# THE INTONATIONAL STRUCTURING OF DISCOURSE

Julia Hirschberg and Janet Pierrehumbert

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974 USA

## ABSTRACT

We propose a mapping between prosodic phenomena and semantico-pragmatic effects based upon the hypothesis that intonation conveys information about the intentional as well as the attentional structure of discourse. In particular, we discuss how variations in pitch range and choice of accent and tune can help to convey such information as: discourse segmentation and topic structure, appropriate choice of referent, the distinction between 'given' and 'new' information, conceptual contrast or parallelism between mentioned items, and subordination relationships between propositions salient in the discourse. Our goals for this research are practical as well as theoretical. In particular, we are investigating the problem of intonational assignment in synthetic speech.

## 1. Introduction

The role of prosody in discourse has been generally acknowledged but little understood. Linguistic pragmaticists have noted that types of **information status** (such as **given/new**, **topic/comment**, **focus/presupposition**) can be intonationally 'marked' [1, 2, 3, 4], that **reference resolution** may depend critically on intonation [5, 6], that intonation can be used to disambiguate among potentially ambiguous utterances [7, 8], and that **indirect speech acts** may be signalled by intonational means [9, 10, 11]. Conversational analysis of naturally occurring data has found that speakers may signal **topic shift**, **digression**, and **interruption**, as well as **turn-taking**, intonationally [12, 13, 14]. And the fact that intonational contours contribute in some way to utterance interpretation is itself unexceptionable [8]. To date, however, identification of the prosodic phenomena involved -- and the proper mapping between these phenomena and their semantico-pragmatic effects -- has been largely intuitive, and the intonational phenomena involved have not been precisely described.

Here, we describe how certain of the resources of the intonational system are employed in discourse. In particular, we discuss how speakers' choice of **pitch range**, **accent**, and **tune** contribute to the **intentional** and **attentional** structuring of discourse -- the way speakers communicate the relationships among their discourse goals and the relative salience of entities, attributes, and relationships mentioned in the discourse.[1] Our findings emerge from an intensive study of a simple example of speech synthesis: the script of a computer-aided instruction system, TNT (Tutor 'n' Trainer) [16], which employs synthetic speech to tutor computer novices in the text editor *vi*. Using the Text to Speech system (TTS) [17], we have been able, by systematic variation of pitch range and by a principled choice of accent and tune, to highlight the structure of the tutorial text and thus to enhance its coherence. While most studies of how intonation is used in discourse have been based solely on examination of intonational contours found in a natural corpus, we have found that intonation synthesis provides a unique opportunity to manipulate the dimensions of variation orthogonally. Thus we can pinpoint factors crucial for a given effect and evaluate various patterns for a given utterance and context.

## 2. The Domain

TNT was designed to teach computer-naive subjects *vi*, a simple UNIX screen-oriented text editor. The tutorial portion provides a brief introduction to word processing, to general features of *vi*, and to the tutor's help facilities; the tutor then guides subjects through a series of learning tasks of graduated difficulty. While the overall task structure is implicit in the tutorial text, the subject can influence the course of the interaction via his/her manipulation of a set of 'helper' keys; these keys provide hints (HINT) and reminders (REMIND) as well as the option of starting a task over again (DO OVER) or suspending the tutorial temporarily (HOLD).

The fact that TNT is explicitly task-oriented,[2] makes it a good test-bed for our purposes. An appropriate segmentation of the text, and a notion of the purpose of each segment and the hierarchical relationships among segments, can be independently determined from the task at hand. Also, certain characteristics of the text presented a particularly interesting challenge for our study. First, the script contains little pronominal reference and very few so-called **clue words** -- words and phrases such as *now*, *next*, *returning to*, *but*, and *on the other hand*, which can identify discourse segment boundaries and relationships among segments, signal interruptions and digressions, and so on [19, 20]. Both of these phenomena (together with intonation) have been identified as important strategies for communicating discourse structure [15, 18, 19]. Their virtual absence from the text presents a convenient opportunity for testing the power of intonation to structure a discourse. Second, while we were not able to isolate points in the text where subjects had special difficulties, we did informally observe certain general problems with **turn-taking**[3] in the tutor -- specifically, it was not always clear when the tutor's *turn* was over -- which we addressed in our synthesis of the text.

## 3. F0 Synthesis

To synthesize the fundamental frequency (f0) contours for the TNT script, we used the intonation synthesis program

1. Grosz and Sidner [15] propose a tripartite view of discourse structure: a **linguistic structure**, which is the text/speech itself; an **attentional structure**, including information about the relative **salience** of objects, properties, relations, and intentions at a given point in the discourse; and an **intentional structure**, which relates **discourse segment purposes** (those purposes whose recognition is essential to a segment achieving its intended effect) to one another.

2. That is, the tutorial is organized around a series of data processing tasks, which the subject is guided through. See [18] for discussion of the characteristics of task-oriented domain discourse.

3. The process by which speakers signal that they have (temporarily) finished speaking and by which hearers interpret such signals [21].

described in [22] in [23, 24]. It permits explicit control over the different dimensions of variation in the intonation system. The dimensions we will discuss here are **phrasing**, **pitch range**, **accent location**, and **tune**. We illustrate each in our synthesis of the introduction to TNT:

```
1.              Hello.
    T150 F.96       H* L L%
2.              Welcome to word processing.
    T150  F.96  H*          H*              L L%
3.              That's using a computer to write letters and
    T136 F.90     H*   H*          H*          H*    H*
                  reports.
                    H*   L L%
4.              Word processing makes typing easy.
    T136 F.96   H*              H*    H*      H*  L L%
5.              Make a typo?
    T125          L*    H*  H H%
6.              No problem.
    T115 F.96   H*  H*    L L%
7.              Just back up,       type over the mistake,
    T115 F.96         H*  H* L H%   H*  H*                L H%
                  and it's gone.
                        H*  L L%
8.              And,      it eliminates retyping.
    T125        H*  L H%      H*            H*    L L%
9.              Need a second draft?
    T115          L*    L*      H*  H H%
10.             No problem.
    T115 F.96   H*  H*    L L%
11.             Just change the first,  and you've got the
    T115 F.87   H*    H*          H*   L H%          H*
                   second.
                    H*     L L%
12.             Today,      the computer will teach you word
    T150 F.96     H* L H%         H*           H*         H*
                  processing.
                             L L%
13.             The computer is new at this,   so be a good
    T136 F.90                    H*       H* L H%          H*
                  student and give it a chance.
                    H*           H*          H*  L L%
14.             We can't answer questions,    if you are
    T136 F.96   H* H*             H*      L H%
                  confused.
                     H*    L L%
15.             We have to let the computer do all the
    T125 F.93              H*           H*       H*
                  teaching.

                    H*      L L%
16.             But if the computer is not working right,     we
    T125 F.87            H*              H*            L H%  H*
                  will help you out.
                         H*       H* L L%
```

**Figure 1.** The TNT Introduction

In Figure 1 and in all figures below, '**T**' indicates the top of the pitch range in Hz, '**F**' indicates amount of compression of the pitch range at the end of declarative phrases, '**H**' and '**L**' indicate high and low tones, '*' indicates a tone's alignment with a stressed syllable, and '*%*' indicates a phrase boundary tone. We discuss these phenomena and our notational system in more detail below.

### 3.1 Phrasing

The first dimension of variation, **phrasing**, may be indicated by a pause, by a lengthening of the phrase-final syllable, and by the occurrence of extra melodic elements on the end of the phrase. Variation in phrasing is illustrated in Figures 2 and 3.[4] In Figure 2, line 8 is produced as a single phrase, whereas in Figure 3, *And* is set off as a separate phrase.

One consequence of this strategy is that *And* becomes more prominent in the second version. Phrasing variation will not be of central concern here. Because of the syntactic simplicity of TNT, there were only a few cases where the phrasing could be varied in interesting ways.

4. Note that phonetic transcriptions given in these and subsequent figures represent the somewhat eccentric output of the TTS system.

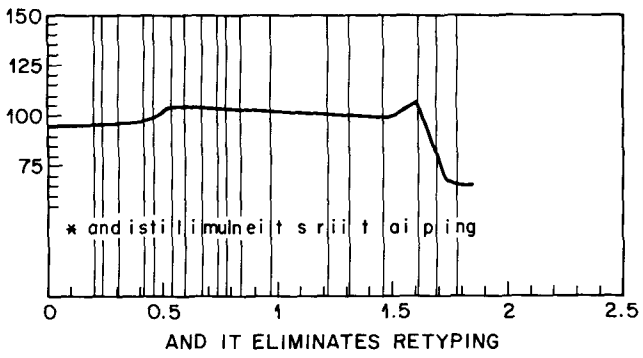



**Figure 2.** One Phrase

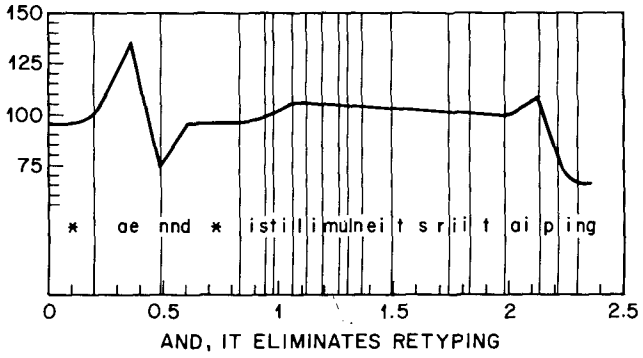



**Figure 3.** Two Phrases

### 3.2 Pitch Range

When a speaker raises his/her voice, his/her overall **pitch range** -- the distance between the highest point in the f0 contour and the speaker's **baseline** (defined by the lowest point a speaker realizes over all utterances) -- is expanded. Thus, the highest points in the contour become higher and other aspects are proportionately affected. Figure 4 shows an f0 contour for line 1 in the script above in the default pitch range used by TTS.

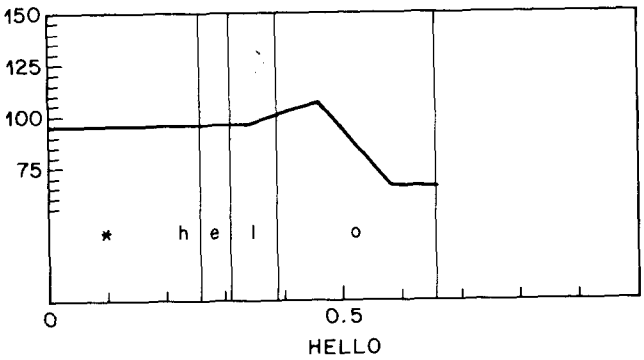



**Figure 4.** TTS Default Pitch Range

Figure 5 shows the contour actually used in synthesizing the TNT script.

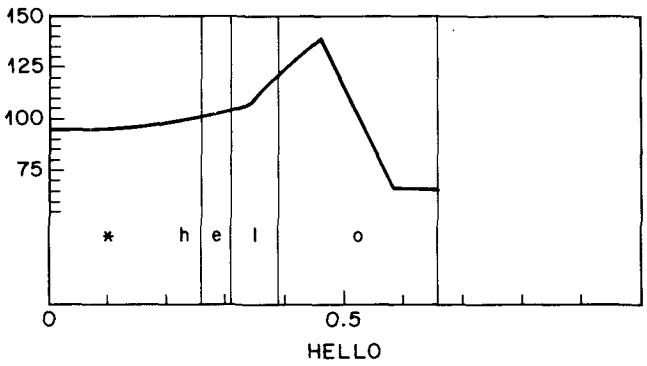



**Figure 5.** Actual Pitch Range

The shape of the actual contour is the same as in Figure 4 but its scaling is different. Changes in pitch range appear to reflect the overall structure of the discourse, with major topic shifts marked by marked increases in pitch range.

In addition to variations in overall pitch range, the intonation system exploits a local time-dependent type of pitch range variation, called **final lowering**. In the experiments reported in [24], it was found that the pitch range in declaratives is lowered and compressed in anticipation of the end of the utterance. Final lowering begins about half a second before the end and gradually increases, reaching its greatest strength right at the end of the utterance. This phenomenon appears to reflect the degree of 'finality' of an utterance; the more final lowering, the more the sense that an utterance 'completes' a topic is conveyed. Contrast Figures 6 and 7.

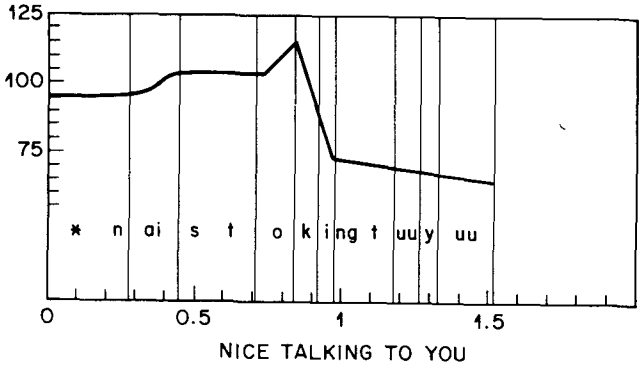



**Figure 6.** With Final Lowering

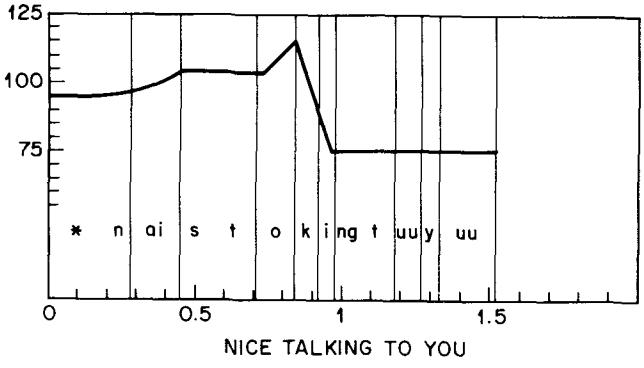



**Figure 7.** Without Final Lowering

In the notational system employed here, **T** represents the **topline**, of a phrase -- the maximal value for the f0 contour in the phrase. **F** expresses the amount of final lowering in terms of the ratio of the lowered pitch range to the starting pitch range. The default value assumed below for **T** is 115 Hz and for **F** is 0.87.

### 3.3 Accent

**Pitch accents**, which fall on the stressed syllable of lexical items, mark those items as intonationally prominent. In line 16, for example, *right* has no pitch accent. If *right* were to be especially emphasized, it would have an accent. (In our notation, the absence of a specified accent indicates that a word is **not** accented; where we wish to highlight this point, we will employ '-' to mark a deaccented word.) The contrasting outcomes are shown in Figures 8 and 9.

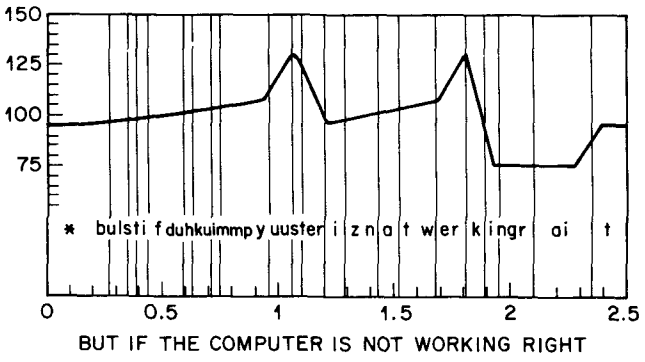



**Figure 8.** *Right* Deaccented

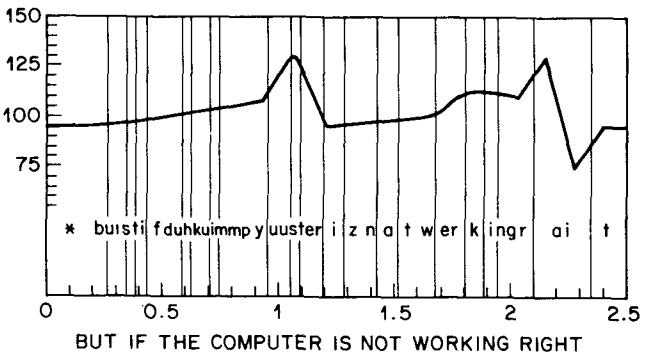



**Figure 9.** *Right* Accented

In the first case, the last f0 peak occurs on *work* and there is a fall to a low pitch on *right*, then a rise at the end of the phrase. In the second case, the entire peak-fall-rise configuration occurs on the word *right*.

There are six types of pitch accent in English [23], two simple tones -- high and low -- and four complex ones. The most frequently used accent, the simple high tone, comes out as a peak on the accented syllable (as, on *right* in Figure 9) and will be represented below as **H***. The 'H' indicates a high tone, and the '*' that the tone is aligned with a stressed syllable. In some cases, we have used a **L*** accent, which occurs much lower in the pitch range than **H*** and is phonetically realized as a local f0 minimum. The accent on *make* in Figure 13 below is a **L***. The other English accents have two tones. Figure 10 shows a version of the sentence in Figures 2 and 3 with a **L+H*** accent substituted for both **H*** accents in the second phrase.

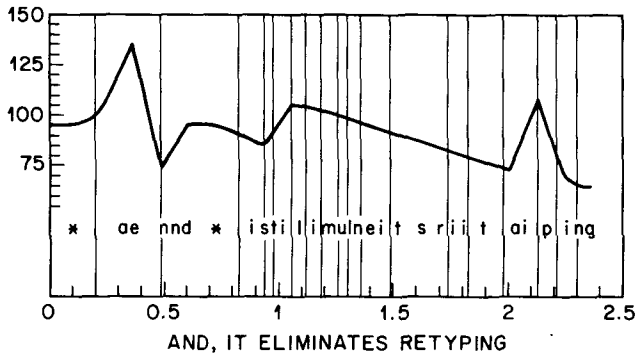



**Figure 10.** An L+H* Accent

Note that there are still peaks on the stressed syllables, but now a striking valley occurs just before each peak.

In our synthesis of the TNT script, we have made extensive use of the type of accent transcribed in [23] as **H*+L**. This accent, like other bitonal accents, triggers a rule which compresses the pitch range on following material in the phrase, a phenomenon known as **downstep** or **catathesis**. For example, a simple contrast between **H* H*** and **H*+L H*+L** is illustrated in Figures 11 and 12 in two versions of the tutorial command to hit the 'remind' helper key -- *Hit remind.*

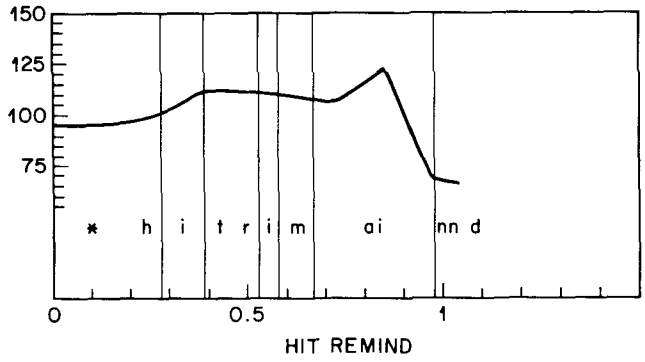



**Figure 11. H* H* L L%**

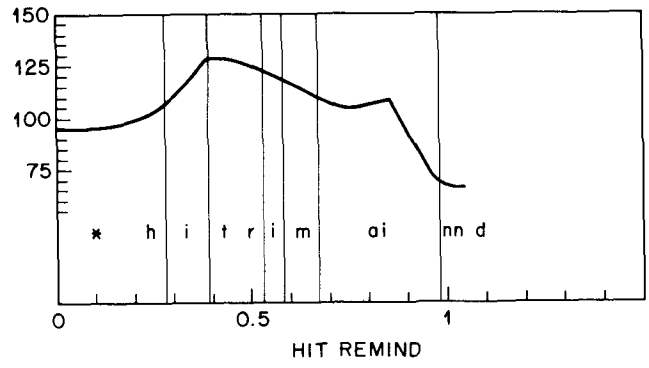



**Figure 12. H*+L H*+L L L%**

We have made particular use of downstepped contours such as this -- i.e., sequences of **H*+L** tones -- which we will term **H*+L sequence** in the discussion below. (See Section 4.3.) The way a speaker is structuring a text helps to determine where pitch accents will fall, as a speaker indicates how referents of accented or deaccented items are related to other items in the utterance or in some larger context.

In addition to pitch accents, each intonational phrase has a **phrase accent** and a **boundary tone**. These two extra tones may be either **L** or **H**. The boundary tone (indicated by '**%**') falls exactly at the phrase boundary, while the phrase accent (indicated by an unadorned **H** or **L**) spreads over the material between the last pitch accent and the boundary tone. Each intonational phrase contains one or more pitch accents, a phrase accent, and a boundary tone.

### 3.4 Tune

A phrase's **tune** or **melody** is defined by its particular sequence of pitch accents, phrase accent, and boundary tone. Thus, **H* L L%** represents a tune with a **H*** pitch accent, a **L** phrase accent, and a **L%** boundary tone. This is an ordinary declarative pattern with a final fall. A interrogative contour is represented by **L* H H%**. The contrast between these two melodies is illustrated in Figures 13 and 14. Figure 13 shows the actual f0 contour for line 5 of the TNT introduction, produced as a question.

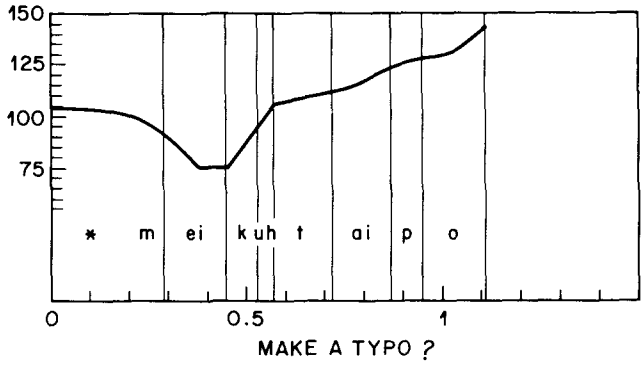



**Figure 13.** Interrogative Contour

Figure 14 shows a declarative pattern for the same sentence.

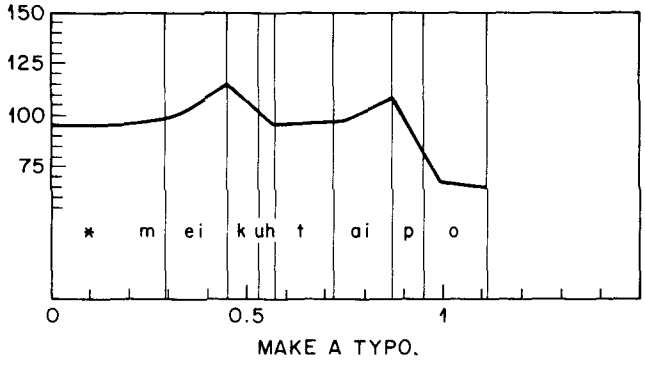



**Figure 14.** Declarative Contour

With the declarative intonation characteristic of imperatives, 5 would probably convey that the hearer was being ordered to produce a typo. Roughly speaking, the tune appears to convey information about speaker attitudes and intentions (as, the speech act the speaker intends to perform) and about the relationship between utterances in a discourse.

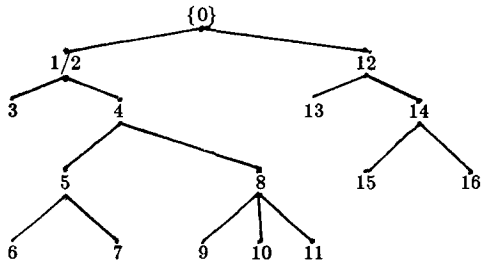
## 4. Intonational and Discourse Phenomena

The major questions underlying our research are: First, what is the relationship between particular intonational phenomena and particular discourse phenomena? For example, what discourse phenomena are associated with changes in pitch range? With the accenting or deaccenting of particular lexical items? With choice of tune? More generally, we also characterize the contributions of these intonational phenomena in terms of the theory of discourse structure developed in [15], by relating intonational contributions to aspects of intentional and attentional discourse structure. Second, how do intonational features such as these interact with one another? Does an expansion of pitch range affect the interpretation of a **rise-fall-rise** contour [25], for example, and if so how? Third, when several discourse features predict conflicting intonational strategies, how is a decision made? When the information represented by a single referring expression, for example, is both 'given' and 'contrastive' -- and thus both deaccentable and accentable -- how is the choice to be made?

### 4.1 Pitch Range Manipulation

Students of discourse commonly observe that discourses often exhibit a hierarchical structure -- into major topics, their subtopics, sub-subtopics, and so on. In task-oriented domains, it has been claimed that this structure reflects the hierarchical structure of a task and its subtasks [18]. So, for example, the TNT introduction above might be segmented as follows (where utterances are labeled by line number):[5]

**Table 1.** Segmenting the TNT Introduction




This bracketing schema defines a discourse segment as any node together with all the nodes it dominates; for example, lines 1-11 form a segment, as do lines 14-16, and so on. An alternative depiction of the hierarchy above would be [{0}[1/2 3 [4 [5 6 7] [8 9 10 11]]] [12 [13] [14 15 16]]].[6] Evidence for such hierarchical segmentation in general is found in instances of pronominal reference to referents linearly distant in the discourse; in such cases, a notion of hierarchical proximity appears plausible.

Previous research [12, 14] has observed that 'topic jump' can be signalled by raised pitch, as well as increased amplitude and markers of self-editing, hesitation, and discontinuity -- and that pauses and changes in rate characterize segment boundaries. In our work with the TNT script, we found that a hierarchical segmentation of discourse can be marked by systematic variation in pitch range, which can signal movement betweeen levels in the segment hierarchy. In addition, by varying the amount of final raising or lowering at the end of phrases, we can indicate the degree of conceptual continuity between one phrase and the next. We have developed algorithms for assigning pitch range and final raising/lowering in terms of the discourse segmentation.

5. We do not claim this is the only possible segmentation, only that it is a plausible one to convey.

6. Note that 1 and 2 are treated as a unit here, although they are synthesized as separate phrases, since it seemed semantically correct.

To illustrate the algorithms, we relate the TNT introduction presented in Figure 1 to its segmentation in Table 1. When the introduction is synthesized using the TTS default pitch range of 75-115 Hz, the topline for each utterance will remain around 115 Hz. However, the hierarchical relationship schematized above among the various segments may be signalled more clearly if the pitch range is varied. In our version of the script, each segment boundary is marked by a variation in pitch range which correlates with the segment's position in the overall discourse. So, major boundaries are denoted by the largest increases, with smaller increases marking subsegment boundaries, and so on. The segment beginning at 1, for example, is marked by raising the f0 topline to 150 Hz; that beginning at 14, by raising the topline to 136 Hz; and that beginning at 15, by raising the topline to 125 Hz.[7] Human speakers do seem to employ a wider spectrum of pitch range variation than we have been able to use in synthesis, however.

We would claim that the appropriateness of changes in pitch range is a function of the segmentation hierarchy -- and is not inherent in the utterance in isolation. Our algorithm for pitch range assignment can in fact enforce one segmentation of a given discourse over another and, in so doing, can disambiguate among potentially ambiguous reference resolutions. For example, *It* in line 7 of Figure 1 coindexes *mistake*, while *it* in line 8 coindexes *word processing*. A simple linear approach to reference resolution (such as [26] ) would have the second coindexical with the previous noun-phrase (np), *mistake*, but a hierarchical approach to discourse structure holds out the possibility that a referent in a segment dominating the current segment may also provide a referent [18], as, in fact, is the case here. While a little thought will make the appropriate referent clear, it is clearer when line 8 is produced with a larger pitch range to signal the beginning of a new subsegment of the segment headed by 4. By so doing, we lessen the possibility that a referent for this *it* will be sought in lines 5-7. The most likely candidate, found in 4, is now both intonationally and conceptually 8's superordinate discourse segment.

While an increase in the pitch range indicates segment boundaries, a decrease in the final lowering effects can indicate the absence of such boundaries, and thus indicate that a given utterance and one which follows it are part of the same segment. So, manipulation of final lowering can also serve to indicate discourse structure, by identifying the internal structure of segments. For example, at one point in the TNT script, the following utterance constitutes an entire discourse segment, so it has the default final lowering (F=0.87); in consequence, the L% tone at the end of *had* will be only 87% as high as it would have been if final lowering had not applied.

```
F.87      Type had.
           H*   H* L L%
```

Compare this with:

```
          Type had.
F.93       H*   H* L L%
          When you're done,      hit changer.
                       H*  L H%        H*   L L%
```

Here, the same utterance is synthesized with less final lowering -- the L% tone at the end of *had*, in particular, will attain 93% of its target height. In this segment, the first line does **not** end the segment. We further propose that the degree of final lowering may correlate with the utterance's position in the discourse hierarchy. Specifically, we suggest that minimal final lowering may indicate a 'push' onto the segment stack and greater degrees of final lowering

7. Our choice of ranges was determined in part by the TTS synthesizer, which tends to sound best when its topline ranges between 115-150 Hz. Preliminary investigation of pitch range changes in human speech indicates that, for male speakers, these choices are reasonable. Note also that it is the relationship among different range levels, not the actual values in Hz, which is important here.

may be associated with 'pops' of this stack. In our synthesis of the TNT text, we have varied degree of final lowering for such 'pops' based upon the level of the segment which this utterance 'completes' (or, equivalently, the level of the segment the next utterance begins). So, to determine the amount of final lowering to assign when synthesizing line 7 in the TNT introduction, we first determine whether it completes a segment (representing a pop) or not (representing a push). If the former, we may note either that it completes the segment begun at line 5 (with a topline of 136 Hz), or that the subsequent segment is begun (by line 8) with a topline of 136 Hz. We assign final lowering of 0.90 when synthesizing line 7 based on either observation; this rather large amount (close to the synthesizer's default maximum of 0.87) conveys a relatively important change of subtopic within the larger discourse segment by indicating rather more disjunction than we would want, for example, between lines 9 and 10.

We are currently testing the associations between pitch range/ final lowering variation and discourse structure proposed above in several ways: by pitch-tracking a large corpus of natural speech,[8] by recording and analyzing subjects reading structured texts, and by asking subjects to perform tasks such as reference resolution from texts synthesized with varying pitch ranges.

### 4.2 Accent Placement

Accent placement, too, can convey information about the structure of a discourse. Traditionally, it has been noted that stress, or accent, can convey information about the focus of an utterance, about given or new information in the discourse,[9] about parallelism, or about contrastiveness. In more general terms, one might say that accent placement appears to be associated with Grosz and Sidner's [15] **attentional** structure -- the salience of discourse entities, properties, relations, and intentions at any point in the discourse. We have particularly noted that the decision to accent or deaccent some item is sensitive to the position of that item in the discourse structure -- that is, just as salience is always determined relative to some particular context, accent placement must be determined with respect to the segment in which the accentable item appears. We take the position that it is the signaling of salience relative to the discourse segment that produces the secondary effects of given-new distinction, topic-hood or contrastiveness, and the favoring of one reference resolution over another.

One of the more common observations about the role of accent placement and the structuring of discourse is that accent can mark some item in the discourse as in **focus** -- i.e., as 'what is being talked about' [28, 29] -- particularly when syntactic or thematic information might predict otherwise. For example, in the following instructions, *erase* is accented in line 2 to indicate that the action of 'erasing' is the focus of the current task.

```
1.      Type hello.
         H*      H* L L%
2.      Next,     let's erase hello.
         H* L H%   H*     H*      - L L%
3.      Hit hint.
         H*  H*  L L%
```

For similar reasons, we accent *hello* in line 1 and deaccent it in line 2.

While focus considerations clearly influence accent placement, determining accent placement solely on the basis of utterance-level focus (as proposed in Gussenhoven [29] and Culicover and Rochemont [30] ) is insufficient. Considerations such as the given/new distinction play an important role.

Speakers typically deaccent given information and accent new information, as when the 'new' information *typing* is accented and the 'old' *word processing* is not in line 3 below:

```
1.      Welcome to word processing.
         H*         H*              L L%
2.      That's using a computer to write letters
          H*  H*          H*       H*   H*
              and reports.
                    H*   L L%
3.      Word processing makes typing easy.
                      H*    H*    H*  L L%
```

Note that these items are marked as 'given' and 'new' within the current segment -- although they may have other status within the larger discourse. Furthermore, items appear 'given' or 'new' not simply because of prior mention (or lack thereof) in a context but via 'physical co-presence', where speaker, hearer, and referents are physically and openly present together; [31] shared world knowledge; or conceptual proximity [1]. For example, the tutor can treat *m* as given in the following text because the student has just (incorrectly) typed *mary*; the character 'm', the student, and the tutor, are thus physically copresent.

```
Oops,    capital m.
 H* L H%  H*     - L L%
```

The new information is that 'm' is to be capitalized. Thus *capital* is accented. Similarly, in the introduction to the tutor presented in Figure 1, we can deaccent *mistake* because it is a super-concept of the previously mentioned *typo*:

```
Make a typo?
 L*     H*  H H%

No problem.
 H*  H*   L L%

Just back up,       type over the mistake,
      H*  H* L H%   H*  H*              L H%
        and it's gone.
                  H*  L L%
```

We also examine how pronominalization interacts with accent placement. Since the ability to pronominalize is itself a standard test of givenness, **prowords**, like other given items, are commonly deaccented. If they are accented, the hearer may draw very different conclusions from an utterance. The following utterance, for example, may well convey an instruction to type the word *something* or even a reprimand for not typing anything yet:

```
Let's begin by typing something.
          H*      H*     H*
```

Since the TNT script employes little pronominalization, we often use deaccenting to 'intonationally pronominalize' repetitions of lexical items.

Accent can also signal that a discourse referent other than that which would be 'most likely' without special accentuation should be sought, as in:

```
1.      We can't answer questions,      if you are confused.
        H*  H*           H*      L H%                H*   L L%
2.      We have to let the computer do all the teaching.
                  H*         H*      H*        H*    L L%
```

Here (and in particular at line 1), *we* is intended to refer to the humans supervising the testing of the tutor, although these humans have **not** previously been mentioned in the script. However, this reference might easily be interpreted as referring to the

8. From interviews collected by A. Kroch and G. Ward and from recordings made of a radio financial advise program by J. Hirschberg and M. Pollack.

9. Prince [27] notes that the 'given/new' distinction has been variously defined as predictable/unpredictable, salient/not salient, shared/not shared knowledge, and proposes a more complex taxonomy of 'assumed familiarity' classifying discourse entities as new, inferrable, or evoked (either textually or situationally). This is closely related -- and often confused with -- the notion of utterance topic/focus.

tutorial system itself. Since pronouns -- as 'given' information -- are commonly deaccented, we accent this one to indicate that an 'unusual' referent should be sought.[10] So, both accent placement and manipulation of pitch range can be used to reorder the list of potential referents for a given referring expression.

Finally, contrastiveness or parallelism may also be communicated via accent. For example, *second* is accented in 3, although it is certainly given in this segment (via mention of *second draft* in 1):

```
1.  Need a second draft?
     L*     L*      L* H H%
2.  No problem.
    H*   H*   L L%
3.  Just change the first,  and you've got the second.
          H*        H*  L H%            H*      H*  L L%
```

Note that, while *second* may be 'given' at the discourse segment level, the decision to accent it is based on contrast within a smaller context, 3. Furthermore, if this function of accent is ignored, contrastiveness may be inferred incorrectly. If we accent *we* in the last line of the tutorial introduction *WE will help you out*, for example, the student would be entitled to infer that others will not be helpful.

We are currently developing algorithms for determining accent placement, based upon the interaction of focus, given/new, parallelism, contrastiveness, and pronominal reference within segment and phrase.

### 4.3 Choice of Tune

It is now widely accepted that the overall melody a speaker employs in an utterance can communicate some semantic or pragmatic information. However, since there are few particular tune types for which we can specify with any confidence just what the meaning might be, it is difficult to generalize about what type of information tunes in general can convey. From those tunes whose 'meaning' seems fairly well understood -- namely, **declarative**, **yes-no question** [23], **surprise/redundancy** [10], **contradiction contour** [33] **rise-fall-rise** [25], and **continuation rise** [34,35] contours -- we propose that tunes convey two sorts of information about discourse.

First, we believe that contours can convey **propositional attitudes**[11] the speaker wishes to associate with the propositional content of an utterance. For example, the speaker may wish to convey that s/he knows **x**, or that s/he *believes* **x**, or that s/he is *uncertain* about **x**, or that s/he is *ignorant of* **x**. In the case of **H*+L** sequences, it appears that a speaker may convey his/her (propositional) attitudes about a hearer's (propositional) attitudes toward an utterances. This tune seems to indicate the speaker's belief that the speech act s/he is performing is superfluous. For example, a speaker may employ it to convey that the propositional content of his/her utterance is already known or would be obvious to the hearer (who, of course, may or may not be attending to it). Note that the speaker may or may not believe that this information is known, in order to wish to convey this meaning. Particularly in pedagogical texts, this contour seems appropriate to introduce straightforward material, as in the following instruction to hit the *remind* key.

```
Remind,    tells you again what to do if you forget.
   H*  L H%  H*          H*          H*           H* L L%

Hit   remind.
H*+L    H*+L L L%
```

However, an **H*+L** sequence is not appropriate in the following similar exchange:

```
Next,    let's erase hello.
 H* L H%  H*     H*       L L%

Hit hint.
 H*  H* L L%
```

In general, such contours do **not** seem felicitous when the utterance conveys information which the speaker believes will be unexpected for the hearer. Here tune choice may reflect attentional as well as intentional aspects of the discourse structure. Like the deaccenting of references to given items, **H*+L** sequence contours seem to convey 'givenness' at a more general level.

Second, we believe that tune can convey the speaker's commitment to some semantico-pragmatic structural relationship holding between the propositional content of utterances (as, that one 'completes' another or is subordinate to another). Many such relations have been proposed in textual analysis [36,37,15]. In the phonological literature, continuation rise has been commonly associated with some sense of 'continuation' or 'more to come' [34]. We have found, however, that this contour can be characterized more precisely as conveying a subordination relationship between the phrase uttered with continuation rise and other utterances in the discourse segment. For example, if the second phrase of line 1 is uttered with continuation rise, then this utterance appears to be subordinated to 2.

```
1.  We can't answer questions,  if you are confused.
     H*  H*           H*    L H%              H*   L H%
2.  We have to let the computer do all the teaching.
              H*        H*      H*        H*    L L%
3.  But if the computer is not working right,   we will help
                  H*          H*          L H%  H*       H*
        you out.
            H* L L%
```

That is, 2 'completes' 1. Without continuation rise on 1, all three utterances will appear to have equal status in the segment. Furthermore, continuation rise is **not** felicitous in all contexts in which the simple sense that 'there is more to come' clearly should be appropriate; for example, continuation rise over 3 -- at the end of the tutorial introduction -- seems quite odd, even though more will clearly follow.

In synthesizing the TNT script, we have employed only a small subset of possible English tunes. Analysis of the 'meaning' of additional tunes is part of our future research. More generally, we must examine how structural relationships conveyed by tunes such as **H*+L** sequence are associated with those conveyed by pitch range.

We have described certain mappings between intonational features and discourse phenomena, associating pitch range variation with the identification of discourse segments and with their internal coherence; accent with types of information status such as topic (focus) and the given/new distinction, with reference resolution and with contrastiveness; and tune choice with the relationships among propositions in the discourse as well as with some propositional attitude the speaker wishes to associate with those propositions. It appears that pitch range and accent placement are most closely associated with a discourse's attentional structure, while tune choice is more closely associated with its intentional structure. However, clearly this picture is too simple. Several intonational features may be used together to create some discourse effect; moreover, in some cases two distinct intonational phenomena seem to produce discourse effects that seem intuitively to be closely related. And sometimes several discourse phenomena may indicate conflicting intonational strategies. These problems are the subject of our future research.

## 5. Discussion

The central thesis of this work is that there are many ways in which intonation helps to structure discourse. By understanding the mapping between intonational phenomena and discourse phenomena, we can enhance both our ability to interpret what

10. The standard example of accentuation influencing pronominal reference resolution in this way is'john hit Bill and then HE hit HIM [32].

11. Propositional attitudes include *knowing, believing, intending, uncertainty,* and *ignorance.*

speakers try to convey and to synthesize speech more effectively. We have described three major intonational phenomena -- pitch range, accent, and tune -- and some of the information they allow speakers to communicate about discourse, demonstrating some links between discourse and intonational phenomena which have not been noted in the literature and refining some notions which have. We also identify major issues which future research on the relationship between discourse and intonation must address, including a more precise mapping between discourse and intonational phenomena, the interaction of intonational phenomena to produce particular discourse effects, and the way conflict between intonational strategies signaled by various aspects of the discourse may be resolved.

We are currently testing and refining our hypotheses by 1) pitch tracking recorded natural discourse to determine pitch range manipulation, and 2) conducting pilot empirical studies of how principled manipulation of pitch range can affect reference resolution. We are also examining in some detail the relationship between pronominalization and deaccenting, pursuant to the development of better accenting algorithms for synthetic speech. Our ultimate goals are practical as well as theoretical. Once we have determined how particular intonational phenomena are related to particular discourse phenomena, the next step is to determine how these findings can be applied to natural-language generation. In particular, how much intonational structuring of generated text can be done automatically? What sorts of information must be represented to support the assignment of rhetorically effective intonation?

ACKNOWLEDGEMENTS

We would like to thank Lloyd Nakatani and Dennis Egan for help with TNT, Barbara Grosz and Candy Sidner for useful discussions, Mary Beckman, Diane Litman, and Ken Church for comments on earlier drafts, and Mark Liberman for assistance with the TTS system and the development of its prosody.

## REFERENCES

[1] Chafe, W., Givenness, contrastiveness, definiteness, subjects, topics, and point of view, in *Subject and topic*, ed. Li, C., Academic Press, New York (1976).

[2] Schmerling, S., Presupposition and the notion of normal stress, *Papers from the Seventh Regional Meeting of the Chicago Linguistic Society*, Chicago, (1971).

[3] Schmerling, S., A re-examination of the notion NORMAL STRESS, *Language* **50** pp. 66-73 (1974).

[4] Wilson, D., and Sperber, D., Ordered entailments: an alternative to presuppositional theories, pp. 229-324 in *Syntax and semantics 11*, ed. Oh, C.-K., and Dinneen, D. A., Academic Press, New York (1979).

[5] Gleitman, L., Pronominals and stress in English, *Language Learning* **11** pp. 157-169 (1961).

[6] Gundel, J., Stress, pronominalization, and the given-new distinction, *University of Hawaii Working Papers in Linguistics* **10(2)** pp. 1-13 (1978).

[7] Jackendoff, R. S., *Semantic interpretation in generative grammar*, MIT Press, Cambridge MA (1972).

[8] Ladd, D. R., *The structure of intonational meaning*, Indiana University Press, Bloomington (1980).

[9] Austin, J. L., *How to do things with words*, Clarendon Press, Oxford (1962).

[10] Sag, I. A. and Liberman, M., The intonational disambiguation of indirect speech acts, *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*, pp. 487-498 Chicago, (1975).

[11] Sadock, J., *Toward a linguistic theory of speech acts*, Academic, New York (1974).

[12] Schlegoff, E. A., The relevance of repair to syntax-for-conversation, pp. 261-288 in *Syntax and semantics 12: Discourse and syntax*, ed. Givon, T., Academic, New York (1979).

[13] Brazil, D., Coulthard, M., and Johns, C., *Discourse intonation and language teaching*, Longman, London (1980).

[14] Butterworth, B., Hesitation and semantic planning in speech, *Journal of Psycholinguistic Research* **4** pp. 75-87 (1975).

[15] Grosz, B. J., and Sidner, C. L., The Structures of discourse structure, 6097, BBN Laboratories Inc. (November 1985). Also appears as CSLI-85-39, as Technical Note #369 from the AI Center, SRI International, and will appear in Computational Linguistics, 1986.

[16] Nakatani, L., Egan, D., Ruedisueli, L., and Hawley, P., *TNT: A talking tutor 'n' trainer for teaching the use of interactive computer systems*, To be presented the Conference on Human Factors in Computing Systems, April 13-17, 1986 (1986).

[17] Olive, J. P., and Liberman, M. Y., Text to speech -- An overview, *J. Acoust. Soc. Am. Suppl. 1* **78**(Fall) p. s6 (1985).

[18] Levy, E. T. and Grosz, B., *Communicating thematic structure in narrative discourse: the use of referring terms and gestures*, PhD thesis, University of Chicago (1984).

[19] Reichman, Rachel, *Getting computers to talk like you and me*, MIT Press, Cambridge MA (1985).

[20] Cohen, R., *A computational model for the analysis of arguments*, PhD thesis, University of Toronto (1983).

[21] Sacks, H., Schlegoff, E., and Jefferson, G., A simple systematics for the organization of turn-taking for conversation, *Language* **50** pp. 696-735 (1974).

[22] Anderson, Mark D., Pierrehumbert, Janet B., and Liberman, Mark Y., Synthesis by rule of English intonation patterns, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2.8.1-2.8.4 San Diego, (1984). Vol. 1

[23] Pierrehumbert, J., *The Phonology and phonetics of English intonation*, PhD thesis, MIT (1980).

[24] Liberman, M., and Pierrehumbert, J., Intonational invariants under changes in pitch range and length, in *Language sound structure*, ed. Aronoff, M., and Oehrle, R., MIT Press, Cambridge (1984).

[25] Ward, G., and Hirschberg, J., Implicating Uncertainty: The Pragmatics of Fall-Rise Intonation, *Language* **61(4)** pp. 747-776 (1985).

[26] Winograd, T., *Understanding natural language*, Academic Press, New York (1972).

[27] Prince, E. F., Towards a taxonomy of given-new information, pp. 223-256 in *Radical pragmatics*, ed. Cole, P., Academic, New York (1981).

[28] Sidner, C. L., *Towards a computational theory of definite anaphora comprehension in English discourse*, PhD thesis, MIT (1979). Also appears as TR 537, MIT AI Lab.

[29] Gussenhoven, C., *On the grammar and semantics of sentence accents*, Foris, Dordrecht, Neth. (1983). Publications in Language Sciences, 16

[30] Culicover, Peter W., and Rochemont, Michael, Stress and focus in English, *Language* **59**(1) pp. 123-165 (1983).

[31] Clark, H. H., and Marshall, C. R., Definite reference and mutual knowledge, in *Elements of discourse understanding*, ed. Joshi, A., Webber, B., and Sag, I., Cambridge University Press, Cambridge (1981).

[32] Lakoff, G., Presupposition and relative well-formedness, pp. 329-340 in *Semantics*, ed. Steinberg, D., and Jakobovits, L., Cambridge University Press, Cambridge (1971).

[33] Liberman, M., and Sag, I., Prosodic form and discourse function, *Papers from the Tenth Regional Meeting of the Chicago Linguistic Society*, pp. 416-427 Chicago, (1974).

[34] Bolinger, D., Intonation and its parts, *Language* **58**(3) pp. 505-533 (1982).

[35] Bing, J., *Aspects of English prosody*, PhD thesis, University of Massachusetts at Amherst (1979). Reprinted by the Indiana University Linguistics Club, 1980

[36] Mann, W. C., Moore, M. A., Levin, J. A., and Carlisle, J. H., Observation methods for human dialogue, RR/75/33, ISI (1975).

[37] McKeown, K., *Generating natural language text in response to questions about database structure*, PhD thesis, University of Pennsylvania (1982).