Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials

Rolf Carlson (1) and Marc Swerts (2)*

(1) CTT, KTH, Sweden

(2) University of Tilburg, The Netherlands

and Universitaire Instelling Antwerpen, Belgium

rolf@speech.kth.se, m.g.j.swerts@uvt.nl

*names in alphabetic order

ABSTRACT

The current paper reports on a study of perceptually based predictions of upcoming prosodic breaks in spontaneous Swedish speech materials. The question tackled here is to what extent listeners are able, on the basis of prosodic features, to predict the occurrence of upcoming boundaries, and if so, whether they are able to differentiate different degrees of boundary strength. To answer these questions, an experiment is conducted in which spontaneous utterance fragments (both long and short versions) are presented to listeners, who are instructed to guess whether or not the fragments are followed by a prosodic break, and if so, what the strength of the break is. Results reveal that listeners are indeed able to predict whether or not a boundary (of a particular strength) is following the fragment.

1. INTRODUCTION

In addition to studies that explore how speakers use prosodic features to demarcate speech units of various sizes (e.g. from the phrase to a complete paragraph), it has been investigated to what extent such prosodic chunking is perceptually relevant for listeners. There is evidence that a listener's processing of incoming speech is indeed facilitated or influenced by the occurrence of prosodic boundary cues (Ostendorf et al. 1990). Similarly, it has been shown that the perceived naturalness of synthetic speech improves if prosodic boundaries are properly generated in the speech output (e.g. Sanderman, 1996). It appears that listeners are not only sensitive to the absence or presence of a boundary, but that it also matters how "strong" the boundary is when it occurs. For instance, a few phonetic studies that focused on the exact nature of the prosodic cues that lead to the perception of a break, consisted of experiments in which listeners were asked to rate the prosodic boundary strength between two words on a given scale (e.g. Dutch: Sanderman, 1996; Swedish: Strangert, and Heldner, 1995; Fant et al. 2000; Hansson, 2002). The results of these studies reveal that perceived boundary strength is heavily dependent on the occurrence of a silent pause, even to the extent that it may overrule the contribution of other parameters such as preboundary lengthening, boundary tone and pitch reset.

The current paper also describes a listener-oriented approach to prosodic boundaries, yet differs from various previous studies in that it will look at possible predictors of such boundaries. This is motivated by the underlying assumption that speakers not only encode prosodic breaks locally at the places where they occur (e.g. in the form of silent pauses), but that they also pre-signal these breaks in advance. This could enable listeners to perceive an upcoming break some time before its actual occurrence. If such predictors[1] indeed exist, this may decrease a listener's cognitive processing load, as they provide an early indication as to which elements in the flow of speech ought to be processed as a whole. We know from previous work on prosody modeling that there are indeed (phonetic) features which presignal upcoming breaks (Grosjean, 1983; Leroy, 1984; Swerts et al., 1994; Klatt, 1979; Ferrer et al. 2002). However, most of these early studies are limited to read-aloud or specifically elicited speech materials, and they do not always clarify how such prosodic predictors relate to potential other linguistic factors which may contain important cues for upcoming breaks (such as syntax) (see Gee and Grosjean, 1983). In as far as research on the perception of spontaneous speech data is concerned, some efforts been done to describe how disfluencies can be predicted (Lickley et al., 1999; Baron et al. 2002).

The current paper reports on a study of perceptually based predictions of upcoming prosodic breaks in spontaneous Swedish speech materials. Questions to be addressed are: Are listeners able to predict the occurrence of upcoming prosodic boundaries? If so, are they able to differentiate different degrees of boundary strength? If so, to what extent is this ability to predict these boundaries based on purely prosodic features? As will become clear in the next section, we have conducted a variant of the gating paradigm, basically an experiment in which spontaneous utterance fragments are presented to listeners, who are instructed to guess whether or not the fragments are followed by a break, and if so what its strength is.

[1] Note that our use of the term "predictor" is somewhat different from the way it is being used in studies that "predict" prosodic boundaries in offline tasks, e.g. for speech synthesis.

2. EXPERIMENT

2.1 Stimuli

All data were taken from a corpus collected within the Swedish GROG project "Boundaries and groupings - the structuring of speech in different communicative situations". The objective of this project is to model the structuring of Swedish speech in terms of prosodic breaks and groupings (Carlson et al., 2002). From this corpus, we selected one recorded interview between a reporter and an in Sweden well known female politician (GS) that was originally broadcast on public Swedish Radio. The interview lasted about half an hour long. We only used speech data produced by the person who was being interviewed. Using a perceptually based protocol for prosodic annotation, the entire interview was prosodically labeled by three independent researchers in the project, who did not take part of the current experiment. The data analysis is discussed in a separate contribution to this conference (Heldner and Megyesi, 2003).

From these materials, we first selected 60 utterance fragments of approximately 2 seconds long. The exact initial cutting point was moved to the nearest word boundary, whereas the final cutting point was fixed. That is, the fragments all preceded the word "och" (and) in their original context, and the fragments were cut right before the silent interval (if any) before that word. The choice to use the word "och" was partly syntactically motivated, given that the fragments then all occurred in comparable syntactic positions before an identical conjunction. In addition, the glottal onset of the first vowel of "och" facilitated cutting the fragment before it in cases where there was no real pause. Also, possible coarticulatory effects are minimized, compared to a situation where always different words would have followed. The conjunction "och" is usually unstressed and mostly realized with a schwa vowel. The fragments differed regarding the presence or absence of a break in between the end of the fragment and the word "och", i.e., as annotated by our independent labelers by a majority voting procedure: in about one third of the cases, there was a strong intervening break, one third of the fragments preceded a weak break and one third was followed by no break at all. From these longer fragments, we then constructed short versions consisting of only the final word of the fragment, leading to 120 stimuli in total.

2.2. Subjects

13 students in logopedics from Umeå university participated as listeners in the experiment on a voluntary basis. They got a movie ticket as an acknowledgement for their participation.

2.3 Experimental procedure

The 120 different stimuli (long and short versions, preceding a strong, weak or no boundary) were mixed and presented sequentially to our listeners via a specifically designed interface, which allows to run perception experiments through the internet using a standard web browser with audio facilities.[2] To minimize possible learning effects, each subject was presented with a differently randomized list of stimuli. Their task was to rate, for each stimulus, on a 5-point scale whether they felt that the fragment preceded no boundary (1), a strong boundary (5), or a boundary having a strength in between these two extremes (2-4). The actual test was preceded by a short introduction which briefly explained a few concepts (such as prosodic boundary) and the actual task. Subjects were also informed that they always had to give an answer, even if they were unsure about their response. No feedback was given on the "correctness" of their responses, and there was no interaction with the experimentors. During the test, subjects could listen as many times as needed to a given stimulus before giving an answer, but they could not return to a previous stimulus after a response had been entered. The experiment was self-paced, and lasted approximately 20 minutes on average.

3. RESULTS

In Table 1 and Figure 1 the results from the perceptually based prediction experiment are presented. The data for the 13 subjects have been grouped according to the labeling by three expert annotators using a majority vote and also according to stimulus fragment size. The overall mean varied between 2,61 for word fragments to 2,97 for the 2 seconds fragments. A repeated-measures ANOVA with between-subjects factors of Boundary type (no boundary vs. weak boundary vs. strong boundary) and Fragment size (one word vs. 2 seconds) revealed significant main effects of Boundary type ($F(2,110)=77.6$; $p<.01$) as well as of Fragment size ($F(1,110)=7.8$; $p<.01$) on the perceived boundary strength. There was no significant interaction between Boundary type and Fragment size. A Games-Howell post hoc test showed that all three boundary types were significantly different from each other ($p<.01$).

Table 1: *Perceived upcoming boundary strength. Grouped according to labeled boundary strength and fragment size. Standard deviation in parenthesis.*

	fragment size	
	one word	2 seconds
no break n = 24 * 13	1,78 (0,96)	2,12 (1,16)
weak break n = 16* 13	2,77 (1,37)	3,07 (1,31)
strong break n = 18* 13	3,59 (1,23)	4,02 (1,09)

[2] http://www.let.uu.nl/~Theo.Veenker/personal/projects/wwstim/doc/en/

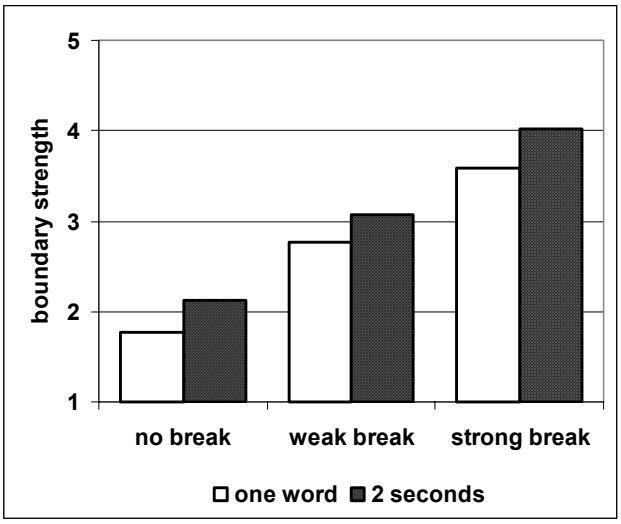

Figure 1. *Perceived upcoming boundary strength. The data is grouped according to labeled boundary strength and fragment size.*

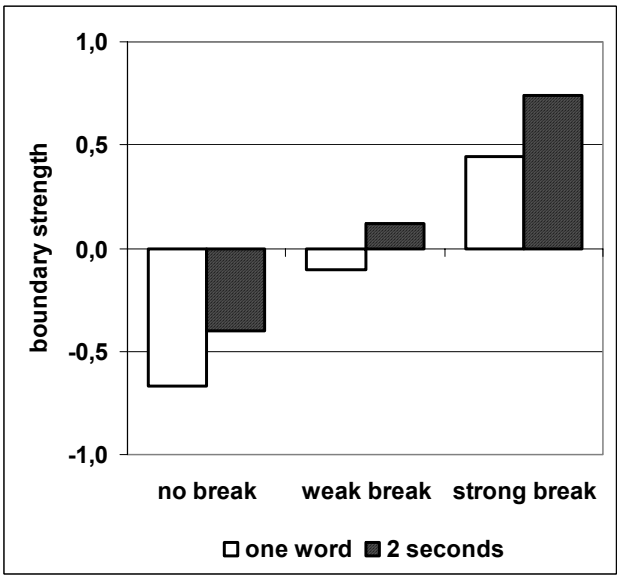

Figure 2. *Perceived upcoming boundary strength (Z-normalized). The data is grouped according to labeled boundary strength and fragment size*

In Figure 2 the data presented in Figure 1 have been Z-normalized using the mean and standard deviation for each subject as normalizing factors. (f(x)=(x-mean)/sd).

Since each word stimulus also can be found as part of a 2 seconds fragment it is possible to correlate the perceptually based prediction of upcoming prosodic breaks based on different sized context. Figure 3 shows that there is a significant correlation (r = ,89) between the two fragment sizes. Only a very weak correlation was found between isolated word duration and break strength prediction. However, exploring a more detailed analysis, using a duration model capturing final lengthening, shows as

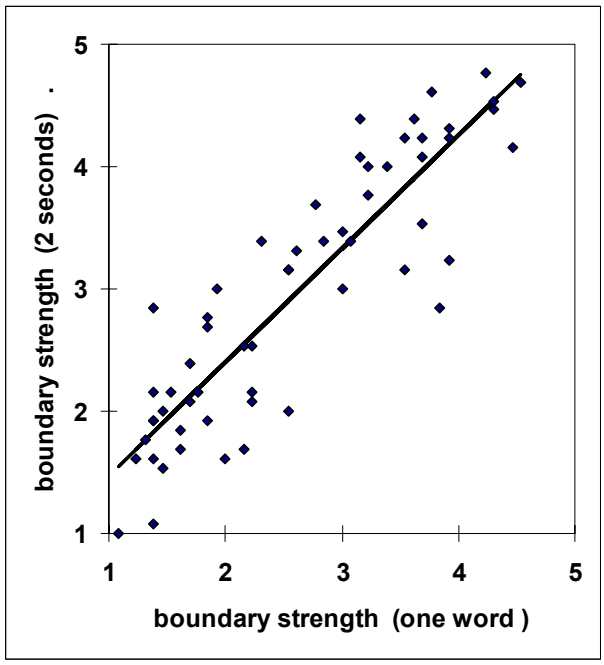

Figure 3. *Correlation between perceived upcoming boundary strength for each word in isolation and in a 2 seconds fragment. Regression coefficient r = 0,89.*

expected promising results for predicting the break strength (Heldner and Megyesi, 2003).

4. DISCUSSION AND CONCLUSION

As already mentioned in the introduction, previous research revealed that the perception of a boundary in the flow of speech is heavily influenced by features that occur at the boundary itself, such as a silent pause, or by features, such as pitch reset, for which you need to be able to have access to both the preceding and the following context of a given boundary. The results of our current study show that a listener is able to also predict a possible upcoming break, based on properties of the preceding context alone. One of the interesting findings is that the responses for the two types of stimuli, namely 2-sec fragments and 1-word stimuli, are not fundamentally different, as is clear from the high correlation between the two sets of responses. Yet, there is an overall difference between the responses in that the longer context has slightly higher, but significant, values for all three classes (no boundary, weak boundary, strong boundary). We speculate that this overall difference is due to the fact that the probability that a listener will predict an upcoming break increases if he or she has been exposed to a longer stretch of speech without a break. Still, the finding that the overall pattern for the two sets of stimuli is essentially the same implies that we cannot conclude that longer context leads to a higher amount of "correct" responses.

At first sight, this may seem a surprising outcome, as one might have expected that the task of guessing an upcoming boundary would be easier for 2-sec stimuli, given that for these stimuli, subjects literally have more speech materials at their disposal for making a decision. Therefore, one could have expected a flatter distribution in the responses for the short, 1-word stimuli. Our contrary finding suggests that the final word contains important prosodic and syntactic features that facilitate the prediction of upcoming breaks. As to the potential prosodic features, it is clear that some of the important boundary predictors may indeed be located in the final word, including features like type of boundary tone preceding the break, final lengthening, loudness patterns and possible effects of voice quality (e.g. the amount of creakiness), whereas other features, like rate of declination, are more characteristic for fragments that are longer than a single word. Similarly, the one word stimuli have some linguistic information in terms of parts-of-speech information which can be of value for the prediction. Some of the one word stimuli were reduced function words, while some content words carried focal stress. Further analysis of these features will shed some more light on this issue.

This leaves us with the question as to what the strength relationship is in cue value between the prosodic and syntactic features for predicting upcoming boundaries. We conjecture that the ability to predict prosodic boundaries is primarily based on acoustic cues, but is probably also supported by syntactic cues. Does this mean that the syntactic structure does not influence the prediction? Perhaps the acoustic features are so important that they are needed to support the decision and can not be over-ruled by a break prediction based only on syntactic features. On the other hand the syntactic structure probably has a predictive power on where a break is placed and acoustically realized. To gain further insight into the pure contribution of prosodic cues, we have planned to perfom future experiments using non-native speakers of Swedish as listeners. This will reduce the impact of syntactic cues.

ACKNOWLEDGMENTS

Marc Swerts is also affiliated with the Fund for Scientific Research – Flanders (FWO - Flanders). We would like to thank the Swedish Radio for making the broadcast material available for analysis; Mattias Heldner, Beata Megyesi and Eva Strangert for the prosodically labeling of the speech material; Theo Veenker for help with setting up the experimental environment; Mattias Heldner for the ANOVA analysis and all members of the GROG team for useful discussions and cooperation.

REFERENCES

[1] Baron, D., Shriberg, E., Stolcke, A. (2002) Automatic Punctuation And Disfluency Detection In Multi-Party Meetings Using Prosodic And Lexical Cues, ICSLP – 2002, September 16-20, 2002, Denver, Colorado USA

[2] Carlson R, Granström B, Heldner M, House D, Megyesi B, Strangert E, Swerts M (2002). Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. Proc of Fonetik 2002, TMH-QPSR, 44: 65-69

[3] Fant, G.; Kruckenberg, A.; Liljencrants, J., (2000.) Acoustic-phonetic Analysis of Prominence in Swedish. In Antonis Botinis (ed.), Intonation, Analysis, Modelling and Technology (Kluwer) 55-86

[4] Ferrer, L., Shriberg, E., Stolcke A. (2002), Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody, ICSLP – 2002, September 16-20, 2002, Denver, Colorado USA

[5] Gee, J. P., Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. Cognitive Psychology, 15, 411-458.

[6] Grosjean, F. (1983) How long is the sentence? Prediction and prosody in the on-line processing of language. Linguistics 21, 501-529.

[7] Hansson, P. (2002): Perceived boundary strength. Proc. ICSLP 2002, Denver.

[8] Heldner M, Megyesi B (2003) Exploring the prosody-syntax interface in conversations, Proc. ICPhS 03

[9] Klatt, D.H. (1979). Synthesis by rule of segmental durations in English sentences, in Frontiers of Speech Communication Research, (Ed. By Lindblom & Ohman,), Academic Press, London.

[10] Leroy, L. (1984) The psychology of fundamental frequency declination. Antwerp papers in linguistics 40, University of Antwerp.

[11] Lickley, R.J., McKelvie, D., Bard, E.G. (1999). Comparing human and automatic speech recognition using word gating. in Proceedings of the ICPhS Satellite meeting on Disfluency in Spontaneous Speech, UC Berkeley, pp. 23-26.

[12] Ostendorf, M., Price, P., Bear, J., Wightman,C.W. (1990), The use of relative duration in syntactic disambiguation, Proc. third DARPA Speech and Natural Language

[13] Sanderman, A. (1996). Prosodic phrasing. Production, perception, acceptability and comprehension. PhD thesis, Eindhoven University of Technology

[14] Strangert, E., Heldner, M. (1995) Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. In PHONUM 3, pp. 85-109. Umeå: Department of Phonetics, Umeå University.

[15] Swerts, M., Collier R., Terken, J. (1994). Prosodic predictors of discourse finality in spontaneous monologues. Speech Communication 15, 79-90.