Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

The Swedish intonation model in interactive perspective

Bruce, G. and Frid, J. and Granström, B. and Gustafson, K. and Home, M. and House, D.

KTH Computer Science and Communication

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 019-022

http://www.speech.kth.se/qpsr

The Swedish intonation model in interactive perspective

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

Abstract

In this paper we discuss some recent extensions of the prosody model used in the research project 'Prosodic Segmentation and Structuring of Dialogue'. In our current modelling of dialogue intonation we are using both model-based resynthesis and text-to-speech. The idea is to be able to regulate the influence of discourse and dialogue structure on prosody and intonation. In the present contribution we report on some aspects of overall F0 trends in dialogues and its implications for the model. We also report on our work to incorporate dialogue-related rules in our text-to-speech system.

Model-based resynthesis

The research reported here is conducted within the ongoing research project 'Prosodic Segmentation and Structuring of Dialogue' supported within the Swedish Language Technology Programme (HSFR/NUTEK). 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 our model-based framework, described in detail e.g. in Bruce et al. (1994), Bruce et al. (1995) we currently label word accents, focal accents, and boundary tones (in a 'tonal' layer), and phrase boundaries (in a 'phrase' layer). Modelling of F0 contours can then be done using these labels together with a set of parameters that control the timing and frequency of accents as well as the overall F0 trend of a phrase.

So far, these parameters have been set independently from phrase to phrase on an experimental basis. We want to be able to govern this parameter setting in a way that incorporates information on discourse and dialogue structure, in order to model the prosodic structure of dialogues. A step in this direction is the introduction of a separate layer of labels which can be used to generate a rule-based set of parameters that will be used in the modelling of a given phrase.

We need to take into account some aspects of the overall characteristics of F0 for a larger stretch of spontaneous speech, e.g., a dialogue, in writing the rules of parameter generation for the discourse layer. One way to gain some insight into this is to utilize the analysis-by-synthesis method described in Bruce et al. (1995). We have employed this method by taking a larger stretch of a dialogue and modelling the F0 contour for each successive phrase in that dialogue using two different approaches:

1) a 'default' contour, where the frequency parameters are held constant for each successive phrase. The only inter-phrasal variation is thus the timing and the identity of the labels. The motivation for using default parameters is that it gives us a similar reference level for all phrases, which can be used for comparisons (see also below).

2) a 'fine-tuned' contour, where the frequency parameters are set individually for each phrase. The values of the parameters are based on measurements in the original F0 contour of that phrase.

The difference is visualized in Fig. 1, which shows the original, the default, and the fine-tuned pitch contours for three successive phrases. Note the similar heights of the peaks in the default curve, whereas in the fine-tuned curve there is variation between each phrase.

Three kinds of comparison can be made from these contours:

a) *original contour - default contour*

This provides a way of normalizing. Since speakers continually vary their register and range, straightforward frequency comparisons are not reliable. However, if the original is

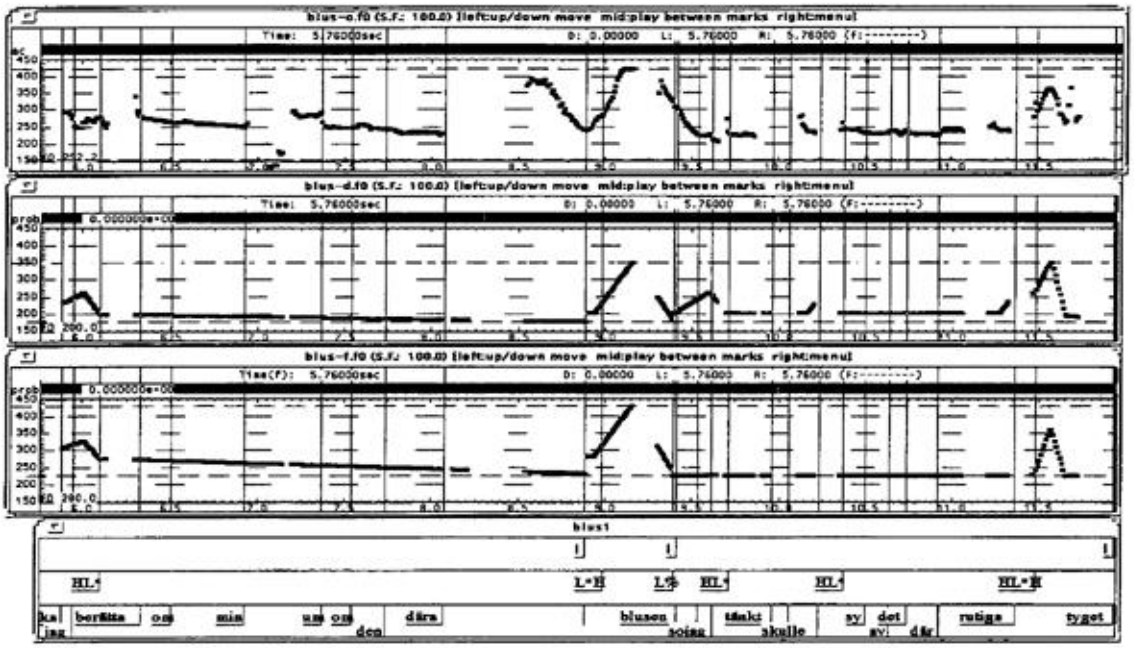

Figure 1. Original, default, and fine-tuned F0 contours of an utterance containing three successive phrases (eng. translation: 'shall I tell you about my about that blouse that I thought I'd sew out of that checkered material'). At the bottom is shown the corresponding transcription in terms of phrasing, prominence (tones) and orthography. The fall at time 8.5 (before 'blusen') is extralinguistic (a clearing of the throat) and is thus not labelled.

compared with a reference contour, it is possible to compare the magnitude of differences.

b) *original contour - fine-tuned contour*

This comparison enables us to inspect: I) how the timings of the modelled accents correlate with the original, II) the differences in the shapes of accents between model and original, and III) what global effect it has to base the frequency tuning on local measurements.

c) *default contour - fine-tuned contour*

The nature of the accent labels and the timing are identical in these two kinds of display. Therefore, they will only differ with respect to F0 (range and register). Since the parameters are held constant from phrase to phrase in the default contours, these contours can be viewed as 'reference contours' against which deviations of the fine-tuning can be matched. It is then possible to localize recurring patterns of deviation that can be matched up with discourse and dialogue structures.

Analysis method

In this study, trends of average F0 for a large number of successive phrases in a dialogue between two female speakers were analysed. The difference between two separate contours of the same phrase, referred to as the difference contour, was estimated by computing the difference in F0 at each point and then calculating the overall mean and standard deviation within each phrase. Means and standard deviations of each successive difference contour were then plotted as a function of their temporal occurrences in the dialogue. This resulted in an overview of the entire analysed dialogue. In total, the dialogue was segmented into 140 phrases.

Results & discussion

Figure 2 shows the standard deviations of the difference contours between default and fine-tuned contours. The magnitude of deviation of each phrase gives us an indication of how to set parameters for later modelling. If we regard the default as a reference, the interpretation is that a larger standard deviation implies a greater difference between the contours. Since the fine-tuned contours have properties of the original, our modelling should aim at approaching the fine-tuned parameters. Thus, a large deviation in a phrase means that parameters should be set so that its F0 contour differs more from the default of that phrase. Large deviations in a phrase also means that the default parameters are inappropriate in modelling that phrase, while a small deviation means that the default parameters yield a good modelling of the phrase.

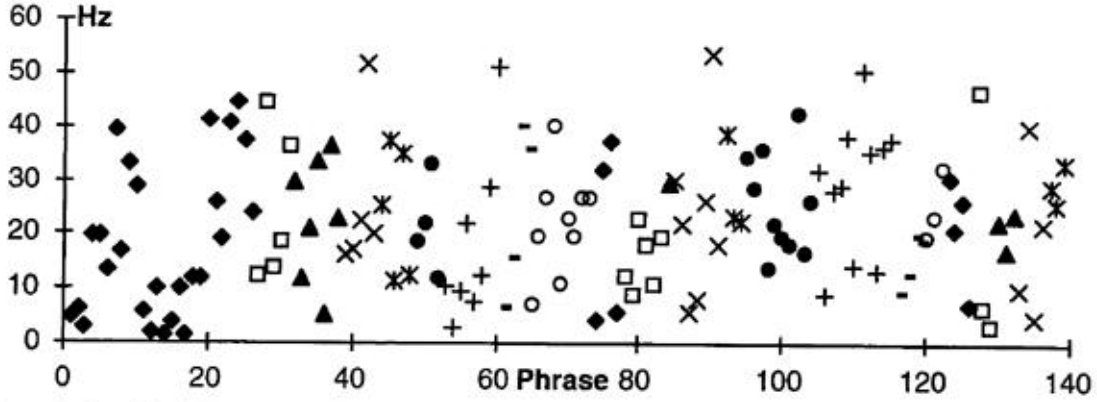

Figure 2. Standard deviations of the difference contours between default and fine-tuned contours for each phrase (each group of identical neighbouring symbols represents one speaker turn).

In figure 2 we see that standard deviation varies greatly through the dialogue. Almost every speaker turn contains phrases with both small and larger deviations. Also, the phrase-to-phrase ratio is very variable.

Explanations for this variation have been sought in the relation between phrase and the accent types (acute, grave, focal) it contains, but no correspondence could be found. We have also examined lexical properties of this dialogue segment, but no systematic variation was found there either. The distribution of the standard deviations must be accounted for by other phenomena, such as dialogue structure and given/new information.

As for the means, if we maintain the view that the fine-tuned contours should be approached, the results are interpreted as follows: a positive mean difference indicates that the default contour is higher than the fine-tuned contour. Parameters should thus be set so as to produce a lower contour. The converse applies to the negative mean differences.

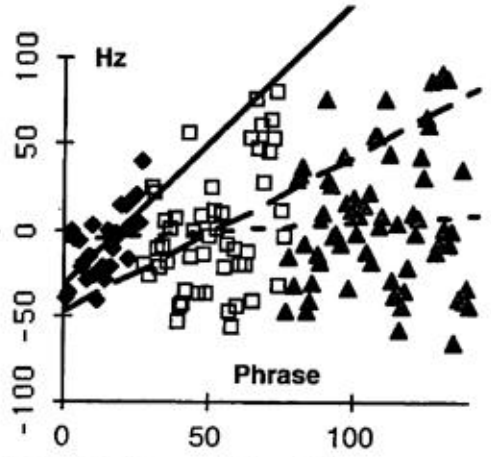

Figure 3. Means of the difference contours between default and fine-tuned contours for each phrase (see text for explanation of grouping and trendlines).

Figure 3 shows the means of the difference for all phrases. The dialogue is split into three different subtopics (by means of a textual analysis), and phrases are grouped after this rough partition of conversational topics. The linear regression of the means of each subtopic is then calculated. The equations of the regression lines give us the possibility to calculate what average F0 to strive for in the modelling. For instance, the line of the first group is described by the equation:

$$y = 1.6514x - 31.749$$

If x is regarded as a phrase's number in temporal order, the negative of y indicates how much the average F0 of a modelled contour of that phrase should differ from the average of the default contour. This works best with a high validity of the regression. The calculation of R^2 (a measurement of the validity of the regression) of each group is given below:

Subtopic 1	0.43856924
Subtopic 2	0.13223696
Subtopic 3	0.00280279

This means that the regression is most valid for the first group, less so for the second, and least valid for the third group. This reflects the fact that the variance between phrases within a group grows with each subtopic. In addition, we see that the regression goes from negative means (fine-tuned is higher than default) in the beginning to positive means (default is higher than fine-tuned) in the end, thus the fine-tuned average F0 decreases within each group of phrases.

In view of these statistical indications, it seems that each subtopic should be modelled separately. If this case study proves to be representative of Swedish dialogues, the phrases in

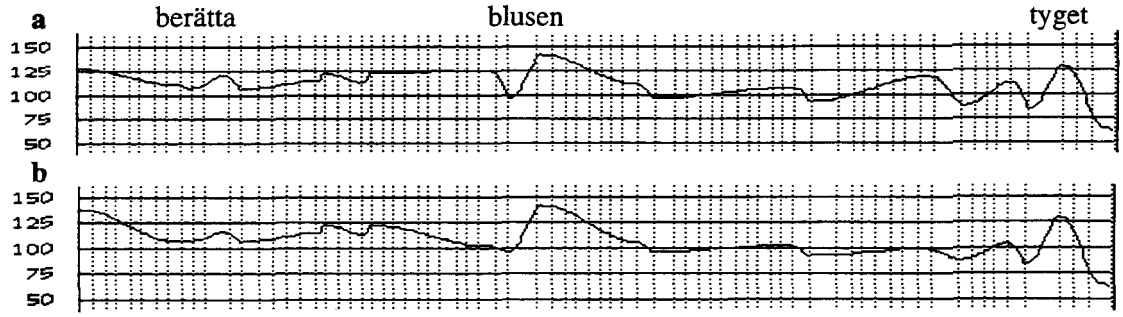

Figure 4. A slightly simplified version of the utterance displayed in Fig. 1. **a** *represents the default prosody of the TTS system with the addition of focus markers.* **b** *is the same sentence with discourse markers inserted in the orthographic text. The text input is:* Ska jag berätta om min mmm om den dära blusen som jag tänkt jag skulle sy av det där rutiga tyget? *The positions of the main content words are indicated above the first illustration.*

the first subtopic should be modelled so as to show a declining F0 trend, while the modelling of the following subtopics should have a successively less marked tilt.

Text-to-speech

In parallel with the analysis-resynthesis method described above we are developing an enhanced version of the KTH text-to-speech (TTS) system. The TTS system was originally designed to provide default prosodic patterns in "neutral" single-sentence utterances. In addition, markers can be inserted into the text to regulate degree of emphasis and focus placement.

The aim is that the enhanced system should also be able to accommodate discourse and dialogue features of the text by taking into account prosodic features observed in both man-man and man-machine dialogues. In order to achieve this, the default synthesis has been extended by introducing a set of parameters that regulate the main prosodic parameters of F0, segment duration and pause duration. With this enhanced system it is thus possible to vary overall pitch level, F0 range on accented syllables (both focal and non-focal), as well as the F0 properties of initial and terminal juncture. In addition, segment durations and the durational relation between vowel and postvocalic consonant(s) can be specified to reflect phonologically relevant rhythmical properties as well as settings associated with individual speakers.

With this system we can reproduce good representations of the kind of prosodic variation that typically occurs in dialogues. Based on studies of the Waxholm database (Bertenstam et al. 1995) and of the speech material used in our analysis-by-synthesis work, we have defined a number of parameter settings which correspond to observed prosodic behaviour. These settings are coded in symbols that are stored in the system's fixed lexicon. By inserting these symbols manually in the orthographic text we can thus generate by text-to-speech prosodic patterns representative of man-man and man-machine dialogues. We are planning to link this facility to an automatic discourse generator, so that relevant codes can be automatically inserted in the input text to be spoken by the TTS system.

Fig. 4 shows the same sentence as that displayed in Fig. 1. Fig. 4a represents the default prosody of the TTS system with the addition of focus markers. Fig. 4b is the same sentence with discourse markers inserted in the orthographic text. These affect the initial juncture and the prefocal stretches leading up to the phrase-final and focally stressed words *blusen* and *tyget*. The text input has been simplified by omitting the noise arising from the clearing of the throat and the pause before this. Note that the F0 register has been transposed to a level more representative of a male speaker. Durational aspects have not been manipulated in this particular example.

References

Bertenstam J, Beskow J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Nord L, de Serpa-Leitao A & Ström N (1995). The Waxholm system - a progress report. In: *Proceedings ESCA Workshop on Spoken Dialogue Systems*, Aalborg, 81-84.

Bruce G, Granström B, Filipsson M, Gustafson K, Horne M, House D, Lastow B & Touati P (1995). Speech synthesis in spoken dialogue research. In: *Proceedings EUROSPEECH 95*, Madrid: 1169-1172.

Bruce G, Granström B, Gustafson K, House D & Touati P (1994). Modelling Swedish prosody in a dialogue framework. In: *Proceedings ICSLP 94*, Yokohama, 1099-1102.