ISCA Archive
http://www.isca-speech.org/archive

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Prediction of Upcoming Swedish Prosodic Boundaries by Swedish and American Listeners

*Rolf Carlson[1], Julia Hirschberg[2] and Marc Swerts[3]**

[1]KTH, Sweden, [2]Columbia University, USA and [3]University of Tilburg, The Netherlands and Universitaire Instelling Antwerpen, Belgium *Names in alphabetic order
rolf@speech.kth.se, julia@cs.columbia.edu and m.g.j.swerts@uvt.nl

Abstract

We describe results of a study of perceptually based predictions of upcoming prosodic breaks in spontaneous Swedish speech materials by native speakers of Swedish and of standard American English. The question addressed here is the extent to which listeners are able, on the basis of acoustic and prosodic features, to predict the occurrence of upcoming boundaries, and if so, whether they are able to distinguish different degrees of boundary strength. An experiment was conducted in which spontaneous utterance fragments (both long and short versions) were presented to listeners, who were instructed to guess whether or not the fragments were followed by a prosodic break, and if so, what the strength of the break was, where boundary presence and strength had been independently labeled. Results revealed that both listening groups were indeed able to predict whether or not a boundary (of a particular strength) followed the fragment, suggesting that prosodic rather than lexico-grammatical information was being used as a primary cue.

1. Introduction

Previous studies have shown that listeners are not only sensitive to the absence or presence of a boundary, but that the strength of the boundary is also important (e.g. Dutch: [10]; Swedish: [6], [11] and [8]). These studies found 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. In addition, we know from previous work on prosody modeling that there are other features such as F0 change, voice quality, and final lengthening which presignal upcoming breaks (e.g. [1], [7] and [12]). These studies are important in that they suggest how listeners may be able to process speech input in real time, while phrases are being produced.

Recently, Carlson and Swerts [5] described a study of listener perceptions of prosodic boundaries in spontaneous Swedish in which stimuli were presented for which pausal cues were unavailable. The specific hypothesis tested in that study was 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 signal these breaks in advance. The general result was that listeners were able to make boundary predictions with considerable accuracy, when compared with hand-labeled breaks.

In the current paper, we will expand upon that study and report on additional studies of non-Swedish speaking listener judgments of the same Swedish data. This new study was undertaken to test whether listeners without access to lexical and grammatical information in the data would exhibit the same ability to identify prosodic boundaries. We will further analyze possible acoustic and prosodic correlates of these judgments.

2. The Experiments

For our studies we conducted a variant of the gating paradigm, in which spontaneous Swedish utterance fragments were presented to listeners, who were instructed to guess whether or not the fragments are followed by a break, and, if so, to rate its strength on a scale from 1 to 5. Our goal was to test whether upcoming boundaries could be identified under conditions in which a) pausal information is not present, and b) lexical and grammatical information is similarly not available. A further goal was to investigate which potential cues might account for listener ability to make such boundaries predictions.

2.1. Database

The speech corpus was selected from one interview of a female politician (GS) that was originally broadcast on public Swedish Radio. The entire interview was prosodically labeled by three independent researchers in the project [9] with respect to boundary presence and strength, with a majority voting strategy used to resolve disagreements.

2.2. Stimuli

60 utterance fragments (each about 2 seconds long) by GS were selected for the experiments. The exact initial cutting point was moved to the nearest word boundary, whereas the final cutting point was fixed. The fragments all preceded the word “och” (and) in their original context, and were cut just before the silent interval (if any) preceding that word. The decision to use the word “och” was partly motivated by syntactic considerations, given that the fragments then all occurred in comparable syntactic positions before an identical conjunction. The fragments differed with respect to the presence or absence of a break between the end of the fragment and the word “och”: in about one third of the cases, labelers found a strong intervening break at the end of the fragment; in about one third they identified a weak break; and in the remainder they judged there to be no break at all after the fragment. From these longer fragments, we then constructed shortened versions consisting of only the final word of the fragment.

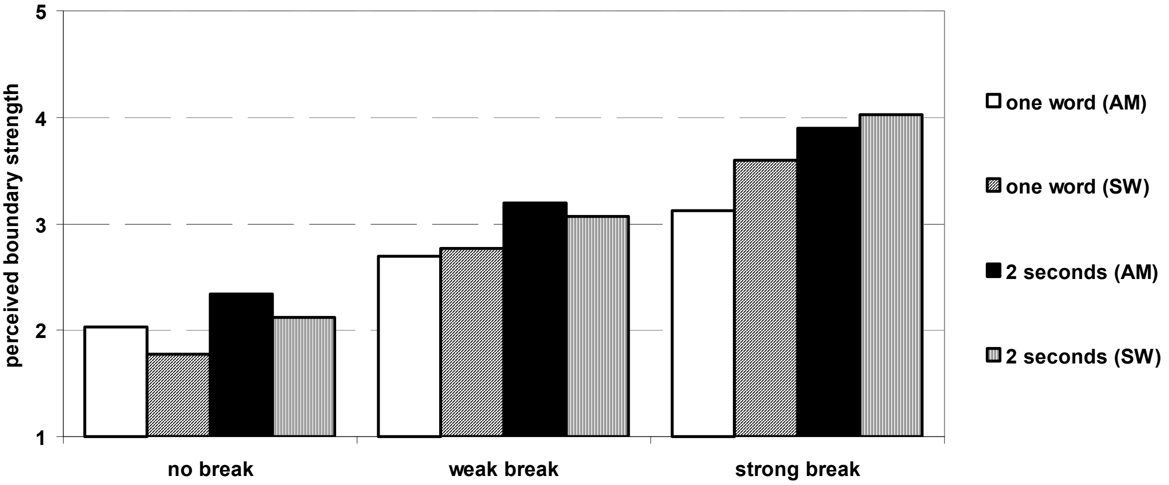

Figure 1. *Perceived upcoming boundary strength (subject scores on a 5-point scale). Data grouped according to expert labeled boundary strength (no,weak or strong break), fragment size and native language*

2.3. Subjects

The Swedish subjects (SW) consisted of 13 students of logopedics from Umeå University, Sweden. It can be assumed that these students also have a good knowledge of English. The American subjects (AM) consisted of 29 staff and students at Columbia University, USA, all native speakers of standard American English with no knowledge of the Swedish language. The second group was chosen to test whether in fact lexical or grammatical information provided the primary cue as to upcoming boundary location and strength. That is, since there is considerable evidence of syntactic correlates of prosodic structure (See for example [2], [3], [13]), perhaps the Swedish subjects in the earlier studies were making use of such cues in their boundary decisions. English-speaking listeners were chosen because of the prosodic phrasing similarities between English and Swedish.

2.4. Perceptual experiment

The 120 different stimuli (long and short versions, preceding a strong, weak or no boundary) were randomized and presented sequentially to our listeners via a specifically designed interface, which allows us to run perception experiments through the internet using a standard web browser with audio facilities. To minimize possible learning effects, each subject was presented with a differently randomized list of stimuli. The task was to rate each stimulus on a 5-point scale according to whether subjects 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 experiment was preceded by a short introduction which briefly explained a few concepts (such as prosodic boundary) and the actual task. No feedback was given on their responses, and there was no interaction with the experimenters. During the test, subjects could listen as many times as they wished to a given stimulus before giving an answer, but they could not return to a previous stimulus after a response had been entered.

3. Results

In Figure 1, judgments are presented in terms of labeled boundary strength, fragment length, and native language. Note that, for the American subjects, as for the Swedish, there is a strong correlation between perceived and labeled boundary strength for both one word and 2 second fragments.

In Figure 2 the same data are grouped only by stimulus length. Interestingly, the one word stimuli receive consistently lower scores compared to the 2 second stimuli. That is, the more speech that subjects were given to judge, the greater was their propensity to hypothesize an upcoming boundary. This result is independent of subjects' native language. A Within-Subjects test shows that there is no significant difference between subjects with different native languages (F(1,110)= 0,05; p < 0,82).

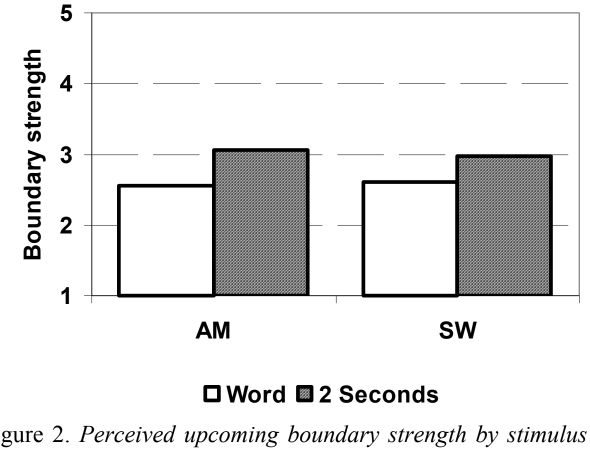

Figure 2. *Perceived upcoming boundary strength by stimulus length. Data grouped according to subject's native language American (AM) and Swedish (SW).*

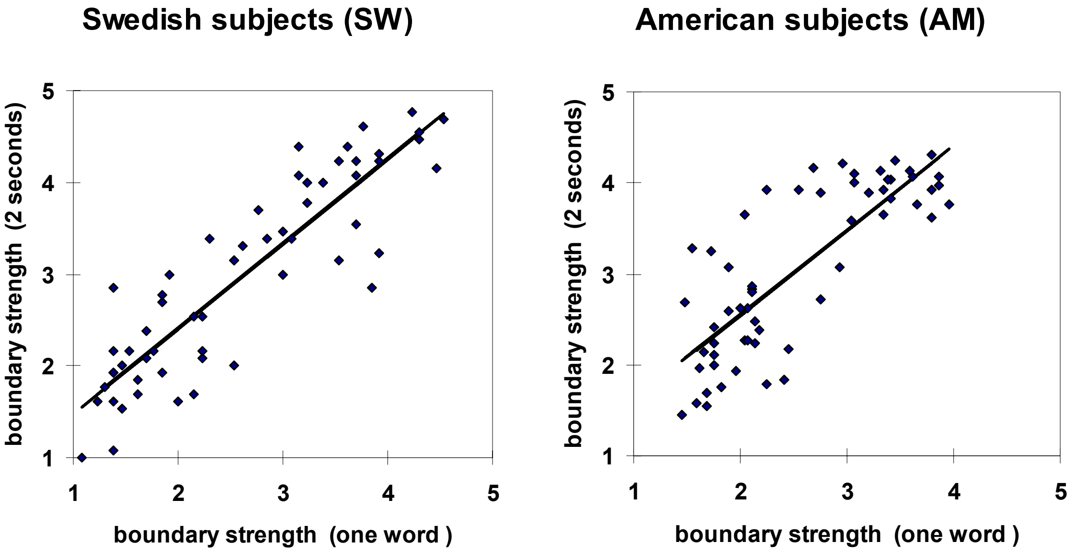

Figure 3. *Correlation between perceived upcoming boundary strength for each word in isolation and in a 2 seconds fragment for the Swedish and American subjects Regression coefficient r = 0,89 (SW) and r= 0,80 (AM)*

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)=73,4; p<.01) as well as of Fragment size (F(1,110)=13,4; p<.01) on the perceived boundary strength. There was no significant interaction between Boundary type and Fragment size. A Tukey HSD post hoc test showed that all three boundary types were significantly different from each other (p<.01).

While the American subjects did exhibit a slightly higher standard deviation in judgments than the Swedish subjects (1.30 vs. 1.19), we can nonetheless conclude that the absence of grammatical and lexical information did not significantly affect listeners ability to make accurate boundary predictions.

Since each short stimulus was also part of a 2 second 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 = 0,89 for the SW subjects and r = 0,80 for the AM subjects) between judgments on the two fragment sizes. So, subjects tended to judge the same boundaries similarly, whether they were given the single word or the longer preceding phrase.

To identify which features of the various stimuli might be influencing subject judgments, we examined some potential acoustic and prosodic cues. Word fragments were acoustically analyzed in terms of presence/absence of final creak, using spectrographic analysis and the median F0 value of the last voiced 50 ms of the word. A small but significant correlation between the final median F0 value and boundary strength was found, (r=0,62; p<.01) for the SW subjects, while the AM subjects show a lower but still significant correlation (r=0,46; p<.01). Other tested F0 cues, such as phrase-final F0 slope, turned out to have less predictive power but still significant (r=0,51; p<.01) for the SW subjects while it was about the same (r=0,49; p<.01) for the AM subjects.

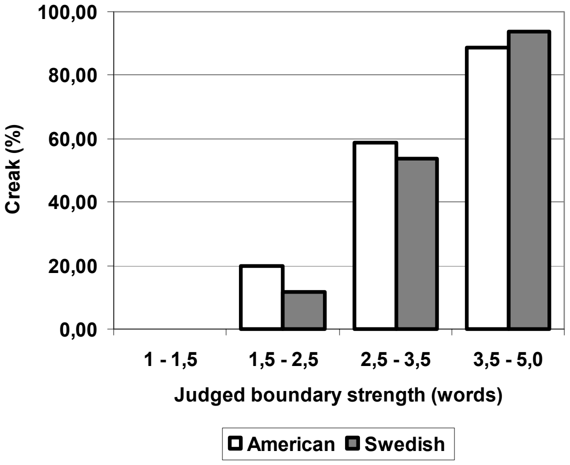

Figure 4. *Number of stimuli with creaky voice (in %) for different judged boundary strength intervals (one word)*

4. Discussion and Conclusion

The results of our current studies show that listeners are in fact able to predict upcoming boundaries based on properties of the preceding word or phrase alone, without access to a following pause. This finding supports a on-line processing model in which listeners can structure the incoming signal into prosodic phrases without the need to process subsequent material. While subsequent pause may serve as a supporting cue for this processing, it does not seem to be primary.

We also find that listeners with no knowledge of Swedish make boundary predictions as accurate as those of Swedish listeners, suggesting that acoustic and prosodic information, can be used in the absence of lexical or grammatical features, to make these decisions. So, if Swedish listeners are indeed making use of explicit lexico-grammatical features, this source of information does ***not*** give them an advantage in their judgments.

One of the intriguing finding of our studies is that for both SW and AM subjects, listeners' predictive ability is independent of the amount of preceding context available to them. Responses for the two types of stimuli, namely 2 second fragments and 1 word stimuli, are quite similar, as is evident from the high correlation between the two sets of responses. While there ***is*** an overall difference between these responses in that the longer context produces significantly higher values for all three classes (no boundary, weak boundary, strong boundary), the overall similarity in listener judgments for the two versions of each stimulus implies that longer context does ***not*** lead to a greater accuracy. This result is rather counter-intuitive, as one might expect that the task of guessing an upcoming boundary will be easier given a larger context. One explanation might be that it is the final word of the fragment that contains the critical acoustic or prosodic features which facilitate the prediction of upcoming breaks. Descriptive studies of intonational phrase boundaries in Swedish and American English in fact support this possibility, finding important boundary predictors located in the final word, including type of boundary tone preceding the break, final lengthening, loudness patterns, and possible effects of voice quality (e.g. the amount of creakiness). Certainly the presence of vocal creak is correlated in our own experiments with subject judgments.

This leaves us with the question of what the role of prosodic and lexico-syntactic features is in predicting upcoming boundaries. Based on our current findings, we conjecture that listeners' ability to predict upcoming prosodic boundaries may be primarily based on acoustic cues. However, there may also be redundancy in the two sources of information. Further, since syntactic structure and lexical choice is strongly correlated with the placement and acoustic realization of prosodic boundaries themselves, the relationship between acoustic-prosodic and lexico-grammatical features may be difficult to tease apart.

5. Acknowledgments

Marc Swerts is also affiliated with the Fund for Scientific Research – Flanders (FWO - Flanders). We would like to thank Theo Veenker for help with setting up the experimental environment and Mattias Heldner for the ANOVA analysis. This work has been carried out within the Swedish project "Boundaries and groupings - the structuring of speech in different communicative situations" (GROG), a project whose overall goals is to model the structuring of Swedish speech in terms of prosodic breaks and groupings [4].

6. 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*, Denver, USA.

[2] Bruce, G., 1995. Modelling Swedish Intonation for Read and Spontaneous Speech, *ICPhS 95*.

[3] Bruce, G.; Granström, B.; Gustafson, K.; House, D., 1993. Prosodic Modelling of Phrasing in Swedish, *ESCA Workshop on Prosody*.

[4] 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. *Fonetik 2002*, TMH-QPSR, 44.

[5] Carlson, R.; Swerts, M., 2003. Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials, *ICPhS 03*.

[6] Fant, G.; Kruckenberg, A.; Liljencrants, J., 2000. Acoustic-phonetic Analysis of Prominence in Swedish. In A Botinis (ed.), *Intonation, Analysis, Modeling and Technology* (Kluwer).

[7] Ferrer, L.; Shriberg, E.; Stolcke, A., 2002., Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody, *ICSLP 2002*, Denver, USA.

[8] Hansson, P., 2003. Prosodic Phrasing in Spontaneous Swedish. *Travaux de l'institut de linguistique de Lund 43*, Dept. of Linguistics and Phonetics, Lund University, Sweden.

[9] Heldner, M.; Megyesi, B., 2003. Exploring the prosody-syntax interface in conversations, *ICPhS 03*.

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

[11] Strangert, E.; Heldner, M., 1995. Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers. *PHONUM 3*,. Umeå: Department of Phonetics, Umeå University. 85-109.

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

[13] Wightman, C.; Shattuck-Hufnagel, S.; Ostendorf, M.; Price, P., 1992. "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries." *Journal of the Acoustic Society of America 91(3)*. 1707-1717.