Local and global prosodic cues to discourse organization in dialogues

Ronald Geluykens and Marc Swerts*
Institute for Perception Research (IPO)
P.O. Box 513, NL-5600 MB Eindhoven

ABSTRACT
It is experimentally investigated to what extent speakers use prosody to signal topic- and turn-boundaries non-ambiguously to a listener. Results show that local and global melodic features are employed to structure both information and interaction.

INTRODUCTION
This paper reports on an investigation into prosodic cues to dialogue structure. In monologue, prosody often signals how the discourse is structured in terms of topical organization (see Swerts & Geluykens, in press). In dialogue, prosody also plays a role in the turn-taking mechanism (Sacks et al 1984). In this paper, we attempt to find out experimentally to what extent these two dimensions, information and interaction, interfere with one another, and how speakers employ their prosodic resources to regulate both dimensions in such a way that they are signalled non-ambiguously to the listener.

Research to date has generally been limited to uncontrolled, spontaneous speech (Brown et al. 1980; Schaffer 1983); to our knowledge, no work exists which tries to investigate the two above-mentioned dimensions independently in a controlled manner. Previous research has also tended to concentrate on local pitch cues (especially falling vs. rising pitch); non-local cues remain virtually uninvestigated. Since more global cues appear to play a role in monologue (Swerts & Geluykens, in press), it seems reasonable to assume that they are also relevant in dialogue discourse.

ACOUSTIC STUDY
Experimental set-up
A series of experiments was set up employing strings of differently coloured geometrical figures (see also Swerts & Collier 1992). Each time two subjects (from a total of ten) were seated in a sound-proof studio, without visual contact, and had to perform three experiments (see also Table 1).

In the first condition (C1; monologue), the speaker had to describe from left to right strings of geometrical figures (as in figure 1), in such a way that the 'topical breaks' between the individual strings became apparent for the hearer; the latter's task was to indicate the perceived breaks on an answer sheet. In condition C2 (dialogue), both subjects acted as speakers and had to produce strings without any internal breaks, and signal to the other participant when their turn was finished; the other speaker then had to take over the floor as soon as s/he felt it was possible. By means of experiments C1 and C2, we wanted to elicit both information ('topic') and interaction ('turn') signals in their purest form.

Figure 1. *Example of a series with strings of geometrical figures (different shadings actually correspond to different colours)*

In condition C3, the two tasks were combined. Speakers had to produce strings of figures (as in C1) and make the topical breaks apparent to the listener; at the same time, they had to indicate (as in C2) when their description was finished, so that the dialogue

partner could take over. Listeners were instructed (i) to transcribe the topic boundaries within each speaker-turn, and (ii) to take over the floor at the appropriate turn boundary. In this way, topic-finality and turn-finality were varied to some extent independently (since topic-finality did not necessarily imply turn-finality).

Table 1. *Overview of experimental set-up for the 3 conditions (C1, C2, C3)*

	speech mode	speaker instruction	hearer instruction
C1	monologue	signal series breaks	transcribe breaks
C2	dialogue (simple)	signal end of turn	take over floor
C3	dialogue (complex)	{signal series breaks	trancribe breaks
		{signal end of turn	take over floor

Auditory analysis of pitch movements

To begin with, we analyzed auditorily the pitch contours in the different discourse locations of the elicited speech, i.e. string-internally, string-finally, and series-finally. It appeared that the contours could most easily be distinguished into those ending in High (H), Mid (M) or Low (L) level of a speaker's register. Results for the distribution of these contours are in Table 2 (representing which of the contours each of the ten speakers used in the majority of cases in the different discourse locations; see also Geluykens and Swerts 1992).

Table 2. *Pitch contours (H, M, L) in various discourse positions*

end of string		-	+	+			-	+				-	+	+
end of series		-	-	+			-	+				-	-	+
C1	M	10	0	0	C2	M	10	0	C3	M	10	0	0	
	H	0	5	0		H	0	2		H	0	9	0	
	L	0	5	10		L	0	8		L	0	1	10	

Table 2 reveals that, in the three conditions, the end of a series is always marked by means of a Low-contour (in C2 and C3 end of series coinciding with a shift in speaking turn) and descriptions of string-internal figures are always provided with an Mid-contour. However, if we look at the string-final (but not series-final) contours, it appears that there is an even distribution of High or Low in C1, but a clear preference for High in C3. Apparently, in the latter condition, a speaker knows that he risks to be interrupted by his partner if he uses a Low.

Table 3. *Mean end-frequencies of different pitch contours [+Sd]*

	monologue	dialogue (simple)	dialogue (complex)
M [total]	+0.08	+0.11	+0.05
M (string ends in H]	–1.71 [2.91]	–2.85 [0.88]	–1.18 [2.41]
M (string ends in L]	+1.56 [2.38]	+1.38 [1.84]	+1.49 [2.04]
H	+7.40 [2.30]	+10.20 [1.17]	+7.92 [2.86]
L	–5.17 [1.56]	–5.56 [1.62]	–4.57 [1.99]

Acoustic measurements

To give acoustic support to our auditory transcriptions, we determined instrumentally the end frequency of each pitch contour. To make comparisons across speakers more easy, the average distance (in semitones) between this end frequency and the speaker's average pitch was calculated. Results are presented in Table 3.

Results show that, in all three conditions, average end-frequencies of string-internal contours (row 1) are very close to average frequencies, and very far removed both from

end-frequencies of L's (row 5) and H's (row 4). This makes end-frequency a reliable indicator of discourse position. Secondly, if one compares internal end-frequencies which occur in strings ending in H (row 2) with those occurring in strings ending in L (row 3), an interesting picture emerges, in that the former have a lower average end-frequency than the latter; this is true in all three conditions. In other words, internal tones are maximally different from the end-tone while staying near the middle of the speaker's pitch range. This has important repercussions, since it would appear that the final position of a figure is, as it were, pre-signalled in the end-frequencies of the prefinal figures. Since end-pitch (fall or rise) is an important cue to discourse location, especially in condition 3, what appears to be the case is a non-local way of signalling whether the ongoing series is going to be turn-final or not. In the following section, we will investigate, among other things, whether this pre-signalling has perceptual cue value for the listener.

PERCEPTUAL EVALUATION

Experimental set-up

In order to evaluate the perceptual cue value of the acoustic characteristics (local and global) discussed in the previous section, we conducted a perception experiment using the speech produced in condition 3 (which was central to our investigation) as input.

Stimuli were prepared in the following manner. From four speakers, we selected strings of 2 up to 5 figures occurring in different discourse positions, viz. turn-initial, turn-medial, and turn-final. We then employed these as stimuli in various formats. First of all, strings were presented in their entirety, including the final pitch contour. Secondly, we chopped off parts of the utterance, starting with the description of the last figure, and continued to do this until some stimuli had only one geometrical figure (the first one) left. Ten test subjects (students and staff at IPO) were then asked to listen to all stimuli, presented in random order, and indicate on a score sheet whether they thought a particular stimulus occurred initially, medially, or finally in a series of at least three of such strings. Subjects could listen more than once to each stimulus.

Results and Discussion

Looking at the results for complete strings, presented in Table 4, it appears that listeners' scores are significantly higher than chance (X^2=217.391, p<.001). Moreover, if one examines scores in more detail, this significance appears to be mainly due to their almost perfect perception of finality versus non-finality. We have therefore reinterpreted the results, conflating the two non-final judgments into one 'non-final' category, as represented in Table 5.

Table 4. *Perceptual evaluation of discourse position within series (complete strings)*

perceived as:	initial	medial	final
initial	**47**	31	2
medial	29	**49**	2
final	4	0	**76**

The differences visualized in table 5 turn out to be highly significant (X^2=205.009, p<.001). In other words, listeners are able to use prosody as a perceptual cue for finality in dialogues. It remains to be seen, however, what it is precisely that causes this result. Do listeners respond solely to the nature of the final pitch contour, or is there something more global, such as relative end-frequencies of internal pitch movements, which they take into account? Results for the incomplete strings, which were constructed to test precisely this potential non-local cue, should give us an answer to this question. Since the relevant distinction appears to be 'final' versus 'non-final' (see above), we have once again conflated the results into these two categories (Table 6).

Table 5. *Perceptual evaluation of finality vs. non-finality (full strings)*

perceived as:	non-final	final	total
non-final	**156**	4	160
final	4	**76**	80
total	160	80	240

Table 6 shows that, even in the absence of final pitch cues, listeners are able to perceive finality of strings well above chance level (X^2=64.496, p<.001). Given the fact that most of these strings were reduced more extensively than by just deleting the final figure, with for some only the first figure remaining, this shows rather convincingly that listeners are able to make perceptual use of non-local prosodic characteristics of the speech signal (though performance does drop compared to full strings, as one would expect, the latter also containing local finality cues).

Table 6. *Perceptual evaluation of finality vs. non-finality (incomplete strings)*

perceived as:	non-final	final	total
non-final	**425**	135	560
final	135	**145**	280
total	560	280	840

CONCLUSION

This limited experimental study has yielded a number of results concerning prosody in dialogue. First of all, it was shown that speakers are able to use prosody in such a way that both the informational and interactional dimension is signalled in a non-ambiguous manner. Speakers make use of local pitch contours to do so, but also appear to pre-signal the type of final pitch movement by means of the end-frequencies of the non-final pitch movements.

In a perception test, the perceptual relevance of both local and non-local cues was investigated. It appears that both local and global prosodic features play a role in the perception of finality: even in the absence of final pitch contours, listeners are still able to predict finality to some extent. This may be due to non-final end-frequencies, but other factors, such as declination and accent distibution, may also play a role (see Geluykens & Swerts 1992).

*both authors are also affiliated with the University of Antwerp (UFSIA and UIA, respectively) and with the Belgian National Science Foundation (NFWO). René Collier is thanked for useful comments on an earlier version of this paper.

REFERENCES

Brown, G., K. Currie & J. Kenworthy (1980)*Questions of intonation.* (London: Croom Helm).

Geluykens, R. & M. Swerts (1992) Prosodic topic- and turn-finality cues. *Proceedings of the Workshop on Prosody in Natural Speech Data,* University of Pennsylvania, August 1992.

Sacks, H., E.A. Schegloff and G. Jefferson (1984) A simplest systematics for the organization of turn taking in conversation. *Language,* Vol. 50, pp. 696-735.

Schaffer, D. (1983) The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics,* Vol. 11, pp. 243-344.

Swerts, M. & R. Collier (1992) On the controlled elicitation of spontaneous speech. *Speech Communication,* Vol. 11, pp. 463-468.

Swerts, M. & R. Geluykens (in press) The prosody of information units in spontaneous monologue. To appear in *Phonetica.*