Intonational Features of Local and Global Discourse Structure

Julia Hirschberg and Barbara Grosz

2D-450, AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974-0636

Division of Applied Sciences
Harvard University
Cambridge MA 02138

1. ABSTRACT

We present results of a study of the relationship between intonational features including pitch range, timing, and amplitude and aspects of discourse structure defined in terms of Grosz and Sidner's (1986) model of discourse. We compare structural labelings of AP news text with prosodic/acoustic features examined from recordings of the same text read by a professional newscaster. We find significant correlations between prosodic/acoustic characteristics and both local and global aspects of discourse structure identified by our labelers. Our results have applications for speech synthesis and, potentially, for speech recognition. (This paper was presented at the DARPA Workshop on Spoken Language Systems, Arden House, February 1992.)

2. INTRODUCTION

The hypothesis that discourse structure is signalled by variation in intonational features such as pitch range, timing, and amplitude has been examined in studies such as [1, 2, 3, 4, 5, 6, 7]. However, as Brown and her colleagues note [2, p. 27]: "... until an independent theory of topic-structure is formulated, much of our argument in this area is in danger of circularity." In this paper we examine the relationship between discourse structure and variation in intonational features using just such an independent model of discourse structure, that proposed by Grosz and Sidner [8] (G&S). We present results of an empirical study comparing intonational features of read text with elements of both the local and global structure of discourse. Our study has immediate application to the generation of appropriate intonational features for synthetic speech, and future applicability to the recognition of discourse structure in speech recognition tasks.

Our corpus consisted of AP news stories recorded by a professional speaker. The intonational features we considered included pitch range, contour, timing, and amplitude. The discourse structural elements we examined at the local level included parentheticals, quotations, tags, and indirect reported speech; at the global level, we studied discourse segmentation — the division of a discourse into constituents that provide the basis for determining discourse meaning. The discourses were labeled by two groups: one group labeled from text; the other group labeled from text while listening to the recorded speech. In this paper, we describe similarities and differences in the segmentations elicited in these two conditions.

Our experiments provide support for three hypotheses. First, instructions can be devised, based on the G&S model, that enable subjects to analyze discourses with considerable similarity. Second, discourse structure is marked intonationally, although the relationship between structure and intonational features is a complex one; a given discourse structural feature may be signaled by several intonational features, either separately or in combination. Third, not every intonational feature which is varied to convey structural information is perceptually salient.

3. SCOPE OF THE STUDY

Although computational theories of discourse make different claims about the basis of discourse structure — e.g. coherence relations [9, 10, 11, 12], syntactic features [13], intentions [8] — all agree that utterances in a discourse group together into segments and that the determination of discourse *meaning* depends crucially on identifying the ways segments fit together. However, discourse segment boundaries do not always align with paragraph boundaries or other orthographic markers in text. And there have been no systematic studies of human labeling of discourse segmentation. As a result, attempts to apply theories of discourse structure have sometimes been frustrated by apparent ambiguities in the structure of a single discourse.

Thus, one goal of our study was to identify similarities and differences among labelers in the segmentation of discourses from text and speech. We wanted to (1) determine whether a set of instructions could be devised that would lead to consistency in segmentation across different labelers and different texts; (2) test the hypothesis that spoken language is less ambiguous than text with respect to discourse segment structure; and (3) identify intonational features that were strongly correlated with discourse structure elements.

We did not, of course, expect all labelings to be identical. Just as a sentence may have multiple parses, a discourse may have several plausible segmentations. The goal of this part of our study was to determine the extent to which segmentations done by different people varied, identify those characteristics of a text that occasioned structural ambiguity, and develop methods for comparing segmentations.

Variation in pitch range has often been seen as conveying 'topic structure' in discourse. Brown et al. [2] found that subjects typically started new topics relatively high in their pitch range and finished topics by compressing their range; they hypothesized that internal structure within a topic was similarly marked. Silverman [3] found that manipulation of pitch range alone, or in conjunction with pausal duration between utterances, enabled subjects to disambiguate reliably potentially ambiguous topic structures. Avesani and Vayra [6] also found variation in range in productions by a professional speaker which appear to correlate with topic structure, and Ayers [7] found that pitch range appears to correlate more closely with hierarchical topic structure in read speech than in spontaneous speech. Duration of pause between utterances or phrases has also been identified as an indicator of topic structure by [2, 1, 6], with longer pauses marking major topic shifts; [4], however, found no such correlation in his data. Amplitude was also found by [2] to increase at the start of a new topic and decrease at the end. And speaking rate has also been investigated [14] as a correlate of structural variation.

Our second goal was to examine the conjecture that speech provides information that enables a listener to identify one of several possible analyses of a discourse as that which a speaker intends to communicate. In their model, G&S propose that discourse be understood in terms of the purposes that underlie it. They argue that three distinct components play a role in discourse structure: the utterances composing the discourse divide into segments forming the LINGUISTIC STRUCTURE; this structure derives from a combination of the INTENTIONAL STRUCTURE, which is a structure of the purposes or intentions underlying the discourse, and the ATTENTIONAL STATE which represents the entities and attributes that are salient during a particular portion of the discourse. Discourses are analyzed as hierarchies of discourse segments. Each segment has an underlying purpose intended by the speaker/writer to be recognized by the listener/reader, the DISCOURSE SEGMENT PURPOSE (DSP). Each DSP contributes to the overall DISCOURSE PURPOSE (DP) of the discourse. For example, a discourse might have as its DP the intention that the listener be informed that there was a plane accident, and individual segments forming that discourse might have as their DSP's intentions that the listener be informed that the plane lost a piece of its tail (an intention contributing information about the accident) and that the passengers were upset (an intention contributing information about the effect of this event). DSP's may in turn be represented as hierarchies of intentions. DSPs **a** and **b** may be related to one another in two ways: **a** DOMINATES **b** if the DSP of **a** is partially fulfilled by the DSP of **b** (equivalently, **b** CONTRIBUTES TO **a**). Segment **a** SATISFACTION-PRECEDES **b** if the DSP of **a** must be achieved in order for the DSP of **b** to be successful. According to this model, part of understanding a discourse is reconstructing the DP, DSPs and relations among them.

We expected differences between the segmentations provided by labelers who labeled solely from text and those who labeled from speech. We also hoped to discover independent, albeit indirect, evidence from intonational variation for the existence of segment boundaries, as well as to provide information about the ways in which intonational features might signal discourse segmentation. In addition to investigating relationships between discourse structure and intonation at the global level, our study examined several local discourse-structural elements.

For spoken language, the determination of discourse structural units at the local level (e.g. identifying parenthetical constituents and quotations) may crucially affect meaning. For example, the sentence '*The government claims the defendants knew that William Parkin a private consultant hired by Teledyne Electronics was paying bribes to Stuart Berlin the Navy official*' may, depending upon how it is uttered, be interpreted to mean that (a) the government claims that the defendants knew X (simple complement); (b) the government claims X, but the defendants knew, X (right-node-raising); or, (c) the defendants knew that the government claims that X (parenthetical) — where X='that William Parkin a private consultant hired by Teledyne Electronics was paying bribes to Stuart Berlin the Navy official'. Because these locally distinct units are often marked orthographically in text, it is presumably easier for readers to agree upon them than on the identification of segment boundaries. Thus, looking for intonational features associated with these local structures minimizes the potential for interlabeler disagreement. As a result, they may provide less equivocal evidence of how speakers use intonational features to convey information about discourse structure.

4. THE EMPIRICAL STUDY

The corpus used in the empirical study consists of three AP news stories, which had been recorded by a professional newscaster from texts available to us. The texts averaged about 450 words in length and the recordings averaged about three and one-half minutes. In this paper, we present our findings for one of these stories (approximately 550 words and four minutes long), as labeled by seven labelers.

4.1. Discourse Segmentation

We developed a set of labeling instructions based on G&S for guiding labelers in segmenting the news stories and identifying various local structural elements. Seven labelers participated in the study. Four (Group T) worked from the text alone. Three others (Group S) labeled from the recording and the text; they were allowed to replay passages as many times as they wished. All of the labelers provided segmentations of one story; three members of Group T and two of Group S also labeled local phenomena for this story. Figure 1 illustrates a sample labeling for this text by one member of Group T. (Note that labelers were allowed to segment according to any division of the text they preferred, although most used the orthographic sentence as their unit of analysis for global structure. The schema presented in Figure 1 identifies only global structure.)

At the global level, we asked labelers to identify segment beginnings and endings and to specify which other segment (if any) the segment was embedded in. In Figure 1, the segments for one labeler are indicated by bracketings of the text; hierarchical relationships among segments are indicated by tabbing. Any unit of analysis (phrase or utterance) in the global segmentations can be described by one of five categories: segment initial sister (**SIS**), segment initial embedded (**SIE**), segment final (**SF**), segment medial immediately following an **SF** utterance — i.e. a POP (**SMP**), or segment medial not following a pop (**SM**). In Figure 1, the first phrases of (a) and (c) illustrate **SIS** utterances; that of (b), (d), and (e) represent **SIE**s; **SF** examples are found at the end of (b), (e), and (f); the first phrase of (f) represents an **SMP**; and all other phrases within the segments (not identified schematically for reasons of space) would represent **SM** units. Differences among utterances in categories **SIE** and **SIS** will not be discussed in this paper; we will refer to them together as segment beginnings (**SBEG**).

Our instructions to labelers for labeling at the global level were cast in terms of the meaning and purpose of the text, because G&S stipulates that intentions are the basic root of discourse segmentation. At the local level, we examined five types of constituents: parentheticals, direct quotations and their tags, indirect reported speech, and speaker attributions for reported speech. We asked both Group T and Group S labelers to mark parentheticals, since these are not always disambiguated orthographically. In addition, we asked Group S to mark direct quotations. Tags and speaker attributions were identified independently by the authors from the text.

4.2. Intonational Features of Discourse Structure

To identify intonational features in the read speech, we labeled the speech for accentuation and phrasing, according to Pierrehumbert's [18] theory of English intonation, using WAVES speech analysis software [19]. We then calculated values for pitch range, as indicated (indirectly) by the fundamental frequency (f0) maximum for the vowel of accented syllables in the phrase;[1] amount of f0 change between phrases, f0(phrase[i])/f0(phrase[i+1]; amplitude, measured within the vowel of the syllable containing the phrase's f0 peak; difference in intensity from prior phrase, measured in decibels (db); contour type; speaking rate, measured in syllables per second (sps); and pausal duration between phrases. We used as our primary unit of analysis Pierrehumbert's phrasal category of intermediate phrase.

Each of these features was then examined as a potential predictor of discourse structure.[2] We compared individual and consensus labelings (i.e. those on which every member of a group agreed) from Group T with those from Group S for direct quotations, tags, indirect reported speech and attributions, parentheticals, and the segment boundaries **SBEG**, **SF**, and **SMP**. Here, we discuss only quotations, parentheticals and segment boundaries.

5. RESULTS AND ANALYSIS

5.1. Discourse Segmentation

We found that discourses can indeed be segmented dependably using our instructions. While no two segmentations were identical, we found no statistically significant difference among six of our seven labelers for labelings of **SBEG** phrases (using Cochran's Q). For **SF** phrases, the seven labelers fell into two groups with no significant difference among members of each group; we hypothesize that each group settled upon a distinct but

[1] Results presented here are based on measurement of f0 maxima for each phrase within the vowel of the syllable containing the phrase's f0 peak. Results from a more conservative measurement at the vowel's amplitude maximum were similar.

[2] In results presented below we have either controlled for phrasal position or performed ANOVAs with both phrasal position and the intonational variable in question as factors, with statistically significant results in each case for the latter.

Figure 1: Sample Segmentation from AP5, Labeler 1

a. [A British Airways Concorde jet with one hundred Americans aboard lost a nine foot piece of its tail today while trying to set a speed record on a world circling journey, but landed safely in Sydney.

b. [William F. Buckley, Jr. and his wife were on board, CBS News reported. The author and commentator had helped organize the trip,which cost each passenger thirty nine thousand dollars.]

c. [A British Airways spokesman said part of the rudder disintegrated while the supersonic jet was flying at forty thousand feet at about fifteen hundred miles an hour nearly twice the speed of sound from Christ church, New Zealand. "It experienced a shudder while over the Tasman Sea that was thought to have been air turbulence," said Stanton.

d. [He said the pilot was unaware of any problem until he was alerted by the control tower at Sydney's Kingsford Smith International Airport.

e. [However, at least one passenger on the one thousand mile flight, which lasted one hour and twenty five minutes, said the plane had shuddered and passengers were tense.]

f. "It was a normal landing, there was no emergency," Stanton said. "The pilot, Capt. David Leney, was told by the control tower that a piece of the tail was missing."]]

...]

plausible interpretation of the text's structure. While we had hypothesized that we might find fewer differences among members of Group S than among Group T labelers, this hypothesis was not in fact borne out. Consistency across Group S labelers was no greater than consistency among all members of the two groups, for labelings of either **SBEG** or **SF**.

Many of the utterances on which labelers disagreed fell into two categories: (1) utterances that might have initiated (or by themselves formed) small separate segments and were thus classified as **SM** by some labelers and **SIE** by others; (2) utterances classified as beginnings by some labelers and **SMP** by others. In the latter case, all of the labelers agreed that there was a discourse break of some kind, but they disagreed about the relationship of the utterance in question to the immediately (linearly) preceding segment; in the following section we provide an analysis of some utterances fell into this class.

5.2. Intonational Correlates of Discourse Structure

Results for our first text are summarized in Table 1. A '+' indicates the row's discourse structural element is characterized by higher values for the column's intonational feature; '-' indicates that the structural element is characterized by relatively low values for the intonational feature. For example, '+' in the 'Pitch Range' column for direct quotations indicates that these phrases are generally higher in range than other phrases.

As shown in Table 1, quoted phrases for Group T were, in general, uttered in a higher pitch range and with less increase in intensity than other phrases; quote-final phrases were produced with a pronounced drop in intensity compared with other sentence-final phrases. Quoted and non-quoted phrases in non-sentence-initial position differed significantly in pitch range (means of 256 Hz vs. 230 Hz; tstat=1.85; df=79; p<.035). Quoted and non-quoted phrases also differed in amount of change from prior phrase in db (1.92 db vs. 5.13 db; tstat=1.71; df=24;p<.05) and between quoted utterance-final and other utterance-final phrases (-5.65 db vs. 1.47 db; tstat=2.87; df=4; p<.025). Comparing these findings with the intonational features of quotations Group S had identified, we found that similar differences in pitch range existed between quoted phrases identified from speech and other phrases, but no significant difference in intensity.

For parentheticals identified by Group T, we also found significant effects for range (195 Hz vs. 258 Hz; tstat=3.67;df=106;p<.001) and for percent change over prior phrase, 81% vs. 107%; tstat=2.29; df=105; p<.02) and intensity (-3.08 db vs. .024 db, tstat=2.04, df=106, p<.025). Our speaker uttered parenthetical phrases in a low pitch range, dropping both pitch and intensity markedly from preceding phrases. Group S's parentheticals were even lower in range (166 Hz vs. 256 Hz; tstat=3.38; df=106; p<.001) than those identified by Group T and exhibited an even more pronounced decrease in pitch (70% vs. 106%; tstat=2.09; df=105; p<.02) and in intensity (-5.10 db vs. 0.13 db; tstat=2.09; df=105; p<0.02). They also were uttered

Table 1: Intonational Correlates of Discourse Features

Discourse Features	Intonational Features Pitch Range	Pitch Range Change	Ampl	Db Change	Prec Pause	Subs Pause	Rate
T:Direct quotes	+			-			
S:Direct quotes	+						
T:Parentheticals	-	-		-			
S:Parentheticals	-	-		-			+
T:SF						+	
S:SF						+	
T:SMP	+				+		
S:SMP	+				+		
T:SBEG	+		+			-	-
S:SBEG					+	-	
T:SBEG+SMP	+					-	
S:SBEG+SMP	+				+	-	

significantly more rapidly than other phrases (6.05 sps vs. 5.06 sps; tstat=1.94; df=106; $p<0.03$).

We take as evidence that these intonational features were used by Group S to identify local structure, the fact that Group S quotations are in general marked more reliably by differences in pitch range than Group T quotations, and that Group S parentheticals are in general marked by larger differences in range and db change than Group T's. We obtained similar results for tags, indirect reported speech, and attributions for such speech.

For global structure, we again found much similarity between intonational features correlated with Group T-identified discourse elements and those correlated with discourse features identified by Group S — with one notable exception which we discuss below. However, for global structures we did **not** find that the intonational features Group S apparently found salient exhibited more pronounced differences over other phrases than Group T-related features.

Intonational features of phrases labeled **SF** and **SMP** by Group T are virtually identical to those for phrases labeled by Group S. For both Group T and Group S **SF**, we find a single intonational correlate, subsequent pause (for T: 1329 msec. vs. 740 msec.; tstat=2.22; df=24; $p<0.02$; for S: 1386 msec. vs. 555 msec.; tstat=3.38; df=17; $p<0.002$). **SF** identified by Group S are followed by only slightly longer average pauses than those identified by Group T — but the ratio of segment-ending pauses to pauses following other sentence-final phrases is greater for Group S. For **SMP**, there is a significant effect for pitch range (340 Hz vs. 296 Hz ; tstat=2.08; df=24; $p<0.025$) and for preceding pause (1329 msec. vs. 603 msec.; tstat=2.66; df=15; $p<0.01$) for Group T. And for Group S we see significant effects for the same factors, pitch range (337 Hz vs. 295 Hz; tstat=2.17; df=24; $p<0.02$) and preceding pause (1386 msec. vs. 698 msec.; tstat=3.04; df=24; $p<0.005$). These findings support similar results in [2, 1, 3].

For **SBEG**, however, while pitch range, amplitude, rate and subsequent pause are significantly correlated with phrases identified by Group T, only preceding and subsequent pause variation distinguishes phrases identified as **SBEG** by Group S. In light of our findings for other global discourse structure elements, we were puzzled at the disparity between our groups with respect to **SBEG**. We were also puzzled that intonational features such as pitch range, which previous production and perception studies had found highly correlated with discourse structure, had no effect on Group S judgments of **SBEG**.

One explanation might be found by examining a superordinate category for **SBEG**. Recall that phrases of the categories **SBEG** (**SIE+SIS**) and **SMP** share the property of **not** being part of the same discourse segment as their preceding phrase; this more general class (**SBEG+SMP**) encompasses shifts to a different segment, some initiating new segments (**SBEG**) and others returning to an embedding one (**SMP**). As we noted above (Section 5.1), labelers often agreed on broader aspects of structure while disagreeing over finer-grained details of the segmentation. In fact, the intonational features characterizing phrases of this more general category for Groups T and S are indeed consistent with the pattern we saw for intonational characteristics of

SF and **SMP**.[3] For **SBEG+SMP** identified by Group T, there are significant effects only for pitch range (336 Hz vs. 294 Hz; tstat=2.41; df=25; $p<0.02$) and subsequent pause (25 msec. vs. 169 msec.; tstat=2.00; df=25; $p<0.03$). These same intonational features are also significantly correlated with **SBEG+SMP** identified by Group S (pitch range: 325 Hz vs. 295 Hz; tstat=1.77; df=25; $p<0.05$; subsequent pause: 30 msec. vs. 183 msec.; tstat=2.34; df=25; $p<0.02$), as is preceding pause (1215 msec. vs. 659 msec.; tstat=2.82; df=24; $p<0.005$). Thus this more general category identifying 'segmentation shifts' yields a comparison of intonational features for Group T and Group S phrases which is consistent with our other findings for global structure, as well as with previous studies of intonation and 'topic shift'.

6. CONCLUSIONS

At both the local and global levels of discourse, we found evidence that structure is associated with intonational variation. Our results provide support for several hypotheses: First, instructions can be developed that enable labelers to produce discourse segmentations with significant similarities. Second, spoken language and written language provide different indicators of discourse segmentation. Third, various intonational features may be employed by a single speaker to convey a given structural element; most elements of discourse structure we examined showed effects for more than one intonational feature. Fourth, although various intonational features may be utilized by a speaker to communicate a single element of discourse structure, only some may be perceptually salient. And, fifth, different configurations of intonational features may be employed to convey the same discourse information in different contexts. For while our aggregate statistics show certain trends, not every token exhibits all these differences.

REFERENCES

1. Lehiste, I. Perception of sentence and paragraph boundaries. In Lindblom, B. and Oehman, S., editors, *Frontiers of Speech Research*, pages 191–201. Academic Press, London, 1979.
2. Brown, G., Currie, K., and Kenworthy, J. *Questions of Intonation.* University Park Press, Baltimore, 1980.
3. Silverman, K. *The Structure and Processing of Fundamental Frequency Contours.* PhD thesis, Cambridge University, Cambridge UK, 1987.
4. Woodbury, A. C. Rhetorical structure in a central alaskan yupik eskimo traditional narrative. In Sherzer, J. and Woodbury, A., editors, *Native American Discourse: Poetics and Rhetoric*, pages 176–239. Cambridge University Press, Cambridge UK, 1987.
5. Woodbury, A. C. Phrasing and intonational tonology in central alaskan yupik eskimo. In Dunne, J., editor, *Mid-America Linguistics Conference Papers*, pages 3–40. University of Oklahoma Press, Norman, 1989.
6. Avesani, C. and Vayra, M. Discorso, segmenti di discorso e un' ipotesi sull' intonazione. In *Att del Convegno Internazionale "Sull'Interpunzione"*, Florence, 1988.
7. Ayers, G. M. Discourse functions of pitch range in spontaneous and read speech. Presented at the Linguistic Society of America Annual Meeting, 1992.
8. Grosz, B. and Sidner, C. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
9. Hobbs, J. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979.
10. Mann, W. and Thompson, S. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.
11. Moore, J. D. and Paris, C. L. Planning text for advisory dialogues. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 203–211. University of British Columbia, June 26-29 1989.
12. Reichman-Adar, R. Extended person-machine interface. *AI Journal*, 22(2):157–218, March 1984.
13. Polanyi, L. A formal model of the structure of discourse. *Journal of Pragmatics*, 12, 1988.
14. Lehiste, I. Phonetic characteristics of discourse. Paper presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan, 1980.
15. Grosz, B. J. The representation and use of focus in dialogue understanding. Technical Report 151, SRI International, Menlo Park Ca., 1977. University of California at Berkeley PhD Thesis.
16. Sidner, C. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse.* PhD thesis, The Massachusetts Institute of Technology, 1979.
17. Grosz, B., Joshi, A., and Weinstein, S. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting*, pages 44–50, Cambridge MA, June 1983. Association for Computational Linguistics.
18. Pierrehumbert, J. B. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.
19. Talkin, D. Looking at speech. *Speech Technology*, 4:74–77, April-May 1989.

[3] Note that **SBEG** significantly outnumber **SMP** phrases when we collapse these categories, so our results do not arise from the latter category dominating the former.