JANET PIERREHUMBERT

TONAL ELEMENTS AND THEIR ALIGNMENT

1. INTRODUCTION

In English, many different melodies are possible on any given word or phrase. Even a monosyllabic word, such as *Anne* can be produced with many qualitatively different melodic patterns, as illustrated in Figure 1. This situation provides a contrast with languages such as Mandarin, in which the tonal pattern is an intrinsic part of the lexical representation. In English, the choice of the melody is not entailed by the choice of words, but rather functions independently to convey pragmatic information. Specifically, it conveys information about how the utterance is related to the discourse and to the mutual beliefs which interlocutors build up during the course of the

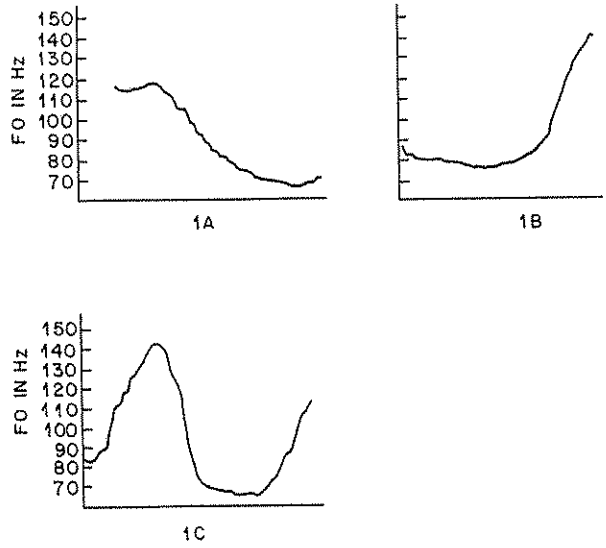

Figure 1. Three different F0 contours for the word Anne. (Reproduced from Liberman and Pierrehumbert 1984)

discourse, as discussed in Ward and Hirschberg (1985) and Pierrehumbert and Hirschberg (1990).

This situation has been recognized from the earliest work on English intonation and it has fostered attempts to phonologically abstract the melody line from the words. Like all phonological abstractions, these efforts have the

M. Horne (ed.), Prosody: Theory and Experiment, 11–36.

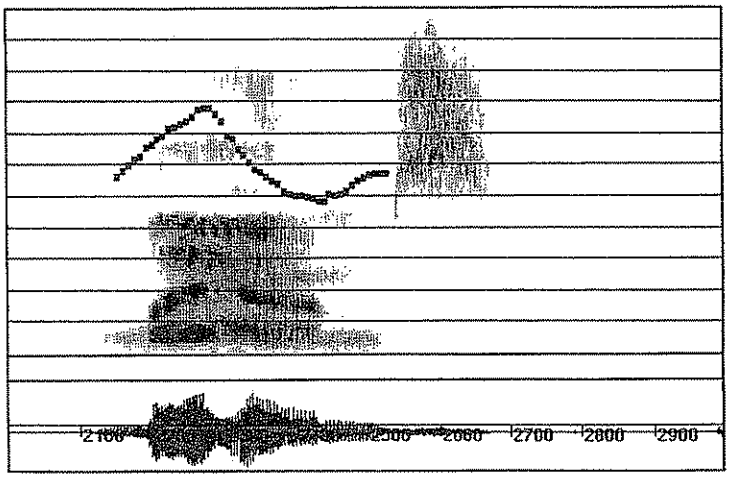

Figure 2a. A declarative pattern with a continuation rise (H L H%) on the word* limb. *The alignment of the contour with the segments can be seen by referring to the spectrogram on which the F0 contour is superimposed.*

goal of mapping out the space of contrasts in a way which treats as the same tokens which are linguistically comparable, and which treats as different tokens which contrast. The problem has not proved to be an easy one. One difficulty is illustrated in Figure 2, which displays F0 contours for a single pattern (a declarative pattern with continuation) on three different words, *limb*, *limo*, *limousine*. Each of the words was produced as the medial element in a list of three items. Although the general character of the pattern is obviously the same (a rise, a fall, and then a smaller rise) the equivalence is not captured in a syllable-by-syllable transcription of F0 levels or changes. The entire pattern is expressed on the first and only syllable of *limb*, with the result that the F0 peak is only part way into the vowel. For *limo*, the first syllable is entirely rising with the peak falling towards the end of the vowel. The F0 on first syllable of *limousine* is also entirely rising, and the rise indeed continues well into the /m/ which serves as onset for the second syllable. Thus, the patterns are only rendered equivalent by a representation which distinguishes the contour itself from the way that the contour is aligned with the syllables.

A second difficulty arises from the character of intonational meaning. Because intonation patterns are not referential (that is, they do not denote objects in the world), their meanings are notoriously slippery. Admittedly, the meanings of many words and morphemes (such as pragmatic particles) are every bit as slippery. The result, however, is that shortcuts to determining phonemic contrast are not available. There is no equivalent in intonational studies to showing someone a picture and asking *"Is this a 'pat' or a 'bat'?"*. To establish the equivalence of some patterns and the contrastiveness of

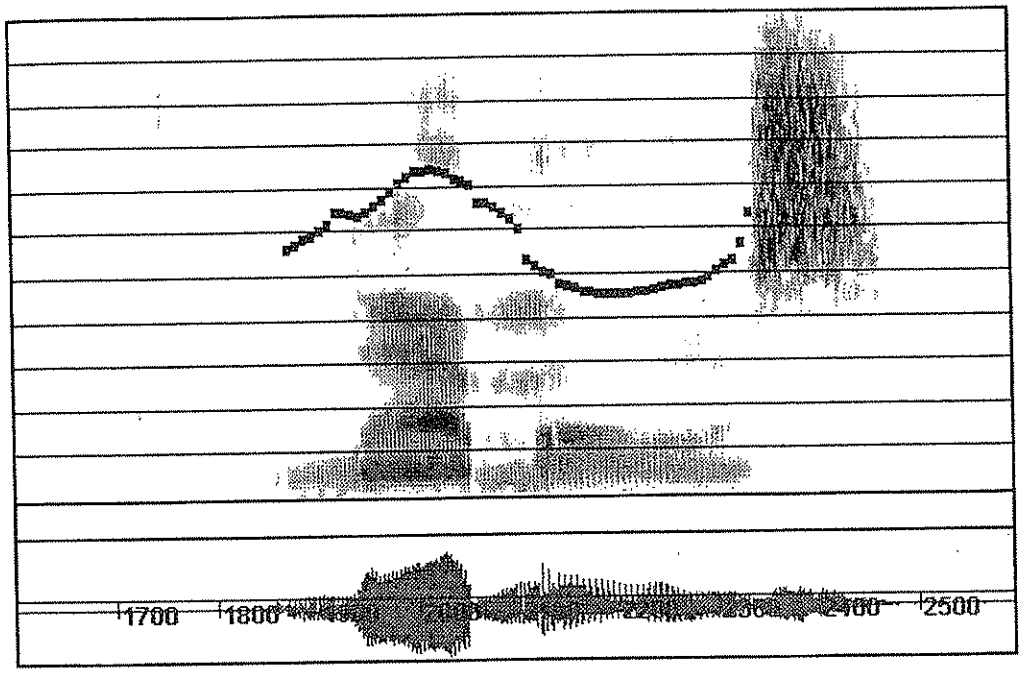

Figure 2b. A declarative pattern with a continuation rise on the word limo.

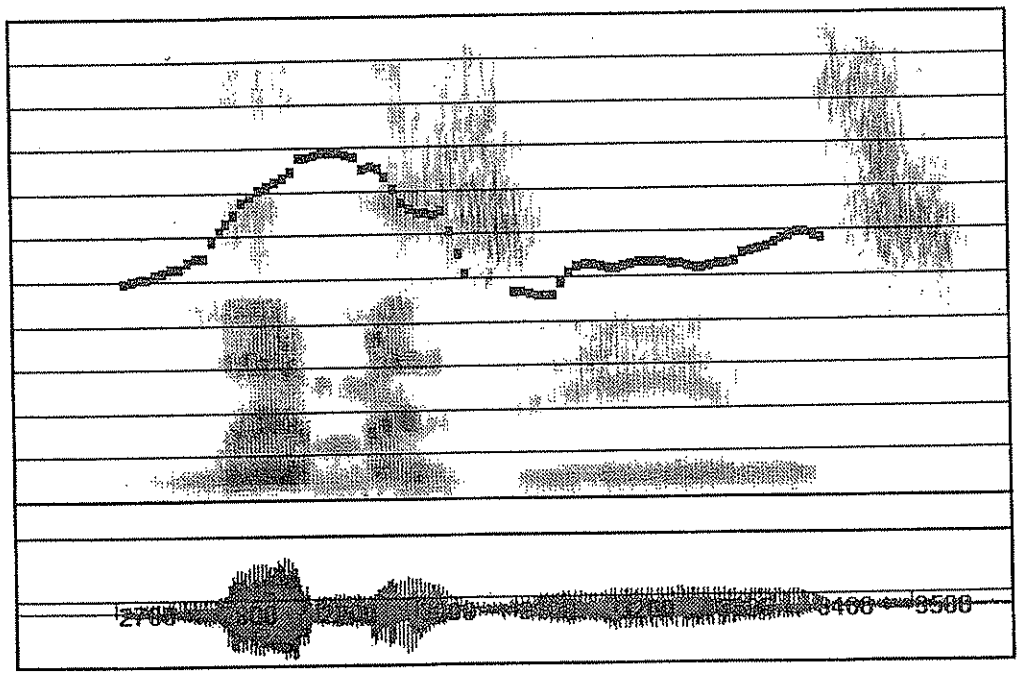

Figure 2c. A declarative pattern with a continuation rise on the word limousine.

others, exacting experiments are required. Even now, only a few of the experiments that would be relevant have actually been carried out.

A last difficulty is that it has often proved difficult to separate conventionalized and quasi-categorical aspects of the intonation pattern from stylistic variation. Indeed some scholars question whether such a separation is possible.

In the early literature, there are two competing approaches to abstracting the intonational pattern from the words. One, due to Trager and Smith (1951) and Pike (1945), decomposes the melody line in terms of tone levels, positing four phonemically distinct levels (L, LM, HM, H). The other, associated with Bolinger (1951, 1958) as well as with most phoneticians of the British school, decomposes melodies in terms of F0 changes or trajectories. Both approaches had advocates up through recent times; for example, Liberman (1975) uses four tone levels as basic units of description, whereas Ladd (1979) uses tone changes. Both approaches have intrinsic advantages and drawbacks. Two-tone models have become a recent standard because they integrate insights from the two approaches while avoiding most of the drawbacks.

An important drawback of four-tone models was already pointed out in Bolinger (1951). It is obvious that the four tones of these models do not denote absolute F0 levels (even within the speech of a single individual), but rather relative position within the pitch range; overall pitch range varies with the speaker's voice level, emotional state, and choice of style. That is, tones are relativized to an F0 space in a way which is reminiscent of vowels in the formant space (see Ladefoged and Broadbent 1957), but which is even more extensive and pervasive. As Bolinger pointed out, the sparsity of tones relative to the rate at which overall pitch range is manipulated leads to pervasive ambiguity in analysis in four-tone models. There is no principled way to distinguish a L LM L pattern produced in a large pitch range from a L H L pattern produced in a small pitch range, and linguists should not, therefore, imagine that listeners actually do so. This criticism is extremely cogent. Theories which enforce spurious distinctions (or distinctions which could not in practice be available) are indeed generally suspect. Two-tone models of intonation substantially reduce the ambiguity of analysis by reducing the inventory of different tonal strings and by making explicit provision for the role played by pitch range in scaling the phonetic outcome.

A further problem with four-tone models, discussed in Pierrehumbert (1980), relates to the existence of stepping patterns, as illustrated in Figure 3. Reserving L for the low termination at the end of the whole pattern, the three steps in this pattern would be described as H, HM, LM. Since only four tones are available in this theory, the implication is that stepping patterns can have only three steps (plus the terminal fall). This implication is false; phrases containing four or even five steps are attested, and the only real limitation appears to be the length of the intonation phrase. But if we countenanced six or

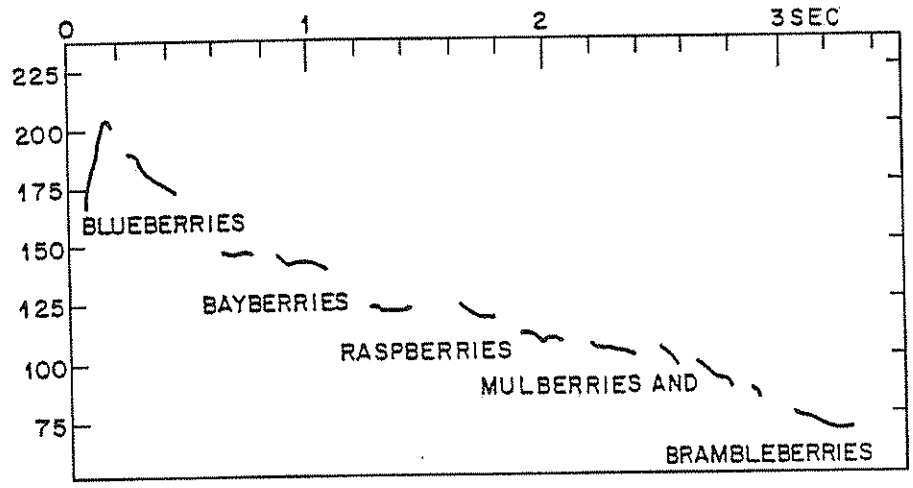

Figure 3. An F0 contour with a series of downsteps (Reproduced from Liberman and Pierrehumbert 1984).

seven different tones in the system, it would have far more tonal distinctions than were needed for any other purpose, and the problem raised by Bolinger would be further aggravated.

The solution to this problem proposed in Pierrehumbert (1980) is to reduce the tonal inventory to H and L and to posit a process of downstep which successively lowers the realization of the H tones. That is, the pattern shown in Figure 3 has the same tonal component repeatedly, with each one stepped down compared to the one before. This solution is inspired by the downstep found in many African tone languages. As pointed out in Anderson (1978), the potentially infinite number of steps in a downstepped sequence can be described as an abstract exponential. Liberman and Pierrehumbert (1984) build on this insight to develop a model of downstep in English in which the process literally is an exponential decay under the correct representation of F0 scaling.

F0-change models also have characteristic weaknesses. The first weakness concerns the issue raised in Figure 2, that of the alignment of the pattern to the segmental material. In order to describe the alignment under different conditions of stress and phrasing, it proves necessary to refer to the endpoints or extrema of the F0 changes. For example, the termination of the pattern illustrated in Figure 2 has a local maximum aligned exactly to the end of the phrase. Any reference to the endpoints of the changes amounts to decomposing the changes into successive F0 levels. Then, too, English has some qualitative distinctions in the positioning of a change in the pitch range. The most uncontroversial are probably the distinction between the "high rise" found in some questions and the "low rise" found in declaratives with continuation, and the distinction between melodies which fall to the bottom of the range and those which fall only part way. Distinguishing these patterns phonologically

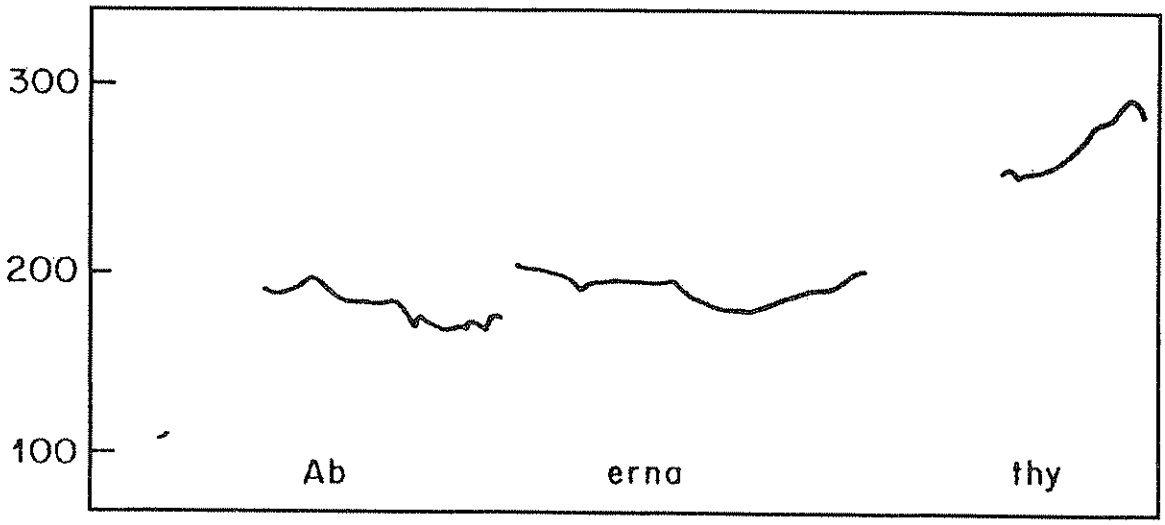

Figure 4. A low rising F0 contour (L L H%) on the word* Abernathy, *illustrating the fact that a pitch accent does not necessarily entail a pitch inflection on the accented syllable.*

requires some way to allude to position in the range, not just to the direction of change. Lastly, as argued in Pierrehumbert (1980), pitch accents can in certain circumstances emerge as F0 levels rather than F0 changes. For example, the nuclear stress on the utterance *Abernathy* whose F0 contour is displayed in Figure 4 has a relatively low F0 value on the syllable /æb/, which is continued directly into a low rising terminal configuration. Despite the impression of prominence on this syllable, there is no movement associated with it as such.

2. TWO-TONE MODELS OF ENGLISH INTONATION

A comprehensive model of English intonation using two tones (L and H) was proposed in Pierrehumbert (1980). It provided a grammar of English melodies and sketched an algorithm for mapping outputs of this grammar into F0 contours. The tonal phonology was revised in Beckman and Pierrehumbert (1986) and the F0 scaling algorithm was revised in Liberman and Pierrehumbert (1984). An intonation synthesis algorithm based on this model is presented in Anderson et al. (1984). This entire body of work builds substantially on Bruce's (1977) model of Swedish accent and intonation. Before presenting it, I would accordingly like to review the contributions of Bruce's work.

2.1 Bruce (1977)

Bruce's (1977) Ph.D. dissertation *Swedish Word Accents in Sentence Perspective* was a unified treatment of accent and intonation in the Stockholm

dialect of Swedish. By examining the F0 contours of utterances which systematically varied accent, stress, and phrasing, Bruce arrived both at a theoretically novel treatment of Swedish melody, and at a method for synthesizing F0 contours. The treatment of the melodic line decomposes it into H and L tones, at a relevant level of abstraction. Some tones originate from the words (via the well-known Accent I – Accent II distinction amongst Swedish words) and others originate from the phrase. The synthesis scheme reconstructs the F0 contour by mapping tones onto F0 targets and interpolating between the targets. When tones are crowded together, some tonal targets are not fully realized; the priority amongst the targets is determined by their phonological status.

The work is distinguished first by its unified treatment of phonology and phonetics. Rather than taking the phonology as given, the work uses phonetic data to clarify phonological issues. Overall, the work is a lesson in the fact that phonology and phonetics cannot be studied separately, but only together.

A second distinctive feature of this work is its abstract view of the basic tonal elements. Much early work on prosody and intonation (such as Fry 1958) takes citation forms of words as basic. Insofar as the intonation of continuous speech was treated at all, it was in terms of concatenation and reduction of word patterns which could have been found in isolation. Bruce, in contrast, adopted the working hypothesis that the "basic" word patterns were abstract patterns whose character would be revealed by examining the full range of variation found when words are produced in different contexts. From his survey of contextual variants, the prenuclear rather than the isolation version of each accent emerges as the most basic form, provided that abundant segmental material gives it a vehicle for full expression. The phonetic form that accents acquire in isolation arises through the interaction of the accent itself with tonal correlates of phrasing; hence the isolation form is more complicated than the prenuclear form. In addition, the postnuclear forms of the accents are phonetically modified by downstepping, while all accents are subject to undershooting in situations of tonal crowding. Because of these effects, no particular tone corresponds to an invariant level or F0 configuration. The abstract use of L tones is particularly striking, because some L tones emerge as valleys in the F0 contour, others as elbows (in downstepped contexts), and others are barely expressed at all.

Bruce's approach also made it possible for him to discover the common denominator of the Swedish Accent I and Accent II patterns – namely, the H L found in both cases – and to show that the critical difference between the two relates to their alignment. For Accent I, the HL is aligned earlier with respect to the segmental material, and for Accent II, the HL is aligned later. The contrast is displayed in Figure 5, reproduced from Figure 28 of Bruce (1977). Perception experiments using synthetic speech with controlled

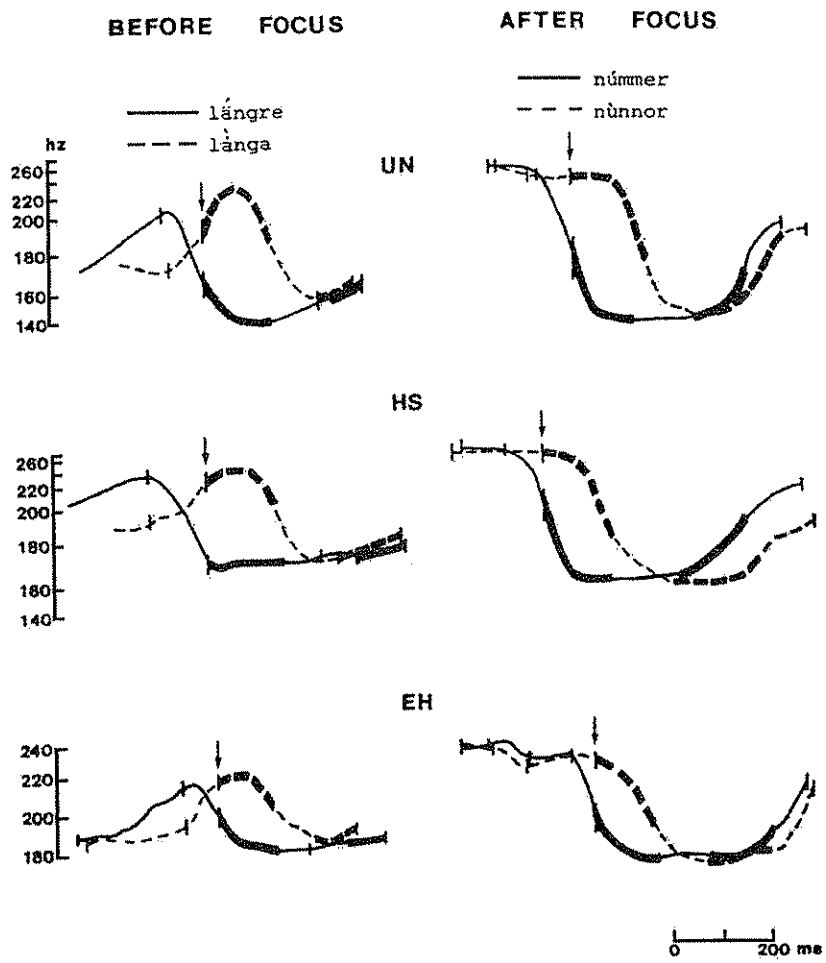

Figure 5. Figure 28 of Bruce (1977), illustrating how Swedish Accent I and Accent II differ in their alignment characteristics. Bruce's figure caption is:"The word accent distinction in non-focal position. F0-contours of accent I- and accent II-words in the second and third positions for three speakers. The line-up point is at the CV boundary of the stressed syllable."

variation of the F0 contours further substantiate his point that relative alignment is the heart of the Accent I/Accent II distinction.

The actual F0 contour on any word emerges from Bruce's model through the interaction of the word accent with tonal features assigned at the phrase level. Specifically, the phrase contributes a H tone which is phonetically manifested towards the end of the word carrying phrasal stress, and the initial and terminal junctures also contribute tonal targets. The citation form is then reconstructed as the form produced in a specific prosodic context – when the word is both phrase-final and bears the main stress of the phrase. The importance of this point cannot be overemphasized. In effect, there is no such thing as an intonation pattern without a prosodic context. The nuclear position

and the phrase-final position are both particular contexts, and as such leave their traces in the intonation pattern.

In Bruce's model, some parts of the F0 contour are more phonologically critical than others. The critical parts are those at which paradigmatic or syntagmatic information is concentrated. The differential importance of different parts of the contour is captured by a modified target-interpolation model for the phonetics. Phonological tones are mapped onto F0 targets, and the targets are connected together to make a continuous F0 contour which includes F0 values for sonorant regions which are not tonally critical. Bruce acknowledges the fact that in situations of tonal crowding, phenomena of undershooting and readjustment may create departures from the capabilities of a pure target-interpolation model. These deviations from the model should not disguise the fundamental insight, which is that some parts of the F0 contour are more critical than others, and that tonal elements are semi-localized in their phonetic manifestations.

Bruce's treatment of tonal alignment provides an important antecedent to the present understanding of licensing. Bruce's work appeared just after the first works in autosegmental and metrical phonology (Leben 1973, Goldsmith 1976, Liberman 1975). Autosegmental phonology was originally motivated by regularities in the tone patterns of African tone languages. To describe these patterns, the theory attributed unaligned melodic strings to the underlying representations of morphemes or words. The surface alignment of the tones was derived by rules for associating the elements of the strings (namely tones) to tone-bearing units. A typical rule mapped the tones left-to-right in one-to-one association with tone bearing units, with some provisions (such as spreading, crowding, and deletion) for situations in which the length of the tonal string and the number of tone-bearing units differ. Liberman's metrical treatment of English intonation differed from autosegmental accounts of tone languages in attributing melodies to phrases rather than to words, a consequence of the nonlexical status of the English melodic line. It also differed in its specific proposal for aligning the tones: the alignment algorithm appealed to hierarchical structure in order to align the tones to the metrically strongest elements in the phrase. However, the general outlines of the two proposals are similar. In both, the entire tonal sequence for a meaningful phonological construct ends up aligned to specific tone-bearing units of that construct through an across-the-board process.

Bruce's treatment of Swedish accent and intonation fits in this mold by distinguishing the morphological source of a tone from the phonological spot where it shows up on the surface. By the morphological source, I mean the domain in which a choice of tone is contrastive. In intonation languages – in which tones are relatively sparse – the underlying domain is typically bigger than the critical phonetic region for the tone. For example, English has the edge of each phrase marked with a tone; in this case the underlying domain is

the intonation phrase, but the surface alignment is to the last syllable or the boundary. An important difference between Bruce's treatment and those just mentioned is that there is more than one source of tones; the eventual melody arises from the interplay of word tones and phrasal tones. Furthermore, tones from each respective source have their own characteristic timing and scaling behavior. This more complicated picture finds an analogy in the present understanding of licensing. Compare, for example, the treatment of syllable-level phonotactics in Goldsmith (1990) and Coleman (1992). These accounts differ in detail, but in both different nodes in the syllable carry different featural properties. The properties at each level have a characteristic temporal scope, and the surface form of the syllable arises from the interplay of the properties at the various levels.

2.2 Pierrehumbert (1980)

Pierrehumbert's (1980) model of English intonation (later revised in Beckman and Pierrehumbert 1986) adopted many of the main insights of Bruce's model of Swedish. Specifically, it described even very complex F0 contours in terms using just two basic tone levels (H and L). It proposed bitonal pitch accents, phonologically located on metrically prominent syllables. Early-aligned accents are phonologically distinguished from late-aligned accents. Relative alignment is indicated notationally by a *; L*+H has L on the stressed syllable with a trailing H, whereas L+H* contrasts by having a H on the stressed syllable with a leading L. The model also distinguished pitch accents from boundary tones on the basis of their characteristic timing behavior. Making this distinction allows the inventory of pitch accents to be the same in prenuclear and in nuclear position. The relatively complex F0 contours found on phrase-final nuclear syllables arise not from a special nuclear inventory, but rather from the crowding of the pitch accent and the boundary tones onto a single syllable. Lastly, a fully explicit but non-trivial phonetic implementation component maps phonological tones onto context-dependent F0 targets. Interpolation and smoothing between targets is responsible for the continuous F0 contour observed. All of these basic features also characterize Bruce's model for Swedish.

English contrasts with Swedish in that the pitch accents are not underlying properties of words. Instead, they are independent pragmatic morphemes which are co-produced with words. The fact that the pitch accents land on the metrically prominent elements of words may be attributed to the general process of entrainment in motor control – this is the process whereby your two hands become synchronized if you pat your head and rub your stomach at the same time.

A further contrast between English and Swedish is that English has more different kinds of pitch accents. Pierrehumbert (1980) proposed seven different

pitch accents: H*, L*, L+H*, L*+H, H+L*, H*+L, H*+H. Beckman and Pierrehumbert (1986) reduced the inventory to six by eliminating the H*+H as a categorically contrastive element. English also has more different boundary treatments than Swedish. In the Beckman and Pierrehumbert model, there are two levels of intonational phrasing, the intermediate phrase and the full intonation phrase. Each of these has a boundary tone, either L or H, although the timing behavior of these tones is rather different. The intermediate phrase boundary tone tends to spread over the entire region from the nuclear accent to the end of the phrase, whereas the intonational boundary tone is more localized right at the phrasal edge. The cross-product of the two choices creates four different post-nuclear configurations for intermediate phrase boundaries which are also intonational phrase boundaries. An optional phrase-initial boundary tone for the intonation phrase is also posited.

Given this inventory of elements, a full grammar of possible patterns is shown in Figure 6. This grammar is graphed as a finite-state network; any path through the network yields a well-formed phrasal melody. The grammar is constructed as if an intermediate phrase could have an indefinitely large number of pitch accents; this is obviously an idealization. In practice, most phrases have one, two, or three pitch accents, and it is extremely unusual to find a phrase with as many as five accents. The responsibility for describing this limitation need not fall on the intonational grammar per se, however; provided that the accents are phonologically constrained to align with metrically strong syllables, then the number of strong syllables in real live phrases naturally limits the number of accents. The grammar is also constructed on the assumption that the pitch accents and boundary tones are found in all possible combinations. As noted above, all combinations of pitch accent with boundary tones are indeed found. However, combinations of different accents within the same phrase are not found to the extent expected, see below.

In the model, tonal sequences are mapped onto F0 targets by locally context sensitive implementation rules. A regular process of interpolation contexts the targets, as discussed above. The realization rules resemble the transformational rules of generative phonology by virtue of applying when their structural description is met, with a structural description being a fragment of a complete tonal analysis. For example, the rule applying to a H% after H applies to any H% in this local context, regardless of what pitch accents may occur in the greater context. They differ in that the output of the rules are not phonological descriptions, but rather parametric phonetic values. In addition, the rules are not ordered in a derivation. Instead, they apply in a "running window" over the phonological description, mimicking the process whereby speakers transform their abstract intentions for an utterance into actual phonetic outcomes with particular physical characteristics.

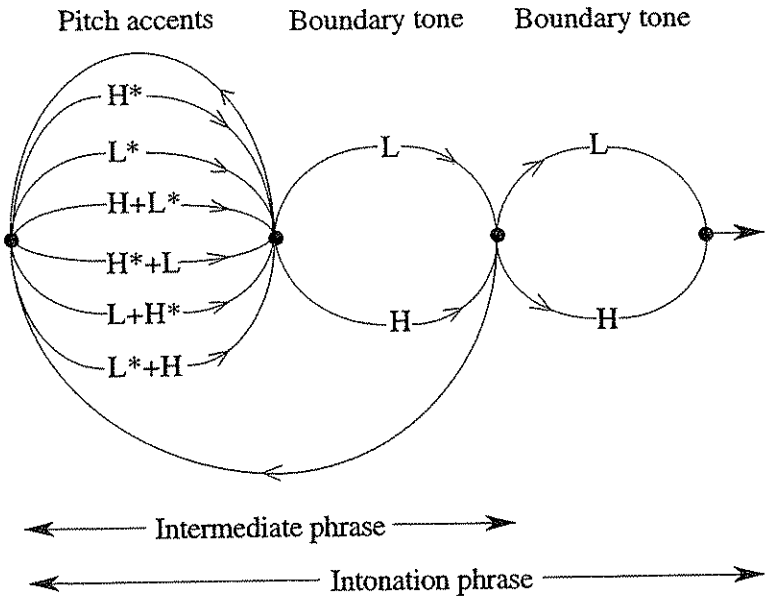

Figure 6. The grammar of English intonation patterns, according to Beckman and Pierrehumbert (1986).

There are three main contextual dependencies in the model. First, the entire space of F0 targets depends on the speaker's choice of pitch range. In the original Pierrehumbert (1980) model, a choice of pitch range was equated with the choice of first peak value. Experimental results in Liberman and Pierrehumbert (1984) and Pierrehumbert and Beckman (1988) led to a revision; choice of pitch range is now understood as selection of an abstract phrase-level parameter which in effect determines the graph paper on which the tones for the phrase are graphed.

The second major context dependency in the model is downstep. In many African tone languages, the second H in a H L H sequence is lower than the first. If the alternating pattern continues (H L H L H...), then each H is lower than the one before, so that a descending staircase results; see Figure 3 above. An analogous effect is found in Japanese, but triggered only by the H+L accent (and not by H and L tones from other sources, even in alternation). Pierrehumbert (1980) proposed that English has a downstep rule affecting H+L H and H L+H sequences. Sequences displaying these particular accent configurations then staircase downwards, in contrast to the bumpy or slightly descending outcome for a sequence of plain H* accents; see Figure 7. Beckman and Pierrehumbert (1986) revised the proposal about the trigger, holding that downstep in English is triggered by two-tone accents. However, the core insight is retained: specifically, the insight that positing a downstep rule for English makes it possible to analyze the many observed F0 target levels for H tones as manifestations of an abstract two-tone system. This proposal is particularly important in analyzing so-called "calling" contours, in

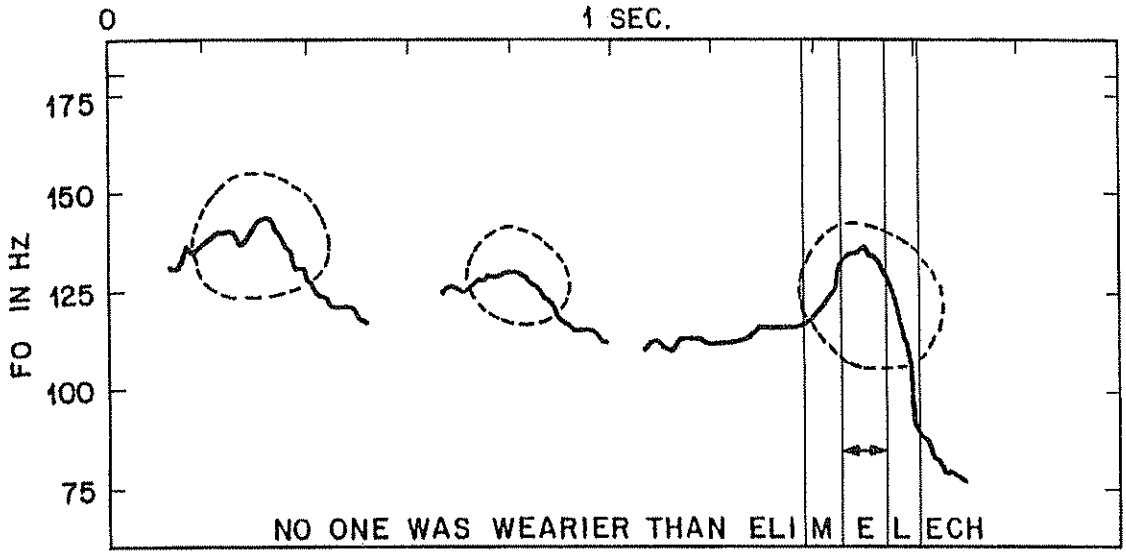

Figure 7. An F0 contour with three H accents, with corresponding F0 peaks circled.*

which the F0 falls after a nuclear peak, but levels out in the middle of the pitch range instead of falling all the way to the bottom. Previously analyzed using an otherwise unsupported M (mid) phrasal tone, these patterns can now be understood as having a downstepped H phrasal tone.

American English is also claimed in this model to have an upstep rule, which applies only to intonation phrase boundary tones (whether L% or H%) after H. This rule is responsible for the fact that F0 contours with a H phrase accent either sustain the same level, or else rise even further at the end (in the canonical yes/no question pattern). Unlike other languages, such as Hungarian, an F0 pattern which rises to a H phrase accent and then falls down again to the boundary does not occur. Some dialects of British English also have the upstep rule, but it appears that others do not. In the dialects which lack the upstep rule, the high-rising questions do not occur, but the rising-falling post-nuclear configuration does occur. As a result, the contrast between a relatively high (or H) and a relatively low (or L) termination after H is still found, and the two-tone decomposition is thus supported.

A complete set of schemata of nuclear/postnuclear configurations is found in Figure 8, reproduced from Pierrehumbert and Hirschberg (1990). Pierrehumbert and Hirschberg also sketch a compositional account of the meanings of the contours. The core concept in this model is the relationship of each phrase to the mutual beliefs as they are built up by interlocutors during a discourse. The H* accent is used to mark focused information which is to be added to the mutual beliefs; the L* accent marks information which is salient but which is for some reason not proposed as an addition. (For example, it may be already present in the mutual beliefs, or it may be doubtful or false.) The L+H accents mark information which is selected from a small domain of

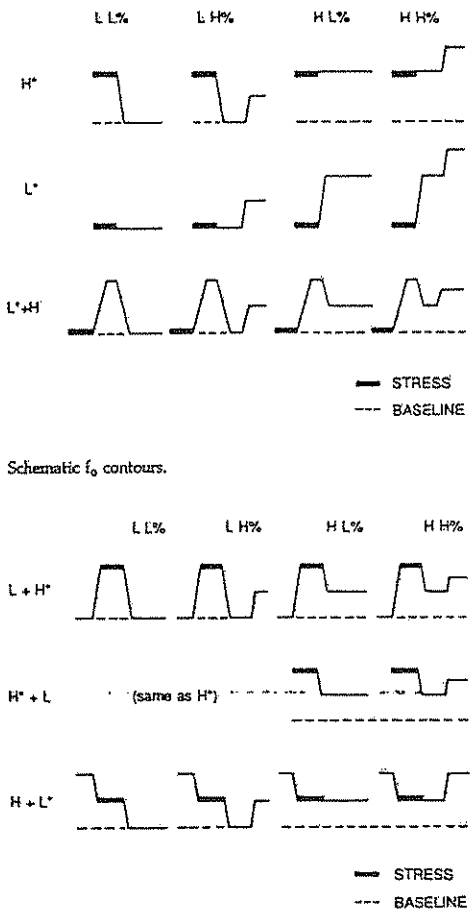

Figure 8. The full inventory of nuclear accents in combination with phrase-final tones, as schematized in Pierrehumbert and Hirschberg (1990).

alternatives, with the L+H* marking an "add" and the L*+H marking a "non-add". The H+L accents represent an instruction to identify a relationship of the information to information which is already mutually believed. The boundary tones differ from the pitch accents in this model in marking the discourse status of the phrase as a whole rather than the status of individual focused elements in the phrase. For a broadly related but competing proposal about the morphemes of the intonational system, see also Gussenhoven and Rietveld (1991).

2.3 ToBI

ToBI (an acronym standing for "Tone and Break Indices") is a standard for transcription of English intonation which was developed and tested by an international group of researchers in the field (see Silverman et al. 1992). The purpose of the standard is to further scientific study of intonation and

technological development by permitting researchers at different laboratories to interpret each other's data and to pool resources in developing on-line databases of prosodically transcribed speech. The immediate antecedents of the standard are Pierrehumbert (1980) and Beckman and Pierrehumbert (1986) (for the decomposition of the melody into L and H tones, organized into bitonal accents and phrasal tones); Ladd (1983) (for the treatment of downstep) and Wightman et al. (1992) for the treatment of juncture. An experiment on inter-transcriber reliability was carried out to validate the system, with 26 independent transcribers analyzing 34 varied utterances. As reported in Pitrelli et al. (1994), the protocols permitted very good reliability, as assessed by the number of transcriber pairs agreeing on the labeling of each word. There was 88% agreement on the presence of tonal elements, 81% agreement on the exact label of tonal elements, and 92% agreement to within one level on the assignment of break indices. This level of reliability is much higher than for previous systems of intonation transcription and will permit the development of shared intonational corpora.

ToBI provides for four parallel channels of transcription. One is the orthographic or phonetic transcription of the words. The second is the melody line, which follows the general outline of Beckman and Pierrehumbert (1986) by providing for monotonal and bitonal accents plus two levels of phrasal tones. Transcribers can also mark the F0 maximum in each phrase, providing a crude but replicable index of the current pitch range. The third channel carries indices describing the strength of the juncture between each two lexical items, ranging from 0 (indicating that cliticization has turned two lexical items into a single prosodic word) to 4 (indicating a maximal, or fully-marked, intonational phrase boundary). The last provides for comments of any type.

Although the standard has obviously been influenced by current theory to a significant extent, it is theory-neutral in several important respects. First, there has been and continues to be controversy about whether downstep in English is predictable from the type and grouping of pitch accents, or whether it is an independent dimension of choice. In ToBI, downsteps are explicitly transcribed (in the style of Ladd 1983) in the hope that researchers will eventually gather enough data to settle this issue. Secondly, Beckman and Pierrehumbert (1986) advanced the hypothesis that non-accentual tones are tightly linked to levels in a hierarchical prosodic structure, with the intermediate phrase contributing one tone (L or H) and the intonation phrase contributing another. For the ToBI reliability trial, Pitrelli developed a transcription parser which enforces this regularity. However, the transcription standard itself provides no impediments to recording junctures which are not synchronized in any particular way to the melodic stream. Anyone with a different theory of how tones and junctures are related could write a different transcription checker to verify the relationship between the channels in the light of their own theory. Lastly, the inventors of ToBI do not claim that the

melody and break index channels exhaust the information relevant to the understanding of intonation. These two channels represent information about which there is a broad consensus in the speech community. The fourth channel can be used to record any further observations about the intonation or prosody that may be of interest. Observations about expressive use of voice quality would be an obvious example.

After ToBI was developed, there arose considerable interest in how it might be applied to other languages. It is important to understand that ToBI is not directly applicable to other languages (or even to some dialects of English) because it is at the level of abstraction of a broad phonemic transcription, or rationalized spelling system, such as those of Korean and Finnish. Just as a broad phonemic transcription for any language must be guided by the phoneme inventory of that language (as revealed by the lexical contrasts), a ToBI-style transcription of the prosody and intonation of any language must be guided by an inventory of its prosodic and intonational patterns.

As such inventories are made, we find many recurring themes in the dimensions of phonetic contrast which are employed in prosodic and intonational systems. Recurrent dimensions include relative F0 height, relative duration, relative alignment, relative force of articulation, and so forth. In a similar way, inventories of vowel systems in the languages of the world reveal the existence of recurrent dimensions such as front/back and rounding. However, equating any particular vowel in one language with a particular vowel in another is highly problematic. For example, although the French high front unrounded vowel is broadly analogous to the English one, its exact degree of height, frontness, and spreading is different, as is its pattern of variation in context. Similarly, we do not expect to see tonal elements literally equated across languages. Instead, the expectation is that broad patterns of contrast and tonal realization might be echoed from one system to the next.

3. EXPERIMENTAL VALIDATION

Some broad features of the two-tone models have been validated by subsequent work; others have been called into question, and others have not been tested experimentally at all.

First, consider the claim that the accent inventory is the same in prenuclear and in nuclear position, with the more complex configurations found in nuclear position being attributable to extra tones originating at the phrasal level. Partial support in favor of this claim is found in experiments by Steele (1986) and by Silverman and Pierrehumbert (1990), who explored the timing of the F0 peak for H* accents in various positions. Previous work (notably, Silverman 1987) had reported that the timing was early in the stressed syllable for nuclear accents and late for prenuclear accents, suggesting that the claimed unity of the

H* accent in these two positions might be illusory. Steele (1986) and Silverman and Pierrehumbert (1990) fleshed out the picture by examining nuclear accents separated from the phrase boundary by varying amounts of material, as well as prenuclear accents separated by varying amounts of material from the nuclear accents. These experiments demonstrated that the two extremes of timing just mentioned actually fall on a continuous gradient. This point was already illustrated by the variable placement of the F0 peaks in Figure 2 relative to the segmental material. Factors contributing to the full gradient include the variable strength of boundaries, the variable sonority and length of the segments, and the variable amount of tonal crowding found in different contexts. With these results, the phonological unity of the prenuclear/nuclear H* accent is supported.

The existence of contours with mixed accent types provided a second line of evidence for the claim that the phonological inventory of accents is the same in nuclear and in prenuclear position. For example, H* and L* accents occur in all (four) possible combinations and orders in two-accent phrases; insofar as all proposed accents occur in all combinations and orders, this provides evidence that each one is an independent phonological unit. The intonational grammar in Beckman and Pierrehumbert (1986) generates 36 accentual combinations for phrases with two accents, and 216 combinations for phrases with three accents, even disregarding the contribution of phrasal tones. Although various combinations of accents are found, nothing like the full set generated by the grammar has ever been documented. For three accent phrases, the typical pattern is either to use the same accent type in all three positions, or else to use one type of accent in both prenuclear positions, and a different type in nuclear position. To understand this observation, it is worthwhile to bear in mind how sparsely languages in general sample the cross-product of the available units. Because of phonotactic constraints, most combinations of phonemes do not represent possible words and most phonotactically possible words do not happen to be real words. Syntactic and semantic constraints have the consequence that most combinations of real words do not constitute potentially observable sentences. Similarly, we need to work out what factors cause gaps in the set of intonation patterns observed. Are there phonotactic factors that are not yet understood? Are phrasal intonation patterns lexicalized as single units, with some being accidentally missing? Are some potential patterns missing because the meanings of the component parts cannot be coherently combined?

A second important feature of the Pierrehumbert (1980) is that it treated the relative alignment of two-tone accents to the segmental material as phonologically contrastive. The two L+H accents differed in which tone controlled the alignment of the whole accent via its affiliation with the stressed syllable. In the L*+H, the L* is phonologically aligned with the stressed syllable and the H falls soon after the L*; in the L+H*, the H* is aligned with the stressed syllable and the L falls soon before it. The key concept – that

English has two distinct relative alignments for the same scooped F0 configuration – was validated by the experiment described in Pierrehumbert and Steele (1989). Pierrehumbert and Steele used LPC analysis and resynthesis to create a set of 15 versions of the phrase *Only a millionaire*, differing only in the relative alignment of the L+H portion of a L+H L H% intonation pattern with the word *millionaire*. These were played 30 times in blocked randomized order to subjects, who were asked to imitate what they heard. The peak alignment in their imitations was measured. Despite the fact that the stimuli had evenly graded peak alignments, two preferred peak alignments were observed in the productions. A related set of experiments by Kohler (1987a, 1987b) also established that relative alignment is contrastive in German. The other relative alignment distinction that Pierrehumbert (1980) proposed for English (H*+L versus H+L*) has never been tested experimentally.

In contrast to relative alignment, phrasal pitch range is claimed by Pierrehumbert (1980) to be a thoroughly gradient reflex of style and discourse structure. One of the pitch scaling experiments discussed in Liberman and Pierrehumbert (1984) provides a partial test of this claim. In the experiment, subjects produced instances of the sentence *Anna came with Manny* with two different focus structures in ten different overall pitch ranges. The desired pitch range was signaled by a number from one to ten written on a note card underneath the sentence. For every subject, the resulting productions showed a smooth gradient of F0 peak values, with no preferred values. In contrast to the Pierrehumbert and Steele experiment, gradient instructions produced gradient results. Thus, phrasal pitch range does not appear to be categorical. More informal confirmation of this point may be found in Hirschberg and Pierrehumbert (1986), who measured all phrasal pitch ranges in a monologue previously recorded for a sociolinguistic study. The phrasal pitch range was found to correspond well to the discourse structure of the monologue, with larger pitch ranges used to mark the introduction of new topics and subtopics.

In the F0 scaling model developed in Liberman and Pierrehumbert (1984), a single parameter stands as the reflex of the phrasal pitch range. It is designated as the "reference line", and represents the F0 value to which an arbitrarily long series of downsteps would assymptote. It is higher than the baseline, which is viewed as a fixed property of the speaker's voice. Related models have also been developed for Japanese (Pierrehumbert and Beckman 1988) and Spanish (Prieto et al. 1996) on the basis of data from the same experimental paradigm, in which speakers are asked to produce intonation patterns involving downsteps at different overall voice levels. In addition, work by Ladd (1993) and Terken (1993) provides further evidence that implicit parameters control the scaling of F0 contours.

The paradigm just summarized does not, however, yield the last word on H tone scaling because it tackles only one source of pitch range variation, namely overall voice level. Other choices of speech style may affect the F0 scaling in

other ways. For example, when people read stories to children using a small voice for a small character, they probably modify the baseline. Further experiments on the full range of stylistic choices would be valuable. In addition, note that F0 is only one of the phonetic parameters affected by overall voice level. Raising the voice also affects the amplitude, the spectral tilt, the degree of coupling of the subglottal system, and other phonetic characteristics. Pierrehumbert (1997) describes a pilot study on the interaction of intonation pattern and overall voice level as determinants of the voice source characteristics. Further work exploring these interactions would lead us to a much fuller understanding of the phonetic correlates of tone. Such understanding would help to address concerns about the ambiguity of F0 contours (which may be less ambiguous when all properties of the signal are considered), and it would also support improvements in the quality of synthetic speech.

The algebraic approach to F0 scaling taken by Liberman and Pierrehumbert (1984) has not been notably successful in describing the scaling of L tones, particularly those near the bottom of the range. One reason may be the complicated articulatory strategies involved in active F0 lowering; see the review in Beckman and Pierrehumbert (1992). Another problem is that what counts perceptually and phonologically as a very low tone may, from a mathematical point of view, not have a well-defined F0 at all. Vocal fold adduction and/or low subglottal pressure can produce irregular movements of the vocal folds which are not periodic at all. A unified treatment of the phonetics of tone is therefore likely to require innovations in the parameterization of the phonetic outcome.

Pierrehumbert (1980) advanced a strong hypothesis about the character of the phonetic implementation rules for tones. For any given target tone, the implementation was held to depend only on the identity and prosodic position of the tone itself, and on the identity and phonetic realization of the preceding tone. The two hallmarks of this "running window" are its strict locality and its temporal asymmetry. The current outcome can depend on actual outcomes in the past, but has no access to actual outcomes for future elements. The model thus presented a very strong contrast to superpositional models of F0 realization, such as Fujisaki (1989), in which the phrasal F0 contour arises through superposition of a phrasal F0 contour with local accent-related F0 contours. A number of experimental studies have undermined the strong claims about locality made in Pierrehumbert (1980). However, the nonlocal effects that have been found are not all amenable to treatment in a strictly superpositional approach.

Detailed studies of tone scaling (such as Liberman and Pierrehumbert 1984, Pierrehumbert and Beckman 1988) indicated the need for implicit phrase-level parameters controlling pitch range. In Pierrehumbert and Beckman (1988), these are treated formally using the conceptual apparatus of attribute

grammars. I have already noted that the implementation of a tone can depend on its prosodic position; in just the same way that the node labels on the hierarchical structure dominating a target tone can be examined by the implementation algorithm, the pitch range parameters carried by these nodes can also be examined. Formally, this is a very simple extension of the original proposal. It obviously tends in the direction of superposition of local and phrasal components; however, there are important differences. One difference is that each pitch range parameter is just a single number and not a time function. Another difference is that the phrasal pitch range parameters are not required to combine with the local components via superposition as such; rather, they figure as values for arguments in a function which computes actual F0 values. They can only be established by fitting an entire model to an entire data set. In principle, they can differentially affect different tones – for example, they might and probably do affect H tones and L tones differently. Straight superposition models do not have this capability, since the term 'superposition' by definition means that components are combined exactly one way, namely by addition on some appropriate scale.

Silverman and Pierrehumbert's findings on tonal alignment provide one example of the need for look-ahead to the upcoming tone in producing the current tone. The phonetic alignment of the tone is shown to be readjusted when another tone is coming up soon; see also the discussion in Bruce (1990). Another example is provided by Laniran and Clements' (1995) study of downstep in Yoruba; they report raising of H tones in anticipation of downsteps. Detailed consideration of final lowering provides further evidence of the need for look-ahead. In Liberman and Pierrehumbert's experiment on downstep, the last step in the sequence was found to be lowered (with reference to the decaying exponential otherwise traced out by the sequence of downsteps); this phenomenon was referred to as "final lowering". How does a target tone "know" that it is last? The hypothesis which fits in best with Pierrehumbert's original model is that exactly nuclear accents undergo the lowering; the question "is this tone in nuclear position" can be answered by examining the prosodic nodes dominating the tone, without the need to refer to flanking material. However, this hypothesis is not really tenable. An unpublished experiment by Pierrehumbert and Liberman found that the amount of lowering varies according to the distance of the nuclear accent from the end of the phrase. Data on Danish collected by Thorsen (1980a, 1980b) plainly show a small but regular effect on the penultimate accent as well as on the last one. Herman (1996) demonstrated the existence of a gradual final lowering effect spanning the last four syllables of the phrase in Kipare, a tone language of the Bantu family. In addition, Herman et al. (1996) have found instrumental support for the observation by Hirschberg and Pierrehumbert (1986) that final lowering depends on the identity of the upcoming boundary tone.

Pierrehumbert and Beckman (1988) effectively weakened Pierrehumbert's previous stance on locality by permitting the realization of any given tone to depend not only on its immediate context, but also on any attribute of any node dominating the tone. For example, a phrase-final H% is an attribute of the intonation phrase node, and as such would be accessible as an influence on the realization of any tone within the phrase. Myers (1996) draws on exactly this capability of the model to describe the tonal realization principles of Chichewa. In general, hierarchical structures provide a way to reencode apparent lookahead in terms of an upward search in the tree structure. However, the model still prohibits phonetic (as opposed to phonological) lookahead. For example, the implementation of a tone could not depend on whether the phonetic outcome for a future final boundary tone would be above or below 140 Hz.

Another strong hypothesis advanced by Pierrehumbert (1980) and Beckman and Pierrehumbert (1986) is that the occurrence of downstep is predictable in English. Pierrehumbert (1980) proposed a downstep rule which was highly analogous to that found in some African languages, in which the second H in a H L H configuration is downstepped relative to the first. Specifically, Pierrehumbert proposed that the second H in a H+ L H or a H L+H configuration is downstepped. According to Beckman and Pierrehumbert, any two-tone accent in English triggers downstep. However, neither of these claims has been substantiated by a large-scale study of naturally occurring speech; substantiation of the claim does require a full inventory of naturally occurring variation, because in any given experimental situation, subjects confine their behavior to a small subset of their full range of capabilities. The alternative to a rule predicting downstep is a phonologically contrastive downstep feature (typically transcribed as !). For example, instead of transcribing H*+L H* L L% for a sequence with a downstepped nuclear accent, the transcription could be H* !H* L L%, attributing the downstep to a distinctive attribute of the nuclear accent itself. This is the solution adopted in ToBI. A drawback of this solution is that the downstep feature is never contrastive in initial position; for example, in phrases with one accent, there is no contrast between H* and !H*. In short, downsteps are stepped down in relation to what came before; it makes no sense to posit a downstep if nothing came before. A defective distribution such as this provides a standard argument for a more abstract analysis. However, the Beckman-Pierrehumbert approach also encounters some problems. In particular, experience to date with ToBI suggests that a L+H accent configuration can be followed either by a downstepped accent, or by one of essentially equal height.

Lastly, the intonational grammar displayed in Figure 6 incorporates the claim that every intermediate phrase has at least one pitch accent. The case in which this claim has been most called into question is that of tags. Figure 9 shows F0 contours for a single phrase with a deaccented object versus an

utterance which has two phrases (under the Beckman and Pierrehumbert (1986) analysis), with the second having a L* L H% contour. Obviously, the main difference between the two is the timing of the F0 contour. However, under the analysis, the word *Manny* has an accent in the second case but not in the first. Fans of abstract analyses see no difficulty with this claim. Others (notably Bing 1979) prefer to suggest that the tag has no accent. If the tag has no accent, then the details of the grammar in Figure 6 need to be modified. However, the general approach survives. To adjudicate between these positions, more work documenting the full range of intonational variation found in tags is needed. The analysis with the accent is more plausible if this accent contrasts with other possible choices of accent in the same position.

4. CONCLUSION

The style of intonational analysis initiated by Bruce for Swedish has now been successfully applied to English as well as to other languages. Hallmarks of this approach are a limited tonal inventory (with two tones sufficing for every language so far studied which has a pitch accent or intonation system rather than a lexical tone system); a clear attribution of tonal properties to different levels of prosodic structure; and explicit, nontrivial principles of phonetic implementation.

These hallmarks emerge from a methodology in which tonal contrasts and their phonetic manifestations are systematically examined in continuous speech. With the perspective provided by continuous speech, we see that isolation forms are not in any sense basic. Instead, they simultaneously display the complexities of initial, final, and nuclear position.

In surveying the intonational system of any language, priority must be placed on identifying the dimensions of contrast in the language. Controlled experiments can be used to identify regularities of timing and F0 scaling, and to distinguish gradient from categorical effects. Recent improvements in transcriptional tools (such as ToBI, with its associated software utilities) also provide unprecedented capability for exploring the full range of intonational variation found in expressive natural speech.

F0 has proved to be an extremely useful parameter for exploring intonation systems. One reason is that it is single-dimensional (facilitating statistical analysis of the data); another is that F0 data can be obtained and analyzed in immense quantity. Yet a third is that it can be naturalistically manipulated to construct stimuli for perception experiments. Bruce (1977) made full use of this situation. The example he set is one reason why the study of intonation is more advanced than the study of any other aspect of continuous speech.

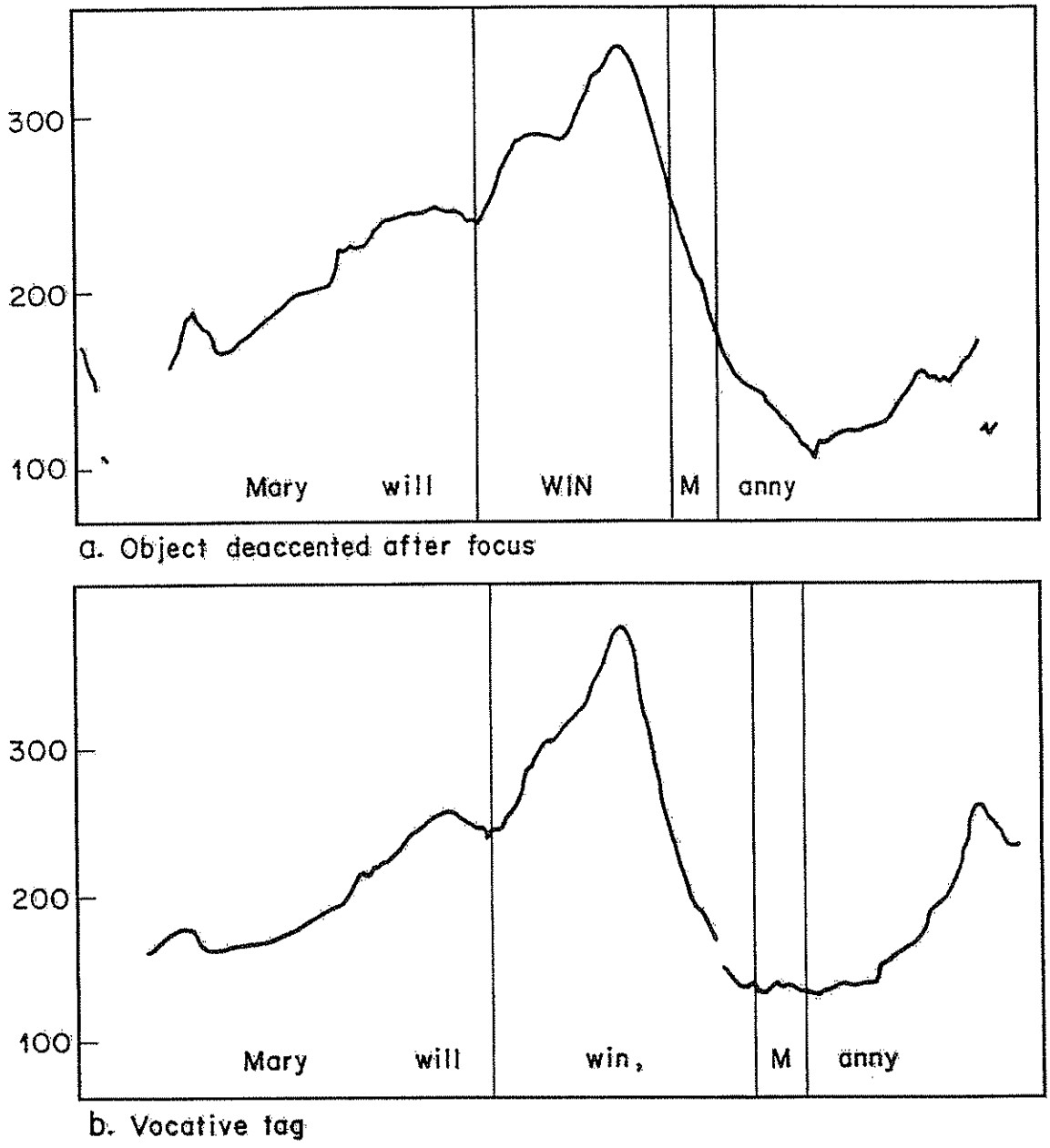

Figure 9a. A H L H% pattern produced on a single phrase in which the word* win *is under focus and* Manny *is deaccented (Reproduced from Beckman and Pierrehumbert 1986). Figure 9b. The highly similar F0 pattern (H* L | L* L H%) in which* win *carries a nuclear accent and* Manny *is a vocative tag standing as a separate intermediate phrase (Reproduced from Beckman and Pierrehumbert 1986).*

REFERENCES

Anderson, M.J., Pierrehumbert, J. and Liberman, M.Y. 1984. Synthesis by rule of English intonation patterns. *Proc. IEEE Congress on Acoustics, Speech, and Signal Processing* I, 2.8.1-2.8.4.

Anderson, S.R. 1978. Tone features. In V.A. Fromkin (ed), *Tone: A Linguistic Survey*. New York: Academic Press, 133-176.

Beckman, M.E. and Pierrehumbert, J. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3, 15-70.

Beckman, M.E. and Pierrehumbert, J. 1992. Tactics and strategies for thinking about F0 variation. In G. Docherty and D.R. Ladd (eds), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge: Cambridge University Press, 387-387.

Bing, J. 1979. *Aspects of English Prosody*. Ph.D. Dissertation, Univ. of Massachusetts at Amherst.

Bolinger, D. 1951. Intonation: Levels versus configurations. *Word* 7, 199-210.

Bolinger, D. 1958. A theory of pitch accent in English. *Word* 14, 109-149.

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

Bruce, G. 1990. Alignment and composition of tonal accents: comments on Silverman and Pierrehumbert's paper. In J. Kingston and M.E. Beckman (eds), 107-114.

Clements, G.N. and Ford, K. 1979. Kikuyu tone shift and its synchronic consequences. *Linguistic Inquiry* 10, 179-210.

Coleman, J. 1992. The phonetic interpretation of headed phonological structures containing overlapping constituents. *Phonology* 9, 1-44.

Fry, D. 1958. Experiments in the perception of stress. *Language and Speech* 1, 125-152.

Goldsmith, J. 1976. *Autosegmental Phonology*. Ph.D. dissertation, MIT. [Published in 1979 by Garland Publishing, New York.]

Goldsmith, J. 1990. *Autosegmental and Metrical Phonology*. Cambridge MA: Blackwell.

Gussenhoven, C. and Rietveld, A.C.M. 1991. An experimental evaluation of two nuclear-tone taxonomies. *Linguistics* 29, 423-449.

Herman, R. 1996. Final lowering in Kipare. *Phonology* 13, 171-196.

Herman, R, Beckman, M. and Honda, K. 1996. Subglottal pressure and Final Lowering in English. *Proc. International Congress of Spoken Language Processing*, vol. 1, 145-148.

Hirschberg, J. and Pierrehumbert, J. 1986. Intonational structuring of discourse. *Proc. 24th Meeting of the Association for Computational Linguistics*, 136-144.

Kingston, J. and Beckman, M.E. (eds), 1990. *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge: Cambridge University Press.

Kohler, K.J. 1987a. Categorical pitch perception. *Proc. XIth International Congress of Phonetic Sciences* (Tallin), vol. 5, 331-333.

Kohler, K.J. 1987b. The linguistic functions of F0 peaks. *Proc. XIth International Congress of Phonetic Sciences* (Tallin), vol. 3, 149-151.

Ladd, D.R. 1979. *The Structure of Intonational Meaning*. Ph.D. dissertation, Cornell University.

Ladd, D.R. 1983. Phonological features of intonational peaks. *Language* 59, 721-759.

Ladd, D.R. 1993. On the theoretical status of 'the baseline' in modeling intonation. *Language and Speech* 36, 435-451.

Ladefoged, P. and Broadbent, D. 1957. Information conveyed by vowels. *Journal of the Acoustical Society of America* 29, 98-104.

Laniran, Y. and G.N. Clements. 1995. A long-distance dependency in Yoruba tone realization. *Proc. XIIIth International Congress of Phonetic Sciences* (Stockholm), vol. 2, 734-737.

Leben, W. 1973. *Suprasegmental Phonology*. Ph.D. dissertation, MIT.

Liberman, M.Y. 1975. *The Intonation System of English*. Ph.D. dissertation, MIT. [Published by Garland Publishing, New York].

Liberman, M.Y. and Pierrehumbert, J. 1984. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle (eds), *Language Sound Structure*. Cambridge MA: MIT Press, 157-233.

Myers, S. 1996. Boundary tones and the phonetic implementation of tone in Chichewa. *Studies in African Linguistics* 25, 29-60.

Pierrehumbert, J. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. dissertation, MIT. Reproduced by Indiana University Linguistics Club, Bloomington.

Pierrehumbert, J. 1997. Consequences of intonation for the voice source. In S. Kiritani, H. Hirose, and H. Fujisaki (eds), *Speech Production and Language*, Speech Research 13. Berlin: Mouton de Gruyter, 111-131.

Pierrehumbert, J. and Beckman, M.E. 1988. *Japanese Tone Structure*. Cambridge, Mass.: MIT Press.

Pierrehumbert, J. and Hirschberg, J. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack (eds), *Intentions in Communication*. Cambridge, Mass.: MIT Press, 271-311.

Pierrehumbert, J. and Steele, S. 1989. Categories of tonal alignment in English, *Phonetica* 46, 181-196.

Pitrelli, J.F., Beckman, M.E. and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. *Proc. International Conference on Spoken Language Processing* (Yokohama), 123-126.

Pike, K.L. 1945. *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Prieto, P., Shih, C. and Nibert, H. 1996. Pitch downtrend in Spanish. *Journal of Phonetics* 24, 445-473.

Silverman, K. 1987. *The Structure and Processing of Fundamental Frequency Contours*. Ph.D. dissertation, Cambridge University.

Silverman, K. and Pierrehumbert, J. 1990. The timing of prenuclear High accents in English. In J. Kingston and M.E. Beckman (eds), 72-106.

Silverman, K., Beckman, M, Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. 1992. ToBI: A standard for labelling English prosody. *Proc. International Conference on Spoken Language Processing* (Banff), vol. 2, 867-870.

Steele, S. 1986. Nuclear accent F0 peak location: effects of rate, vowel, and number of following syllables. *Journal of the Acoustical Society of America* 80 Supplement 1, S51.

Terken, J. 1993. Baselines revisited. Reply to Ladd. *Language and Speech* 36, 453-459.

Thorsen, N. 1980a. Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish. *Annual Report of the Institute of Phonetics, University of Copenhagen* 14, 1-29.

Thorsen, N. 1980b. A study of the perception of sentence intonation – evidence from Danish. *Journal of the Acoustical Society of America* 67, 1014-1030.

Trager, G.L. and Smith, H.L. 1951. *An Outline of English Structure* 41. Norman OK: Battenburg Press.

Ward, G. and Hirschberg, J. 1985. Implicating uncertainty: the pragmatics of fall-rise. *Language* 61, 747-776.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P.J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91, 1707-1717.

Department of Linguistics, Northwestern University, Evanston, IL, USA