Intonation Analysis and Synthesis with Reference to Swedish

Gunnar Fant and Anita Kruckenberg

Department of Speech, Music and Hearing, KTH, Sweden
gunnar@speech.kth.se

Abstract

The present report reviews findings about F0 patterns and in specific the realisation of the Swedish accent 1 and accent 2 tone patterns. We have developed a novel system for normalizing F0 contours which allows the handling of male and female data in a common frame. It also facilitates the sorting out of individual patterns from a norm. For this purpose we have defined a semitone scale with a fixed reference. As in the Fujisaki model [10], we have employed a superposition scheme of adding local F0 modulations to prosodic phrase contours, but with different shaping algorithms. The influence of the syntactic frame, and of word prominence and its relation to the single peak of accent 1 and the dual peak of accent 2 has been quantified. Some language universal traits, such as time constants and typical shapes of local F0 patterns, are discussed. The perceptual smoothing of local F0 contours has been illustrated in a simple experiment which relates to the concept of an auditory time constant. Our Swedish prosody modules have ensured a high quality in synthesis and a robustness in performance with respect to uncertainties in text parsing. Modifications for English and French prosody have provided promising results.

1. Introduction

Over an extended period of time we have been engaged in studies of the prosodic realisation of prose reading. A broad range of topics has been covered [5-9]. We have studied temporal structures, junctures and pausing, prominence scaling and co-variation of acoustic parameters related to prominence. These include voice source characteristics and a survey of the foundation of prosody in speech production. In recent years, a major part of our work has been devoted to intonation modelling, and attempts to tie together bits and pieces of acquired knowledge in text-to-speech synthesis.

As a result we now have a platform for quite efficient realisation of prosody in text-to-speech systems. At present we are engaged in the difficult and partly illusive problem of predicting prosodic grouping from syntactical text analyses. As a background we have a database of several subjects' reading. We are intrigued by the existence of some quite stable aspects of intonation contours and accent modulation, as well as by the large individual variability of prosodic grouping and pausing. A conclusion is that deviations from an ideal syntactical parsing norm may not be crucial as long as they reflect alternative speaking styles within an acceptable frame.

Our intonation modelling is in part based on the established principle of superposing local accent modulations on underlying prosodic base curves. Compared to earlier systems for Swedish [2-3, 10, 12] it presents several novelties, such as normalization of F0 data in frequency and time, which has made it possible to derive representative averages from a group of mixed male and female speakers and to structure individual variations.

The prosody model incorporates duration and the covariation of F0 and duration as a function of accent type, predicted prominence and location within a sentence. Our rules have been tested in an Mbrola frame. Intensity coding has not yet been incorporated, but we have access to relevant data.

2. The Swedish tone accents

In Swedish we have two distinct tone accents, also referred to as word accents, 1 and 2, [1]. Dialectal variations exist. Here we are concerned with the standard norm typical of the region around Stockholm. There exists a rather limited number of word pairs in which the tonal pattern distinguishes meaning. A classical example with traditional accent notations is “ánden” (the duck) indicating a rise in the accented syllable, whereas in the accent 2 word “ànden” (the spirit) there is a fall.

However, the main importance of the accent distinction is to preserve a correct pronunciation. In connected speech more than half of the content words carry accent 2, which dominates in di-and polysyllabic words.

With our notations, essentially derived from the canonical work of Bruce [1], the accents of disyllabic words carry modulation contours

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

Unaccented syllables are labelled Lu. L* and Ha define two sample points in the voiced part of an accent 1 primary syllable. H pertains to the preceding

unstressed syllable, which may be absent in sentence initial position. It is of secondary importance only. When present, it acts as a possible reference point for connecting to the following low point L*. All monosyllabic words carry accent 1.

In accent 2 the sample points H* and L in the primary syllable are followed by a high point Hg in the next or a following syllable. In compound words Hg is located in the final constituent. Between L and Hg there might then occur one or more Lu syllables.

Increasing prominence of accent 2 words is only in part related to the size of the H*L fall, which saturates at a moderate stress level, at which Hg takes over as the major stress correlate. The major role of the H*L fall is to signal the identity of accent 2.

From a phonological point of view it has been claimed that Swedish has only one marked accent, i.e. accent 2. The Hg peak of accent 2 has the same function as Ha of accent 1, i.e. to signal prominence. The prominence peak is thus the main constituent of accent 1. The specific and marked feature of accent 2 is the H*L fall.

Our data analysis supports this view, but we have found reasons to retain the basic model of Bruce [1]. An inconsistency in our use of his symbols is, that in case of a low prominence where L* and Ha are ill defined, we assign them to segmental positions. As a result Ha may be lower than L*.

3. Frequency and time normalization

Our main corpus for intonation analysis and modelling derives from 3 males and 2 females reading a two-minute long passage of a novel. In order to derive representative average intonation contours we have employed a system of frequency and time normalization.

A basic requirement is the semitone scale. We have introduced a fixed semitone scale with the unit St defined by

$$St = 12[\ln(Hz/100)/\ln 2] \qquad (1)$$

which attains the reference value of St=0 semitones at 100 Hz, St=12 at 200 Hz and St=-12 at 50 Hz.

The conversion from semitones to frequency is accordingly

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

The semitone scale preserves the main shape of male and female intonation contours. The first stage of the normalization is to subtract a subject's average F0 in St units from his or her St contour, which provides a common base for males and females. In our study we found long-time-average St values of +9,5 and +7 for the two females and -1, 0 and+1 respectively for the three males. Translated into Hz these correspond to 173 and 150 for the females and 94, 100 and 106 Hz respectively for the males.

The next step in the normalization pertains to the time scale. Individual differences in utterance length are removed. This is accomplished by a sampling of F0 data limited to one or two measures per syllable. Accented syllables receive two measures, L* and Ha for accent 1 and H* and L for accent 2. All other syllables receive one measure only, which applies to the secondary peak Hg of accent 2, the initial H of accent 1 and all unstressed syllables, denoted Lu. These discrete data points are now connected by smoothed lines, which constitutes the normalized intonation contour

We now have an efficient tool for comparing individual speakers within the same frequency and time frame. A general observation is the small inter-subject spread in accent modulation depth, in specific of the H*L fall in the primary syllable of an accent 2 word, which is of the order of one semitone only. In addition, normalized H* values constitute rather stable anchor points for the upper bound of an intonation contour, with an inter-subject standard deviation of somewhat less than two semitones. This is illustrated in Figure 1.

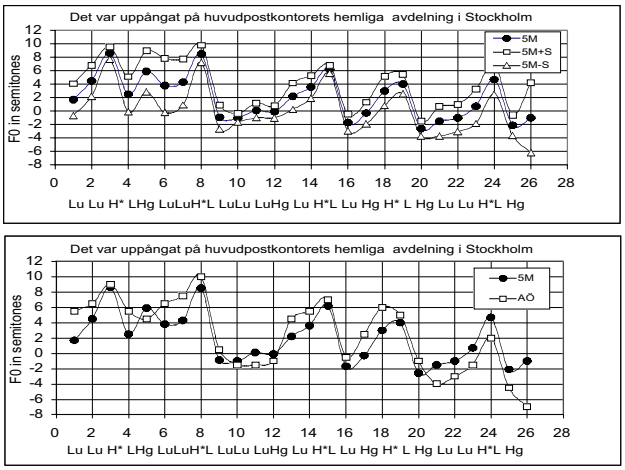

Figure 1. *Above, mean of five subjects' sampled intonation contours and the mean plus and minus a standard deviation. Below, a comparison of our reference female subject AÖ and the mean of the five subject*s.

The relatively large spread in the early part of the sentence, to be seen in the upper part of the figure, is explained by some individuals inserting a juncture before a long preposition phrase.

The lower part contains the mean curve of the five subjects and our reference female speaker, AÖ. Except for a higher initial value and a lower final value, she has a higher starting point and a lower final value. A general conclusion is, that except for individual global gestures, the main trends within a sentence are the same for males and females. Our normalization procedure has been quite successful. Observe the similarities in overall

declination rates, as well as in absolute levels of normalized F0. No adjustments were made for this particular sentence.

The close-up examples of Figure 2 illustrate samples of individual variations. The two speakers in the top graph differ only in the size of the accent 2 prominence peak Hg. In the lower graph one of the speakers shows an accent 1 H L*Ha pattern with a high Ha typical of normal prominence while the other subject shows a declining HL*Ha sequence, indicating low prominence.

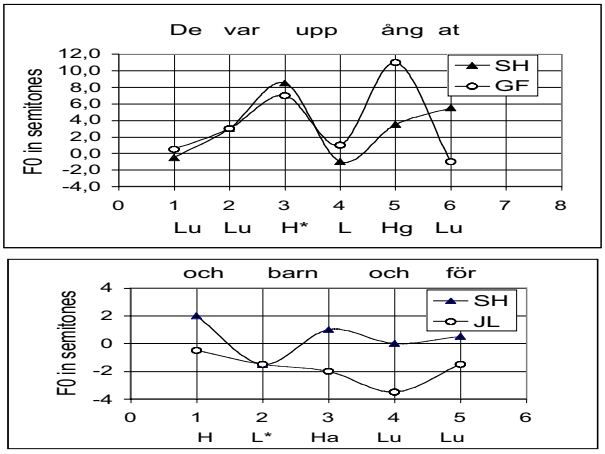

Figure 2. *Individual variations of accent 2 above and accent 1 below.*

4. Synthesis

Accent 1 and accent 2 modulations are superimposed on smooth base curves, each tailored to a specific onset, main level and declination within a prosodic group. As exemplified in Figure 3, a complete sentence may contain a single or a number of such modules.

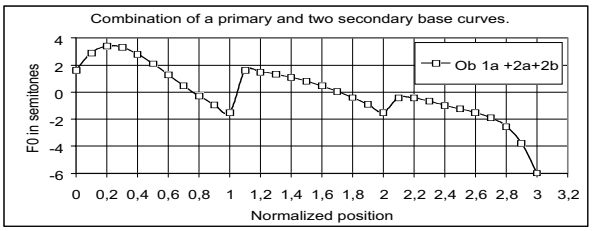

Figure 3. *Example of successive base contours in a long sentence divided into three major prosodic groups.*

The F0 reset at boundaries is related to pause length. Accent modulations as well as base curves are derived from a statistical analysis of our five subjects, three males and two females. This has involved a non linear regression analysis of the variation of F0 values with respect to predicted prominence, quantified by our continuous scale factor RS, and the position within a prosodic group and a sentence. These relations are expressed with up to fourth power polynomes. This should provide a closer tie to actual speech data than the Fujisaki method with second order filter step responses.

Figure 4 shows an example of a normalized F0 contour of the average of our five speakers' spoken data, and the corresponding contour predicted from our general rules. There is a close agreement. The average departure is of the order of 1,5 semitones only, which covers accent modulation as well as more global features related to intonation modules.

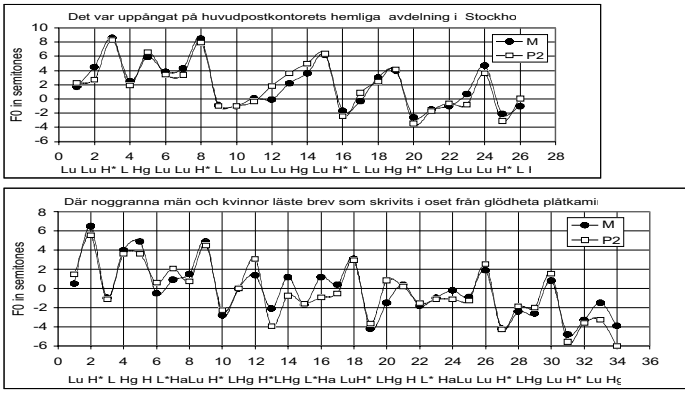

Figure 4. *Predicted F0, P, and measured F0, M.*

Our synthesis by rule for Swedish is judged to have a prosody superior to other systems demonstrated up till now. The Mbrola frame borrowed from Babel-Infovox is of a high quality. The segmental concatenation accounts for some unavoidable audible degradations, but these are at a low level.

5. A note on auditory integration of pitch

In our Mbrola synthesis, intonation contours and accent modulations are straight line approximations of the true continuous F0 curves of real speech. It is remarkable that the percept is all the same quite convincing. We have made a test comparing two realisations of the Swedish accent 2 word [ànna], in which the second syllable, the vowel [a], attains a typical F0 peak. One is programmed from our rules with straight line F0 components, the other with cosine approximations of each line, see Figure 5, which also shows the normalized representation.

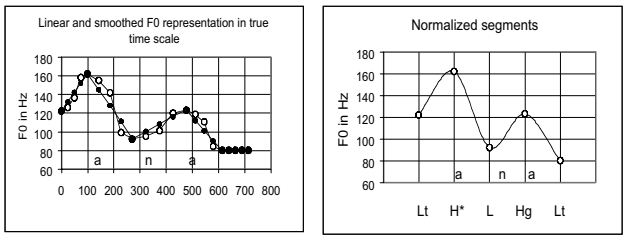

Figure 5. *Straight line and cosine approximation of F0 in the accent 2 word "Anna". To the right the normalized intonation contour.*

No major differences could be detected between the two versions, but careful listening revealed that the linear approximation had a pitch of about one half semitone below that of the more gradually shaped contour.

The tonal percept thus appears to follow rules of auditory integration, which could be modelled in analogy to linear or maybe nonlinear systems. A guess would be a time constant of the order of 50-100 ms.

6. Multi-language applications

We have recently attempted to transfer experience from our Swedish rule system to synthesis of English and French. Our preliminary results are quite promising, especially with respect to French prosody, which is illustrated in Figure 18.

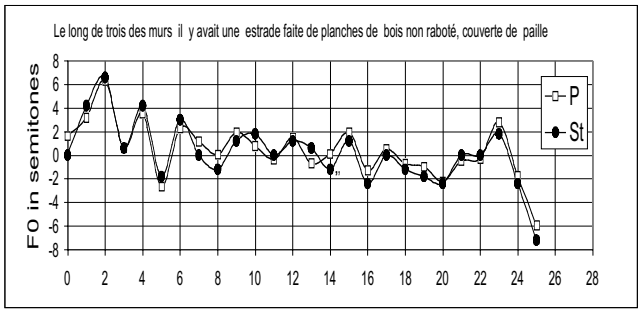

Figure 6. *Measured, St, and predicted F0, P. The text is:"Le long de trois des murs il y avait une estrade faite de planches de bois non raboté couverte de paille."*

This is a normalized graph of a spoken sentence and a prediction from tentative rules. The close match is to some extent influenced by the analysis-by-synthesis performed on the training material, but our tentative rules have functioned well also in other sentences.

For French we have introduced a modified version of our Swedish accent 1, which accounts for the typical iambic pattern of word intonation within a prosodic group. The final rise, typical of sentence internal prosodic groups, can generally be introduced without a specific intonation module by a high RS value in the last content word. Sentence final groups have the same or larger declination towards a low F0 than in Swedish, and the pre-pause lengthening is more apparent.

7. Conclusions

The FK text-to-speech prosody rules for Swedish have functioned remarkably well. We have, in a relatively short time, performed tentative transfers to French and English, but these have to be followed by a more detailed analysis. Our modular tools appear to have a language universal significance, and can be adjusted for language specific needs.

The system of frequency and time normalization and the introduction of a continuously scaled prominence factor RS determining F0 as well as duration is unique. It could have applications also in comparative studies, to sort out individual intonation patterns from a norm.

Our data are in fair agreement with the language universal rules suggested in [5]. One example is the dimension of "hat patterns", i.e. of the duration and peak height of a single prominent F0 peaks modelled as a triangle. Although this works fairly well in synthesis there is much evidence that single intonation peaks are bell shaped with smooth onsets and offsets. In our Swedish data, [8], we find symmetrical structures of 200-300 ms base length and a height of 5-10 semitones depending on the degree of prominence. The time constant of rise and decay is thus of the order of 100-150 ms. It appears to reflect language independent physiological constraints. Rise and decay may have different time constants associated with different muscular engagements, [11].

8. 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] Collier, R., 1991 Multi-language intonation synthesis. *Journal of Phonetics* (1991) 19,61-73.

[5] Fant G., 1997. The voice source in connected speech. *Speech Communication* 22: 125-139.

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

[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] Fant, G., Kruckenberg, A.; Gustafson K.; Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish, *Speech Prosody 2002, Aix en Provence*. Also in *Fonetik 2002, TMH-QPSR* 2002.

[10] 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.

[11] Fujisaki, H., Tomana, R., Narusawa, S.; Ohno, S.; Wang, C., 2000. Physiological mechanisms for fundamental frequency control in standard Chinese. ICSLP 2000, SS(01)-3, 1-4.

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