INTONATION AND THE INTENTIONAL STRUCTURE OF DISCOURSE

J. Hirschberg D. Litman J. Pierrehumbert

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

G. Ward

Department of Linguistics
Northwestern University
Evanston IL 60201

ABSTRACT

It is increasingly recognized in Natural Language Processing that intonation makes a significant contribution to the communication of discourse structure. However, the correspondence between particular intonational features and specific aspects of discourse structure is only beginning to be understood. In this paper, we show how tune, phrasing, accent, and pitch range can combine to convey information about the nature of speaker intentions and about the relationship among those intentions. Our findings reveal new sources of linguistic information for research in plan inference and discourse understanding, and allow us to make more sophisticated use of intonational variation in synthetic speech.

1. Introduction

It is increasingly recognized in Natural Language Processing that intonation makes a significant contribution to the communication of discourse structure. However, the correspondence between particular intonational features and specific aspects of discourse structure is only beginning to be understood. In [3] we proposed a tentative mapping between what Grosz and Sidner [1] term the attentional and intentional structures of discourse, and intonational features such as pitch range, accent, phrasing, and tune. In this paper, we extend this work, focussing on the role of intonation in communicating intentional structure. We show how intonational features can combine to convey information about the nature of speaker intentions and about the relationship among those intentions. In particular, we describe how pitch range is used to communicate discourse structure; how phrasing and accent indicate cue phrase interpretation for words like *now, moreover, and finally,* which can convey relationships between intentions; and how tune conveys information about speaker intentions and the relationship between those intentions. Our findings reveal new sources of linguistic information for research in plan inference and discourse understanding, and allow us to make more sophisticated use of intonational variation in synthetic speech. As a working framework, we adopt Grosz and Sidner's [1] model of discourse structure and Pierrehumbert's system of intonational description [6].

1.1 Intentional Structure

Orosz and Sidner [1] propose a tripartite view of discourse structure: a linguistic structure, which is the text/speech itself; an attentional structure, which includes information about the relative salience of objects, properties, relations, and intentions at any point in the discourse; and an intentional structure, which relates discourse segment purposes (DSPs) - whose recognition is essential to a segment achieving its intended effect — to one another. Each DSP contributes to the overall discourse purpose (DP) of the discourse. DPs and DSPs are intentions whose satisfaction represents the main purpose of a discourse or segment, e.g. "Intend that an agent believe some fact" or "Intend that an agent believe that one fact supports another." While all DSPs by definition must contribute to the DP, DSPs may also be related to one another in one of two ways: First, DSP1 is said to contribute to DSP2 when DSP1 provides part of the satisfaction of DSP2; in this case DSP2 is said to dominate DSP1. Second, DSP1 is said to satisfaction-precede DSP2 whenever DSP1 must be satisfied temporally before DSP2. These relations thus impose two partial orderings on DSPs in a discourse: a dominance hierarchy and a satisfaction-precedence ordering.

1.2 Dimensions of Intonational Variation

In Pierrehumbert's [6] system of intonational description, intonational contours are described as sequences of low (L) and high (H) tones in the F0 (fundamental frequency) contour. A well-formed intermediate phrase consists of one or more pitch accents, which are aligned with stressed syllables (syllable alignment is indicated by *) on the basis of the metrical pattern of the text, plus a simple H or L which characterizes the phrase accent. The phrase accent spreads over the material between the last pitch accent of the current intermediate phrase and the beginning of the next - or the end of the utterance. Intonational phrases are composed of one of more such intermediate phrases plus a boundary tone, which may also be H or L and is indicated by '%'. It falls exactly at the phrase boundary.

A phrase's tune, or melody, is defined by its particular sequence of pitch accent(s), phrase accent(s), and boundary tone. For example, an ordinary declarative pattern with a final fall is represented as H* L *L%* - that is, a tunc with H* pitch accent(s), a L phrase accent, and a *L%* boundary tone. An interrogative contour is represented as L* H H% - L* pitch accent(s), H phrase accent and H% boundary tone. The contrast between these two melodies is illustrated in Figures 1 (declarative) and 2 (interrogative), for the sentence *Bill doesn't drink because he's unhappy.**

Intermediate and intonational phrases can be identified by pausing and phrase-final syllable lengthening as well as the extra melodic elements of phrase accent and boundary tone present at the end. Variation in phrasing is illustrated by comparing Figure 1 (a single phrase) with Figure 3 (two phrases).

* These and subsequent examples were synthesized using the Bell Labs Text to Speech System [5].

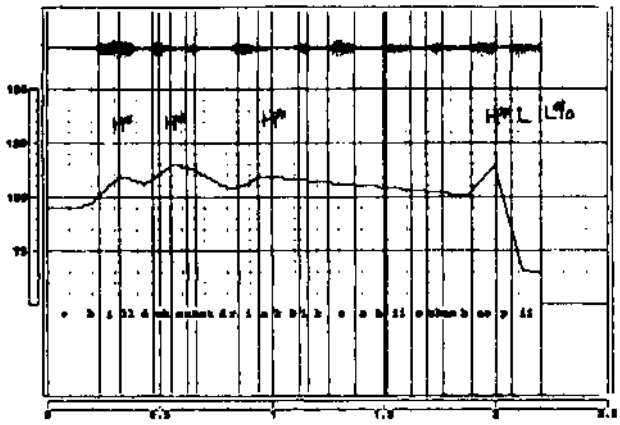

Figure 1. Declarative Contour

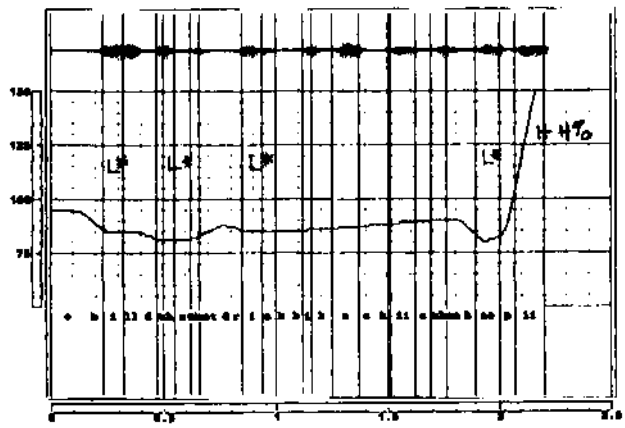

Figure 2. Interrogative Contour

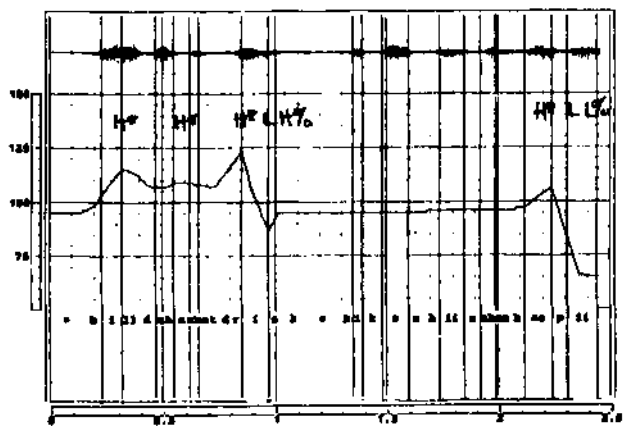

Figure 3. Two Phrases

Pitch accents, which fall on the stressed syllable of lexical items, mark those items as intonationally prominent. There are six types of pitch accent in English [6], two simple tones ~ high and low - and four complex ones. The high tone, the most frequently used accent, comes out as a peak on the accented syllable (as, on *Bill* in Figure 1) and is represented as H*. The 'H' indicates a high tone, and the '*' that the tone is aligned with a stressed syllable. L* accents occur much lower in the pitch range than H* and are phonetically realized as local fO minima. The other English accents are composed of two tones. For example, figure 4 shows a version of the sentence in Figure 1 with a L*+H accent substituted for the H* accent on *Bill.*

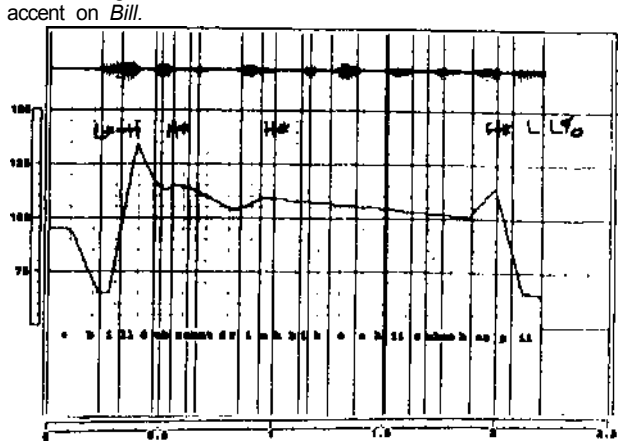

Figure 4. A L*+H Accent

Note that the peak now occurs just after the stressed syllable.

When a speaker's voice is raised, the overall pitch range - the distance between the highest point in the fO 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 affected proportionally. In both cases, the shape of the actual contour is the same, but its scaling is different.

In addition to variations in overall pitch range, the intonation system exploits a local time-dependent type of pitch range variation, called final lowering. 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.

2. Pitch Range and Topic Structure

The topic structure of a discourse includes the initial segmentation prerequisite to the identification of DSPs as well as the relationships that hold among DSPs. In [3], we proposed that speakers can signal this structure by manipulating pitch range and final lowering ratios. When speakers increase their pitch range from one utterance to the next, they can signal varying degrees of topic change. Degree of final lowering in an utterance can be used to signal the 'level' of topic which that utterance concludes; maximum final lowering signals the conclusion of major topics, for example. In both cases, of course, it is the relationship of pitch ranges and final lowering ratios employed, rather than any absolute values, that is at issue.

While we developed these hypotheses in the course of synthesizing prepared text,* similar hypotheses were developed independently by Silverman [10], who verified them empirically. Silverman tested subjects on potentially ambiguous texts such as the following, synthesizing them to reflect alternate paragraph structures by pitch range and final lowering manipulation:

This building company offers several different schemes for double- glazing. The cheapest is acrylic sheeting. You pay by the square metre, plus the mounting clips. Installation is extra. The most expensive systems are the "slimline" and "royal" schemes. Prices include sealed glass units, and draught-proof frames. *All materials are delivered free within Cambridge.*

Depending upon the structuring of the paragraph, *All materials...* may apply only to "the most expensive systems" - or it may begin a new paragraph, and so apply to both cheap and expensive systems. For this and five other texts and for 20 listeners, Silverman found that subjects' judgments about paragraph structure (elicited by questions such as "For which schemes are all materials delivered free within Cambridge?") followed the prosodic structuring 70.4% of the time.*'

We plan to record subjects reading similar texts to test whether they manipulate international features to disambiguate potentially ambiguous anaphors. From pitch tracks of these recordings, we hope to determine whether speakers communicate different anaphora resolutions for a given text by manipulating pitch ranges over the text or other intonational features such as accenting, speech rate, and pausal duration.

3. Phrasing, Accent, and Cue Phrases

Cue phrases [8] are linguistic expressions - such as *and, first, first of all, for example, therefore* ~ that may be used to explicitly convey information about the attentional or intentional structure of a discourse. For example, *therefore* can indicate that a new DSP is dominated by a previous DSP, while *first* can indicate relationships of dominance as well as of satisfaction-precedence. In general, cue phrases also have 'non-cue' interpretations. The problem of how to distinguish between the two has been little addressed in the literature. Consider the cue phrase *first.* In Example 1, *first* is used to convey information about the intentional structure:***

1.

Tony: I have a couple of questions uh *first* I
Harry: Fire away

In particular, *first* indicates the start of a sequence of DSP's dominated by the DSP representing discussion of Tony's questions.

In contrast, consider the use of *first* in Example 2:

2.

Harry: Well as far as the 10 in the savings account goes, take 9 out of there and put it in a money market fund. As far as the CDs are concerned, the *first* one comes due when-give me a date.

Here, *first* is used as a modifier and does *not* provide explicit information about the intentional structure of the discourse. In other words, while cue phrases *may* be used to communicate discourse structure, they may also be used to different effect.

While the examples above might be syntactically distinguishable, other cases are not, as in 3.

3.

Ron: Right well 11 read various books on taxes myself and I was under the impression that uh when you get investigated that they have to tell you *first* uh I don't know if ...

Here, *first* might begin a new DSP *(First uh, I don't know...).* Yet, from the recorded speech it seems clear that, instead, *first* ends the previous clause *...have to tell you first.**

Our findings from a pilot study of the cue phrase *now* in recorded natural speech [2], show that cue and non-cue usages of linguistic expressions can be distinguished internationally. In the one hundred instances of *now* we examined, cue and non-cue uses patterned distinctly in terms of accent and composition of intonational phrase. Non-cue uses were always part of larger intermediate or intonational phrases, and were usually accented - always with a H* or complex accent. In contrast, cue uses were either 1) part of larger intonational phrases and deaccented or accented with L*; or 2) separate intermediate or intonational phrases accented with L* or H*. Also, all cue uses either appeared in the initial position of intonational or intermediate phrases, or in positions preceded only by other cues. The few non-cue *nows* appearing in this position were always distinguished from cue *nows* by their H* accent.

Thus, in our study, non-cue *now* was always distinguishable from cue *now* by a combination of accent type, position in intonational/intermediate phrase, and overall composition of phrase. We are currently examining other cue phrases in this regard, as well as extending the set of intonational features we are analyzing to include relative pitch range of target and surrounding intonational phrases, duration of cue phrase, and type of contour of the target phrase. We will also examine how cue and non-cue use in written text can be related to usage in speech.

4. Tune and Speaker Intention

Tune provides information about two aspects of intentional structure. In [3] we proposed that tunes can convey the relationships among DSPs. Here, we describe work on how tune can help in the identification of DSPs.

A given sentence may be uttered with different tunes to convey different meanings, and a given tune may induce different interpretations in different contexts. In the case of indirect speech acts, for example, a yes-no question contour on "Can you pass the salt?" will be more likely to elicit a direct yes-no response, while a declarative pattern will be more likely to be interpreted as a request to pass the salt [9]. However, a yes-no question contour over "My name is Mark Liberman?" did *not* convey a request for a yes-no response - but, rather,

* The text of TNT, a talking tutor for the Unix screen-oriented text editor vi [4].

• These results were statistically significant. Subjects failure to follow prosodic cues in other cases might be due to semantic (Mas of particular paragraphs or to variable subject sensitivity to such cues.

*Examples 1-3 are taken from a radio call-in program, "Speaking of Your Money," taped the week of 1 February 1982 [7].

Note that example 3 is presented without the transcriber's punctuation. While orthography may help to disambiguate cue from non-cue usage in written text - if only by marking sentence boundaries - in cases such as 3, the sentence boundary itself can only be distinguished intonation ally. Hence it is intonation, rather than syntax or surface position, that provides the necessary disambiguation in speech - and which is reflected orthographically in text.

whether that name was known to the hearer [6]. Clearly, a more sophisticated notion of what tunes convey and how they interact with other aspects of the linguistic context is required.

One tune that we have been investigating is the L*+H L H% contour [11]. This contour can be used to convey both *uncertainty,* i.e. It's *not* the case that the speaker believes a scale or scalar* is appropriate', and *incredulity,* i.e. It *is* the case that the speaker believes a scale or scalar is inappropriate'. The former is illustrated by A's response in 4, and the latter by B's subsequent exclamation:

4.
B: Did you do well on the midterm?
A: I got a B.
B: You got a B!

Here, A conveys uncertainty about whether getting a B constitutes 'doing well' and B responds with incredulity about that uncertainty.

Both incredulity and uncertainty can be subsumed by the abstraction *lack of speaker commitment to the appropriateness of an evoked scale or scalar value.* For any speaker S and any scale or scalar x, we can say that S is uncommitted to the appropriateness of x whenever (a) S believes x is inappropriate, or, (b) S does *not* believe x is either appropriate or inappropriate (i.e., S *doesn't know whether* x is appropriate or not]). Now, we can say that S is incredulous about the appropriateness of x just in case (a) is true. And, we can say that S is uncertain about the appropriateness of x whenever is true. So, lack of speaker commitment -- (a) or (b) -- subsumes both incredulity (a) and uncertainty (b). The incredulity and uncertainty interpretations appear distinguishable in terms of other intonational features -- rate and pitch range in particular; we plan to test the contribution of these factors empirically.

If, then, tune can convey such propositional attitudes, it can thereby convey information about a discourse's intentional structure. Specifically, tune can provide information from which DSPs can be inferred. For L*+H L H%, the knowledge that S believes x inappropriate plus the assumption that, by speaking, S intends to convey this, permits the hearer to infer that S intends to convey that x is inappropriate. So, in 4, A intends B to believe A uncertain about the appropriateness of the value she has supplied on the grading scale given the question 'Did you do well?'; B, in turn, intends A to believe that B believes that the grade does *not* constitute doing well on that scale. What remains is to discover how other tunes determine and reveal the intentional structure of discourse.

5. Discussion and Future Research

In this paper we have shown some of the ways intonation contributes to the intentional structuring of discourse. We have described research on the relationship between pitch range and discourse structure, the way phrasing and accent indicate cue phrase usage, and the information tune conveys about speaker intentions. We are continuing this investigation in the following ways: 1) extending our analysis of the intonational features of cue phrases; 2) investigating the interaction of pitch range manipulation and the use of cue phrases to signal topic structure; 3) conducting empirical investigations of the intonational disambiguation of indirect speech acts 4) conducting perceptual studies of L*+H L H% to determine whether differences in pitch range and rate favor one interpretation over another; 5) developing a compositional theory of tune meaning.

* Scales are defined as partially ordered sets and scalars as elements of those sets.

REFERENCES

1. Grosz, B. and Sidner, C. Attention, intentions, and the structure of discourse. *Computational Linguistics 12,* 3 (1986), 175-204.
2. Hirschberg, J. and Litman, D. Now let's talk about now: identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting,* Association for Computational Linguistics, Stanford, 1987.
3. Hirschberg, J. and Pierrehumbert, J. The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting,* Association for Computational Linguistics, New York, 1986, pp. 136-144.
4. Nakatani, L., Egan, D., Ruedisueli, L., and Hawley, P. TNT: A talking tutor 'n' trainer for teaching the use of interactive computer systems. In , Conference on Human Factors in Computing Systems, April 13-17, 1986.
5. Olive, J.P. and Liberman, M.Y. Text to speech - An overview. *Journal of the Acoustic Society of America, Suppl. 1 78,* Fall (1985), s6.
6. Pierrehumbert, J.B. The phonology and phonetics of English intonation. PhD Thesis, Massachusetts Institute of Technology, 1980.
7. Pollack, M.E., Hirschberg, J., and Webber, B. User Participation in the Reasoning Processes of Expert Systems. MS-CIS-82-9, University of Pennsylvania, 1982. A shorter version appears in the AAAI Proceedings, 1982.
8. Reichman, R. *Getting computers to talk like you and me: discourse context, focus, and semantics.* MIT Press, Cambridge MA, 1985.
9. Sag, I.A. and Liberman, M. The intonational disambiguation of indirect speech acts. In *Papers from the Eleventh Regional Meeting,* Chicago Linguistic Society, Chicago, 1975, pp. 487-498.
10. Silverman, K. Natural prosody for synthetic speech. PhD Thesis, Cambridge University, 1987.
11. Ward, G. and Hirschberg, J. Implicating uncertainty: the pragmatics of fall-rise intonation. *Language 61,* 4 (1985), 747-776.