A New Approach to Intonation Analysis and Synthesis of Swedish

Gunnar Fant[1,2], Anita Kruckenberg[1], Kjell Gustafson[1,3], Johan Liljencrants[1]

[1] Department of Speech, Music and Hearing, KTH, Sweden
[2] Department of Languages, University of Skövde, Sweden
[3] Babel-Infovox AB, Stockholm, Sweden
{gunnar; anita; kjellg; johan}@speech.kth.se, kjell.gustafson@infovox.se

Abstract

The main body of our study derives from the processing of 5 subjects' reading of a corpus from a Swedish novel. Two of the subjects were females. Intonation contours on a log frequency scale have been sampled and normalised to eliminate differences in mean tonal level and duration. As a result, intonation patterns across speakers are brought out revealing individual performances as well as group average data. A second part of our study has been to develop rules for predicting intonation contours and associated acoustic parameters from a superposition model. Syntactically determined prosodic sentence and phrase contours with associated juncture specifications are selected as a basic frame. Local word accent 1 and accent 2 modulations are added. These as well as phoneme durations are quantified with respect to a continuously scaled and lexically determined prominence parameter, Rs, and with respect to context and position within an utterance.

1. Introduction

Over the years a number of studies of various aspects of speech prosody have been carried out by Fant and Kruckenberg and collaborators, [4-8]. A major part has been concerned with duration and aspects of timing, such as rhythmical trends. Another substantial domain of studies has been the voice source in connected speech and its relation to aspects of speech production, [8]. More recently our efforts have been directed towards intonation analysis and modelling. Accordingly we are now in a position to tie together the bits and pieces of our knowledge base into a structured system for text-to-speech applications. We have recently been able to test the validity of our system in a text-to-speech environment, using Mbrola Tools [11].

The novelty of our approach can be stated in the following terms:

(1) Consistent use of a log-frequency display of F0 data in terms of the departure from a reference of 100 Hz, which attains the value of St=0 semitones. It is used in all illustrations in the present article. The conversion equations are

$$Hz = 2^{St/12}\ 100 \qquad (1)$$

$$St = 12[\ln (Hz/100)/\ln 2] \qquad (2)$$

In our experience the span of F0 modulations for males and females is closely the same on the semitone scale. Our multi-parameter speech processing system developed by Liljencrants provides a standard display of 2 mm per semitone in F0.

(2) It is possible to normalise the data for each speaker by reference to his or her particular mean F0 level.

(3) Furthermore, we have adopted a temporal normalisation. It involves the sampling of an intonation contour of a sentence by a fixed number of measures, one or two per syllable depending on accentual assignments.

(4) These normalisations in frequency level and time make possible a quantitative derivation of the essentials of an intonation contour and facilitates inter-speaker comparisons and specifications of group average data.

(5) Another novelty, as far as text-to-speech applications are concerned, is that our continuously scaled prominence parameter Rs may be adopted for the control of any acoustic parameter, in the first place F0 and duration.

As a consequence, our stress assignments are not limited to discrete phonological categories such as focal versus non-focal status.

(6) Prediction rules are derived from the entire corpus of read texts in a process that singles out base curves for major prosodic groups with associated junctures, separate from superimposed accentual modulations.

2. Analysis

Our main corpus for intonation analysis and modelling derives from the reading of 3 males and 2 females of a two-minute long passage from a novel. As a guide for F0 normalisation we plotted individual declination contours based on the average of unaccented syllables at five locations within an utterance. As shown in Fig. 1 they pertain to samples centered at relative positions of 0.1, 0.3, 0.5, 0.7 and 0.9.

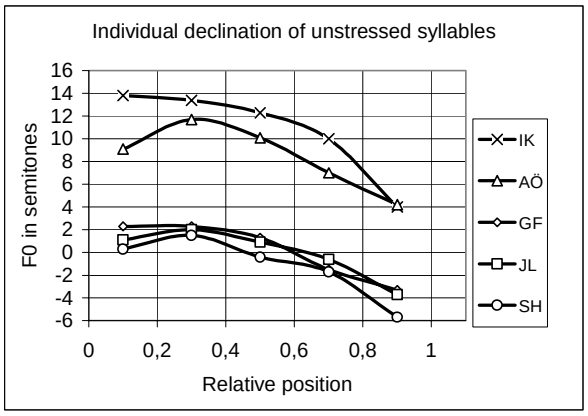

Figure 1: Individual declination contours of two females above and three males below. Frequency in semitones, St, versus 100 Hz.

Individual scale factors were also calculated from the average of all unaccented syllables. For the females we noted values of 9.5 and 7 St and for the males, -1, 0 and +1 St respectively. These values conform closely with what could be expected from Fig. 1. Furthermore, a calculation of a total mean, including accented syllables, gave almost the same individual scale factors. It may thus be concluded that unaccented syllables occupy F0 positions halfway between the highs and lows of accented syllables.

2.1 The Swedish word accents

The traditional notation for an accent 1 word such as "ánden" (the duck) indicates a rise in the accented syllable whereas the accent 2 word as in "ànden" (the spirit) shows a fall. In the canonical model of Bruce [1] an accent 1 domain should incorporate, if present, an unstressed syllable immediately before the accented syllable. An accent 2 domain contains a secondary stress in a region of the next or a following syllable. Thus with our notations

Accent 1 H L* Ha
Accent 2 H* L Hg

Unaccented syllables are denoted Lu. Ha and Hg have the same origin in production. In traditional vocabulary they are carriers of sentence or focal accent. In our continuously scaled system they show a distinct increase with the prominence parameter Rs.

The accent 2 low point L occupies a position at the floor of an intonation contour and thus below the Lu region. The accent 1 L* seldom reaches the same low level as the accent 2 L.

Fig. 2 provides a systematic view of the word accents at a low and at a high prominence level.

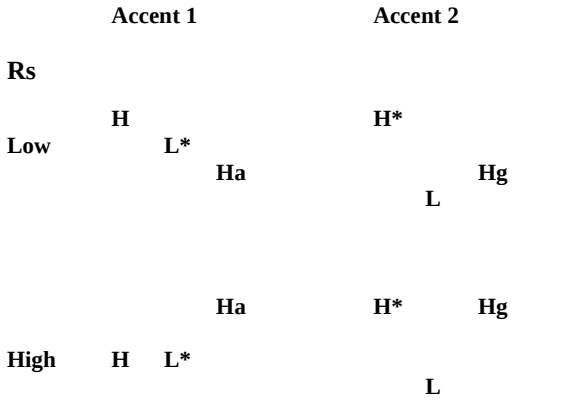

Figure 2: Schematic view of accent 1 and accent 2 constituents at a high and a low prominence level.

The seeming anomaly of the weak accent 1 Ha being lower than the preceding L* reflects our specific convention of sampling L* at the left and Ha at the right hand part of the vowel. We retain a correct relation at a high prominence level. The accent 1 H is of little importance only, and shows considerable individual variations. It can usually be substituted by the unaccented component, Lu.

The relative prominence of an accent 2 word is largely conveyed by its secondary accent peak, Hg. With increasing prominence the level of the primary peak H* saturates at Rs=22, whilst Hg continues to rise. In Swedish we often tie together a group of words into a compound attaining an accent 2 pattern. A word like "huvudpostkontoret" (the main post office) would thus attain our prosodic transcript H*LLu Lu Lu H*L Lu. It occurs in one of the sentences we have synthesised.

2.2 Individual variations

Fig. 3 contains illustrative samples of accentual patterns extracted from our corpus. Observe the relative height and leftward shift of the Hg peak of subject GF in the top part of Fig. 3 associated with high prominence, and the delayed and smoothed-out Hg domain of subject SH. The lower part of the figure illustrates the two accent 1 prototypes discussed in connection with Fig. 2.

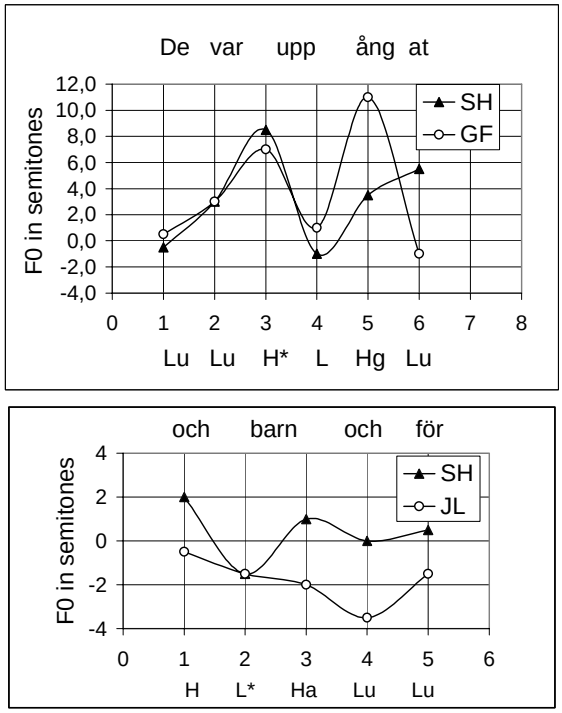

Figure 3: Examples of accent 2 and accent 1 realisation.

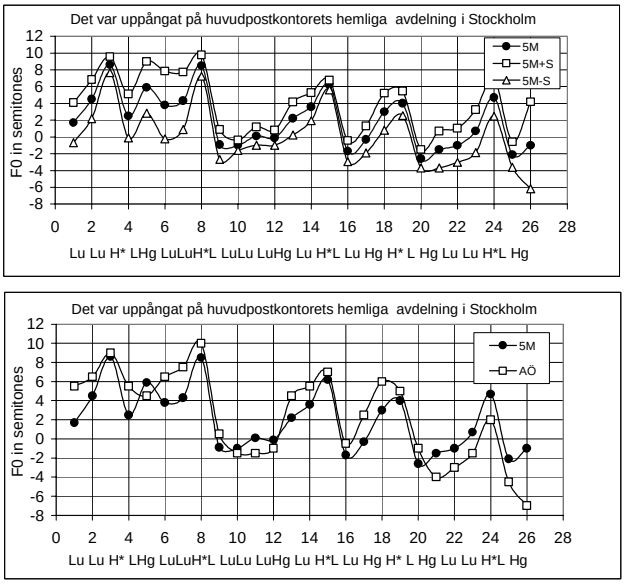

Figure 4: Above, mean of five subjects' sampled intonation contours and the mean plus and minus a standard deviation. Below, a comparison of female subject AÖ and the mean of the five subjects.

These two graphs illustrate the success of the normalisation, which brings out common features as well as regions of relative large and relative small individual variability. To the former belongs the word "uppångat", a close-up of which appeared in Fig. 3, and to the latter the H*L fall of accent 2 words. There is also an overall good agreement between the female subject AÖ and the five speakers' mean. Typical of AÖ is a somewhat higher sentence initial F0 as well as a lower sentence final F0 and a relative suppression of accent 2 secondary peaks, Hg.

3. Prediction of intonation contours

Our superposition model makes use of two different base contours, one for the initial clause of a sentence, which is also used for a complete sentence if no major juncture can be expected to occur. The other base contour is used for clauses or phrases within a sentence separated by pauses of the order of 200–400 ms. A possible example is illustrated in Fig. 5 comprising a full sentence allotted one primary and two secondary base contours.

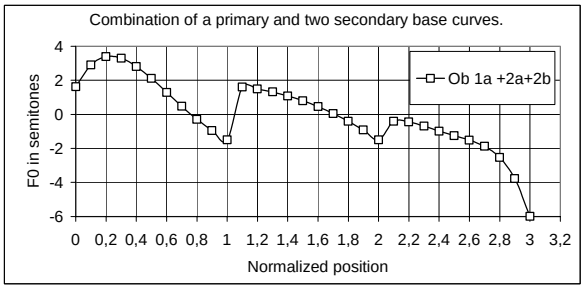

Figure 5: Example of successive base contours in a long sentence divided into three major prosodic groups.

We have found that the extent of the F0 reset at a juncture is linearly related to the duration of the pause. Thus, after a complete sentence where we would expect a pause close to 1000 ms the base contour attains a reset of 7,5 semitones.

Accentual modulations to be superimposed on a base contour have been derived as a function of prominence Rs and relative position within the contour. This involves a normalisation to a time scale from 0 to 1. Short major prosodic groups thus attain the same total F0 fall as longer groups, which is found to be in fair agreement with our data. Equations for deriving data points are somewhat different for primary and secondary prosodic groups.

Predictions are carried out in two steps. In the first round, as implied by our notational system, each syllable attains two data points if accented, and otherwise one point. In the second round we apply rules to ensure continuities and coherence within phrase and accent domains. In many cases this is accomplished by deleting an Lu point. One example is the carry-over of non-extreme Ha and Hg high points into a following unstressed syllable. Another is that an Lu immediately following an accent 2 L point inherits the L value. Special rules apply to pre- and post-focal positions.

4. Text-to-speech realisation

We have tested the intonation model in an Mbrola diphone synthesis environment. Duration rules were taken from an earlier database of one subject reading a much larger part of the novel text. These data matched data from our intonation corpus very well.

Special care was taken to predict pauses and other junctures with associated F0-reset routines from syntactic structures and possible semantic interpretations. Such relations are by no means simple. One and the same potential juncture may be realised in a variety of modes depending on the speaker, Fant, Nord and Kruckenberg [4].

Our approach has been to limit the choice to one out of four categories mainly defined by pause duration, 1000 ms after a complete sentence, 400 ms after a major clause and 200 ms or 75 ms at lower levels of the hierarchic structure.

It was found that duration and juncture rules were of considerable importance to the naturalness. One indication is the close relation between prominence Rs and duration data, supplementing the co-varying rules of F0 modulations as a function of Rs.

Rs was lexically determined from word class. The major categorisation is between content words and function words. Rules for syntactic and possible semantic modifications are included.

The results of our synthesis experiments have been quite promising. An indication is the close fit between the experimentally determined average intonation contours of the five speakers and what has been predicted by rules. This is illustrated in Fig. 6.

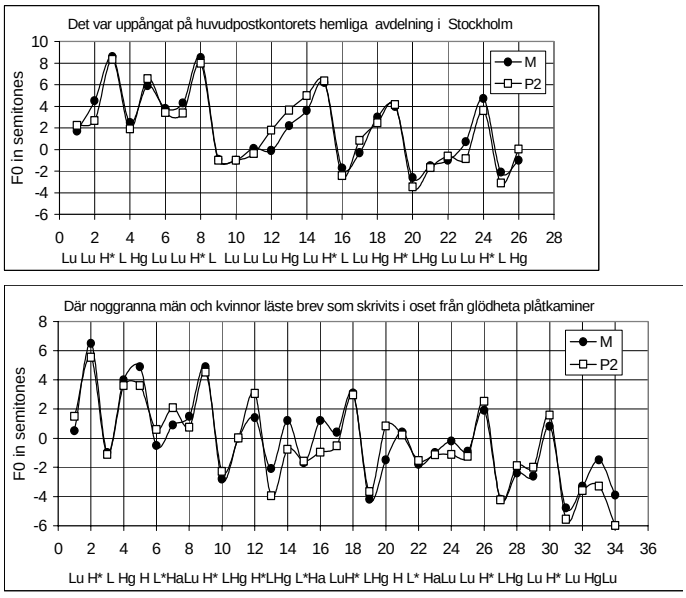

Figure 6: Predicted and measured F0, sentence 1.

5. General discussion

Our intonation model, now established after a substantial period of earlier studies [4-8], is more detailed than those of current systems for Swedish text-to-speech synthesis [2, 3]. As in the Fujisaki model with special reference to Swedish [9], we have adopted a superposition model based on a log-frequency scaling of F0, but we employ a more direct quantitative coding of accentual modulations.

Our model should find use not only in synthesis but also in descriptive analysis as a supplement to the traditional modelling of Gårding [10]. The F0 normalisation and sampling procedure allows a data reduction retaining essentials of intonation contours.

6. References

[1] Bruce, G., 1977. *Swedish Word Accents in Sentence Perspective*. Lund, Gleerup.

[2] Bruce, G.; Filipsson, M.; Frid, J.; Granström, B.; Gustafson, K.; Horne, M.; House, D., 2000. Modelling of Swedish Text and Discourse Intonation in a Speech Synthesis Framework. In Antonis Botinis (ed.), *Intonation. Analysis Modelling and Technology*. Kluwer Academic Publishers, 291-320.

[3] Carlson, R.; Granström, B., 1973. Word accent, emphatic stress, and syntax in a synthesis-by-rule scheme for Swedish. *STL-QPSR*, 2-3/1973, 31-35.

[4] Fant, G.; Nord, L.; Kruckenberg, A., 1986. Individual Variations in Text Reading. A Data-Bank Pilot Study. *STL-QPSR*, 4/1986, 1-17.

[5] Fant, G.; Kruckenberg, A., 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, 2/1989, 1-83.

[6] Fant, G.; Kruckenberg, A.,1994. Notes on Stress and Word Accent in Swedish, *STL-QPSR*, 2-3/1944, 125-144. Also published in *Proc. Int. Symp. on Prosody, 18 Sept 1994, Yokohama*, 19-36.

[7] Fant, G.; Kruckenberg, A.; Liljencrants, J., 2000. Acoustic-phonetic Analysis of Prominence in Swedish. In Antonis Botinis (ed.), *Intonation. Analysis, Modelling and Technology*. Kluwer Academic Publishers, 55-86.

[8] Fant, G.; Kruckenberg, A.; Liljencrants, J.; Hertegård, S., 2000. Acoustic phonetic studies of prominence in Swedish. *TMH-QPSR*, 2/3 2000, 1-52.

[9] Fujisaki, H.; Ljungqvist, M.; Murata, H., 1993. Analysis and modelling of word accent and sentence intonation in Swedish. *Proc. 1993 Intern. Conf. Acoust. Speech and Signal Processing*, vol. 2, 211-214.

[10] Gårding, E., 1989. Intonation in Swedish. *Working papers.* Lund University Linguistics Department 35, 63-88. Also in (eds.) Daniel Hirst and Alberto Di Cristo. *Intonation Systems.* Cambridge University Press 1998, 112-130.

[11] Dutoit, T.; Pagel, V.; Pierret, N.; Bataille, F.; van der Vrecken, O., 1996. The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proc. ICSLP 96*, Philadelphia, vol. 3, pp. 1393-1396.