# LUND UNIVERSITY

## Global features in the modelling of intonation in spontaneous Swedish

Bruce, Gösta; Filipsson, Marcus; Frid, Johan; Granström, Björn; Gustafson, Kjell; Horne, Merle; House, David

# GLOBAL FEATURES IN THE MODELLING OF 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

In this paper, we present a model for the analysis and synthesis of intonation in spontaneous conversations in Swedish. The model is implemented in a computational environment, making it possible to generate F0 contours, which can be imposed on a speech waveform using the PSOLA technique. Testing and exploration of this analysis-by-synthesis system on spontaneous conversational speech have led to recent extensions in the analytical framework, as well as in the actual realisation of the prosodic transcription as F0 contours. In earlier versions of the model, phrase boundaries were treated similarly for all phrases, but now we make a distinction between a minor and a major boundary. Different intonational characteristics of the type of phrase boundary have consequently been implemented. We have also introduced a rule of downstepping, which is realised by successively lowering the valleys after the pitch gesture of an accented syllable.

## 1. INTRODUCTION

The work presented here is performed within the framework of the research project 'Prosodic structuring and segmentation of intonation in spontaneous speech'. This paper we will report on some attempts to extend our intonation model with analytical tools to capture global aspects of intonation in spontaneous speech. This is done in order to be able to model F0 for several consecutive and semantically coherent phrases.

### 1.1. The intonation model

In several previous papers ([1], [2]) we have presented work based on an intonation model for Central Standard Swedish that involves categorisation of accentuation (prominence) and phrasing (grouping) of an utterance.

The categories are expressed using tonal turning points (H / L) with association to stressed syllables or boundaries and is the result of a perceptual analysis (see Table 1).

| Prosodic category | Label |
|---|---|
| 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 |
| Terminal juncture | L%, H%, LH% |
| Minor phrase | \| |
| Major phrase (turn) | \|\| |

### 1.2. Analysis-by-synthesis

An important part of our work has been the testing of the intonation model through analysis-by-synthesis. The tonal transcription has two functions: it constitutes a phonological representation of the analysed utterance and is also the input to a resynthesis module, where it is supplemented with phonetic rules that take care of the specific timing and appearance of F0 events, as well as concatenation between them. For the resynthesis we are using an implementation of the PSOLA synthesis algorithm in the ESPS/Waves+ environment ([1], [3], [4]).

The analysis-by-synthesis system presupposes that the speech waveform is divided into discrete units of different lengths. For each unit, certain parameters can be set in order to represent other intonational features and to explicitly state the physical identity of F0 events. In a speech synthesis perspective, the values of these parameters can be determined automatically for each unit using different methods, e.g., by applying models of topic structure and/or the Given/New distinction. Such models have been discussed in other papers (e.g. [5]).

## 2. ANALYSIS OF GLOBAL INTONATION

In our previous work, the model has mainly been applied to single utterances and lab-like speech. As our attention

gradually has turned toward spontaneous speech and dialogues, the model has also been divided into smaller portions. In a more elaborated version. When a speaker produces longer stretches of speech, as in a dialogue, the speaker may structure his/her utterances in a manner that may be related to e.g. the topic structure and the interactive character of the speaker situation. This will often be reflected in the pitch used by the speaker, and this is what we term global aspects of intonation.

## 2.1. Register and Range

In analysing spontaneous speech, we assume that any length of speech can be divided into smaller portions. In doing so, our model uses two units: major and minor phrases. One way of analysing global aspects of intonation is to represent them by means of two parameters that inherently belong to each phrase: Register and Range. In model terms, Register denotes the starting level of a phrase's intonation, roughly corresponding to the F0 level of the valley preceding the first accented syllable. Thereafter it is often gradually lowered at each accent as a result of downstepping. Another parameter represents the speaker's floor, or lowest possible level of F0, and limits the lowering of Register.

By Range we mean the height of a pitch gesture of an unfocused accent, starting from the Register level. This corresponds to Ladd's concept *tonal space*, which is defined as: "...a subset of the overall speaking range which is available for realising tonal distinctions at any given point in the utterance." ([6], p. 73). As the range of the pitch gesture usually is higher for a focal accent, we use two separate parameters for the two levels of prominence.

## 2.2. Analysis

These global features are analysed for each intonational phrase, alongside with the analysis of prominence and grouping. We currently perform the analysis of global features according to a model recognising three Register levels (High, Mid and Low) and four Range categories (High, Mid, Low and Flat). The register can also be Decreased, which indicates a more gradual (less than a full shift to a lower level) decrease of register. We thereby combine absolute and relative approaches to phrase intonation. The absolute levels (High, Mid and Low) are absolute compared to other phrases, but relative to the speaker. The Decreased level is relative to the preceding phrase.

The choice between an absolute or relative label is related to the grammatical and semantic content, as well as the topical coherence of the phrases. The stronger the relation between the phrases, the more likely the relative label is used. The question of how to measure the topical coherence is naturally an intricate matter, involving many aspects of semantic and grammatical relationships. At this stage we rely on a few loosely defined guidelines mixed with impressionistic analyses [5].

The result is a detailed and powerful notational device, that can cover the intonation over several phrases. Together with a transcription of accentuation and grouping, it can produce very accurate close-copy models of phrase intonation.

## 3. MEASUREMENTS OF THE GLOBAL CHARACTERISTICS OF F0

Another aspect of our work has been to investigate some acoustic features of the global F0 characteristics of conversations through measurements of the local Min, Mean and Max F0 of phrases in dialogues. For phrases in isolation, it is a well known and often occurring phenomenon of that F0 gradually goes down, even when there is no phonological indication of this. The perceptual impression over the course of a number of consecutive phrases is that a similar lowering occurs, especially if they are semantically coherent ([7], [8], [9]). This is also visible when inspecting F0 contours. Our expectations are thus to find a reflection of this lowering in the F0 measurement points within a speaker turn.

## 3.1. Method

In order to investigate this, we first analysed the dialogues according to our model in order to determine where the phrase boundaries were. Resynthesis was performed to validate the analysis. Then, a computer program calculated the F0 Min, Mean and Max for each phrase. The Means were calculated using all voiced portions of the utterance. Technical errors of the F0 analysis, such as octave jumps, for example due to creaky voices, could be removed by setting a local F0 range for each phrase, thus excluding the data points outside this range.

## 3.2. Results

The following tables show the summed-up differences in Min, Mean and Max F0 of a turn-final phrase and a turn-initial phrase in three different dialogues. Both the total for all turns within each dialogue, and the total divided by the number of speaker turns are shown. The latter shows the average decrease per dialogue turn of each measurement parameter. All figures are in Hz.

The first table represents a conversation among four participants.

| | Min | Mean | Max |
|---|---|---|---|
| Total | -386.0248 | -626.0764 | -772.957 |
| per turn (n=21) | -18.38213 | -29.81316 | -36.80747 |

The second table represents a conversation between two participants.

| | Min | Mean | Max |
|---|---|---|---|
| Total | -132.0834 | -44.1708 | -189.9978 |
| per turn (n=15) | -8.80556 | -2.94472 | -12.66652 |

The third and fourth tables represent two different speakers in the same conversation. Speaker 1:

| | Min | Mean | Max |
|---|---|---|---|
| Total | -172.1085 | -133.4626 | -28.7536 |
| per turn (n=20) | -8.605425 | -6.67313 | -1.43768 |

Speaker 2:

| | Min | Mean | Max |
|---|---|---|---|
| Total | -273.633 | -334.7105 | -319.907 |
| per turn (n=15) | -18.2422 | -22.31403 | -21.32713 |

## 3.3. Discussion

As can be seen in the tables, there are differences both between dialogues and between speakers, both regarding the extent of lowering and in the manner in which Min, Mean and Max interact with each other. However, a common trend is that the summed-up phrase differences are negative in all dialogues and for all measurement parameters. This is interpreted as a trend that speakers use a lower F0 turn-finally than turn-initially.

A possible explanation of the 'negative' trend within each turn is that the number of accents are fewer the later in the turn a phrase is. This would result in fewer potential F0 peaks. However, even though this has not been explicitly examined, our impression of the distribution of accents over the phrases in a turn does not support that explanation.

In view of these statistical indications, it seems necessary that F0 contours should be modelled so that corresponding measurements of the modelled F0 contours should approach these results, i. e. a lower Max, Mean and Min F0 in turn-final phrases than in turn-initial ones. Listeners presumably expect this, and it is plausible that a similar lowering in our modelling of F0 contours is beneficial for the listener.

## 4. INCORPORATING GLOBAL FEATURE IN THE MODEL

The mechanics of the F0 generation system has b[een] developed to take these suggestions into account. T[he] model can now be used to generate an F0 contour multiple phrases, corresponding to one major phrase t[hat] consists of one or more minor phrases. The initial mi[nor] phrase always gets a specific value, (dependent whether it has Low, Mid or High Register), whereas following phrases get Decreased register. A ma[jor] phrase could correspond to a speaker turn, but must necessarily do so.

### 4.1. Major and minor boundaries

As stated earlier, phrase boundaries are classified either minor or major, and different F0 characteristics each phrase boundary marker have been implemented minor boundary is marked by a fall to a level somewh[at] between the starting register and the speaker's floor. register reset is made for the next phrase, instea[d] continues on the same level. This results in a lo[w] initial F0 than the preceding phrase. A major phr[ase] boundary is marked by a fall to the preceding phrase. A major phr[ase] boundary, which is initial in the following m[ajor] phrase, starts at a register level specified for that phras[e].

Two phrases with a minor boundary between then thus modelled so that the parameter values of the sec[ond] phrase is lower compared to the first. When more t[wo] two phrases follow each other, the result is a grad[ual] lowering for all the phrases, since each phrase will s[...] at a register level equal to the preceding one's f[loor] level.

### 4.2. Downstepping

Each accent is modelled according to the principle downstepping, where the valley succeeding the accen[t] lowered compared to the valley preceding the acc[ent] Thereby the register decreases successively at e[ach] accent. This is done recursively, so that the post-acc[ent] drop becomes smaller for each accent (see also [6] 75).

### 4.3. Discussion

These extensions of the model capture the gl[obal] characteristics of spontaneous speech better than previous model. The modified model can capture c[...] where both register and range are decreased ove[r] number of minor phrases, with resetting occurring major phrase boundary. This pattern can be related to information structure of a phrase; there may, example, be a difference in intonational pattern betw[een] phrases with a content that refers backward to preceding phrase and a difference in a content that refers to the cas[e]
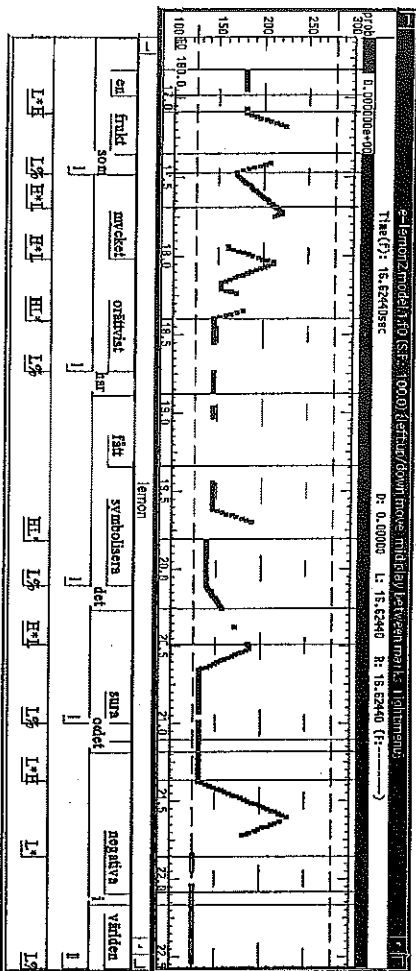
Figure 1. Modelled F0 contour of the utterance 'en frukt som mycket orättvist har fått symbolisera det sura och det negativa i världen' (trans.: '*a fruit that very unfairly has come to symbolise what is sour and negative in the world*'). The three tiers show the orthographic, the phrasal, and the tonal transcriptions.

## 5. EXAMPLE

An example of an F0 contour generated by the model is shown in Figure 1. The utterance is taken from the four-speaker conversation, and is spoken by a male speaker, who has a rather high voice. The parameters are set so as to copy his acoustic characteristics as closely as possible. Note the asymptotic decrease in the downstepping of the pitch accents.

## 6. REFERENCES

[1]   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.

[2]   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.

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

[4]   G. Möhler and G. Dogil, "Test environment for the two level model of Germanic prominence", *Proceedings EUROSPEECH 95*, vol. 2, 1019-1022, Madrid, 1995.

[5]   G. Bruce, M. Filipsson, J. Frid, B. Granström, K. Gustafson, M. Horne and D. House, "Text-to-intonation in spontaneous Swedish", to appear in *Proceedings EUROSPEECH 97*, Rhodos, 1997.

[6]   D. R. Ladd, "Intonational Phonology", University Press, Cambridge, 1996.

[7]   R. Van den Berg, C. Gussenhoven and A. C. M. Rietveld, "Downstep in Dutch: Implications for a model" In G. J. Docherty and D. R. Ladd, editors, *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 335-359, University Press, Cambridge, 1992.

[8]   G. Bruce, "Textual aspects of prosody in Swedish", *Phonetica* 39, 274-287, 1982.

[9]   N. Grønnum, "The Groundworks of Danish Intonation: An Introduction", Museum Tusculanum Press, University of Copenhagen, 1992.