TEXT-TO-INTONATION IN SPONTANEOUS SWEDISH

Gösta Bruce, Marcus Filipsson*, Johan Frid*, Björn Granström**, Kjell Gustafson**, Merle Horne* & David House* (names in alphabetical order)*

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

ABSTRACT

This paper deals with a number of aspects of intonation in spontaneous dialogues in a language technology perspective. The key topics to be addressed are: I) the analysis of global intonation and its interaction with textual structure, II) the implementation of global and textual aspects of discourse intonation in an analysis-by-synthesis environment. We present models for the analyses of intonation and textual content in spontaneous conversations in Swedish. The models are implemented in a computational environment, making it possible to generate F0 contours, which can be imposed on a speech waveform using the PSOLA technique. The result is a text-to-intonation system, where textual and lexical analyses automatically generate hypothetical intonation contours, which can be evaluated through resynthesis, and eventually be used in a text-to-speech system.

1. INTRODUCTION

The work presented here is performed within the framework of the research project 'Prosodic segmentation and structuring of dialogue', whose goals are to increase the understanding of how prosody is used in interactive speech, and to develop a prosodic model based on this understanding.

Our current research has concentrated on implementing a system for the generation and consequent synthesis of F0 contours, on the basis of analyses of intonation and textual content. The intonational component of the system is based on a model of Swedish intonation that accounts for:

- word and focal accents
- juncture tones
- downstepping
- shifts in register and range

The textual analysis concentrates on dependencies between propositions. For instance, we have observed effects of contextual factors in the expression of F0. Successive phrases that impressionistically share a common theme have been observed to exhibit a slowly decreasing intonational register, possibly in order to signal the coherence between them. To capture this, we have developed a rule system that uses relational notions and devices such as co-ordination of utterances, repetition of morphemes, lexical-semantic relations such as meronymy as well as pronominalisation, in order to predict the register and range configurations for utterances. This has also been incorporated as a component of the F0-modelling computer program.

2. ANALYSIS OF GLOBAL INTONATION

Our intonational model represents global aspects of intonation by means of two parameters: *Register* and *Range*. Register refers to a base level of an utterance's intonation, roughly corresponding to the F0 level of the unaccented portions of an utterance. By Range we mean the height of a pitch gesture, starting from the register base level.

We have attempted a global analysis according to a model recognising three register levels (High, Mid and Low) and four range categories (High, Mid, Low and Flat) for each intonational phrase (cf. [1]). The register can also be Decreased, which indicates a more gradual (less than a full shift to Low) decrease of register. The absolute values of the categories are speaker-dependent and are determined after examination of production data of each individual speaker. Thus, different speakers will have different values for e.g. High register.

This results in a detailed and powerful notational device that can account for the intonation over several phrases. Together with a transcription of accentuation and grouping, it can produce very accurate close-copy modelling of phrase intonation. However, in a perspective of generating intonation for successive utterances, a mechanism that allows unrestricted variation as to the sequences that these categories are allowed to appear in would have too much descriptive power, and would perhaps also allow physiologically implausible configurations. Therefore, it is necessary to constrain the sequences of range and register which should be allowed.

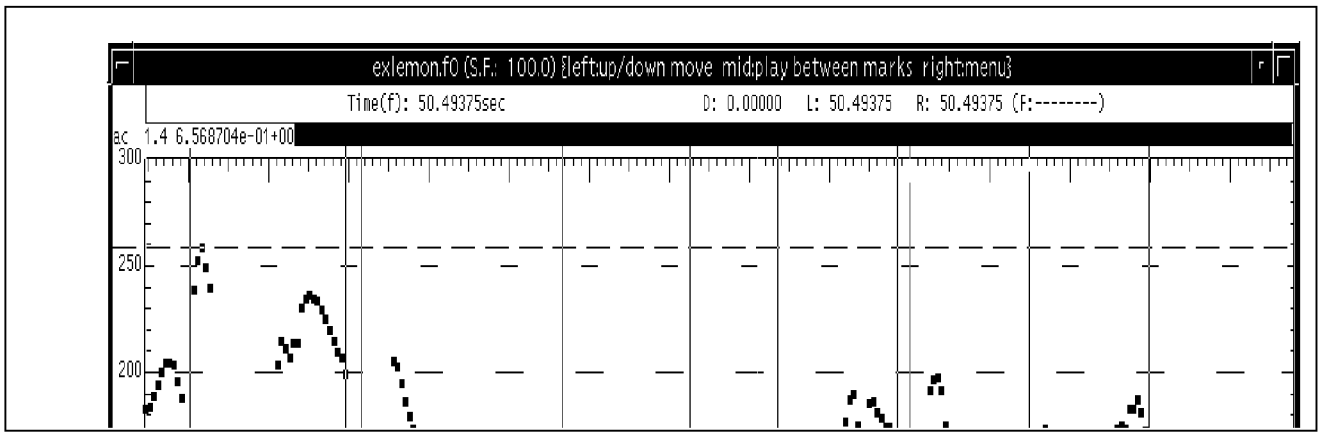

Figure 1. An utterance with two phrases, where the second one uses a lower register. See Figure 2 for a translation. The vertical lines in the lower tier symbolise major (‖) and minor (|) phrase boundaries.

As a working paradigm, we have chosen a very simplified version using only two combinations of the register and range categories: either Mid or Decreased register, combined with Mid range. These are roughly interpretable as, respectively, the 'raised' and 'non-raised' categories used by Brown, Currie and Kenworthy in [2].

3. TEXT ANALYSIS AND GENERATION OF INTONATION LABELS

The idea that forms the main hypothesis in our work on textual influence on dialogue and discourse prosody is that the structuring of conversational topics, by means of signalling topic coherence (continuation) and topic boundaries (shifts), has important reflections in the intonation of utterances. Earlier attempts at this approach have been reported on in [3] and [4], where we introduced a lexical-semantic analysis of the referent structure in a discourse. This accounts for the given/new structure of referents, and thereby the accentuation/deaccentuation in noun phrases. There are, however, relationships between other constituents in a discourse that can contribute to the overall structuring of the coherence of conversational topics. This can be illustrated by studying the intonation of the utterance in Figure 1. It consists of two phrases, that exhibit similar syntactic structures and discourse functions, giving an opinion on something. In this speech segment, we would like to model the second phrase with a lower register than the first phrase, i.e., (in our model) with Decreased register. But there is nothing in the noun phrases that we can utilise to relate these phrases. There is, however, the similar syntactic structure, the conjunction 'och' (and) and the repeated attribute 'jättebra' (very good).

In order to capture features of discourse coherence that are not dependent on the lexical-semantic structure of discourse referents, an extended method of analysis has been developed. It can be described as a hybrid of grammatical and semantic roles, as well as lexical-semantic structure. The initial step is an analysis of the lexical elements in a text, according to which they are classified in different categories. A description of the categories follows below.

Adverbials of Time, Manner and Space - they give the setting of the events being talked about.

Subject - the noun phrase with the referent that often performs an action (i.e., agent).

Verb - the notional component of a verb complex, that denotes the action, state or event.

Object - the noun phrase with the referent that often undergoes an action (i.e., patient).

Phrasal attributes - text units altering or evaluating the relationships between subjects, verbs and objects

Phrase starters, Phrase enders - transitional elements, conjunctions, relative pronouns that occur at the beginning and/or end of phrases

The categories form a partitioning of the dialogue text into analysis units, corresponding to grammatical functions. Whenever a new main verb is found a new analysis unit is formed. If any text unit is analysed as a category that already exists in an analysis unit, a new analysis unit is also formed.

Dialogue

EH hur funkar de då
EM det gör att *em* det hjälper till med matsmältn när man *eh* alltså smälta maten det är jättebra att starta dagen så och jättebra att väcka magen på det sättet

(Translation:)

EH how does it work then
EM it makes *um* it helps with the food-digest when you *uh* I mean digesting food it´s really good to start the day like that and very good to wake up your stomach that way

Analysis

PHRASE STARTER	SUBJECT	NOTIONAL COMP. OF VERB	ATTRIB.	OBJECT	PHRASE ENDER	LABEL
hur	det	funka			då	Mid
	det	göra			att	Mid
	det	hjälpa till med		matsmältn		Mid
när	man					Decr.
alltså		smälta		maten		Decr.
	det	är (=vara)	jättebra		att	Mid
		starta		dagen	så	-
och			jättebra		att	Decr.
		väcka		magen	på det sättet	-

Figure 2. Example of analysis of a dialogue fragment.

Relative pronouns and conjunctions also mark a new unit, since they often coincide with an intonational phrase boundary. The infinitive marker 'att' (to) is associated with the end of a new analysis unit [cf. 5]. However, the following unit is not assigned an intonation label of its own if it lacks a subject.

Performing this kind of analysis on spontaneous speech is by no means as trivial as it may appear here, but these categories are believed to represent a reasonable blend of grammatical detail and analytical ease. We do not claim to have found the ideal set of constituents, but these categories have emerged as a working system for our purposes.

Connections between categories in different analysis units are then identified by means of a set of rules, e.g. the lexical-semantic relations.

The connections are then used to generate the intonational labels that were introduced earlier. The basic principle of label assignment is: as long as there is an element in an analysis unit that has a connection with an element in the immediately preceding analysis unit, it gets the label Decreased, but if there is no connection it gets Mid. If there is a textual link between two consecutive units, their coherence is signalled by using a non-resetting phrase intonation. If there is no obvious semantic relation, a boundary is signalled by resetting the register in the second phrase. In the example in Figure 1, the second phrase will be assigned the label Decreased, because of the repeated attribute 'jättebra', which is lexically identical to the attribute in the first phrase. Figure 2 shows an example of the analysis of a dialogue fragment. Units containing speech errors, mistakes, and hesitations are not counted as previous units but are disregarded. In these cases the previous unit is thus defined as the one prior to the unit containing the error, or hesitation.

4. IMPLEMENTATION OF THE INTONATION MODEL AND RESYNTHESIS

Our intonation model combines global (register and range) and local (accentuation and phrasing) features. The accentual analysis is performed according to a model recognising two levels of prominence: accented and focused, as well as the two Swedish word accents on both levels. We also distinguish high or low boundary tones. This system is thoroughly described in [6], and here we will describe the new features in the model. We

have incorporated rules to model downstepping and have also included the prosody-independent text analysis that allows us to model shifts in register from one utterance segment to another. The input to the F0-modelling system is a prosodic transcription. This could be based either on an analysis of a pre-recorded (spontaneous) dialogue or on a text to be synthesised, for which prosodic labels have been generated.

Phrase boundaries are classified as either minor or major, and different intonational characteristics of phrase boundary marking have been implemented. A minor boundary is marked by a fall to a level somewhere between the starting register and the speaker´s floor. No register reset is made for the next phrase, and it continues on the same level. A major phrase boundary is marked by a fall to the speaker´s floor as well as a reset to the starting register of the next phrase. We have also included a rule of downstepping. Downstepping occurs after a focal accent in a phrase, and continues after a minor phrase boundary, unless the next phrase also contains a focal accent. A major boundary, however, breaks the downstepping. Downstepping is realised by lowering the valley after an accent compared to the valley before the accent. Thereby the register decreases successively towards a major boundary.

This presupposes that the speech waveform is divided into discrete units of different lengths. For each unit, a number of parameters can be set in order to represent intonational features. These parameter values can be derived from the textual analysis, allowing us to generate an F0 contour for any given utterance. In the generation of F0-contours the labels representing global intonation have been incorporated in our model. Each Mid is taken to represent a major phrase boundary, and each following Decreased represents the minor phrase boundaries within a major phrase. The F0 contours are resynthesised using the PSOLA technique ([7], see also [8]), which lets us use the new F0 contour, which leaves all other signal information intact.

The rules of the textual analysis predict boundaries on a grammatical-functional basis. The correlation between grammatically determined clause boundaries and prosodically analysed phrase boundaries is not always perfect, and this causes some mismatches in the resynthesis approach, since some phrase boundaries may appear in other places than in the original. This problem is, however, limited to resynthesis.

The generation-resynthesis system enables us to evaluate the auditive analyses and the models, to enhance them further, and add the influence of other types of analysis. The system could also be used as the intonation component of a text-to-speech-system, where intonation contours can be generated from an automatic analysis of text and lexical lookup of accents. In our studies of intonation, we also use a hybrid, where we mix the analytical and automatic approaches.

5. DISCUSSION

Applied to a working man-machine dialogue system, one of the future tasks is to include our improved prosody model, both in the recognition and in the synthesis part. The need for improved prosodic models, and good implementations of these, has become increasingly recognised in recent years in the area of man-machine dialogue systems. On the speech-recognition side, better prosodic models can contribute to better interpretations on the syntactic, pragmatic and contextual levels, as well as more accurate disambiguation on the lexemic level; on the speech synthesis side, it is well recognised that natural-sounding prosody increases the intelligibility of synthetic speech in addition to reducing the sense of fatigue that may occur in human participants not accustomed to synthesised speech. As feedback from this work we can expect to be able to fine-tune our model even further.

6. REFERENCES

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

[2] G. Brown, K. Currie & J. Kenworthy, *Questions of intonation*, London: Croom Helm, 1980.

[3] G. Bruce, J. Frid, B. Granström, K. Gustafson, M. Horne and D. House, *Proceedings of Nordic Prosody VII*, Joensuu, in press.

[4] G. Bruce, M. Filipsson, J. Frid, B. Granström, K. Gustafson, M. Horne, D. House, B. Lastow and P. Touati, "Developing the modelling of Swedish prosody in spontaneous dialogue", *Proceedings ICSLP 96*, vol. 1, 370-373, Philadelphia, 1996.

[5] M. Horne and M. Filipsson, "Computational extraction of lexico-grammatical information for generation of Swedish intonation", *Progress in Speech Synthesis*, J. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), 443-457, New York: Springer, 1996.

[6] G. Bruce, B. Granström, M. Filipsson, K. Gustafson, M. Horne, D. House, B. Lastow and P. Touati, "Speech synthesis in spoken dialogue research", *Proceedings EUROSPEECH 95*, vol. 2, 1169-1172, Madrid, 1995.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication* 9, 453-467, 1990.