A33 reviewed earlier

A33

5

"Gesture" in prosody: comments on the paper by Ladd

BRUCE HAYES

Ladd's paper is a beautiful example of how experimental work can, with serious thought, be made to bear on the abstract system of phonetic rules and representations. In the case at hand, he has shown how the Gussenhoven–Rietveld effect, initially just a minor data puzzle, bears importantly on a much larger issue, that of how pitch is scaled by the system of phonetic rules.

5.1 Ladd's arguments

Ladd takes on the claim that pitch range is subject to "Free Gradient Variability"; i.e. that over a window no larger than pitch-accent size, the speaker is free to select a local pitch range. We might characterize this as positing a "beast within"; this creature monitors our speech constantly, assessing how much it cares about what we are saying at that instant, and adjusts the pitch range of our voices accordingly.

The beast metaphor should not be dismissed as absurd. For example, Bolinger (1986, 1989) has suggested quite explicit analogies between intonation and more primitive forms of vocal expression.

Ladd subjects Free Gradient Variability to three criticisms:

1. It fails to explain the Gussenhoven–Rietveld effect. When the effect occurs, listeners must interpret increases in the pitch of the first of two peaks as increases in the prominence attached to the second peak. Under Free Gradient Variability, listeners surely would interpret increases in the pitch of the first peak more directly; i.e. as increases in the prominence of the first peak.

   As a corollary, Ladd notes that the *overriding* of the Gussenhoven–Rietveld effect when the second peak is especially high implies the existence

of a H+* accent, i.e. an Overhigh which is interpreted by listeners as especially prominent no matter what the context.

2. Ladd and others (4.2.1) have found quite subtle hierarchical effects relating the pitch levels of H* accents to syntactic and discourse grouping. It is hard to see how such effects could be discerned by listeners if the beast were constantly adjusting pitch up and down at the same time.

3. If pitch range is freely variable, it is unlikely that pitch-scaling experiments (e.g. the "Anna came with Manny" experiment of Pierrehumbert 1980, Liberman and Pierrehumbert 1984) could produce such beautifully clean mathematical relations between pitch targets. Presumably, the random variations of Free Gradient Variability would overwhelm such patterns in noise.

Arguments (1) and (2) seem particularly compelling; we return to (3) later. I also agree with Ladd's point concerning research strategies: we are better off doing without Free Gradient Variability unless it is firmly shown to be necessary. A theory lacking it makes more precise predictions, and forswearing Free Gradient Variability serves as inducement to explain apparently "random" variation in the height of pitch peaks.

5.2 Objections to H+

My disagreement with Ladd centers on his proposal that English intonation should involve a *phonological* category H+, found in the "Overhigh" pitch accent H+*. I think this proposal has a number of drawbacks, to be outlined below. Moreover, I will suggest later that there is a plausible alternative to H+ that retains Ladd's insights while permitting a simpler phonological system.

5.2.1 Phonemic opposition or continuum?

To begin, it seems likely that in setting up the category H+, Ladd is phonemicizing a continuum. Consider figure 5.1, which shows pitch tracks of myself pronouncing "The melon was yellow", with four different degrees of special emphasis on *yellow*. It is possible that there is an identifiable line on this continuum, but this is certainly not proven at this stage. Note that the pitch of *melon* does not vary much, suggesting that what was being varied was not the *overall* pitch range of the utterance.

The problem becomes worse if we consider *monosyllables* pronounced with the putative H+* versus H*: here, it becomes even harder to imagine

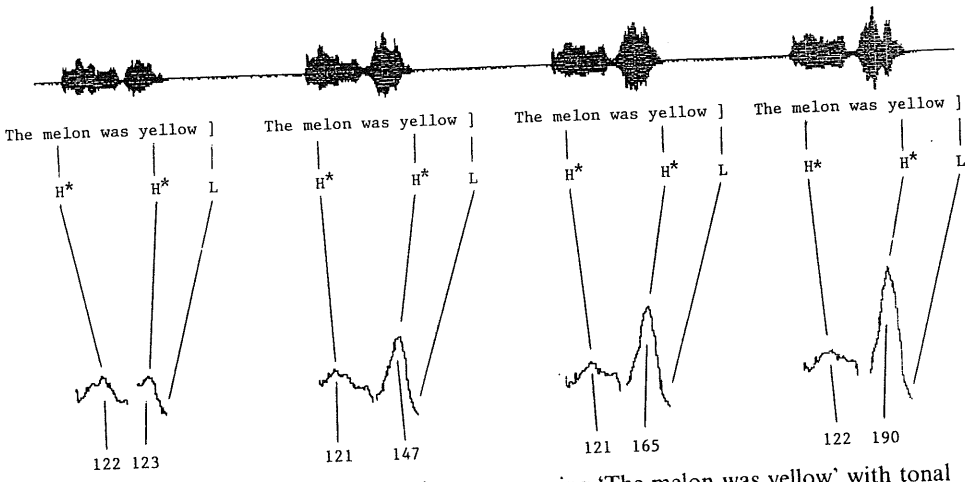


Figure 5.1. Pitch track of a single speaker pronouncing 'The melon was yellow' with tonal sequence shown and four different degrees of special emphasis on *yellow*.

that a phonological opposition is present; cf. the continuum of pronunciations presented by Liberman and Pierrehumbert (1984: 159).

True phonological contrasts can be supported by their effect on the perceptions of native speakers: phonetic differences tend to be far more perceptible when they cross phonemic boundaries (Werker and Tees 1984, and references cited there). In fact, experiments of the appropriate type have been carried out for intonation, with positive results (Pierrehumbert and Steele 1987; Kohler 1987). I believe the existence of H+ could be best defended with perceptual evidence of this sort.

5.2.2 *The mix-and-match problem*

H+ amplifies what I will call the *mix-and-match* problem: if we set up a number of basic primitives (pitch accents, boundary tones) as the elements of the intonational system, to what extent can they be freely combined in actually occurring intonations? A striking aspect of the two-height system of Pierrehumbert (1980) is that *every* logically possible tonal sequence is a possible intonation, at least as far as nuclei (in the sense of Ladd 1980) are concerned. My estimate is that if we add H+ to the basic inventory of H and L, the logical possibilities will be less fully instantiated. (This is a point made in a similar connection by Beckman and Pierrehumbert 1986.)

For example, I believe there is no H+ analogue of the so-called "calling contour" of Ladd (1978). Such an analogue would resemble (1a), but with a higher starting point.

(1)

a. Johnny!
H* !H = [dʒɑːː ni:::]

b. *Johnny!
H+* !H = *[dʒɑːː ni:::]

(In these examples and below, downstep is transcribed !H for convenience; no claims are intended concerning the controversy over how downstep is to be represented; cf. Ladd 1983, Beckman and Pierrehumbert 1986).

It also appears that H+ does not occur as a boundary tone. The boundary tones of English are limited to the binary contrast of L versus H, and variation in the pitch scaling is determined by pragmatics (Pierrehumbert and Hirschberg 1990: 279).[1]

In general, it appears that adopting H+ would force us to add to the grammar various rules that sharply limit its distribution. This would wipe out much of the progress that has been made on the mix-and-match problem.

5.2.3 *Undergeneralization*

A third problem for H+ is *undergeneralization*. Reverting briefly to the beast metaphor, it seems that the beast makes his views clear by exaggerating the prosodic contrasts that are already present. Consider, for instance, the natural way of emphasizing the word *yellow* when *The melon was yellow* is pronounced as a question. Figure 5.2 shows four tokens from my own speech; tonal representations are schematic. Note that: (a) the pitch of the L* accent doesn't vary much. This accords with previous observations (Liberman and Pierrehumbert 1984; 218–219). (b) The H boundary tone varies greatly with degree of emphasis.

The point is that H+ will not solve the problem of expressing extra emphasis in this intonation, since there is no high accent in nuclear position. The tone that gets shifted up is a boundary tone. But the semantics of final boundary tones (Pierrehumbert and Hirschberg 1990) are characteristically not affiliated with individual words, but with whole phrases. In essence, the H boundary tones in figure 5.2 are raised "accidentally," to accommodate emphasis on the word *yellow*.

When we turn to downstepped contours, illustrated in figure 5.3, the situation looks worse. It appears that a speaker cannot vary the height

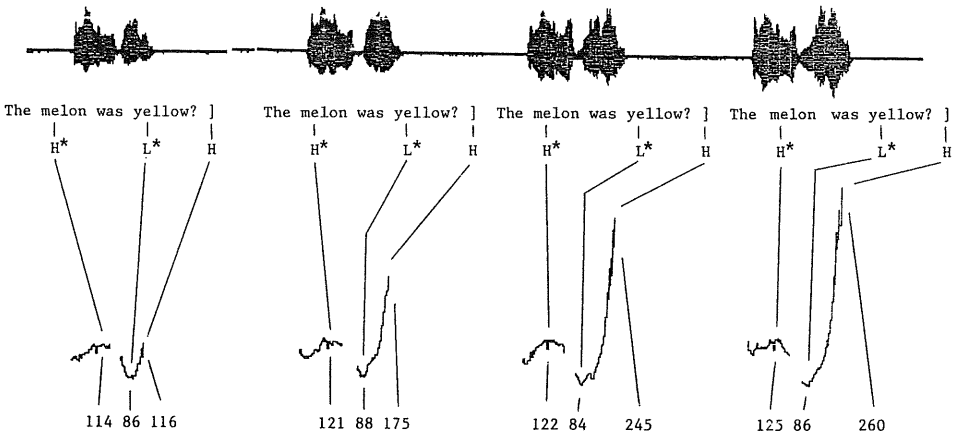


Figure 5.2. Pitch track of a single speaker pronouncing 'The melon was yellow' with tonal sequence shown and four different degrees of emphasis on *yellow*.

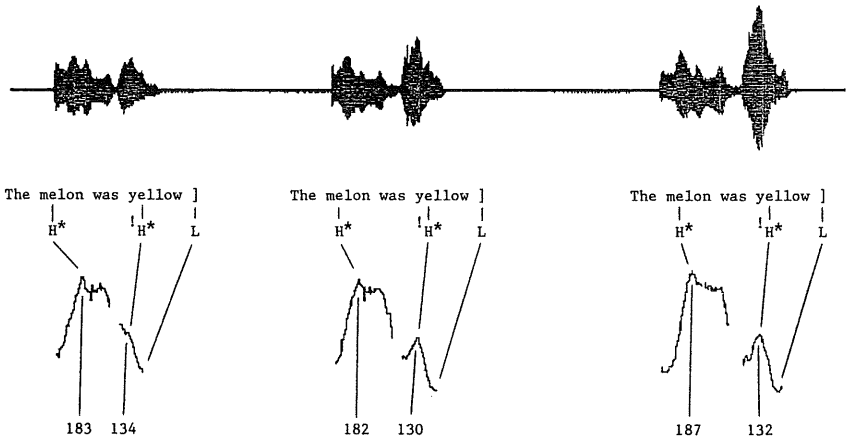


Figure 5.3. Pitch track of a single speaker pronouncing 'The melon was yellow' with tonal sequence shown and three different degrees of special emphasis on *yellow*.

of $^{!}H^{*}$ very much, given the existence in the system of both higher (H^{*}) and lower (L^{*}) targets, a point made by Ladd (1990a: 39). Therefore, the beast can only adjust amplitude and perhaps duration to mark special emphasis. This is probably why it is harder to produce as many degrees of emphasis in this contour, and I only produced three examples here instead of four.[2]

In general, it seems that positing an H+* accent only addresses part of the problem of how individual words or phrases are emphasized.

To summarize, the H+ hypothesis suffers from the lack of evidence for a true phonological contrast, from loss of system symmetry, and from insufficient generality as an account of emphasis.

5.3 A paralinguistic alternative

What is needed is an analysis that preserves Ladd's insights about the Gussenhoven–Rietveld effect without causing problems with the phonology. Such an account is possible if we consider more seriously the nature of the "beast" alluded to above. Suppose that systematic behavior in pronunciation can be the result of either regularities in the linguistic system or regularities in what I will call the *gestural system*. Within the linguistic system I include the word sequence and its phonemic, syntactic and semantic structure, the stress pattern, phonological phrasing, and the intonational tune. Within the gestural system I would include the communicative elements that accompany spoken language: gesture as conventionally construed, body movements not ordinarily considered to be gesture, facial configuration, gaze, and so on.

Suppose also that certain *vocal* elements are part of the gestural system. Crucially, these include pitch range, though I would also include expressive adjustments of duration, amplitude and voice quality. This basic distinction between language and paralanguage is a long-standing one; see for instance Ladd's quotation (this volume, pp. 43–44) from Trager and Smith (1951), or Stockwell *et al.* (1956).

Kendon (1975, 1980) has made intensive studies of gesture in the more conventional sense of body movement. His method was to record conversations on motion pictures along with a sound track. The film is then examined one frame at a time, with all body motions, down to the smallest finger movements, recorded according to the frames during which they begin and end.

An article resulting from this work that is of great interest to intonational phonologists is Kendon (1972). Here, Kendon focused on the alignment in time between body movements and the linguistic structure of the speakers' utterances. Although Kendon's results include various complexities, the overall picture can be summarized as follows. Body movements during speech are not aligned haphazardly, but coincide with crucial landmarks in the linguistic signal. In particular, they can be aligned with (a) stressed syllables (typically the ending point of the movement falls on or just before the stress); (b) the boundaries of linguistic units. In addition, a phenomenon we might refer to as "spreading" is widely found: a body part will take on a particular configuration, and change it synchronously with a linguistic boundary.

Students of modern intonational theory (as developed by Liberman 1975; Bruce 1977; Pierrehumbert 1980; Ladd 1983, and others) should find this a strikingly familiar pattern. Intonational tones are characteristically divided into pitch accents, which align with stressed syllables, and boundary tones,

which align with the edges of linguistic units. Moreover, the spreading of boundary tones has also been observed. Even the loose "semantics" of body gestures is reminiscent of the semantics of intonational tones, as comparison of Kendon's paper with Pierrehumbert and Hirschenberg (1990) shows.

Kendon (1975) provides additional evidence for the alignment of body gestures with linguistic structure: listeners can synchronize their gestures without seeing each other, provided they are both listening to the same speaker.

The upshot is this: the gestural "beast" is more sophisticated than we might have thought, in that it *knows the grammar*. Kendon's work suggests that the boundary between language and paralanguage can sometimes be startlingly thin.

Now, if we extend the notion of gesture to include vocal gestures like pitch-range expansion, another point emerges: the beast characteristically respects grammatical contrasts. For example, in Finnish, with phonemic vowel length, gestural lengthening of short vowels is avoided (Prince 1980, citing L. Carlson). Similarly, the absence of pitch-range expansion during the production of downstepped pitch accents (figure 5.3 above) plausibly reflects the need to preserve the phonological contrast of L^* versus $^!H^*$ versus H^*.

In light of the apparent linguistic sophistication of paralanguage, it seems plausible to write paralinguistic *rules*, of a rough and gradient nature, which can refer in their structural descriptions to linguistic information. Here is a conjectured rule for the part of the gestural system controlling pitch range:

(2) *Vocal Emphasis*

To emphasize a constituent, exaggerate (to the desired degree) the phonetic correlates of stress within that constituent. Conditions:

a. Do not override the stress contour of the utterance.
b. Do not override the $H^* \sim {}^!H^*$ contrast.

By "phonetic correlates of stress" I mean anything available, including amplitude, duration, and pitch range. Increases in pitch range will make themselves felt not just on pitch accents, but on boundary tones, as in figure 5.2.

It should be pointed out that (2) is a gradient rule; one can provide as little extra emphasis to a word as one likes. Hence exaggerated H^* is not a different category from H^*; i.e. there is no phonemic contrast. It is in this sense that our proposal is not the same thing as introducing a H+ tone.

Provision (2a) is the crucial part of the hypothesis. Following Chomsky and Halle (1968), Selkirk (1984), and others, I assume a set of rules which assigns a stress contour to an utterance, based on focus, old versus new

information, linear order, and constituency. The output of these rules, i.e. a set of relative-prominence relations among syllables, is assumed to be one of the phonological configurations that is treated as inviolate by the gestural system. The upshot is that only words bearing the main, nuclear stress are eligible for gestural enhancement, since enhancement elsewhere would perturb the realization of stress.

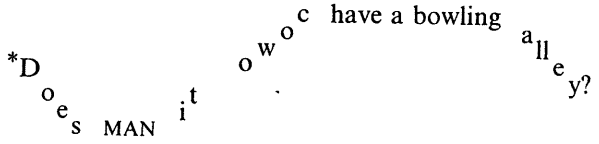
Within his system, Ladd says basically the same thing (4.3.4): "Gradient pitch-range variability can be a property of an individual accent only when the accent is both (a) nuclear, and (b) overhigh [= H+]." We modify this statement in two ways: the hypothesis doesn't need phonemic H+ in order to work, and the limitation to *nuclear* accents is a natural consequence of a more general principle, namely that gestural vocalizations tend not to override phonological distinctions. With these changes, Ladd's scenario for interpreting the experiments (4.3.2) still goes through.

5.4 Against going overboard: Is all intonation gestural?

A potentially disturbing aspect of our effort to write rules for gesture is a blurring of the distinction between language and paralanguage: is perhaps *all* of intonation gestural?

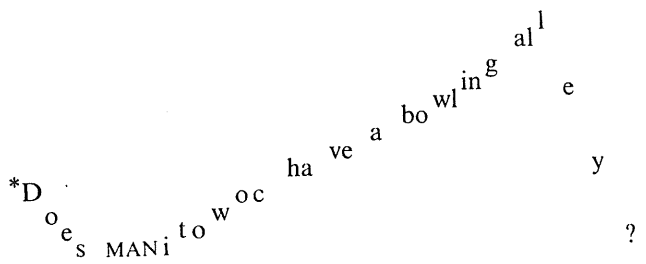
While this issue does not seem settled, I presently believe that the language/paralanguage boundary is real, and that intonation falls on the side of language. Intonation systems are amenable to phonemic analysis using a small number of contrasting categories. It seems unlikely that this is true of gesture, where the number of entities is enormous, and the notion of contrasting categories seems less useful. Moreover, intonation systems incorporate a clear and fairly rigid criterion of well-formedness. For example, (3) (modeled after Pierrehumbert 1980: 2.36C) strikes me as a clearly ill-formed English intonation:

(3)



Such well-formedness judgments are language-specific, as Ladd (1990b) has argued. Thus (4) is bizarre in English, but fine in Bengali, where this shape is the normal one for yes/no questions (Hayes and Lahiri 1991).

(4)



While gestural systems may be culture-specific, it seems unlikely that the notion of well-formedness can be defined with such clarity as it can in language (Kendon 1980: 223).

5.5 Pitch–height experiments

Ladd raises question (3) above (p. 65): how is it that pitch-scaling experiments such as Liberman and Pierrehumbert (1984) can produce such clean results? His own explanation is very simple: there is no Free Gradient Variability, so such experiments access linguistically defined pitch-scaling relations, unperturbed by local emphasis. In the view presented here, the answer cannot be so simple: we have proposed to exclude gestural emphasis on nonnuclear accents, but since nuclear accents can vary freely, and are included in most pitch-scaling experiments, we still need an explanation.

The suggestion here is again based on attributing some sophistication to the gestural beast: on request, people can control gestural behavior precisely. For example, in repeating the sentence *Anna came with Manny* several dozen times, they have no reason to give varying emphasis to the nuclear accent on different occasions; and unless they do so as a way of combatting boredom, they will not. What emerges directly reflects the phonological scaling.

There is anecdotal evidence that at least some speakers possess exquisite control over their vocal gestural patterns. This comes from Liberman and Pierrehumbert's (1984) important research on the pitch-scaling system of English. Given a particular pitch-scaling model, Liberman and Pierrehumbert's experiments might be thought of as a means of searching for the "fundamental constants" of English intonation, much as constants are determined by experiment in the physical sciences. The intonational constants investigated were: (a) the downstep constant, which scales a downstepped tone with respect to the preceding tone; (b) the final lowering coefficient, which lowers pitch range at the end of a phrase; and (c) the answer/background ratio, which scales focused new information.

The values for the constants that were found differed from subject to subject (see example 5).

(5)

	MYL	JBP	DWS	KXG
Downstep constant	**0.59**	**0.62**	0.68	(not measured)
Final lowering coefficient	**0.68**	**0.68**	0.77	0.59
Answer–background ratio	**1.66**	**1.63**	1.33	1.59

What is striking is the close resemblance between MYL and JBP (the authors of the study) versus their divergence from DWS and KXG. I bring this out not to claim any problem with the results; rather, I believe that the "constants" are in fact under gestural control. For example, the variable nature of final lowering, and its use to convey discourse structure, is discussed by Hirschberg and Pierrehumbert (1986) and Pierrehumbert and Hirschberg (1990). The answer/background ratio is probably determined by the value employed in the Vocal-Emphasis rule (2).

If these "constants" vary gesturally, then subjects must choose arbitrarily which values to use in an experiment. It should not surprise us if two authors, accustomed to enacting intonation contours in each other's presence, should have arrived at a tacit agreement on the values to use in making the recordings. The Liberman/Pierrehumbert results are well supported by their *qualitative* agreement with the data; we should not expect the quantitative values to agree, since the subjects choose them.

This example suggests an answer to Ladd's original question: in the arbitrary context of an experiment, speakers can exercise close control over their gestural systems, yielding clean experimental results.

5.6 Conclusion

In the study of phonological rules, paralanguage can usually be ignored, as it impinges little on the categorial entities with which phonologists deal. But in the study of phonetic rules, paralanguage often interacts closely with the linguistic system. In these comments, I have argued for dealing explicitly with paralanguage, as a rule system distinct from but closely connected to the linguistic system. A potential benefit is that our conception of the linguistic system can be more tightly constrained. In particular, Ladd's striking interpretation of the Gussenhoven–Rietveld effect is more persuasive, I think, when recast in a theory that distinguishes linguistic and paralinguistic rules.

Notes

1 There is one other case: Ladd's footnote 4 proposes the existence of H+ as a nonnuclear pitch accent. For discussion and an alternative account, see Hayes (1992).

2 The downstepped nuclear accents in the more emphatic versions of figure 5.3 have a peak shape, rather than the smooth shoulder of the least emphatic version. The pitch rises in the peaks are not especially audible, and I believe the peaks are in fact the acoustic result of heightened subglottal pressure on the nuclear-stressed syllable, rather than being phonologically specified.

References

Beckman, M. and J. Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3: 255–309.

Bolinger, D.L. 1986. *Intonation and its Parts*. Stanford, CA: Stanford University Press.

1989. *Intonation and its Uses*. Stanford, CA: Stanford University Press.

Bruce, G. 1977. *Swedish Word Accents in Sentence Perspective*. Lund: Gleerup.

Chomsky, N. and M. Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Hayes, B. 1992. "Gesture" in prosody. Ms. UCLA, Los Angeles, CA. (Distributed, together wtih Ladd's chapter, as an Occasional Paper of the Linguistics Department, University of Edinburgh.)

Hayes, B. and A. Lahiri. 1991. Bengali intonational phonology. *Natural Language and Linguistic Theory* 9: 47–99.

Hirschberg, J. and J. Pierrehumbert. 1986. The intonational structuring of discourse. In *Proceedings of the Twenty-fourth Annual Meeting*. New York: Association for Computational Linguistics.

Kendon, A. 1972. Some relationships between body motion and speech: An analysis of an example. In Aron Siegman and Benjamin Pope (eds.) *Studies in Dyadic Communication*. New York: Pergamon Press.

1975. *Studies in the Behavior of Social Interaction*. Bloomington: Indiana University and Lisse: Peter de Ridder Press.

1980. Gesticulation and speech: Two aspects of the process of utterance. In M.R. Key (ed.) *The Relationship of Verbal and Nonverbal Communication*, 207–227.

Kohler, K. 1987. Categorical pitch perception. *Proceedings of the Eleventh International Congress of Phonetic Sciences*, 5: 331–333. Tallinn: Academy of Sciences of the Estonian SSR.

Ladd, D.R. 1978. Stylized intonation. *Language* 54: 517–540.

1980. *The Structure of Intonational Meaning: Evidence from English*. Bloomington: Indiana University Press.

1983. Phonological features of intonational peaks. *Language* 59: 721–759.

1990a. Metrical representation of pitch register. In J. Kingston and M. Beckman (eds.) *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge: University Press, 35–57.

1990b. Intonation: emotion vs. grammar. *Language* 66: 806–816.

Liberman, M. 1975. The intonational system of English. Ph.D. dissertation, MIT.

Liberman, M. and J. Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle (eds.) *Language Sound*

Structure: Studies Presented to Morris Halle by his Teacher and Students. Cambridge, MA: MIT Press, 157–233.

Pierrehumbert, J. 1980. The phonology and phonetics of English intonation. Ph.D. dissertation, MIT.

Pierrehumbert, J. and J. Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan and M. Pollock (eds.) *Intentions in Communication.* Cambridge, MA: MIT Press, 271–312.

Pierrehumbert, J. and S. Steele. 1987. How many rise–fall–rise contours? *Proceedings of the Eleventh International Congress of Phonetic Sciences*, 3: 145–148. Tallinn: Academy of Sciences of the Estonian SSR.

Prince, A. 1980. A metrical theory for Estonian quantity. *Linguistic Inquiry* 11: 511–562.

Selkirk, E.O. 1984. *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge, MA: MIT Press.

Stockwell, R.P., J.D. Bowen and I. Silva-Fuenzalida. 1956. Spanish juncture and intonation. *Language* 32: 641–645.

Trager, G.L. and H.L. Smith. 1951. *An Outline of English Structure.* Norman, OK: Battenburg Press.

Werker, J. and R.C. Tees. 1984. Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America* 75: 1866–1878.