, in terms of Beckman & Pierrehumbert 1986). A dip, though not essential for this effect, might well cooperate with it, by suggesting a downward shift in the speaker’s pitch range that lowers listeners’ expectations for the height of the subsequent peak and renders the actual peak correspondingly more prominent. If this analysis, involving structural partitioning of the utterance in the prosodic domain, is generally on the right track, all else being equal, it might be more attractive—at least for proponents of the “direct-mapping approach”—to also think of the interpretational counterpart of what we now call $L+H^*$ in terms of a structural notion rather than a semantic feature, by taking $L+H^*$ to encode the boundary between nonprojecting, narrow focus and the sentence background on its left (mediated by the rules governing the position of word stress and thus alignment of

the starred tone); in other words, to lean toward the second of the two views on contrastiveness sketched out in section 1 above, which regards contrastiveness as a necessary property of narrow focus. (On the right the edge of the focused constituent is already conveyed by the position per se of the nuclear accent, in conjunction with the rules of focus projection; see, e.g., Selkirk, forthcoming.) Further, the placement of this prosodically marked boundary might selectively support some focus-theoretical frameworks but not others. For instance, it seems worth exploring whether this boundary always corresponds to that between the background and the focus of an overt or covert focalizer as defined in a recent formalization of the Praguian framework of Topic-Focus Articulation (see Partee, Hajicova & Sgall 1994); it does for the discourse fragments under consideration in the present study, but these involve comparatively simple focus structures that would receive a similar analysis in a variety of other frameworks as well. Should such further explorations turn out to strengthen the proposal advanced here, that L+H* (indirectly) demarcates the edge of a distinct semantic/pragmatic element, the results of this study could be seen as one of a growing number of observations suggesting the existence of prosodic edge markers for non-prosodic constituents.

Acknowledgments

This work was supported by NIH grant DC01798 to the second author. The first author has also been sponsored by the U.S.-Czech/Slovak Science and Technology Joint Fund in cooperation with the Ministry of Education of the Czech Republic under project no. 92058. We are indebted to Lisa Selkirk, Bob Ladd, and Chuck Clifton for valuable suggestions regarding experimental design and data analysis.

References

- Beckman, M., and G. Ayers. 1994. 'Guidelines for ToBI Labeling', version 2.0. Distributed together with audio recording and printed F0 tracks of the ToBI corpus by the Ohio State University, Columbus.
- Beckman, M.E. and J.B. Pierrehumbert. 1986. 'Intonational Structure in Japanese and English.' *Phonology Yearbook* 3, 255-309.
- Bolinger, D. 1989. *Intonation and Its Uses*. Stanford University Press, Stanford.
- Chafe, W. 1976. 'Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View', in C.N. Li (ed.), *Subject and Topic*, Academic Press, New York, 25-55.
- Couper-Kuhlen, E. 1984. 'A New Look at Contrastive Intonation', in J. Watts and U. Weidmann (eds.), *Modes of Interpretation: Essays Presented to Ernst Leisi*, G. Narr, Tuebingen, 137-158.
- Dommelen, W.A. van. 1994. 'Does Dynamic F0 Increase Perceived Duration? New Light on an Old Issue', *J. of Phonetics* 21:367-386.
- Eady, S.J., and W.E. Cooper. 1986. 'Speech Intonation and Focus Location in Matched Statements and Questions', *J. Acoust. Soc. Am.* 80:402-415.
- 't Hart. 1991. 'F0 Stylization in Speech: Straight Lines versus Parabolas', *J. Acoust. Soc. Am.* 90:3368-70.
- Klatt, D.H., and L.C. Klatt. 1990. 'Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers', *J. Acoust. Soc. Am.* 87:820-857.
- Kohler, K.J. 1991. 'Terminal Intonation Patterns in Single-Accent Utterances of German: Phonetics, Phonology, and Semantics', in K.J. Kohler (ed.), *Studies in German Intonation*, Arbeitsbericht (AIPUK), Institut fuer Phonetik und digitale Sprachverarbeitung, Universitaet Kiel.
- Kohler, K.J., and R. Gartenberg. 1991. 'The Perception of Accents: F0 Peak Height versus F0 Peak Position', in K.J. Kohler (ed.), *Studies in German Intonation*, Arbeitsbericht (AIPUK), Institut fuer Phonetik und digitale Sprachverarbeitung, Universitaet Kiel.
- Ladd, D.R. 1980. *The Structure of Intonational Meaning*, Indiana University Press, Bloomington.
- Ladd, D.R. 1994. 'Constraints on the Gradient Variability of Pitch Range, Or, Pitch Level 4 Lives!', in P.A. Keating (ed.), *Phonological Structure and Phonetic Form* (Papers in Laboratory Phonology 3), Cambridge University Press, Cambridge, 43-63.
- Ladd, D.R., J. Verhoeven, and K. Jacobs. 1994. 'Influence of Adjacent Pitch Accents on Each Other's Perceived Prominence: Two Contradictory Effects', *J. Phonetics* 22:87-99.
- Nabelek, I.V., A.K. Nabelek, and I.J. Hirsh. 1970. 'Pitch of Tone Bursts of Changing Frequency', *J. Acoust. Soc. Am.* 48:536-553.
- Partee, B.H., E. Hajicova, and P. Sgall. 1994. 'Topic-Focus Articulation, Tripartite Structures, and Semantic Content', unpublished manuscript, University of Massachusetts, Amherst, and Charles University, Prague.

- Pierrehumbert, J.B. 1980. *The Phonology and Phonetics of English Intonation*, Ph.D. dissertation, MIT.
- Pierrehumbert, J.B. and S.A. Steele. 1989. 'Categories of Tonal Alignment in English.' *Phonetica* 46:181-196.
- Pitrelli, J. 1994. 'Evaluating ToBI Inter-Transcriber Consistency', presentation at the 3rd ToBI Workshop, Boston University, August 1994.
- Selkirk, E. (forthcoming) 'Sentence Prosody: Intonation, Stress, and Phrasing', in J. Goldsmith (ed.), *Handbook of Phonological Theory*, Blackwell, Oxford.
- Sgall, P., E. Hajicova, and J. Panevova. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Reidel, Dordrecht.
- Taglicht, J. 1982. 'Intonation and the Assessment of Information', *J. Linguistics* 18:213-230.
- Verhoeven, J. 1994. 'The Discrimination of Pitch Movement Alignment in Dutch', *J. of Phonetics* 22:65-85.

Department of Linguistics
South College
University of Massachusetts
Amherst, MA 01003

bartels@linguist.umass.edu
kingston@linguist.umass.edu