COMBINING STATISTICAL AND PHONETIC ANALYSES OF SPONTANEOUS DISCOURSE SEGMENTATION

Marc Swerts
Institute for Perception Research (IPO), Eindhoven, The Netherlands

ABSTRACT

This paper presents a method to study prosodic features of discourse structure in unrestricted spontaneous speech. Past work has indicated that one of the major difficulties that discourse prosody analysts have to overcome is finding an independent specification of hierarchical discourse structure, so that one avoids circularity. Previous studies have tried to solve this problem by constraining the discourse or by basing segmentations on a specific discourse theory. The current investigation explores the possibility of experimentally determining discourse boundaries in unrestricted speech. In a next stage, it is investigated to what extent boundaries obtained in this way correlate with specific prosodic variables.

INTRODUCTION

It is intuitively clear that most discourse exhibits structure in that it consists of larger-scale information units, or discourse segments. Those segments can be viewed as building a discourse hierarchy, since segments may be embedded in others: for instance, someone may talk about his holidays, with subtopics on hotel, food, and so on. In practice, however, it is often difficult to specify exactly the boundaries of those higher-level units and their mutual relationships. This poses a serious methodological problem to investigators who study linguistic, e.g., prosodic, correlates of discourse structure. Therefore, one would like to obtain ideally an 'independent' specification of information structure so as to avoid circularity. That is, it needs to be guaranteed that the junctures in the information flow are determined independently from prosodic considerations [1]. In the literature, one sees basically two solutions to overcome this problem.

In a first line of research, the problem is somewhat circumvented by looking at construed speech materials. One group of researchers has looked at read-aloud texts with predetermined paragraph boundaries (e.g., [9], [3], [6]); similarly, others have focused on tightly constrained types of spontaneous speech, by experimentally eliciting discourse in such a way that it becomes easily segmentable in consecutive information units ([8], [7]). In this way, prosodic features of discourse segments can be adequately investigated. These studies are limited, though, in that the structures investigated are controlled, but overly simple. It remains to be seen to what extent the findings can be generalized to more complex discourse.

The second approach is more theory-based in an attempt to motivate segmentations on the basis of explicit models of discourse structure. In studies such as [2] and [4], both within the Grosz and Sidner framework, 7 subjects were instructed to segment a set of monologues, using speaker intention as a criterion. It turns out that there is considerable variation between labelers as no two segmentations are the same. In particular the specification of hierarchical relationships between segments appears to be difficult. Therefore, it is decided in these studies either to concentrate on only those structural features agreed upon by all labelers [2], or retain those boundaries assigned by at least 4 out of 7 labelers [4].

This paper addresses another approach, partly inspired by Rotondo ([5]). Instead of taking the variance between labelers as a disadvantage, it rather exploits it to specify hierarchically different discourse boundaries. In contrast with earlier studies, it takes the segmentations of relatively many labelers to arrive at this goal. Basically, boundary strength is then computed as the proportion of subjects agreeing on a given break. In the following, it will be illustrated how this method offers a useful alternative to already existing procedures.

METHOD

The speech materials used in this study consisted of 12 spontaneous monologues (Dutch): 6 painting descriptions produced by 2 female speakers, MM and LK, amounting to 46.5 minutes of speech in total.

In a task that was individually performed, 38 subjects were instructed to mark paragraph boundaries in transcriptions of the monologues presented without interpunction or specific layout to indicate paragraph structure. Subjects were told to draw a line between the word that ended one paragraph and the one that started the next paragraph. No explicit definition of a paragraph was given. There were two conditions: half of the subjects could listen to the actual speech (SP condition), whereas the other half could not (TA condition). The reason to have both these conditions was to gain insight into the added value of prosody.

A typical example of part of a text is given below, followed by a literal translation in English. The two digits between round brackets represent the boundary strenght estimates, computed as the proportion of subjects indicating that there was a break, for the SP and TA condition, respectively. For sake of presentation, boundaries of strength 0 in either of the two conditions are only given when there is a stronger break in the other condition.

> het is echt een paard dat [uh] over iets heel springt heel heel snel (**0.26; 0.11**) de man die d'r opzit die zit ook helemaal in zo'n gebogen [uh] [uh] ruitershouding met zijn billen omhoog en zijn [uh] hoofd (**0;0.05**) in de manen van het paard (**0.95;0.16**) het paard is wit (**0.11;0**) [uh] ruiter is is [uh] rozig rood (**0.53;0.79**)
>
> (it is really a horse that [uh] jumps across something very very fast the man who sits on it he really sits also in such a bent over [uh] [uh] rider's position with his bum in the air and his [uh] head stuck in the mane of the horse the horse is white [uh] rider is is [uh] pinkish red)

As can be seen, the breaks between word clusters may vary between relatively weak ones (e.g., 0.05) to relatively strong ones (0.95). In total, the two monologues gave 889 'minimal units', i.e., sequences of words not seperated by any of the labelers.

RESULTS

Comparing SP with TA

A rough idea of the differences for the two conditions can be derived from figure 1, which shows the boundary strenght values for one typical monologue. The figure shows that the two

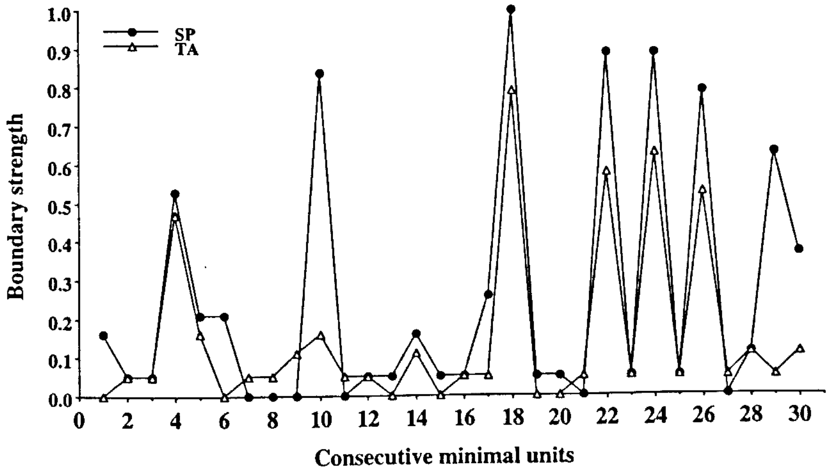

Figure 1: Boundary strength values for consecutive minimal units in SP and TA condition (further explanations in text)

experimental conditions give similar results, but segmentation is clearer in the text-with-speech case. The stronger breaks are more fully pronounced in the SP condition in the sense that proportionally more subjects agree on a paragraph transition. Also, although not visible in figure 1, some passages in the monologue receive different segmentations (indicating structural ambiguity) in the text-alone condition; such sections appear to be unambiguous when subjects have access to speech. This shows that prosody can limit the text interpretation.

Phonetic analyses

Given these observations, the speech was phonetically analyzed to explore potential relationships between boundaries obtained by the Rotondo method and prosodic features. Only the phonetic correlates of the boundaries in the SP condition are studied. Inspired by the literature, measurements include pitch range and pause, taking F0 maximum and silent interval as the respective acoustic correlates. The distribution of different boundary tones was also investigated. To have a unit of analysis, the monologues were transcribed by an independent labeler, who had to mark both the boundaries of phrases plus the sort of boundary tone at their respective ends (see below).

Pitch range In any given phrase, the highest F0 peak in an accented syllable was taken as a measure of pitch range ([2]). The results are given in Table 1 with the median values for pitch range in phrases following a boundary of a particular strength. The strength estimates are clustered into 5 groups, i.e,. one cluster containing values for breaks on which up to 25% of the labelers agreed, the next cluster having agreements between 25 and 50%, etc. Phrases within minimal units, i.e., units not subdivided by any of the labelers, were taken as a seperate category, since they formed a relatively large group. The median, rather than the mean, was taken as a more conservative measure to obtain a rough estimate, since the data in the different clusters were not always normally distributed. (A similar procedure is followed in Tables 2 and 3.) From Table 1 it can be seen that pitch range covaries with the depth of a discourse break. Pitch tends to become higher when the phrase follows a stronger boundary. This is true not only as an overall result, but also for the two speakers seperately.

Table 1: Median values of pitch range (in Hz) for speakers LK, MM seperately and pooled as a function of the different clusters of boundary strength

	0	0- .25	.25- .50	.50- .75	.75- 1
LK	228	231	238	245	252
MM	221	231	239	238	245
Both	225	231	238	244	249

Pause Pauses were measured as silence intervals of at least 1 second long. Results are shown in Table 2, giving the median length of pauses preceding a boundary of a particular strength. It reveals a similar tendency as with the data for pitch range. Looking at the overall data, it appears that pauses gradually become longer as a function of the strength of the discourse boundary, although the details for the two speakers seperately are somewhat more complex: the lengths for LK somewhat level off for the .75-1 cluster, whereas there is a sudden increase there for MM.

Table 2: Median values of pause (in s) for speakers LK, MM seperately and pooled as a function of the different clusters of boundary strength

	0	0- .25	.25- .50	.50- .75	.75- 1
LK	0	0	1.4	2.5	2.0
MM	0	0	1.9	1.6	4.5
Both	0	0	1.4	2.0	3.2

Boundary tone Finally, the distribution of different types of boundary tones was investigated. Originally, the transcriber was instructed to mark phrases as ending in a low, mid or high boundary tone. In this paper, the latter two categories are collapsed into one, i.e., non-low boundary tones. Table 3, giving the proportion of low boundary tones, shows that the chance that such a tone will occur becomes higher as a function of the strength of the boundary.

Table 3: Proportion of low boundary tones for speakers LK, MM seperately and pooled as a function of the different clusters of boundary strength

	0	0- .25	.25- .50	.50- .75	.75- 1
LK	.04	.18	.24	.27	.40
MM	.10	.29	.40	.38	.42
Both	.07	.22	.30	.31	.41

DISCUSSION

The contribution of this paper is primarily methodological in that it presents a technique to analyse hierarchical discourse structure and its potential phonetic correlates in unrestricted discourse. It is a useful alternative to existing methods, as it is general and reproducible. A major disadvantage, however, is that the boundary strength measure (ideally) requires a large amount of subjects.

As for the prosodic results, it is interesting to see that prosodic variables such as pitch range, pause length and number of low boundary tones increase continuously with boundary strength at the discourse level. This is similar to prosodic phrasing results below the level of the sentence ([10]).

Of course, the features studied in this paper are not the only potentially interesting ones. In particular, preliminary observations suggest that transitions between major information units are accompanied by hesitation phenomena, such as filled pauses that point towards planning processes. These constitute an interesting area for further research. Also, future work will have to determine in what ways the experimentally based discourse boundaries correspond to junctures predicted by discourse theories.

ACKNOWLEDGMENTS

M. Swerts is also affiliated with the University of Antwerp (UIA). Thanks are due to R.-J. Beun for providing the speech materials, to E. Blaauw for transcribing them, and to R. Collier and A.A. Sanderman for commenting upon an earlier version of this paper.

REFERENCES

[1] Brown, G., Currie, K. and Kenworthy, J. (1980): *Questions of intonation*. London: Croom Helm.

[2] Grosz, B. and Hirschberg, J. (1992): "Some intonational characteristics of discourse structure," *ICSLP 92*, pp. 492-432.

[3] Lehiste, I. (1975): "The phonetic structure of paragraphs," *Structure and process in speech perception* (ed. by Nooteboom and Cohen), pp. 195-206.

[4] Passonneau, R.J. and Litman, D.J. (1993): "Intention-based segmentation: human reliability and correlation with linguistic cues," *ACL-93*.

[5] Rotondo, J.A. (1984): "Clustering analyses of subjective partitions of text," *Discourse Processes* 7, pp. 69-88.

[6] Sluijter, A. and Terken, J. (1994): "Beyond sentence prosody: Paragraph intonation in Dutch," *Phonetica* 50, pp. 180-188.

[7] Swerts, M. and Geluykens, R. (1994): "Prosody as a marker of information flow in spoken discourse," *Language and Speech* 37, pp. 21-43.

[8] Terken, J. (1984): "The distribution of pitch accents in instructions as a function of discourse structure," *Language and Speech* 27, pp. 269-289.

[9] Thorsen, N.G. (1985): "Intonation and text in Standard Danish," *JASA* 77, pp. 1205-1216.

[10] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P.J. (1992): "Segmental durations in the vicinity of prosodic phrase boundaries," *JASA* 91 (3), pp. 1707-1717.