Prosodic Modelling of Phrasing in Swedish

Gösta Bruce*, Björn Granström**, Kjell Gustafson** and David House*
(names in alphabetical order)
*Lund University, Dept. of Linguistics and Phonetics, Helgonab. 12, S-22362 Lund
**KTH, Dept. of Speech Comm. and Music Acoustics, Box 70014, S-10044 Stockholm

ABSTRACT

In this contribution we will give a summary report on a project on prosodic phrasing in Swedish. Examples of three different methodological approaches will be given: speech analysis, speech synthesis and prosodic parsing. The results include expansions to the model of Swedish prosody.

INTRODUCTION

The starting point for our research effort within phrasing and prosody of Swedish has been our judgement that we possess a fair amount of knowledge about prominence relations and accentuation but that we know relatively little about prosodic grouping and phrasing. It is widely recognised, however, that grouping – involving the double aspect of coherence (connective) signalling and boundary (demarcative) signalling – is one of the main functions of prosody.

This contribution summarizes the work within a recently completed co-operative project between Lund University and KTH. Our primary goal has been to attain new knowledge about phrasing and prosody in Swedish. Our focus of interest here is in particular the grouping of words into prosodic phrases and domains of similar size. The problems to tackle concern questions in both phonology and phonetics. The main phonological issue is to try to understand what structure could be assumed for prosodic phrasing, particularly what types of prosodic phrase can be identified as relevant domains between a 'prosodic word' and a 'prosodic utterance'. The basic phonetic issue is to find out what speech variables (F0, duration, intensity, phonation type, pausing, etc.) and combinations of them can be used to signal phrasing. Basically three different methods are being utilised within the project: speech analysis, speech synthesis and prosodic parsing.

SPEECH PRODUCTION/ANALYSIS

The first method is the collection and analysis of speech production data. It involved the construction and recording of specially designed test material produced by one Stockholm speaker, as well as the selection and recording of suitable read text passages by several speakers (laboratory speech). The first set consisted of 22 sentences repeated three times, typically occurring as minimal pairs, where the location of the sentence internal clause boundary (here represented by a comma) was varied. These sentences were, for the most part, syntactically ambiguous, and designed to give us an idea about phrasing strategies and to enable us to easily test these strategies in the text-to-speech framework. One such pair is presented below:

När pappa fiskar, stör Piper Putte. (When daddy is fishing, Piper disturbs Putte)
När pappa fiskar stör, piper Putte. (When daddy is fishing sturgeon, Putte peeps)

Considerable variation in the signalling of phrasing was observed (Bruce, Granström and House, 1992) including: 1) the use of duration only in clause/phrase boundary signalling, 2) signalling of coherence by deaccentuation, 3) coherence signalling through the use of a "hat pattern", and 4) in longer sentences with greater syntactic complexity, the combined use of duration and F0 cues, including the absence vs. presence of F0 downstepping.

An extension of the above speech material has also been investigated within the project (Bruce, Granström, Gustafson and House, 1991). The distinctive feature of these sentences (also minimal pairs, repeated ten times) is the absence vs. presence of an internal clause boundary. An example pair is:

Lärarna backar för pojkarnas sparkar. (The teachers back away from the boys' kicks)
Lärarna backar, för pojkarna sparkar. (The teachers back away because the boys kick)

While a connective F0 downstepping was a characteristic feature of a sentence without an internal boundary, a fair degree of variation was found in the production of sentences containing a boundary. This variation can be summarised by the following strategies: 1) a boundary cue comprising marked pre-boundary lengthening preceding a small physical pause, with F0 downstepping indicating coherence, 2) the addition of a focal accent to the accent before the boundary, accompanied by a terminal F0 fall and moderate pre-boundary lengthening, and 3) an upstepping F0 pattern initially — interpretable as an extra emphasis for contrast on the first accent — and a relatively wide F0 range on the accent before the boundary accompanying moderate pre-boundary lengthening.

In most of our 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, resulting in either a grouping of 2+3 accents or 3+2 accents:

B. Fast man offrade bonden, och löparen hälsade kungen.
(But we sacrificed the pawn, and the bishop greeted the king)
L. Fast man offrade bonden och löparen, hälsade kungen.
(Though we sacrificed the pawn and the bishop, the king greeted us)

These two sentences, as well as an ambiguous version of them, were recorded three times. 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 'bonden' as a deep vs. shallow F0 valley. Interestingly, there is no corresponding F0 difference after the 3rd accent 'löparen'. The main durational difference can be seen as a pre-boundary lengthening after the 2nd and 3rd accent respectively depending on the phrasing. This sentence type was also used in a formal perception test using text-to-speech synthesis (see next section).

SPEECH SYNTHESIS AND PERCEPTION

The second method employed was 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 (Carlson, Granström and Hunnicutt, 1990). Text-to-speech synthesis was used to test the phrasing strategies observed in the production data above (Bruce, Granström and House, 1992). The default realisation of clause boundary signalling in the text-to-speech system, including a silent pause, appears to be unambiguous, but too strong and hence unnatural in many contexts. Several alternative boundary signalling strategies that also have the function of disambiguating the synthetic versions of the potentially ambiguous sentences in our speech material were identified and explored.

Some of the strategies concerning the relationship between tonal and durational cues were tested perceptually using the KTH rule synthesis where subjects can interactively vary duration and tonal parameters by moving a point on the computer screen. A sentence pair similar to sentences B and L above was used in a formal perception test where

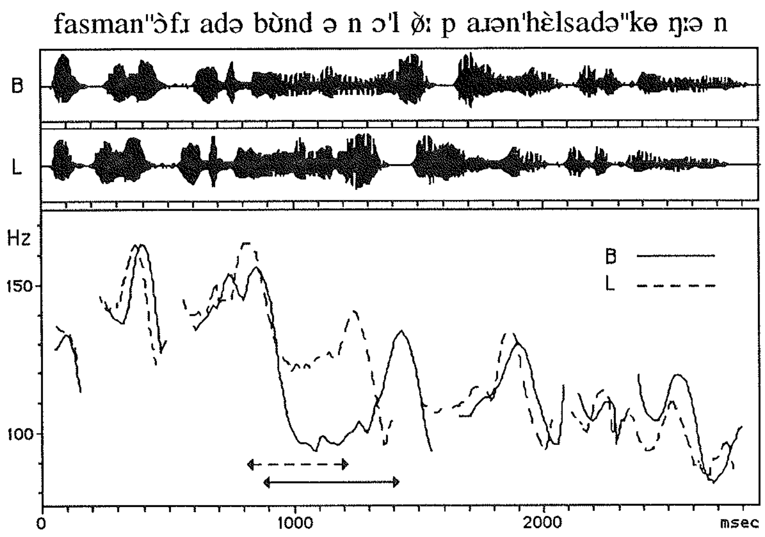

Figure 1. *Waveforms and fundamental frequency contours of the sentence "B" (solid line) and "L" (dashed line). Arrows indicate the domain used for parameter manipulation of synthetic versions in the perception experiment.*

subjects were asked to determine an optimal position for each interpretation and a line of ambiguity across the screen. Results of this test, presented in Bruce, Granström, Gustafson and House (1993a), clearly indicate that both duration and F0 are effective as phrasing signals. Although some of the individual listeners tended to have a cue preference (F0 or duration), the interaction between these two cues seems to be one of complementation rather than comprising a primary and a secondary cue. Furthermore, 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 are also consistent with our more specific observation from production data 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.

PROSODIC PARSING

The third method used in this project was prosodic parsing directed towards the recognition of phrasing. The first stage of the prosodic recognition method is the use of a human recogniser, an expert reader of an acoustic record of speech for the identification of potential phrases of an 'unknown' speech signal (cf. House and Bruce, 1990). Based on the knowledge used by the expert reader for prosodic parsing, the subsequent procedure is then to teach the computer to make an automatic analysis of prosodic phrases. Generally, we believe that the prosodic parser is particularly suitable for testing hypotheses about the interaction between different speech variables for the expression of prosodic phrasing.

The prosodic parsing experiment carried out in this project involved recording and analysing longer text passages read by two different speakers. In the experiments an expert 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. The results are then compared to two independent, auditively based transcriptions of the readings (Bruce, Granström, Gustafson and House, 1993b). There is fair agreement between the transcribers (~80%) and the expert is frequently able to identify the boundary. The mean absolute boundary locations identified across all experiments by the expert was slightly more than 70%, and by relaxing the criterion to +/- one word it increased to 85%. Sometimes the strength of the boundary is not agreed on. Results from these experiments will form the basis for the formulation of automatic recognition rules for phrases which can be integrated into an automatic prosodic parsing system.

CONCLUDING REMARKS

Our modelling of prosodic phrasing involves both coherence and boundary marking. We assume that successive prosodic words are typically grouped into prosodic phrases. This means that the prosodic phrase is related to the accentual structure. Thus specific combinations of tonal gestures for accentuation can signal coherence within a prosodic phrase. Boundary signals for a prosodic phrase can be either a separate gesture before the first accent (initial juncture) or after the last accent (terminal juncture), or they can coincide with an accentual gesture at the beginning or at the end of a phrase.

In the continuation of this project we want to widen the scope and direction of research in two different ways. Firstly we intend to cover not only the grouping function (e.g. phrasing) but also the prominence function of prosody (e.g. accentuation). We consider these to be the two main and basic functions of prosody. We specifically intend to study the interaction between phrasing and accentuation. These categories seem to be easy to separate from each other in theory but are frequently conflated in the practical situation. Secondly we would like to study how prosodic grouping and prominence are exploited in a dialogue context.

ACKNOWLEDGEMENT

The present work is in part carried out under a contract from the Swedish Language Technology Programme.

REFERENCES

G. Bruce, B. Granström and D. House (1992), "Prosodic phrasing in Swedish speech synthesis", in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoît and T.R. Sawallis (eds.), pp. 113-125. Elsevier Science Publishers B.V., Amsterdam.

G. Bruce, B. Granström, K. Gustafson and D. House (1991), "On prosodic phrasing in Swedish", *PERILUS XIII*, pp. 35-38. Inst. of Linguistics, University of Stockholm.

G. Bruce, B. Granström, K. Gustafson and D. House (1993a), "Interaction of F0 and duration in the perception of prosodic phrasing in Swedish", in *Nordic Prosody VI*, B. Granström and L. Nord (eds.), Almquist and Wiksell, Stockholm.

G. Bruce, B. Granström, K. Gustafson and D. House (1993b), "Phrasing strategies in prosodic parsing and speech synthesis", in *Proc. Eurospeech '93, European Conf on Speech Comm and Technology*, Berlin, September 21-23, 1993.

R. Carlson, B. Granström and S. Hunnicutt (1990), "Multilingual text-to-speech development and applications", in *Advances in speech, hearing and language processing*, W. Ainsworth (ed.), pp. 269-296. JAI Press, London.

D. House and G. Bruce (1990), "Word and focal accents in Swedish from a recognition perspective", in *Nordic Prosody V*, eds. K. Wiik and I. Raimo, pp. 156-173. Turku University.