PHONETIC RULES FOR A PHONETIC-TO-SPEECH SYSTEM

A.A. Sanderman and R. Collier***
**KPN Research P.O Box 421, 2260 AK Leidschendam, The Netherlands*
E-mail: A.A.Sanderman@research.kpn.com
***Institute for Perception Research, P.O. Box 513, 5600 MB, Eindhoven, The Netherlands*
E-mail: collier@ipo.tue.nl

ABSTRACT

In our previous research we investigated the demarcative function of prosody at the sentence level and the importance of this information for listeners in terms of perception, acceptability and ease of comprehension. In this paper we investigate if the results obtained for utterances spoken in isolation can be generalised to utterances spoken in context.

1. INTRODUCTION

From previous research (de Pijper & Sanderman, 1994; Sanderman & Collier, 1995) we know that speakers use variable combinations of prosodic cues to mark systematically boundaries of different strength in (complex) sentences. Therefore, a speaker primarily adapts the duration of pauses and secondarily chooses a particular type of melodic boundary marker and/or declination reset.

The relevance of the prosodic features pause, melodic boundary marker, declination reset and preboundary lengthening for listeners was shown in several ways:

1. The prosodic features used by a speaker are perceptually relevant, which is evident from the fact that listeners systematically assign different Perceptual Boundary Strength (PBS) values to word boundaries differing in these prosodic characteristics. Untrained listeners were well able to score the perceived boundary strength on a 10-point scale. The agreement between listeners was very high and they were not biased by syntactic and/or semantic information (de Pijper & Sanderman, 1994; Sanderman & Collier, 1995).
2. The implementation of the prosodic features in synthesised utterances resulted in a strong improvement of the perceived quality of the synthetic speech, which was determined on the basis of the acceptability scores of listeners (Sanderman & Colllier, 1996).
3. Properly phrased utterances facilitate the comprehension process. It appeared that listeners were able to answer a question faster when the utterance that contained the answer was well-phrased (Sanderman & Collier, 1997).

However, the research described above concentrated on utterances spoken in isolation. Since in natural speech and in dialogue systems most of the sentences occur in larger discourse units, we want to know how speakers use prosodic features to highlight the structural make-up of sentences occurring in different positions (initial, medial, final) in a context. It is not unthinkable that a speaker uses different boundary-marking strategies. Brubaker (1972) found that utterances near the end of a paragraph were read out faster than those occurring earlier, and Caspers (1994) reported that in fast speech some melodic boundary markers are deleted and the number of different pitch configurations is reduced. Beside this, we want to know if the relationship between PBS and prosodic characteristics remains the same as in isolated utterances. It is possible that the speaker uses the same prosodic characteristics in marking boundaries, but that the perception of listeners changes due to their interaction with variable speech rate.

2. METHOD

To give an answer to these questions several texts and isolated utterances were spoken by a professional speaker. Every text contains one target sentence, which is placed in initial, medial and final position of the text (IniText, MedText, FinText) and in final position of the first paragraph (FinPar). The isolated utterances match the four types of context utterances (with respect to the written form):
IniIso = 16 isolated utterances that are identical to the 16 IniText utterances,
MedIso = 16 isolated utterances that are identical to the 16 MedText utterances,
FinParIso = 16 isolated utterances that are identical to the 16 FinPar utterances,
FinIso = 16 isolated utterances that are identical to the 16 FinText utterances.

The isolated utterances and the utterances coming from different positions in the text were presented to 17 untrained listeners in 4 sessions. They were asked to express the PBS of the word boundaries of the utterances on a 10-point scale. The whole procedure was carried out in the same way as in de Pijper & Sanderman (1994) and Sanderman & Collier (1995).

The pause durations were measured and the pitch contour types and the declination resets were determined (for the exact procedure see de Pijper & Sanderman (1994) and Sanderman & Collier (1995)). Also, the speech ratte was determined for the text utterances and the isolated utterances. The speech rate is the total duration of the sentences, without the pause duration, divided by the number of syllables.

Table I: The speech rate in syllables per second and the s.d. for the utterances spoken in context (IniText, MedText, FinPar, FinText) and the matched utterances spoken in isolation (IniIso, MedIso, FinParIso, FinIso)

	syll/s	s.d.
IniIso	5.08	0.31
IniText	5.20	0.31
MedIso	4.99	0.26
MedText	5.51	0.25
FinParIso	5.13	0.31
FinPar	5.54	0.28
FinIso	5.18	0.42
FinText	5.58	0.39

3. RESULTS

The speech rate of the utterances read in isolation and in context are given in Table I.
From analyses of variance we know that the mean rate in syllables per second for IniIso and IniText do not differ significantly ($F_{(1,30)}$ = 1.20, p > .05). Whereas the rates of MedIso and MedText differ significantly ($F_{(1,30)}$ = 33.10, p < .0001), and also for FinParIso-FinPar ($F_{(1,30)}$ = 15.42, p < .001) and FinIso-FinText ($F_{(1,30)}$ = 7.81, p < .01). From this we can conclude that non-initial utterances were spoken faster than initial ones.

In Table II the number of observations for the different types of melodic markers used are given for the context and the matched isolated utterances.
The likelihood ratio chi-square (I-test of Spitz (1961) reveals no significant difference for IniText-IniIso (*I* = 11.69, p = .04, df = 5), FinPar-FinParIso (*I* = 6.80, p > .05, df = 5) and FinText-FinIso (*I* = 1.38, p > .05, df = 5), but does show one for MedText-MedIso (*I* = 16.78, p < .01, df = 5). This significant effect can be attributed mainly to the fact that the speaker uses the contour **1E** more often in the isolated utterances (MedIso). This tendency can also be observed in FinPar-FinParIso and FinText-FinIso.

Table III is similar to Table II, except that now the different categories of pauses are given. Again we do not observe much difference in the numbers when we compare the isolated versions with the context versions. This is confirmed by the results of the *I*-test of Spitz: IniText-IniIso (I = 2.53, p > .05, df = 4), MedText-MedIso (I = 1.43, p > .05, df = 5), FinPar-FinParIso (I = 4.36, p > .05, df = 5) and FinText-FinIso (I = 1.94, p > .05, df = 4).

Table II: Number of observations for the different types of melodic markers for the utterances spoken in context (IniText, MedText, FinPar, FinText) and the matched utterances spoken in isolation (IniIso, MedIso, FinParIso, FinIso).
* Melodic markers in terms of the grammar of Dutch intonation as described in 't Hart, Collier and Cohen (1990).

	Melodic markers*					
	0	**1∅**	**1E**	**1E2**	**1A2**	**12**
IniIso	30	2	12	3	10	6
IniText	28	4	11	0	19	1
MedIso	17	3	12	5	12	2
MedText	22	4	3	0	15	7
FinParIso	23	5	11	2	11	4
FinPar	25	1	6	1	15	8
FinIso	20	5	8	3	12	6
FinText	23	5	5	3	10	8

Table III: Number of observations for the six pause categories in ms for the utterances spoken in context (IniText, MedText, FinPar, FinText) and the matched utterances spoken in isolation (Inilso, Medlso, FinParlso, Finlso).

Melodic markers*						
	0	<100	<200	<300	<400	>399
Inilso	42	12	3	4	2	0
IniText	47	10	4	1	2	0
Medlso	31	12	2	2	4	0
MedText	31	11	2	2	4	1
FinParlso	34	13	2	3	4	0
FinPar	36	12	0	4	3	1
Finlso	32	16	2	2	2	0
FinText	30	13	4	3	4	0

Also, analyses of the observed frequencies for the resets showed that our professional speaker does not use more or fewer resets or lowerings in utterances spoken in isolation than he does for those in context (see Table IV).
From the listening experiment it appeared that the difference in PBS values in isolated utterances and in utterances in context is very small when the speaker uses the same combinations of cues. In Figure 10 is given the correlation between the PBS values for the two sorts of utterances.

Table IV: Number of observations for the types of reset for the utterances spoken in context (IniText, MedText, FinPar, FinText) and the matched utterances spoken in isolation (Inilso, Medlso, FinParlso, Finlso).

	No reset	Reset	Lowering
Inilso	58	5	0
IniText	56	7	0
Medlso	45	6	0
MedText	43	6	2
FinParlso	50	6	0
FinPar	44	10	2
Finlso	48	6	0
FinText	46	6	2

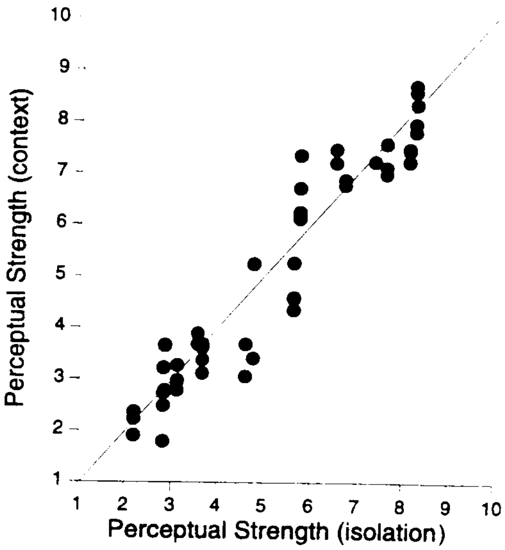

Figure 1: Scattergram of perceptual boundary strength values for utterances spoken in isolation and in context for the professional speaker

4. DISCUSSION AND CONCLUSION

From the results of this experiment it can be concluded that the professional speaker uses the same prosodic characteristics to mark boundaries in utterances spoken in context and in isolation. In general, the same types of melodic markers are used, except that there is a tendency for the pitch contour **1E** to be used more often in isolated utterances. Furthermore, the number and the duration of the pauses do not differ, nor does the use of resets. Also, the same combinations of prosodic characteristics do not result in different judgements of the PBS by the listeners.

In the introduction of this experiment we suggested that the behaviour of the speaker and the listeners could be different for utterances in context and in isolation, as a result of the differences in speaking rate between these types of utterances. The results show no difference in the behaviour, although there are clear differences in speaking rate: text-initial utterances were spoken as fast as utterances spoken out in isolation, whereas the non-initial ones were spoken faster. However, these differences in speaking rate are not nearly as large as the ones between normal and fast speech reported by Caspers (1994). This can be the reason that we observed the same speaker and listener behaviour in isolated and context-embedded utterances.

Since the results obtained for utterance spoken in isolation can be generalised to utterances spoken in context, we can also conclude that rules as developed for utterances spoken in isolation (Sanderman & Collier, 1996) can also be used for utterances spoken in context.

5. REFERENCES

't Hart, J., Collier, R. and Cohen, A. (1990). *A perceptual study of intonation: An experimental phonetic approach to speech melody* (Cambridge University Press).

De Pijper, J.R. and Sanderman, A.A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues, *Journal of the Acoustical Society of America, 96 (4)*, 2037-2047.

Sanderman, A.A. and Collier, R. (1995). Prosodic phrasing at the sentence level, in: *Producing Speech: Contemporary issues for Katherine Safford Harris,* edited by F. Bell-Berti and I.J. Raphael (American Institute for Physics, New York), p. 321-332.

Sanderman, A.A. and Collier, R. (1996). Prosodic rules for the implementation of phrase boundaries in synthetic speech*, Journal of the Acoustical Society of America, 100,* 3390-3397.

Sanderman, A.A. and Collier, R. (1997). Prosodic phrasing and comprehension. *Submitted to Language and Speech.*