DEVELOPING THE MODELLING OF SWEDISH PROSODY IN SPONTANEOUS DIALOGUE

Gösta Bruce, Marcus Filipsson*, Johan Frid*, Björn Granström**, Kjell Gustafson**, Merle Horne*, David House*, Birgitta Lastow* & Paul Touati* (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

The main goal of our current research is the development of the Swedish prosody model. In our analysis of discourse and dialogue intonation we are exploiting model-based resynthesis. By comparing synthesized default and fine-tuned pitch contours for dialogues under study we are able to isolate relevant intonation patterns. This analysis of intonation is related to an independent modelling of topic structure consisting of lexical-semantic analysis and text segmentation. Some results from our model-based acoustic analysis are presented, and the implementation in text-to-speech-synthesis is discussed.

1. INTRODUCTION

The object of study in the project *Prosodic Segmentation and Structuring of Dialogue* is the prosody of spontaneous dialogue in a language technology framework [1,2]. The ultimate goal of our research is the development of a more powerful prosody model. In our study we are employing a methodology containing analysis of discourse/dialogue structure (independent of prosody), prosodic analysis - both auditory analysis in the form of prosodic transcription and acoustic-phonetic analysis (based on F0 and waveform information) - as well as speech synthesis (model-based resynthesis, text-to-speech). These different analysis types involving both symbol and signal information and model-based resynthesis are combined and synchronized with each other in the same ESPS/Waves+ environment. The labelling used (symbol information) consists of several tiers: an orthographic tier, a tonal tier, a boundary tier, a discourse referent tier, and a textual segmentation tier (see also Figure 1). In our work we are exploiting speech material from the national Swedish prosodic database under development. The dialogues under study cover true spontaneous conversations, spontaneous but more restricted dialogues, and read dialogues from scripts.

2. MODEL-BASED RESYNTHESIS

We are exploiting model-based resynthesis as a tool in our analysis of the prosody of dialogues. Prosodic characteristics of speech are analyzed auditorily according to a prosodic model [1,3]. In this model, we label prominence levels (word accents and focal accents), boundary tones, and phrase boundary strengths (minor and major). The transcription labels and their temporal alignment with acoustic events are supplemented with phonetic rules for the specific timing of pitch targets, interpolation between them, and parameter values that control the local and global pitch register and range. This results in F0 specifications that are used as inputs for the resynthesis, for which we use an implementation of the PSOLA technique [4].

The analysis-by-synthesis method is used in the acoustic analysis of F0 trends. We thereby model F0 contours by altering the following parameters: F0 register (start and end values) and F0 range (regular and focal values), cf. [5]. These parameters are used to model F0 contours of phrases using two different approaches:

1. a 'fine-tuned' contour, in which the parameter values are set on the basis of measurements of the original F0 contour of a phrase, cf. [6]. This provides us with a simple description of the optimal frequency parameters for a particular phrase. Phrases are resynthesized and perceptually evaluated.
2. a 'default' contour, where the parameter values are held constant, which results in similar values for all phrases. The only inter-phrasal variation is thus the timing and the identity of the prosodic labels.

By comparing these two kinds of contours, we are able to see deviations from the current prosodic model, which does not contain regulation of discourse prosody, e.g., topic structure, dialogue features, etc. The places of major deviation need additional modelling that can govern the setting of parameter values appropriately [7]. The idea is to relate an analysis of topic structure to intonational patterns in dialogues and extract possible correlations. On the basis of the results, a model of the textual influence on intonation is created, i.e. rules for parameter value generation are constructed. This modelling of textual influence is implemented alongside the auditory analysis. The result is an extended prosody model, where we control both local (accents and boundary tones) and global (F0 trends within phrases, relations between phrases) aspects of F0. Resynthesis is then again utilized to evaluate the results. Figure 1 is an illustration of the methodology.

3. TOPIC STRUCTURE MODELLING

3.1. Lexical-semantic relations

A topic structure can be thought of as the result of the lexical, grammatical and semantic/pragmatic parameters that interact to create the text/conversation. A discourse is said to be ‘coherent’ if the topic structure is clear and easy to follow. The sentences or utterances are logically linked to each other and

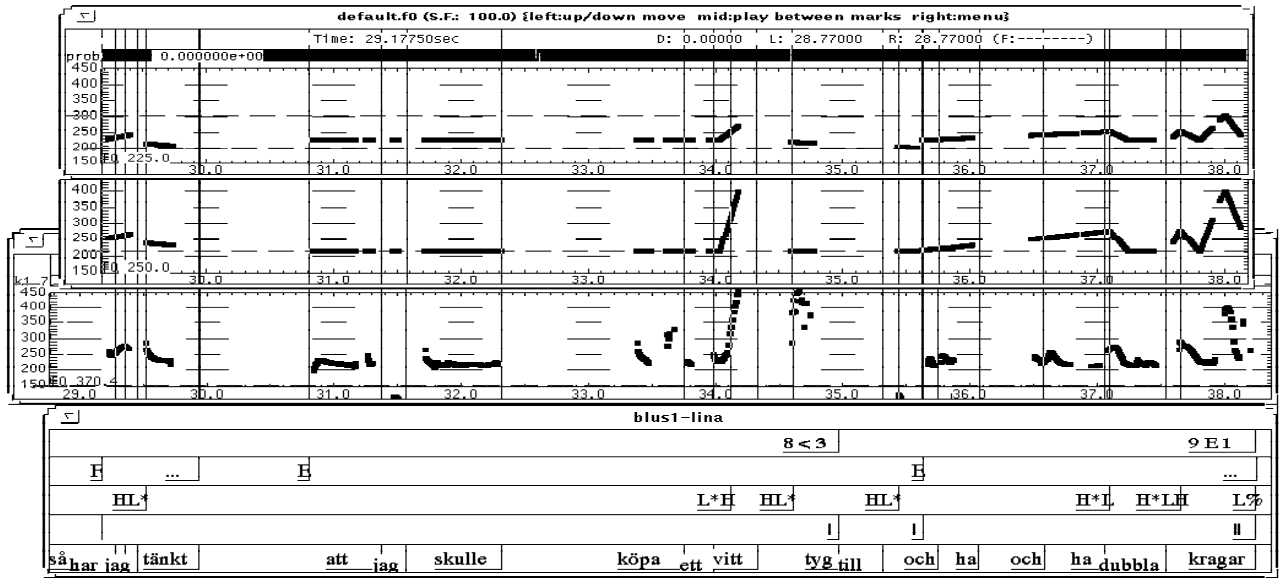

Figure 1: From top to bottom: Default, fine-tuned, and original F0 contours of the utterance '*then I thought I´d buy white material and have and have a double collar*' . Below is shown the corresponding labelling for lexical analysis, text segmentation analysis, prominence plus boundary tones (Hs and Ls and their combinations), phrasing (minor = '|', major = '||'), and orthography.

are relevant in the context where they are used. According to [8], coherence is created by the use of a number of 'cohesive devices'. One of these devices is coreference or cospecification, since, in order to know if one is still speaking or writing about the same topic, there must be some way of referring back in the discourse to a referent that has been mentioned earlier in the text/conversation. Content words are related to each other by morphological identity and lexical semantic relationships (synonymy, hyponymy, and partonymy/meronymy) [9]. In [10] it was shown how these relations can be tracked computationally in a linguistic preprocessor to a text-to-speech system. The information on cospecification can then be used in the F0 generating component in order to appropriately assign focal and nonfocal word accents.

Although the tracking of these lexical relations was initially developed for predicting accent assignment within a restricted domain, it is possible to extend the modelling of lexical semantic relations to cover more domains and thus describe larger 'semantic frames' that define prototypical scenarios, institutions, etc. (cf. [11]). One can imagine these frames as networks where there are connections between referents in different semantic fields.

In order to illustrate this type of frame analysis, we will show the lexical structure of the frame 'Recipe for a Hot Tuna Fish Sandwich' as it develops in a dialogue from a Radio Sweden program where the guest is asked to present a favourite recipe for the listening audience. A translation of an excerpt from the dialogue is presented in (1):

(1) **Guest**: (breathing) uh I'm not really any kind of experienced cook uh but I do have a recipe for a hot sandwich which I've in fact develop developed a little (breathing) uh it's a tuna fish sandwich and you make it like this you have white bread for example (breathing) and on that you put a mishmash of uh it makes rather a lot but you have to have a can of tuna fish, a package of crème fraiche just tuna fish in water uh is good otherwise there's so much fat (breathing) and then a package of crème fraiche and then about a third of a jar of mayonnaise preferably light mayonnaise there too since there's crème fraiche in (it) (breathing) and then just a tiny dab of mustard it can be strong mustard

Interviewer: French ; **Guest**: Yes ; **Interviewer**: Scanian...

Guest: uh and then a lot of chopped leeks and a dash of Italian salladspice and then you stir all that up together nice and even you know (breathing) and then I think it's a good idea if you let it sit and rest a couple of hours so that the taste spreads through the whole thing and then you slap it on those slices of uh bread and then in the oven with them and if you want you can put...

As can be seen in Figure 2, we are including inferences of the type *cook← recipe* in the frame model as well as the traditional lexical semantic relations of hyponymy, partonymy and synonymy, so as to capture all the lexical relationships between referents (see also [12] for the importance of modelling inference in algorithms for discourse segment boundary detection).

In order to include this discourse information in the database, we have developed a method for labelling the lexical relations described above. These labels are inserted on a separate discourse-structure level in the Waves environment along with the prosodic labels. Following in (2) are the labels used:

(2)

x	The discourse referent (DR) x is not related to any other DR
= x	The DR x is morphologically identical to or a pronominalized form of a preceding DR
x = y	The DR x is a synonym of the DR y
x < y	The DR x is a hyponym of the DR y
x > y	The DR x is a hyperonym of DR y
x E y	The DR x is part of the DR y
x Z y 1 ... y n	The DR x is the sum of the DR's $y_1 ... y_n$
[x]	The DR x is a superordinate (non-basic) term.
x ← y	The DR x is inferable from the DR y

These lexical relations have subsequently been used to explain some unexpected patterns of accentual downtoning in the data. Downstepping of word accents, for example, has been observed to occur in a number of cases of lexically 'new' information in this dialogue [13]. This is not what one would expect since new information is generally accentually highlighted. However, it has been seen that this downstepping correlates with certain aspects of lexical semantic structuring. Accentual downtoning is associated with lexically new information in two environments: the first is when the new information is realized by a superordinate 'non-basic' word such as *mishmash* ([x]) and second when the new information is a specification which is in some sense non-central to the development of the topic, such as *tuna fish in water* (x<yEz) i.e. *tunafish in water* (x) is a specification (<) of *tunafish* (y) which is in turn a part of (E) the *mishmash* (z).

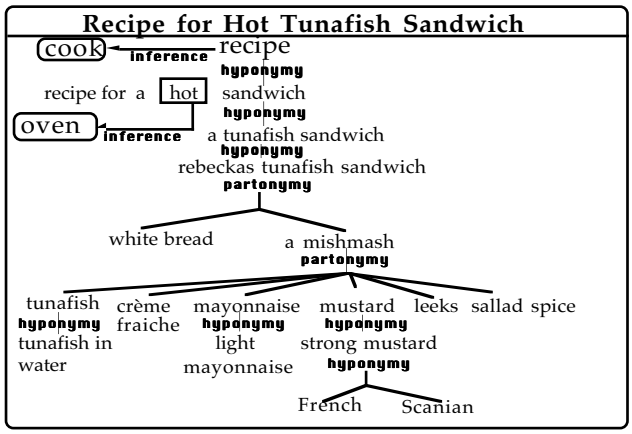

Figure 2: Frame semantic topic structure representation for dialogue fragment in (1).

3.2. Text segmentation

There exists a variety of coding systems for text analysis, based on different principles, e.g., the Initiative/Response system proposed in [14]. We are currently developing a textual topic anlysis model that is based on a combination of the lexical-semantic analysis and functional grammar. A strictly textual approach is chosen since it avoids the circularity of including prosodic features in the definition of discourse structure.

The verbatim transcriptions of spontaneous dialogues are divided into segments. The lexical items are classified according to a simple model of functional grammar. The categories used are Subject, Predicate/Verb, Object and Attribute. Segmentation of the text is performed so that each segment maximally contains one subject and one (compound) verb. 'Connector' words (and, or, if, so, etc.) are used to locate the boundaries between segments. This means that a segment contains one subject, one (compound) verb, and the objects and attributes related to them. Classification of the segments is based on the word items that segments contain and their properties according to the "Lexical Relations" described in the previous section. The following categories are used:

Initial (I), segment containing a DR not related to any other previous DR.

Expansion (E), segment containing a DR related to a previous DR by hyponymy, partonymy, implication or summation.

Continuation (C), segment containing coreference or iteration of previous referent; or a DR related to a previous DR by morphology or synonymy; or by pronominalization, coordination (elliptic subject), or by a correlate; or an assignment of an attribute to a DR.

Follower (F), segment not containing referential material, objects or attributes.

Summary (S), evaluational expressions, judgements and opinions.

Mistake/Reparation (M/R), speech errors and the announcement of them ("... I almost said").

4. MODEL-BASED ACOUSTIC ANALYSIS

In this section some results of an analysis of a conversation between two female speakers of Swedish are presented. The part under study deals with the sewing of a blouse and includes such aspects as what material to use and where to find patterns (intuitively regarded as sub-topics). Textual and lexical-semantic analyses were performed for this dialogue in order to obtain segmentation and topic classification of the segments (see example in Figure 1). The fine-tuned method described above was used to model an F0 contour for each segment, and the parameter values used in that modelling were examined.

It has long been known that speakers make use of F0 trends over long stretches of speech (e.g. [15,16,17]). In order to measure for similar effects in our data, we examined the successive Focal Range parameters for each segment. The results are plotted in Figure 3. It can be seen that the Focal range parameter exhibits a global acoustic characteristic with ranges decreasing towards the end. One possible explanation for the high focal ranges in the beginning could be that they introduce previously unused information that will constitute the frame for the conversation to follow. They are realized as extra salient in order to signal that they should have a prominent status in the mind of the listener. The decreasing saliency of the focal accents in the following part of the

dialogue could be analyzed in the same way; when the frame is set, speakers emphasize new information to a lesser degree. As the context grows, concepts can be made less salient as more and more can be inferred from previous context.

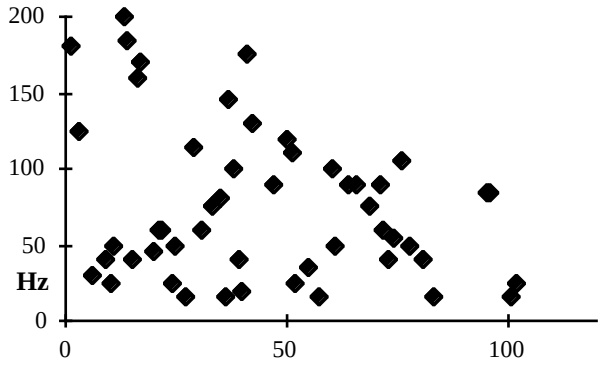

Figure 3: Fine-tuned focal ranges in Hz for each successive segment of a section of a dialogue.

5. TEXT-TO-SPEECH SYNTHESIS

We have developed the KTH text-to-speech (TTS) system in such a way that we can use it to study prosodic aspects of discourse and dialogue in parallel with the analysis-resynthesis method described above. By specifying a number of prosodically relevant parameters, we can vary F0, as well as segment and pause duration, in a systematic way and study the effects of these manipulations. After achieving prosodically good resynthesis of utterances in our databases, we have selected a number of parameter settings which, encoded in orthographic segments, can be inserted manually in the text input to the TTS system. The next stage, which we are now working on, is to include this in a working man-machine dialogue system, the Waxholm system [18].

6. ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK). Gayle Ayers was a guest researcher in Lund during several months during 1993 and 1994 and has contributed to the project.

7. REFERENCES

1. Bruce, G., Granström, B., Gustafson, K., House, D. and Touati, P. “Modelling Swedish prosody in a dialogue framework”, *Proceedings ICSLP 94*, (Yokohama) Vol. 3, 1099-1102, 1994.

2. Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D. and Touati, P. “On the analysis of prosody in interaction”, to appear in *Computing Prosody*, Springer-Verlag, Berlin, 1996.

3. Bruce, G., Granström, B., Filipsson, M., Gustafson, K., Horne, M., House, D., Lastow, B. and Touati, P. “Speech synthesis in spoken dialogue research”, *Proceedings Eurospeech 95* (Madrid) Vol. 2, 1169-1172, 1995.

4. Moehler, G. and Dogil, G. “Test environment for the two level model of Germanic prominence”, *Proceedings Eurospeech 95* (Madrid) Vol. 2, 1019-1022,1995.

5. Gårding, E. “On parameters and principles in intonation analysis”, *Working Papers* 40, 25-47, Lund University, Department of Linguistics, 1993.

6. t'Hart, J., Collier, R. and Cohen, A. *A perceptual study of intonation*, Cambridge U.P., Cambridge, 1990.

7. Bruce, G., Frid, J., Granström, B., Gustafson, K. and House, D. “The Swedish intonation model in interactive perspective”, *Proceedings Fonetik 96* (Nässlingen), TMH-QPSR Vol.2, 1996 (in press).

8. Halliday, M. and Hasan, R. *Cohesion in English*, Longmans, London, 1976.

9. Cruse, D. *Lexical semantics*, Cambridge U.P., Cambridge, 1986.

10. Horne, M., Filipsson, M., Ljungqvist, M. and Lindström, A. ‘‘Referent tracking in restricted texts using a lemmatized lexicon: implications for generation of prosody’’, *Proceedings Eurospeech ´93* (Berlin) Vol. 3, 2011-2014, 1993.

11. Metzing, D. ‘‘Frame representations and lexical semantics’’. In Eikmeyer, H-J. och Rieser, H. (eds.) *Words, worlds and contexts*, 320-42, de Gruyter, Berlin, 1981.

12. Litman, D. and Passonneau, R. ‘‘Combining multiple knowledge sources for discourse segmentation’’, *Proceedings, 33 ACL* (Association for Computational Linguistics), 1995.

13. Ayers, G., Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D. and Touati, P. “Modelling intonation in dialogue”, *Proceedings ICPhS 95* (Stockholm), Vol. 2, 271-281, 1995.

14. Sinclair, J. and Coulthard, M. *Towards an analysis of discourse: The English used by teachers and pupils*, Oxford University Press, London, 1975.

15. Lehiste, I. “The phonetic structure of paragraphs”, In A. Cohen and S. Nooteboom (eds.) *Structure and Process in Speech Perception*, 195-206, Springer-Verlag, Berlin, 1975.

16. Bruce, G. “Textual aspects of prosody in Swedish”, *Phonetica*, Vol. 39, 274-287, 1982.

17. Swerts, M. *Prosodic features of discourse units*. Doctoral dissertation, TU Eindhoven, 1994.

18. 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. and Ström, N. “The Waxholm system - a progress report”, *Proceedings ESCA Workshop on Spoken Dialogue Systems* (Aalborg), 81-84, 1995.