Studies in Natural Language Processing
Branimir K. Boguraev, *Editor*

This series publishes monographs, texts, and edited volumes within the interdisciplinary field of computational linguistics. It represents the range of topics of concern to the scholars working in this increasingly important field, whether their background is in formal linguistics, psycholinguistics, cognitive psychology, or artificial intelligence.

Also in this series:

Memory and context for language interpretation by Hiyan Alshawi
The linguistic basis of text generation by Laurence Danlos
Natural language parsing edited by David R. Dowty, Lauri Karttunen, and Arnold Zwicky
Relational models of the lexicon by Martha Walton Evens
Computational linguistics by Ralph Grishman
Semantic interpretation and the resolution of ambiguity by Graeme Hirst
Reference and computation by Amichai Kronfeld
Machine translation edited by Sergei Nirenburg
Semantic processing for finite domains by Martha Stone Palmer
Systemic text generation as problem solving by Terry Patten

Challenges in natural language processing

Edited by
MADELEINE BATES
and
RALPH M. WEISCHEDEL
BBN Systems and Technologies
Cambridge, MA

CAMBRIDGE
UNIVERSITY PRESS

9 Prosody, intonation, and speech technology

JANET PIERREHUMBERT

9.1 Introduction

The purpose of this chapter is to explore the implications of some facts about prosody and intonation for efforts to create more general and higher quality speech technology. It will emphasize parallels between speech synthesis and speech recognition, because I believe that the challenges presented in these two areas exhibit strong similarities and that the best progress will be made by working on both together.

In the area of synthesis, there are now text-to-speech systems that are useful in many practical applications, especially ones in which the users are experienced and motivated. In order to have more general and higher quality synthesis technology it will be desirable (1) to improve the phonetic quality of synthetic speech to the point where it is as easily comprehended as natural speech and where it is fully acceptable to naive or unmotivated listeners, (2) to use expressive variation appropriately to convey the structure and relative importance of information in complex materials, and (3) to model the speech of people of different ages, sexes, and dialects in order to support applications requiring use of multiple voices.

Engineers working on recognition have a long-standing goal of building systems that can handle large-vocabulary continuous speech. To be useful, such systems must be either speaker-independent or speaker-dependent; if speaker-dependent, engineers must be trained using a sample of speech that can feasibly be collected and analyzed. Present systems exhibit a strong trade-off between degree of speaker independence on the one hand and the size of the vocabulary and branching factor in the grammar on the other. As the vocabulary size increases, the extent of the acoustic differences between words decreases, on the average, and it becomes more likely that productions of different words by different speakers will be confused with each other. Similarly, the more words that are grammatically possible at any particular point in the sentence, the greater the risk of confusion. Even speaker-dependent systems are far from the desired level of generality. In addition, systems will need a vastly enhanced ability to recover and manipulate semantic and pragmatic information. In understanding

I would like to thank Alex Waibel and Stephen Levinson for stimulating discussions concerning the capabilities of HMM recognizers.

the speech of other people, we make many leaps of inference, as in the following examples:

1. I'm parked on 52nd St. (I == my car)
2. Can you tell me who is authorized to sign a PO? [= Please tell me . . .]

It is widely acknowledged that systems that cannot make such inferences will strike users as maddeningly literal-minded in all but the simplest exchanges.

In view of these goals, I would like to highlight two strategic problems presented by prosody and intonation as they function in speech. The first is the challenge that the allophonic effects of prosody and intonation present for training procedures. These are the procedures whereby a representative sample of utterances is collected and analyzed in order to construct a statistically optimal model of all the utterances that a speech system will handle. The second is the problem of formalizing what prosody and intonation mean. It is clear that human listeners can use prosody and intonation to make inferences about a speaker's goals and attitudes, and can use these inferences to make their own conversational contributions appropriate and useful. Machines will not be able to do this until a more explicit theory of the meaning of prosody and intonation is discovered.

Section 9.2 discusses training procedures in relation to both synthesis and recognition. It will summarize the present state of theory in the representation of prosody and intonation, and provide examples of allophonic effects that would pose a problem for present training procedures. Section 9.3 will turn to the issue of intonational and prosodic meaning.

9.2 Prosody, intonation, and allophony

9.2.1 Training procedures and why they are an issue

The success of Hidden Markov Models (HMMs) in speech recognition demonstrates the power of effective training procedures. HMMs use a transition network (or a hierarchy of such networks) to describe the utterances they can recognize. An introduction to the method is provided in Rabiner and Juang (1986). Jelinek (1985) discusses how it has been applied in a 5,000-word vocabulary recognition system at IBM. Levinson (1985) develops the mathematical relationship between HMMs, template matching, and stochastic parsing.

As Levinson points out, some linguistic regularities, such as coarticulation across word boundaries, are systematically omitted by HMMs. Nonetheless, HMMs outperform systems that are based on attempts to implement linguistic theory without using a statistical training method. This fact indicates the power of statistical training. However, the property of HMMs that makes them easy to train – the assumption that transitional probabilities are statistically independent – also limits their ability to capture generalizations. Consider what happens in

the vicinity of a single node in the network – for the sake of concreteness, let us say that this node is a spectral section representing the noise burst for /t/. The network can have a single such node if what can follow the burst does not depend on how it was reached. If there is such a dependence, then the network must have more than one /t/ burst node, and must segregate from each other the sequences of transitions that exhibit the dependence.

In fact, however, these dependencies are the norm. The theory of prosodic representation in phonology, as sketched in Section 9.2.2, is based on the finding that different levels of grouping each control both aspects of phonological well-formedness and details of pronunciation. For example, we find effects of syllable structure, word structure, and intonational phrasing. HMMs implicitly model the objective consequences of syllable and word structure by constructing a separate model for each word, as pointed out in Levinson (1985). This is one reason for their success. However, effects that encompass more than one word are not modeled. Any statistically important effects that are not effectively modeled can be expected to degrade the overall performance. Although the tremendous redundancy of some domains (e.g., connected digit recognition) can make this loss of information affordable, systems handling large vocabulary continuous speech will need to exploit the available information as effectively as possible. Furthermore, the implicit treatment of word level prosodic effects does not lead to the same efficiency in representation and training that a more explicit treatment might make possible. For example, a human listener who observed a velar fricative in the word "foggy" would be able to infer that other words with an intervocalic /g/ in a falling stress environment could also be pronounced with a fricative. An HMM system would need to acquire this information for each word separately. For large-scale systems, the ability to make relevant generalizations across the possible variants of different words may prove crucial to efficient training.

It is not widely recognized that training is also a central issue for progress in speech synthesis by rule. The text-to-speech systems we now have reached commercial potential only after many years of work by speech scientists, and companies like AT&T, DEC, and Infovox all have large development teams devoted to improving their speech synthesis and incorporating it into applications. Of course building a system for the first voice is the hardest, and the creation of comparable systems for additional speakers, dialects, or languages is considerably expedited by the lessons learned in developing the first, and by the feasibility of adapting a considerable portion of the software. This fact has indeed been demonstrated both by the multiple voices available for the DEC synthesizer, and by the rule compilers described in Hertz (1982) and Granstrom and Carlson (1986). The Granstrom and Carlson compiler has supported the relatively rapid development of commercial synthesis systems for many languages. Nonetheless, a considerable amount of work in descriptive phonetics is involved in the creation of each new synthetic voice, even at the current state-of-the-art level of

quality. The work will be considerably greater as we aim for fully fluent and natural quality. This is the case because voices, dialects, and languages differ from each other in every aspect of the sound structure, not in just some particular such area as the phoneme inventory or the pattern of coarticulation between adjacent phonemes.

Let us consider some examples. As Fourakis and Port (1986) have shown, the detailed timing of the nasal-fricative transition in words like "tense" differs according to dialect; for American speakers the velum closes before the release of the tongue, but not for South Africans. American speakers ordinarily flap the /t/ in words like "butter", but others aspirate. Words such as "aluminum" and "elementary" have different phoneme sequences and stress patterns in British and American speech. The Received Pronunciation, Anglo–Irish, and Scots dialects of English differ in their phrasal intonation (Bolinger, 1989). In Pierrehumbert and Talkin (in press), one subject used a generally breathy voice following the main stress of an utterance, whereas the other used a creaky voice. This difference in overall voice quality in turn had ramifications for the segmental allophony. Pitch range and voice quality are used conventionally by speakers as markers of social identity; Finnish men favor a gravelly voice quality and Russian women in positions of authority use a much higher pitch range than their American counterparts.

Building speech synthesizers that incorporate such differences amounts to creating a comprehensive quantitative model of the sound structure for each voice. This is something that phoneticians have not yet accomplished even once; Klatt's model of his own voice, as incorporated in his synthesis rules, may be considered the most complete such model to date. To achieve this goal, an alternative must be found to carrying out innumerable phonetics experiments, each involving innumerable measurements. In short, it will be necessary to find some way of acquiring quantitative descriptions semiautomatically. This will mean finding ways to use a large sample of speech to set the parameters of a general phonetic and phonological model. Let us now give some idea of what such a model looks like.

9.2.2 *Sound structure and its phonetic correlates*

A traditional view of sound structure contrasts speech segments with suprasegmentals. The string of segments arises from a sequence of local paradigmatic distinctions (for example, the contrast between "pat" and "bat" or between "pat" and "pad"), is taken to be phonetically expressed in properties of the speech spectrum. All nonlocal distinctions, whether syntagmatic or paradigmatic, are grouped together as suprasegmentals. For example, suprasegmentals are taken to include both stress (which is a syntagmatic feature since it describes the relative strength of syllables) and the paradigmatic distinction between rising

and falling phrasal melodies. The phonetic domain of the suprasegmentals is taken to be fundamental frequency (f0), amplitude, and duration.

Although this view is implemented in current text-to-speech systems and underlies proposals for the use of prosody and intonation in speech recognition (Lea, 1980; Waibel, 1988), it is not supported by the results of research in linguistic phonetics. On the one hand, effects of segment type on f0, amplitude, and duration are both substantial and well established (see Lehiste, 1970; Steele, 1985; Silverman, 1987; and literature reviews in these works). On the other hand, prosody and intonation have large effects on the speech spectrum. The effects of syllable structure (the smallest unit of prosodic structure) are particularly well accepted. Randolph (1989) shows position within the syllable to be a stronger statistical predictor of stop allophony than the local phonemic context. In addition, experiments have demonstrated various spectral effects of stress, phrasing, and intonation pattern (Harris, 1978; Monsen, Engebretson, and Vemula 1978; Gobl, 1988; Pierrehumbert, 1990; Pierrehumbert and Talkin [in press]).

Therefore, the view of sound structure that will be adopted here does not contrast segments with suprasegmentals. Instead, it draws a contrast between content and structure. Content, which is taken to cover all paradigmatic distinctions whether local or nonlocal, is phonetically expressed in terms of relative positions along dimensions of articulatory control, and therefore in the corresponding acoustic parameters. Structure covers the grouping and strength of elements of the content. It has an indirect phonetic expression, by influencing the extent and timing of articulatory gestures, and consequently by helping to determine which particular values of acoustic parameters realize each paradigmatic distinction in each particular case.

As an example of "content", consider the contrast between /p/ and /b/. It involves a contrast in laryngeal articulation. Phonetic parameters reflecting this contrast include the spectrum during the stop gap and right after the release, the duration of the voiceless region following the release, and the f0 when the voicing begins. Similarly, the intonational contrast between Low and High tone (L and H) (which is found, among other places, in the difference between a terminal declarative and one with a continuation rise) also involves a contrast in laryngeal articulation. The laryngeal articulation is reflected not only in f0 but also in the source spectrum, and therefore in the speech spectrum.

Intonational phrasing is an example of "structure". That is, an intonation phrase specifies that certain words are grouped together, and that one of these words is the strongest prosodically. The intonation phrase does not in and of itself specify any part of the content. Rather, it provides opportunities to make choices of content. For each intonation phrase, the speaker selects not only the words, but also a phrasal melody. The phonetic manifestations of the various elements of the content depend on their position with respect to the intonational phrasing. It

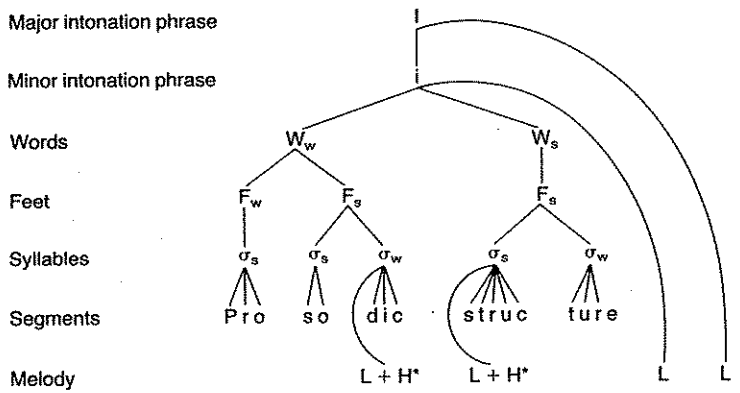

Figure 9.1. A phonological representation for the phrase "Prosodic structure".

is well known that syllables are being lengthened before intonation phrase boundaries. Phrasing affects the actual f0 values achieved for L and H tones; H tones marking phrase boundaries are subject to a different scaling rule than phrase medial H tones (Pierrehumbert, 1980). Pierrehumbert and Talkin (in press) also report that intonational phrasing affects the voice onset time for stops, even when there is no pause at the phrase boundary.

Figure 9.1 illustrates how content and structure are represented in a modern phonological framework. Details of the representational scheme are taken from Pierrehumbert and Beckman (1988). The figure shows the representation for the phrase "Prosodic structure", produced with a particular declarative melody that is often used to present information in a contrastive light. Two streams of content are produced simultaneously: the segments arising from the word choice and the tones comprising the melody. These two streams are coordinated by their links to a hierarchical structure with well-defined levels: the syllable, the foot, the word, and minor phrase, and the major phrase. Many phonological theories advocate more levels, such as a subsyllabic unit of the "mora" (Hyman, 1985; McCarthy and Prince, 1990) or a level of the "Phonological phrase" above the word but below the minor phrase (Selkirk, 1984), but the exact number of levels will not be important here. Each node at a given level dominates one or more nodes of the next lower level; one node within each grouping is singled out as the strongest or most prominent one. For example, each foot begins with a strong syllable. The strongest foot in the word is the one containing the main stress of the word. The timing of the tones with respect to the phoneme string follows from which syllables they are linked to.

The phonetic expression of any particular element of the content depends on what it is, what the neighboring content is, and its relation to the prosodic structure. All levels of prosodic structure, even the highest, are demonstrated

experimentally to play a part in controlling details of pronunciation. The discussion here will concentrate on effects of prosodic structure that are both nonlocal and gradient, because these present a strategic problem for future training methods. I view this problem as the central one, because present technology has a certain level of success in handling effects that are gradient but local, or nonlocal but qualitative. Local gradient effects are presently handled by encoding detailed phonetic properties in detailed whole-word models. This method performs well for systems of small to moderate vocabulary size, although it has a poor ability to represent or infer generalizations across words, as discussed above. It also neglects local effects that cross word boundaries. Hierarchically organized networks can handle the type of nonlocal qualitative constraints that arise from, e.g., sentence grammar; see discussion in Jelinek (1985) and Levinson (1985). In this approach, the nodes of one network each represent networks at a lower level. One network may represent possible sentences by specifying which words can follow each other; the word nodes are then expanded into detailed acoustic networks in which each node is a spectral section. Although it has not yet been done, nothing in principle prevents the same approach from being applied to prosodic trees.

However, there is at present no obvious method of combining these two approaches (detailed acoustic models and network layering) to handle nonlocal gradient effects. Hierarchical organization of networks is intrinsically qualitative. On the other hand, whole-phrase templates (that is, statistically adequate acoustic representatives of all possible phrases the system might encounter) would be prohibitively large and numerous for anything but extremely limited systems. Even a system that recognizes telephone numbers will be built on the assumption that telephone numbers can be decomposed down to the word level. Using separate templates for all intonation phrases in telephone numbers (that is, all possible sequences of three to four digits) would mean recording and processing a statistically significant sample of each of more than 10,000 items, and then evaluating these as unrelated alternative candidates during recognition. Such an approach would be out of the question for a domain of moderate complexity, permitting, say, phrases of one to five words constructed over a vocabulary of 1,000 items.

At the same time, it is clear that a general treatment of nonlocal gradient effects would subsume local or qualitative effects as subcases. An approach that handles nonlocal dependencies can view local dependencies as particularly small nonlocal dependencies. Similarly, qualitative effects can be viewed as instances of gradient effects that are restricted to just a few values.

9.2.3 *Some nonlocal effects of prosody and intonation*

In this section, we discuss three examples in which prosody and intonation affect the pronunciation of speech segments. In the examples, the structure of the whole phrase determines how some particular portion of it is pronounced.

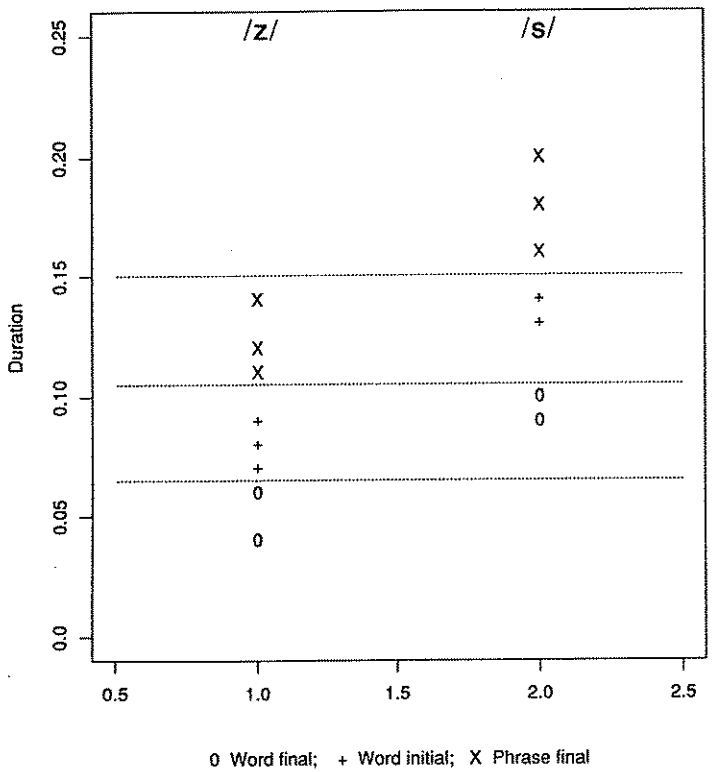

Figure 9.2. Durations of /s/ and /z/ in an illustrative set of utterances from one speaker. The phonemes are found word-initially, word-finally (but not phrase-finally), and phrase-finally.

The first example is based on a small data set that was collected for illustrative purposes. In the data set, produced by a single speaker, the phonemes /s/ and /z/ occurred word-initially, word-finally (but not phrase-finally), and phrase-finally. Figure 9.2 shows the durations of the /s/s and /z/s, with different plotting characters used for the different prosodic contexts. As is evident in the figure, there is a substantial overlap between the /s/ durations and /z/ durations when prosodic position is ignored. However, in each individual position, the /z/s are shorter than the /s/s. This is a typical illustration of the concept of "relational invariance" discussed in Fant (1987). Phonetic properties of phonemes are much better separated statistically when context (including both neighboring content and prosodic position) is taken into account than when it is ignored. Even when context drastically shifts the phonetic realizations of both members of a contrasting pair, the paradigmatic contrast between the two is still usually expressed.

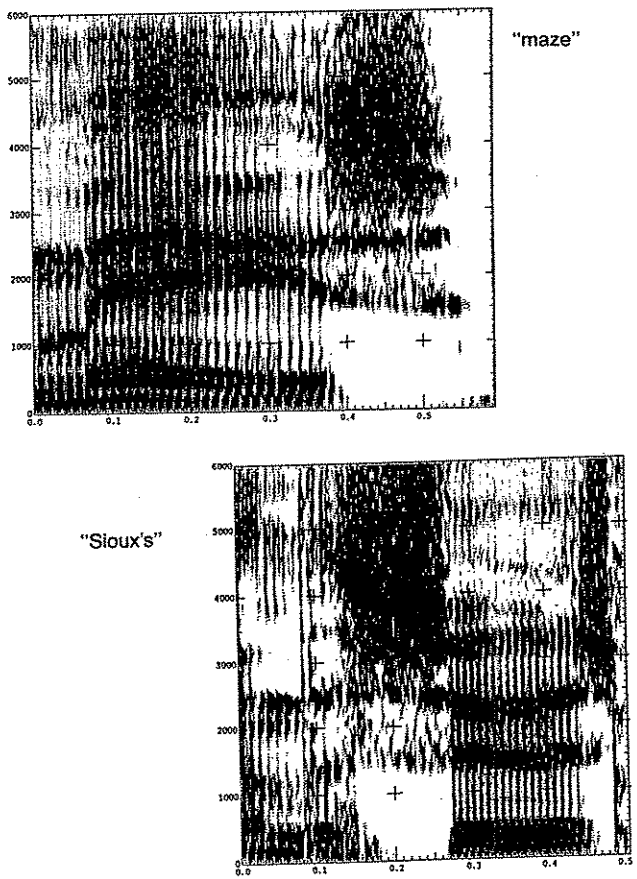

Figure 9.3. Spectrograms for /s/ in "Sioux's" versus phrase-final /z/ in "maze".

A critical phonetician might question whether the duration measurements in Figure 9.2 give a misleading impression by ignoring differences in the spectrum of /s/ and /z/. Figure 9.3, comparing phrase-final /z/ in "maze" with word-initial /s/ in "Sioux's", indicates that spectral characteristics for the two phonemes overlap statistically just as durations do. In particular, in these particular utterances, both fricatives have exactly two pitch periods of voicing, because the /z/ is typically devoiced phrase-finally.

The second example concerns the effects of intonation on vowel spectra. Figure 9.4 displays LPC spectral sections for the schwa in "tuba", spoken by a single speaker in a single recording session, in the middle of a declarative

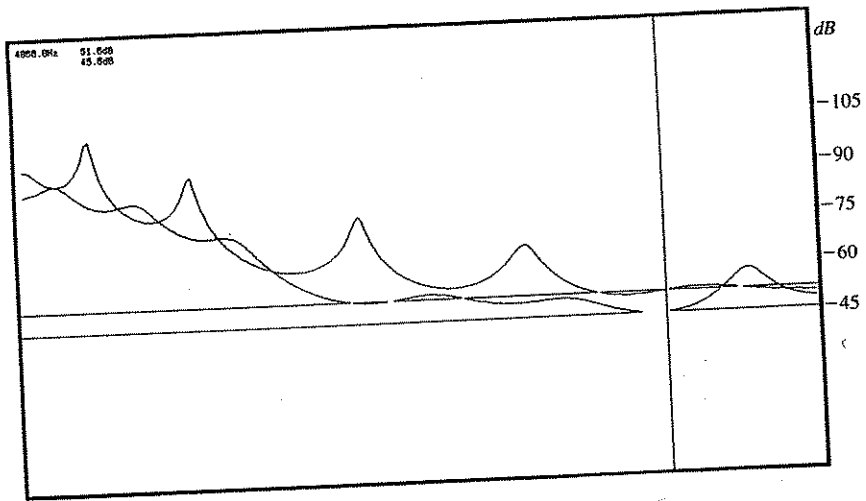

Figure 9.4. LPC spectra for the schwa in "tuba", occurring phrase-medially in a declarative and phrase-finally in a question. The same speaker produced both utterances.

sentence versus at the end of a yes/no question. The sentence intonation is related to an extremely high f0 value at the end of the question, as well as to a soft and breathy voice quality. Laryngeal models (Ishizaka and Flanagan, 1972; Titze and Talkin, 1979) in combination with analytical studies of the acoustic consequences of source variation (Ananthapadmanabha, 1982; Ananthapadmanabha and Fant, 1982) lead to the prediction that the source characteristics at the end of the question should raise the formats above the values in the declarative, in addition to affecting the bandwidths and overall spectral shape. As the figure shows, this effect can be quite substantial, indeed every bit as great as formant differences that distinguish vowels. In addition, the spectral prominence of the fundamental in the high breathy voice creates the potential for confusing it with a formant.

The pronunciation of /h/ in continuous speech provides a third example. The data given here are drawn from Pierrehumbert and Talkin (in press). Their experimental materials varied the position of /h/ relative to the word prosody (e.g., "hawkweed" vs. "Omaha"), and also the position of the target words relative to the phrasal prosody. In (3), the target word "hawkweed" has the main stress of the sentence, whereas in (4) it follow the main stress that is on the state name because of the constrastive stress.

3. Is it Oklahoma hawkweed?
4. It's Alabama hawkweed, not Oklahoma hawkweed.

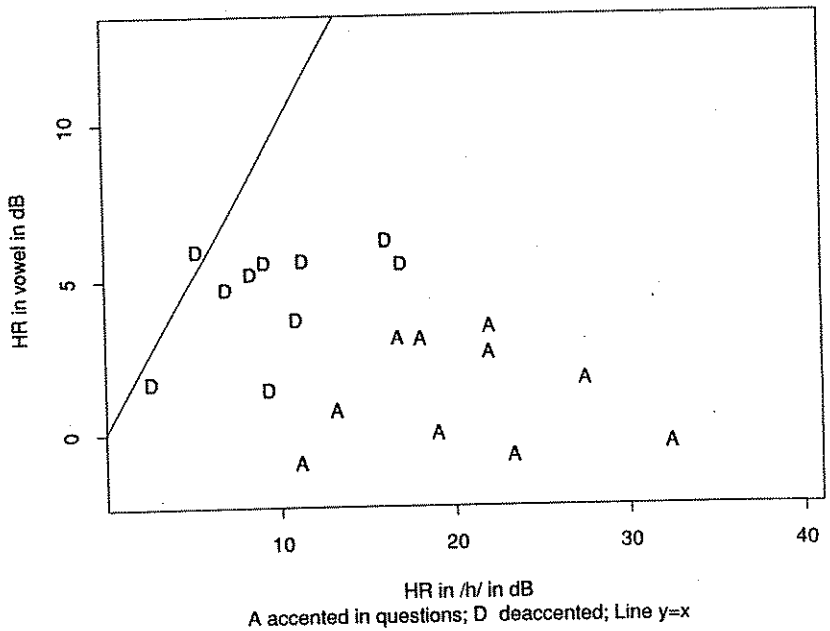

Figure 9.5. HR in "hawkweed" and "hogfarmer," subject DT. Harmonic ratio in /h/ plotted against harmonic ratio in the following vowel, for word-initial /h/. Plotting characters contrast the cases where the target word is accented in a question and where it is deaccented following a focused word.

Note that both (3) and (4) have low tones at the target location. This was an important aspect of the experimental design. By keeping the fundamental frequency below one-third of the first formant value, it was possible to obtain some indications of the source characteristics (or characteristics of the glottal waveform) without inverse filtering, a problematic procedure for breathy sounds.

The measure whose behavior is plotted in Figures 9.5 and 9.6 is the harmonic ratio, defined as the difference on dB between the energy at the fundamental and the energy at the next harmonic. This is an index of the degree of vocal fold spreading, and is expected to be greater for more /h/-like sounds and less for more vocalic sounds. In the figures, the harmonic ratio during /h/ is plotted against the ratio during the following vowel. A diagonal line in each figure represents $y = x$, or the case in which the vowel and the /h/ have the same value and are accordingly neutralized as seen through this measure. The degree of contrast between the /h/ and the vowel can therefore be related to the perpendicular distance from this line. A line perpendicular to the $y = x$ diagonal is also drawn for reference.

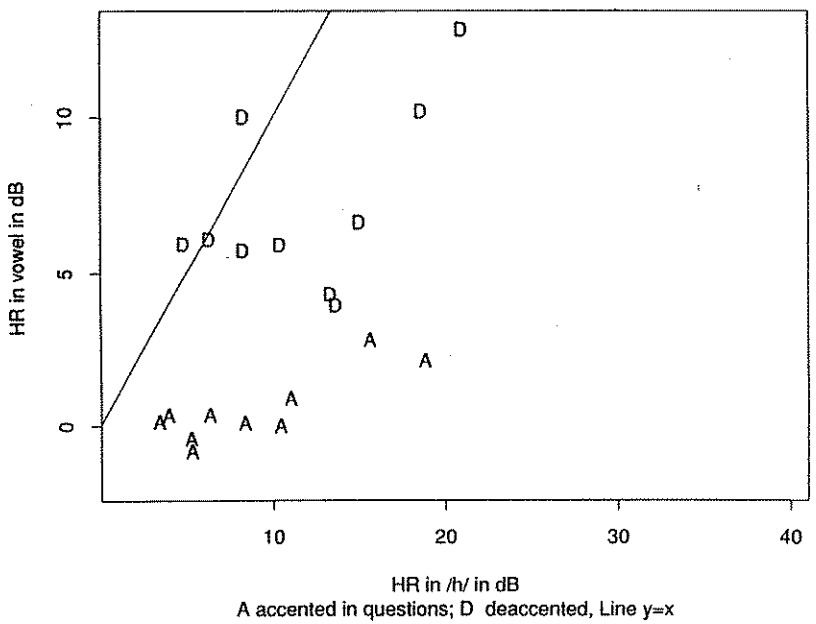

Figure 9.6. HR in "tomahawk" and "Omaha," subject DT. Harmonic ratio in /h/ plotted against harmonic ratio in the following vowel, for word-medial /h/ beginning a syllable without main word stress. Plotting characters contrast the cases where the target word is accented in a question and where it is deaccented following a focused word.

Figure 9.5 shows the outcome for stressed word-initial /h/ ("hawkweed" and "hogfarmer"), contrasting the cases where the word is accented in a question such as (3) and the cases where it is deaccented following a focus, such as (4). The sentence prosody has a major effect on the degree of contrast between the /h/ and the following vowel, with the accented tokens showing the consequences of stronger articulation, especially of the /h/.

Figure 9.6 shows the outcome for word-medial /h/ when it begins a syllable with subordinate stress or no stress ("tomahawk" and "Omaha"). Again, a contrast is found between accented and deaccented words. However, here the main effect of accent is to shift production of both the /h/ and the vowel in a more vocalic direction (toward the lower left corner of the graph), without a substantial effect on the degree of contrast between the /h/ and the vowel.

The comparison of Figures 9.5 and 9.6 shows that word and sentence prosody interact to determine how /h/ is pronounced. The combined data also have a nontrivial amount of overlap, in the range of 0 to 10 dB, between HR values for /h/ and those for vowels. However, the /h/s and the vowels are well distinguished from each other in context, as indicated by the fact that only a few data points fall on or above the y = x line.

The phoneme /h/ was chosen for study by Pierrehumbert and Talkin (in press) because characterizing its source was a relatively tractable problem, in comparison to phonemes that have an oral constriction as well as a distinctive laryngeal articulation. However, both the phonetics literature and informal observation indicate that prosodic structure has both widespread and large influences on source characteristics. For voiceless stops, the alternation between aspirated and glottalized variants under the control of syllable structure and stress is well known. The quantitative extent of such effects depends on the phrasal prosody; for example, Pierrehumbert and Talkin also report that the voice onset time for /t/ (in "tomahawk") is approximately doubled at an intonation phrase boundary even in the absence of a pause. Voiced stops range phonetically from voiceless unaspirated stops to voiced fricatives, with /d/ even having a sonorant allophone, the flap. Similarly, the weak-voiced fricatives can be produced as stops in a strong position and as sonorants in a weak one. Many speakers have an overall shift in voice quality after the main stress, with some adopting a breather quality and others a creaky one. In addition, in utterance final position we find a reduction in subglottal pressure and a tendency toward devoicing.

9.2.4 *Consequences for speech technology*

The examples discussed in Section 9.2.3 indicate that it is impossible to recognize the speech segments without recognizing the prosody and intonation. It is also impossible to recognize the prosody and intonation without recognizing the segments. For example, what counts as long /z/ would count as a short /s/. Thus a judgment about whether a particular fricative region was in phrase-final position (and had accordingly undergone phrase-final lengthening) would depend on what phoneme that region was taken to represent.

This situation causes some speech engineers to throw up their hands, asking "Which comes first, the chicken or the egg?" This sense of being at an impasse has its source in the assumption that speech processing must recover some aspect of the representation first, or bottom-up. A considerable effort has been put into analysis schemes that attempt to carry out bottom-up classification robustly; that is, without falling into the confusions or errors that can readily arise from the statistical overlap of phonemes taken out of context.

Note that there is no impasse as far as the human mind is concerned; it apparently recognizes the phonemes and the prosody together. Furthermore, in other areas we actually have a technology for knitting together local information into a coherent overall structure, and that is parsing. One forte, indeed a raison d'être, of parsing technology is its ability to handle nonlocal dependencies. For example, sentence (5) is ill-formed as it stands, but well-formed in the larger

context in (6), and any serious proposal about sentence parsing provides a mechanism for dealing with this dependency.

5. *You put in the basket.
6. The pie that you put in the basket was delicious.

What we need to do for speech, then, is to develop parsers that can handle the observed nonlocal dependencies. The relevance of parsing technology to sound structure has already been established by Church's work on parsing syllable structure from a fine phonetic transcription, as discussed in Church (1983) and (1987). This important demonstration has been recently followed up by Randolph (1989), whose syllable parser takes as input a manually defined collection of predicates on the speech signal. The parsers we need would work from parameterizations of the speech signal that are automatically computable, rather than manually specified. They need to handle all levels of prosodic structure, as they interact with each other, instead of only the lowest level. In addition, they need to be trainable; that is, we need methods using a transcribed corpus to set the statistical parameters of a general grammar.

9.3 Prosody, intonation, and meaning

The example of a phonological representation given in Figure 9.2 above showed two streams of content, the phonemes and the intonation pattern. This section describes what intonation is like and how it relates to sentence prosody. It sketches what kind of information intonation and sentence prosody convey. The sketch will give an idea of what theoretical problems must be solved before a computationally tractable formal treatment becomes available.

9.3.1 The English intonation system

English has a large variety of different intonation patterns. The same words can be produced with many different patterns, with different semantic and pragmatic meanings. Figure 9.7 shows f0 contours for a few different renditions of the phrase "another orange". The patterns are labeled according to the transcription system developed in Pierrehumbert (1980) and modified in Beckman and Pierrehumbert (1986). The patterns are made up of pitch accents (which align with stressed syllables) and boundary tones, which mark phrasal edges regardless of stress. The pitch accents can consist of one or two tones, and the diacritic * is used to mark the tones that fall on the stress; in a bitonal accent the unstarred tone falls in the immediate vicinity of the starred tone, either on the same syllable or on a nearby one. A stressed syllable will lack a pitch accent if it belongs to a word that is not prominent in the phrase (for example, because it contains given information), but each phrase must have at least one pitch accent somewhere. Boundary tones are assigned for two levels of phrasing, which coincide in the

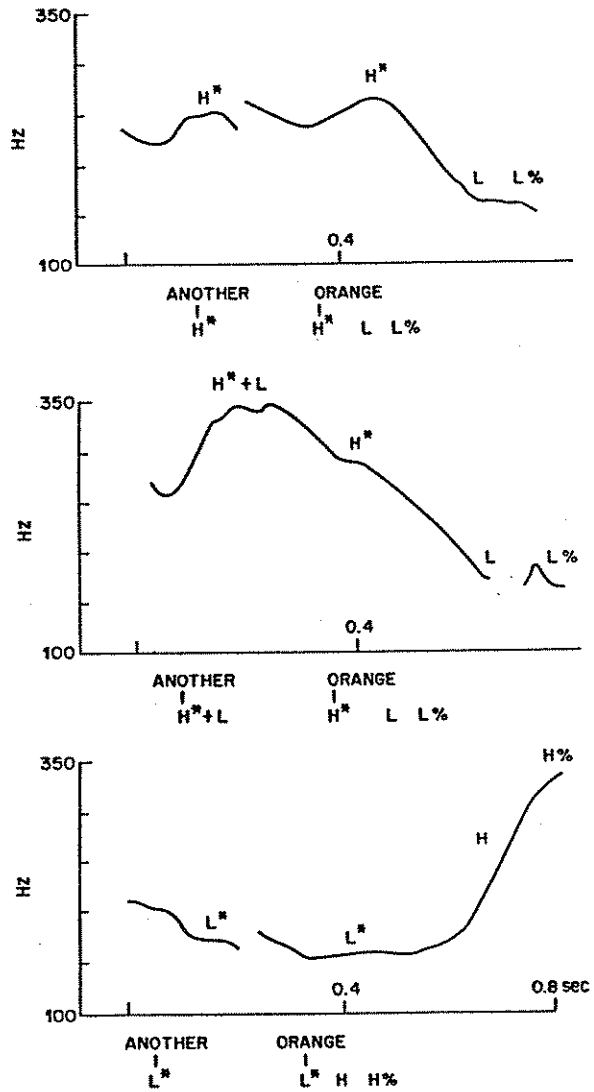

Figure 9.7. F0 contours for the phrase "another orange", produced with a variety of intonation patterns. Transcriptions are according to the system developed in Pierrehumbert (1980) and Beckman and Pierrehumbert (1986).

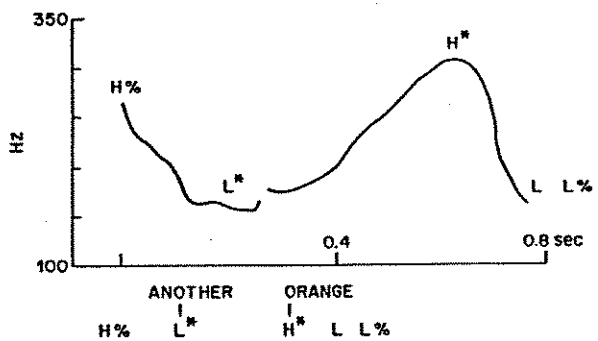

Figure 9.7. *(cont.)*

very simple materials in the figure. The diacritic % marks the boundary tone for the stronger phrase boundary (the intonation phrase boundary).

As the figure indicates, the pitch accents in a phrase can differ from each other. Furthermore, most of the possible combinations of pitch accents and boundary tones are attested. A compositional treatment of intonational meaning developed in Pierrehumbert and Hirschberg (1990) relates these different choices to different elements of pragmatic meaning.

It is important to distinguish what the intonation pattern is from where it goes. The same pattern can be aligned with the words in different ways, depending on the phrasing and on the phrasal stress as influenced by focus. Figure 9.8 shows the H* L H% pattern (a declarative intonation with a continuation rise) assigned to the same text in two different ways. In the first, the main stress falls on "vitamins"; in the second it falls on "Legumes". This is a difference in sentence prosody – the grouping and prominence of words in the sentence – but not a difference in intonation. The second version might arise in a dialogue that establishes the later words in the sentence as given information, e.g.,

7. Tell me some good sources of vitamins.
8. LEGUMES are a good source of vitamins.

9.3.2 Focus

Figure 9.8 brought out the phonological distinction between what the intonation pattern is and where it goes. This phonological distinction goes along with a distinction in the type of meaning conveyed. This section discusses meaning differences related to differences in the location of pitch accents. Discussion of what intonation patterns proper mean will follow in Section 9.3.3.

In the linguistics literature, words like "LEGUMES" in example 7 are described as focused. These elements are marked with pitch accents and other

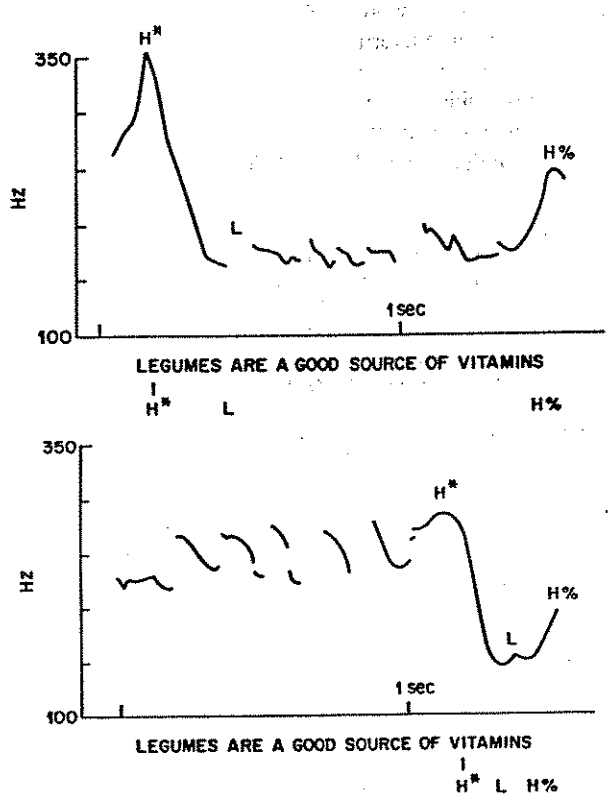

Figure 9.8. F0 contours for a declarative pattern with a continuation rise, produced on a sentence with the main phrase stress in two different places. Stress on "legumes" would be appropriate if previous context had already introduced the rest of the information in the sentence.

phonetic manifestations of emphasis, such as long and fully produced stressed vowels. In the natural language processing literature "focus" is used in a different sense. Discourse entities are taken to be in focus if they belong to a set of entities that have been (directly or indirectly) evoked in the discourse, and that are presumed to be salient for the conversational participants. (See Grosz and Sidner, 1986.) Because the words for already salient discourse entities are often (though not necessarily) deaccented, these two differing usages can lead to confusion. I will use "focus" in the sense it is used in linguistics.

Focus has a conspicuous relationship to the distinction between old and new information, with new information tending to receive focus. A heuristic developed by Silverman for assigning pitch accents in a text-to-speech system brings out this relationship while revealing its limitations. The system maintained a

buffer of the forty most recent content words (modulo affix stripping). Content words on the list were deaccented. Silverman also states that function words were deaccented everywhere, but the reader infers from his examples that this rule was restricted to monosyllabic function words, with others treated like content words. In the following text, capitals indicate accented words, lowercase deaccented words, except that italics are used to mark deaccented content words.

9. In the MINDS of MANY PEOPLE, the EFFECTS of AGING are REGARDED as NEGATIVE. At BEST, OLDER people are SEEN as MORE CAREFUL and as having *more* EXPERIENCE and KNOW-HOW, but RESISTANT to CRITICISM and INSTRUCTION, and SLOW to LEARN. At WORST, ANYONE OVER THIRTY is seen as LESS MENTALLY ABLE, less HEALTHY, and DECLINING GENERALLY. CURRENT SCIENTIFIC OPINION is FAR less PESSIMISTIC. IT is TRUE that TESTS SHOW a *decline* in *mental ability* as *people* GROW *older*, but this *decline* is NOT SIGNIFICANT until they REACH their MID FORTIES or EVEN EARLY SIXTIES.

Although the beginning of this passage is acceptable, a very unfortunate series of deaccentuations occurs toward the end as more discourse entities are evoked. Note that the second instance of the word "decline", for example, could readily be accented even though it repeats a word in a prior clause in the same sentence. Many examples indicate that repeated reference is not enough to trigger deaccenting, if the role of the referring expression changes. Example (10), a classic, is due to Lakoff (1971).

10. John called Bill a Republican, and then HE insulted HIM.
11. Now we have to turn left.
    – No, we turn RIGHT rather than LEFT.

However, it is unclear exactly what changes in role can cause an expression to be accented even when it is not new. Hirschberg and Ward (in press) propose that the pronouns in (10) are accented because they appear with new case features; although Bill is already mentioned in the discourse, he is new as an AGENT. This suggestion appears to be a step in the right direction, because it takes focus not to apply to words per se, but rather to words as they function in sentences and discourses. However, Hirschberg and Ward have not empirically evaluated or even proposed a treatment of all cases of accent on previously mentioned words. In addition, a formal theory is needed for cases in which words not previously mentioned are deaccented because they are sufficiently inferable.

Focus also contributes to meaning by affecting the scope of adverbs, as in the following examples analyzed in Rooth (1985).

12. a. I only said that Carl likes HERRINGS.
    b. I only said that CARL likes herrings.
13. a. A u usually follows a Q.
    versus
    b. A U usually follows a q.

(12a) is false if I said that Carl likes some other type of fish; (12b) is false if I said that someone else likes herring. Although (13b) is true, (13a) is false because most u's follow some other letter than q. Rooth develops a technical treatment of such phenomena within the framework of Montague semantics. However, this treatment does not handle the phenomena sketched in (9) through (11), and a comprehensive formal treatment of focus is still not available. Filling this need will be an important step toward the creation of computer programs that either provide or understand prosody.

9.3.3 *Information conveyed by tunes*

Listeners infer from the intonation pattern information about the speaker's attitude, his goals in making his utterance, and the type of speech act he intends to perform. For example, Sag and Liberman (1975) discuss cases in which intonation can fairly reliably serve to suggest that a syntactic yes/no question is really an indirect request.

However, further investigation has shown that the basic meanings of intonation patterns cannot be taken to be speaker attitudes or types of speech acts. In different contexts, the same pattern can be associated with different attitudes or speech acts and likewise different patterns can be associated with the same attitudes or speech acts. Therefore, this information must be inferred from basic meanings as they interact with context. The general character of the basic meanings for intonation patterns is brought out by two highly important case studies.

Ladd (1978) investigated the so called "vocative" pattern, which has an f0 peak on the stressed syllable and then falls to a mid value, sustained to the end of the utterance. This pattern is transcribed H*+ L H L% in Pierrehumbert's notation, and is often used for calling out to someone, e.g.,

14. Christopher! Your lunch!

Ladd shows that this pattern is not really a vocative. It cannot be used for calling out in a true emergency; it can be used on sentences other than vocatives. One example he discusses is

15. Watch out for the crevasse!

Mountain climbers warning of a possible emergency would hardly use the H*+L H L% pattern on this sentence; however, it might be used by the abominable snow mother giving her child a reminder as he departs for school. Ladd suggests that the basic or general meaning of the contour is to mark information that should be in some way expected or known.

Two studies by Ward and Hirschberg (1985, 1988) analyzed superficially diverse uses of the rise-fall-rise, or L*+H L H% contour. This pattern is commonly used to convey either uncertainty or incredulity, as in the second and third turns of the following conversation, respectively:

16. Did you take out the garbage?
    Sort of.
    Sort of!

The first study, dealing only with the "uncertainty" reading of the contour, proposes that the contour is appropriate when a scale of alternatives is available (because of the preceding material in the discourse or the utterance that carries the contour itself). The contour conveys uncertainty with respect to the scale. This uncertainty could be about whether the value of the accented item on the scale is appropriate, about whether the scale is the particular scale that is appropriate, or about whether a scale is appropriate at all. The second study unifies the "uncertainty" and "incredulity" readings by proposing that the contour conveys lack of speaker commitment with respect to the scale. It also demonstrates experimentally using LPC hybridization that the same melody is actually used in both cases.

This work provides strong evidence that different melodies do not directly convey truth, speaker attitude, or emotion. A person might use the contour on information that is true if some aspect of the social situation makes him uncertain about asserting it. On the other hand, on the "incredulous" reading, he could use it to mark information that he wishes to imply to be false. The first "sort of" in (16) is polite and even timid; the second, which uses the very same melody, is angry and assertive. Thus deductions about matters such as truth, anger, politeness, or assertiveness are made from the basic meaning of the contour and from all other information in the context, including the meaning of the words, the voice quality, and the relationship between the speakers.

Following up this approach in their compositional pragmatics for tune meanings, Pierrehumbert and Hirschberg (1990) claim that pitch accents in general convey information about how the accented item is related to the mutual beliefs being built up during the course of a conversation. Meanings for all six pitch accents are proposed. For example, the L* marks information whose status is in doubt:

17. Your name is Mark Liberman?
    L* H H%

It can also suggest that the information is already known; compare (18b) ("For your information . . . ") with (18c) ("As you should already be aware . . . ")

18. a. Let's order the Chateaubriand for two.
    b. I don't eat beef.
       H* L H%
    c. I don't eat beef.
       L* L H%

The L* is also used on information that is extrapropositional. One example is

The L* is also used on information that is extrapropositional. One example is the use of L* on "now" when used as a discourse marker. In Litman and Hirschberg's (1990) study of tapes from a radio talk show, they found that L* was much more common on instances of the word "now" used to mark organization of the discourse than it was on instances of the same word used as a temporal adverb. Postposed vocatives provide a second example of L* on extrapropositional information, according to the analysis in Beckman and Pierrehumbert (1986).

19. Your lunch is ready, Sam.
    L* L H%

We see from these examples that intonation conveys pragmatic information that has a central function in successful dialogue and is poorly marked in ordinary text. Because of this, analysis of intonation has the potential for playing an important part in systems that engage in dialogue with people. Any effective use of intonation for this purpose, however, will require substantial progress in circumscribing and formalizing the pragmatic deductions that human listeners perform with such virtuosity.

9.3.4 Will f0 assist word recognition?

Researchers who have little experience with intonation sometimes hope that, if fully exploited, it could resolve ambiguities that are presently found to be problematic. However, the sketch in Section 9.3.3 makes it clear that f0 cannot be expected to substantially assist word recognition in unrestricted running speech. This is the case because the melody is not provided by the lexical choice, but rather functions as a separate simultaneous channel of information.

In order to appreciate this point, consider the phonological reasons for a syllable to exhibit a high f0 value.

It is stressed and has a H* pitch accent.
It is in the domain of a H boundary tone.
It is in between two H tones.

Reasons for it to exhibit a low f0 value are:

It is stressed and has a L* pitch accent.
It is in the domain of a L boundary tone.
It is in between two L tones.

That is, any f0 value is phonologically compatible with any stress pattern, and

the primary determinant of f0 is the phrasal melody rather than any aspect of the word.

In view of this situation, the probability distributions for f0 in relation to stress presented in Waibel (1988) should not be unsurprising. They show that the f0 values for unstressed syllables are not in general very different than those for stressed syllables. F0 might provide useful information about stress in extremely restricted domains that effectively constrain the intonation patterns used.

9.4 Conclusion

This chapter has discussed two areas in which our present scientific understanding of prosody and intonation has ramifications for the future of speech technology. First, we considered the problems posed by the prosodic and intonational control of allophony, emphasizing the nonlocal long distance dependencies that pose the most serious challenges. The lesson from these examples is that future technology will need to integrate the representational insights of modern phonology with the effective training procedures now available only for HMMs. Second, we laid out some types of pragmatic information that are conveyed by intonation. Because this pragmatic information is both central to effective dialogue and poorly marked in text, intonation has the potential for playing an important role in interactive systems. However, considerable progress in formalizing pragmatic inference will be necessary to make this possible.

In the discussion of these two cases, speech recognition and speech synthesis have been treated together. From a strategic point of view, these fields face many of the same problems. Coordinating work more tightly in these areas would help assist progress in speech technology in general.

References

Ananthapadmanabha. (1982). "Truncation and Superposition," STL-QPSR 2-3/1982, 1–17, Royal Institute of Technology, Stockholm.

Ananthapadmanabha and G. Fant. (1982). "Calculation of True Glottal Flow and its Components," *Speech Communication*, 1, 167–184.

Beckman, M., and J. Pierrehumbert. (1986). "Intonational Structure in Japanese and English," *Phonology Yearbook 3*, 255–310.

Bolinger, D. (1989). *Intonation and its Uses*. Stanford University Press.

Church, K. (1983). *Phrase-structure Parsing: A Method for Taking Advantage of Allophonic Constraints*. Ph.D. Dissertation, MIT. Distributed by Indiana University Linguistics Club, Bloomington.

Church, K. (1987). "Phonological Parsing and Lexical Retrieval." In U. Frauenfelder and L. Tyler, Eds., *Spoken Word Recognition* (a Cognition Special Issue), MIT Press, Cambridge. 53–70.

Fant, G. (1987). "Interactive Phenomena in Speech Production," *Proc. of 11th International Congress of Phonetic Sciences*.

Fourakis, M. S. and R. Port. (1986). "Stop Epenthesis in English," *J. of Phonetics*, 14, 197–221.

Gobl, C. (1988). "Voice Source Dynamics in Connected Speech," STL-QPSR 1/1988, Royal Institute of Technology, Stockholm, 123–159.

Granstrom and Carlson. (1986). *Linguistic Processing in the KTH Multi-lingual Text-to-Speech System*, International Conference on Acoustics, Speech, and Signal Processing. 45(1)1–4.

Grosz, B. and C. Sidner. (1986). "The Structures of Discourse Structure," *Computational Linguistics*, 12(3).

Harris, K. (1978). "Vowel Duration and its Underlying Physiological Mechanisms," *Language and Speech*, 21(4), 354–361.

Hertz, S. (1982). "From Text to Speech with SRS," *J. Acoust. Soc. Am.*, 72(4), 1155–1170.

Hirschberg, J. and G. Ward. (in press). *Accent and Bound Anaphora*.

Hyman, L. (1985). *A Theory of Phonological Weight*. Foris Publications, Dordrecht.

Ishizaka, K. and J. L. Flanagan. (1972). "Synthesis of Voiced Sounds from a Two-mass Model of the Vocal Cords," *Bell Syst. Tech. Jour.*, 51, 1233–1268.

Jelinek, F. (1985). "The Development of an Experimental Discrete Dictation Recognizer." *Proceedings of the IEEE*, 73(11), 1616–1625.

Ladd, D. R. (1978). "Stylized Intonation," *Language*, 54, 517–540.

Lakoff, G. (1971). "Presupposition and Well-formedness." In *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*, 329–340. Cambridge Univ. Press.

Lea, W. A. (1980). "Prosodic Aids to Speech Recognition." In W. A. Lea, ed., *Trends in Speech Recognition*. Prentice-Hall, Englewood Cliffs.

Lehiste, I. (1970). *Suprasegmentals*. MIT Press.

Levinson, S. (1985). "Structural Methods in Automatic Speech Recognition." *Proceedings of the IEEE*, 73(11), 1625–1646.

Litman, D. and J. Hirschberg (1990). "Disambiguation Cue Phrases in Text and Speech." *Proceedings of COLING 90*, Helsinki.

McCarthy, J. and A. Prince. (1990). "Foot and Word in Prosodic Morphology: The Arabic Broken Plural," *Natural Language and Linguistic Theory*, 8, 209–238.

Monsen, R. B., A. M. Engebretson and N. R. Vemula. (1978). "Indirect Assessment of the Contribution of Subglottal Air Pressure and Vocal-fold Tension to Changes of Fundamental Frequency in English," *J. Acoust. Soc. Am.*, 64, 65–80.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*, MIT Ph.D. dissertation. Available from Indiana University Linguistics Club, Bloomington, Indiana.

Pierrehumbert, J. (1990). "A preliminary study of consequences of intonation for the voice source." In *Quarterly Progress and Status Report, Speech Transmission Laboratory*, Royal Institute of Technology, Stockholm.

Pierrehumbert, J. and M. Beckman. (1988). *Japanese Tone Structure, Linguistic Inquiry Monograph Series 15*, MIT Press, Cambridge.

Pierrehumbert, J. and J. Hirschberg. (1990). "The Meaning of Intonation Contours in the Interpretation of Discourse." In Cohen, Morgan, and Pollack, eds., *Intentions in Communication*. SDF Benchmark Series in Computational Linguistics, MIT Press, Cambridge.

Pierrehumbert, J. and D. Talkin. (in press). "Lenition of /h/ and glottal stop," in Ladd and Doherty (eds.), *Papers in Laboratory Phonology II*, Cambridge University Press, Cambridge.

Rabiner, L. and B. H. Juang. (1986). "An Introduction to Hidden Markov Models." *IEEE ASSP Magazine* (1) 4–16.

Randolph, M. (1989). *Syllable-based Constraints on Properties of English Sounds*. Ph.D. dissertation, MIT.

Rooth, M. (1985). *Association with Focus*. Ph.D. dissertation, U. Mass., Amherst.
Sag, I. and M. Liberman. (1975). "The Intonational Disambiguation of Indirect Speech Acts." *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*, Chicago Linguistic Society, University of Chicago. 487–497.
Selkirk, E. O. (1984). *Phonology and Syntax: The Relation Between Sound and Structure*. MIT Press, Cambridge.
Silverman, K. (1987). *The Structure and Processing of Fundamental Frequency Contours*. Ph.D. dissertation, Cambridge University.
Steele, S. (1985). *Vowel Intrinsic Fundamental Frequency in Prosodic Context*. Ph.D. dissertation, University of Texas at Dallas.
Titze, I. R., and D. T. Talkin. (1979). "A Theoretical Study of the Effects of Various Laryngeal Configurations on the Acoustics of Phonation," *J. Acoust. Soc. Am.*, 66, 60–74.
Waibel, A. (1988). *Prosody and Speech Recognition*. Pitman, London.
Ward, G. and J. Hirschberg. (1985). "Implicating Uncertainty; The Pragmatics of Fall-Rise Intonation," *Language* 61, 747–776.
Ward, G. and J. Hirschberg. (1988). "Intonation and Propositional Attitude; the Pragmatics of the L*+H L H%." *Proceedings of the 5th Eastern States Conference on Linguistics*, 512–522.

PART VI

Conclusion