

# Cues to upcoming Swedish prosodic boundaries: subjective judgment studies and acoustic correlates

Rolf Carlson<sup>a</sup>, Julia Hirschberg<sup>b</sup>, and Marc Swerts<sup>c\*</sup>

<sup>a</sup>KTH, Sweden, <sup>b</sup>Columbia University, USA and <sup>c</sup>University of Tilburg, The Netherlands and Universitaire Instelling Antwerpen, Belgium \*Names in alphabetic order

## Corresponding Author:

Rolf Carlson <rolf@speech.kth.se>  
Dept. Speech, Music and Hearing, KTH (Royal Institute of Technology)  
Lindstedsvägen 24. 5th floor  
SE-100 44 Stockholm, Sweden  
phone: +46 8 790 7568  
fax: +46 8 790 7854  
<http://www.speech.kth.se/~rolf>

## Abstract

Studies of perceptually-based predictions of upcoming prosodic boundaries in spontaneous Swedish speech, both by native speakers of Swedish and of native speakers of standard American English reveal marked similarity in judgments. We examined whether Swedish and American listeners were able to predict the occurrence and strength of upcoming boundaries in a series of web-based perceptive experiments. Utterance fragments (in both long and short versions) were selected from a corpus of spontaneous Swedish speech, which was first labeled for boundary presence and strength by expert labelers. These fragments were then presented to listeners, who were instructed to guess whether or not they were followed by a prosodic break, and if so, what the strength of the break was. Results revealed that both Swedish and American listening groups were indeed able to predict whether or not a boundary (of a particular strength) followed the fragment. This suggests that acoustic and prosodic, rather than lexico-grammatical and semantic information was being used by listeners as a primary cue. Acoustic and prosodic correlates of these judgments were then examined, with significant correlations found between judgments and the presence/absence of final creak and phrase-final F0 level and slope.

**Keywords: prosodic boundaries, prosody perception**

## 1. Introduction

Earlier studies have demonstrated not only that listeners can readily detect the presence or absence of prosodic boundaries but that they are also able to distinguish between different *degrees* of boundary strength (e.g. Dutch: Sanderman, 1996; Swedish: Fant et al., 2000 and Hansson, 2003; English: Price et al., 1990; Wightman et al., 1992). Such research has also found that the perception of boundary and of degree of boundary strength is heavily dependent on the occurrence of a silent pause at the perceived boundary location. While other factors, such as a rise or fall in f0, change in intensity, lengthening of the syllable preceding the perceived boundary, or glottalization over the preceding syllable or syllables, may also influence boundary perception (Baron et al., 2002; Ferrer

et al., 2002; Grosjean, 1983; Klatt, 1979; Leroy, 1984; Swerts et al., 1994), most studies have found the presence of some amount of pause before the boundary to be the single most salient cue to boundary perception, even in the absence of other cues. However, the presence of these additional features is still of considerable importance in suggesting how listeners may be able to process speech input in real time, while phrases are being produced --- before a pause occurs.

Most of the early studies are limited to read-aloud or specifically elicited speech materials, and they do not always clarify how 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).

Recently, Carlson and Swerts (2003) 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. More recently, the current authors extended this study to English listeners' perceptions of the Swedish data, finding that, even in the absence of possible lexical and syntactic information, non-Swedish speakers could predict upcoming boundaries on a par with native Swedish speakers.

In the current paper, we will expand upon these studies and report on additional analyses of the acoustic and prosodic correlates of listener 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 were followed by a prosodic boundary, and, if so, to rate its strength on a scale from 1 to 5. Our goals were twofold: We first wanted to test whether upcoming boundaries can be identified reliably by native speakers of a language, when pausal information is not present. We additionally wanted to determine whether such identification could take place when the lexical and grammatical information available to native speakers for the preceding phrase was not available; that is, could acoustic/prosodic information alone be used to identify upcoming boundaries before the production of a pause. If indeed native and non-native speakers could make such boundary predictions reliably, we next wanted to know which potential cues might account for their predictions.

### **2.1. Speech Stimuli**

The stimuli we used for our experiments were selected from an interview given by a female politician (GS), which was originally broadcast on public Swedish Radio. The entire interview was prosodically labeled by three independent researchers in the project (Heldner and Megyesi, 2003) with respect to boundary presence and strength, with a majority voting strategy used to resolve disagreements.

From this corpus, 60 utterance fragments (each about 2 seconds long) spoken by GS were selected for the experiments. 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 exact initial cutting point was placed at the nearest word boundary 2 second before the final cut point. 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" and the strength of that break. Approximately one third of the stimuli

were selected from locations in which our labelers had found a strong prosodic boundary at the end of the fragment. About one third were taken from contexts where the labelers had identified a weak boundary. And the remainder of the stimuli were selected from context in which the labelers had found no boundary after the fragment. Finally, from each of these 2-second fragments, we constructed a short version, consisting of only the final word of the fragment. Thus, we produced a total of 120 tokens for the perception experiments.

Figure 1. Perceived upcoming boundary strength. Data grouped according to labeled boundary strength, fragment size and native language

## **2.2. 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. This group was chosen to determine whether native speakers of a language can identify upcoming prosodic boundaries in their own language reliably, with no pausal cues, and, if so, how much speech material preceding the location to be judged they required in order to make reliable judgments. 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. This group was selected to test whether in fact lexical or grammatical information provides a necessary cue to upcoming boundary location and strength, or whether listeners can rely upon acoustic and prosodic information alone. Since there is considerable evidence that there are strong syntactic correlates of prosodic structure (See for example Bruce, 1995; Bruce et al., 1993; Price et al., 1990; Strangert and Heldner, 1995; Wightman et al., 1992), if native speakers can indeed predict upcoming boundaries in their language reliably, it may be because they make use of lexical or syntactic information. English-speaking listeners were chosen for this phase of the study because of the prosodic phrasing similarities between English and Swedish.

## **2.3. Perceptual experiments**

For the perception experiments, our 120 stimuli (long and short fragments, preceding a strong boundary, weak boundary 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 given to subjects was to rate each stimulus on a 5-point scale from 'no boundary at all follows this fragment' (1) to 'a strong boundary follows this fragment' (5). The actual experiment was preceded by a short introduction which briefly explained a few concepts (such as prosodic boundary) and the actual task.

**Introduction:**

*Thank you for participating in this test. We are studying how people make "breaks" between words. For example, speakers can put pauses in their utterances or can signal otherwise that there is some boundary between two consecutive words.*

*In this experiment, you will be presented with spoken utterance fragments (Swedish) that are either 2 seconds long or that consist of only one word.*

*These fragments could look like this:*

*2-second fragment: när man tog avstånd naturligtvis ...*

*long word: paragrafen ...*

*short word: den ...*

*In this experiment we would like you to judge on how strong a break will follow these fragments, for instance after the words "naturligtvis", "paragrafen" and "den" in the examples above. You will need to express your judgment on a 5-point scale. If you think there will be strong break after the last word, then you respond with 5. If you feel there will be no break after the last word, then you respond with 1. The rest of the scale you can use to mark the in between categories. We ask you to always give an answer, even if you are unsure about your answer.*

No feedback was given on subjects' responses, and there was no interaction whatsoever with the experimenters. During the test, subjects could listen as many times as they wished to a given stimulus before making a judgment, but they could not return to a previous stimulus after a response had been entered.

### 3. Results

#### 3.1. Perceptual Judgments

Results of our perception experiments showed that both sets of listeners could make reliable decision about upcoming boundaries from both the short and the long fragment stimuli.

Figure 1 presents judgments 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. Figure 2 shows the same data grouped only by stimulus length. Interestingly, the one word stimuli receive consistently lower scores compared to the 2 second stimuli. So, the more speech that subjects were given to judge, the greater was their tendency to hypothesize an upcoming boundary. This result is independent of subjects' native language. A Within-Subjects comparison shows no significant difference between Swedish and English-speaking subjects ( $F(1,110) = 0,05$ ;  $p < 0,82$ ).

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).

Both boundary type and fragment size thus influenced subject judgments. 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 subjects' perception of 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 English-speaking subjects *did* exhibit a slightly higher standard deviation in judgments compared to the Swedish subjects (1.30 vs. 1.19), the fact that they also did predict upcoming boundary strength quite accurately indicates to us that the absence of grammatical and lexical information does not significantly affect listeners ability to make accurate boundary predictions. Thus it is at least possible to predict upcoming boundary strength from acoustic and prosodic information alone.

Since each short stimulus was also part of a 2 second fragment, it is possible to correlate the perceptually based prediction of upcoming prosodic boundaries based on difference in amount of speech available for judgments. 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 phrase as preceding context.

### 3.2. Acoustic and Prosodic Correlates

We next attempted to identify which features of the fragments rated in our perceptual experiments might be influencing subject judgments. To this end, we examined several potential acoustic and prosodic cues, including glottalization over the final word in the fragment and changes in  $f_0$  at the end of the fragment.

Word fragments were acoustically analyzed in terms of presence/absence of final creak (glottalization), using spectrographic analysis and the median  $F_0$  value of the last voiced 50 ms of the word. Figure 4 shows that indeed fragments with more creak were more likely to be judged to precede a stronger boundary. Also, we found a small but significant correlation between the final word's median  $F_0$  value and perceived boundary strength, ( $r=0,62$ ;  $p<.01$ ) for the SW subjects, while the AM subjects show a lower but still significant correlation ( $r=0,43$ ;  $p<.01$ ) (Also shown in Figure 5). Other  $F_0$  cues we examined, such as phrase-final  $F_0$  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. So, it would appear that several acoustic and prosodic cues may account for listeners' judgments of boundary strength, with creak and mean  $f_0$  of fragments' final word appearing to be the strongest candidates.

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

Figure 5. Correlation between perceived upcoming boundary strength and the  $F_0$  median (Hz) final 50 ms for a) Swedish listeners and b) American listeners.

Heldner and Megyesi (2003) showed, when analyzing the original material from same speaker (GS), that words and word-final rhymes before weak boundaries were considerably longer than before strong boundaries, thus indicating relatively more final lengthening. Strangert (2004) reports that the same pattern is evident across the different sizes of chunks found in the data. There is a

general trend, in particular at boundaries judged as weak, of increasing lengthening the less words in the chunk.

#### 4. Pilot experiment using non American or Swedish listeners

In addition to the two user groups, American and Swedish, a third group (NN) was tested in a pilot experiment using the same procedure, Figure 6. This group consisted of seven students of Chinese origin also studying at Columbia University, USA. The result for these students showed similar pattern for the longer (2 seconds) stimuli compared to the American and Swedish subjects but they showed less ability to make the break prediction based on only one word stimuli. As can be seen also, the America listeners showed a tendency also to make more accurate predictions from longer stimuli. The results suggests that the linguistic background might have an influence on how prosodic features can be exploited for break prediction.

*Figure 6.* Perceived upcoming boundary strength. Data grouped according labeled boundary strength and native language and to fragment size a) 2 seconds and b) word

#### 5. Discussion

Our studies of Swedish and English speakers' judgments of upcoming boundaries in spontaneous Swedish speech 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. Furthermore, reliable judgments can be made by listeners without access to lexical or syntactic information from the speech. These findings support a on-line processing model in which listeners structure incoming speech into prosodic phrases without the need to process subsequent material in the input signal. While subsequent pause may serve as a supporting cue for this processing, it does not seem to be primary. And while lexico-grammatical information may also be useful to listeners in segmenting spoken input into prosodic phrases, it is not necessary to that segmentation.

An unexpected additional finding from our experiments is that, for both SW and AM subjects, listeners' predictive ability is independent of the amount of preceding context available to them: Judgments of the short (1 word) and longer (2s) are quite similar, as is evident from the high correlation between the two sets of responses. While the longer context does produce 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. We had originally hypothesize that the task of guessing an upcoming boundary would be easier, given a larger context. One explanation for this counter-intuitive result might be that it is the final word of the fragment that in fact 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 and  $f_0$  change are correlated in our own experiments with subject judgments.

This leaves us with the question of what the roles of prosodic vs. lexico-syntactic features may be in normal speech processing. 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.

## 6. 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 goal is to model the structuring of Swedish speech in terms of prosodic breaks and groupings Carlson et al., 2002.

## 7. References

- 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.
- Bruce, G., 1995. Modelling Swedish Intonation for Read and Spontaneous Speech, Proc. ICPhS 95.
- Bruce, G., Granström B., Gustafson K. and House, D., 1993. Prosodic Modelling of Phrasing in Swedish, Proc of an ESCA Workshop on Prosody.
- 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.
- Carlson R., Swerts M., 2003. Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials, Proc. ICPhS 03.
- Fant G., Kruckenberg A., Liljencrants J., 2000. Acoustic-phonetic Analysis of Prominence in Swedish. In A Botinis (ed., Intonation, Analysis, Modeling and Technology (Kluwer).
- 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.
- Gee, J. P., F. Grosjean, 1983. Performance Structure: A Psycholinguistic and Linguistic Appraisal. *Cognitive Psychology* 15. 411-458.
- Grosjean, F., 1983. How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistics* 21, 501-529.
- 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.
- Heldner M., Megyesi B., 2003. Exploring the prosody-syntax interface in conversations, Proc. ICPhS 03.
- 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.
- Leroy, L., 1984. The psychology of fundamental frequency declination. *Antwerp papers in linguistics* 40, University of Antwerp.
- 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.

- Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., Fong, C., 1990. The Use of Prosody in Syntactic Disambiguation, *Journal of the Acoustical Society of America* 90 (6), 2956-2970.
- Sanderman, A., 1996. Prosodic phrasing. Production, perception, acceptability and comprehension. PhD thesis, Eindhoven University of Technology.
- 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.
- Strangert, E., 2004. Speech chunks in conversation: Syntactic and prosodic aspects. *Proc. Speech Prosody 2004*, Nara, 305-308.
- Swerts M., Collier R., Terken J., 1994. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication* 15, 79-90.
- Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P., 1992. Segmental Durations in the Vicinity of Prosodic Phrase Boundaries, *Journal of the Acoustic Society of America* 91 (3). 1707-1717.

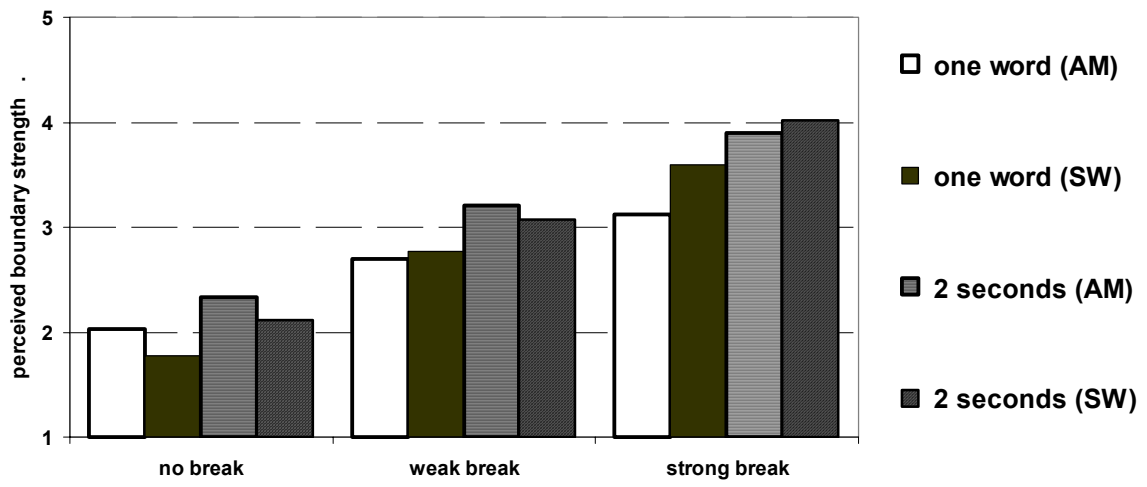
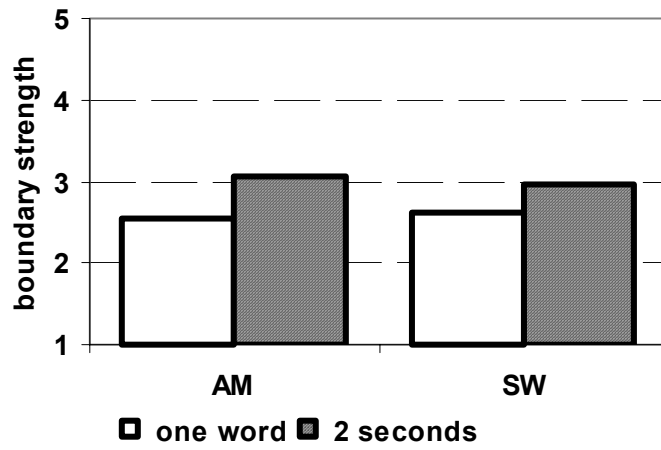
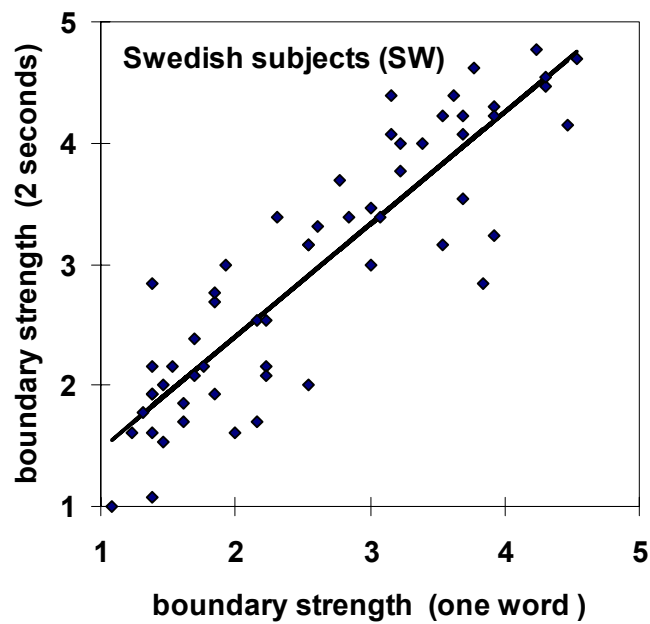
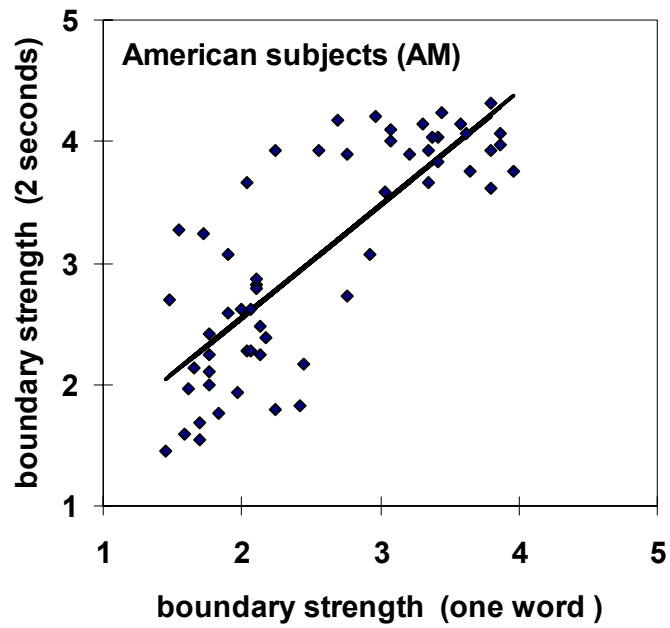


Figure 1. Perceived upcoming boundary strength. Data grouped according to labeled boundary strength, fragment size and native language



*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)

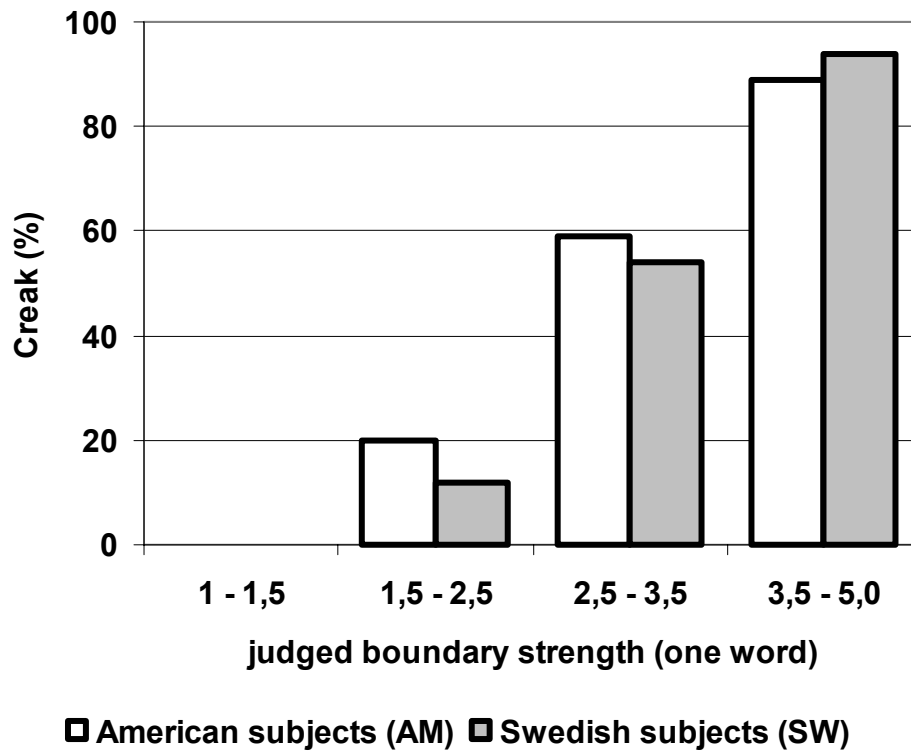


Figure 4. Number of stimuli with creaky voice (in %) for different judged boundary strength intervals (one word).

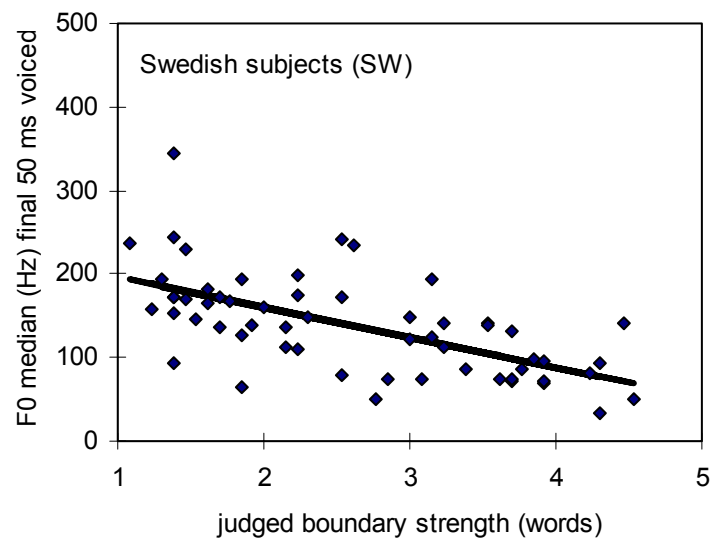
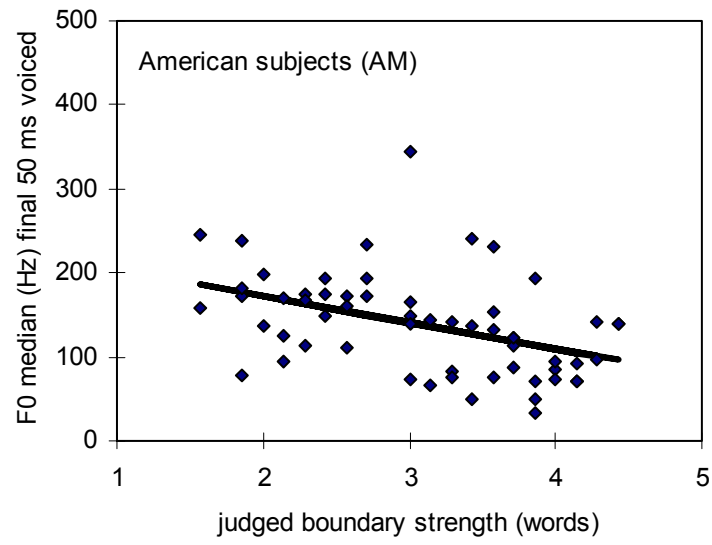


Figure 5. Correlation between perceived upcoming boundary strength and the F0 median (Hz) final 50 ms for a) American subjects and b) Swedish subjects.

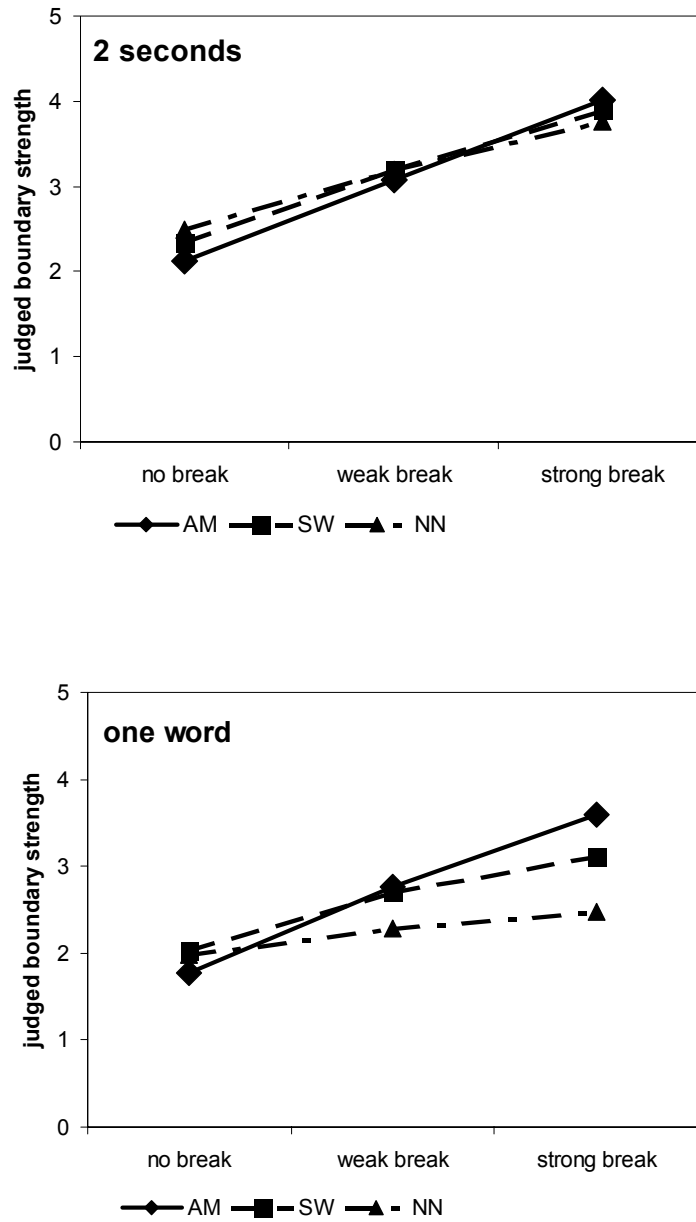


Figure 6. Perceived upcoming boundary strength. Data grouped according labeled boundary strength and native language and to fragment size a) 2 seconds and b) word