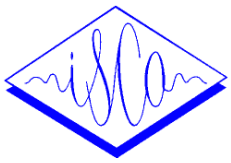

ESCA Workshop on
Spoken Dialogue Systems
Vigsø, Denmark
May 30 - June 2, 1995

Conveying Discourse Structure through Intonation Variation

Julia Hirschberg, Christine H. Nakatani, Barbara J. Grosz

AT&T Bell Laboratories, Harvard University, Harvard University

1. ABSTRACT

This study represents ongoing research into the relationship between intonational variation and discourse structure. Our goals are to examine correlations between prosodic and acoustic variation in spontaneous elicited speech and read speech and the discourse structure which subjects assign to the textual material of this speech. We will employ our findings in improving the naturalness of intonational variation in text-to-speech.

2. Introduction

Previous studies of the acoustic and prosodic correlates of discourse structure have suggested a number of ways in which discourse level meaning can be conveyed by acoustic-prosodic properties such as pitch range and pausal duration (Avesani and Vayra, 1988; Ayers, 1992; Brown, Currie, and Kenworthy, 1980; Lehiste, 1979; Silverman, 1987) (cf. (Woodbury, 1987)), amplitude (Brown, Currie, and Kenworthy, 1980), speaking rate (Lehiste, 1980), and intonational prominence (Brown, 1983; Terken, 1984). Most such studies have relied on either intuitive analyses of notions such as topic-structure, or operational definitions of discourse-level properties, such as paragraph markings as indicators of discourse segment boundaries. However, an independently-movivated theory of discourse was used in (Hirschberg and Grosz, 1992; Grosz and Hirschberg, 1992) to identify intonational correlates of discourse structure, where discourse structural elements were determined by trained subjects following (Grosz and Sidner, 1986). This paper describes results of current research on the intonational correlates of discourse structure, based on a new corpus we have been collecting, the Boston Directions Corpus. This new data collection effort is aimed at addressing certain deficiencies observed during the earlier studies and at expanding our analysis to compare spontaneous with read speech.

The research reported here was supported by grants NSF IRI 9009018 and NSF IRI 9308173 from the National Science Foundation.

3. The Boston Directions Corpus

The Boston Directions Corpus comprises elicited monologues produced by multiple non-professional speakers, who were given written instructions to perform a series of increasingly complex direction-giving tasks. Speakers first explained simple routes such as getting from one station to another on the subway, and progressed gradually to the most complex task of planning a round-trip journey from Harvard Square to several Boston tourist sights. The speakers were provided with various maps, and could write notes to themselves as well as trace routes on the maps. For the duration of the experiment, the speakers were in face-to-face contact with a silent experimental partner (a confederate) who traced on her map the routes described by the speakers. The speech was subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted; subjects returned several weeks after their first recording to read the transcribed speech. Both sets of recordings were then acoustically and prosodically labeled using the ToBI labeling convention (Pitrelli, Beckman, and Hirschberg, 1994).

Discourse segmentations were obtained from three subjects labeling from text alone (group T) and three labeling from speech and text (group S). Each group of subjects was provided with a set of instructions for labeling in the Grosz & Sidner framework, a framework each labeler was already familiar with. The text for each task was presented to labelers as a sequence of text strings, each corresponding to an INTERMEDIATE PHRASE, according to (Pierrehumbert, 1980). Subjects were essentially asked to segment the discourse, identifying the underlying purposes for each segment and the hierarchical relationships among segments. Percentages for consensus labels (labels on which all labelers in the group agreed) for segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF) are given in Table I.

Two interesting trends have emerged from the current study. First, in contrast to our earlier findings,(Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992) group S segmentations differed significantly from those of group T. Table I shows that

TABLE I
Percentage of Consensus Labels by Segment Boundary Type

READ SPEECH (N=130)				
	Seg-initial (SBEG)	Seg-final (SF)	Segment-medial (SCONT)	All types
Text alone (T)	17%	17%	7%	38%
Speech & Text (S)	26%	24%	26%	75%
SPONTANEOUS SPEECH (N=145)				
	Seg-initial (SBEG)	Seg-final (SF)	Segment-medial (SCONT)	All types
Text alone (T)	17%	16%	16%	46%
Speech & Text(S)	25%	23%	33%	78%

listening to speech while segmenting produceed more consensus boundaries for both read and spontaneous speech than did segmenting from text alone. When the read and spontaneous data were pooled, labelers from text and speech agreed upon significantly more SBEG boundaries ($p<.05$, $\chi=4.5$, df=1) as well as SF boundaries ($p<.02$, $\chi=6.3$, df=1) than labelers from text alone. Further, it is not the case that segmenters from text simply chose to place fewer boundaries in the discourse; if this were so, then we would expect a high number of SCONT consensus labels where no SBEGs or SFs were identified. Instead, we find that the number of consensus SCONTs was significantly higher for the labelings from text and speech than for labelings from text alone, for read ($p<.001$, $\chi=11.7$, df=1) and spontaneous speech ($p< 1.5 \times 10^{-9}$, $\chi=36.6$, df=1). These factors combined to yield significantly higher percentages of consensus labels overall, for both read ($p< 1.8 \times 10^{-9}$, $\chi=36.1$, df=1) and spontaneous speech ($p< 1.4 \times 10^{-8}$, $\chi=32.2$, df=1). We conclude that aspects of the speech signal can help disambiguate among alternate segmentations of the same text, and thus the availability of speech critically influences the outcome of discourse structure analysis.

The second trend concerns a somewhat surprising effect of speaking style on segmentation, namely that of read versus spontaneous speaking modes. Spontaneous speech is generally claimed to exhibit less reliable prosodic indicators of discourse structure than read speech (cf. (Ayers, 1992)). Yet, in our corpus, spontaneous speech actually produced significantly more SCONT consensus labels than did read speech, for groups S and T combined ($p<.004$, $\chi=8.7$, df=1). The higher overall percentages of consensus labels for spontaneous speech are attributable to this difference in SCONT labelings.

We also examined the following acoustic and prosodic correlates of consensus labelings of intermediate phrases labeled as SBEGs and SFs: f0 maximum and average f0; rms maximum and average; speaking rate; and duration of preceding and subsequent pauses. We compared segmentation labels not only for group S versus group T, but also for spontaneous versus read speech. As noted, while intonational correlates for segment boundaries *have* been identified in read speech, they have been observed in spontaneous speech rarely and descriptively. In contrast, we have found strong correlations for consensus SBEG and SF labels for groups S and T in both spontaneous speech and read speech.[1] Results on consensus SBEG and SF labels for labelers who labeled from text alone and for labelers labeling from text and speech are shown in Tables II and III.

Given group T segmentations, we found significantly higher f0 maximum and average f0, and rms maximum and average, and shorter subsequent pause, for both spontaneous and read speech; for read speech we also found significant correlations for preceding pauses. Given group S segmentations, we found significantly higher maximum and average f0, higher maximum rms, longer preceding and shorter succeedings pauses for read and spontaneous speech, and higher average rms as well for read speech. Results on consensus SF labels were as follows: given group T segmentations, we found significantly lower average f0 and rms maximum for both read and spontaneous speech, and lower rms average and subsequent pause in addition for read speech. Given group S segmentations, we found lower average f0, rms maximum, rms average, shorter preceding pause and longer subsequent pause for both read and spontaneous speech,

[1] T-tests were used to test for statistical significance of difference in the means of phrases, e.g. beginning and not beginning segments. Results reported are significant at the .025 level or better.

TABLE II
Acoustic/Prosodic Correlates of Consensus Labelings from Text Alone

	F0 Max	Avg F0	RMS Max	Avg RMS	Rate	Preceding Pause	Subsequent Pause
Read SBEG	higher	higher	higher	higher		longer	shorter
Spon SBEG	higher	higher	higher	higher			shorter
Read SF		lower	lower	lower			longer
Spon SF		lower	lower				

TABLE III
Acoustic/Prosodic Correlates of Consensus Labelings from Text and Speech

	F0 Max	Avg F0	RMS Max	Avg RMS	Rate	Preceding Pause	Subsequent Pause
Read SBEG	higher	higher	higher	higher		longer	shorter
Spon SBEG	higher	higher	higher			longer	shorter
Read SF	lower	lower	lower	lower		shorter	longer
Spon SF		lower	lower	lower		shorter	longer

and in addition, lower f0 maximum for read speech.

While these results now hold for only a single speaker, they are quite promising. We may hypothesize that speakers can convey structural information about a discourse in spontaneous, as well as in read speech. We may also hypothesize that this structural information is in fact signalled at least in part by prosodic and acoustic information, since discourse labelings produced while listening to speech correlated with more acoustic-prosodic features than labelings from text alone. Certain acoustic-prosodic features such as preceding pause, for example, appear to have been made use of in segmentation decisions for group S. Our continuing analysis of this corpus will test the generality of these trends against more data, including speech from multiple speakers and discourse segmentations produced by naive subjects.

In our parallel study of intonational prominence in this corpus, a total of 173 noun phrases in the read speech for the five direction-giving discourses were analyzed for lexical form (e.g., proper names, definite/indefinite NPs), grammatical function (e.g., subject, direct object, prepositional object), surface-order position (e.g., sentence-initial, medial, final), position in major intonational phrase (e.g., phrase-initial, medial, final), and accentuation (unaccented or pitch accent type). Similar to the pilot study findings, 23% of noun phrases are accentually reduced, i.e. bearing fewer pitch accents than the citation-form.[2] Preliminary analysis indicates that lexical form and grammatical function are significant factors in determining accentuation, with names being less reduced than full NPs, and subjects less reduced than objects.[3] Surface-order and intonational phrase position were not significant.

As in earlier work (Nakatani, 1993), we found that the simple notion that references to given entities in the discourse should be accentually reduced fails to predict accentuation, since reintroductions of discourse-old entities are often accented. Interestingly, it was not the case either that repetitions of the same referring expressions were accentually reduced any more than were alternate lexical expressions referring to discourse-old entities. In the case of repeated referring expressions, the second mention is usually unreduced in intonational prominence when a discourse segment boundary intervenes between it and the first mention. Thus, accentual reduction cannot be considered an epiphenomenon of lexical givenness; if it were so, then lexical repetitions should simply be reduced. Rather, discourse structure interacts with lexical and other factors to constrain the deac-

[2]The AT&T Text-to-Speech System was used to determine the majority of citation-form accent assignments. Two native speaker judgments were used for items not in the TTS lexicon.

[3]Chi-square tests were used to test significance. Results reported are at the .02 level or better.

centuation of given information.

Finally, certain cases of accentual reduction did not arise previously in the narrative pilot study. In the Boston Directions Corpus, head nouns are frequently deaccented in full NPs with accented adjectival modifiers. This phenomenon typically occurs when the speaker contrasts two referential tokens of the same type (e.g. RED line SUBWAY vs. GREEN line subway, RIGHT TURN vs. ANOTHER right turn). However, the two tokens are not confined to the same discourse segment. This poses a problem for the global focusing mechanisms in (Grosz and Sidner, 1986), which assumes that entities in sister segments cannot be simultaneously in global focus. We will explore whether limited relaxations of this assumption, such as considering the most recently popped focus space to be in non-immediate global focus, can account for our cases of accentual reduction. A similar relaxation was necessary to account for the deaccentuation of object proper names in the narrative study.

4. Conclusions

In sum, results from our earlier studies of intonation and discourse suggested that discourses can be segmented reliably, that intonation is used by speakers to convey information about structure at the discourse level, and that the relationship among intonational features and discourse elements is more complex than previous studies have suggested. Current analyses of the Boston Directions Corpus have supported these hypotheses, and have also uncovered important effects of speaking style and segmentation methodology on the ability to obtain reliable analyses of discourse structure. Contrary to expectations, we found that discourse structure analysis is most robust for spontaneous speech labeled from speech and text together. Findings of these corpus-based studies in sum suggest that looking at spoken language can lead to improvements in the descriptive and computational adequacy of theories about discourse structure as well as intonational meaning.

ReferencesReferencesReferences

Avesani, Cinzia and Mario Vayra. 1988. Discorso, segmenti di discorso e un' ipotesi sull' intonazione. In *Atti del Convegno Internazionale "Sull'Interpunzione"*, Florence.

Ayers, Gayle M. 1992. Discourse functions of pitch range in spontaneous and read speech. Presented at the Linguistic Society of America Annual Meeting.

Brown, G. 1983. Prosodic structure and the given/new distinction. In D. R. Ladd and A. Cutler, editors, *Prosody: Models and Measurements*. Springer Verlag, Berlin, pages 67–78.

Brown, G., K. Currie, and J. Kenworthy. 1980. *Questions of Intonation*. University Park Press, Baltimore.

Grosz, B. and J. Hirschberg. 1992. Discourse structure and intonation. Israeli Conference on Theoretical Linguistics, Bar Ilan, June.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hirschberg, J. and B. Grosz. 1992. Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, pages 441–446, Harriman NY, February. DARPA, Morgan Kaufmann.

Lehiste, I. 1979. Perception of sentence and paragraph boundaries. In B. Lindblom and S. Oehman, editors, *Frontiers of Speech Research*. Academic Press, London, pages 191–201.

Lehiste, I. 1980. Phonetic characteristics of discourse. Paper presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan.

Nakatani, C. 1993. Accenting on pronouns and proper names in spontaneous narrative. In *Proceedings of the European Speech Communication Association Workshop on Prosody*, Lund, Sweden, September.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.

Pitrelli, John, Mary Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *Proceedings of the Third International Conference on Spoken Language Processing*, volume 2, pages 123–126, Yokohama. ICSLP.

Silverman, K. 1987. *The Structure and Processing of Fundamental Frequency Contours*. Ph.D. thesis, Cambridge University, Cambridge UK.

Terken, J. 1984. The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech*, 27:269–289.

Woodbury, Anthony C. 1987. Rhetorical structure in a central Alaskan Yupik Eskimo traditional narrative. In J. Sherzer and A. Woodbury, editors, *Native American Discourse: Poetics and Rhetoric*. Cambridge University Press, Cambridge UK, pages 176–239.