

PHRASING STRATEGIES IN PROSODIC PARSING AND SPEECH SYNTHESIS

Gösta Bruce*, Björn Granström**, Kjell Gustafson** and David House*
(Names in alphabetic order)

*Department of Linguistics and Phonetics, 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 some experiments and results from a research project called Prosodic Phrasing in Swedish in which the overall aim has been to investigate and model prosodic aspects of phrasing. A series of prosodic parsing experiments are reported where a prosody expert was given the task of identifying prosodic phrases solely on the basis of a visual representation of unknown spoken text passages showing the waveform, intensity and fundamental frequency. The results were then compared to two independent, auditive based transcriptions of the readings. The comparison demonstrated a close relationship between the two types of judgments, indicating that a rather reduced acoustic representation can serve as the basis for prosodic parsing. Discrepancies encountered were often the result of an interrelationship between phrase-boundary gestures and accentual gestures. Preliminary guidelines for the prosodic parsing of phrasing are proposed. The dependence between phrasing and accentuation is further explored in a speech synthesis framework, and the influence of focal accentuation on the phrasing impression is discussed.

Keywords: Prosody, Phrasing, Parsing, Synthesis

1. INTRODUCTION

The intention of this contribution is to report on some experiments and results concerning prosodic parsing and synthesis carried out 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. The overall aim of the project has been to investigate and model prosodic aspects of grouping and phrasing. 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]).

Concerning the grouping function of prosody, our approach has been to actively look not only for boundary signals (demarcative cues) but also for coherence signals (connective cues). Therefore in our search for phonetic signals of phrasing, coherence cues are deemed to be as important as boundary cues.

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] and is 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]. The 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 [8][9]. 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, we will describe a series of preliminary experiments where a human recognizer performed the task of a prosodic recognizer. The results are compared to auditive based transcriptions. Findings from these experiments are used as a basis for the development of strategies for phrase recognition. One problem area here is the interplay between phrasing and accentuation. This has been explored in a speech synthesis framework. Informal testing demonstrates the possible dependence of focal accentuation on phrasing.

2. PARSING EXPERIMENT

Speech material

Four short informative text passages concerning coffee and spices and containing from 62 to 143 words each were read several times by two female speakers of Stockholm Swedish. The readings were recorded on tape with one fluent reading

(no hesitation pauses or repetitions) of each text by each speaker being selected for analysis. Each recording was approximately 30 to 60 seconds in duration. The recordings were analyzed using the Lupp program [10] on a Macintosh II. Visual representations of the text passages were created showing the waveform, intensity and fundamental frequency. These representations were then mounted together making one large sheet for each text passage and speaker.

Method

The visual representations of the texts were presented to an expert in prosody (one of the authors) who had no knowledge of the content of the texts. His task was to identify prosodic phrases and mark major and minor phrase boundaries. Major phrase boundaries corresponded roughly to prosodic utterances and were most often characterized by a physical pause in the speech signal. Minor boundaries corresponded roughly to prosodic phrases and were generally not characterized by a physical pause.

Independent auditory transcriptions of the texts were then carried out by another one of the authors and the expert reader. Their task was the same as above, i.e. to identify prosodic phrases and mark major and minor boundaries. Percentage agreement scores were calculated based on the discrepancy between the individual transcriptions.

The two transcriptions for each text were combined into a single key transcription containing all transcribed boundaries. Cases of transcriber disagreement between major and minor boundaries were classified as major boundaries in the key transcription. This transcription then served as a base-line for the evaluation of boundary and phrase identification by the expert reader. Percentage scores for 'boundaries identified' were calculated on the basis of discrepancies between the transcribed boundaries (total, major and minor), and the visually identified boundaries. Percentage scores for 'phrases identified' were calculated where phrase identification agreement was defined as cases where the discrepancy between the transcribed and visually identified boundary location was not greater than one word. Those cases where the difference was more than one word contributed to the score called 'boundary errors'.

Finally, several interview sessions were held with the expert reader. The objective of these sessions was to render explicit the criteria used by the reader when making the identification decisions.

Results

Results for three of the text readings are presented in Tables 1-3. Agreement between transcribers was roughly 80% with, on the average, greater agreement for major boundaries than for minor boundaries. Transcribed boundaries also differed between the two speakers as seen in Tables 2 and 3.

Table 1. Identification results for text 1, speaker 1.

Type of boundary/phrase:	total	major	minor
Number of transcribed boundaries (key transcription)	43	17	26
Transcribers' agreement	91%	82%	85%
Boundaries identified	63%	94%	35%
Phrases identified	77%	94%	54%
Boundary errors	7%	6%	12%

Table 2. Identification results for text 2, speaker 1.

Type of boundary/phrase:	total	major	minor
Number of transcribed boundaries (key transcription)	19	12	7
Transcribers' agreement	84%	75%	57%
Boundaries identified	79%	83%	57%
Phrases identified	89%	83%	71%
Boundary errors	16%	0%	16%

Table 3. Identification results for text 2, speaker 2.

Type of boundary/phrase:	total	major	minor
Number of transcribed boundaries (key transcription)	23	9	14
Transcribers' agreement	74%	78%	64%
Boundaries identified	78%	100%	60%
Phrases identified	87%	100%	73%
Boundary errors	22%	0%	33%

The mean absolute boundary locations identified across all experiments by the expert was slightly more than 70% for all boundaries while phrase identification reached nearly 85%. Major boundaries received a considerably higher identification score than did minor boundaries. Results for all three texts were fairly uniform with the exception of a low (35%) identification score for minor boundaries in text 1 (Table 1).

The interview sessions revealed the following main criteria for the recognition of phrasing. A peak or a plateau associated with what appears to the expert to be a combination of focal and first post-focal accents provides a strong primary cue for coherence. A second coherence cue is downstepping for post-focal accents.

A strong boundary cue is provided by an extra low tonal level at the boundary followed by the resetting of F0 after the boundary. An additional boundary cue is a fall from a peak followed by a rise to a new peak where both peaks are regarded as focal peaks by the expert. Major boundaries were almost always accompanied by a physical pause.

Boundary errors and missed boundaries were mainly the result of accentual gestures being interpreted as boundaries and vice-versa.

Discussion

In general, the results of the visual recognition experiments demonstrate that considerable phrasing information is available in a rather reduced acoustic representation of the speech signal. The criteria used for recognition also confirm the results of our previous production and perception studies using sets of syntactically ambiguous sentences comprising minimal pairs [11][12][13].

The visual representation and hence the criteria used for recognition was essentially tonal in nature with the exception of the major boundary pauses. For automatic recognition purposes, durational and spectral cues would probably improve the recognition score, especially for the minor phrases.

The division of the texts into minor phrases is, however, not a trivial question as evidenced by the discrepancies between the auditory transcriptions especially in text 2 (Tables 2 and 3). In many cases the disagreement was not so

much over what constituted a boundary as to whether the boundary was a major phrase or a minor phrase boundary. Additional work is planned in this area related to work on the transcription of break indices in English [14].

The fact that phrase identification received higher scores than did boundary identification points to the importance of coherence cues. In several instances where the absolute location of the boundary was not identified by the expert reader, coherence cues provided information enabling the reader to identify the phrase. These results give rise to the concept of the boundary area. In cases where explicit boundary cues were missing or weakened due to, for example, the influence of an accentual constraint, coherence cues could be used to point to a general boundary area.

Finally, the errors resulting from the interplay of accentual and phrasal gestures indicate an area of importance in understanding the nature of phrasing in Swedish. In the following section, speech synthesis is used to explore this area.

3. SYNTHESIS: PHRASING AND ACCENTUATION

Interplay between phrasing and accentuation

Our experience from prosodic parsing experiments in Swedish thus shows that there is a possible confusion risk between accentual gestures and phrase boundary gestures. It seems theoretically uncontroversial to assume, however, that the accentuation of words as an expression of their salience is a distinct phonetic category, separate from the corresponding grouping of words into phrases (phrasing). While in real-life situations we can occasionally find examples where phrasing is independent of accentuation, it is probably much more often the case that phrasing and accentuation are at least partly interdependent.

One possible, extreme situation of the relationship between phrasing and accentuation - one where phrasing can be varied without a necessary, concomitant change in accentuation - is illustrated by the test paradigm used in a recent perceptual experiment [12][13]. The current IPA symbolization, where [°] means a word in focus, has been used:

(1a) Fast man offrade bonden, och löparen hotade kungen.
[fast man °ɔ̃fradə 'bɔ̃ndən || ɔ̃ 'lɔ̃parən 'hɔ̃tadə 'kɔ̃ŋən]
(but we sacrificed the pawn, and the bishop threatened the king)

(1b) Fast man offrade bonden och löparen, hotade kungen.
[fast man °ɔ̃fradə 'bɔ̃ndən ɔ̃ 'lɔ̃parən || 'hɔ̃tadə 'kɔ̃ŋən]
(though we sacrificed the pawn and the bishop, the king threatened us)

The results of our synthetic testing show that the signalling of coherence or boundary between the two words at the potential boundary by means of F0 and duration in combination can be done without necessarily affecting the accentuation (prominence level) of these words.

At the other extreme, we can identify situations where phrasing and accentuation are typically interdependent. One case in point, where phrasing is largely dependent on accentuation, is the use of deaccentuation of a word leading to a unit accentuation (and thus to a particular phrasing) and

giving room for only one interpretation (2b) in a sentence pair like the following:

(2a) När pappa fiskar, stör Piper Putte.
[næɪ 'pɑ̃pɑ̃ 'fɪ skɑ̃ || 'stœɪ 'pɪpəɪ 'pø̃tœ]
(When daddy is fishing, Piper disturbs Putte.)

(2b) När pappa fiskar stör, piper Putte.
[næɪ 'pɑ̃pɑ̃ 'fɪskɑ̃ 'stœɪ || 'pɪpəɪ 'pø̃tœ]
(When daddy is fishing sturgeon, Putte peeps.)

Another kind of dependence between phrasing and accentuation is exemplified by tonal gestures having a double function, i.e. both for phrasing and accentuation. Below are shown the main features of the intonation model implemented in our text-to-speech synthesis for Swedish:

Table 4. Features of the intonation model

CATEGORY	TURNING POINTS
unaccented	-
accent I	H L*
accent II	H* L
focal accent I	(H) L* H
focal accent II	H* L H
focal accent II compound	H* L ... L* H
initial juncture	L
terminal juncture	L

The implication of this modelling of intonation is that we can identify a few cases of combined tonal gestures for phrasing and accentuation. While the default case would be separate tonal gestures for the signalling of phrase boundaries and accentuation, the following examples of tonal gestures carrying both functions are not uncommon. A phrase where the initial syllable has focal accent I typically starts with a pitch rise (LH) serving the double function of a focal accent and an initial juncture. Correspondingly, a phrase ending for example in a non-focal accent II will be characterized by a pitch fall (H*L) carrying the functions of both word accent and terminal juncture (cf. [15]).

Focal accentuation and phrasing

Against this background it is reasonable to ask whether it is possible to give the impression of a particular phrasing with a specific combination of accentuations only, i.e. without explicit boundary signalling such as pre-boundary lengthening and fall to a bottom F0 level. One possible hypothesis would then be that focal accent (under certain conditions) can be perceived as also being phrase initial. In an informal test with synthetic speech (with no explicit, internal phrase boundary signalling) using the test paradigm as in (1) above, the following variations were performed. A synthetic version of this test utterance with 'löparen' in focus (but with no focus on 'kungen') gave a clear phrasing impression with a boundary before the focal word ('löparen') (cf. Figure 1A).

Another version of the same test utterance with instead 'hotade' in focus (but otherwise identical to the first version) gave, contrary to our hypothesis, the same phrasing impression (boundary before 'löparen') as in the first version (cf. Figure 1B). The reason for this somewhat unexpected phrasing impression may be found in the pitch

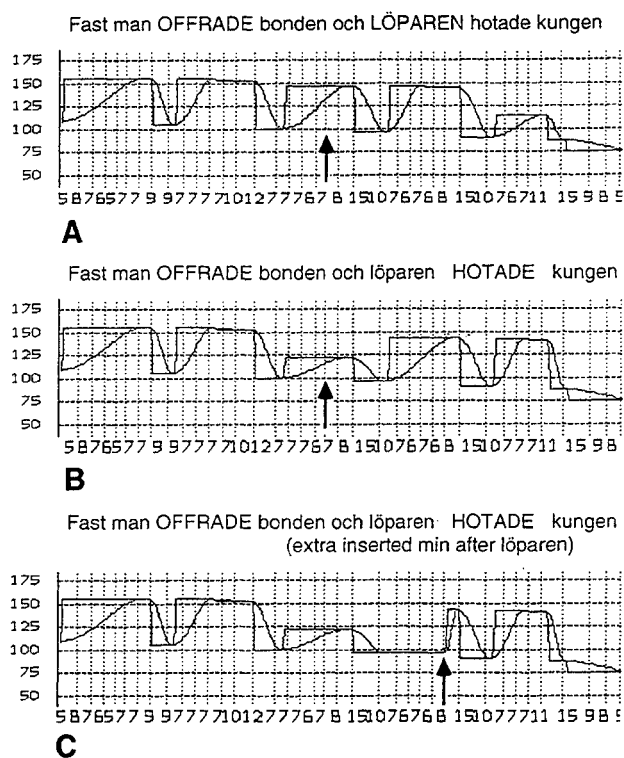


Figure 1. Synthesis parameter plots for the utterance used in the synthesis testing. Arrows indicate perceived boundary locations. Capitals indicate focal accent. Segment durations in csec. are indicated on the X-axis. The Y-axis is F0 in Hz.

concatenation in the non-focal word 'löparen': a relatively early timed, slow rise from L to H*.

In a third version of the test utterance with the same focussing as in the second version ('hotade' in focus) but instead with a later timed, abrupt F0 rise, thus giving an extended word accent L in 'löparen', the impression of phrasing is clearly altered to a boundary before 'hotade' (cf. Figure 1C). We take this informal synthetic testing as preliminary evidence for the potentially important role of concatenation for coherence or boundary signalling, which should be further explored in a more formal perceptual experiment.

In another parallel test sequence containing the same three synthetic versions of the test utterance, the accent I word 'damen' ['dæmən] 'the queen' was substituted for the accent II word 'löparen', as a variation on the same theme. The phrasing impression in the three new versions (with an accent I word in the critical position) remained the same as in the versions having an accent II word instead.

4. CONCLUSIONS

The preliminary results of the parsing experiments indicate the usefulness of both tonal boundary cues and coherence cues for recognition of phrasing. Furthermore, these results and those of the synthesis tests demonstrate the importance of an understanding of the interrelationship between accentual and phrase boundary gestures for both recognition and synthesis.

The findings also point to important areas of further research. For prosodic parsing and recognition, durational and spectral cues should also be exploited. More rigid criteria for the identification of major and minor boundaries and tests for intertranscriber reliability will aid in the evaluation of prosodic recognition.

Finally, by extending our speech data to a dialogue context, we intend to increase our understanding of how prosodic aspects of speech are used in a communicative situation. This increased knowledge will enable us to create a more comprehensive prosody model.

ACKNOWLEDGEMENT

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK).

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] 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.
- [9] D. House. *Tonal perception in speech*. Lund University Press. 1990.
- [10] L. Eriksson. "New phonetic programs for Macintosh", *Working Papers* 36, pp. 73-80. Dept. of Linguistics and Phonetics, Lund University. 1990.
- [11] 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.
- [12] G. Bruce, B. Granström, K. Gustafson, and D. House. "Aspects of prosodic phrasing in Swedish", *Proceedings ICSLP 92*, 1: pp. 109-112, Banff, Canada. 1992.
- [13] G. Bruce, B. Granström, K. Gustafson, and D. House. "Interaction of F0 and duration in the perception of prosodic phrasing in Swedish", *Proceedings Nordic Prosody 6*. KTH, Stockholm. In press.
- [14] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert and J. Hirschberg. "TOBI: a standard for labeling English prosody", *Proceedings ICSLP 92*, 2: pp. 867-870, Banff, Canada. 1992.
- [15] G. Bruce. *Swedish word accents in sentence perspective*. Lund: Gleerup. 1977.