



Towards an enhanced prosodic model adapted to dialogue applications

G. Bruce*, B. Granström**, K. Gustafson**, M. Horne*, D. House* & P. Touati*
(names in alphabetical 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

This paper discusses the need for an enhanced model of Swedish prosody for use in dialogue systems. A description is given of the components of the model which is being developed as part of the project *Prosodic Segmentation and Structuring of Dialogue*. The model is based on parameters derived from observations of both man-man and man-machine dialogues. The main extension in relation to our previous model lies in the introduction of continuous parameters which affect the phonetic realization of the discrete parameters of our original model. These continuous parameters are related in complex ways to features of the discourse situation.

1. INTRODUCTION

The research reported here is conducted within the project *Prosodic Segmentation and Structuring of Dialogue* [1]. The object of study in the project is the prosody of dialogue in a language technology framework. 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: *Language Technology for Spoken Dialogue Systems* [2] (the Waxholm project), *Intonation in Restrictive Texts: Modelling and Synthesis* [3] and *Interaction in Speech between Prosody, Syntax, Semantics and Pragmatics* [4].

2. GOAL AND METHODOLOGY

The primary goal of the project *Prosodic Segmentation and Structuring of Dialogue* is to increase our understanding of how the prosodic aspects of speech are exploited interactively in dialogue and on the basis of this increased knowledge to be able to create a more powerful prosody model which is capable of making phonetic distinctions that are communicatively relevant in a dialogue situation. 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

In this contribution we concentrate on our research related to the development of an enhanced prosodic model, for use in a man-machine dialogue context.

In our work 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, simulated dialogues using text-to-speech synthesis and man-machine dialogues are also exploited in our study of dialogue prosody.

3. THE NEED FOR AN ENHANCED PROSODIC MODEL IN AUTOMATIC SPEECH RECOGNITION AND TEXT-TO-SPEECH SYSTEMS

Until fairly recently both speech recognition and speech synthesis have operated with rather narrow contextual windows of analysis. In the case of speech synthesis systems, a window size of one sentence has been typical, and in speech recognition, one utterance consisting of possibly one complete sentence but often limited to just one word in size has been typical. Recent years have, however, seen the development of integrated systems where relatively large contexts can be successfully analyzed during the speech recognition part of a man-machine dialogue. An example of this is the project *Language Technology for Spoken Dialogue Systems*.

Up until now, prosody has played only a minor role in affording contextual cues in the speech recognition components of such systems, and in the speech synthesis components, prosodic differentiation based on contex-

tual cues has been very limited. In a man-machine dialogue context it is important that the synthesis is capable of generating relevant context-dependent prosodic distinctions. For instance, in a language like Swedish where accentual defocussing is commonly used as one means of signalling old/given information, a lack of a defocussing strategy in a man-machine dialogue situation may lead to misunderstandings on the part of the human participant.

As part of the Language Technology Programme, The Prosodic Segmentation project is able to make use of the work being conducted in the other projects within the programme. We are benefiting especially from cooperation with the projects *Language Technology for Spoken Dialogue Systems* and *Intonation in Restrictive Texts: Modelling and Synthesis*. This cooperation entails a two-way effect of benefits: our project is able to draw on the work done in those projects in terms of both theoretical results and material for use in our own analysis work, and in turn the practical results we are producing by way of an enhanced prosody model and rules for the automatic generation of distinctive and contextually relevant prosodic patterns can be tested within the framework of the other projects. The database of the Waxholm project includes samples of spontaneous speech produced in a man-machine dialogue situation. This constitutes important material for the study of the prosody produced spontaneously by humans in such situations. Relevant aspects of this material that we are studying are related to:

- text/discourse segmentation (topics, focus)
- dialogue structuring (feedback, turns, initiative/response)
- attitudinal/emotional factors

4. ANALYSIS OF DIALOGUES

The point of departure for our work on a prosodic model is our present model which goes back to Bruce [5] and which is in essence implemented in our TTS system. To achieve our goals within this project we are studying a wider range of speech situations than those on which the present model was based, in particular speech produced in dialogue situations of different kinds, both man-man and man-machine dialogues.

In one of our studies of man-man dialogues we have, among other things, investigated the differences that exist between spontaneous and read speech [6]. This has been done by having the participants of a spontaneous dialogue re-enact the same dialogue a few weeks later from a (slightly edited) written transcript. This method, which has been successfully used in a number of other studies, e.g. [7], [8], allows us to identify more easily

factors that may be presumed to be characteristic of spontaneous dialogue, as opposed to read speech. The two versions of the dialogue are, not surprisingly, audibly clearly distinct. One of the observations of this study has been that the acted version exhibits a more coherent speaking style as compared to the spontaneous version; this, on the other hand, exhibits a more lively, interactive style. A broad prosodic transcription of the two versions reveals some interesting differences in accentuation and focus locations as well as in phrasing, some of which appear to reflect stable differences between the two speaking styles. The most striking differences seem to be related to phrasing. There is a tendency for a phrase, as signalled by pitch and other cues, to accommodate more words in the read version than in the spontaneous one. 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 from our study that characteristically different pitch patterns are powerful tools in signalling both the textual aspects (in particular topic structure) and the feedback dimension of dialogue structure. The complexities of the use of prosody in spontaneous speech can be illustrated by the following example: in an area of transition between two conversational topics, one of the speakers signals the topic shift by using a marked increase of F0 in the last utterance of the first topic. The effect is that, although this utterance belongs textually to the first topic, and thus points backwards, prosodically it is grouped with the next utterance, and thus points forwards. In the read version, on the other hand, the topic shift is signalled by a typical, marked increase in F0 at the discourse boundary.

Compared to the prosody of real-life human dialogues, that of the human participants in the man-machine dialogue situations that we have studied exhibits on the whole a rather limited degree of prosodic variation. This pertains to both F0 range, focus assignment, and tempo. It seems likely that this is, at least in part, related to the limitations in prosodic variation on the part of the machine partner. The main notable exception is when the machine does not understand or misunderstands what the human says. Typically, the human then exhibits a wider F0 range, and commonly other features characteristic of agitated or angry speech, such as increased intensity. One variant involves a widened range in the focus domain, combined with a narrowed range in a higher than average register. This kind of signalling could be used as a cue to switch to a human operator in a practical machine-man system.

5. DEVELOPING THE PROSODY MODEL

An adequate prosody model must be able to explain and generate the different configurations of prosodic parameters that we observe in our spoken dialogue analyses, and a model and theory of spoken dialogue must be able to relate such prosodic patterns to features of the dialogue situation. Since our current prosody model is based on monologue situations with a one-sentence context window, it is our goal in the present project to develop it for use in multi-sentence-window applications and specifically to be relevant in dialogue situations.

The enhanced model under development uses the same basic building blocks as our standard model. The main extension lies in the fact that a number of gradational elements are being added, which relate to the phonetic realization of the elements of the standard model, which are basically phonological and discrete in nature.

The scope of the model is the major phrase unit, which may consist of two or more minor phrases. The division of the major phrase into minor phrases is manifested in the model by either boundary signals at the junction between them or by signs of cohesion within one or more of them.

Each phrase can be divided into a number of domains: initial juncture domain, (optionally) prefocal domain, focal domain, (optionally) postfocal domain, terminal juncture domain. As a variant, a non-focal main domain may take the place of a prefocal + focal domain.

Within each domain, each major event, i.e. in practice each turning point, is independently variable with respect to both the value and the timing of F0.

The components of the enhanced model fall into two groups,

- A discrete elements**
- B gradational elements**

The components of the standard model belong, for the most part, to group A.

The discrete elements of group A are mainly phonological/linguistic or pertain to grammatical, semantic and textual structure. These elements are typically binary.

The parametric values of the gradational elements of group B mainly reflect paralinguistic factors, such as the emotional state or the regional background of the speaker, his or her attitude in the speaking situation, or other pragmatically determined factors, related for instance to the dialogue situation. The structure of the model is shown in table 1. Each of the elements of the model can be parametrically varied, independently within each of the domains referred to earlier.

The durational parameters affect the rhythmic structure both within and across metric feet. The timing parameters affect the turning points of the basic F0 events, something which is crucial to achieving authentic sounding synthesis, and which it is hypothesized offers important cues in the perception of the nuances of natural speech. Pauses can be specified with respect to whether they are silent or filled, and whether they affect the durational relationships of the previous segments, and if so, to what extent.

There exists a close relationship between the elements in the two groups, in that those in group A are manifested and specified in terms of those in group B, that is, it is typical of the elements of B that they pertain to the *realization* of those of A.

Although the scope of the model is restricted to the major phrase, it is part of our wider work to examine the parametrical realizations within the major phrase as seen in the wider context of a dialogue situation. Here, too, the phonetic realization of initial or terminal juncture or signs of cohesion or lack of cohesion across a major phrase boundary may function as means to signal the structure of a succession of major phrases in a dialogue setting.

Table 1 Components of the enhanced model

A discrete elements	
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
B Gradational elements	
F0 phenomena	
	F0 range F0 register general direction of F0 movement (slope) timing of F0 events
Duration	
Voice source characteristics	
Reduction phenomena	

6. LINKING THE PROSODY MODEL TO THE PARAMETERS OF DIALOGUE STRUCTURE

We have initially studied dialogue structure and the prosody of dialogues independently, as far as that is possible in practice, in order to avoid the danger of circularity.

Having established, on the one hand, the prosodic parameters - both discrete and gradational - and, on the other, the relevant categories of dialogue structure and human dialogue management, the next task is to establish the link between the two.

To do this we have implemented the enhanced prosody model in an experimental speech synthesis system. In this we have defined special prosodic markers which can be inserted in the text. These markers represent particular settings of the gradational prosodic parameters of the enhanced model so that it is possible to select virtually any point in the multi-dimensional space created by these parameters. They can represent complex combinations like "reduced F0 range at a raised level prefocally, with extended F0 range in focus, followed by a rising terminal juncture". The combinations chosen generate patterns that we observe in our speech database and which we hypothesize are relevant indicators of dialogue-related prosodic behaviour. This setup can be used to experimentally test the validity of our hypotheses of how prosody is used specifically in dialogue situations in terms of emotional states like anger and indifference, and dialogue structure phenomena like turn-taking and feedback seeking.

Later we will advance one step further and use the enhanced model in experiments to test our hypotheses on the roles and functions of prosody in a dialogue setting. By coupling the enhanced synthesis to the Waxholm dialogue system we can study the prosody of both the automatically generated speech and that of the human participants in this new environment. One hypothesis which can then be tested is that the limited range of prosodic variation exhibited by the human speakers in our man-machine material is correlated to the limitations in prosodic variation and expressiveness on the part of the machine dialogue partner.

Looking further into the future, the implementation of this model in an automatic dialogue system will facilitate the production of appropriate prosody once it has the means to make contextually based choices among a repertoire of prosodic possibilities.

The model is intended to be one of both production and perception of human speech. This means that it should be designed to function equally well in a speech recognition system and in a text-to-speech system.

7. ACKNOWLEDGMENT

This work has been supported by grants from The Swedish National Language Technology Programme. We would like to thank Marcus Filipsson and Birgitta Lastow who have assisted us in our work on the man-machine dialogues.

8. REFERENCES

- [1] G. Bruce, B. Granström, K. Gustafson, D. House & P. Touati, "Modelling Swedish prosody in a dialogue framework", Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP 94), pp. 1099-1102. Yokohama 1994.
- [2] M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, R. Lindell & L. Neovius, "An experimental dialog system: WAXHOLM," Proceedings of Eurospeech '93. pp 1867-1870, 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, pp. 81-84. Department of Linguistics, Uppsala University, 1993.
- [5] G. Bruce, Swedish word accents in sentence perspective, Lund: Gleerups, 1977.
- [6] G. Bruce, "Modelling Swedish intonation for read and spontaneous speech". To appear in the Proceedings of ICPHS, Stockholm 1995.
- [7] E. Gårding, "Prosodiska drag i spontant och uppläst tal", G. Holm (ed.), Svenskt talspråk, pp. 40-85, Uppsala: Almqvist & Wiksell. 1967.
- [7] Ayers, G., "Discourse functions of pitch range in spontaneous and read speech", OSU Working Papers in Linguistics, Vol. 44, pp. 1-49, 1994.