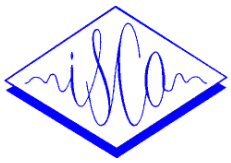

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

MODELLING SWEDISH PROSODY IN A DIALOGUE FRAMEWORK

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

*Dept of Linguistics and Phonetics, Helgonabacken 12, S-22362 Lund, Sweden
**Dept of Speech Comm. and Music Acoustics KTH, Box 70014, S-10044 Stockholm, Sweden

ABSTRACT

The research reported here is conducted within the recently initiated project 'Prosodic Segmentation and Structuring of Dialogue'. The object of study in the project is the prosody of dialogue in a language technology framework. The specific goal of our research is to increase our understanding of how the prosodic aspects of speech are exploited interactively in dialogue – the genuine environment for prosody – and on the basis of this increased knowledge to be able to create a more powerful prosody model. In this paper we present an overview of project design and methods.

INTRODUCTION

The vast majority of phonetic studies of prosody have until quite recently been centered upon relatively stereotypic settings in the phonetics laboratory, so-called laboratory speech. In this type of speech material experimental control is high, as relevant parameters can be varied and studied systematically, while the degree of naturalness is often instead quite low. The construction of prosody models currently being used in text-to-speech systems is typically based on the analysis of prosody from such laboratory speech material. Even today there exist few phonetic studies of the prosody of spontaneous speech and dialogue, i.e. the kind of context where prosody has its main function and use. The reason for this bias is to be found in the relative complexity of prosody. Spontaneous speech and dialogue offer such a richness of prosodic variation that its study can be said to presuppose a fundamental understanding of prosody in the more controlled context of laboratory speech.

The object of study in the recently initiated project *Prosodic Segmentation and Structuring of Dialogue* is the prosody of dialogue in a language technology framework [1]. The project represents cooperation between Phonetics at Lund University and Speech Communication at KTH, Stockholm and is part of the Swedish Language Technology Programme. Related projects within the Language Technology framework are *Intonation in Restrictive Texts: Modelling and Synthesis* [2], [3], *Interaction in Speech between Prosody, Syntax, Semantics and Pragmatics* [4], [5] and also *Language Technology for Spoken Dialogue Systems* [6], [7]. The focus of our present contribution will be on research methodology.

BACKGROUND

Research within our project *Prosodic Segmentation and Structuring of Dialogue* is based on earlier work on prosody from different perspectives. One starting point is research conducted within the project *Contrastive Interactive Prosody* (KIPROS) at Lund supported by the Bank of Sweden between 1988-90. The object of study of KIPROS was dialogue prosody in a contrastive perspective in French, Greek and Swedish. We conducted three types of analysis: analysis of dialogue structure, auditory (prosodic) analysis, and acoustic-phonetic analysis. This project was our first large-scale confrontation with spontaneous speech and dialogue and comprised exploratory testing of the prosody model which was based on experience from extensive work with laboratory speech (see also [8]). The focus of the KIPROS project was largely on methodology, which resulted in the development of tools and conventions for prosodic transcription of Swedish and French [9], [10]. Experience from the project also made apparent the main difficulties involved in analyzing spontaneous speech where experimental control is low.

The second point of departure for the current project is work carried out within the project *Prosodic Phrasing in Swedish* which was also a joint effort between Phonetics in Lund and Speech Communication at KTH, Stockholm, and part of the Language Technology Programme 1990-93. Our cooperation relates to two different research traditions: work in Lund aimed at developing a model for Swedish prosody and work in Stockholm directed towards the development of the prosodic component of a text-to-speech system. The main orientation of this project was directed towards studying how prosody signals phrasing, i.e. grouping of words into phrases. The Prosodic Phrasing project represented a return to the phonetics laboratory and more controlled conditions in the form of analyses of read speech [11], [12], [13].

GOAL AND METHODOLOGY

The primary goal of the new project is to increase our understanding of how the prosodic aspects of speech are exploited interactively in dialogue - the genuine environment for prosody - and on the basis of this increased knowledge to be able to create a more powerful prosody model. To be able to achieve this goal the following methodology is being employed:

- analysis of dialogue structure (independent of prosody)
- auditory analysis in the form of prosodic transcription
- acoustic-phonetic analysis (based on F0 and waveform information)
- speech synthesis (text-to-speech)

We are exploiting speech material from the national Swedish prosodic database under development. The dialogues under study cover true spontaneous conversation, spontaneous but more restricted and well controlled dialogues, as well as acted dialogues from scripts. Artificially spliced dialogues, dialogues simulated using text-to-speech synthesis and man-machine dialogues (cf. [6], [7]) are also exploited in our study of dialogue prosody.

An important methodological starting point of our work is to initially consider prosody and dialogue as potentially independent. This means that we consider it possible to first make separate analyses of prosodic categories and dialogue structuring. Only later are the prosodic analysis and the analysis of dialogue structure combined in order to find potentially interesting connections. Therefore, we do not a priori anticipate that a particular question intonation is always used by the speaker taking a strong initiative in a conversation, or that the introduction of a new conversation topic is necessarily signalled prosodically by the speaker.

ANALYSIS OF DIALOGUE STRUCTURE

The ultimate goal for prosody research within language technology is to be able to combine phonetic knowledge about prosody with linguistic and other contextual information. It is therefore important that the analysis of the dialogue structure itself is carried out independently of prosody. We have been working with three basic, interactive dimensions, namely textual aspects, turn regulating aspects and aspects of initiative/response structure.

The *textual* aspect concerns division into conversation topics involving grouping into 'speech paragraphs' [14]. This applies to discourse both in the form of dialogue and monologue. It is clear that prosody plays an important role in signalling topic structure, even if different studies show different types of relationships.

Aspects of *initiative/response* structure, i.e. traditionally questions and answers, concern the contribution of the speakers to the development of the dialogue through taking or refraining from taking initiative, responding to initiatives and making reference to what has been said (cf. [15]). Prosody plays an important role here, although it is clear that there is a considerable degree of freedom in the way that it is used to signal this aspect of dialogue structure.

The *turn regulating* aspect involves e.g. taking, keeping, yielding and competing for the floor in a dialogue (cf. [16]), which may be partly overlapping with initiative / response structure but still potentially distinct. It is apparent that this aspect is signalled by different means (verbal, non-verbal, prosodic). The exact contribution of prosody here is not fully understood.

In addition to the above, there is also a *feedback* dimension, indicating the way in which speakers give and seek feedback in a dialogue. Feedback giving (back-channelling) is often noted in dialogue studies while the speaker's feedback seeking (seeking feedback from the listener) has not been given as much attention. It is possible that the feedback dimension can be seen as a subdivision of the initiative/response structure, although we have chosen to regard it as a separate dimension for the present time. We believe that prosody plays an important role in signalling both feedback giving and seeking.

Other interactive dimensions which can easily be expressed prosodically are the signaling of *attitudes / emotions* and *rhetoric activity* [17].

AUDITORY ANALYSIS

An independent auditory analysis of prosody is made in the form of a prosodic transcription. This transcription is tied to the orthographic representation of the dialogue and thus contains symbolization of selected prosodic categories.

We have witnessed a marked increase in interest in transcription, including prosodic transcription, during the last five year period. One important reason for this newborn interest arises from new needs for annotation of large speech databases. A starting point was the 1989 IPA Convention in Kiel for the revision of the International Phonetic Alphabet, the first substantial revision in 40 years. The new version of the IPA (cf. [18]) does not, however, contain any specific symbolization of discourse prosody.

Another example of this transcription wave is ToBI (Tones and Break Indices), a system which has recently been developed for the prosodic transcription of American English [19]. This transcription system provides symbols mainly for prominence and grouping. An innovation in ToBI is the combined auditory and acoustic (F0, waveform) analysis, where both types of information are integrated in the prosodic notation.

Unlike ToBI we have chosen to rely on a purely auditory analysis. Our transcription is intended to be phonological rather than a narrow phonetic transcription. The prosodic transcription that was developed within the KIPROS project consists of symbols for the following prosodic categories: prominence, grouping, pausing, pitch range and boundary tones. There are, however, other potentially interesting categories such as voice intensity, voice quality and speech tempo which have not been included. In the KIPROS transcription system IPA symbols for prominence, grouping and pausing are used as well as special symbols for pitch range and boundary tones. IPA symbolization of prominence, grouping and pausing is abstract but well established and relatively simple, while pitch range and boundary tones are represented by more iconic and transparent symbols (cf.[9]).

Prosodic base transcription (IPA)

SYMBOL	CATEGORY	
	prominence	
(no symbol)	unstressed	(= no prominence)
ˌcv	stressed [no distinct. acc. I / II]	(= weak prominence)
'cv	accented [accent I]	(= strong prominence)
'cv̀	accented [accent II]	
"cv	focussed [accent I]	(= extra strong prominence)
"cv̀	focussed [accent II]	
	grouping	
(no symbol)	coherence	(e.g. phrase internal)
cv | cv	weak boundary	(e.g. prosodic phrase)
cv || cv	clear boundary	(e.g. pros. utterance)
cv ||| cv	extra clear boundary	(e.g. speech paragraph)

('cv' represents any syllable ['c' = consonant,'v' = vowel])

A vital issue for the construction of a national Swedish prosody database within the Language Technology framework is the choice of prosodic transcription conventions. Discussions of this issue have resulted in an agreement whereby a base module for prosodic transcription based on the IPA comprises a standard for the phonological symbolization of the categories prominence and grouping as shown above (see further [20]).

In addition to this base module different prosody projects within Language Technology are expected to create their own modules according to existing needs. In our new Prosodic Segmentation project we intend to add a module containing symbols from the KIPROS transcription system. Moreover we have also begun development of a module for tonal analysis using notation not unlike ToBI.

ACOUSTIC-PHONETIC ANALYSIS

Our acoustic-phonetic analysis comprises standard F0 extraction and spectral information in addition to the speech waveform. The analysis is carried out in the ESPS/Waves environment which includes transcription and labeling in multiple tiers [21]. This enables an automatic processing of possible relationships between, for example, prosodic and discourse categories.

An important part of the analysis of F0 is the intonation model which has been developed from extensive studies of laboratory speech (cf. [22], [23]). The intonation model involves categorization with respect to accentuation (prominence) and phrasing (grouping), including boundary signalling and other intonation features. The categories are expressed using tonal turning points (H / L) with association to stressed syllables or boundaries. The main features of the intonation model are given in the following table.

Representation of prosodic categories in the intonation model

PROSODIC CATEGORY	TONAL TURNING POINTS
unaccented	-
accent I	HL*
accent II	H*L
focussed accent I	HL* H
focussed accent II	H*L H
focussed accent II (compound)	H*L...L* H
initial juncture	L / H
terminal juncture	L / H

Accent I and accent II are critically timed in relation to foot boundaries, i.e. stressed syllables. In our analysis the two word accents appear to have a distinctively different timing of the same accentual gesture (H(igh), L(ow)) relative to the stressed syllable, accent I being timed earlier than accent II. Thus accent as a higher prominence level than just stress is cued mainly by pitch, although an accented foot is usually also longer than an unaccented foot.

An important grammatical as well as prosodic distinction in Swedish is the one between simplex and compound words. A compound consists of (at least) two feet (stress groups), where only the first foot is accented, while a simplex word consists of only a single foot (stress group). While focal accentuation is primarily determined by semantics and pragmatics (given/new), focal accent is typically also a default choice for a word in a phrase final position. Phonetically, focal accentuation is marked by a more complex accentual gesture, an extra H after the HL for (word) accent. The focal accent H is executed in the same foot (stress group) as the accent HL for a simplex word, while it occurs in the final foot of a compound word. This extra pitch prominence is usually accompanied by increased duration of the word in focus.

Generally, the initial juncture (boundary signal) of a prosodic phrase involves a LH gesture. This LH gesture can be either a separate gesture before the first accent of the phrase or coincide with an accentual gesture, e.g. with the LH of an initial, focal accent. The terminal juncture (boundary signal) of a phrase instead involves a HL gesture. Correspondingly, this HL gesture can be either a separate gesture, e.g. after a phrase final focal accent, or coincide with the HL of a (non-focal) accent gesture. In longer phrases (with more than two accented words), two post-focal accents within the same phrase will typically occur in a downstep, i.e. the terminal HL gesture can be regarded as being executed in two successive steps, while instead two pre-focal accents of a phrase are characterized by the absence of downstep. This tonal signalling of coherence and boundaries for phrasing is also accompanied by temporal signalling as well as by other correlates.

SPEECH SYNTHESIS

In previous studies we have mainly been focussing on duration and fundamental frequency as correlates of prosodic structure. This is not in neglect of other correlates, but only reflects the general belief that these two dimensions comprise the most common and robust cues. Other glottal adjustments, resulting in voice source changes such as spectral tilt, amplitude, irregular voicing and aspiration could all be part of signalling a prosodic function. With the new GLOVE synthesiser it is now possible to model many voice source characteristics [24], but this has been utilized only to a limited extent [25]. General phonetic reductions could also serve a prosodic function, especially in relation to prominence. The application or lack of application of phonological rules have been observed as a cue to phrasing in our earlier studies [26]. The new possibilities will be used in the present project.

In the text-to-speech system we will exploit the many ways of interacting with rules and parameters [27]. It is possible to use the entire system or to run separate rule components. Most of the project work will be done on experimental versions of the phonetic rule component. In this case the input text is not restricted to regular phonetic transcriptions but could also be symbols indicating e.g. syntactic structure and degree of emphasis or symbols triggering experimental rules, i.e. information that could be supplied by the higher levels of a dialogue system. There are special facilities to interactively change rule variables. For example, the location and height of a fundamental frequency peak can be modified during rule execution. Another possibility is to display and modify parameters after rule execution but before parameter interpolation and synthesis. On this level it is possible to refer to a natural production by superimposing a computer generated spectrogram on the parameter display. In this way the synthesis by analysis method and the rule synthesis method are combined in one research environment. Another interesting way in which we can use the combined speech analysis/synthesis environment is to use the label files, with phonetic transcriptions and segment durations, as input to the phonetic component of the text-to-speech program. In this way we can study effects of other prosodic rules selectively. Coherence in the prosodic expression will be

one objective of the more comprehensive model proposed in this project.

ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK). Gayle Ayers, Dept of Linguistics, Ohio State University was a guest researcher in Lund (autumn 1993) and has contributed to the project.

REFERENCES

[1] G. Bruce, B. Granström, K. Gustafson, D. House & P. Touati, "Preliminary report from the project 'Prosodic segmentation and structuring of dialogue'",*Working Papers 43, Fonetik -94,* pp. 34-37. Department of Linguistics, Lund University, 1994.

[2] M. Horne, M. Filipsson, C. Johansson, M. Ljungqvist & A. Lindström,(compound) "Improving the prosody in TTS systems: Morphological and lexical-semantic methods for tracking "new" vs. "given" information", *Working Papers 41, Proceedings of an ESCA workshop on prosody*, pp. 208-211. Department of Linguistics, Lund University, 1993.

[3] M. Horne, "Generating prosodic structure for synthesis of Swedish intonation", *Working Papers 43, Fonetik -94,* pp. 72-75. Department of Linguistics, Lund University, 1994.

[4] E. Strangert, E. Ejerhed & D. Huber, "Clause structure and prosodic segmentation", *RUUL 23, Fonetik -93,* 81-84. Department of Linguistics, Uppsala University, 1993.

[5] E. Strangert & M. Heldner, "Prosodic labelling and acoustic data", *Working Papers 43, Fonetik -94,* pp. 120-123. Department of Linguistics, Lund University, 1994.

[6] M. Blomberg, R. Carlson, K. Elenius, B. Granström, S. Hunnicutt, R. Lindell & L. Neovius, "An experimental dialogue system: Waxholm", *RUUL 23, Fonetik -93*, pp. 49-52. Department of Linguistics, Uppsala University, 1993.

[7] R. Carlson & S. Hunnicutt, "Dialog management in the Waxholm system", *Working Papers 43, Fonetik -94,* pp. 46-49. Department of Linguistics, Lund University, 1994.

[8] E. Gårding, "Prosodiska drag i spontant och uppläst tal", In G. Holm (ed.), *Svenskt talspråk*, pp. 40-85. Uppsala: Almqvist & Wiksell, 1967.

[9] G. Bruce & P. Touati, "On the analysis of prosody in spontaneous dialogue", *Working Papers* 36, pp. 37-55. Department of Linguistics, Lund University, 1990.

[10] G. Bruce & P. Touati, "On the analysis of prosody in spontaneous speech with exemplifications from Swedish and French", *Speech Communication* 11, pp. 453-458, 1992.

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

[12] G. Bruce, B. Granström, K. Gustafson & D. House, "Interaction of F0 and duration in the perception of prosodic phrasing in Swedish", In B. Granström & L. Nord (eds.), *Nordic Prosody VI*, pp. 7-22. Stockholm: Almquist & Wiksell International, 1993a.

[13] G. Bruce, B. Granström, K. Gustafson & D. House, "Prosodic modelling of phrasing in Swedish", *Working Papers 41, Proceedings of an ESCA workshop on prosody*, pp. 180-183. Department of Linguistics, Lund University, 1993b.

[14] G. Brown, K. Currie & J. Kenworthy, *Questions of intonation*, London: Croom Helm, 1980.

[15] P. Linell & L. Gustavsson, *Initiativ och respons. Om dialogens dynamik, dominans och koherens*, Studies in Communication no. 15. University of Linköping, 1987.

[16] A. Cutler & M. Pearson, "On the analysis of prosodic turn-taking cues", In C. Johns-Lewis (ed.), *Intonation in Discourse*, pp. 139-155. London: Croom Helm, 1986.

[17] P. Touati, "Overall pitch and direct quote-comment structure in French political rhetoric", *RUUL 23, Fonetik -93*, pp. 98-101. Department of Linguistics, Uppsala University, 1993.

[18] I.P.A, "Report on the 1989 Kiel convention", *Journal of the International Phonetic Association* 19 (2), pp. 67-80, 1989.

[19] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, P. Price, J. Pierrehumbert & J. Hirschberg, "TOBI: A standard for labeling English prosody", In *Proceedings of the Second International Conference on Spoken Language Processings* 2, pp. 867-870. Banff, Canada, 1992.

[20] G. Bruce, "Prosodisk strukturering i dialog", To appear in *Svenskans Beskrivning 20.* Umeå University, 1994.

[21] G. Ayers, "Discourse functions of pitch range in spontaneous and read speech", To appear in *Working Papers from the Dept of Linguistics.* Ohio State University, 1994.

[22] G. Bruce *Swedish word accents in sentence perspective*, Lund: Gleerups, 1977.

[23] G. Bruce, & B. Granström. "Prosodic modelling in Swedish speech synthesis", *Speech Communication* 13, pp. 63-73, 1993.

[24] R. Carlson, B. Granström & I. Karlsson, "Experiments with voice modelling in speech synthesis", *Speech Communication,* Vol. 10, pp. 481-490, 1990.

[25] B. Granström & L. Nord, "Neglected dimensions in speech synthesis", *Speech Communication* 11, pp. 459-462, 1992.

[26] G. Bruce & B. Granström, "Modelling Swedish intonation in a text-to-speech system", *Proc. of Fonetik-89*, STL-QPSR 1/1989, 17-21. Department of Speech Communication and Music Acoustics, KTH, Stockholm, 1989.

[27] R. Carlson, B. Granström & 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.