Prosodic features at discourse boundaries of different strength

Marc Swerts[a)]
Institute for Perception Research (IPO), P.O. Box 513, NL-5600 MB Eindhoven, The Netherlands

(Received 22 November 1995; revised 17 June 1996; accepted 27 August 1996)

This paper presents the design and the evaluation of 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 as to avoid 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 first 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: the features pause, pitch range, and type of boundary tone are studied as a function of discourse structure. © *1997 Acoustical Society of America.* [S0001-4966(97)01701-3]

PACS numbers: 43.71.Es, 43.70.Fq, 43.70.Hs, 43.70.Bk [RAF]

INTRODUCTION

It is intuitively clear that a coherent discourse exhibits a certain syntax, comparable to that at the sentence level. This implies that spoken or written texts consist of larger-scale information units, or discourse segments, which can be viewed as building a hierarchical structure. That is, segments may contain or may be embedded in others: for instance, someone may talk about his holidays, with subtopics on hotel, food, etc. The structure is generally reflected in the distribution of particular linguistic phenomena such as anaphora (Grosz and Sidner, 1986; Reichman, 1985; Webber, 1988; Geluykens, in press) or discourse particles (Shiffrin, 1987; Passonneau and Litman, 1993). It has also its consequences for prosody, which is investigated in this paper, since such features as speech melody, pause and tempo are likely candidates to highlight the make-up of a spoken text.

However, investigators who study linguistic, for instance, prosodic, correlates of discourse structure are confronted with a serious methodological problem. Ideally, one would like to obtain an ''independent'' specification of junctures in the information flow and of the mutual relationships between segments. It is of course known that prosody may influence subjects' perceptions of discourse structure, so that it is legitimate that investigators use it as one source of information to determine discourse structure. But it should be guaranteed that the identification of boundaries does not solely depend on prosodic considerations, to minimize the risk of circularity. In order to make them useful for prosody research, the criteria developed to mark discourse also need to be explicit and reproducible, and need to be more than one individual's intuitive notion of information structure.

In practice, however, it is unclear to what extent such autonomous and reproducible macro-labeling can be achieved on the basis of existing discourse theories. Although these may start from legitimate assumptions, taking, for instance, coherence relations (Mann and Thompson, 1986) or topicality (van Kuppevelt, 1995; Brown and Yule, 1983) as a basic organizing principle of discourse, they often lack a degree of explicitness to enable (naive) subjects to reliably label a given text in terms of structural features. Coding schemes to annotate dialogue structure in terms of utterance functions have been developed and tested in the HCRC Map Task Corpus (Carletta *et al.*, 1996a, 1996b) and the VERBMOBIL project (Reithinger, 1995; Jekat *et al.*, 1995). However, there are virtually no empirical investigations, except for a few in the Grosz and Sidner framework (see below), to test to what extent subjects are able to consistently identify ''discourse segments'' using theory-based guidelines. On the other hand, story grammars such as the ones by Thorndyke (1977) and Mandler and Johnson (1977) do present generative grammars, with explicit rewrite rules defining units and relationships, but those schemata are very much genre specific, i.e., typical for a particular type of narrative prose, so that their usefulness as a general system for discourse labeling remains questionable.

There have been a few attempts reported in the literature to overcome the problem of independently specifying discourse structure. In a first line of research, the whole issue is somewhat circumvented by looking at constructed speech materials. One group of researchers has looked at read-aloud texts with predetermined paragraph boundaries (for instance, Thorsen, 1985, 1986; Lehiste, 1975; Sluijter and Terken, 1994; Brubaker, 1972; Bruce, 1982); 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 (Terken, 1984; Swerts and Collier, 1992; Swerts *et al.*, 1994; Swerts and Geluykens, 1994; Venditti and Swerts, 1996). In this way, prosodic features of discourse segments can be adequately investigated. It was indeed found that larger-scale information units have an impact on the distribution of accents, and prosodic characteristics such as pause, globally declining melodic patterns, local boundary tones, and durational variation. These studies are limited, though, in that the structures investigated, while being controlled, are overly simple. An intrinsic danger, therefore, is that the role of prosody as a means to signal information structure is exaggerated, since there are no ''disturbing'' factors of interac-

[a)]Electronic mail: swerts@natlab.research.philips.com

tive or attitudinal nature. It is thus an open question to what extent findings from these experimental studies can be generalized to more complex types of discourse.

The second approach is theory-based, in an attempt to motivate segmentations on the basis of an explicit model of discourse structure, i.e., the one proposed by Grosz and Sidner (1986). In studies by Grosz and Hirschberg (1992; see also Hirschberg and Grosz, 1992) and Passonneau and Litman (1993), seven subjects were instructed to segment a set of monologues, using speaker intention as a criterion. It turns out that there is some variation between labelers as ''no two segmentations were identical'' (Hirschberg and Grosz, 1992, p. 443). In particular the specification of hierarchical relationships between segments appears to be difficult. It is decided in these studies either to concentrate on only those structural features agreed upon by all labelers (Grosz and Hirschberg, 1992), or retain those boundaries assigned by at least four out of seven labelers (Passonneau and Litman, 1993). The studies report that the resulting discourse segments have clear prosodic features in terms of pitch range (Grosz and Hirschberg, 1992) and pause (Grosz and Hirschberg, 1992; Passonneau and Litman, 1993). However, from these investigations it cannot be decided whether prosody can also be exploited to directly signal the level of embeddedness of a given segment. Passonneau and Litman (1993) have limited their study to a purely linear intention-based segmentation task, since more complex segmentation was thought to be too cumbersome, given the average length of their narratives. Grosz and Hirschberg (1992), on the other hand, do try to conduct an empirical study of hierarchies, but do not really use these hierarchical representations in the phonetic analyses afterward, since boundaries at different levels of embeddedness are collapsed to represent one category.

There have been a few more efforts in the mean time to explore the possibilities of determining discourse structure within the framework of Grosz and Sidner (Hirschberg *et al.*, 1995; Nakatani *et al.*, 1995). Even a manual with instructions for segmenting discourse has been produced, which is based on this particular theory and which is meant for ''naive'' subjects, but this guide still needs to be further evaluated (Nakatani *et al.*, 1995). Despite all these previous attempts, the current paper for two reasons introduces yet another method to establish discourse structure. First, although newer versions of the labeling manual will probably lead to more consistent results, it will probably not yield complete agreement on discourse structure, certainly when subjects are confronted with difficult, incoherent texts (Condon and Čech, 1996). Second, there are currently no generally accepted techniques available to assess the reliability of different discourse labelings. Carletta (1996) criticizes the measures used in many other discourse studies to evaluate the agreement between labelers, and proposes a so-called kappa statistic as an alternative.

Therefore, to arrive at an independent discourse labeling in view of prosody research, this paper addresses another approach, which—as will become clear in the next section—exploits the fact that subjects vary in the way they agree on discourse segmentations. It is partly inspired by Rotondo (1984), although he only considered labelings of texts, and not speech or transcribed speech. The goals of our study are twofold: first, it will be explored to what extent the experimentally oriented method offers a useful alternative to already existing procedures; second, it is investigated to what extent boundaries obtained in this way correlate with specific prosodic variables. More specifically, the paper studies whether prosody may play a role in signaling hierarchical discourse structure.

I. METHOD

Rotondo (1984) starts from the idea that from group segmentation data one can derive a hierarchical discourse structure. When subjects, for instance, are asked to mark where one ''complete thought'' has ended and another one begins, they will expectedly have some uncertainty about what constitutes such a unit. Consequently, such a task will lead to inconsistencies among subjects, since some will put a lot of boundaries, whereas others only a few, or in different places. It is logical to assume, however, that there will be less disagreement about stronger breaks since they present clearer transitions in the flow of information. It can thus be expected that more people will feel inclined to mark these as a boundary. Therefore, instead of taking the variance between labelers as a disadvantage, one can rather exploit it to specify hierarchically different discourse boundaries. To arrive at this goal, the segmentations of relatively many labelers are needed. Basically, boundary strength is then computed as the proportion of subjects agreeing on a given break, assuming that the degree of certainty about a given discourse boundary reflects its level of embeddedness. At the same time, the task of simply having to mark the boundary between two consecutive units is relatively easy and less time consuming, compared to a technique where subjects are asked to express boundary strength on, for instance, a ten-point scale, which was previously used to study prosodic phrasing at the sentence level (de Pijper and Sanderman, 1994).

The speech materials used in the present study were taken from a larger corpus of 30 spontaneous Dutch monologues (Beun, 1991). These were originally collected in an experiment in which subjects had to describe a set of paintings. It is important to notice that the resulting narratives were not controlled in terms of discourse structure, since, in principle, speakers had the freedom to organize their descriptions as they liked. Twelve monologues, i.e., six descriptions produced by two speakers (MM and LK), amounting to 46.5 min of speech in total, were selected for further analysis as they were comparable in length (3 to 5 min), and the speakers were both female, using about the same pitch range (which could facilitate prosodic comparisons afterward).

Those selected descriptions were used in a task that was individually performed by 38 subjects, who were instructed to mark paragraph boundaries in transcriptions of the monologues. These were presented without punctuation 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.

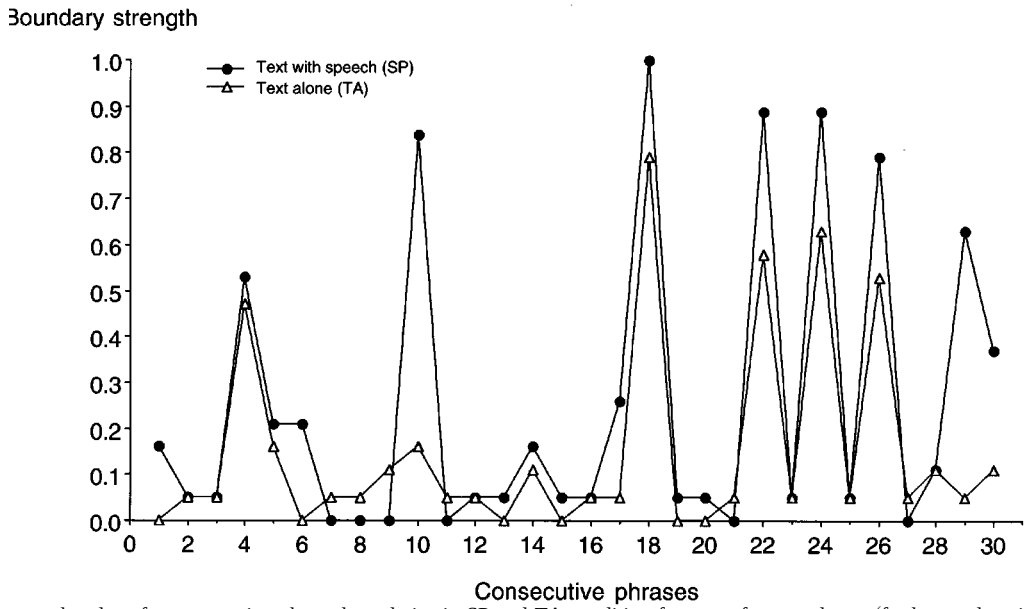

FIG. 1. Boundary strength values for consecutive phrase boundaries in SP and TA condition for part of a monologue (further explanations in text).

There were two conditions: half of the subjects could listen to the actual speech (SP condition), whereas the other half only saw the written version (TA condition). The reason to have both these conditions, a procedure also used by Grosz and Hirschberg (1992), was to gain insight into the added value of prosody. The underlying hypothesis is that prosody helps the listener to better understand the discourse so that speech is comparably less ambiguous than text alone.

A typical example of part of a text is given below, followed by a literal translation in English. The two digits in round brackets represent the boundary strength 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 heen 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 from relatively weak ones (for instance, 0.05) to relatively strong ones (0.95).

In principle, as subjects were not given any constraints on how they should segment the discourse, every word boundary was a potential candidate for a breaking point between two paragraphs. But from the previous example, it is clear that one can distinguish some ''minimal units,'' i.e., sequences of words not separated by any of the labelers. It appeared that almost all the boundaries of those units coincided with prosodic phrase boundaries, although minimal units could contain more than one phrase; therefore, in the rest of this paper, the prosodic phrase will be taken as a unit of analysis. To this end, prosodic phrases were marked by one experienced phonetician (other than the author), who did not have access to the segmentation results of the 38 labelers. She was asked to assign both the boundaries of phrases and the kind of boundary tone (high, mid, or low) at their respective ends. No explicit definition was given of either prosodic phrase or boundary tone. Her labelings resulted in a total of 889 phrases for the two monologues (535 for LK and 354 for MM). In taking the prosodic phrase as an unit of analysis, the study becomes similar to that of Passonneau and Litman (1993) and Hirschberg and Grosz (1992) who restricted subjects to placing boundaries between prosodic phrases.

To check whether the proportion of subjects agreeing on a break can truly be seen as a measure of boundary strength, one monologue of speaker MM was used as input for a small comparison test. The monologue, split into consecutive phrases, was presented to 20 subjects, none of whom had participated in the previous labeling experiment. They were asked to assign a number to each transition between two consecutive phrases using a ten-point scale, with ''1'' meaning ''very weak boundary'' (word boundary) and ''10'' meaning ''very strong boundary'' (paragraph boundary). As in the previous test, half of the subjects could listen to the speech (SP condition), whereas the other half had but the text to label the discourse (TA condition). It appears that this method of establishing the depth of a discourse boundary gives results which are very comparable to the one based on proportions of subjects. The correlations between the ratings on the ten-point scale (averages of ten labelers) and the proportions, are highly significant, both in the SP ($r=0.75$, $n=30$, $p<0.0001$) and in the TA condition ($r=0.70$, $n=30$, $p<0.0001$). In other words, the assumption that the degree of

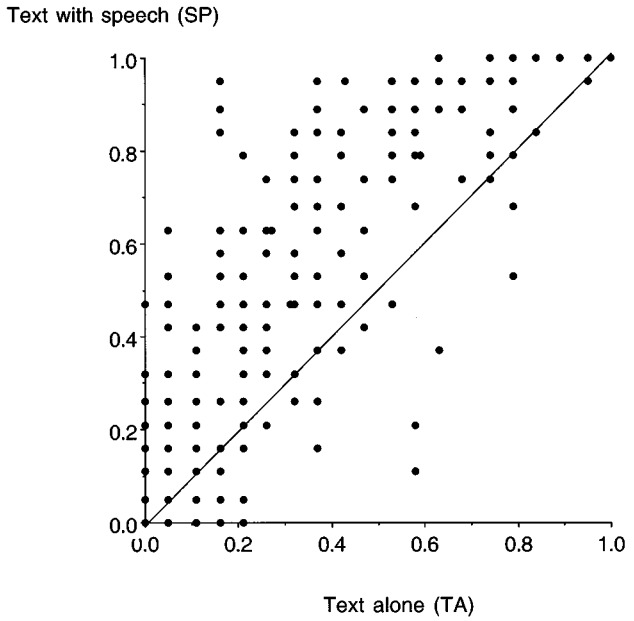

FIG. 2. Scattergram of boundary strength values in the SP and TA condition.

certainty about a given discourse boundary reflects its level of embeddedness appears to be valid.

II. RESULTS

A. Comparing text-with-speech (SP) and text-alone (TA)

A first idea of the segmentation results and of the differences for the two conditions can be derived from Fig. 1, which shows the boundary strength values for part of one typical monologue by speaker MM. The figure reveals that the two experimental conditions produce results which have both similarities and differences. First inspection reveals that the two conditions are comparable in the sense that they both have major breaks in similar positions; they are different in that segmentation is clearer in the text-with-speech case: stronger breaks are more clearly visible in the SP condition as proportionally more subjects agree on a paragraph transition. Also, focusing on the passage around phrase number 10 in Fig. 1 shows that in the text-alone condition, different subjects may share the impression that within a ''window'' of a few phrases, there is a break, but they disagree on the exact location of this boundary. Such sections which receive different segmentations in the TA condition (indicating structural ambiguity) often appear to be unambiguous when subjects have access to speech.

This is supported quantitatively in Fig. 2, which gives a scattergram of the values obtained in the SP and the TA condition. The Pearson correlation coefficient between the responses is relatively high: $r=0.89$ ($n=889$, $p<0.001$) which shows that there is a strong association between the two conditions. This is even true when the 0 boundaries are left out of the comparison, as they form an overwhelming class: $r=0.82$ ($n=252$, $p<0.001$). The data points in this figure that fall on the diagonal represent the scores that were identical in the two conditions. As a general picture, however, one can observe that most of the scores are located above this line, meaning that in the SP condition systematically more subjects agree on a paragraph transition.

As a particular data point in Fig. 2 may contain more than one value, it is not visible from this plot how many values fall within a certain range. This becomes obvious, however, when looking at Fig. 3 which, for the two conditions, gives a distribution of the strength estimates. These are clustered into five groups, i.e., one cluster containing values for phrase boundaries which up to 25% of the labelers marked as a paragraph transition, 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 separate category, since they formed a relatively large group. First, the figure reveals—as expected—that in both conditions you get relatively few ''strong'' boundaries and relatively many ''weak'' ones. At the same time, it appears that the SP condition compared with the TA condition has comparatively more values in the extreme clusters, i.e., the 0 boundaries and the ones of strength >0.75. Apparently, when subjects have access to speech, they are surer about both strong and zero boundaries, whereas in the TA condition, the scores are more diffuse. In other words, this indicates that people have clearer segmentations in the SP condition. This is supported with a chi square test, which shows that the two distributions are statistically different ($\chi^2=40.36$, $p<0.001$, $df=4$): the effect is mainly due to the differences in the <0.25 and the >0.75 classes.

B. Phonetic analyses

The previous section has brought to light that segmentation is clearer when subjects have access to speech. To explain this, one need to learn more about how particular utterances are actually spoken, because this is information subjects lack when they have but the text to label the discourse. In this section, the prosodic structure of the monologues is studied, in order to find out whether it can (partly) account for the differences between the two conditions.

The speech was phonetically analyzed to establish 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, since prosody obviously could not have played a role in the other condition. Measurements include pitch range and pause, taking F_0 maximum[1] and silent interval as the respective acoustic correlates. The distribution of different boundary tones was also investigated, as transcribed by an independent labeler (see above). The choice of these prosodic variables was inspired by the literature, since these were mentioned as being important phonetic structuring devices. Potentially interesting prosodic markers of discourse structure, such as final lengthening (Lehiste, 1979) and variation in speech tempo (Brubaker, 1972), were not investigated, because phonetically segmented and annotated versions of the monologues were not available.

1. Pause

A first major device reported in the literature to mark boundaries of different kinds is pause (for instance, Swerts

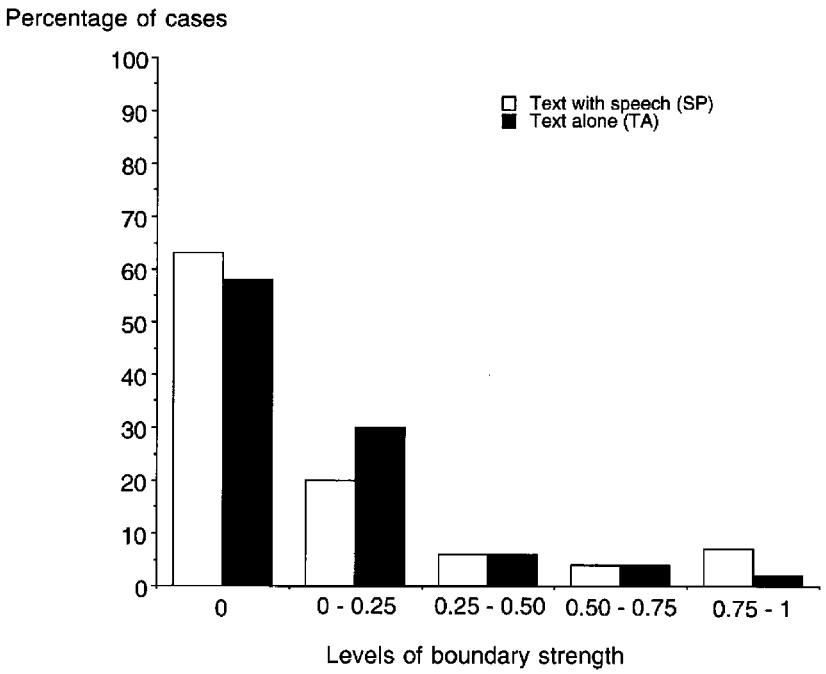

FIG. 3. Percentage of responses for different degrees of boundary strength in SP and TA condition.

and Geluykens, 1994; Grosz and Hirschberg, 1992; de Pijper and Sanderman, 1994; Passonneau and Litman, 1993). Structural breaks are generally accompanied by a significant silent interval. It can therefore be expected that the breaks given by the current method are also signalled by pause. More specifically, it is explored whether differences in the strength of a break are marked by differences in the length of the pause. Strangert (1993) found in Swedish spoken news bulletins that the difference between phrases, sentences, and paragraphs is reflected in the length of the silent interval at the boundary.

The place and length of pauses were taken from the original database (Beun, 1991), in which only those pauses were retained of at least 1 s long. Additionally, shorter silent intervals were measured manually by the author with the minimum value set to 0.25 s.

There is a moderate, but very significant correlation between the boundary strength values and the pause durations: $r=0.63$ ($n=889$, $p<0.0001$), showing that there is trend for longer pauses to be associated with stronger breaks. Table I gives the average boundary strength values for each of six pause duration categories, for both speakers pooled and separately. An analysis of variance reveals a significant overall difference between the six pause categories ($F_{(5,883)}=184.35$, $p<0.0001$). *Post-hoc* Scheffé tests showed that all the conditions differed from each other significantly ($p<0.05$).

2. Pitch reset

Previous studies, at various levels of linguistic analysis (for instance, de Pijper and Sanderman, 1994; Swerts and Geluykens, 1994), have reported that breaks between information units, can be highlighted by means of melodic discontinuities. Pitch tends to decline in the course of an utterance, but is generally reset at junctures in the information flow. Therefore, it was hypothesized that there will be a strong tendency for a declination reset at major breaks given by the Rotondo method. It was also explored whether the depth of a break is reflected in the amount of resetting. Swerts *et al.* (1996) found for Swedish spontaneous and read-aloud monologues a systematic difference in pitch reset between units differing in depth, i.e., phrases and utterances, with a weaker reset in the former than in the latter.

The pitch resets were measured in two steps. First, in any given phrase, the highest F_0 peak in an accented syllable at the vowel's amplitude maximum was taken as a measure

TABLE I. Mean boundary strength values (and respective standard deviations and number of observations) associated with different degrees of pause length (in seconds) for speakers MM and LK, pooled and separately.

	Both			MM			LK		
Pause length	$\bar{x}$	s.d.	n	$\bar{x}$	s.d.	n	$\bar{x}$	s.d.	n
No pause	0.01	0.04	447	0.01	0.04	188	0.01	0.05	259
<1 s	0.08	0.16	198	0.08	0.18	76	0.08	0.15	122
1–2 s	0.26	0.32	159	0.23	0.29	48	0.28	0.33	111
2–3 s	0.37	0.28	24	0.24	0.32	8	0.44	0.25	16
3–4 s	0.58	0.29	30	0.44	0.30	14	0.69	0.24	16
>4 s	0.78	0.24	31	0.78	0.22	20	0.77	0.30	11

TABLE II. Mean boundary strength values (and respective standard deviations and number of observations) associated with different degrees of pitch reset (in semitones) for speakers MM and LK, pooled and separately.

	Both			MM			LK		
Pitch reset	$\bar{x}$	s.d.	n	$\bar{x}$	s.d.	n	$\bar{x}$	s.d.	n
<-4.11	0.02	0.05	19	0.02	0.04	9	0.02	0.07	10
$-4.11--2.07$	0.07	0.19	95	0.06	0.18	40	0.07	0.19	55
$-2.07--0.03$	0.07	0.19	310	0.08	0.19	118	0.07	0.18	192
-0.03–2.01	0.11	0.23	277	0.09	0.23	111	0.12	0.23	166
2.01–4.05	0.37	0.38	101	0.32	0.39	39	0.40	0.36	62
>4.05	0.40	0.38	25	0.39	0.39	12	0.41	0.39	13

of pitch range. This measure was taken rather than the actual F_0 peak, since the latter sometimes constituted a somewhat exaggerated value, certainly in cases where a prominence-lending pitch rise combined with a melodic boundary marker (Grosz and Hirschberg, 1992; Beckman and Ayers, 1994; Menn and Boyce, 1982). Second, the distance (in terms of semitones) was measured between the pitch range values before and after a given boundary of a particular strength. If there was an upward reset, this measure gives a positive number, and vice versa.

The correlation between the boundary strength values and the pitch reset data is low, but significant ($r=0.35$, $n=827$, $p<0.0001$). Table II lists the average boundary strength values for different degrees of pitch reset, again for both speakers pooled and separately. To create six categories, the reset data were clustered taking steps of 1 s.d. An analysis of variance showed a significant overall effect of pitch reset on boundary strength ($F_{(5,821)}=33.13$, $p<0.0001$). *Post-hoc* Sheffé comparisons showed that the resets are clustered in two groups, the four weakest versus the two strongest degrees of reset.

3. Boundary tone

The last prosodic variable investigated was a type of boundary tone. Several studies (Brown *et al.*, 1980; Swerts and Geluykens, 1994; Swerts *et al.*, 1994; Swerts *et al.*, 1994) have shown that speakers can signal information structure by means of different boundary tones. That is, phonologically low tones tend to occur at the end of discourse segments, whereas the high ones are generally found inside such units. Accordingly, the hypothesis here is that low tones are more typical in the final position of a major discourse segment, and conversely, high ones are more typical of more shallow breaks.

Originally, the transcriber who marked phrase boundaries (see above) was instructed to also mark them as ending in a low, mid, or high boundary tone. No specific framework was given. In this paper, the latter two categories are collapsed into one, i.e., nonlow boundary tones, because the transcriber complained about the difficulty to reliably distinguish between the high and mid tones. In this way, the prosodic labeling becomes identical to the distinction made by Brown *et al.* (1980).

Table III gives the average boundary strength values for the two types of low boundary tone, again for the two speakers pooled and separately. An analysis of variance shows that the two types are significantly different ($F_{(1,887)}=87.03$, $p<0.0001$).

C. Prediction of discourse structure

The previous section has shown that there are indeed systematic correspondences between the break indices obtained by the Rotondo method and prosodic variables. It is now explored to what extent the discourse structure can really be predicted by means of these prosodic variables. It is not expected that one can fully explain the segmentation results on the basis of the pausal and melodic cues alone, since subjects to a large extent appear to score similarly in the SP and TA conditions (see r values given above). The task given to the labelers essentially was a semantic one, as they actually were asked to mark information units for which basically only the content is needed. The hypothesis is that prosody may help the job of segmenting the text as it supports particular segmentation hypotheses on the part of a listener.

In order to find out more about the significance of prosodic variables for the segmentation task, a general linear model was fit to the scores from the SP condition using a method of least squares regression. The following additive model, not including interactions,

$$\text{boundary}=\text{pause}+\text{reset}+\text{tone}, \tag{1}$$

with pause and range clustered into six levels and with two types of boundary tone (low and nonlow) can account for 58% of the variance, with significant effects for each of the prosodic parameters (see Table IV). If the two-way and

TABLE III. Mean boundary strength values (and respective standard deviations and number of observations) associated with type of preceding boundary tone for speakers MM and LK, pooled and separately.

	Both			MM			LK		
Boundary tone	$\bar{x}$	s.d.	n	$\bar{x}$	s.d.	n	$\bar{x}$	s.d.	n
nonlow	0.09	0.22	761	0.08	0.21	290	0.10	0.22	471
low	0.31	0.36	128	0.27	0.35	64	0.36	0.37	64

TABLE IV. Results of a general linear model fit on the segmentation data in the SP condition.

Pause	$F_{(5,815)}=195.82, \quad p<0.0001$
Reset	$F_{(5,815)}=20.52, \quad p<0.0001$
Boundary	$F_{(1,815)}=30.98, \quad p<0.0001$

three-way interactions are added to the model, the explained variance increases to 65%. If the model is simplified to each of the prosodic factors separately, it turns out that pause can explain 51% of the variance, pitch reset 17% and boundary tone 9%.[2]

From the above, it can be concluded that prosody may have functioned as a cue to the labelers, but that it is clearly not the only factor. Obviously, subjects have probably relied extensively on the content of the texts presented. Since this was the only information for subjects in the TA condition, their scores, clustered into five categories (see above), were added as a parameter to the additive model in (1). The reason to include these responses is to gain insight into the relative importance of the semantic cues. In doing so, the explained variance increases to 85%. Including two-way interactions results in an explained variance of 90%. A fit on the text cues separately gives a value of 77%. So one may conclude that the textual cues are clearly predominant for the labelers, but that the contribution of the prosodic variables studied here is certainly not negligible.

III. DISCUSSION

The contribution of this paper is primarily methodological in that it presents a technique to analyze 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 minor disadvantage, however, is that the boundary strength measure (ideally) requires a large amount of subjects. Of course, as Rotondo (1984) already remarked before, this method gives a somewhat temporary solution since ultimately one would like to have an accurate theory of discourse processing to get an adequate representation of text structures. Consequently, a logical follow-up to the current work might be to try and determine in what ways the experimentally based discourse boundaries correspond to junctures predicted by discourse theories.

There are, however, a few drawbacks which need to be solved in the future, although they are probably not typical for this approach alone. First, some of the assumptions somewhat implicitly underlying this technique is that topic changes are abrupt in the sense that they occur locally between a phrase ending one information unit and another starting the next one. The technique so far does not capture more gradual and smooth transitions between discourse segments. Similarly, the method does not account for the fact that labelers may be confused about the exact spot at which a boundary occurs, but they may agree that within a certain ''region'' there is change of information unit. Both problems may, however, be solved by applying some sort of window on the data.

As for the prosodic results, it is interesting to see that prosodic variables such as pitch reset, pause length, and proportion 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: Wightman *et al.* (1992) found that preboundary lengthening appears ''to occur in several gradations corresponding to some of the perceptually distinct break indices'' (p. 1714); de Pijper and Sanderman (1994) also report that the strength of prosodic boundaries in spoken utterances perceived by listeners is not simply a binary feature as their continuous perceptual boundary measure covaries with such features as pause, pitch and declination reset in a seemingly gradient manner. From the current study, a similar picture emerges for phrasing at the macrolevel, because prosody has been shown to mark boundaries of segments at multiple levels of embeddedness.

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. This is in agreement with the claim by Chafe (1980) that significant breaks in the flow of information lead to conspicuous speech disfluencies, indicating that these are points where a speaker is planning a lot to verbalize his ideas. Similarly, Brubaker (1972) found that initial utterances of a paragraph tend to be uttered at a slower speaking rate than other utterances, being consistent with a ''reduction of uncertainty hypothesis,'' i.e., that speaking rate increases and pause duration decreases as uncertainty about the remaining content of the paragraph decreases (see also Swerts *et al.*, 1994). An interesting area for future research, therefore, is to explore whether discourse structure achieved by the Rotondo method has some psycholinguistic relevance. Given the claims by Chafe (1980), a speaker-oriented hypothesis is that hesitations will tend to occur at major breaks in the flow of information. First analyses on the monologues investigated here (Swerts *et al.*, 1996) indeed show that at least one type of disfluencies, i.e., filled pauses, may carry information about discourse structure, in that stronger breaks in the flow of information are more likely to co-occur with filled pauses than weaker ones. Moreover, filled pauses at stronger breaks tend to be segmentally and prosodically different from other ones, and more often have preceding and following silent pauses.

The Introduction remarked on the difficulty of obtaining independent specification of junctures in the information flow and of the mutual relationships between them. If such an independent specification could be achieved, the contribution of prosodic features could then be analyzed by studying correlations between the features and the independently established junctures. This article has introduced a new method to achieve such independent specification; this is based on equating boundary strength with the frequency of junctural judgments by subjects. The strong correlations found between the strength of junctures thus obtained and the strength of prosodic marking confirms the efficacy of the methodology and opens up a new path of research.

ACKNOWLEDGMENTS

M. Swerts is also affiliated with the University of Antwerp (UIA) and with the Flemish Fund for Scientific Research (F.W.O). Thanks are due to R.-J. Beun for providing the speech materials, to E. Blaauw for transcribing them in terms of prosodic phrases, and to R. Collier and A. A. Sanderman for commenting upon an earlier version of this paper.

[1]The fundamental frequency (F_0) is taken as the acoustic correlate of pitch and was determined by the method of subharmonic summation (Hermes, 1986).

[2]Note that these percentages, given our general linear model fitting procedure, do not have to add up to the value of the model with the prosodic parameters combined.

Beckman, M. E., and Ayers, G. (**1994**). ''Guidelines for ToBI labelling. Manuscript and accompanying speech materials,'' Ohio State University. [Obtain by writing to tobi@ling.ohio-state.edu]

Beun, R. J. (**1991**). Transcripties spontaan gesproken monologen [Transcriptions of spontaneously produced monologues], IPO report 792.

Brown, G., and Yule, G. (**1983**). *Discourse Analysis* (Cambridge U.P., Cambridge).

Brown, G., Currie, K., and Kenworthy, J. (**1980**). *Questions of Intonation* (Croom Helm, London).

Brubaker, R. S. (**1972**). ''Rate and pause characteristics of oral reading,'' J. Psycholinguist. Res. **1**, 141–147.

Bruce, G. (**1982**). ''Textual aspects of prosody in Swedish,'' Phonetica **39**, 274–287.

Carletta, J. (**1996**). ''Assessing agreement on classification tasks: the kappa statistic,'' Computational Linguist. **22** (2), 249–254.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Newlands, A., Doherty-Sneddon, G., and Anderson, A. (**1996a**). ''HCRC Dialogue Structure Coding Manual,'' HCRC/TR-82, Human Communication Research Centre, University of Edinburgh.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Newlands, A., Doherty-Sneddon, G., and Anderson, A. (**1996b**). ''Dialogue Structure Coding and Its Uses in the Map Task'' (in preparation).

Chafe, W. L. (**1980**). *Pear Stories: Cognitive, Cultural and Linguistics Aspects of Narrative Production* (Ablex, Norwood, NJ).

Condon, S. L., and Čech, C. G. (**1996**). ''Discourse Coding Reliability: Problems and Possibilities'' (in preparation).

Geluykens, R. (**in press**). *The Pragmatics of Discourse Anaphora in English: Evidence from Conversational Repair* (Mouton de Gruyter, Berlin).

Grosz, B., and Hirschberg, J. (**1992**). ''Some intonational characteristics of discourse structure,'' Proc. International Conference on Spoken Language Processing, Banff, Canada, October 1992, pp. 492–432.

Grosz, B., and Sidner, C. L. (**1986**). ''Attentions, intentions, and the structure of discourse,'' Computational Linguist. **85**, 363–394.

Hermes, D. J. (**1986**). ''Measurement of pitch by subharmonic summation,''J. Acoust. Soc. Am. **83,** 257–264.

Hirschberg, J., and Grosz, B. (**1992**). ''Intonational Features of Local and Global Discourse Structure,'' *Proc. of the Speech and Natural Language Workshop* (DARPA, Harriman, NY), pp. 441–446.

Hirschberg, J., Nakatani, C., and Grosz, B. (**1995**). ''Conveying discourse structure through intonation variation,'' Proc. ESCA Workshop on Spoken Dialogue Systems: Theories and Applications, pp. 189–192, Vigsø, Denmark, May/June 1995.

Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., and Quantz, J. J. (**1995**). ''Dialogue Acts in VERBMOBIL,'' Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, Verbmobil Report 65.

van Kuppevelt, J. (**1995**). ''Discourse structure, topicality and questioning,'' J. Linguist. **31**, 109–147.

Lehiste, I. (**1975**). ''The phonetic structure of paragraphs,'' in *Structure and Process in Speech Perception*, edited by A. Cohen and S. Nooteboom (Springer-Verlag, Berlin), pp. 195–206.

Lehiste, I. (**1979**). ''Perception of sentence and paragraph boundaries,'' in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, London), pp. 191–201.

Mandler, J. M., and Johnson, N. S. (**1977**). ''Remembrance of Things Parsed: Story Structure and Recall,'' Cognitive Psychol. **9**, 111–151.

Mann, W., and Thompson, S. (**1986**). ''Relational Propositions in Discourse,'' Discourse Processes **9**, 57–90.

Menn, L., and Boyce, S. (**1982**). ''Fundamental frequency and discourse structure,'' Language Speech **25**, 341–383.

Nakatani, C., Grosz, B., and Hirschberg, J. (**1995**). ''Discourse structure in spoken language: Studies on speech corpora,'' Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation. Stanford, March.

Nakatani, C., Grosz, B. J., Ahn, D. D., and Hirschberg, J. (**1995**). ''Instructions for Annotating Discourses,'' Technical Report Number TR-21-95. Center for Research in Computing Technology (Harvard University, Cambridge, MA).

Passonneau, R. J., and Litman, D. J. (**1993**). ''Intention-based segmentation: human reliability and correlation with linguistic cues,'' Proc. ACL-93, Ohio State University: Association for Computational Linguistics.

de Pijper, J. R., and Sanderman, A. A. (**1994**). ''On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues,'' J. Acoust. Soc. Am. **96,** 2037–2047.

Reichman, R. (**1985**). *Getting Computers to Talk like You and Me: Discourse Context, Focus, and Semantics* (Bradford, Cambridge).

Reithlinger, N. (**1995**). ''Some experiments in Speech Act Prediction,'' Proc. AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Stanford, March.

Rotondo, J. A. (**1984**). ''Clustering analyses of subjective partitions of text,'' Discourse Processes **7**, 69–88.

Shiffrin, D. (**1987**). *Discourse Markers* (Cambridge U.P., Cambridge).

Sluijter, A., and Terken, J. (**1994**). ''Beyond sentence prosody: Paragraph intonation in Dutch,'' Phonetica **50**, 180–188.

Strangert, E. (**1993**). ''Speaking style and pausing,'' PHONUM **2**, 121–137.

Swerts, M., and Collier, R. (**1992**). ''On the controlled elicitation of spontaneous speech,'' Speech Commun. **11**, 463–468.

Swerts, M., and Geluykens, R. (**1994**). ''Prosody as a marker of information flow in spoken discourse,'' Language Speech **37**, 21–43.

Swerts, M., Collier, R., and Terken, J. (**1994**). ''Prosodic predictors of discourse finality in spontaneous monologues,'' Speech Commun. **15**, 79–90.

Swerts, M., Bouwhuis, D. G., and Collier, R. (**1994**). ''Melodic cues to the perceived ''finality'' of utterances,'' J. Acoust. Soc. Am. **96**, 2064–2075.

Swerts, M., Strangert, E., and Heldner, M. (**1996**). ''F_0 declination in spontaneous and read-aloud speech,'' Proc. International Conference on Spoken Language Processing, Philadelphia, October 1996, pp. 1501–1504.

Swerts, M., Wichmann, A., and Beun, R.-J. (**1996**). ''Filled pauses as markers of discourse structure,'' Proc. International Conference on Spoken Language Processing, Philadelphia, October 1996, pp. 1033–1036

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

Thorndyke, P. W. (**1977**). ''Cognitive Structures in Comprehension and Memory of Narrative Discourse,'' Cognitive Psychol. **9**, 77–110.

Thorsen, N. G. (**1985**). ''Intonation and text in Standard Danish,'' J. Acoust. Soc. Am. **77**, 1205–1216.

Thorsen, N. G. (**1986**). ''Sentence intonation in textual context—supplementary data,'' J. Acoust. Soc. Am. **80**, 1041–1047.

Venditti, J., and Swerts, M. (**1996**). ''Prosodic cues to discourse structure in Japanese,'' Proc. International Conference on Spoken Language Processing, Philadelphia, October 1996, pp. 725–728.

Webber, B. (**1988**). ''Discourse deixis: Reference to discourse segments,'' Proc. of the 26th Annual Meeting, pp. 113–122, Buffalo, Association for Computational Linguistics.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (**1992**). ''Segmental durations in the vicinity of prosodic phrase boundaries,'' J. Acoust. Soc. Am. **91**, 1707–1717.