Prosody by Rule in Swedish with Language Universal Implications

Gunnar Fant and Anita Kruckenberg

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

Abstract

The FK text-to-speech prosody rules for Swedish are outlined. They cover all levels including prosodic grouping from syntactical analysis. It is a superposition system with local accentuations superimposed on modular F0 patterns of specific rise and decay patterns in successive prosodic groups. F0 in semitones and segmental durations are calculated as a function of lexically determined prominence and position. Speaker normalisation in frequency and time allow the pooling of male and female data in the analysis stage. The main architecture has been successfully tested in French and English synthesis.

1. Introduction

This is a follow-up of our report on intonation analysis and synthesis of Swedish, [5] presented at the speech prosody meeting in Aix-en-Provence 2002. We have developed a synthesis by rule system in an Mbrola diphone environment. Our ambition has been to contribute to the overall state of the art of producing natural prosody based on our previous work during a 15 year period, see [2-6].

As in the Fujisaki model [7], F0 is processed on a log scale and accent modulations are superimposed on sentence contours divided into modules. Specific to our system is the prediction of these components from lexical and syntactical requirements. The overall prosodic grouping within a complete sentence is structured in terms of a number of successive intonation modules, usually 1 to 4 within a sentence. These represent major prosodic constituents with specific main F0 level, onset and decay contours. They are associated with certain F0 reset characteristics at boundaries and are coordinated with patterns of pausing and final lengthening.

Our inventory of intonation modules and superimposed accent modulations have been modelled from a database of five subjects' reading of a two minute long passage from a novel. This has involved non-linear regression analysis of F0 parameters specific for the Swedish accent 1 and 2 with respect to our continuously scaled prominence parameter RS and the position of the accent within a sentence or a major prosodic group.

The same general procedure also applies to segmental data. Duration of speech sounds are calculated by rules according to assigned prominence, phoneme class and position within a syllable and a prosodic group.

Additional data on associated intensity and voice source properties within a sentence are available but have not yet been applied to synthesis. They are of secondary importance only. F0 and duration are primary.

A study of individual variations in prosodic organisation of a sentence and of pausing [6] has revealed a considerable spread. A general experience from tests with systematic variations in synthesis has revealed a considerable tolerance for departures from our tentative norms, providing variations are held within basic constraints. These findings also suggest means of categorising prosodic variability as a component in individual performance and in requirements for different text materials and reading situations.

The success we have had in developing Swedish prosody rules raised the question whether basic parts of the architecture might be of relevance to other languages, and if so, what modifications would be needed. Preliminary attempts of application to English and French have been made. These capitalize from a detailed acoustic phonetic study of our standard Swedish prose text translated into English and French [3]. It was largely concerned with temporal structures and has provided us with a databank of segmental durations and also with some F0 data.

Results from synthesis have been quite promising, especially for French text-to-speech. As illustrated at the recent ICPhS in Barcelona, an interesting outcome is the possibility of a transfer of language rules, e.g. synthesizing an English text read with a French sound inventory and French prosody, and conversely, a French text read with an English sound inventory and prosody. Advanced text-to-speech rule systems may accordingly find use as a tool in second language teaching.

2. Normalized intonation contours

Initial print-out of F0 data from a sound recording as well as data processing is made on a log frequency scale. We have introduced a 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 St value in a read passage from his or her St contour which adds to a common base for males and females. In our study we found speaker long time average St values of +9,5 and +7 for the two females and -1, 0 and+1 respectively for the three males.

The next step in the normalisation 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 and all other syllables one.

3. The Swedish tonal accents

In Swedish we have two contrasting tonal accents, accent 1 and accent 2. With notations essentially derived from the work of Bruce [1] they define modulation contours

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

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

In accent 2 the sample points H* and L in the primary syllable are followed by a high point Hg, which is not strictly syllable bound. In compound words it is located in the final constituent.

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.

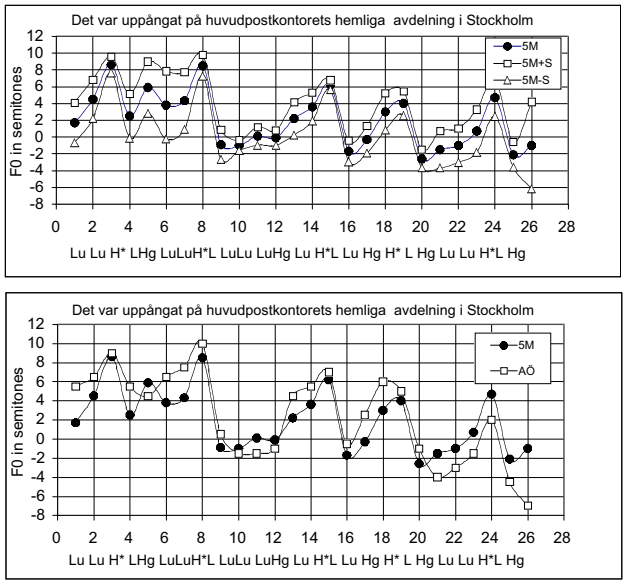

Figure1: *Above, mean of five subjects' sampled intonation contours and the mean plus and minus a standard deviation. Below, a comparison of our reference female subject AÖ and the mean of the five subjects.*

We now end up with a sequence of discrete F0 points. In order to restore continuity in the final presentation of an F0 contour we employ an automatic smoothing of data in the Excel program. This is illustrated in Figure 1

This type of display is 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, normalised 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.

Furthermore, except for individual global gestures, the main trends of declination within a sentence are the same for males and females. Our normalisation procedure has been quite successful.

4. Prosodic grouping

A complete sentence of moderate size may be produced as a single prosodic group with certain rise and decay characteristics in F0, or as a succession of groups [5] each carrying an intonation module, i.e. a base curve upon which accent modulations are superimposed. At present we employ four different modules depending on position. In sentence final positions before full stops they receive an extra F0 lowering.

In boundary regions, i.e. at junctures, there is always a final lengthening with or without a proper pause. Pause durations tend to be proportional to the F0 reset. Typical of the juncture at a sentence boundary defined by a full stop in the text is an F0 reset of 7 semitones and a pause close to 1000 ms. At the juncture between two main clauses the F0 reset is of the order of 3,5 semitones combined with a pause of about 400 ms [5]. At lower syntactical levels there exist a large number of possible combinations of syntactical constituents where smaller values of F0 resets and pause duration may be expected. However, in our experience, individual variations in reading are excessively large. We have also noted a systematic difference between our prose reading and news reading over the radio in which sentence pauses were of the order of 500 ms [6].

A syntactic parser exploited to its full capacity to define junctures may generate unnatural breaks, which should be avoided. Work along these lines has led us to predict junctures and pauses on a probability basis and with respect to the size of constituents. Thus, in our collected data, the boundary before a subordinate clause was only to 30 % realized by a pause and to 70 % by terminal lengthening only. Similar values were observed for noun phrases. Before and after a preposition phrase 92% of all junctures were realized by final lengthening only. On the other hand, we encountered relatively large pauses between major constituents of a complex noun phrase.

Our five subjects had quite similar pause durations, close to 1000 ms at a full stop, but showed a large spread in the number and total time allotted to pausing within a complete sentence [6].

5. Word and syllable prominence

A unique feature of our system is that both F0 and duration are controlled by a continuously scaled prominence parameter labelled RS.

It was originally derived from listening tests in which subjects were asked to assess syllables and words of our standard prose texts in a scale from 0 to 30. As a guide average values around RS=10 for unstressed syllables and RS=20 for stressed syllables was suggested. We found that words received about the same RS as the dominating syllable in the word.

Content words averaged RS=19. Numerals and adjectives topped the scores with RS=21 followed by nouns RS=20, verbs and adverbs RS=17.

In the class of function words the scores showed rather small variations around a mean value of RS=11 ranging from 12,5 for pronouns to 10,5 for auxiliary verbs.

Highly reduced syllables, usually articles, were graded in the range RS=3-8. However, function words may be raised in prominence and content words reduced.

Accentuation, i.e. the presence of a significant F0 modulation, is limited to RS>14.

Focal accentuation is usually confined to RS levels above 23 [4]. Special rules have been adopted for pre- and postfocal reduction [4].

6. Duration

Our modelling of temporal structures is based on a separate databank of measured durations of phonemes and syllables structured by sequential constraints and prominence levels. Specific values are derived from the RS parameter by linear regression within a frame of reference values for stressed and unstressed positions. All phonemes within a syllable share the same RS. The definition of syllable boundaries is thus of some importance.

The reverse process, i.e. predicting RS values of a syllable from the duration of its constituent phonemes in text reading, provides a means of estimating the perceptual salience of duration as a cue to perceived prominence. We have found a high correlation between a sequence of RS estimates from listening and a sequence of RS estimates from durations alone.

7. Synthesis

Our prosody rules have been tested in an Mbrola diphone environment with access to the Infovox database for conversion from text to phonetic form with word class and accent tagging. In addition, we use their database of one or several reference speakers for phoneme to sound conversion on a diphone basis. These have been spoken with monotone pitch, which is a requirement for insertion of specific intonation patterns.

As outlined in the previous sections the programming employs the following steps.

1. A prosodic grouping in terms of a sequence of intonation modules with associated pauses and related boundary conditions is carried out.

2. Each syllable is assigned an RS value.

3. Accented syllables are given two F0 data points and other syllables one point.

4. Syllable positions within a prosodic group are converted to relative positions in a scale from 0 to 1. This normalization is motivated by the tendency of the total F0 decay within a prosodic group to be independent of its length.

5. F0 of unstressed syllables are placed on the contours of intonation modules. Accented syllables attain F0 modulations superimposed on an intonation module.

6 These are derived from RS values and relative position within an intonation module.

7 Duration of sound segments are calculated.

8. Special rules apply to positioning of F0 points within the temporal frame.

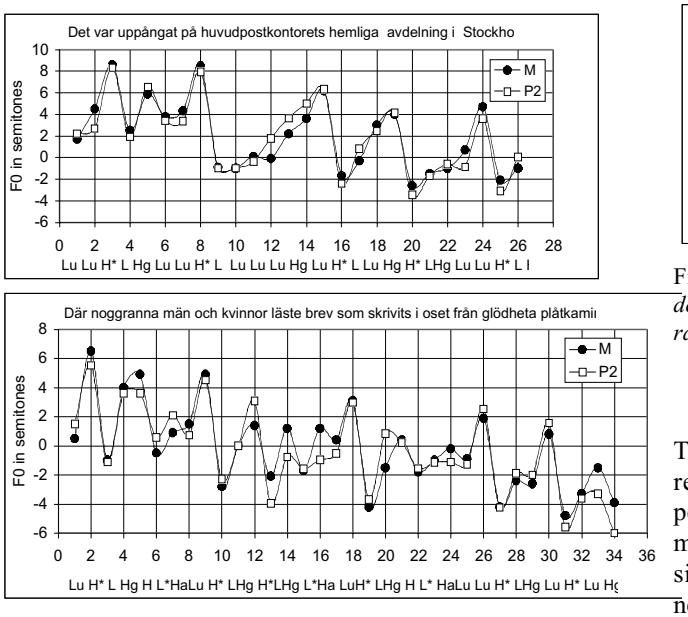

Figure 2:*Predicted F0, P2 and measured F0, M.*

8 Results

Figure 2 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.

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 3.

It pertains to a spoken sentence and a prediction from tentative rules. The very close tie 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.

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 in French. 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 declination towards a low F0 as in Swedish.

In British English a substantial fall of the F0 contour is frequently found also in non-final prosodic groups. Special rules apply to compound stress.

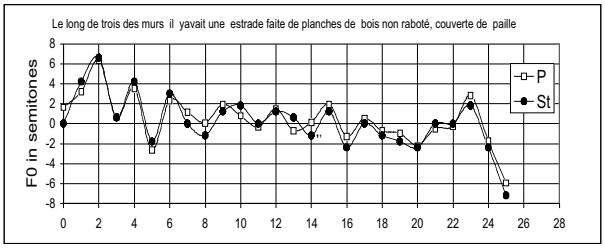

Figure 3: *Measured, St, and predicted F0, P."Le long de trois des murs il y avait une estrade faite de planches de bois non raboté couverte de paille".*

9. Conclusions

The FK text-to-speech prosody rules have functioned remarkably well. We have in a relatively short time performed tentative transfers to French and English. The modular tools appear to have a language universal significance and can be adjusted for language specific needs.

At the oral session we intend to demonstrate synthesis in the three languages and what happens in a switching of language codes, simulating a Frenchman speaking English without a proper insight in the sound inventory and prosody, and also the converse, of an Englishman speaking French.

Systematic manipulation of synthesis rules of this type could have applications in second language teaching.

10. References

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

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

[3] Fant, G.; Kruckenberg, A.; Nord, L., 1991. Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* 19, 1991, 351-365.

[4] 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.

[5] 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, 283-286*. Also in Fonetik 2002, *TMH-QPSR,* 44-2002.

[6] Fant, G.; Kruckenberg, A.;Barbosa Ferreira, J., 2003. Individual variations in pausing. A study of read speech. *Phonum 9, Umeå University*, 193-196.

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