

## ASPECTS OF PROSODIC PHRASING IN SWEDISH

Gösta Bruce\*, Björn Granström\*\*, Kjell Gustafson\*\* and David House\*+

\*Department of Linguistics and Phonetics, Lund University, Helgonabacken 12  
S-223 62, Lund, Sweden

\*\*Dept of Speech Communication and Music Acoustics, KTH, Box 70014,  
S-100 44, Stockholm, Sweden

### ABSTRACT

This paper reports on an ongoing research project called 'Prosodic Phrasing in Swedish', the object of which is to investigate prosodic phrasing and grouping in Swedish. Different methods exploited within the project are the analysis of speech production data, the use of text-to-speech synthesis and the use of speech recognition (prosodic parser). Production data from specially designed test material for Swedish has shown that tonal and temporal cues are combined to signal differences in phrasing. To test this analysis a perceptual experiment was conducted using the KTH rule synthesis where duration and F0 could be changed interactively. 12 listeners participated in the test with the task of identifying optimal positions for two distinct interpretations of an ambiguous test sentence as well as a line of ambiguity. The results of the experiment confirm our initial analysis and provide interesting individual variation indicating different perceptual strategies which may also be related to speaking habits. Longer texts are used in prosodic parsing experiments where the task is to identify prosodic phrases.

### INTRODUCTION

The intention of the present paper is to give an overview of current research within a project called 'Prosodic Phrasing in Swedish'. The project is a cooperative effort between Phonetics at Lund and Speech Communication at KTH, Stockholm and is part of the Language Technology Programme in Sweden. Our main orientation is towards basic research: to gain new knowledge about phrasing and prosody as it appears in Swedish. There are also potential applications of our research within systems for speech synthesis and recognition. For other recent studies of prosodic phrasing and grouping in Swedish see [1][2].

The general problems we are dealing with relate to both phonetics and phonology. One of the foremost phonetic problems involves investigating what speech variables and combinations of them (F0, duration, intensity, phonation type, pausing, etc.) can be used to signal phrasing including the possibility of a hierarchy among these cues. The main phonological issue involves understanding what structure could be assumed for prosodic phrasing, particularly what factors (syntax, semantics, unit length, etc.) govern the grouping of words into phrases in speech. A related question is what types of prosodic phrases can be identified as relevant domains between a 'prosodic word' and a 'prosodic utterance' (for a discussion see for example [3][4][5]).

Our basic approach to the study of phrasing in spoken language is governed by recognizing that grouping in general, and specifically phrasing, is characterized by its two aspects: coherence (connective signalling) and boundary (demarcative signalling). Therefore in our search for phonetic signals of phrasing, coherence cues are deemed to be as important as boundary cues.

+ Names in alphabetic order

Three different methods are being exploited in the project. The first method is the collection and analysis of speech production data, ranging from specially designed test utterances to suitable read text passages (laboratory speech). This method also includes the processing of speech data in the KTH speech data base [6]. The speech production data method is being used primarily to generate hypotheses about important speech properties in phrasing.

The second method is the use of text-to-speech synthesis for the testing of hypotheses about the signalling of prosodic phrasing. In the KTH text-to-speech system there are several ways of interacting with rules and parameters [7]. This synthesis method can be used both for the testing of hypothesized generalizations on new, unrecorded speech material and for the design of more specific perception tests.

The third method is prosodic parsing directed towards the recognition of phrasing. We believe that the prosodic parser is particularly suitable for testing hypotheses about the interaction between speech variables for the expression of prosodic phrasing.

In this paper, the speech production data is exemplified by the analysis of some specially designed test sentences. Results of this analysis serve as a basis for a perception experiment using speech synthesis. The description of this experiment comprises the main body of this paper. Finally the use of prosodic parsing will be presented in outline form.

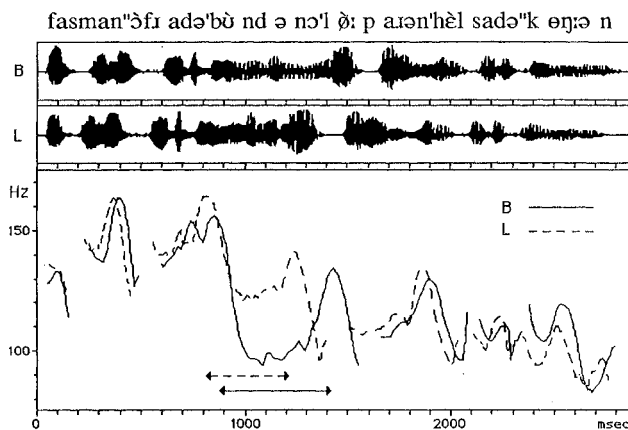
### SPEECH PRODUCTION DATA

Our initial approach was to collect and analyse production data from a series of syntactically ambiguous sentences comprising minimal pairs where the sentence internal boundary was varied. Our analysis showed that several different phrasing strategies were exploited to disambiguate the sentences. These strategies involve different combinations of F0 and duration cues which contribute to coherence and boundary signalling [8][9][10].

In most of this test material, phrasing and accentuation are partly interdependent, as deaccentuation is often used as a coherence cue for the division into phrases. In one type of sentence, however, accentuation stays the same, while phrasing is varied. The following sentence pair is used to illustrate this subset, where the characteristic difference is the location of the internal boundary (here represented by a comma) resulting in either a grouping of 2+3 accents or 3+2 accents:

B. Fast man offrade bonden, och löparen hälsade kungen.  
[ fast man "ɔ̃fradə 'bündən ɔ̃ 'löpərən hɛlsadə "køŋ:ən ]  
(but we sacrificed the pawn, and the bishop greeted the king)

L. Fast man offrade bonden och löparen, hälsade kungen.  
[ fast man "ɔ̃fradə 'bündən ɔ̃ 'löpərən hɛlsadə "køŋ:ən ]  
(though we sacrificed the pawn and the bishop, the king greeted us)



**Figure 1.** Waveforms and fundamental frequency contours of two natural productions of the sentence *Fast man offrade bonden och löparen hälsade kungen* with a phrase boundary after *bonden* (solid line) and *löparen* (dashed line). Arrows indicate the domain used for parameter manipulation of synthetic versions in the perception experiment.

A male Stockholm Swedish informant recorded these two sentences as well as an ambiguous version of them three times, altogether nine examples. These examples were tested informally using five listeners who were given the task of identifying them (forced choice) as either divided after *bonden* (B) or *löparen* (L). One typical and clearly identified version of each sentence is illustrated in Figure 1.

It is clear that both tonal and temporal cues are combined to signal the difference in phrasing. The notable F0 difference occurs after the 2nd accent ['bùndən] as a deep vs. shallow F0 valley. There is no corresponding F0 difference after the 3rd accent ['lø:pa:rən]. The main durational difference can be seen as a pre-boundary lengthening after the 2nd and 3rd accent respectively depending on the phrasing.

#### SYNTHESIS AND PERCEPTION

In this study some of the strategies of phrasing concerning the relationship between tonal and temporal cues have been tested perceptually using a version of RULSYS (the KTH rule synthesis) [7]. One feature of this system is the possibility of interactively changing the synthesis by moving a point on the computer screen. The corresponding X and Y parameters are used in the synthesis rules, and can, depending on the rule system, be made to affect the synthesis in different ways.

For the present test, we used the above test sentence, but changed the verb *hälsade* (greeted) to *hotade* ['hù:ta:də] (threatened) as being more suited to the chess context. The subjects were asked to listen to the way the test sentence was altered by moving the cursor around the screen. They were given two tasks for each experiment:

- \* to determine an optimal position for interpretation B and an optimal position for interpretation L, i.e. positions where the sentence can only be interpreted as having the boundary after *bonden* or *löparen* respectively without sounding exaggerated or too unnatural.

- \* to establish a 'line of ambiguity' across the screen such that readings along the line could be equally well interpreted as either B or L.

#### Description of the test

Three versions of the synthesis rules were used, corresponding to three different interactions of local segmental duration and fundamental frequency. In all three

tests, cursor movement along the X-axis varied the duration of phrase final segments, while F0 was varied by cursor movement along the Y-axis. In tests 1 and 2 duration and F0 were variables in the word *bonden* only, whereas in test 3 they were controlled simultaneously in both *bonden* and *löparen*, but in opposite directions. Test 2 employed an expanded F0 range, but was otherwise similar to test 1. The specific design of the testing in terms of duration and F0 variation was governed by the analysis of production data as illustrated above.

The manipulation of duration affected the segments between the stressed vowel of the test word and the following potential phrase boundary in all three tests. The F0 variation affected the low point of the falling contour associated with the stressed vowel having word accent 2 (grave accent), which in the variety of Swedish modelled in the synthesis is manifested by a fall starting near the beginning of the stressed vowel [11]. It was hypothesized that a reading with a shallow F0 valley as part of an F0 downstepping would be perceived such that the two words (*bonden* and *löparen*) belonged to the same phrase, whereas a very marked fall on the stressed vowel to a low F0 as a break in the downstepping trend would cue a boundary after the word containing the fall. Similarly, a lengthening of the segments following the stressed vowel was expected to cue a phrase boundary whereas no lengthening or a shortening was hypothesized to contribute to a cohesive effect.

The variation in duration was between 67% and 150% of the default ('neutral') duration of the segments in question, in the word *bonden* for tests 1 and 2, while they varied between 58% and 142% for test 3 in the manipulated segments of both *bonden* and *löparen* (for symmetry reasons). The variation in F0 was between 50 Hz and 133 Hz in tests 1 and 3. The expanded range in test 2 was between 75 Hz and 200 Hz. The maximum value resulted in a level pitch at the point of F0 manipulation.

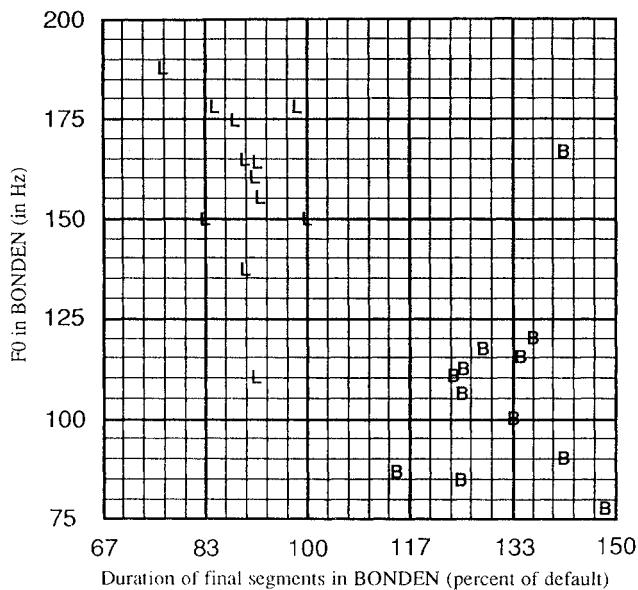
#### Test procedure

12 native speakers of Swedish were selected to act as subjects; all were staff members at KTH. The subjects listened to the synthesis through earphones. The computer screen displayed a square divided into 25 smaller squares of equal size. A cursor could be moved around the screen by means of the arrow keys in ten equal steps in each of the smaller squares. The subjects were asked to mark their answers on a sheet of paper which had a layout similar to that of the screen. The different tests were given to the subjects according to a rotated order design. There was no time limit specified. Most subjects finished the three tests within a total of 30-60 minutes.

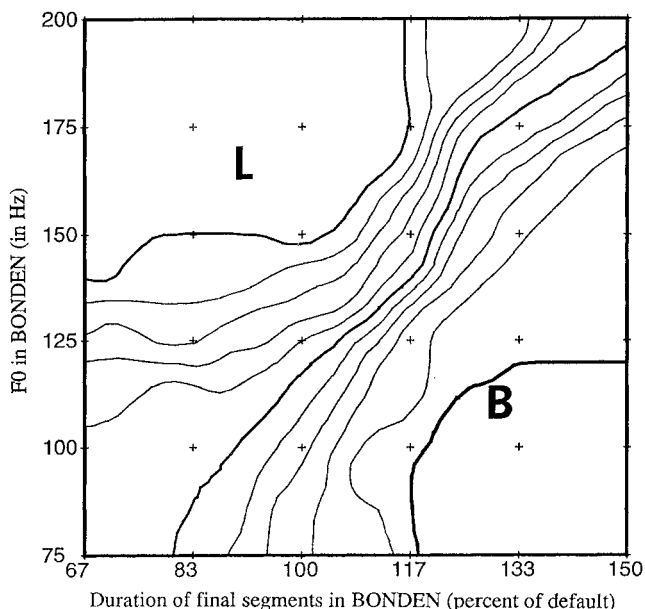
#### Results

In the present paper we will focus on test 2 as representative of the three tests. The results of all three tests show on the whole very similar trends. Figure 2 shows the individual, optimal points in the response square from test 2. "L" and "B" mark the responses for a boundary directly after *löparen* and *bonden* respectively. Although we can observe some response overlap in the F0 dimension (but none in the duration dimension), the responses demonstrate a clear clustering tendency for both L (upper left) and B (lower right).

In Figure 3, the pooled results of the same test are displayed. The median, optimal points are indicated by "L" and "B", and the lines represent contours of equal percentage, 10% apart, based on "lines of ambiguity". Both duration and F0 appear to be effective as phrasing signals. It is obvious that the fundamental frequency and duration differences found in natural speech serve the expected function in our synthesis experiment. A shallow F0 valley after the 2nd accent ['bùndən] combined with relatively short segment durations between the



**Figure 2.** Individual results of the perception experiment, test 2, showing optimal positions for synthetic versions with the phrase boundary after *löparen* (L) and *bonden* (B). The X-axis represents the duration of the segments following the stressed vowel of *bonden* in percent of the default duration. The Y-axis represents the F0 at the end of the stressed vowel of *bonden*.



**Figure 3.** Results of the perception experiment, test 2, showing median, optimal positions for synthetic versions with the phrase boundary after *löparen* (L) and *bonden* (B). The lines are contours of equal percentage, 10% apart, based on "lines of ambiguity". Heavy lines mark off areas of 100% response for *löparen* (upper left), 100% response for *bonden* (lower right), and the 50% line (middle). X and Y axes as in Fig. 2.

stressed vowel and the potential phrase boundary indicates that the two words *bonden* and *löparen* belong to the same phrase, while a fairly deep F0 valley combined with relatively long segment durations instead favours the interpretation of the two words as belonging to different phrases.

The individual spread is obvious from Figure 2 and is also observed in the individual ambiguity lines. This is reflected in Figure 3 where the contours represent mean percentage votes for the different interpretations 10% apart. These curves are rather close together in the centre of the graph between the optimal points, but very spread out in the periphery.

### Discussion

In order to better understand how the interactions of local segment duration and fundamental frequency can affect the perception of phrasing, the values of the manipulated test word have to be seen in relation to an overall F0 and duration plot of the whole test sentence. This can be done by referring to Figure 4 which displays F0 contours and segment durations of synthetic versions of the test sentence, where parameters are taken from median positions for phrase boundary at *löparen*, ambiguous phrasing, and boundary after *bonden*, as indicated in the Figure.

Although several possible ways of interpreting the results present themselves and should be pursued through a formal testing, the findings of the perception experiment generally do not appear to contradict a more global interpretation of tonal and temporal patterns. The durations and F0 values of the test word seem to be judged in relation to both what precedes and what follows for the same speech parameter in the test utterance. Thus the results could also be consistent with our more specific hypothesis presented above that a shallow F0 valley as part of an F0 downstepping pattern has a connective function signalling coherence within a prosodic phrase, while a deep F0 valley as a break in the downstepping trend has a demarcative function signalling a phrase boundary. See also [12] for a recent treatment of the perception of tonal patterns in speech.

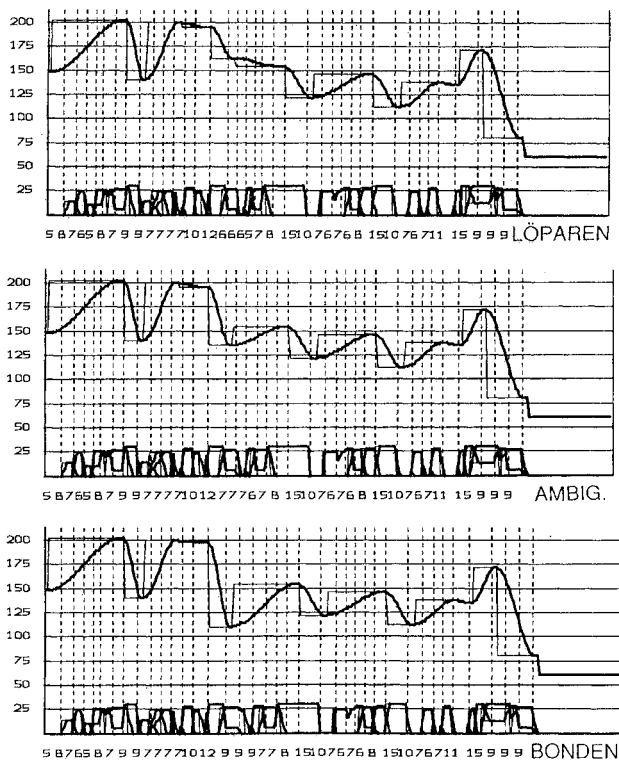
The question of a more global or a more local interpretation of the results from the perception experiment has not been settled yet, however. For a recent discussion of this see [13]. It is noteworthy that the results of test 3, where the subjects manipulated the parameters at both potential boundaries, are not significantly different from those of the other two tests. We take this to indicate that marking the boundary clearly enough after *bonden* makes a boundary marker after *löparen* redundant to some extent. This is in line with our observations above that in some of the productions of the reference speaker the main differences in terms of F0 are associated with the first of the two potential boundaries.

### Individual strategies

To account for the observed individual spread both in the location of the optimal points and in the establishing of an "ambiguity line", one possible interpretation is that subjects employ quite different listening strategies. One reason for the divergence of the responses in the bottom left and top right areas of Figure 3 may be that F0 and duration in these areas give counter-indications, resulting in a perceptual conflict which the subjects have resolved in different ways. Assessing the effect of the combination of those F0 contours and durational relations that they felt were unreasonable or impossible the subjects were forced to make a choice, and we hypothesized that they would make this choice on the basis of what for them would be the most important contribution to the boundary marking, which meant ignoring the less favourable contribution of the other parameter.

We find some support for this view in the individual behaviour of the subjects. Although it is clear that the majority of the listeners rely more equally on both F0 and duration, some of the ambiguity lines show a marked horizontal trend,

fasman"ɔ̄fɪa dɛ'bʊndɔ̄nɔ'lɔ̄p aɾɔ̄n'hʊ: t adə"k ɔ̄j: ɔ̄n



**Figure 4.** Synthesis parameter plots for the sentence used in perception test 2 where parameters are taken from median positions for phrase boundary after *löparen*, ambiguous phrasing, and boundary after *bonden* as indicated in Figure 3. Segment durations in csec. are indicated on the X-axis. The Y-axis is the parameter value in Hz for the upper curve (F0) and the relative level in dB for the lower source amplitude curves.

indicating that for those subjects duration was (relatively) unimportant in establishing the different readings. In others we find lines that are close to vertical, indicating that for those subjects it was duration that mattered most in determining the different readings. We find the same trends in regard to the establishing of optimal locations for B and L, and for some subjects these trends point strongly in the same direction for both the ambiguity lines and the location of optimal points. We conclude from this that some of the subjects are more "duration minded" while others are more "F0 minded" in tackling tasks like the ones in our perception experiment.

One possible explanation is that the listeners may have judged the experimental stimuli on the basis of how they themselves would have produced sentences with the same meanings. To test the hypothesis that the apparent duration-vs. F0-mindedness might reflect differences in the actual production of the subjects, we asked two of the subjects, who in the perception experiment were clear examples of the two types of listeners, to take part in a reading session. We asked them to read the test sentence in the three different ways that we had specified in the perception experiment.

In analysing the readings of the two subjects, we found evidence that the duration-minded subject made greater use of duration than did the F0-minded subject. This is in line with our hypothesis above. We did not, however, find any similar evidence in the recorded material regarding the use of F0.

## PROSODIC PARSING

The third method used in the project involves recording and analysing longer text passages read by two different speakers. These recordings (each about one minute in length) form the basis for a series of prosodic parsing experiments. In the experiments an expert mingogram reader is given the task of identifying prosodic phrases solely on the basis of a visual representation of the text showing the waveform, intensity and fundamental frequency (cf. [14]). The results are then compared to two independent, auditive based transcriptions of the readings.

Preliminary results indicate that the phrasing strategies used in our short test sentences and described above are also used by the different speakers in the longer text passages. These strategies can, in a majority of cases, be identified by the expert reader resulting in a visually identified prosodic phrase corresponding to auditive perceived phrasing. Results from the mingogram reading experiments will form the basis for the formulation of automatic recognition rules for phrases which we intend to integrate into an automatic prosodic parsing system.

## ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme. The representation of percentage contours in Figure 3 was created by Johan Liljencrants.

## REFERENCES

- [1] E. Gårding and L. Eriksson. "Perceptual cues to some Swedish phrase patterns - a peak shift experiment", *STL-QPSR* 1/1989, pp. 13-16. Royal Institute of Technology, Stockholm, 1989.
- [2] E. Strangert. "Pausing in texts read aloud", *Proceedings of the Twelfth International Congress of Phonetic Sciences*, 4: pp. 238-241. Aix-en-Provence, France, 1991.
- [3] E. Selkirk. *Phonology and syntax: the relation between sound and structure*. Cambridge, Mass: the MIT Press, 1984.
- [4] M. Nespor and I. Vogel. *Prosodic phonology*. Dordrecht: Foris, 1986.
- [5] J. Pierrehumbert and M. Beckman. *Japanese Tone Structure*. Cambridge, Mass: the MIT Press, 1988.
- [6] R. Carlson, B. Granström, and L. Nord. "The KTH speech data base", *Proceedings of the ESCA workshop on speech input/output assessment*, pp. 1.3.1-1.3.4, 1989.
- [7] R. Carlson, B. Granström, and S. Hunnicutt. "Multilingual text-to-speech development and applications", in W. Ainsworth (ed.), *Advances in speech, hearing and language processing*, pp. 269-296. London: JAI Press, 1991.
- [8] G. Bruce and B. Granström. "Modelling Swedish prosody in text-to-speech: phrasing", in K. Wiik & I. Raimo (eds.), *Nordic Prosody V*, pp. 26-35. Phonetics Department, Turku University, 1990.
- [9] G. Bruce, B. Granström and D. House, "Strategies for prosodic phrasing in Swedish", *Proceedings of the Twelfth International Congress of Phonetic Sciences*, 4: pp. 182-185. Aix-en-Provence, France, 1991.
- [10] G. Bruce, B. Granström, K. Gustafson, and D. House. "Prosodic phrasing in Swedish", *Working Papers* 38, pp. 5-17. Dept. of Linguistics and Phonetics, Lund University, 1991.
- [11] G. Bruce. *Swedish word accents in sentence perspective*. Lund: Gleerup, 1977.
- [12] D. House. *Tonal perception in speech*. Lund University Press, 1990.
- [13] N. Grønnum. *The groundworks of Danish intonation: an introduction*. Museum Tusulanum Press, University of Copenhagen, 1992.
- [14] D. House and G. Bruce. "Word and focal accents in Swedish from a recognition perspective", in K. Wiik & I. Raimo (eds.), *Nordic Prosody V*, pp. 156-173. Phonetics Department, Turku University, 1990.