PERCEIVED BOUNDARY STRENGTH

Petra Hansson

Department of Linguistics and Phonetics
Lund University, Sweden
petra.hansson@ling.lu.se

ABSTRACT

In this paper, results from a perception experiment on perceived prosodic boundary strength in spontaneous speech are presented. The analysis of the listeners’ responses reveals a good agreement in rating the strength of prosodic phrase boundaries on a visual analogue scale (VAS), and a strong correlation between perceived boundary strength and pause duration.

1. INTRODUCTION

In this paper, results from a perception experiment are presented. The perception experiment has two purposes: 1) to relate perceived boundary strength to three known cues for prosodic phrasing in Swedish (pausing, F0 reset and final lengthening), and 2) to try to find empirical evidence to support the established division into two phrasal categories in Swedish. In the present paper, we will focus primarily on the former of these two aims.

1.1. Perception of prosodic phrasing in Swedish

The perception of prosodic phrasing in Swedish has already been investigated to some extent. A previous perception experiment [1] e.g. has shed light on the interaction of duration and F0. Horne, Strangert and Heldner [2] have investigated the relationship between different prosodic category boundaries (prosodic words, prosodic phrases and prosodic utterances) and final lengthening and pause duration in production data. However, to our knowledge, few attempts have been made to investigate the perception of prosodic boundary strength in Swedish.

1.2. Perceived boundary strength and the prosodic hierarchy

As discussed by Ladd [3], differences in boundary strength are only possible under the Strict Layer Hypothesis if they reflect differences of boundary type. According to the Strict Layer Hypothesis, any unit at a given level of the hierarchy consists exclusively of units at the next lower level of the hierarchy. In the tonal transcription system for Swedish [4], a distinction is made between two boundary strengths and two phrasal categories: prosodic phrases (which are delimited by weak boundaries) and prosodic utterances (which are delimited by strong boundaries). Although convincing evidence has been put forward to support the distinction made between three degrees of prominence in the intonation model for Swedish (stress, accent and focus), the validity of assuming two phrasal categories has not been demonstrated empirically.

Wightman, Shattuck-Hufnagel, Ostendorf and Price [5] note that, the ‘intonational phrase’ (corresponding loosely to the ‘prosodic phrase’ in the Swedish intonation model) is a prosodic constituent that is widely accepted among researchers. It can be defined as a group of words in an utterance which is delimited in some way as a larger unit of phrasing. A category upon which fewer researchers agree is the ‘intermediate phrase’ (which is delimited by a phrase accent), a level of phrasing between the prosodic word and the intonational phrase.

The intermediate phrase is generally not assumed to be a relevant phrasal category in Swedish. Although in many dialects of Swedish, a phrase accent is generally found phrase-finally (prior to the boundary tone), this is not always the case. The phrase accent is found after the word accent fall in words in focal position. Focally accent words are, however, not always perceived to be phrase-final, and therefore the presence of a phrase accent is not indicative of a phrase boundary. An example of this can be found in the evaluation of the Swedish base prosody transcription system, undertaken by Strangert and Heldner [6]. In the evaluation, nine expert transcribers (experienced phoneticians and speech researchers specializing in prosody) were asked to mark prominences and boundaries in a short read speech material (233 words). The focal accents on which the largest number of labelers agreed (7 of the 9 labelers) were found on words (libyska ‘Libyan’ and bombplanen ‘the attack planes’) after which none of the labelers perceived phrase boundaries.

As regards a higher-level phonological constituent, Wightman et al. [5] mention phonetic effects with a possibly larger domain than the intonational phrase, but note that they can be argued to relate to discourse structure. Nevertheless, in the modeling of Swedish intonation, the ‘prosodic utterance’, a higher-level phonological constituent is assumed. All prosodic phrases leading up to the next strong boundary in the stream of speech have to be considered parts of one and the same prosodic utterance under the Strict Layer Hypothesis. In spontaneous speech, this often causes conflicts because of the expectation that a prosodic utterance not only is defined by the strong boundaries delimiting it, but also by its internal prosodic structure, i.e. by some sort of coherence-signaling between the phrases in the utterance (such as tonal coupling [7]). If it can be shown that listeners exhibit good agreement in rating more than two boundary strengths, there may be reasons to look for another way of accounting for differences in boundary strength than with the above mentioned differences of boundary types.
2. METHOD

2.1. Stimuli

In order to investigate the relationship between perceived boundary strength and the prosodic cues for prosodic phrasing, a total of 50 short speech fragments were chosen from the SweDia 2000 database [8]. Within the research project SweDia 2000 (Phonetics and Phonology of the Swedish Dialects around Year 2000), both spontaneous speech and words and phrases elicited with a number of specific research questions in mind have been recorded. In the present study, speech extracted from the spontaneous part of the database has been used. Speech from five female and five male subjects, all from Skåne, the southernmost region of Sweden, have been used (five speech fragments from each speaker). The recordings were made with a sampling rate of 44.1 kHz and 16 bit resolution, and have been transferred digitally to a Sun workstation and stored as ESPS/Waves+™ files. The sampling frequency has subsequently been converted to 16 kHz.

All speech fragments (typically one or two utterances long) contain at least one prosodic phrase boundary. In the perception experiment, the listeners were presented with an orthographic transcription of the utterance(s) in which the boundary of interest was marked with a ‘/’, as in (1).

(1) Sen tittar jag på lite såpor, / och så älskar jag musik.

‘Then I watch some soaps, / and I love music.’

All boundaries to be judged by the listeners occurred either between complete sentences or between clauses, i.e. in syntactically motivated positions. In spontaneous speech, it is sometimes hard to determine whether two speech fragments are best described as two separate sentences introduced by discourse markers such as and or but or as two coordinated main clauses. Therefore, no attempts have been made to restrict the test to only include sentence boundaries.

2.2. Listeners’ task

The listeners were presented with a list of the 50 utterances/utterance pairs, and placed in front of a computer screen. Only the last word before the boundary of interest was displayed on the computer screen, see Figure 1. The subjects were instructed to read each utterance/utterance pair and listen to the recording of it. The recordings could be played and replayed by clicking on a button on the screen. The subjects’ task was to indicate how strongly marked they perceived the boundary in the recording. The whole experiment lasted roughly 20 minutes. The order of presentation of the utterances was randomized and different for all listeners. A trial run containing three recordings preceded the actual test.

2.3. Visual Analogue Scale

The listeners indicated the boundaries’ strength on a so-called Visual Analogue Scale (VAS). The VAS has been used in the measurement of clinical phenomena (such as pain) since the 1920s [9], but it is suitable for measuring a variety of subjective non-clinical phenomena as well. Often the VAS is presented to subjects on paper. Its most common form is a 100 mm horizontal line. The subjects respond by placing a mark through the line at a position that represents their current perception of a given phenomenon (in this case a given boundary’s strength) between the labeled extremes of the scale (in this case between ingen gräns ‘no boundary’ and mycket stark gräns ‘very strong boundary’, see Figure 1). The VAS is then scored by measuring the distance (in mm) from one end of the scale to the subject’s mark on the line. The level of measurement represented by VAS data is usually assumed to be interval or ratio [9].

![Figure 1: In the computer program Judge][10], the listeners rated the strength of given boundaries by adjusting a scrollbar.](image)

In the present perception experiment, a computer-based VAS has been used. The advantages of computer-based VAS scales are discussed in detail in [10].

Eighteen subjects have completed the perception test (five phoneticians and speech researchers, eight students (all with some knowledge of prosodic labeling) and five naïve listeners).

2.4. Measurements and normalization

Three properties of the boundaries in the stimuli have been investigated: the pause duration (in 33 of the 50 speech fragments, the boundary was associated with a silent interval), the amount of final lengthening (or rather the difference in articulation rate between the penultimate and final word in the prosodic phrase preceding the boundary), and the reset of F0 across the boundary.

The duration of the pause may be perceived differently depending on whether the speaker talks fast or slowly. Therefore, in addition to measuring the actual duration of the pauses, the pause durations were also related to the speakers’ speaking rates (the pause’s length was divided by the average length of a syllable in the two prosodic words preceding the pause).

Since the inherent duration of a phone is known to be the largest source of variation in segmental duration, when measuring final lengthening one would like to measure the difference between the duration of a given segment and the mean duration of that specific phone type. In [5], normalized duration, a method that measures the duration of a segment as the number of standard deviations from the mean duration of the phone contained in the segment, is developed. However, in order to obtain the means and standard deviations of a speaker’s phones, a fairly large amount of speech data needs to be segmented and labeled. Since we have no speech recognizer available to us for segmenting and labeling our south Swedish data automatically, we chose another method for measuring final lengthening in the present investigation. We measured and compared the articulation rate in the two last words in the phrase. Since the measured articulation rate (number of
syllables divided by the word’s duration) is greatly affected by the size of the word (or rather the number of unstressed syllables it contains) all stimuli were chosen in such a way that the prosodic phrases before the boundaries of interest ended with two two-syllable prosodic words. The difference in articulation rate was normalized for differences in articulation rate between speakers and recordings by dividing it by the average articulation rate in the two words being compared.

The F0 reset was measured as the difference in F0 between the last accent peak of the phrase before the phrase boundary of interest and the first accent peak in the following phrase (in Hz and in ST).

3. RESULTS AND DISCUSSION

3.1. General characteristics of the phrase boundaries in the stimuli

Two thirds of the examined boundaries (33 of 50) were associated with a pause (a silent interval). The pauses ranged in duration from 225 ms to 2847 ms. The F0 resets across the phrase boundaries ranged from a 186 Hz large reset of F0 to a lowering of F0 across the phrase boundary of 52 Hz. The normalized difference in articulation rate between the phrase final and penultimate word ranged from 1.0 (corresponding to a change in articulation rate from 5.4 syllables per second to 1.8 syllables per second) to –0.3 (corresponding to a change from 4.1 syllables per second to 5.6 syllables per second).

None of the three variables investigated are related to each other. There is no statistically significant correlation between the duration of the pause, the size of the F0 reset and/or the change in articulation rate. Apparently speakers do not maximize the use of several cues simultaneously to increase the strength of a boundary. This finding is partly consistent with the trading relation between pause duration and final lengthening reported on in Horne, Strangert and Heldner [2]. Out of the 50 stimuli, 14 are associated with a pause, a (positive) reset of F0 and a reduction of the articulation rate (indicating final lengthening, FL). Another 23 are associated with two of the above mentioned cues, 12 with only one cue and one with none of the cues, see Figure 2.

The prosodic phrase boundaries described above were chosen to suit our perception experiment; they are all syntactically motivated and they are all preceded by two two-syllable prosodic words in phrase final position. They are not a random sample of prosodic phrase boundaries, and we will therefore not make any attempts to generalize the observed prosodic patterns (i.e. the lack of a relationship between pause duration, amount of final lengthening and F0 reset, and the way in which the different cues are combined) to other circumstances. It is possible that prosodic phrase boundaries in e.g. other syntactic positions exhibit different patterns than those described here.

3.2. Results of the perception experiment

Statistical analyses reveal that the listeners agree very well in their perceptual judgments. The Pearson correlation coefficient of each pairwise combination of listeners was significant at the .001 level. Therefore, we have chosen to follow Sanderman [11] and pool the scores of the listeners for each boundary and calculate a mean to obtain an estimate of each boundary’s perceived strength (in mm). No apparent differences were found between the phoneticians’, students’ and naïve listeners’ perceived boundary strengths (PBS), so their scores were pooled together. In Figure 3, the mean perceived boundary strength of the stimuli types (the different cue combinations) are given.

Since each possible combination of cues was not represented by the same number of stimuli in the test (the experiment was not designed to test the effect of number of cues upon perceived boundary strength), some caution needs to be taken when interpreting the results shown in Figure 3. However, one pattern is clearly observable. Boundaries not associated with a pause are perceived as much weaker than those that involve pausing. Somewhat surprisingly, the number of cues used to mark a particular boundary does not appear to be very important for how strongly it is perceived. It is not clear from these results that a larger number of cues would lead to the perception of a stronger boundary, as perhaps expected. There is e.g. no evident difference in perceived boundary strength between those boundaries that are associated with all three cues (pause, F0 reset and reduction of articulation rate, n=14) and those boundaries only associated with a pause and a reduction
of the articulation rate (n=14). Figure 3 may give the impression of a clear division into two boundary strengths (providing evidence to support a distinction between two phrasal categories): a weak boundary with a perceived boundary strength of 0-10, and a strong boundary with a perceived boundary strength of 50-70. However, this was not the case. The perceived boundary strength (in mm) of the 50 stimuli ranges from 1 to 89 (see Figure 4 below), with a mean of 42 and a standard deviation of 30.

In order to investigate a possible relationship between the three variables pause duration, F0 reset and change in articulation rate and perceived boundary strength, Pearson correlation coefficients were calculated. The correlation between pause duration and perceived boundary strength proved very strong. Both the actual pause duration (r=.91, r²=.83, p<.001) and the normalized pause duration (r=.88, r²=.78, p<.001) correlate strongly with perceived boundary strength. As shown in Figure 4, the longer the pause is, the stronger the listeners perceive the boundary. The only clear exceptions are pauses longer than 2 seconds, which are not perceived to be stronger than pauses of 1.5 to 2 seconds.

However, F0 reset proved not to be related to perceived boundary strength, and adding the variable ‘change in articulation rate’ to the prediction of PBS scores made a negligible contribution (r=.92, r²=.85).

4. FUTURE WORK

Although no strong relationships were found between amount of F0 reset and perceived boundary strength or between reduction in articulation rate and perceived boundary strength, these are both cues that may be important for the perception of boundary strength for some listeners. In order to investigate this possibility, the present study will be completed with more detailed analyses of individual listeners’ responses to the stimuli. Furthermore, other ways of measuring and normalizing the three cues for prosodic phrasing will be sought. The perception experiment reported on in this paper has two purposes. Here, we have only dealt with the relationship between perceived boundary strength and three relevant cues to prosodic phrasing. The main finding is that pause duration is the strongest cue for perceived boundary strength. The importance of pause duration for perceived boundary strength is currently being investigated in more detail in a second perception experiment. In the second experiment, the listeners are asked to judge the same stimuli as those used in the first experiment, but the pauses occurring in the stimuli have been manipulated and set to have the same length. Unlike the listeners in the first experiment, the listeners in the second experiment do not agree in their perceptual judgments. Apparently, other possible cues to boundary strength are not perceptually relevant when they do not correlate with pause length. The second aim of the perception experiment presented in this paper, the implications the listeners’ good agreement in rating boundary strength may have on the established distinction between two boundary strengths and two phrasal categories, remains to be explored.

5. REFERENCES


