

SPEECH SYNTHESIS IN SPOKEN DIALOGUE RESEARCH

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

* Dept of Linguistics and Phonetics, Helgonabacken 12, S-22362 Lund, Sweden
{gosta.bruce | marcus.filipsson | merle.horne | david.house | birgitta.lastow | paul.touati}@ling.lu.se
** Dept of Speech Comm. and Music Acoustics, KTH, Box 70014, S-10044 Stockholm, Sweden
{bjorn | kjellg}@speech.kth.se

ABSTRACT

This paper concerns the use of speech synthesis as a tool in investigating dialogue structure and in modelling prosody for use in dialogue-based text-to-speech systems. We report on two different approaches used in our efforts to investigate how prosody is used as a communicative tool in dialogue situations. Our work has relevance for practical applications of both text-to-speech and speech recognition technology.

1. INTRODUCTION

Our current work is directed towards the prosody of dialogue within the project *Prosodic Segmentation and Structuring of Dialogue* [1], which is part of the Swedish Language Technology Programme. In this work we are studying both man-man and man-machine dialogues. The work has both a theoretical side, the basic study of dialogue structure (turn regulating mechanisms, feedback seeking, etc.) and its relevance for prosody, both of natural speech and in speech synthesis; and a practical side, the implementation of rules to produce adequate prosody in text-to-speech (TTS) used in a dialogue system.

TTS systems have in the past mainly been used to model read speech. As the analysis window of most current systems is quite narrow, the possibility of exploiting pragmatic factors and such textually relevant features as dialogue structure has been very limited.

This situation makes imperative the study both of dialogue structure itself and of the functions and phonetic manifestations of prosody in dialogues. The functions of prosody in dialogues range from purely linguistic ones to the signalling of speaker attitudes and of emotions such as joy, anger and frustration [2]. It is our aim, during our current work, to make a contribution in each of these areas.

2. THE PROSODY MODEL

The model of Swedish prosody used in this project is an enhanced version of that developed by cooperation between Lund and KTH and implemented in our TTS system [3].

The model identifies a number of discrete categories with associated labels. The model recognizes two levels of prominence, for each level of prominence the distinction between the two word accents in Swedish, initial and terminal boundary tones, and two degrees of grouping

(see Table 1). The system of labelling is similar to that used in ToBI [4].

Table 1. Discrete prosodic categories for Swedish and ToBI-like labels used in the prosodic model and transcription system. The star () indicates the location of the stressed syllable and the percent sign (%) the group boundaries.*

<i>Prosodic category</i>	<i>Label</i>
Accent I	HL*
Accent II	H*L
Focal accent I	(H)L*H
Focal accent II	H*LH
Focal accent II compound	H*L...L*H
Initial juncture	%L
Terminal juncture	L%, LH%
Minor phrase	
Major phrase	

In addition to these discrete categories we operate with a range of gradational elements. These have the dual function of supplying a tool for the experimental manipulation of the phonetic realization of the synthesized utterances and of enabling us to model non-categorical variation, for instance related to individual speaking styles and to dialogue situations. These are shown in Table 2. For a further description of the enhanced prosody model, see [5].

Table 2. Gradational elements of the prosody model

F0 phenomena:
F0 range
F0 register
general direction of F0 movement (slope)
timing of F0 events
Duration
Voice source characteristics
Reduction phenomena

3. TWO METHODS OF SYNTHESIS

Synthesis is an efficient tool for research into spoken dialogue. It can be profitably used in the study of aspects of dialogue itself, it can be used to test hypotheses of a variety of kinds, for instance regarding what categories are required to adequately represent the prosody of dialogue, and it can be used experimentally in an effort to develop TTS systems that are better able to represent

spoken dialogue, once answers have been found to the more basic, theoretical questions.

We use speech synthesis in all these different ways. Our data consist of both man-man and man-machine dialogues. In our work we are following two different approaches. The first is a rule synthesis approach based on our text-to-speech system. The second tool, which is still under development, is primarily aimed at testing intonation models on natural sentences using PSOLA resynthesis techniques [6, 7]. For this work we are exploiting the ESPS/Waves environment.

3.1. A parametric approach

The first approach uses a parametric extension to our existing synthesis system which is based on the RULSYS development environment. On the basis of observations in our speech material we have defined a set of prosodic parameters and implemented these in the TTS system. By manipulating the parameter values we can generate F_0 and durational patterns closely resembling those of our speech material. This parametric model basically specifies the phonetic properties of the prosody of utterances. Linked with this is a mapping procedure whereby relevant phonological and discourse-related categories can be mapped to specific settings of the phonetic parameters. This parameter-based approach allows us to test perceptual properties of the different parameters. It is easy to specify and model new patterns when such are observed in the speech material. We can also model differences that are due to factors other than strictly phonological ones, for instance such as are due to speaker atti-

tudes and emotions and regional variation. Above all, in the context of dialogue modelling, it is possible to specify prosodic variation that can be attributed to the dialogue situation and to model this variation.

The parameters allow us to specify, among other things, F_0 baseline, F_0 topline, F_0 high (H and H*), F_0 low (L and L*), F_0 slope, and the phonetic details of initial and terminal juncture. In addition there are parameters regulating the exact timing of specified types of turning points, relative to a default time value.

In practical terms, the parameters are manipulated by markers inserted in the text. On the basis of production data we can define the parameter values of markers corresponding to specific phonological and non-phonological prosodic categories. By inserting these in orthographic text we can produce prosodic effects beyond those that can be generated by the default rules of the model. We are planning to link this system to an automatic dialogue management system (the Waxholm system [8]) where such phonological markers can be selected automatically on the basis of dialogue structure information. Initially we have been studying the prosodic realization of the human participants and relating this to features of the dialogue situation.

Figure 1 A shows the F_0 trace of an utterance in the database of the Waxholm project. This utterance exhibits a range of features of a dialogue-related kind, as well as ones related to other situational and individual aspects. For instance, the utterance starts in a high register, and pre-focally the F_0 range is quite narrow. Note also the gentle final rise in F_0 . This kind of 'open-ended' intona-

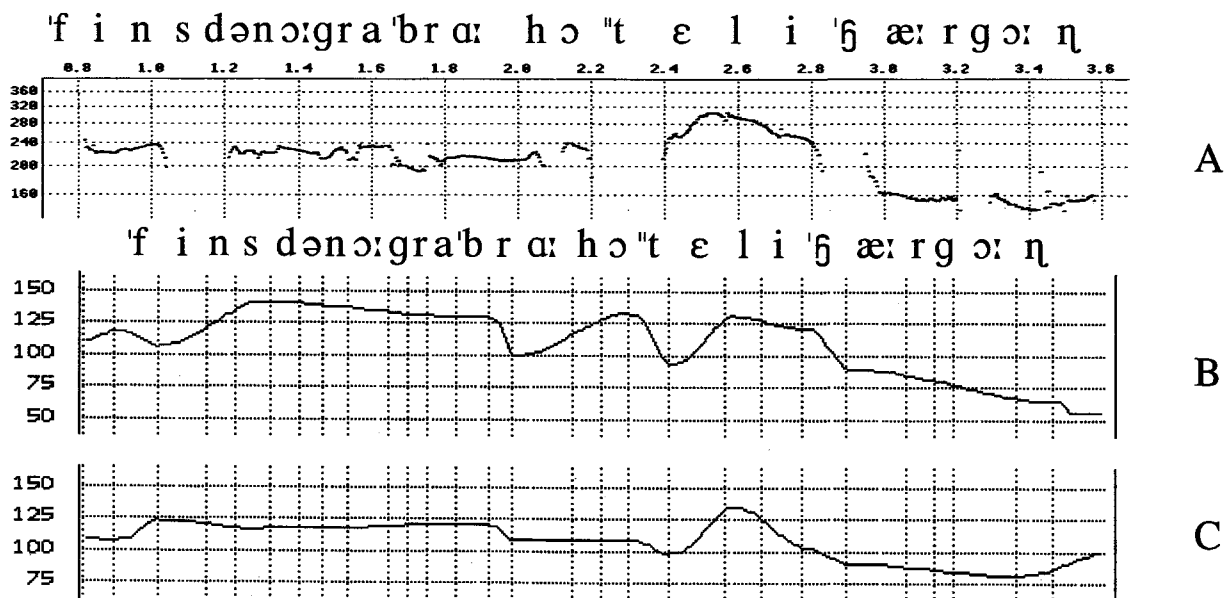


Figure 1 Human-produced utterance (A) and synthesis using the parametric approach based on this utterance.

B: Text input with information about focal accent. C: Text input includes dialogue-related markers. The F_0 register used is that for our standard male synthesis. Note that microprosodic effects are not displayed in B and C and that durational aspects of the human-produced utterance are not included in the synthesis. The utterance is "Finns det några bra hotell i skärgården?" (Are there any good hotels in the Archipelago?)

tion is characteristic of many of the utterances in our database, and seems to be used to signal politeness and other 'positive' attitudes in the dialogue situation. The pattern is particularly characteristic of this speaker and can therefore be employed as part of a characterization of her speaking style.

Figure 1 B and C shows the F_0 traces of two synthetic versions of the same utterance. That of 1 B has been generated from orthographic text with information about focus placement as the only extra addition. The input for 1 C included just three markers, indicating:

- the level of emphasis to affect the focussed word.
- the F_0 characteristics of the prefocal domain and the F_0 slope for the whole utterance
- the details of the utterance-final F_0 gesture

With this moderate amount of hand editing we achieve a high degree of similarity with the prosodic shape of the original as seen in the diagrams and as confirmed by informal listening.

3.2. Resynthesis from a prosodic transcription

Our other approach is to use an analysis/resynthesis procedure. Our point of departure is a categorization and symbolization of the basic, prosodic functions, prominence and grouping.

The transcription of the dialogues is made by an expert, based on a purely auditory analysis. The transcription results in a sequence of boundary and tonal labels (see Figure 2). The alignment of the tonal labels is with the CV boundary of the stressed syllable, and the alignment of the boundary labels is with the start and end

points of the speech or group boundaries. The labels for the orthographic words are right-aligned in the space that they occupy.

The prosodic transcription discussed here thus represents a phonological analysis of prominence and grouping. It also constitutes the input to the resynthesis module. In the implementation of the intonation model, the prosodic information contained in the transcription has to be supplemented with phonetic rules which will take care of the more specific timing of prominences, pitch level and range (including focus realization), F_0 drift (downdrift, downstep, upstep), as well as the interpolation between turning points, as discussed below.

The use of the analysis/resynthesis method in the present framework has a double purpose. The first, more direct goal is to verify/falsify the prosodic transcription. This will give us feedback on the correctness of the transcription and will reveal any incorrect auditory judgements about focus placement, prominence levels, phrasing and the like. The second, more long term goal is to use the analysis/resynthesis tool for developing our intonation model in a dialogue prosody framework, as a first step towards improving the generalized model used in the text-to-speech system. As our prosodic transcription covers only prominence and grouping, other aspects of intonation are thus not explicitly modelled in our current resynthesis.

The process of generating an F_0 contour involves transforming the phonological prosodic transcription of the analysis into a set of labels marking the location of the turning points of the F_0 curve. The basic labels of this transform are H and L, which can be followed by * or by

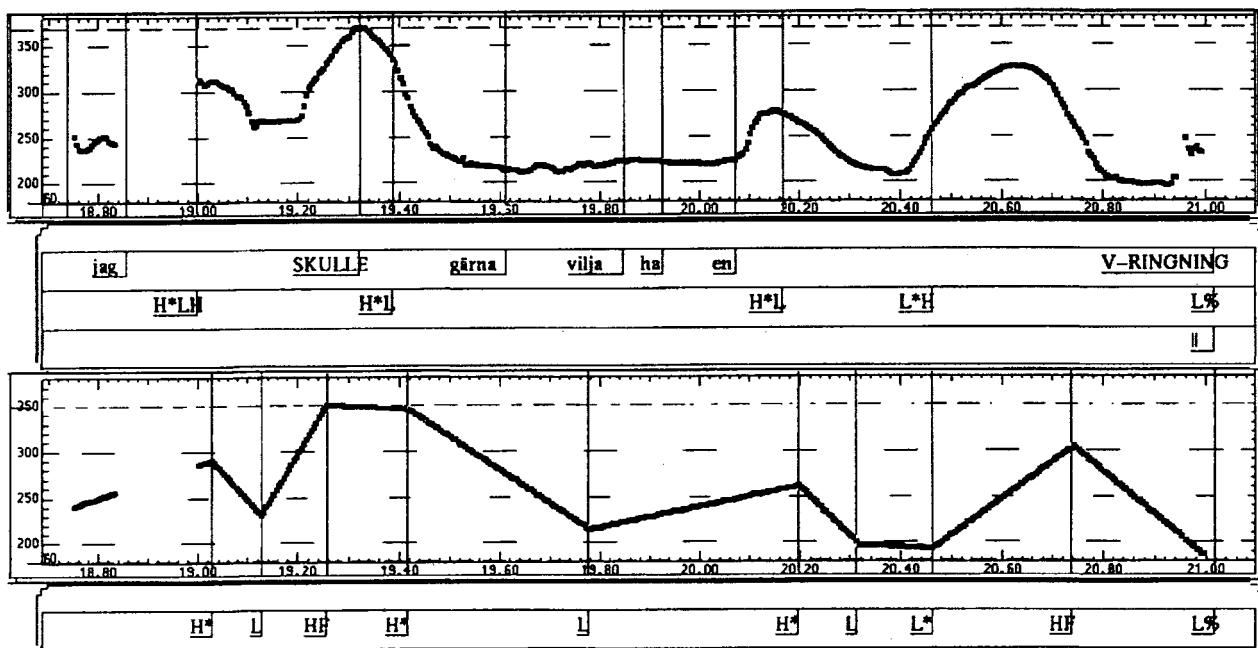


Figure 2 An example using the resynthesis approach on an utterance from a spontaneous dialogue: an original F_0 contour (upper part) and the model-generated F_0 contour (lower part). The utterance is "Jag skulle gärna vilja ha en V-ringning" (I would quite like to have a V-neck)

% to mark the presence of an accent or a boundary. The transformation is achieved by means of simple rules. It is important to note that the only input to the system is comprised of the transcription labels and their time alignment. This gives us a segmentation of the speech signal into stretches corresponding approximately to stress groups or feet (aligned with the CV boundary of the stressed syllable). We have no other segmental information available, nor any information about voiced/voiceless distinctions. This, of course, limits the amount of information which can be included in the rules for placing the turning points, because we cannot refer to e.g. vowels or syllables as points of reference in the speech signal. In general, the starred (*) turning point is placed at the location, or very near (30 ms after) the location of the transcription label. Preceding H's are placed a fixed number of milliseconds (30) before the location of the label. Succeeding turning points are spaced equally between the locations of the current transcription label and the next label. This solution may seem to be *ad hoc* but is not without motivation in production data. In a study by Bruce [9] variability in the timing of the pitch gesture for focal accent relative to segmental references was demonstrated. Instead, disregarding segmental references and using the beginning of the stress group as the line-up point, there appeared to be a high degree of constancy in the timing of the whole focal accent gesture. It should be noted that the actual numbers in milliseconds of the implementation are at this point chosen as test values partly based on earlier work on tonal stylization [10, 11].

In addition to the basic rules for the transformation of transcription labels, an additional set of rules has been developed to capture different phenomena we have found in our production data.

Finally, there is a large set of option rules or settings in the system. These options control such things as range, floor, downdrift and height of focal gestures and are selected experimentally to test specific hypotheses. Relevant options are: base level (Hz), floor (Hz), range (Hz), focal multiplier, maximum time for a rise within a label, maximum time for a fall within a label and maximum time for a rise between two labels.

Next, when all turning points are placed properly, a contour is generated by interpolation. At the moment we are using simple linear interpolation, but other options will be exploited, such as cosine functions and splines.

In the top part of Figure 2 the original F_0 contour of a test sentence can be seen. Below are an orthographic transcription and a prosodic transcription with the tonal and the boundary tiers. Further below is a model-generated F_0 contour of the same sentence, with the turning points as they have been placed by the algorithm.

4. DISCUSSION

A prosodically sophisticated TTS system needs to use, explicitly or implicitly, a prosodic labelling system. This is an intermediate step in the process of generating prosodically relevant acoustic parameters from a text. In order to arrive at appropriate labels different methods

can be employed. We are following two lines of research in order to approach this goal. With both methods we can model synthetic speech on production data. This allows us to test hypotheses on dialogue structure and on pragmatic and situational aspects of speech. By being coupled to automatic dialogue systems the parametric approach will allow us to explore the prosodic aspects of both the human and the machine partner in automatic spoken dialogue system. The results emerging from our work will be of relevance not only for the construction of text-to-speech systems applicable to dialogue situations, but also for work on the prosodic basis of automatic speech recognition systems.

ACKNOWLEDGEMENTS

This work has been supported by grants from The Swedish National Language Technology Programme.

The authors are very grateful to Grzegorz Dogil and Gregor Möhler of IMS, University of Stuttgart, for sharing their implementation of PSOLA synthesis with us, and for their invaluable assistance in porting it to our platform.

REFERENCES

- [1] Bruce, G., Granström, B., Gustafson, K., House, D. and Touati, P. (1994), "Modelling Swedish prosody in a dialogue framework", Proc. ICSLP 94, pp. 1099-1102, Yokohama.
- [2] Cudd, P.A., Hunnicutt, S., Arthur, J., Granström, B., Aguilera, S., Waernulf, B., Dalsgaard, P. and Wilson, G. (1995), "Voices, Attitudes and Emotions in Speech Synthesis", In Placencia Porrero and Puig de la Bellacasa (eds.) The European Context for Assistive Technology. Proceedings of the 2nd TIDE Congress, pp. 344-347. Amsterdam: IOS Press.
- [3] Bruce, G. and Granström, B. (1993), "Prosodic modelling in Swedish Speech synthesis", Speech Communication 13, pp. 63-73.
- [4] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992), "ToBI: a standard for labelling English prosody", Proceedings of ICSLP 92, pp. 867-870, Edmonton.
- [5] Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D. and Touati, P. (1995), "Towards an enhanced prosodic model adapted to dialogue applications", Proceedings of ESCA Workshop on Spoken Dialogue Systems, Vigsø, pp. 201-204, Aalborg.
- [6] Moulines, E. and Charpentier, F. (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication 9, 453-467.
- [7] Moehler, G. and Dogil, G. (1995), "Test environment for the two level model of Germanic prominence", to appear in Proceedings Eurospeech '95, Madrid.
- [8] Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993): "An experimental dialog system: WAXHOLM", Proceedings of Eurospeech '93. pp. 1867-1870.
- [9] Bruce, G. (1986), "How floating is focal accent?", In Gregersen and Basbøll (eds.) Nordic Prosody IV, pp. 41-49, Odense University Press.
- [10] House, D. (1990), Tonal perception in speech. Lund University Press, Lund.
- [11] House, D. and Bruce, G. (1990), "Word and focal accents in Swedish from a recognition perspective", In Wiik and Raimo (eds.) Nordic Prosody V, pp. 156-173. Turku University.