MODELLING SWEDISH INTONATION FOR READ AND SPONTANEOUS SPEECH

Gösta Bruce
Department of Linguistics and Phonetics, Lund, Sweden

ABSTRACT

Our approach to the study of dialogue prosody involves analysis of dialogue structure, prosodic analysis (auditory and acoustic-phonetic) and synthesis. A comparison of spontaneous and read speech shows variation in global pitch level and range to reflect topic structure more stereotypically in read speech, and variation in local pitch range on a focussed item to be a signal of feedback seeking and interaction in spontaneous speech.

INTRODUCTION

The focus of attention of our current prosody research is the prosody of spontaneous speech and dialogue. The framework for this prosody research is the project *Prosodic segmentation and structuring of dialogue* centered around the description of Swedish [1]. This project is a cooperation between phonetics at Lund and speech communication at KTH, Stockholm with support within the second phase of the Swedish Language Technology Programme 1993 - 1996.

The ultimate goal of the project is to create a more powerful prosody model, in particular for the description of intonational patterns of spontaneous speech and dialogue. In order to achieve this goal we will have to increase our understanding of the structuring of spontaneous dialogue, particularly how intonation contributes to the development of such a dialogue.

Background

Our current prosody model is based on two decades of prosody research within the project group with experience mainly from read, laboratory speech. Thus our modelling of intonation in Swedish is typically based on the analysis of speech material elicited from informants who have been asked to perform specific tasks in a phonetics laboratory environment.

The intention behind this laboratory setting has not been the study of the prosody of reading or even less so elocution as a particular speaking style. Instead the idea has been to create a situation simulating natural speech to a reasonable extent while taking advantage of the laboratory setting with its high degree of experimental control and the possibility of varying test parameters one at a time.

The drawback is the somewhat artificial situation with albeit phonetically well balanced speech material (particularly in terms of microprosody) but often semantically unusual utterances and pragmatically strange situations. These constraints will put fairly heavy demands on the informants and their acting skills. Even if the informant turns out to be a good actor, the recorded utterance will clearly be an instance of prepared, read speech and will probably be recognized as such, even if the more or less pronounced non-sense character of the utterance is disregarded and only its prosody is considered.

A case in point is our way of eliciting phonetic prominence (focus) in different positions of a test utterance. The method exploited is a question - answer paradigm, where a variable context utterance (Question) is used to elicit a particular word in focus (or a whole phrase) of a fixed response utterance (Answer) [2].

Example
Question: Vad vill man lämna för några nunnor? (What nuns do they want to leave?)
Answer: Man vill lämna några LÅNGA nunnor. (They want to leave some TALL nuns.)

By the use of sonorant consonants and vowels of approximately the same degree of opening, differences in F0 that may be due to microprosodic variation are neutralized. Such a test paradigm presents a kind of simulation of a small portion of a very simple dialogue, where the informant has been instructed to perform both the roles of asking and answering.

Until fairly recently there have been relatively few phonetic studies of prosody in spontaneous speech and dialogue, i.e. the kind of situation where prosody has its proper function and usage. The reason for this state of affairs is to be looked for in the relative complexity of prosody. Spontaneous speech and dialogue offer such a richness of prosodic variation that its study seems to require a basic understanding of prosody in the more controlled context of laboratory speech.

THE PROSODY MODEL

The fundamental question addressed in my paper is: what kind of differences in terms of intonation and pitch patterns are typically found when we compare natural, spontaneous speech with prepared, read speech of the type referred to above?

In order to be able to find an adequate answer to this question, we will have to understand what counts as a difference, i.e. we will have to devise some kind of measure of variability. An instrument that may help us to this effect is our prosody model which has been implemented in a text-to-speech system for the generation of synthetic F0 contours and timing patterns [3]. Thus one reference for our current research effort within dialogue prosody and spontaneous speech is the Swedish prosody model, based on experience from prepared speech in the phonetics laboratory simulating natural speech as described above.

A question related to the one raised above will then be: how suitable is the intonation model which has been built upon our knowledge about prepared, laboratory speech also for the description of natural, spontaneous speech, and further how could the model be accommodated and elaborated to encompass the pitch patterns of spontaneous dialogue?

A fundamental assumption behind our modelling of prosody is that prosody can express a number of different, communicative functions, although the relationship between a certain function and its phonetic expression is typically indirect and quite complex.

Our modelling of the intonational structuring has been particularly concerned with the basic functions: grouping (signalling of coherence and boundary) into prosodic words, compounds, phrases, utterances and paragraphs, and prominence (foregrounding and backgrounding) of words and phrases, as well as their interaction [3]. A basic component of the intonational model is the tonal inventory in terms of tonal turning points (H, L, their combinations and diacritic symbols for alignment with prominences and boundaries) used as a phonological / abstract phonetic notation for both prominence relations and grouping. The input string to the model is then a tonal transcription (symbolic notation) on which the phonetic implementation rules operate to create the output F0 contour.

RESEARCH METHODOLOGY

One fundamental distinction between prepared, read speech and unrehearsed, spontaneous speech is the amount of planning involved. The on-line planning of spontaneous as opposed to the completely preplanned, read speech will typically result in well known prosodic differences such as the number and distribution of pauses, more variation in speech tempo, voice quality and voice intensity as well as repetitions, false starts and corrections characteristic of spontaneous speech. What effects this difference in planning has specifically on intonation and choice of pitch patterns is less clear, however.

A true spontaneous dialogue is often described as an act of negotiation between the two (or more) interlocutors. Important aspects of the structure of the dialogue are at least the following: textual aspects [topic structure, semantic focus], initiative / response structure, feedback seeking [are you with me?, do you see what I mean?] and giving ['mm', 'yeah'] as well as turn regulating [keeping, yielding, taking and struggling for the turn]. Our analysis of the structure of a dialogue attempts to take these aspects

into account as well as other aspects like signalling of attitudes and rhetoric activity, which seem to be represented in all kinds of speech to a varying degree.

An important starting point for our work is to initially regard dialogue and prosody as independent. This means that we assume that it is convenient at first to make an analysis of the structuring of dialogue and a corresponding analysis of prosodic categories. Only then is a coupling made between the prosodic analysis and the dialogue analysis which enables the establish-ment of possible, interesting inter-connections and correlations. Therefore, we do not a priori assume that there would be, for example, a special question intonation obligatorily used by a speaker taking a strong initiative in a conversation, or that the introduction of a new conversation topic necessarily needs to be signalled prosodically.

Our prosodic analysis is divided into an auditory analysis in the form of a prosodic transcription and an acoustic-phonetic analysis. The prosodic transcription relevant here is a broad, phonetic analysis containing symbolization of both prominence and grouping. It consists of two parts. The first part is the tonal tier for the notation of two levels of prominence (accented, focussed), including the lexically determined distinction between the two word accents in Swedish, as well as junctures (initial and terminal boundary tones). The second part is the grouping tier with two levels of phrasing (minor and major phrase, corresponding to prosodic phrase and prosodic utterance respectively) being symbolized. Our transcription is reminiscient of the ToBI transcription system [4], but unlike ToBI we rely exclusively on an auditory analysis.

The acoustic-phonetic analysis is based on F0 and waveform information, whereby both global features (in terms of, for example, F0 level and F0 range) and local features (in terms of direction and timing of F0 events) are taken into consideration and interpreted in our current prosody model.

tonal structure

accented	accent I (HL*)	
	accent II (H*L)	
focussed	accent I ([H]L*H)	
	accent II (H*LH)	
	compound (H*L...L*H)	
juncture	initial (%L; %H)	
	terminal (L%; LH%)	

grouping

boundary	minor	|
	major	‖

The three types of analysis - analysis of dialogue structure, auditory analysis, acoustic-phonetic analysis - involving both symbol and signal information are combined and synchronized with each other in the same ESPS/Waves+ environment. The labelling used (symbol information) consists of an orthographic tier (marking the end of words), a tonal tier (symbols of tonal structure), a boundary tier (symbols of grouping), dialogue structure tier (hierachical topic structure), and a miscellaneous tier (with extralinguistic and other information).

An important part of our research methodology is the use of speech synthesis. In addition to the text-to-speech synthesis as described above, one method currently being developed is the implementation of our intonation model in the ESPS/Waves+ environment which will be used as an analysis-by-synthesis tool. The input is the prosodic transcription with information about type and time location of tonal turning points. This information (with few segmentation marks) together with phonetic rules according to our intonation model are fed into a modifed version of the ESPS / Waves+ synthesizer. The synthesis module will be exploited both to verify the prosodic transcription and to develop the prosody model itself.

COMPARING INTONATION IN READ/SPONTANEOUS SPEECH

In the present paper we do not intend to give any conclusive answer to the question about typical differences between read and spontaneous speech. Instead we will try to come up with a few hypotheses about the specific, interactive contribution of intonation, features that would be typically lacking in prepared, non-interactive speech, as well as features typical of the intonation of both read and spontaneous speech.

Experimental design

One way of tackling this question experimentally is to make a direct comparison between the original, spontaneous version of a section of a dialogue and a corresponding, read version of the same portion of speech (cf. for example an early study by Gårding [5] and a recent study by Ayers [6]).

In our material, the original dialogue is a friendly conversation between two adult speakers, a daughter and her mother, who were talking spontaneously for around 13 minutes about partly predetermined topics. The read version is an acted dialogue involving the same speakers who took part in the spontaneous conversation. In the read version, they simulated their own, original dialogue from scripts. The recording of the acted version took place a few weeks after the original recording. The section from the dialogue which we have chosen for specific comparison and illustration in the present paper and which appears particularly interesting from a dialogue prosody point of view lasts for about a minute. In this section a transition from one conversational topic to another takes place (the very end of the topic 'the trip to Berlin' and the initial part of 'the blouse'), where the daughter takes the initiative and is thedominating speaker in the initial, mainly monologue part of this topic (Figure 1).

D: ja fast vi hinner inte så mycket	D: yes but we can't do all that much
M: nej	M: no
D: på	D: in
M: nej	M: no
D: en och en halv dag	D: a day and a half
M: Tiergarten och	M: Tiergarten and
D: men det ska bli jättespännande i alla fall	D: but it'll be tremendously exciting anyway
M: mm	M: mm
D: mm	D: mm
M: ja	M: yeah
D: ska jag berätta om min om den dära blusen som jag tänkte jag skulle sy av det där rutiga tyget	D: shall I tell you about my about that blouse that I thought I'd sew out of that checkered material
M: mm	M: mm
D: mm det är ju så jätterutigt så jag tror det blir jättekonstigt om man bara syr en vanlig skjorta	D: mm it's really so very checkered so I thought it would be really strange if I just sewed an ordinary shirt
eller ja det blir inte jättekonstigt men det är så fint tyg så det är synd om man inte gör nåt av det då	or well it wouldn't be really strange but it's such nice material so it would be a shame if I didn't do something with it
så har jag tänkt eh att jag skulle köpa ett vitt tyg till och ha dubbla kragar	so I thought ah that I would buy some white material and have a double collar
M: mm	M: mm
D: eh så jag ska ha en jag skulle gärna vilja ha en v-ringning och sen så eh lite snibbig krage så här	D: ah so I'll have a I'd really like to have a v-neck and then ah sort of pointed collar like this

Figure 1. Extract from spontaneous dialogue: friendly conversation between mother [M] and daughter [D]; Swedish original to the left and English translation to the right.

Tentative findings and discussion

The two versions of the dialogue section are, not surprisingly, auditively clearly distinct. The prepared, non-interactive but coherent character of the acted version of the dialogue as opposed to the characteristic interactive, on-line planning of the spontaneous version is quite striking. While we can assume that this impression is at least partly due to differences in pausing, variation in speech tempo, voice intensity, and voice quality as well as in the degree of reduction / elaboration, our specific task here is to try to isolate the contribution of pitch patterns and intonation.

The prosodic transcription, which is the broad, auditory analysis described above (basically a phonological analysis of accentuation and phrasing) displays apparent similarities between the two versions investigated. We can observe some differences in accentuation and focus locations as well as in phrasing between the versions. Some of these differences are clearly 'accidental', while others are probably more stable differences between speaking styles. In our search for regularities we will have to neutralize for differences between the versions in for example focus placement and phrasing that are clearly optional and not dependent on the specific speaking style.

Judging from the prosodic transcription, the most stable difference between the versions appears to be in phrasing. In the read version there is a tendency for a phrase to accomodate more words than in the spontaneous version as signalled by pitch and other cues. This may be thought of as due to the difference in planning between the speaking styles. The chunking into smaller units characteristic of the spontaneous speech is likely to be a reflection of the on-line planning. It is clear, though, that the broad, auditory categories used in our prosodic transcription do not reflect the apparent difference in intonation between the two speaking styles.

In our phonetic analysis of prosody we will concentrate on how the pitch patterns of the dialogue section may reflect two potentially important aspects of dialogue structure, namely the textual aspects (in particular topic structure) and the feedback dimension.

A number of studies have shown how variation in global F0 range reflects the hierarchical organization of a discourse and the segmentation into topics or subtopics, for example [7], [8], [9], [10], [11], [12]. An expansion of F0 range at the beginning and a compression of F0 range values towards the end of a text unit is typical. This global downtrend over a text unit has been modelled, by way of extrapolating from similar phenomena occurring over the course of a single utterance, as declination [10], downstep [13] or initial raising / final lowering [11].

The study by Grønnum-Thorsen [10] is particularly instructive. In text units containing four utterances the first utterance has higher F0 values and the last utterance has lower F0 values, while the two medial utterances tend to cluster around the same, intermediate F0 values. The generalization may be that the beginning of a new topic / speech paragraph is signalled by pitch raising and the end by pitch lowering, while the ongoing speech in between has no particular textual signalling by pitch.

Figure 2 shows F0 contours of the transitional phase between the major topics ('the trip to Berlin' and 'the blouse') for the read and spontaneous versions of the dialogue, including two utterances before and one utterance (consisting of two prosodic phrases) after the topic shift. In the acted, read version the major topic shift is clearly signalled by means of F0. A marked shift from a fairly low F0 level and compressed F0 range to a high F0 level (increase by half an octave) and wide F0 range at the discourse boundary ties in with cited and expected relations. The decrease in F0 level and F0 range from the first to the second phrase of the utterance beginning the new topic is also apparent.

Figure 3 shows another example utterance containing two prosodic phrases representing the beginning of a subtopic somewhat later in the dialogue. Also here the read version displays

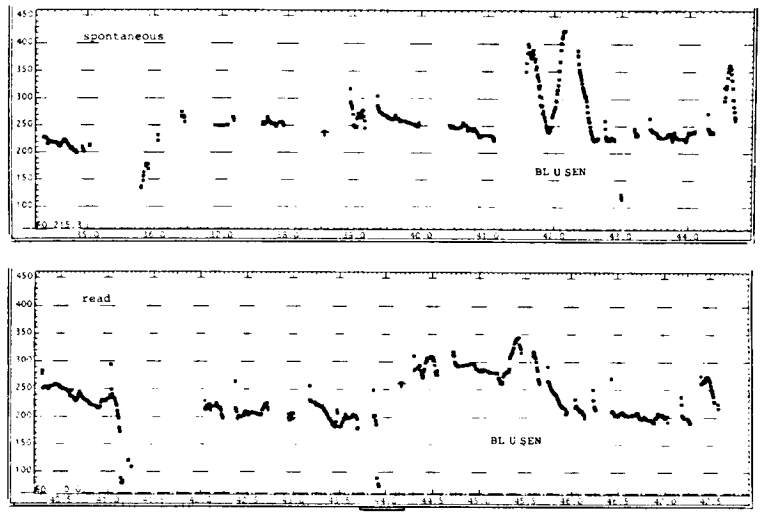

Figure 2. The effect of major topic shift. F0 contours of the same utterances from the spontaneous version (upper part) and the read version (lower part) by speaker [D]. The arrow indicates the time location of the topic shift. Focal word is in capital letters.

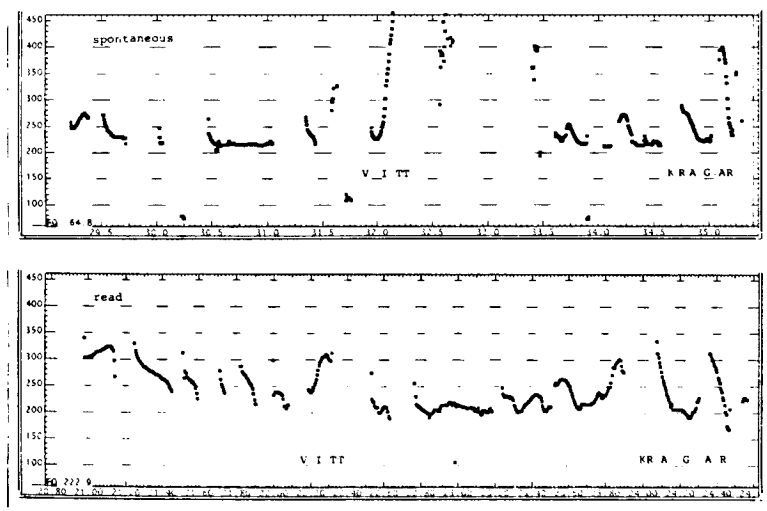

Figure 3. The effect of interactivity. F0 contours of the same utterance from the spontaneous version (upper part) and the read version (lower part) by speaker [D]. Focal words are in capital letters.

higher F0 values of the first phrase as compared with the second phrase of the utterance, reflecting the further textual organization of a subtopic.

In the spontaneous version (see Figures 2 and 3) we also find signs of textual organization in the distribution of pitch patterns including the major topic shift. Also here it is primarily global F0 level and F0 range which seem responsible for this organization. But this signalling appears to be less marked and maybe less stereotypic than in the acted version. Similar findings are reported in a corresponding study of American English read and spontaneous speech by Ayers [6].

A case in point is the signalling of the major topic shift in the spontaneous version. Unlike the read, acted version, the first utterance of the new topic 'the blouse' begins on approximately the same F0 level as the final utterance of the old topic, albeit with a wider F0 range. The last utterance of the topic 'the trip to Berlin' is namely characterized by a clear increase in F0 level as compared with the immediately preceding utterance. It functions prosodically as a transition utterance, both rounding off the old topic and, as it were, anticipating the topic shift, and also as a turn-keeping signal. Another exemplification from spontaneous dialogue of an utterance constituting a transition between major conversation topics, which textually belongs to the old topic but prosodically is also clearly affiliated with the new topic, is found in [14].

The difference in interactivity between the original, spontaneous version and the acted, read version is reflected in the feedback dimension. The seeking of feedback by the speaker having the turn seems to be one characteristic of the spontaneous dialogue which is typically lacking in the acted, prepared dialogue.

In the spontaneous version as exemplified in Figures 2 and 3 there appears to be a particularly wide, local F0 range on the two focussed items in each Figure (affecting mainly the focal H), while the corresponding items in the acted, read version have a moderately wide F0 range. The difference in range between the read and spontaneous versions for the focussed items exemplified amounts to about half an octave.

One possibility is that the difference in F0 values is related to differences in interactivity, specifically in the feedback dimension. The concentration of feedback seeking to certain focussed items is evidenced by the fact that it is at these points in time that feedback is also given through the use of support items of the 'mm', 'yeah' type. The particularly wide F0 range on the focussed items thus seems to be a reflection of the speaker seeking feedback from the listener.

This is reminiscient of the high rising tone sequence (H H%) found for spontaneous as opposed to read speech in American English by Ayers [6], what Sacks and Schegloff call 'try marker', and what Clark and Schaefer term a 'trial constituent' [15].

It should be pointed out, however, that in the spontaneous speech studied the exploitation of an extra wide F0 range on focussed items appears to be quite variable for successive phrases and utterances, apparently depending on the speaker's need for feedback.

CONCLUSION

Our comparison of the two versions of the dialogue section examined here serves as an illustration of possible intonational differences between read and spontaneous speech. The most apparent differences from this comparison are summarized here.

Pitch appears to play a major role in the signalling of textual organization, topic structure and division into speech paragraphs. This is evidenced in the spontaneous and acted versions which both display variation in global pitch level and pitch range as a reflection of this organization. A possible difference between read and spontaneous speech may be that in read speech textual organization is more stereotypic and rigidly marked. The initial raising of pitch level and range of the first phrase / utterance, intermediate values in between, and the final lowering of the last phrase / utterance of a text unit may represent the reading stereotype.

A marked increase in local pitch range specifically on focussed items may serve as a means of seeking feedback from your interlocutor. This is a feature characteristic of spontaneous speech, while it appears to be absent in acted, read speech.

These and other features of intonation are currently being modelled in our prosody model and tested for their assumed significance as signals of read and spontaneous speech.

ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK). I want to acknowledge the cooperation with my colleagues in the research project 'Prosodic segmentation and structuring of dialogue', namely Björn Granström, Kjell Gustafson, Merle Horne, David House and Paul Touati. Gayle Ayers, Dept of Linguistics, Ohio State University was a guest researcher in Lund for several months during 1993 and 1994 and has contributed to the project. She prepared the recordings and the analysis of the dialogue investigated.

REFERENCES

[1] Bruce, G., B. Granström, K. Gustafson, D. House and P. Touati (1994), "Modelling Swedish prosody in a dialogue framework", *Proceedings of ICSLP 94*, pp. 1099-1102, Yokohama.

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

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

[4] Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg (1992), "ToBI: a standard for labelling English prosody", *Proceedings of ICSLP 92*, pp. 867-870, Edmonton.

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

[6] Ayers, G. (1994), "Discourse functions of pitch range in spntaneous and read speech", *OSU Working Papers in Linguistics*, Vol. 44, pp. 1-49.

[7] Lehiste, I. (1975), " The phonetic structure of paragraphs", A. Cohen and S. Nooteboom (eds.), *Structure and Process in Speech Perception*, pp. 195-206, Berlin: Springer-Verlag.

[8] Brown, G., K. Currie and J. Kenworthy (1980), *Questions of intonation*, London: Croom Helm.

[9] Bruce, G. (1982), "Textual aspects of prosody in Swedish", *Phonetica*, Vol. 39, pp. 274-287.

[10] Thorsen, N. (1985), "Intonation and text in Standard Danish", *JASA*, Vol. 77, pp. 1205-1216.

[11] Hirschberg, J. and J. Pierrehumbert (1986), "The intonational structuring of discourse", *Proceedings of the 24 Meeting of the Association of Computational Linguistics*, 136-144, New York.

[12] Silverman, K. (1987) *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. Thesis, Cambridge UK: Cambridge University.

[13] Berg, R. van den, C. Gussenhoven and T. Rietveld (1992), "Downstep in Dutch: Implications for a model", G. Docherty and R. Ladd (eds.) *Papers in Laboratory Phonology II. Gesture, Segment, Prosody*, pp. 335-359, Cambridge University Press.

[14] Bruce, G, P. Touati, A. Botinis and U. Willstedt (1988), " Preliminary report from the KIPROS project", *Working Papers*, Vol. 33, 23-50, Lund University: Dept. of Linguistics and Phonetics.

[15] Clark, H.H. and E.F. Schaefer (1989), "Contributing to discourse", *Cognitive Science*, Vol. 13, pp. 259-294.