On the Perception of Pitch Accents

by

Kevin L. Landel

B.S., Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, California
1983

SUBMITTED TO THE MEDIA ARTS AND SCIENCES SECTION
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

MASTER OF SCIENCE

AT THE MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1988

Signature of the Author

....

Kevin L. Landel
Media Arts and Sciences Section
June 6, 1988

Certified by

....

Chris Schmandt
Principal Research Scientist
Thesis Supervisor

Accepted by

....

Stephen A. Benton
Chairman
Departmental Committee on Graduate Students

On the Perception of Pitch Accents

by

Kevin L. Landel

Submitted to the Media Arts and Sciences Section on June 6, 1988 in partial fulfillment of the requirements of the degree of Master of Science at the Massachusetts Institute of Technology

Abstract

A series of experiments was performed to determine the thresholds at which perturbations to pitch contours became important to the semantic interpretation of utterances. The perception of pitch accents was controlled by manipulating the accents' relative magnitudes and surrounding prosodic environment. The perturbations investigated were: the separation of **H** and **L** pitch accents, the presence vs. the absence of pitch accents, and the effects of pitch range and prominence on the perception of pitch accents. The intent is to suggest the pertinent information which needs to be addressed for the implementation of speech systems that make better use of prosody.

Thesis Supervisor: Chris Schmandt
Title: Principal Research Scientist

This work was supported by the Defense Advance Research Projects Agency (DARPA) contract number N00039-86-C-0406.

For Tom

Contents

List of Figures

Part I

Introduction

Chapter 1

Motivation

We've all heard the phrase, "It's not what you say, it's how you say it." This bit of common sense is almost obvious, but proves to be extremely difficult to implement when constructing interactive speech systems. Many designers have sidestepped this issue of *prosody*, concentrating on *what* is being said over *how* it is being said, with the justification that "no prosody is better than bad prosody." If systems are to take full advantage of the power of communication inherent in speech, they need to be designed to handle more diverse prosodic contexts. Studies are beginning to show that presentation is just as important as content [18,5,2], and that presentation and content often cannot be separated.

Not all acoustic properties are important to the prosodic intent of an utterance. Some may be perceptible but meaningless, while others may go by unnoticed yet have a large effect on the semantics of an utterance. In order to include better prosodic rules, one needs to know which prosodic attributes are important, and to what extent they need to be manipulated to achieve a desired effect. A better understanding of the thresholds at which prosodic attributes become perceptually salient will help the designer of a speech recognition or generation device implement rules that take into account the important prosodic attributes, while ignoring those that lack significance. Perhaps Dennis Klatt says it best:

"If one makes a spectrogram of a sentence produced by a text-to-speech system, and compares it with a sentence read by the person whose speech formed the basis for system development, it is easy to see ways in which the two acoustic patterns differ. It is less easy to tell whether individual differences are perceptually important, ***but if one has some idea of discrimination limits, the perceptual salience of various speech cues,*** and the articulatory basis of acoustic discrepancies, then good guesses can be made as to the specific rules needed in the future..." [8] (emphasis mine).

This document describes a series of experiments which investigate the points at which perturbations to pitch contours become perceptually salient and affect the interpretation of the meaning and intent of an utterance. Using the framework for the phonology of pitch accents put forth by Pierrehumbert [15,16,13], pitch contours were manipulated in order determine the magnitudes at which pitch excursions become recognizable as pitch accents; a change in the perceived sequence of pitch accents indicates a change to the implied meaning of an utterance.

Chapter 2

Background

The discussions which follow rely heavily on the assumption of the reader's familiarity with the concept of *pitch accents* put forth by Janet Pierrehumbert [15]. In order to acquaint the reader with these concepts, a brief overview of the subject is covered here. This will be little more than a paraphrase of the excellent review given in [14]; for a more thorough treatment, the reader is directed to that text.

2.1 Intonation, pitch, and pitch contours

Before beginning with a discussion of the phonology of pitch accents, it is important to start with an understanding of the terms that are going to be used throughout this paper. The investigations described herein are investigations of *intonation*. Intonation is the use of pitch, duration and energy for prosodic effect, that is, to change the implied meaning of an utterance beyond what is meant by the lexicon alone [4].

Pitch is defined in the literature as the *perceived* the frequency at which a speaker's vocal cords are vibrating during voicing. The actually *measured* frequency of vibration is called the *fundamental frequency*, or F0. Pitch and F0 will differ in the cases where, for example, the vocal cords are not vibrating periodically, disrupting the determination of an F0, yet a fundamental frequency is still perceived, or where the measured frequency of vibration differs from that which is perceived [11,9]. For example, the use of "creaky-voice" at the end of an utterance is often still perceived with a "correct" pitch [10].

My use of the terms pitch and fundamental frequency will differ from that used in the literature. I will use these terms interchangeably to mean the fundamental frequency of vocal fold vibration. This will not cause any difficulties since the perceived frequency and the actual frequency will always be the same throughout each of the experiments described.

Finally, the variation of fundamental frequency over time is called a *pitch contour.* Pitch contours are often represented graphically.

This paper is concerned with perturbations to pitch contours, and the points at which these perturbations become large enough to affect the meaning or intent of an utterance. Pitch excursions that carry meaning can be represented by *pitch accents.* The theory of the use of pitch accents that will be used here was put forth by Pierrehumbert in [15].

2.2 Pierrehumbert's phonology

Pierrehumbert's theory of intonational phonology is based on the concept that semantic excursions in a pitch contour, that is, peaks and valleys that carry intonational meaning, are realized through the use of *pitch accents.*

2.2.1 Pitch accents, phrase accents, and boundary tones

Phonologically, pitch accents are comprised of one or a combination of two "tones" – **H** and **L**. **H** and **L** are only specified as relative to each other: that is, an **L** is lower than an **H** would be in the same context.

Pitch accents are always associated with stressed syllables, and must be aligned with them.[1] This is annotated with a star, as in **H*** or **L***. In the case where an excursion is too abrupt to be attributable to a single pitch accent, a *bitonal* is used. Bitonals are transcribed as **H+L** or **L+H**. Again, a star is used to denote the alignment of the accent with the syllable. For example, an excursion which reached a peak before the main stress of the syllable, but then dropped to a local minimum at the point of peak energy would be transcribed as **H+L***.

[1] Pierrehumbert defines this alignment as falling on the stress of the syllable. I take this to mean that the maximum or minimum pitch excursion occurs at the same time as the peak energy of the syllable.

A variation of the pitch accent is the ***boundary tone***. Boundary tones control the pitch at the onset and offset of a phrase. These are always comprised of a single tone, and is differentiated from a pitch accent proper by a percent sign: **H%** or **L%**. Their behavior is the same as for regular pitch accents except for alignment – the ***target pitch*** is reached at the end of voicing for phrase offset; at the beginning of a phrase, F0 at the onset of voicing begins at the target pitch.[2]

Finally, there is a pitch accent which has no alignment whatever. Called a ***phrase accent***, this tone governs the pitch between the last pitch accent proper and the final phrase boundary. The phrase accent is also always a single tone and, because of its floating nature, carries no "star".

The grammar of pitch accents

Any number of pitch accents can occur within an intonational phrase, but the minimum tonal sequence is one pitch accent followed by a phrase accent and then a boundary tone. The choice of pitch accents, phrase accents and boundary tones used is free. The behavior of the pitch contour between the accents and boundary tones is governed by the surrounding tones. The amount of interaction between tones is dependent upon their time-frequency separation, with the constraint that a given target is independent of succeeding tonal elements, but may be influenced by at most the immediately preceding one.

2.2.2 Prominence

Finally, the targets reached during the pitch excursions are subject to scaling by ***prominence***. Prominence accounts for differences in target pitches not attributable to the presence or absence of pitch accents, and not caused by segmental effects. For example, if a speaker wanted to use pitch to put extra emphasis on a word that already carried a pitch accent, she would scale its pitch by increasing the accent's prominence. Prominence is required by the phonology because pitch accents alone are not sufficient to specify the target of a

[2] Boundary tones at the beginning of phrases are rarely if ever notated. The pitch at the beginning of a phrase is often the same as that of the preceding phrase, and so is specified by the preceding phrase's final boundary tone.

pitch excursion. Utterances with the same pitch accent sequence may still have different pitch contours, and hence different meanings, because the target pitches on the accents are controlled by the attitudes of the speaker and the relation of the phrase to others in the discourse.

2.2.3 A clarification

At this point it is necessary for me to make a clarification: in the remainder of this document I will use the term ***pitch accent*** generically to mean any semantic perturbation of a pitch contour, be it due to pitch accent proper, phrase accent, or boundary tone, unless a distinction is made explicitly. This is done to simplify the discussion of the relationships between pitch and intonation.

Part II

Investigations

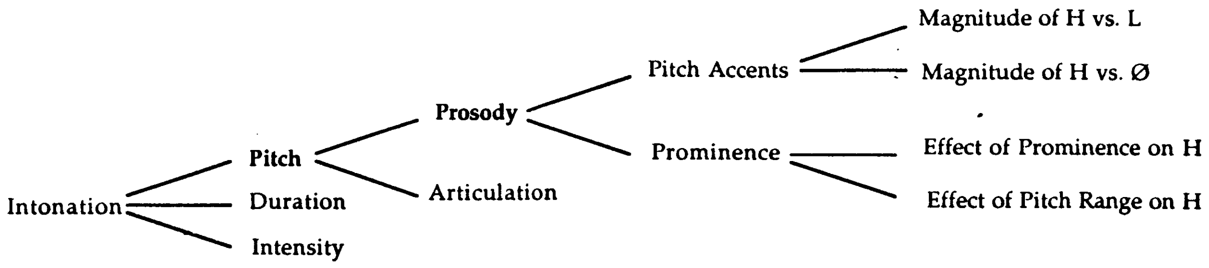

Figure 2.1: Perturbations to pitch contours have various causes.

A successful investigation of prosody requires *both* analyses of generated prosody for the formulation of hypotheses, and verification of these hypotheses through perceptual experiments. Analyses of subjects' speech may uncover some attributes which vary systematically, but without verification on a perceptual level, these variations may be completely irrelevant to the generation of prosody.

In her theory of intonational phonology, Pierrehumbert [15] has identified the variations to pitch that are pertinent to the generation and perception of prosody, but has not differentiated these variations from those which may arise by other means.

The causes of pitch contour perturbations can be broken down as follows (see figure 2.1). Perturbations can be due to prosody, that is, as a result of a deliberate action to change the attentional, intentional, or semantic content of an utterance, or as a consequence of articulation. Prosodic effects can be further decomposed into those arising from the use of pitch accents, and those resulting from (Pierrehumbert's definition of) prominence. Perturbations caused by articulation include vowel intrinsic pitch (IF0), and the effects due to other segmental influences, such as consonant environment [7,21].

Pierrehumbert's theory provides a framework and language with which to investigate and discuss prosodic events. Her treatment of pitch contours as resulting from a series of pitch accents provides a strong base from which to carry out investigations into the use of pitch from a *perceptual* point of view.

The investigations which follow explore the point at which the perceived cause of a perturbation changes from non-prosodic to prosodic. They are based on the premise that pitch excursions that carry meaning, i.e. pitch accents, will be larger than those that don't, and that exploring the point at which pitch excursions become large enough to be meaningful is the same as exploring the point at which pitch accents become perceptible.

In the first investigation, the difference between an **H** accent and an **L** accent is determined by asking listeners to judge whether semantically ambiguous, single-syllable words could be questions or statements. In the second investigation, the magnitude of an **H** accent alone is determined by asking listeners to determine the main topic of a sentence. The third investigation explores the effect of prominence on the magnitude of pitch accents by asking listeners to judge whether a sentence in which the prominence on the first pitch accent is varied could be a question or a statement. The last investigation explores the effect of a listener's expectation of pitch range on the perception of an **H** accent. This is done by presenting listeners with pairs of sentences, the first of which calibrates the listeners' expectation of the pitch range that will be used in the second sentence. The listeners are then asked to judge whether the second sentence is a question or not.

Chapter 3

Comparison of H and L accents

Pierrehumbert's model of intonational grammar [15,16], in all its completeness, still does not define how high an **H** is or how low an **L** is other than to explain that an **H** is higher, and an **L** lower than some reference level (and subject to scaling via prominence). One would like to know the magnitude of an F0 excursion that will cause a pitch accent to be perceived.

3.1 Introduction

Pitch contours carry the majority of the prosodic information used in speech [3]. This information can be represented through the use of pitch accents, which mark the salient excursions in a pitch contour. If one can determine whether a given excursion is semantically salient, then one can determine whether a pitch accent exists at this location. Unfortunately, it is difficult to determine the semantic salience of a pitch excursion; one must not confuse the salience of an excursion with its perceptibility. It may be possible to hear the difference in the height of an excursion without this difference carrying any meaning. In order to be attributable to a pitch accent, an excursion must carry meaning.

One intonational cue that is highly correlated with its meaning is the final rise in pitch associated with questions. This rise is not infallible, however. Utterances which have pitch contours that seem to rise at the end may or may not be questions. For example, lists of words are often recited with pitch rises at the end of each word, and the ends of utterances which may be continued often have what is called a continuation rise. Of particular interest is the intonation of paraverbals and single-syllable words. Such utterances are commonly used during backchannel communication [25]. For example, the affirmative ***uh-huh*** often has rising intonation similar to the question ***what street?***, and ***yeah*** is often intonationally similar to ***where?***. Yet paraverbals lacking lexical content can still be correctly interpreted as a question or as a statement. The same is true of semantically ambiguous words. Even without context, it is possible for listeners to understand that such an utterance was a question or a statement. What prosodic cues influence this decision?

When segmental and speaker-specific influences are eliminated,[1] the shape of the pitch contour is the most important cue for determining the intent of the utterance. One other factor cannot be neglected, however. If similar utterances are heard in succession, there may be a tendency to compare a given utterance to one which preceded. Such a comparison would undoubtedly influence the determination of the intent of the utterance.

This experiment investigates how the shape of the phrase-final pitch contour influences the perception of an utterance as a statement or as a question. Control groups were used to investigate how prior exposure to the stimuli might influence the subjects' responses. An analysis of the results in terms of pitch accents [15] indicates the change in pitch required to differentiate an **H** accent from an **L**.

This experiment is very similar to the one performed by Hadding-Koch and Studdert-Kennedy in 1964 [4].

3.2 Description

Single-syllable, semantically ambiguous words were generated with a family of pitch contours that ranged from falling to monotone to rising at the phrase boundary. The subjects' task was to decide whether each utterance was a question, a statement, or not distinguishable.

[1] Other factors besides intonational and semantic context may affect question/statement perception: speaker baseline and range; phrase-final F0 excursions that span multiple syllables vs. those that span only one; and whether the listener adjusts for IF0 effects. These factors will not be considered in this investigation.

Stimuli presentation was under the subjects' control, via computer. Each stimulus could be heard as many times as desired, but only one "vote" per stimuli was accepted. Subjects' responses were recorded via computer, in a forced-response format.

Three groups of subjects were presented with these stimuli; one in which the stimuli were ordered by decreasing target pitch, one in which the stimuli were ordered by increasing target pitch, and one in which the stimuli were randomly ordered. Within each group, the order of stimuli presentation was the same from subject to subject. It was felt that the effects due to the order of stimuli presentation could not be eliminated by presenting each subject with a different sequence of stimuli, because of the small sample size. It was decided that the best plan of attack was to provide the opportunity for these effects to assert themselves in a predetermined manner, by presenting each subject with the same sequence of stimuli, and to test for significance if ordering effects did occur.

The first two groups were used as controls to account for cases in the third group in which a subject's responses may have been influenced by preceding stimuli. These stimuli represent the worst case in which experience with a particular contour would influence the point at which a change in judgment from question to statement, or vice versa, would occur. It was expected that the group who listened to the stimuli that were ordered by decreasing target pitch, that is, a sequence of utterances which start out as questions and end up as statements, would switch from question judgements to statement judgements earlier in the sequence, that is, at a lower higher pitch, than those who heard the stimuli ordered by rising targets. The size of this difference can be used as a measure of the strength of the influence of ordering.

Utterances with falling pitch contours had the pitch accent notation **H* L L%**, while those with rising contours had the accents **H* H H%** [15]. Comparing the points at which the majority of subjects perceive question contours when given the **H* H H%** sequence to those who perceive statement contours when presented with **H* L L%** will give an indication of the change in pitch required to differentiate an **L** accent from an **H**.

This investigation differs from the 1964 experiment [4] in three important ways. First, the stimuli are fundamentally different. The stimuli from the 1964 experiment consist of the bisyllabic utterance *for Jane*, while those in this experiment are monosyllabic. Bisyllabic utterances pose a problem to an investigation such as this because stress placement affects the shape of the contour. The utterance *for Jane* can have stress placed on either word. If stress were placed on the first word, then the excursion which cues a question will move from the last pitch accent (actually the boundary tone) to the first pitch accent.[2] Since the first pitch accent is the same for both questions and statements, ***prominence*** will become a factor in determining the difference between questions and statements. Limiting the stimuli to monosyllabic utterances ensures that the transition from statement to question (or vice versa) is cued solely by a change in pitch accent type.

Secondly, the contours presented by Hadding-Koch and Studdert-Kennedy consisted of a rise-fall-rise (or rise-fall-fall)[3] while those of the present experiment are simply -rise (or -fall). These two contours may be interpreted in completely different ways.

Finally, the 1964 study was an investigation into the effect of the height of the initial rise on the question/statement transition, which is more similar to the study described in chapter 5 than the present experiment. Nevertheless, it will be interesting to compare the results from the 1964 study to those obtained here.

3.2.1 Method

Stimuli

An analysis of the author's intonation in single-syllabic questions and statements showed that F0 remains fairly constant until the the energy reaches its peak value. An F0 excursion begins at this point of peak energy, and continues until the desired target is reached (see figure 3.1). In cases in which segmental constraints do not permit enough time to reach the desired target pitch, the excursions begin earlier (see figure 3.2). Note that even though the excursions start earlier in the shorter utterance, the duration of the excursions are still

[2] The effect of moving the pitch accent forward in the stimulus was actually described by Hadding-Koch and Studdert-Kennedy, though not explained. They said that questions may be distinguished by a comparatively high F0 throughout the utterance. The explanation of this phenomenon is straightforward under the phonology posited by Pierrehumbert. Unfortunately, such phonologies did not exist at the time of their study.

[3] Rise-fall-rise can be realized with a variety of pitch accents: **H*+L H H%, L+H* L H%, L*+H L H%, H L H%**, etc.

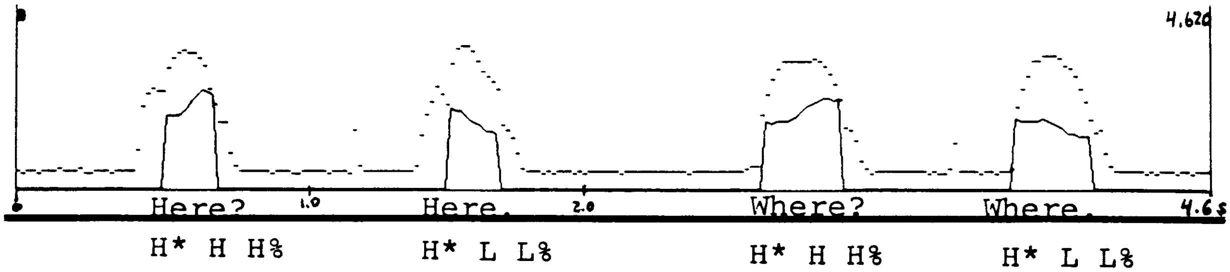

Figure 3.1: Pitch (solid line) and energy (dashed line) of the utterances *Here?* and *Here.*, *Where?* and *Where.*

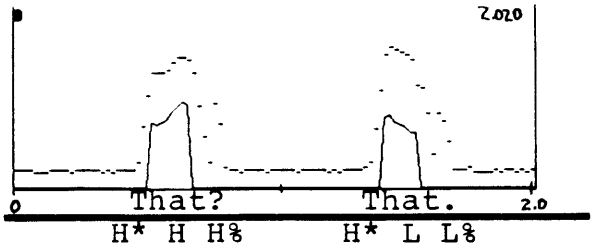

Figure 3.2: Pitch and energy of the utterances *That?* and *That.*

shorter than those in the longer utterances. This means that given the same target pitch, the excursions in the shorter utterances will have steeper slopes than those in the longer utterances. This can be seen for the words *there* and *that* in figure 3.3. *There* exhibits normal pitch behavior, with the pitch excursion beginning at the energy peak. The pitch excursion in *that* begins before the energy peak due to the stop that shortens the voiced portion of the word.

Using this model, stimuli were generated from the words *there* and *that.* To produce each contour, an LPC encoding was made of each word spoken as a question and as a statement (see figure 3.3). Statement contours were generated from the statement version of each word, and question contours from the question version, so that naturalness due to utterance duration and location of the energy peak was minimally affected.

Each excursion began at the same point as in the original utterance, and followed a

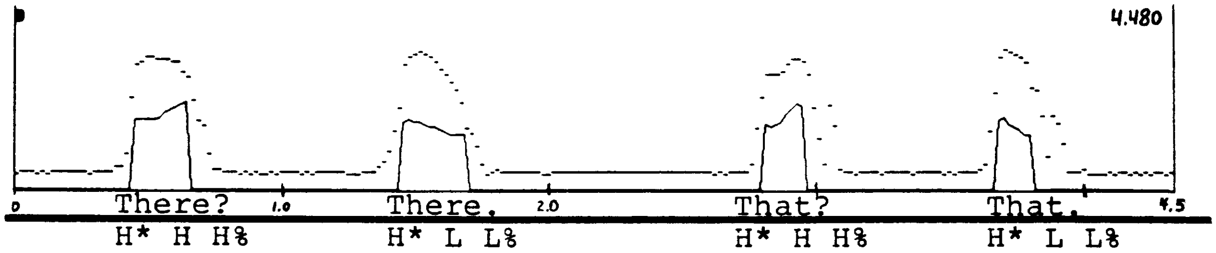

Figure 3.3: Pitch and energy of the utterances *There?*, *There.*, *That?* and *That.*

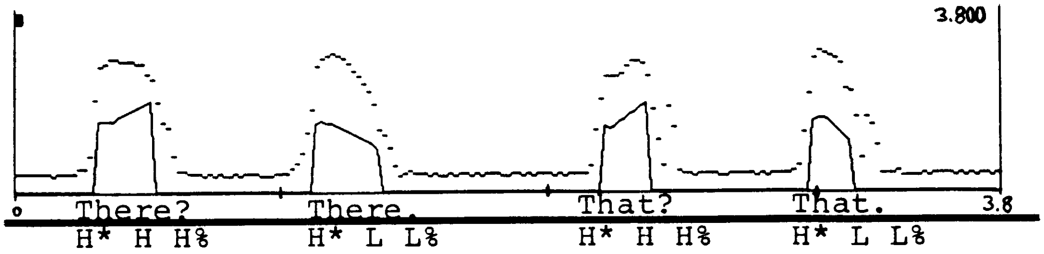

Figure 3.4: Examples of the stimuli *There?*, *There.*, *That?* and *That.* The target pitches are six semitones above monotone for the questions, six semitones below monotone for the statements.

linear trajectory to the target pitch at the termination of voicing.[4] The beginning of the excursion was at 117 Hz for *There* spoken as a question, and at 112 Hz when it was spoken as a statement. The beginning of the excursion was at 105 Hz for *That* spoken as a question, and at 123 Hz when it was spoken as a statement. The target pitch ranged from six semitones below monotone to twelve semitones above monotone, in single semitone steps. The duration of the excursion was 140 milliseconds for *There* spoken as a question, 180 milliseconds for *There* spoken as a statement, 140 milliseconds for *That* spoken as a question, and 100 milliseconds for *That* spoken as a statement. Figure 3.4 shows some sample stimuli.

Note that all of the question contours have the pitch accent sequence **H* H H%**, and that all of the statement contours have the sequence **H* L L%**; differences between contours with like sequences is due solely to prominence.

Stimuli collection and generation

The stimuli were recorded on analog cassette tape and then encoded in LPC. The recorder was a Nakamichi MR-2 with the following settings: Bias: normal; Equalization: 120us; Noise Reduction: Dolby-B. The microphone was a Shure SM12A noise-canceling headset, running through a Shure M267 mixer.

The stimuli were encoded and resynthesized in LPC with a Texas Instruments TI Speech

[4] Other methods of realizing the excursion could have been used: the target could have remained the same as in the original, with the beginning point of the trajectory being manipulated, or the beginning *and* endpoints could have been manipulated, with the trajectories pivoting around a stationary midpoint. All of these methods can be considered equivalent in the case of a single-syllable utterance, as long as naturalness is not affected by the height of the targets. The method used was chosen because of its ease of implementation.

processing card running on an IBM PC. Pitch modification was performed via Pitchtool, an interactive LPC editing facility.

Environment

Each subject was seated in a comfortable chair, approximately 12 to 24 inches from a Sun workstation. The subjects wore a pair of Koss KC-180 headphones, run from a Sansui AU-3900 amp.

After prompting the subjects for their initials, the computer displayed a Sun-window showing four buttons: **Play/Repeat**, **Question**, **Statement**, and **Can't Tell**. Once initials were entered, the subject interacted with the computer solely through the use of the mouse: **Play/Repeat** caused a stimulus to be presented; **Question**, **Statement**, and **Can't Tell** recorded the subjects' responses.

The subjects were instructed that clicking on **Play/Repeat** would cause one of the two stimuli to be heard, and that their task was to determine whether the stimulus could be a statement, a question, or was indeterminate. They were told that they could click on **Play/Repeat** as many times as necessary to make a determination, but that it was usually best to base their judgements on their first impressions.[5] In addition, a suggestion was made to envision each stimulus in a sentence such as "You put it *there*?".

The control subjects were informed that the stimuli were ordered, and how they were ordered. They were told that the investigator was interested in the point at which the stimuli became indeterminate. For example, the first control group was told that the stimuli would start out as questions, become indefinite, and end up as statements, and that the investigator was interested in the point where the questions became indefinite, and where the indefinite stimuli became statements.

All subjects were told that only one response was accepted per stimulus. In the event that a mistake was made (for example clicking on **Question** when they meant to click on **Statement**), they were told to just go on to the next stimulus, and that they needn't worry about the erroneous response.

[5] The number of times a stimuli was repeated was not recorded; this facility was included solely to reduce performance pressure on the subjects.

3.2.2 Data analysis

The data from each group was analyzed independently.

At each target pitch, the number of question, statement, and undecided responses were summed and converted into percentages. This was done for each word independently (see figures 3.7, 3.8 and 3.9), and for both words combined (see figures 3.5, 3.10 and 3.11).

Analysis by word was done to ensure that differences in contour shape due to the duration of voicing was not affecting the perception of the crossover points. Since perception of pitch accents may have to do with pitch trajectories rather than pitch differences, it might be expected that utterances which lack the required time to reach a target might still have the same trajectory as longer utterances. In this case, the target pitch reached in the shorter utterance would be lower (in the case of a question) than that reached in a longer utterance. The importance of trajectory vs. target was determined by comparing the slope of the excursions as well as the final target values. If the contours on both words exhibit the same slope, then trajectory is the predominant prodosic cue, but if they have the same target pitches, then pitch differences are more important to the perception of accent type.

The crossover points, where approximately 50% of the subjects switch from question to statement judgments (or vice versa), in the responses made by the control groups were compared. It was expected that the subjects would make the crossover earlier in the sequence when the stimuli were ordered than when they were randomized. Those who started out making a certain judgement would be comparing the current stimuli to the one just heard, and so would need a smaller excursion to change their judgments from one interpretation to another.

3.3 Results

Sixty-nine subjects participated in the study; fifty-four heard stimuli that were presented in random order, eight heard stimuli that were ordered by falling target pitch, and seven heard stimuli that were ordered by rising target pitch. All subjects were presented with all the stimuli. Each subject made one response per stimulus.

Initial analysis of the data showed significant scatter in the responses of the group who

Combined Responses
Total number of subjects: 48

	number of votes			percent of votes		
Target	qst	stmt	und	qst	stmt	und
-6	0	96	0	0	100	0
-5	0	96	0	0	100	0
-4	0	96	0	0	100	0
-3	0	90	6	0	94	6
-2	1	88	7	1	92	7
-1	0	89	7	0	93	7
0	1	76	19	1	79	20
1	14	57	25	15	59	26
2	23	49	24	24	51	25
3	39	36	21	41	38	22
4	59	17	20	61	18	21
5	69	12	15	72	13	16
6	86	6	4	90	6	4
7	92	0	4	96	0	4
8	95	0	1	99	0	1
9	96	0	0	100	0	0
10	96	0	0	100	0	0
11	96	0	0	100	0	0
12	96	0	0	100	0	0

Figure 3.5: Tally of the responses made by the subjects who heard the stimuli in randomized order, for both words combined.

received the randomized stimuli. Further analysis indicated that this scatter was due to a small number of subjects, who are discussed in section 3.3.4. These subjects (six out of sixty-nine) will be excluded from the following analyses, leaving forty-eight who heard the randomized stimuli.

3.3.1 Combined analysis

Figure 3.5 shows the responses made by the subjects who heard the stimuli in randomized order. The responses for both words have been combined.

Figure 3.6 shows a plot of the combined responses made by the subjects whose stimuli

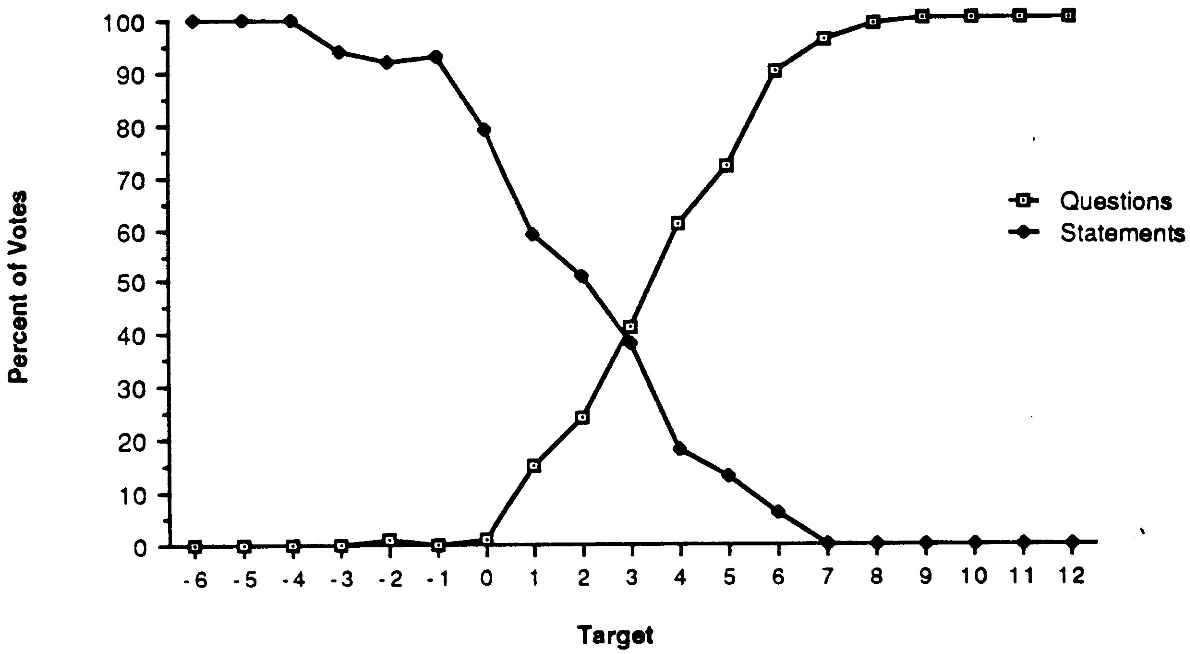

Figure 3.6: Plot of the combined responses made by the subjects who heard the stimuli in randomized order.

were presented in random order. The cross-over from questions to statements is near three semitones above monotone, with approximately 40% question judgements, 40% statement judgements, and 20% undecided.

Between five semitones above monotone and approximately one semitone below monotone, statement judgements make a fairly linear progression of approximately 13% per semitone; at one semitone below monotone, approximately 92% of the subjects have made statement judgements. Between monotone and approximately six semitones above monotone, question judgements make a fairly linear progression of approximately 13% per semitone; at six semitones above monotone, approximately 90% of the subjects have made question judgements.

At the standard interquartile breakpoint of 75% [22], approximately five semitones separate question judgements from statement judgements. This suggests that approximately five semitones are required to differentiate an **H** accent from an **L**.

This result differs from Hadding-Koch and Studdert-Kennedy's finding (at the lowest prominence on the initial rise) in which three to four semitones separate an **H** from an **L**.

3.3.2 Analysis by word

Figure 3.7 shows the sum of the responses made by the subjects who heard the stimuli in randomized order, for each word. The number in the column labeled **Target** indicates the number of semitones above or below monotone (labeled zero, indicating neither rise nor fall in pitch from the beginning of the pitch excursion to the target). Figures 3.8 and 3.9 show the responses made by the subjects who heard the ordered stimuli.

Between three and six semitones separate question judgements from statement judgements at the standard interquartile breakpoint for each word. The mean distance between question and statement judgements is approximately five semitones (with a standard deviation of one).

It appears as if the slope of the pitch excursion is affecting interpretation of the contours. The interquartile breakpoint for questions is lower for the word *that* (the shorter utterance) than for the word *there*, and for statements is higher for the word *that* than the word *there*. That is, a smaller excursion is required in the shorter utterances. This is because the slope of the shorter utterances is similar to that of the longer ones, so a trajectory with a shorter duration will end up at a smaller target.

The effect of slope was not determined for the control groups because it was uncertain what the interaction of stimuli ordering and utterance duration would be.

3.3.3 Controls: The effect of prior exposure

An examination of the data from the control subjects (figures 3.10 through 3.13) demonstrates that the format in which the stimuli were presented permitted successive stimuli to affect one another. As expected, the subjects crossed over from one interpretation to the

Analyzed Responses
Total number of subjects: 48

Target	percent of votes					
	'there'			'that'		
	qst	stmt	und	qst	stmt	und
-6	0	100	0	0	100	0
-5	0	100	0	0	100	0
-4	0	100	0	0	100	0
-3	0	90	10	0	98	2
-2	2	90	8	0	94	6
-1	0	94	6	0	92	8
0	2	81	17	0	77	23
1	17	58	25	13	60	27
2	31	48	21	17	54	29
3	40	35	25	42	40	19
4	81	13	6	42	23	35
5	85	8	6	58	17	25
6	90	8	2	90	4	6
7	94	0	6	98	0	2
8	100	0	0	98	0	2
9	100	0	0	100	0	0
10	100	0	0	100	0	0
11	100	0	0	100	0	0
12	100	0	0	100	0	0

Figure 3.7: Percentage of the responses made by the subjects who heard the stimuli in randomized order, for each word.

Analyzed Responses
Total number of subjects: 8

Target	percent of votes					
	'there'			'that'		
	qst	stmt	und	qst	stmt	und
-6	0	100	0	0	100	0
-5	0	100	0	0	100	0
-4	0	100	0	0	100	0
-3	0	100	0	0	100	0
-2	0	100	0	0	100	0
-1	0	100	0	0	100	0
0	0	100	0	0	100	0
1	13	88	0	0	100	0
2	13	88	0	13	88	0
3	13	75	13	25	75	0
4	25	50	25	38	25	38
5	38	50	13	63	13	25
6	50	38	13	75	13	13
7	75	0	25	100	0	0
8	100	0	0	88	0	13
9	100	0	0	100	0	0
10	100	0	0	100	0	0
11	100	0	0	100	0	0
12	100	0	0	100	0	0

Figure 3.8: Percentage of the responses made by the subjects whose stimuli were ordered by falling target pitch.

Analyzed Responses
Total number of subjects: 7

Target	percent of votes					
	'there'			'that'		
	qst	stmt	und	qst	stmt	und
-6	0	100	0	0	100	0
-5	0	100	0	0	100	0
-4	0	100	0	0	100	0
-3	0	86	14	0	100	0
-2	0	71	29	0	100	0
-1	0	71	29	0	100	0
0	0	71	29	0	57	43
1	57	14	29	29	29	43
2	86	14	0	29	29	43
3	100	0	0	71	14	14
4	100	0	0	71	14	14
5	100	0	0	86	14	0
6	100	0	0	100	0	0
7	100	0	0	100	0	0
8	100	0	0	100	0	0
9	100	0	0	100	0	0
10	100	0	0	100	0	0
11	100	0	0	100	0	0
12	100	0	0	100	0	0

Figure 3.9: Percentage of the responses made by the subjects whose stimuli were ordered by rising target pitch.

Combined Responses
Total number of subjects: 8

Target	number of votes			percent of votes		
	qst	stmt	und	qst	stmt	und
-6	0	16	0	0	100	0
-5	0	16	0	0	100	0
-4	0	16	0	0	100	0
-3	0	16	0	0	100	0
-2	0	16	0	0	100	0
-1	0	16	0	0	100	0
0	0	16	0	0	100	0
1	1	15	0	6	94	0
2	2	14	0	13	88	0
3	3	12	1	19	75	6
4	5	6	5	31	38	31
5	8	5	3	50	31	19
6	10	4	2	63	25	13
7	14	0	2	88	0	13
8	15	0	1	94	0	6
9	16	0	0	100	0	0
10	16	0	0	100	0	0
11	16	0	0	100	0	0
12	16	0	0	100	0	0

Figure 3.10: Tally of the combined responses made by the subjects whose stimuli were ordered by falling target pitch (control group 1).

other earlier in the sequence instead of later. A comparison of figure 3.12 and 3.13 shows this best: the crossover point for the first control group is at five semitones above monotone, and at one semitone above monotone for the second control group.

This outcome could have several explanations: anticipation of the crossover, acclimatization to the stimuli, or comparisons with preceding stimuli. Anticipation of the crossover could have occurred because the control subjects were told that the point at which the stimuli seemed to change from one type of utterance to the other was of importance; they may have been especially attentive to this threshold and made the crossover at the earliest possible opportunity. Another possibility is that the subjects became acclimatized to the

Combined Responses
Total number of subjects: 7

Target	number of votes			percent of votes		
	qst	stmt	und	qst	stmt	und
-6	0	14	0	0	100	0
-5	0	14	0	0	100	0
-4	0	14	0	0	100	0
-3	0	13	1	0	93	7
-2	0	12	2	0	86	14
-1	0	12	2	0	86	14
0	0	9	5	0	64	36
1	6	3	5	43	21	36
2	8	3	3	57	21	21
3	12	1	1	86	7	7
4	12	1	1	86	7	7
5	13	1	0	93	7	0
6	14	0	0	100	0	0
7	14	0	0	100	0	0
8	14	0	0	100	0	0
9	14	0	0	100	0	0
10	14	0	0	100	0	0
11	14	0	0	100	0	0
12	14	0	0	100	0	0

Figure 3.11: Tally of the combined responses made by the subjects whose stimuli were ordered by rising target pitch (control group 2).

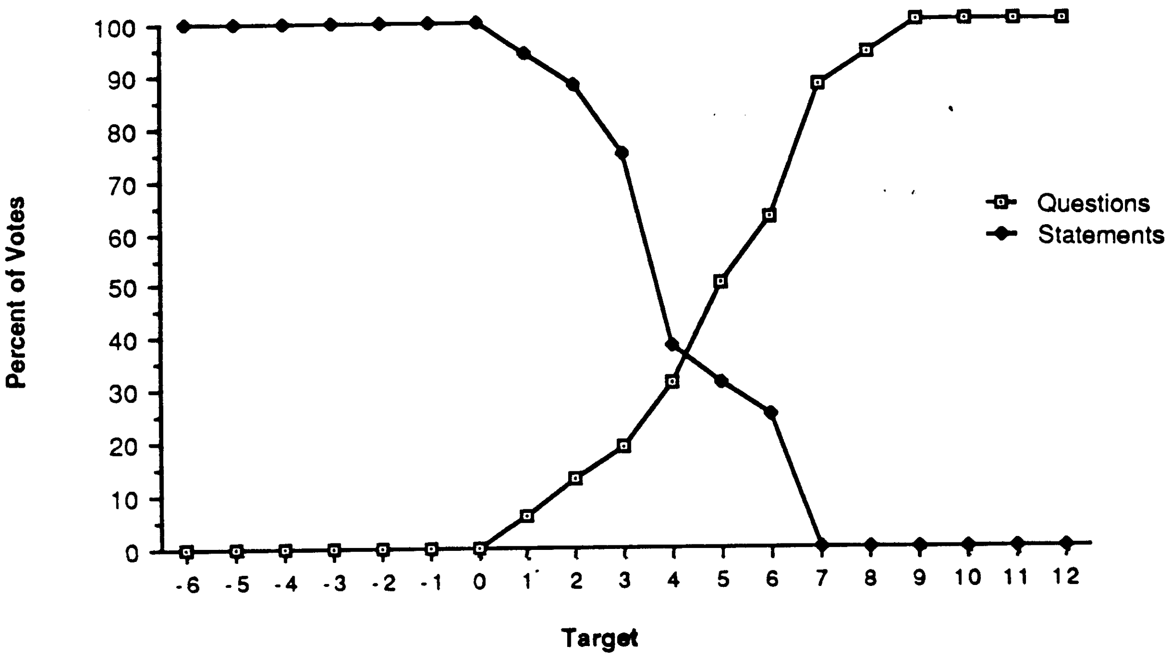

Figure 3.12: Plot of the combined responses made by the subjects whose stimuli were ordered by falling target pitch (control group 1).

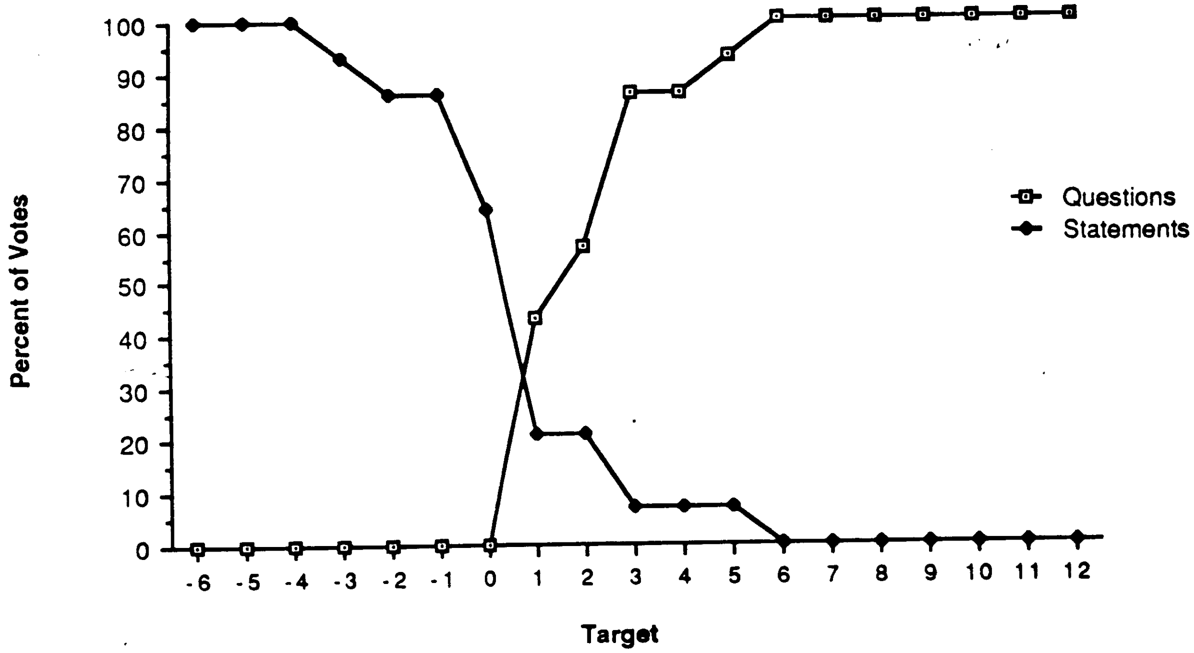

Figure 3.13: Plot of the combined responses made by the subjects whose stimuli were ordered by rising target pitch (control group 2).

stimuli. This would cause them to be more responsive to changes that affect the categorization of the stimuli; a smaller than normal change would be more easily perceptible, so the subjects would make the crossover at an earlier point in the sequence. Finally, the subjects may have been comparing each stimulus with its immediate predecessor. This may have allowed the subjects to detect a change in pitch earlier in the sequence, again because a smaller pitch change would be more easily perceptible.

Comparisons with prior stimuli are of concern because its effects may be accentuated in the randomized sequences. Situations in which pairs of stimuli are separated by a large jump in their target pitches may be more affected by comparisons than stimuli which have only a small jump between them. The randomized sequences will have more large jumps than the ordered sequences.

The effect of comparisons with prior stimuli on the separation of **H** and **L** accents can be seen by combining the data from both of the control groups and comparing them with the data from the randomized group. Combining the control groups nullifies the effects due to ordering; if the separation in the randomized group is the same as that in the combined control groups, then the effects of stimuli ordering will be shown to be insignificant.

As can be seen in figure 3.14, this is indeed the case. The difference between an **H** and an **L** is approximately five semitones, the same as for the randomized stimuli, and the crossover (50%) point for both is at three semitones above monotone.

3.3.4 "Anomalous" judgements

A small number of subjects (six out of 69) have unusual responses. Because of the nature of their responses, these subjects were removed from the preceding analyses. A cutoff was placed at four semitones below monotone for statements and at nine semitones above monotone for questions; those subjects who didn't consistently make statement judgments below four semitones or question judgments above nine semitones were placed in a separate group and analyzed independently. The cutoff points were chosen by looking at the responses made by the control groups (see figures 3.12 and 3.13). When these are combined (figure 3.14) it can be seen that 100% of the subjects make statement judgments below four semitones and 100% make question judgments above nine semitones.

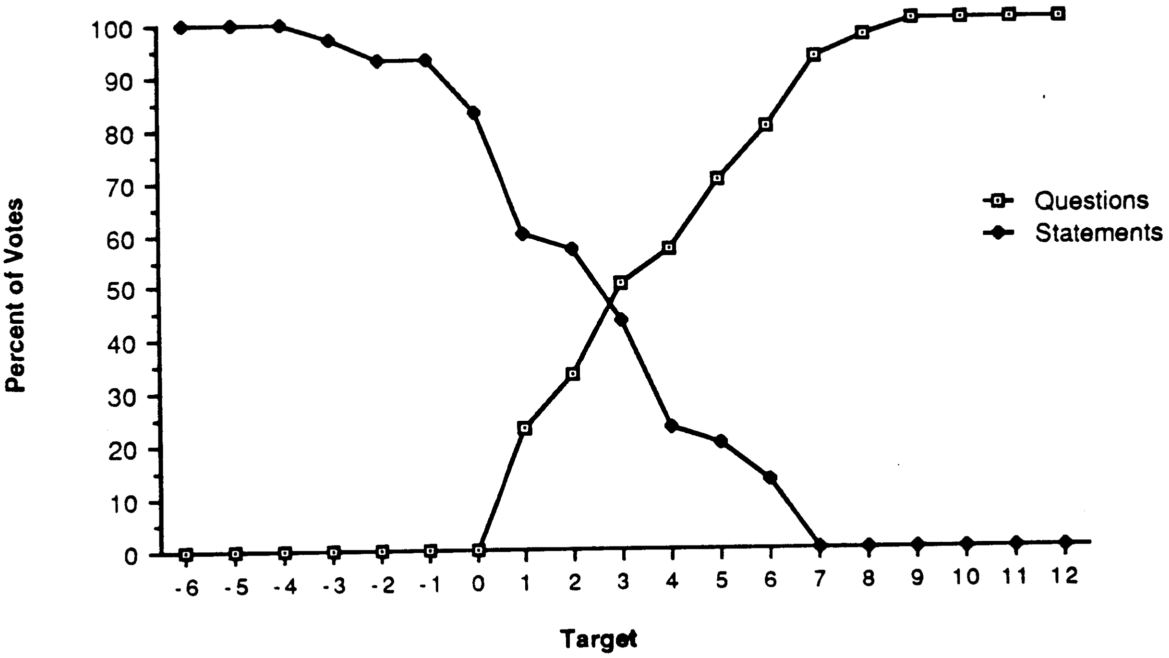

Figure 3.14: Plot of the responses made by both of the controls groups, combined.

Individual Responses
Subject: pjd

Target	number of votes					
	'there'			'that'		
	qst	stmt	und	qst	stmt	und
-6	0	1	0	0	1	0
-5	0	1	0	0	1	0
-4	0	1	0	0	1	0
-3	0	1	0	0	1	0
-2	0	1	0	0	1	0
-1	0	1	0	0	1	0
0	0	1	0	0	1	0
1	0	1	0	0	0	1
2	0	0	1	0	0	1
3	0	0	1	0	0	1
4	1	0	0	0	0	1
5	1	0	0	1	0	0
6	1	0	0	1	0	0
7	1	0	0	1	0	0
8	1	0	0	1	0	0
9	1	0	0	1	0	0
10	1	0	0	1	0	0
11	1	0	0	0	1	0
12	1	0	0	1	0	0

Figure 3.15: The anomalous response at the eleventh semitone above monotone in the word *that* might be due to a mis-selected response.

The way in which these subjects responded appear anomalous in three ways. Some (two out of six) appear as isolated judgements that differ significantly from the trend exhibited by their neighbors; figure 3.15 shows one such occurrence. Such judgements might be due to a mistaken vote; perhaps the subject clicked on **Statement** when he meant to click on **Question**.

Other unusual judgements (four out of six) appear normal except at extreme contour trajectories. The responses shown in figure 3.16 are particularly interesting because of the high number of **Can't Tell** judgements made at very low target pitches, while not one **Can't Tell** judgment was made at the transition from statements to questions; this subject

Individual Responses
Subject: jsw

Target	number of votes					
	'there'			'that'		
	qst	stmt	und	qst	stmt	und
-6	0	1	0	0	1	0
-5	0	0	1	0	0	1
-4	0	1	0	0	1	0
-3	0	1	0	0	0	1
-2	0	1	0	0	1	0
-1	0	1	0	0	1	0
0	0	1	0	0	1	0
1	0	1	0	0	1	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	1	0	0	1	0	0
5	1	0	0	1	0	0
6	1	0	0	1	0	0
7	1	0	0	1	0	0
8	1	0	0	1	0	0
9	1	0	0	1	0	0
10	1	0	0	1	0	0
11	1	0	0	1	0	0
12	1	0	0	1	0	0

Figure 3.16: The high incidence of **Can't Tell** judgements below three semitones below monotone might be due to this subject's stricter definition of the attributes that define a statement.

appeared to be fairly certain about which contours implied statements and which implied questions, but indecisive as the target pitches dropped.

Perhaps this subject misunderstood the instructions, and chose to interpret the statements so narrowly as to exclude utterances with especially low target pitches. For example, the subject may have interpreted these utterances as assertions, which he deemed to be different from simple statements. An interview with this subject did not clarify whether this was indeed the case.

Finally, there is the possibility that the "anomalous" responses are due to the ordering of the randomized stimuli. This is described in the next section.

Order effects

The effects of stimuli ordering are shown by the percentage of the anomalous responses that are preceded by stimuli of the same or different categorical type. Three of the four subjects whose responses cannot be explained as mistakes chose **Can't Tell** instead of **Statement** when the targets were at four semitones below monotone or lower. The stimuli which correspond to these responses were all preceded by stimuli whose target pitches were at seven semitones or above. Two of the four subjects (50%) selected **Can't Tell** when the preceding target was at seven semitones (with the given target was at four below monotone); three of the four (75%) selected **Can't Tell** when the preceding target was at twelve semitones (with the given target at five below). The last subject's responses did not follow this trend. His responses can best be explained as due to misinterpreted instructions.

It is evident that the majority of the subjects whose responses were "anomalous" were basing their judgments on a comparison of preceding stimuli, not by acclimatization.

Combining anomalies with the randoms group

When the responses of the group who heard the randomized stimuli are combined with those which were anomalous (figure 3.17), the difference between an **H** and an **L** is approximately five semitones.

3.4 Summary

After accounting for anomalous data caused by mis-entered responses, misinterpreted instructions, and effects due to the influence of preceding stimuli, the data show fairly clear breakpoints at six semitones above monotone for questions and at one semitone below monotone for statements. At the *interquartile* breakpoint of 75%, five semitones above monotone induce question judgements and monotone induces statement judgements. Since the question contours have the pitch accent sequence **H* H H%** and the statement contours have the sequence **H* L L%**, the five semitone difference in target pitch corresponds to the level of discrimination between **H** and **L** pitch accents.

The difference between the question and statement judgments (at the 75% point) is

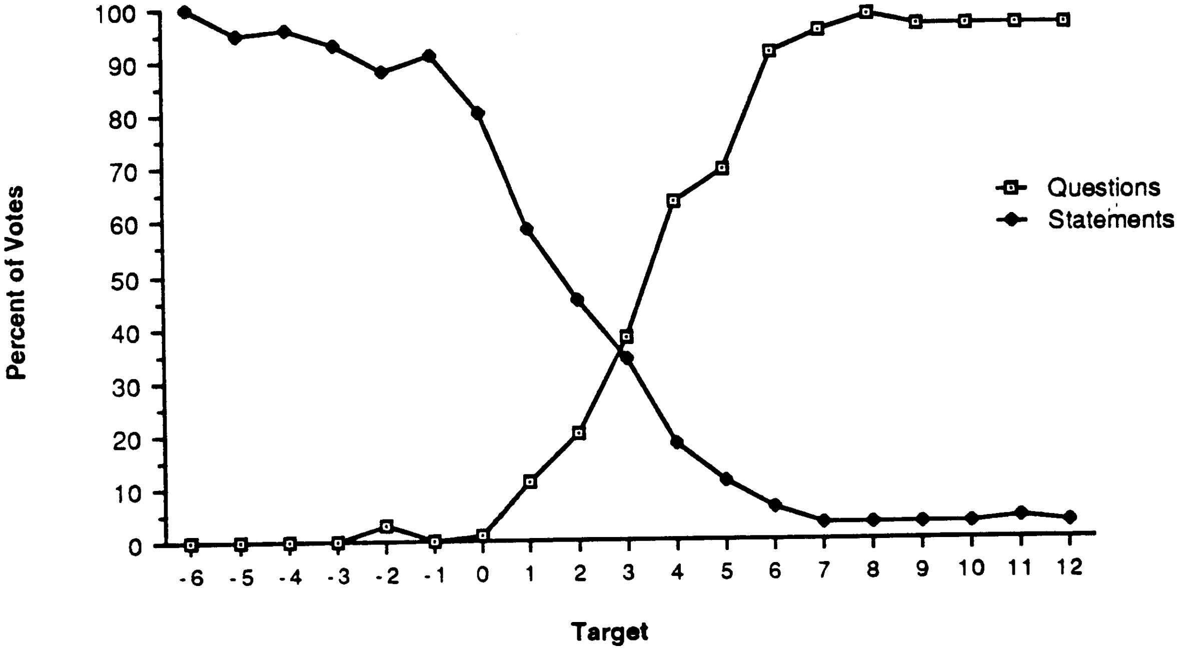

Figure 3.17: Plot of the responses made by the subjects whose stimuli were randomized, combined with those who made anomalous responses.

approximately five semitones whether the stimuli are of short duration or long, whether they are presented in randomly or ordered by target pitch, or whether they are combined by duration or order of presentation.

The five semitone separation at the 75% point differs from that found by Hadding-Koch and Studdert-Kennedy. They found a three to four semitone separation.

The slope of the pitch excursion appears to affect the magnitude at which the pitch accents are perceived. A smaller excursion is required for the shorter utterance.

Successive stimuli appear to interact with one another, probably due to comparisons made by the subjects. When the stimuli are ordered, the crossover from question to statement (or vice versa) occurs earlier in the sequence than when the stimuli are randomized.

A small number of subjects made "anomalous" judgements, probably due to comparisons between successive stimuli. There was a high correlation of anomalous responses to pairs of stimuli that differed markedly by categorical type, such as questions with high targets followed by statements with low targets. However, these judgements did not significantly affect the results; the separation of **H** and **L** remained five semitones whether or not the anomalies were included in the analyses.

3.5 Discussion

The difference in these results from those found by Hadding-Koch and Studdert-Kennedy can be attributed to the difference in contour types used in the two studies. The rise-fall-rise may be semantically different from the simple -rise depending on the location of the initial rise. In addition, the height of the initial rise may affect the point of transition of the judgements from question to statement (or vice versa). This possibility is explored in chapter 5.

The incidence of unusual data responses could have been eliminated with a more careful experiment design. A better interface would have included an "oops" button to allow subjects to correct mis-entered responses. This format was not used because mistakes were expected to be infrequent and recognizable. Such a button would not have complicated the interface, would have removed one possible source of anxiety, and would have resulted in cleaner data.

Misinterpreted instructions could have been avoided with careful wording. The **Can't Tell** button should probably have been labeled "Indeterminate". The description of **H* L L%** contours as "statements" was probably also a mistake. This interpretation of this term is too explicit; a more general term that covers a wider range of contours might be "assertion".

The influence of preceding stimuli can be reduced by careful ordering of the randomized stimuli so as to eliminate large jumps in target values, presenting each subject with a truly random sequence of stimuli (determined via latin square [22]), or using a memory load technique to separate the stimuli (such as presenting pairs of utterances, the first of which has a standard, non-varying contour, and the second with the modified contours). Each has some shortcomings, however.

Eliminating large jumps in target values would be difficult because, prior to this experiment, there existed no data which defined how large is "large". Averaging out ordering effects by giving every subject a different sequence of stimuli would only be valid if the number of subjects was quite large. Finally, pairs of contours may still exert an influence on one another, possibly even more so than randomly presented contours.

Chapter 4

Presence vs. absence of pitch accents

The previous experiment has shown that approximately five semitones separate the perception of a question from a statement. When viewed in terms of Pierrehumbert's phonology, this implies that five semitones separate an **H** pitch accent from an **L**.

While this is an important finding, the cases in which a speaker must choose either an **H** or an **L** are relatively few; more often, a speaker has the option of using no accent at all. While it may be argued that the difference between no accent and an **H** or **L** accent is simply half the difference between an **H** and an **L**, this conclusion cannot be directly drawn from the preceding experiment. The next experiment will address this issue.

Also explored will be the effect of the magnitude of preceding pitch accents on the perception of those which follow. This was suggested by the previous experiment (section 3.5), in which preceding stimuli appeared to affect succeeding responses.

4.1 Introduction

Speakers emphasize words using a combination of pitch, duration, and loudness. Listeners use these cues to varying degrees as well; pitch changes have the largest effect on the perception of emphasis, followed by duration, and then loudness [3]. Emphasizing one word over another can affect the scope, meaning, or intended interpretation of the utterance by

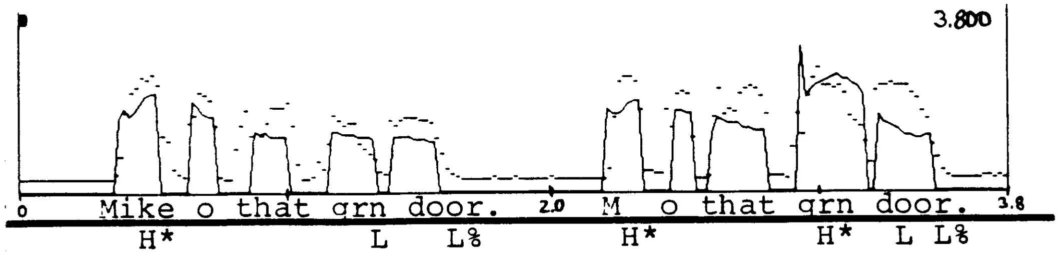

Figure 4.1: Pitch and energy of the sentence ***Mike opened that green door*** spoken twice, first with the pitch accent sequence of sentence (A), then with the sequence of sentence (B).

signalling that this word is somehow important. Conversely, it is possible to determine which word was emphasized by asking listeners about the meaning, scope or interpretation of the utterance. This method of determining emphasis lends itself well to investigating the effects of intonation because it avoids problems that may result from asking listeners about the intonation explicitly.

The change in pitch required to differentiate the presence of an **H** accent from the absence of an accent can be determined by altering the pitch of a particular word, and asking listeners about their interpretation of the resulting utterance. This experiment investigates the minimum pitch change required for a listener to determine whether or not a given word carries an **H** pitch accent. The effect of surrounding pitch accents is also explored.

4.2 Description

The sentence ***Mike opened that green door*** was generated with the peak pitch on the word ***green*** varying. The stimuli had the following pitch accent transcriptions (see figure 4.1):

```
(A)  Mike opened that green door.
      H*                     L    L%

(B)  Mike opened that green door.
      H*                H*   L    L%
```

The sentence was presented in the context of being an answer to one of two questions: (A) ***Who opened that green door?*** or (B) ***Which door did Mike open?*** The subjects' task was to decide which of the two questions best fit the presented answer. Those who heard

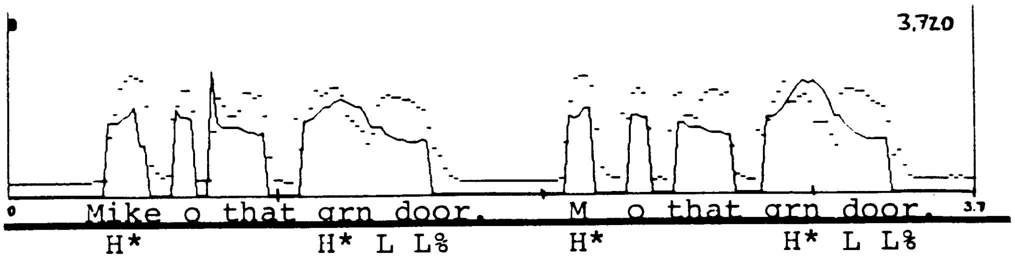

Figure 4.2: Pitch and energy of the sentence *Mike opened that green door*, spoken twice. The pitch accent sequence for both is that of sentence (B). Note how the pitch contour peaks near the center of the word *green*. The pitch contours are unmodified.

the utterance *Mike opened that* GREEN *door* were expected to pick question (B), while those who didn't were expected to interpret the utterance as MIKE *opened that green door* and therefore pick question (A).[1] The point at which the majority of subjects picked question (B) was the point at which the word *green* changed from being unaccented to carrying an **H** accent.

The stimuli were randomly ordered. Stimuli presentation was under the subjects' control. Each stimulus was presented twice. Each stimulus could be heard as many times as desired, but only one response per presentation was accepted. Subjects' responses were recorded in a forced-choice format.

4.2.1 Method

Stimuli

An analysis of the author's intonation while speaking the sentence with different pitch contours showed that F0 excursions during the word *green* have a parabolic shape that peaks near the center of the voiced portion of the syllable (see figure 4.2). Using this model, the sentence was resynthesized in LPC with a variety of pitch contours.

To produce each contour, LPC encodings were made of sentence (A) with varying amounts of emphasis on the word *Mike*, and of sentence (B) with varying amounts of emphasis on the word *green* [23] (see figure 4.3). To produce a natural sounding utterance with a minimal amount of prominence on either of the pitch accents, the first two syllables

[1] The grammar for pitch accents requires that the word *Mike* receive an accent when no words stand out as acoustically prominent.

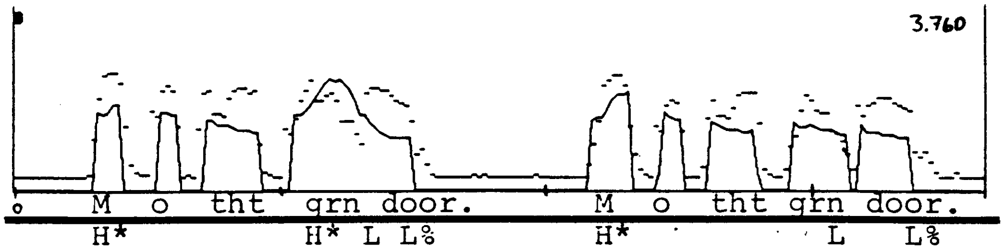

Figure 4.3: Pitch and energy of the sentence with emphasis on on the word *green* and with emphasis on the word *Mike*.

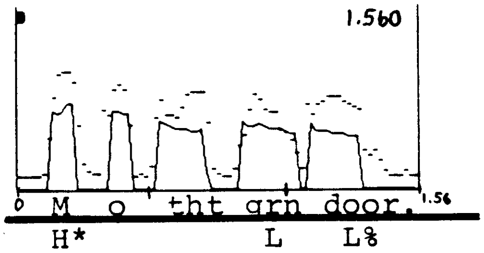

Figure 4.4: Pitch and energy of the reference sentence, assembled from the minimal prominence portions of two sentences.

from sentence (B) were inserted in place of those in sentence (A) (see figure 4.4). This sentence was used as the reference upon which changes to the contour were performed to produce the stimuli.

Each contour was synthesized with a pitch trajectory beginning at the same point as in the reference utterance, peaking at the target pitch in the center of voicing, and ending with the original pitch at the termination of voicing. The target pitch ranged from zero to six semitones above the reference, in single semitone steps. Figure 4.5 shows some sample stimuli.

To demonstrate the effect of the prominence of preceding pitch accents, a series of stimuli were generated with the word *Mike* emphasized a fixed amount, and with the emphasis on *green* varying as described above. In order to determine how much to raise the pitch on *Mike*,

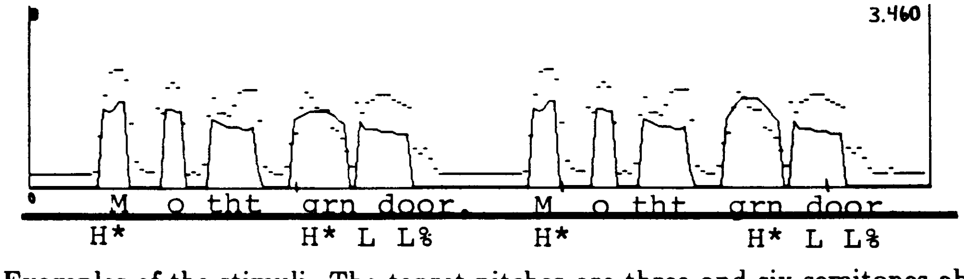

Figure 4.5: Examples of the stimuli. The target pitches are three and six semitones above the reference.

a pilot study was done in which a few subjects were presented with stimuli with varying amounts of emphasis on both *Mike* and *green*. It was determined that approximately three semitones were required for the subjects to perceive an accent on *green* when *Mike* was unaccented. When this value was tried as a target on *Mike*, however, it was found that an accent was never perceived on *green*. Consequently the target for the fixed emphasis on *Mike* was decreased to two semitones above the reference.

Stimuli collection and generation

The stimuli were encoded directly in LPC with a Texas Instruments TI Speech processing card running on an IBM PC. The microphone was a Shure SM12A noise-canceling headset, running through a Shure M267 mixer. Pitch modification was performed via Pitchtool, an interactive LPC editing facility.

Environment

Each subject was seated in a comfortable chair, approximately 12 to 24 inches from a Sun workstation, in an acoustically-treated room. The stimuli were re-synthesized in real-time, run through an Audioarts Engineering 4200A parametric equalizer to eliminate noise generated by the speech card, and amplified by a Crown D-75. A pair of Rogers LS3/5A speakers were placed on either side of the subject at approximately 1.5 meters. The volume was adjusted according to the subjects' preference.

After prompting the subjects for their initials, the computer displayed a Sun-window showing four buttons:

- **Play/Repeat**, which caused a stimulus to be presented;
- **WHO opened that green door?**, which recorded the subjects' responses;
- **WHICH DOOR did Mike open?**, which recorded the subjects' responses; and
- **Back up**, which allowed the subjects to correct a mistaken response.

Once initials were entered, the subject interacted with the computer solely through the use of the mouse.

The subjects were instructed that clicking on **Play/Repeat** would cause one of the stimuli to be heard, and that their task was to determine whether the stimulus could be an answer to one of the two questions displayed on the buttons in the Suntool. They were told that they could click on **Play/Repeat** as many times as necessary to make a determination, but that it was usually best to base their judgements on their first impressions.[2]

All subjects were told that only one response was accepted per stimulus. In the event that a mistake was made, they were told that they could repeatedly click on **Back up** to go back as far as necessary to correct the response that they had made.

4.2.2 Data analysis

At each target pitch, the number of responses for each question were summed and converted into percentages. These were plotted for the sentences in which the word *green* had the various target values, and for the sentences in which the word *Mike* carried the various targets. The plots for the word *green* in which the word *Mike* was emphasized were compared to those in which the word *Mike* was neutral. Similarly, the plots for the word *Mike* in which the word *green* was emphasized were compared to those in which the word *green* was neutral.

Comparisons were based on analyses of variance and linear regression. A non-zero linear regression with a high correlation coefficient will show that the perception of the presence of the pitch accent is dependent on the magnitude of the target pitch. An analysis of variance will verify that this result is significantly different from that which might be obtained by random chance. The analysis of variance will also show that the responses for the emphasized words are significantly different from those for the neutral words (summed over the target pitches), and that the interaction of target pitch and emphasis is significant.

It was expected that a lower target would be required to perceive an accent on *green* when the emphasis on *Mike* was neutral, as well as on *Mike* when *green* was neutral. The same targets were expected for both cases.

[2] The number of times a stimuli was repeated was not recorded; this facility was included solely to reduce performance pressure on the subjects.

4.3 Results

Nineteen subjects participated in the study. Each subject responded to each stimulus twice, for a total of thirty-eight responses per stimulus. Each subject heard all combinations of emphasis on *Mike* and *green*.

Figure 4.6 shows the percent of subjects who perceived a stronger accent on the indicated word, under the conditions shown. The first pair of responses are the percentage of subjects who perceived a stronger accent on the word *Mike*, when the word *green* was emphasized by two semitones, and when it was neutral. The second pair of responses are the percentage of subjects who perceived a stronger accent on the word *green*, when the word *Mike* was emphasized by two semitones, and when it was neutral. The number in the column labeled **Target** indicates the number of semitones above the reference (labeled zero, indicating neutral intonation on the indicated word) reached during the pitch excursion on the indicated word.

Figures 4.7 and 4.8 show this data plotted. Figure 4.7 shows that the percentage of subjects who perceived a stronger accent on the word *green* while *Mike* was emphasized, and while *Mike* was unemphasized. Using the standard interquartile breakpoint of 75% [22], it can be seen that approximately two semitones were necessary for the perception of the accent when *Mike* was unemphasized, but that this increased to more than three semitones when *Mike* was emphasized. At four semitones above the reference, approximately 92% of the subjects perceived a stronger accent on the word *green* when *Mike* was not emphasized, while it took more than six semitones to induce more than 90% of the subjects to perceive a stronger accent when *Mike* was emphasized.

The plot of the percentage of subjects who perceived a stronger accent on the word *Mike* (figure 4.8) shows that much more than six semitones are required to overcome an emphasis placed later in the sentence.

A linear regression performed on this data (see figure 4.6) shows that it is highly correlated, with a non-zero slope, and an analysis of variance shows that this is significant ($F = 465.460$ at $p < 0.001$). Analysis of variance also shows that emphasis on the word *Mike* does not significantly affect the perception of an accent on the word *green*, although

Percentage Responses
Number of subjects: 19

Target	Number of responses			
	Perceived accent on: 'Mike'		Perceived accent on: 'green'	
	with emphasis on 'green'	without emphasis on 'green'	with emphasis on 'mike'	without emphasis on 'mike'
0	21	97	5	11
1	18	95	37	50
2	34	97	55	74
3	42	100	68	76
4	42	100	82	92
5	55	97	82	92
6	63	97	89	92
Slope	7.4	.25	13	12 (18)*
Corr. coef.	.97	.29	.94	.88 (.94)*

*The number in parentheses is the result of the linear regression for the points from target 0 through target 4 only.

Figure 4.6: Percentage of subjects who heard a stronger accent on the indicated word, under the neutral and emphasized conditions.

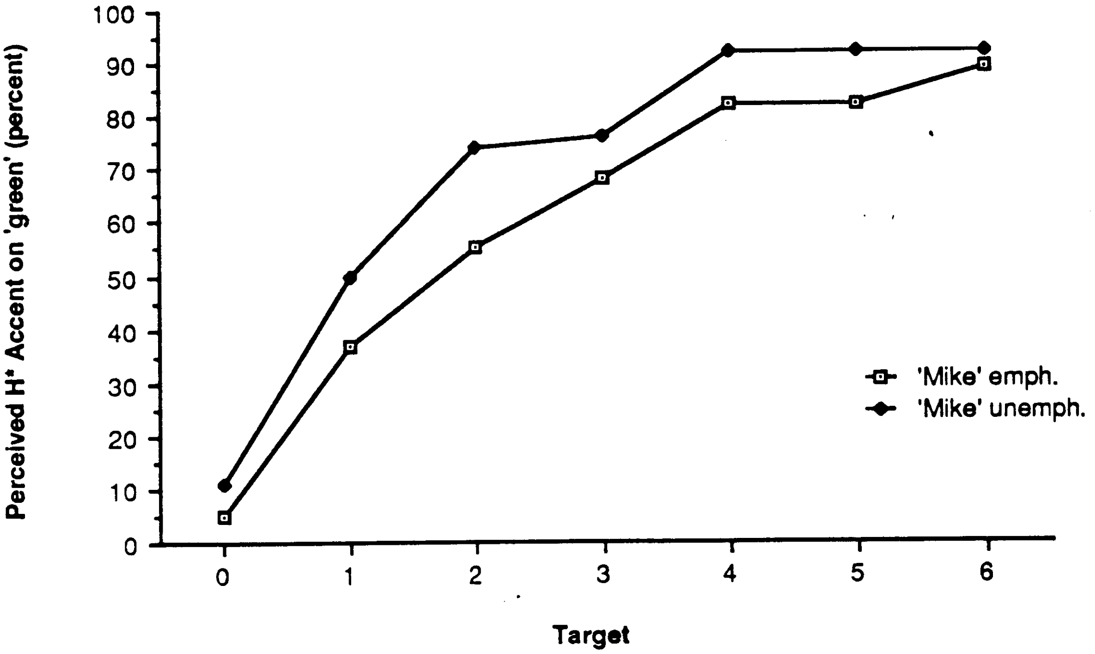

Figure 4.7: Percentage of subjects who heard a stronger accent on the word *green*, under the neutral and emphasized conditions.

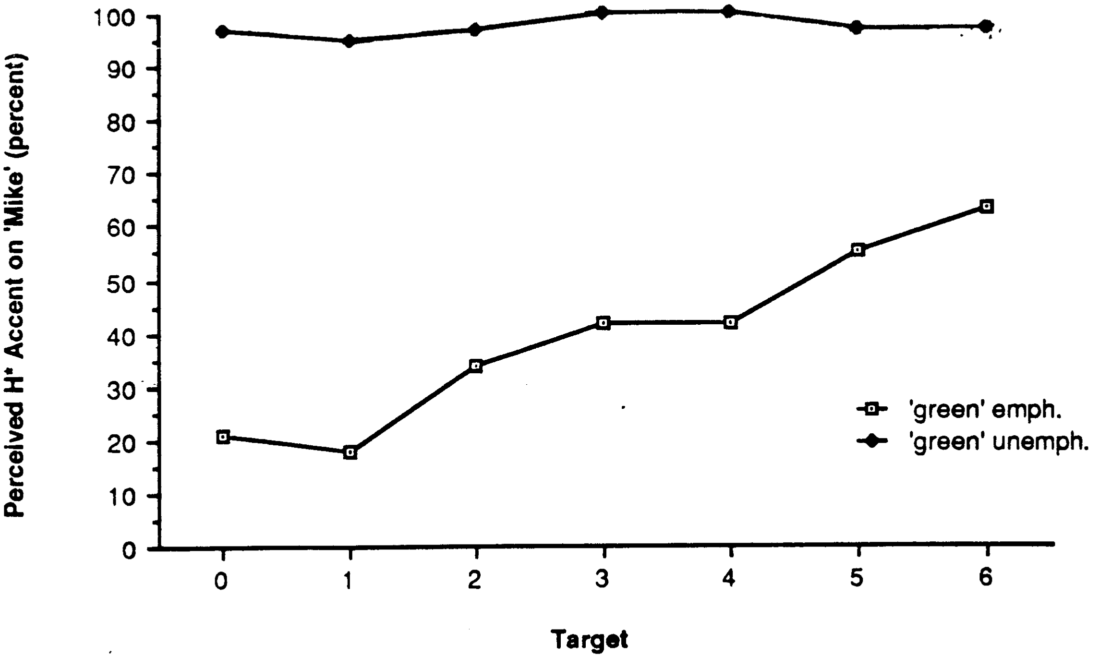

Figure 4.8: Percentage of subjects who heard a stronger accent on the word *Mike*, under the neutral and emphasized conditions.

there is a non-zero tendency toward this result ($F = 3.630$ at $p < 0.073$); however, the perception of an accent on the word *Mike* is significantly affected by emphasis on the word *green* ($F = 60.129$ at $p < 0.001$). The interaction of target pitch with emphasis is also significant ($F = 4.632$ at $p < 0.001$).

4.4 Discussion

The effect of target pitch on the perception of the presence of a pitch accent is confirmed by the non-zero slope and high correlation of the data, coupled with the high significance found by the analysis of variance. The small slope and low correlation of the responses of the perceived accent on *Mike* when *green* was unemphasized occurred because lack of emphasis on the word *green* implies that the pitch accent sequence is **H* L L%**; the subjects always heard an accent on *Mike* simply because there was no accent for them to hear on the word *green*.

Analysis of variance confirms the suspicion that surrounding pitch accents do indeed affect the perception of a given pitch accent. The tendency toward the effect of surrounding pitch accents is shown by the F ratio for the amount of emphasis on *Mike* summed over the target pitches on *green*, and by the significant F ratios found for the emphasis on *green* summed over the targets on *Mike*.

Note that the difference in excursion required to induce the same number of subjects to perceive a pitch accent is greater in the *Mike* sentences than the *green* sentences; the F ratio only suggests a tendency for the *green* sentences, but shows a high significance for the *Mike* sentences. This is counter-intuitive; Pierrehumbert asserts that targets are chosen by speakers without look-ahead [14], yet the effect of the succeeding pitch accent is larger than that of the preceding one. Perhaps this is due to a lowering of the topline as an utterance progresses [17]. This lowering permits a smaller pitch change near the end of an utterance to convey a larger prominence than the same pitch change would at the beginning of the utterance; hence the larger effect of prominence at the end of the sentence on the perception of the accent at the beginning of the sentence.

In conclusion, the results suggest that those sentences in which the word *green* was neutral had the pitch accent sequence **H* L L%**, while those in which the pitch on the word ***green*** was raised more than three semitones above its neutral value had the accent sequence **H* H* L L%**. The increase in pitch caused the perception of the extra accent on the word ***green***. This implies that the minimum excursion required for the perception of an **H** pitch accent is approximately two to three semitones.

4.5 Other observations

Using randomly-ordered stimuli on the small sample size did not appear to affect the outcome of the results. Also, the use of the **Back Up** button prevented substitution errors of the type experienced in the previous investigation.

4.6 Summary

It has been shown that a pitch change of approximately two to three semitones is required to perceive the presence of an **H** pitch accent, when surrounding pitch excursions are kept to a minimum.

As surrounding pitch excursions were increased, the required pitch change also increased. When the preceding pitch accent was increased in prominence by two semitones, the pitch change required to perceive the **H** accent increased to more than three semitones.

Finally, the effect of succeeding pitch accents was larger than expected. This is explained as a consequence of the lowering of the topline as an utterance progresses, which permits a smaller pitch change near the end of the utterance to convey a larger prominence than would be the case nearer the beginning.

Chapter 5

Effects of preceding prominence on pitch accent perception

5.1 Introduction

The previous experiment indicated that the height of a preceding pitch accent can influence the perception of the presence or absence of a given accent. For example, a preceding accent with a high prominence may cause the accent in question to be perceived as unaccented unless its pitch is raised accordingly.

This experiment investigates how the height of the target pitch of a given pitch accent needs to change as the height of a preceding accent is changed, in order for the excursion to still be interpreted as a pitch accent.

5.2 Description

The sentence *Mike put it there?* was generated with the prominence on the word *Mike* and the target pitch on the word *there* varying. [1]

The stimuli had the following pitch accent transcriptions (see figure 5.2):

[1] Here I specifically mean to use Pierrehumbert's definition of *prominence* [14], see section 2.2.2, in order to disambiguate the change in target pitch on the word *there* from the change in the target on the word *Mike*.

```
(A)  Mike put it there.
       H*            L   L%

(B)  Mike put it there?
       H*            H   H%
```

The subjects' task was to decide how the sentence would be punctuated, if it were to be transcribed. A forced-choice response format was used; the subjects could pick either a question mark (? or !?), period (.), exclamation point (!), or ellipsis (...). The responses were divided into questions and non-questions; those who chose question marks had perceived an **H%** phrase boundary, as in sentence (B). As the prominence on the first pitch accent rose, it was expected that it would take a higher target pitch for an **H%** to be perceived.

These types of punctuation were chosen because they represent the majority of the different interpretations that the pitch contours could have. It was felt that providing a larger repertoire of responses than the simple question/statement used in chapter 3 would eliminate the problem described in that chapter in which subjects misinterpreted the meaning of the "statement" button. The two question marks correspond to a definite, meaningful terminal rise in pitch. The combination of exclamation point and question mark (!?) corresponds to a higher target pitch, such as would be used to imply incredulity, than would be used for a simple question. The period and exclamation point correspond to a final lowering of pitch. An exclamation point may be realized with a terminal rise-fall. Ellipsis is used to mark the contours that fall between a question and a statement. Since these contours would not have a significant terminal fall or rise, for example as in a continuation contour [9], they would not be interpreted as a question and would consequently not carry an **H%** boundary tone.

The stimuli were randomly ordered. Stimuli presentation was under the subjects' control, via computer. Each stimulus was presented twice. Each stimulus could be heard as many times as desired, but only one response per presentation was accepted. Subjects' responses were recorded via computer.

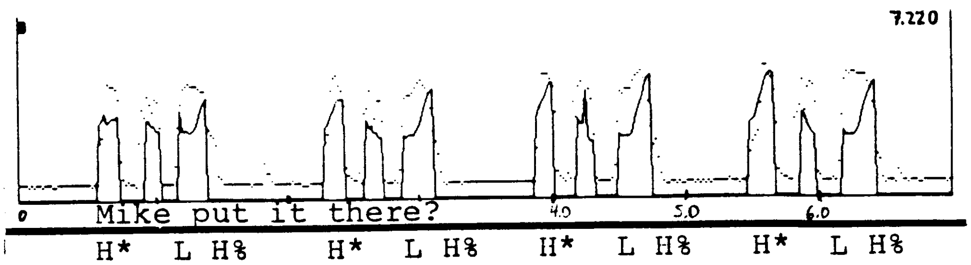

Figure 5.1: Pitch and energy of the sentence with four levels of emphasis on the word *Mike*. The peak pitch on the word *Mike* is 146 Hz in the first sentence, four semitones above this in the second sentence, eight above in the third, and ten semitones higher in the last sentence.

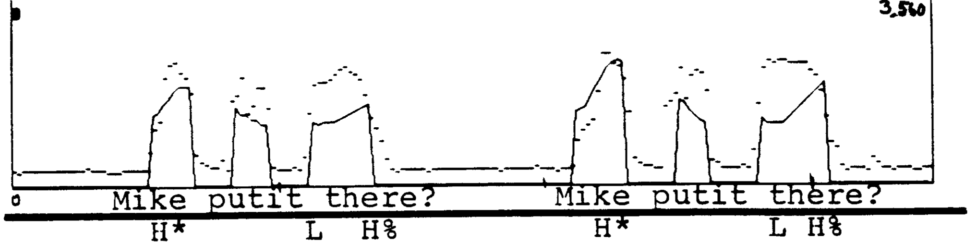

Figure 5.2: Examples of the stimuli. The target pitches are five and eleven semitones above monotone. The first sentence has low emphasis on the word *Mike*; the second has high emphasis on the word *Mike*.

5.2.1 Method

Stimuli

To produce each contour, four LPC encodings were made of the sentence with different levels of emphasis on the word *Mike* [23] (see figure 5.1). In the first of these sentences, the peak pitch on the word *Mike* is at 146 Hz. In the second sentence, the peak pitch is approximately four semitones (186 Hz) above this; the third increases to eight semitones (229 Hz), and the fourth to ten semitones (258 Hz). [2]

The contours on the word *there* were then modified by imposing a linear trajectory on the pitch, beginning at the same point as the pitch excursion began in the original utterance and continuing to the target pitch at the termination of voicing (see chapter 3). The target pitch ranged from one to 19 semitones above monotone, in odd semitone steps. This produced 40 stimuli, each of which were presented twice, making a total of 80 responses per subject. Figure 5.2 shows some sample stimuli.

[2] The speaker's baseline is approximately 88 Hz.

Stimuli collection and generation

The stimuli were recorded on analog cassette tape and then encoded in LPC. The recorder was a Nakamichi MR-2 with the following settings: Bias: normal; Equalization: 120us; Noise Reduction: Dolby-B. The microphone was a Shure SM12A noise-canceling headset, running through a Shure M267 mixer.

The stimuli were encoded and resynthesized in LPC with a Texas Instruments TI Speech processing card running on an IBM PC. Pitch modification was performed via Pitchtool, an interactive LPC editing facility [12].

Environment

Each subject was seated in a comfortable chair, approximately 12 to 24 inches from a Sun workstation, in an acoustically-treated room. The stimuli were re-synthesized in real-time, run through an Audioarts Engineering 4200A parametric equalizer to eliminate noise generated by the speech card, and amplified by a Crown D-75. A pair of speakers, Rogers LS3/5A Monitors, were placed on either side of the subject at approximately 1.5 meters. The volume was adjusted according to the subjects' preference.

After prompting the subjects for their initials, the computer displayed a Sun-window showing six buttons:

- **Play/Repeat**, which caused a stimulus to be presented;
- **Mike put it there? (!?)**, which recorded the subjects' responses;
- **Mike put it there.**, which recorded the subjects' responses;
- **Mike put it there!**, which recorded the subjects' responses;
- **Mike put it there...**, which recorded the subjects' responses; and
- **Back up**, which allowed a subject to correct a mistaken response.

Once initials were entered, the subject interacted with the computer solely through the use of the mouse.

The subjects were instructed that clicking on **Play/Repeat** would cause one of the stimuli to be heard, and that their task was to determine how the stimuli were punctuated, should they want to transcribe each sentence. They were told that they could click on

Play/Repeat as many times as necessary to make a determination, but that it was usually best to base their judgements on their first impressions.[3] In addition, the interpretation of a question mark was described as suggesting an amount of uncertainty, for example disbelief or incredulity as well as a simple question. Ellipsis was described as if the statement were incomplete, but was about to be completed, for example "Mike put it there... and then picked it up again." It was suggested that the subjects try to imagine such a scenario each time a stimulus was presented.

All subjects were told that only one response was accepted per stimulus. In the event that a mistake was made, they were told that they could repeatedly click on **Back up** to go back as far as necessary to correct the response that they had made.

5.2.2 Data analysis

At each target pitch on the word *there* and amount of prominence on the word *Mike*, the number of responses for each question response were summed and converted into percentages. A plot of percent question responses vs. prominence of the word *Mike* was made for each target pitch on the word *there*. An analysis of variance was performed to determine the significance of the effect of prominence on the magnitude of the excursion required for the perception of an **H%**.

It was expected that a lower number of subjects would perceive an **H%** accent on the word *there* (i.e. perceive the utterance as a question) as the prominence on the word *Mike* went up.

5.3 Results

Ten subjects participated in the study. Each subject responded to each stimulus twice, for a total of twenty responses per stimulus.

Figure 5.3 shows the percent of subjects who perceived an **H%** accent on the word *there* (i.e. those who interpreted the sentence as a question) vs. those who perceived an **L%** (all other responses). The columns labeled **Prominence 1** through **Prominence 4** correspond

[3] The number of times a stimuli was repeated was not recorded; this facility was included solely to reduce performance pressure on the subjects.

Percentage Responses
Total number of subjects: 10

Target	Percent of responses							
	Prominence 1		Prominence 2		Prominence 3		Prominence 4	
	H%	L%	H%	L%	H%	L%	H%	L%
0	0	100	0	100	5	95	15	85
1					5	95	10	90
2	0	100	5	95				
3					15	85	15	85
4	65	35	10	90				
5					20	80	25	75
6	80	20	30	70				
7					45	55	35	65
8	100	0	45	55				
9					50	50	35	65
10			75.	25				
11					70	30	40	60
12			100	0				
13					85	15	70	30
Slope	14		8		6		4	
r	.95		.96		.98		.92	

Figure 5.3: Percentage of subjects who perceived the indicated accent on the word *there*, for each of the indicated amounts of prominence on the word *Mike*.

to the different amounts of prominence used on the word *Mike*; Prominence 1 is the lowest, and Prominence 4 the highest. The number in the column labeled **Target** indicates the number of semitones above monotone (labeled zero, indicating no pitch change) reached at the end of the sentence.

Figure 5.4 shows this data plotted as percent of subjects who perceived an **H%** boundary tone vs. target pitch on the word *there*. Each curve corresponds a particular level of prominence on the word *Mike*.

Figure 5.5 shows this data plotted as percent **H%** perception vs. level of prominence on the word *Mike*. Each curve corresponds to a particular target pitch on the word *there*.

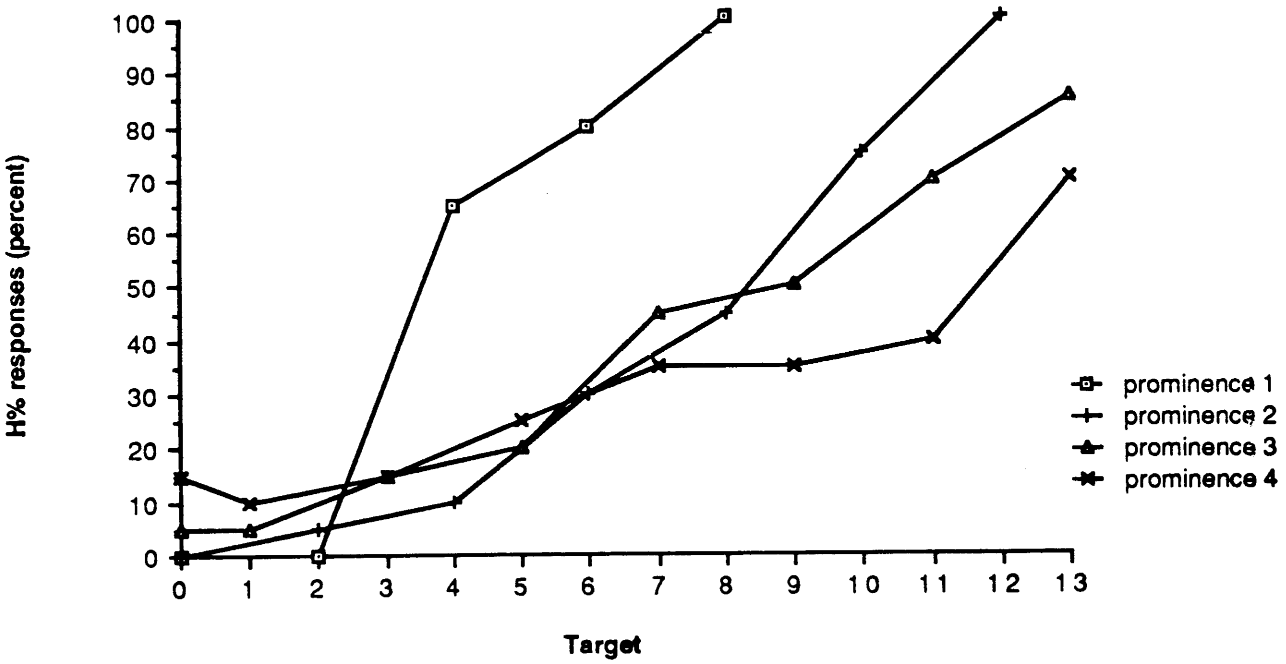

Figure 5.4: Plot of the percentage of subjects who perceived an H% boundary tone vs. target pitch on the word *there*. Each curve corresponds to a particular level of prominence on the word *Mike*.

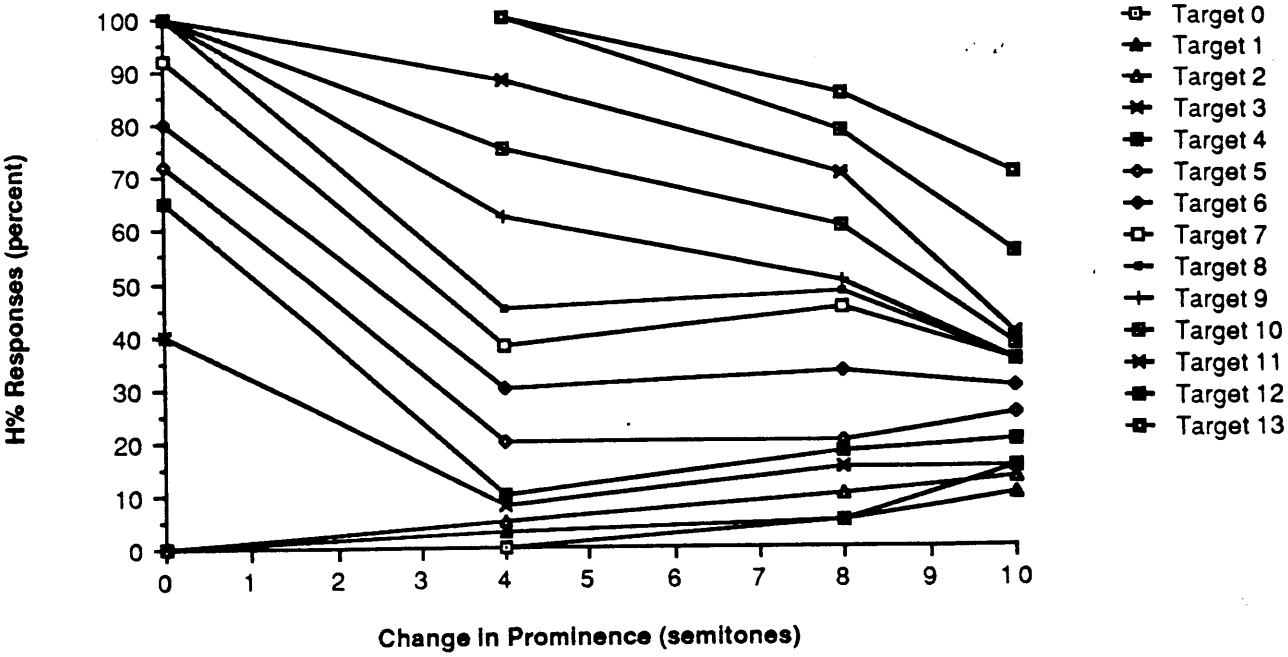

Figure 5.5: Plot of the percentage of subjects who perceived an **H%** boundary tone vs. amount of prominence on the word *Mike*. Each curve corresponds a particular target on the word *there*.

The abscissa lists the level of prominence in semitones; the sentence with the lowest peak pitch on the word *Mike* (prominence one) is the reference, and so is marked at zero. The sentence with prominence level two is four semitones above the reference, prominence three is at eight semitones above the reference, and prominence four is at ten.

The increase in prominence required to induce the subjects to perceive the accent is suggested by the plot. At prominence one, 75% of the subjects perceived an **H%** at a target of five semitones; ten semitones were required at prominence two, about an octave at prominence three, and more than thirteen semitones at prominence four. However, an analysis of variance showed that this effect of prominence was insignificant ($F = 1.604$ at $p < 0.211$).

5.4 Discussion

The results do not clearly show that prominence on a given pitch accent can affect the perception of a succeeding accent. This is puzzling; the plot shows the effect quite clearly, but the analysis of variance suggests otherwise. This result is probably due to the small number of subjects. If the same data resulted from a larger subject pool, the analysis of variance would show a significant effect.

The data show a similarity in the responses at low target values, but a divergence at higher targets (see figure 5.4). This *interaction* [22] between prominence and target value, in which the effect of prominence at higher target values may be more pronounced than at lower targets, can also be tested with analysis of variance. The analysis of variance shows that there is a significant interaction between prominence and target pitch ($F = 8.764$ at $p < 0.001$).

The data show a tendency for the transition from the perception of **L%**'s to **H%**'s to rise dramatically from around four semitones at prominence one to around ten at prominence two, and then to drop again and level off at five to seven semitones at prominence three and four (see figure 5.6). This occurs for four out of the ten subjects. This high percentage suggests that this is not just an anomaly due to the small sample size.

Interviews with the subjects revealed that at high prominence on the word *Mike*, the intonation was confusing and oftentimes meaningless. In these cases, the subjects would listen only to the intonation on the last word and entirely ignore the beginning of the sentence. It appears that listeners who cannot interpret the intonation will revert to a "default" threshold for detecting pitch accents. The data fits with this explanation quite well; the experiment performed in chapter 3 showed that single-word utterances require a pitch excursion of approximately five semitones in order to be perceived as a question.

The lack of a plausible intonational context for the increase in prominence on the word *Mike* may have caused the pitch contour to be interpreted as due to pitch range rather

Individual Responses
Subject: -bc

Target	Number of responses							
	Prominence 1		Prominence 2		Prominence 3		Prominence 4	
	H%	L%	H%	L%	H%	L%	H%	L%
0	0	2	0	2	0	2	0	2
1					0	2	0	2
2	0	2	0	2				
3					0	2	0	2
4	2	0	0	2				
5					1	1	1	1
6	2	0	0	2				
7					1	1	2	0
8	2	0	0	2				
9					1	1	0	2
10			2	0				
11					1	1	0	2
12			2	0				
13					2	0	2	0

Figure 5.6: Example of a subject who may have been confused at high prominences on the pitch accent on *Mike*

than a simple increase in emphasis on this particular word due to, for example, contrastive stress. Perhaps the subjects became confused because their expectation of the speaker's use of pitch range was being calibrated by only one pitch accent.

If subjects were in fact interpreting the increase in prominence as an increase in pitch range, then this experiment was designed improperly. A better test of the effect of pitch range would use more pitch accents to calibrate the subjects' expectations, and would separate the calibration from the contour being evaluated. This can be done by using two sentences instead of one for each stimulus. The first sentence would contain several pitch accents of high prominence, scaled according to pitch range. The targets would be chosen so that interpretations such as contrastive stress would not be possible. The second sentence would vary only at the boundary tone. This design would ensure that pitch range would be the only parameter influencing the subjects' interpretation of the contour.

5.5 Summary

This experiment has attempted to show that the height of preceding pitch accents can affect the perception of pitch accents that come later in an utterance.

This outcome is suggested by the data, but not proven conclusively, possibly because of the small subject pool. However, a significant interaction is shown, in which the effect of prominence is more pronounced at higher target values than at lower ones.

Also suggested by the data is the tendency for listeners who have difficulty interpreting the intonation to revert to a "default" threshold for detecting pitch accents. This default threshold is at approximately five semitones.

Finally, the possibility is discussed that the subjects were interpreting the utterances on the basis of pitch range rather than pitch accent prominence. A better experiment is described that addresses this possibility. This experiment is presented in the next chapter.

Chapter 6

Effects of pitch range expectation on pitch accent perception

6.1 Introduction

Pitch range is the extent to which a speaker varies pitch over a given utterance. A larger pitch range will have higher peaks and deeper valleys in the pitch contour than a smaller range. Speakers use pitch range for various purposes. An increased pitch range often accompanies a heightened emotional state of the speaker, such as happiness or anger [1,24]. It may also serve a discoursal role. Topic shifts and beginnings of units similar to paragraphs are often started off with an increase in pitch range; this decreases until the next topic shift or end of the paragraph-like unit [6].

How might the realization of a semantically salient pitch excursion (i.e. a pitch accent) be influenced by an increase in pitch range? One might imagine that such a pitch excursion must also be increased so as not to be masked by the exaggerated peaks and valleys in the rest of the pitch contour.

The previous experiment showed that the prominence of a given accent can affect the perception of a succeeding accent within the same sentence. Whether this conclusion can be generalized to imply that expected pitch range affects the perception of pitch accents is another matter. An increase in prominence on one pitch accent may not be the same as an increase in pitch range used by a speaker; a single pitch accent may not be enough to cause a listener to recalibrate her expectation of the pitch range used by the speaker. An unusually high prominence on a single pitch accent may simply be misunderstood, rendering the whole pitch contour semantically meaningless.

This experiment, designed to better control for this possibility, investigates how the height of the target of a given pitch accent needs to change as the expected pitch range of the utterance is varied, in order for the excursion to continue to be interpreted as a pitch accent. Stimuli were comprised of two sentences in order to separate the contour used for calibration from that being evaluated. The first sentence, used for calibration, carried pitch accents which were all scaled according to pitch range. The second sentence, which bore the contour being tested, varied only at the boundary tone.

6.2 Description

The following sentences were generated with the pitch accent transcriptions shown (figure 6.1 shows a typical contour). They are listed in order of increasing use of pitch range.

```
1)  (a) Joan painted her car.         (b) She made it pink?
         H*               H* L L%          H*         H* H H%

2)  (a) Mike went to the store.       (b) He bought beef?
         H*                H* L L%         H*        H* H H%

3)  (a) Ann is a programmer.          (b) She uses 'C'?
         H*        H*      L L%            H*       H* H H%

4)  (a) Sue bought a newspaper.       (b) She bought the Globe?
         H*           H*      L L%         H*               H* H H%
```

Each stimulus consisted of both sentence (a) and (b). Sentence (a) always had the same pitch contour; it was used as a "calibration" sentence, to cause the subjects to expect a certain pitch range in the succeeding sentence. It was expected that a subject who anticipated a larger pitch range would require a higher pitch contour trajectory in order to perceive the excursion as a pitch accent.

This was tested with sentence (b); the target pitch of the final boundary tone in sentence (b) was changed for each stimulus. The subjects' task was to decide how the second sentence (b) would be punctuated, if it were to be transcribed. A forced-choice response format was used; the subjects could pick either a question mark (?), period (.), exclamation point (!), or ellipsis (...). The responses were divided into questions and non-questions. Those who chose question marks had perceived an **H%** phrase boundary; those who chose anything else perceived an **L%**. It was expected that it would take a higher target pitch for an **H%** to be perceived as the pitch range of sentence (a) rose.

The stimuli were randomly ordered. Stimuli presentation was under the subjects' control, via computer. Each stimulus was presented twice. Each stimulus could be heard as many times as desired, but only one response per presentation was accepted. Subjects' responses were recorded via computer.

6.2.1 Method

Stimuli

To produce the different pitch ranges for the (a) sentences, each one was spoken four times, with increasing amounts of prominence on the accented words. One recitation of each sentence was chosen such that there were four different sentences, each with a different use of pitch range.

In the first of these sentences, the peak pitch on the first pitch accent was at 157 Hz, and the peak pitch on the second pitch accent was at 129 Hz. Sentence two had peaks of 200 and 157 Hz, resp; sentence three at 200 and 195 Hz, and sentence four at 229 and 205 Hz, resp. As a measure of the relative amounts of pitch range used in each of these sentences, the average of the two peaks was taken and compared to sentence one. Sentence one's average pitch is 143 Hz; sentence two's is four semitones above this, at 178.5 Hz; sentence three's is six semitones above (197.5 Hz), and sentence four's is seven semitones above the reference, at 217 Hz.

To produce the different contours for the (b) sentences, each one was spoken once. The contours on the last word were then modified by imposing a linear trajectory on the

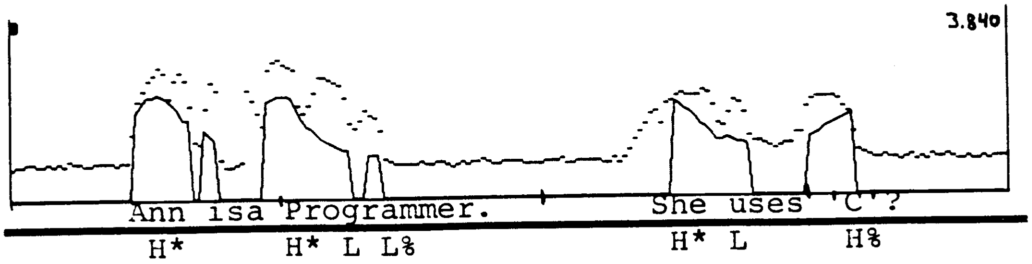

Figure 6.1: Pitch and energy of sentence 3(a), *Ann is a programmer.*, and 3(b), *She uses 'C'?*

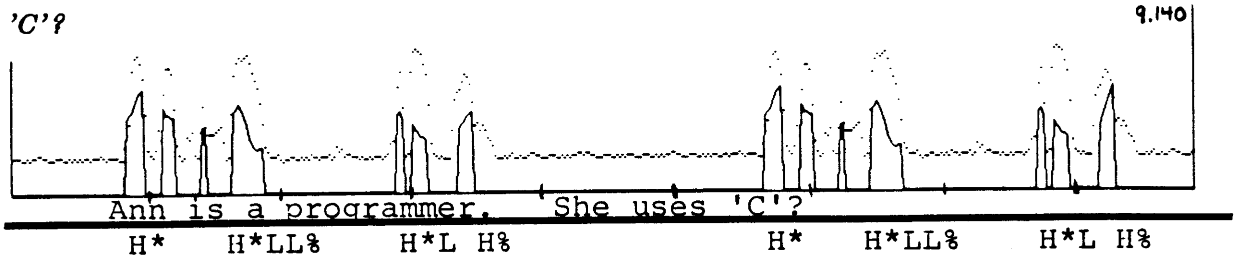

Figure 6.2: Examples of the stimuli. The target pitches are five and eleven semitones above monotone.

pitch, beginning at the same point as the pitch excursion began in the original utterance, and continuing to the target pitch at the termination of voicing, as in chapter 3. The target pitch ranged from one to 19 semitones above monotone, in odd semitone steps. This produced 40 stimuli, each of which were presented twice, making a total of 80 responses per subject. Figure 6.2 shows some sample stimuli.

Stimuli collection and generation

The stimuli were recorded on analog cassette tape and then encoded in LPC. The recorder was a Nakamichi MR-2 with the following settings: Bias: normal; Equalization: 120us; Noise Reduction: Dolby-B. The microphone was a Shure SM12A noise-canceling headset, running through a Shure M267 mixer.

The stimuli were encoded and resynthesized in LPC with a Texas Instruments TI Speech processing card running on an IBM PC. Pitch modification was performed via Pitchtool, an interactive LPC editing facility [12].

Environment

Each subject was seated in a comfortable chair, approximately 12 to 24 inches from a Sun workstation, in a acoustically-treated room. The stimuli were re-synthesized in real-time, run through an Audioarts Engineering 4200A parametric equalizer to eliminate noise generated by the speech card, and amplified by a Crown D-75. A pair of speakers, Rogers LS3/5A Monitors, were placed on either side of the subject at approximately 1.5 meters. The volume was adjusted according to the subjects' preference.

After prompting the subjects for their initials, the computer displayed a Sun-window showing six buttons:

- **Play/Repeat**, which caused a stimulus to be presented;
- **Question Mark? (!?)**, which recorded the subjects' responses;
- **Period.**, which recorded the subjects' responses;
- **Exclamation Point!**, which recorded the subjects' responses;
- **Ellipsis...**, which recorded the subjects' responses; and
- **Back up**, which allowed the subjects to correct a mistaken response.

Once initials were entered, the subject interacted with the computer solely through the use of the mouse.

The subjects were instructed that clicking on **Play/Repeat** would cause one of the stimuli to be heard, and that their task was to determine how the stimuli were punctuated, if they were to transcribe each sentence. They were told that they could click on **Play/Repeat** as many times as necessary to make a determination, but that it was usually best to base their judgements on their first impressions.[1] In addition, the interpretation of a question mark was described as suggesting any amount of uncertainty, for example disbelief or incredulity as well as simple questioning. Ellipsis was described as if the statement were incomplete, but was about to be completed, for example "Mike bought beef... fruit, and vegetables." It was suggested that the subjects try to imagine such a scenario each time a stimulus was presented.

All subjects were told that only one response was accepted per stimulus. In the event that a mistake was made, they were told that they could repeatedly click on **Back up** to go back as far as necessary to correct the response that they had made.

[1] The number of times a stimuli was repeated was not recorded; this facility was included solely to reduce performance pressure on the subjects.

6.2.2 Data analysis

At each target pitch and pitch range, the number of responses for each question were summed and converted into percentages. A plot of percent question responses vs. pitch range used in the first sentence was made for each target pitch of the final boundary tone. An analysis of variance was performed to determine the significance of the effect of pitch range on the magnitude of the excursion required for the perception of an **H%**.

It was expected that a lower number of subjects would perceive an **H%** accent on the last word (i.e. perceive the second sentence as a question) as the pitch range of the first sentence increased.

6.3 Results

Ten subjects participated in the study. Each subject responded to each stimulus twice, for a total of twenty responses per stimulus.

Figure 6.3 shows the percent of subjects who perceived an **H%** accent on the last word of sentence (b) (i.e. those who interpreted sentence (b) as a question) vs. those who perceived an **L%**. The columns labeled **Pitch Range 1** through **Pitch Range 4** correspond to sentence pair 1 through 4, that is, each column corresponds to a particular pitch range used in each of sentence (b) (sentence (a) is the same for each column). The number in the column labeled **Target** indicates the number of semitones above monotone (labeled zero, indicating no pitch change) reached at the end of sentence (b).

Figure 6.4 shows this data plotted as percent of subjects who perceived an **H%** boundary tone vs. target pitch on the last word of sentence (b). Each curve corresponds to the particular amount of pitch range used in sentences one through four.

Figure 6.5 shows this data plotted as percent **H%** perception vs. amount of pitch range used. Each curve corresponds to a particular target pitch reached on the last word of the (b) sentences. The abscissa lists the relative pitch range used in the (a) sentences, in semitones. Sentence one corresponds to the points plotted at zero semitones (because it is the reference against which the other sentences are compared); sentence two is four semitones above the reference, sentence three is six above, and sentence four is seven semitones above the

Percentage Responses
Total number of subjects: 10

Target	Percent of responses							
	Prominence 1		Prominence 2		Prominence 3		Prominence 4	
	H%	L%	H%	L%	H%	L%	H%	L%
1	10	90	0	100	0	100	0	100
3	35	65	0	100	5	95	0	100
5	50	50	15	85	5	95	25	75
7	65	35	10	90	0	100	10	90
9	85	15	15	85	20	80	30	70
11	100	0	60	40	40	60	60	40
13	100	0	80	20	75	25	60	40
15	95	5	85	15	100	0	75	25
17					95	5	90	10
19					100	0	90	10

Figure 6.3: Percentage of subjects who perceived the indicated accent on the last word of sentence (b), for each of the indicated pitch ranges.

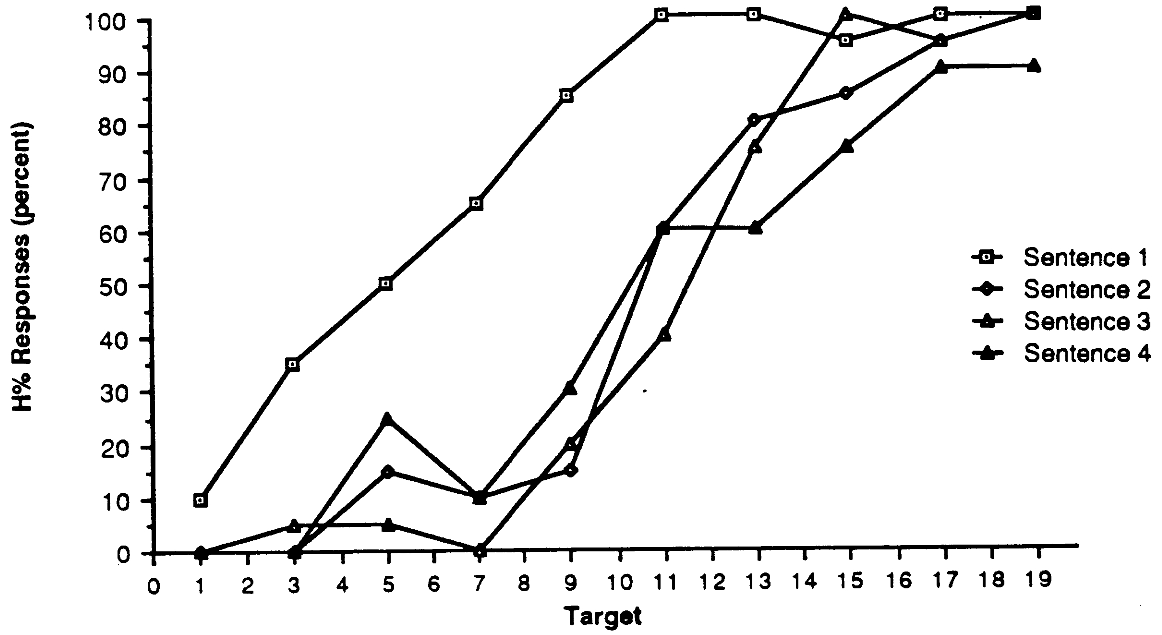

Figure 6.4: Plot of the percentage of subjects who perceived an H% boundary tone vs. target pitch on the last word of the (b) sentences. Each curve corresponds to the particular amount of pitch range used in sentences one through four.

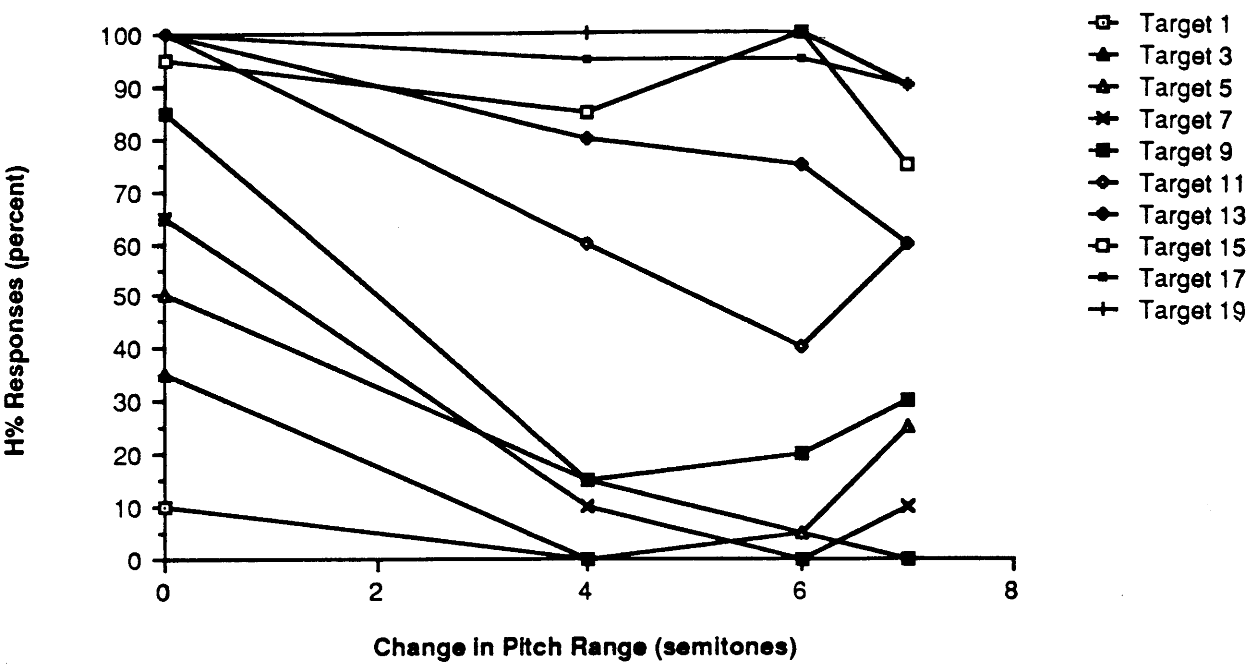

Figure 6.5: Plot of the percentage of subjects who perceived an **H%** boundary tone on the last word of sentence (b), vs. the pitch range used in sentence (a). Each curve corresponds to a particular target reached on the last word.

reference.

It is evident that the target pitch of the boundary tone needs to be increased as the expected pitch range increases, in order for the subjects to perceive an **H%** accent. At pitch range one, 75% of the subjects perceived an **H%** at eight semitones; twelve semitones were required at pitch range two, thirteen at pitch range three and fifteen at pitch range four. This increase in target required as the pitch range went up is confirmed by the analysis of variance; F is 13.306 at $p < 0.001$.

6.4 Discussion

The results show that pitch range expectation, once established, can continue across sentence boundaries and affect the perception of a pitch accent in the next sentence.

The effect of expected pitch range is much stronger than the effect of prominence discussed in chapter 5. Although the same number of subjects participated in both investigations, the effect of prominence failed to reach significance, while the effect of expected pitch range is clearly significant.

Close analysis of the data shows a tendency for a high incidence of **H%** judgements at the relatively low target of five semitones (see figure 6.4). This occurs for four out of the ten subjects; figure 6.6 is an example of such an occurrence. This high percentage suggests that this is not just an anomaly due to the small sample size.

This tendency can be explained in terms of *naturalness*. All of the second (b) sentences were originally generated with minimal use of pitch range and a final excursion of four to six semitones. Since the excursion required in neutral sentences has been shown to be approximately six semitones (see chapter 5), one would expect that, all other things being equal, more subjects would make **H%** judgements at approximately five semitones. It is possible that, even though the subjects' pitch-range expectations were being manipulated by the first sentence of the pair (sentence (a)), cases in which the utterances were most natural might have elicited an **H%** judgement.

The data support this hypothesis. The least effect on the judgements at five semitones is at the smallest expected pitch range, and the largest effect is at the largest pitch range. One would expect the largest effect to correlate with the largest pitch range because the utterance with the largest pitch range has the highest possibility of being misinterpreted or misunderstood than the lower ranges.[2] In fact, if we ignore the data from pitch range four, the data are actually quite orderly.

Eliminating pitch range four from the analysis may be valid from another point of view as well: this data is skewed when compared to all of the other pitch ranges. That is, below

[2] While care was taken to ensure that sentence (4a) sounded natural prosodically, the high pitch range used might not have actually sounded so because of the prosodic context in which the stimuli were embedded. It may well be that in another context, the high pitch range would have been completely legitimate.

Individual Responses
Subject: mjs

Target	Number of responses							
	Prominence 1		Prominence 2		Prominence 3		Prominence 4	
	H%	L%	H%	L%	H%	L%	H%	L%
1	0	2	0	2	0	2	0	2
3	0	2	0	2	0	2	0	2
5	2	0	1	1	0	2	2	0
7	1	1	0	2	0	2	1	1
9	2	0	0	2	0	2	1	1
11	2	0	0	2	0	2	2	0
13	2	0	2	0	1	1	2	0
15	2	0	2	0	2	0	2	0
17					2	0	1	1
19					2	0	2	0

Figure 6.6: Example of a subject who made a large number of H% responses at five semitones.

ten semitones, the data is similar to that which would be expected from a pitch range between one and two, and above ten semitones as if the expected pitch range were four or higher (see figure 6.4). Pitch range one through three, on the other hand, are nearly parallel, in that almost all of the points in pitch range one are above those of pitch range two, which are nearly all above pitch range three.

The skew of pitch range four can be explained as follows: below ten semitones, the effect of naturalness (described above) takes precedence over expected pitch range; above ten semitones, the excursion required to influence the subjects into an **H%** judgement may become so steep that the contour begins to sound unnatural. In addition, the short duration of voicing during the word *Globe* in sentence (4b) eliminated a plateau in pitch which often accompanies such steep trajectories, decreasing the chance for the proper recognition of the final target.

In order to eliminate such problems, the pitch range in sentence (4a) should not be so high as to sound unnatural in the given prosodic context, and care should be taken to ensure that the final word not end in an unvoiced consonant, especially a stop. A listening test should be performed to ensure that the stimuli are as natural as possible prior to their inclusion in such an experiment. In addition, subjects should be provided with a means to qualify their responses with a rating of the naturalness of the stimuli, in order to provide justification for conclusions such as those drawn above.

6.5 Summary

It has been shown that pitch range expectation affects the perception of pitch accents. As the pitch range expected by the listeners increases, the target pitch required for the perception of a pitch accent also increases.

As in chapter 5, a "default" threshold near five semitones tends to be preferred by subjects under less than ideal conditions. In this case, the unnaturalness of the stimuli appears to be influencing the subjects' responses toward this default.

Finally, suggestions are made that would eliminate unnatural-sounding stimuli and their associated problems. Stimuli must be generated that permit sufficient time for pitch excursions to be executed and recognized. Pre-tests should be carried out to ensure that the stimuli are not unusual, and a means for subjects to qualify their responses should be provided.

Chapter 7

Summary and Future Directions

7.1 Summary

The experiments that have been performed have investigated the pitch excursions required to cause the perception of pitch accents in various prosodic environments.

It was found that an excursion of approximately five semitones can cause a change in the perception of the identity of a pitch accent, from an **H** to an **L**. This value seems to be a kind of "magic number" that appears in a variety of prosodic environments; not only is this the magnitude of the excursion required when there are no other prosodic or lexical cues to the presence or absence of a pitch accent, but it seems to arise in cases in which the prosodic or lexical cues are misleading or confusing. Perhaps this is the "natural" excursion required for pitch accent perception; listeners tend to prefer this value when other cues break down.

When not too extreme, the prosodic environment was shown to affect the magnitude of this excursion. When the prominence of a pitch accent is increased, the excursion required on the next accent also appears to increase.[1] This influence extends to the use of pitch range as well; when the prominence on many preceding pitch accents causes a listener to expect a higher pitch range, then the magnitude of a pitch excursion that signals a pitch accent must also increase.

[1] This result is not proven conclusively here, due to the small number of subjects who participated in this experiment.

Also demonstrated was the effect of lowered topline described by Pierrehumbert [15], in which the pitch range is compressed throughout the course of an utterance. As a consequence of this, a smaller excursion is required to convey the presence of a pitch accent at the end of a sentence than at the beginning. This result implies that listeners expect and compensate for this lowering of topline.

7.2 Future Directions

In addition to prosodic expectation, subjects' expectations with respect to the experimental environment also affected the outcome of the investigations. Care must be taken in the design and presentation of experiments to eliminate the possibility of misinterpreted instructions and mis-entered responses, as well as interactions between succeeding stimuli. As time went on, these experiments were better designed to account for situations like this. Instructions given to the subjects were more clear, concise, and consistent. The "Back-up" button not only reduced subject anxiety, but provided higher confidence in the data collected. The "Can't Tell" button was removed in later experiments because of the possibility for misuse. In retrospect, it probably should have been left in at least as an opportunity for subjects to qualify their responses. For example, the button might have been renamed "Unsure", and the subjects would have been able to use this button in addition to their choice of response. A more standard approach would have been to include a separate facility for the subjects to indicate their confidence in their responses, perhaps though a numerical scale of some sort. This would certainly have given more information regarding responses in the instances where pitch excursions were too extreme.

As always, one inquiry begets ten more. These investigations just barely scratch the surface of the nature of the perception of pitch accents. A more thorough coverage of the perception of pitch needs to include an exploration of the effects due to articulation and their interaction with prosody. Intrinsic pitch, the effect of consonant environment, and the combination of the two, as well as the magnitude of semantic pitch excursions that are not represented by changes in the pitch accent sequence are some areas that should be investigated.

The effect of IF0 on the magnitude of the excursion needed for the perception of the presence or absence of pitch accents should be one of the first investigations to pursue. Significant literature exists on the intrinsic pitch of vowels [23,7,21] that can be used to design experiments similar to the ones described here. For example, Reinholt Petersen's findings that IF0 is smaller in unstressed syllables than stressed ones, and that this effect decreases during the course of an utterance [19] should be correlated with the interactions of IF0 on the perception of pitch accents.

The effect of consonant environment on pitch contours is another area that should be explored. House and Fairbanks' findings concerning the effect of consonant environment on the pitch of vowels [7] needs to be verified and then incorporated into the studies. Then the effects of the combination of consonant environment and IF0 should be investigated.

A slight digression away from pitch accents is also appropriate. Since the point of these experiments is to determine which pitch excursions have semantic significance, it makes sense to investigate how meaning changes as a pitch contour is changed in ways other than the addition or removal of pitch accents. For example, the difference between a simple question and an incredulous question must be due to prominence, if the lexical content and the pitch accent sequence are unchanged. Investigating the amount of prominence change required to cause a change in meaning is the same as determining the *just-noticeable difference* (JND) of intonation [20]. Knowing the JND of intonation would have helped prevent the unnatural-sounding stimuli discussed in chapters 5 and 6. The exceedingly high prominence and pitch range used in these stimuli occurred because the step size from low to high prominence was chosen arbitrarily. If the JND were known then a better step size might have been chosen.

Once the JND of pitch contours is determined, its magnitude should be compared to the size of other pitch perturbations such as IF0, consonant environment, and pitch accent excursions. The effects of these perturbations on JND's should also be investigated.

Finally, all of these experiments need to be repeated with emphasis on their effects on **L** accents. Each of the experiments performed or described have dealt primarily with the perception of **H** pitch accents. In order to fully investigate the perception of intonation in terms of Pierrehumbert's phonology, both **H** and **L** pitch accents need to be considered.

It is hoped that the investigations described herein have shed some light on the relation of the phonology of intonation and the perception of intonation, and that this information will prove useful in future endeavors in the implementation of speech understanding systems.

Chapter 8

Acknowledgements

I am deeply grateful to the following people, without whom this thesis would never have been completed.

To Tod Machover, for his undying support and enthusiasm; Barry Vercoe, who took his advisorship to heart; and Chris Schmandt, who kept me to the path, yet gave me the opportunity to follow my instincts.

I'd also like to thank my readers, Professors Campbell Searle, Barry Vercoe, and especially Dennis Klatt, for their invaluable insight and advice.

And of course, the wonderful people of the Media Lab: To Janet and Janette, who kept me from going insane, Megan and the tbc, who helped me go insane when I needed it; Joe for shared frustration, DC for shared secrets, and Kristy for sharing the Thesis Monster with me; Bill for being so sane, and Paul for understanding distance; Steph for making me laugh, helping me cry, and showing me the beauty in things; JH for the treks, ML for the insight, and Kenji for the scum cool times. Thanks to all of you, and all those too numerous to mention!

I made it!!

-- KAEL

This work was supported by the Defense Advance Research Projects Agency (DARPA) contract number N00039-86-C-0406.

Bibliography

[1] Gary Collier. *Emotional Expression*. Lawrence Erlbaum Associates, 1985.

[2] Kristy L. Dowers. *The effects of intonation th the comprehension of connected speech.* Thesis, MIT Media Lab, 1988.

[3] D. B. Fry. Experiments in the perception of stress. *Language and Speech*, 1(2):126–152, 1958.

[4] K. Hadding-Koch and M. Studdert-Kennedy. An experimental study of some intonation contours. *Phonetica*, 11:175–185, 1964.

[5] J. Hirschberg and D. Litman. Now let's talk about now: identifying cues phrases intonationally. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, pages 163–171, July 1987.

[6] Julia Hirschberg and Janet Pierrehumbert. The intonational structure of discourse. In *Proceedings of the Association for Computational Linguistics*, pages 136–144, July 1986.

[7] Arthur S. House and Grant Fairbanks. The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.*, 25(1):105–113, January 1953.

[8] Dennis H. Klatt. Review of test-to-speech conversion for english. *Journal of the Acoustic Society of America*, 82(3):737–793, 1987.

[9] Klaus J. Kohler. F0 in the perception of lenis and fortis plosives. *Journal of the Acoustic Society of America*, 78:21–32, 1985.

[10] John Laver and Robert Hanson. Describing the normal voice. In Darby, editor, *Speech Evaluation in Psychiatry*, pages 51–78, Grune and Stratton, Inc., 1981.

[11] Wayne A. Lea. Prosodic aids to speech recognition. In Wayne A. Lea, editor, *Trends in Speech Recognition*, chapter 8, Prentice Hall, 1980.

[12] Michael McKenna. *Pitchtool, an interactive, graphical tool for editing digitized speech pitch.* Annual report to NTT, MIT Media Lab, 1987.

[13] J. Pierrehumbert and J. Hirschberg. The meaning of intonation contours in the interpretation of discourse. In *Plans and Intentions in Communication*, SDF Benchmark Series in Computational Linguistics, MIT Press, forthcoming.

[14] Janet B. Pierrehumbert. Automatic recognition of intonation patterns. In *Proceedings of the 21st conference*, pages 85–90, Association for Computational Linguistics, 1983.

[15] Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, MIT, Dept of Linguistics, 1980.

[16] Janet B. Pierrehumbert. Synthesizing intonation. *Journal of the Acoustic Society of America*, 985–995, Oct 1981.

[17] Janet B. Pierrehumbert and M. Liberman. Course notes. CSLI, June 1984.

[18] David B. Pisoni, Howard C. Nusbaum, and Beth G. Greene. Perception of synthetic speech generated by rule. *Proceedings of the IEEE*, 73(11):1665–1676, Nov 1985.

[19] N. Reinholt Petersen. *Variation in inherent F0 level differences between vowels as a function of position in the utterance and in the stress group.* Technical Report 13, Institute of Phonetics, University of Copenhagen, 1979.

[20] Thomas Rossing. *The Science of Sound.* Addison-Wesley, 1982.

[21] Kim E. A. Silverman. *Natural prosody for synthetic speech.* PhD thesis, Cambridge Universtity, 1987.

[22] Joan Gay Snodgrass, Gail Levy-Berger, and Martin Haydon. *Human Experimental Psychology.* Oxford University Press, 1985.

[23] Shirley A. Steele. Interaction of vowel f0 and prosody. *Phonetica*, 43:92–105, 1986.

[24] Carl E. Williams and Kenneth N. Stevens. Emotions and speech: some acoustical correlates. *JASA*, 52(4 (Part 2)):1238–1250, 1972.

[25] Victor H. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting*, pages 567–578, Chicago Linguistics Society, 1970.