Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Individual and contextual variations of prosodic parameters

Fant, G. and Kruckenberg, A.

journal: TMH-QPSR
volume: 48
number: 1
year: 2006
pages: 005-009

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

Individual and contextual variations of prosodic parameters

Gunnar Fant & Anita Kruckenberg

Abstract

This is a summary of variabilities and co-variation of prosodic parameters found in our studies of text reading and in the development of text-to-speech synthesis. In addition to F0, duration and intensity, the survey includes aspects of voice production and perception. The role of sub-glottal pressure is discussed. Speech parameters have been correlated with our continuously graded prominence parameter RS. Individual variations in pausing and the realisation of prosodic boundaries have been studied.

1. Introduction

During the last 20 years we have contributed to the analysis and synthesis of the prosodic structure of read speech, mainly directed to the Swedish language [1-11]. Studies of rhythmical structures and comparisons of French, English and Swedish have been undertaken and also studies of poetry reading.

At an early stage, our major emphasis was on duration, but we gradually included rather detailed studies of intonation patterns which opened the way to a project on Swedish text-to-speech synthesis, the FK system, now realized within an Mbrola frame.

2. The FK model

Our ambition has been to develop a model of prosodic realisation, covering major stages of transformation from linguistic structure to acoustic parameters. In the first place, we have been concerned with F0 and duration measures, which are employed in Mbrola synthesis, but we have also attempted to relate the modelling to sub-glottal pressure and to intensity and voice source parameters.

A main impression from our synthesis is that an appropriate control of F0 and duration alone secures a high degree of naturalness in prosodic realisation. Our ambition is to contribute to a wider view of the co-variation of all parameters involved**.**

Our modelling of F0 contours is based on normalized data from 3 male and 2 female subjects' reading of a text of 2 minutes' length, which has provided insights in both typical intonation patterns and individual variations.

Special attention has been devoted to prosodic boundary cues in terms of F0 patterns, pauses and pre-pause lengthening. We find large individual variations of prosodic grouping, i.e. where boundaries are inserted and their realisation. Within limits, there is accordingly a considerable tolerance of deviations from an established set of parsing rules in speech synthesis.

A most important decision has been to base all F0 measures on a semitone scale instead of Hz [3-11], which allows a normalization of male and female data of varying pitch levels to be transposed to a common reference. Moreover, we have adopted an absolute semitone scale, labelled St, which has a zero value at 100 Hz.

We have introduced a continuously scaled parameter RS for syllable prominence, ranging from 0 to 30. It was derived from rather extensive perception tests [2]. This particular scaling has been adopted in [15]. Unstressed syllables and function words attain RS values averaging 11, and stressed syllables attain values of the order of 20 or more. Word prominence was found to equal the prominence of the main stressed syllable of the word [2].

Lexical default values have been determined. We found averages of RS = 21 for nouns, 22 for adjectives, 19 for verbs, 16 for adverbs, and 12 for auxiliary verbs. Extreme values are 30 for emphatically raised prominence and RS = 0 for complete deletion. Focal accentuation is found in the range of Rs = 23-27. Pre- and post-focal reductions are of the order of minus 3 RS units [4-5].

The continuity of RS parameter variations ensures a phonetically more realistic approach

than the traditional phonological concepts of focus and sentence accent. More than one word may be emphasized in a sentence.

In our scheme, F0 and duration co-vary with RS and position within a prosodic group. F0 attains an accent modulation, specific for each of the Swedish tonal accents 1 and 2. These are superimposed on one or more prosodic base curves within a sentence.

Our rules for modelling of duration derive from a 10-minute-long reading of a single trained speaker.

Several of our earlier publications [2-10] have dealt with aspects of the co-variation of F0, duration, intensity and sub-glottal pressure with prominence and also of individual variations. Here follows a systematic overview with data on prosodic grouping and pauses included.

2. Intonation modelling

A normalization with respect to a subject's average pitch with F0 expressed in semitones has proved to be most effective. But speakers differ in reading speed and short time temporal variations. Accordingly, as an act of normalization, we have substituted time by the relative positions of data points within a sentence or a prosodic group.

Following Bruce [12] and Bruce et al.[13], with some modifications, we have adopted a notation system of H, L*, Ha for accent 1 and H*, L, Hg for accent 2. Unaccented syllables are denoted Lu. In synthesis, they have default positions within a prosodic base curve, but they may be excluded or attain F0 values adjusted with respect to adjacent accent points, usually as a carry-over of an intonation gesture.

In compound accent 2, the secondary stress, indicated by Hg, is located on a following word stem. We have adopted simplified but specific sampling rules. The accent 2 high point H* is set at the left boundary of the vowel. The same place is chosen for the low point L* of accent 1. The low point L of accent 2, which indicates the endpoint of the F0 fall, does not have a segmental reference. It is timed by a fast intonation gesture. In synthesis, L is placed 150 ms after the H* point.

The accent 1, H syllable is assigned to the unstressed syllable next before Ha. It has the role of a starting point towards the L*. It shows great variability and is of minor importance.

Figure 1 shows how accent 1 and accent 2 reference points vary as a function of RS from the lower limit of an accented syllable, RS=15, to the highest value, RS=30. These data pertain to sentence medial position.

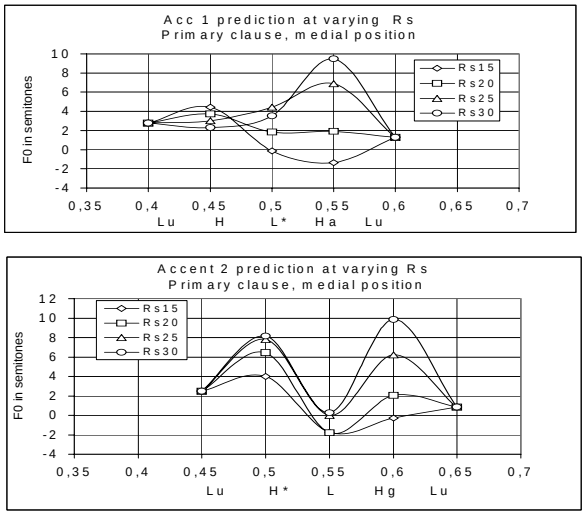

Figure 1. Accent 1 and accent 2 parameters as a function of prominence RS.

The accent 1 parameter H shows a small range of F0 variation only and fuses with L* at higher prominence levels. At low prominence levels, Ha takes the role of a pronounced minimum. This is a consequence of the segmental constraints of our notation system. In the range from RS=20 to RS=30, both Ha and Hg increase by 8 semitones, but H* saturates at RS=25.

2.1. Average and individual intonation patterns

Frequency and time normalized intonation patterns are exemplified in Figure 2. They show the spread within the group of three males and two females and a comparison of our reference female speaker AÖ with the group mean. Some general traits are apparent. There is an overall declination of F0 within the sentence. Rather stable points are found in the height of accent 2 H* peak amplitudes, where the standard deviation among speakers is of the order of 1,5 semitones only.

There is an overall uniformity of speaking behavior across subjects and gender. An exception is the great spread of data points in the word "uppångad", which reflects varying prominence, and to some degree also a varying prosodic boundary marking. More detailed comments are found in [5, 10].

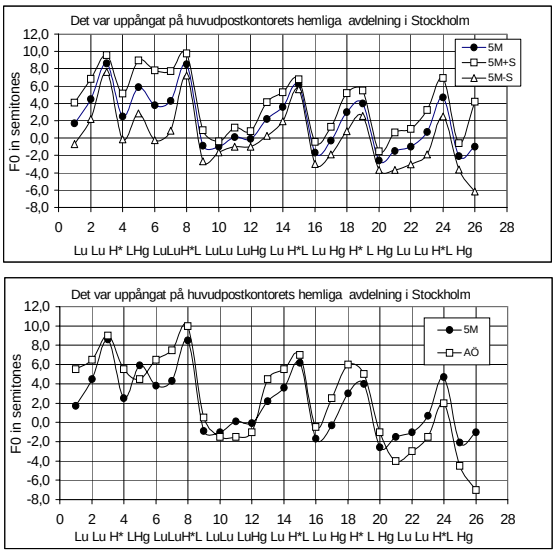

Figure 2. Normalized F0 contours. Above, the mean of the five speakers and the mean plus and minus one standard deviation. Below, the mean and one of the female speakers, AÖ. Sentence translation: "It was steam-opened at the main post- office secret department ".

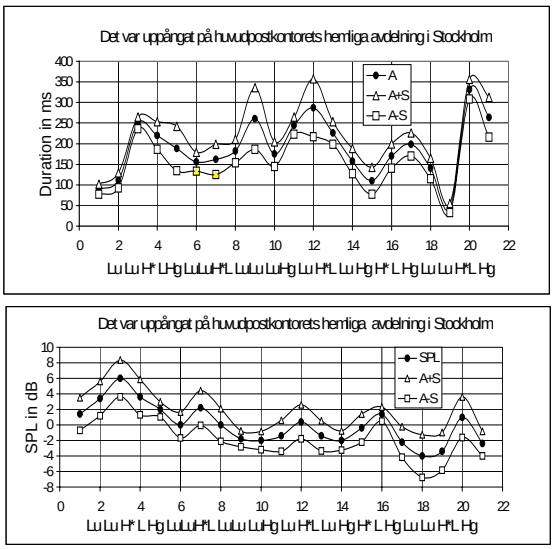

Figure 3. Group averages of syllable duration, above, and sound pressure level, SPL, below. Plus minus one standard deviation have been added.

It is of interest to note that the intonation contour of the female subject AÖ follows the main trend of the group, except for a higher start and a lower sentence final level.

3. Duration and intensity

Data on syllable duration and sound pressure levels are shown in Figure 3. Syllable duration averages 190 ms and has a spread of the order of +/- 25 ms, except for two extreme values, which reflect two subjects' final lengthening. There is no apparent trend of declination within the sentence.

This is contrary to the distribution of the SPL data, which shows a typical down-drift towards the end of the sentence. There are five enhanced regions. These data have been normalized with respect to each subject's mean level, and thus retain relational patterns. The standard deviation of SPL data is of the order of 2 dB.

4. Co-variation of speech parameters

A basic theme of prosodic analysis is the co-variation of F0, duration and intensity. Can one grade their relative importance, and how do they vary with prominence and within a prosodic frame? Are there specific individual patterns? What is the role of sub-glottal pressure? We have dealt with these problems in several publications [3-6, 8-11].

4.1. Regression analysis

A useful tool is regression analysis. However, some caution must be taken with respect to the frame of sampling. A consequence of the positional dependency within a major prosodic group is that it does not pay to construct regression graphs as an average from all data within a sentence. To cite Roman Jakobson, [16]: "Relations must be studied "Ceteris paribus", i.e. in the same context. A syllable in the late part of a sentence may have the same prominence RS as one in an earlier location but a lower SPL (Figure 3).

Accordingly, most of our regression analysis has been confined to the early part of a sentence. Studies of the vowel [a] reported in [4, 8] provided the following data of the correlation coefficients R^2 with respect to prominence as summarized in Table 1.

Table 1: Regression coefficient R^2

Duration	0,80
Subglottal pressure Psub	0,70
Voice source scale factor Ee	0,60
Intensity SPL	0,78
Intensity SPLH	0,82
SPLH- SPL	0,87
Joint duration and SPLH	0,90

SPLH is the high frequency pre-emphasized sound pressure level [4-8]. The difference, SPLH-SPL serves as a measure of spectral tilt. The highest score $R^{2.}$ = 0.90 was found by combining SPLH-SPL with duration. Thus, adding an additional strong co-varying predictor does not improve the prediction score. This appears to be a general finding.

4.2. Co-variation within a sentence

Relations between RS and accent 1 and accent 2 parameters were illustrated in Figure 1. As stated in section 2, the perceived prominence of an accent 2 word is largely carried by the peak F0 in the secondary syllable, denoted Hg. At increasing prominence, above RS=20, the primary peak H* saturates, whilst Hg continues to rise, at RS=30 reaching 10 semitones above the baseline for unstressed syllables. The same range is found for the accent 1 peak Ha. Above RS=20, there is also a tendency of saturation of both intensity parameters and of duration.

The lower part of Figure 4 illustrates the co-variation of F0 and SPL from the sentence in Figures 2 and 3. Within the lower half of a speaker's available F0 range, SPL increases by about 1 dB per semitone. However, as F0 increases towards a focal prominence, RS>22, the saturation of SPL sets, in while F0 continues to rise.

Phoneme durations are found to vary with respect to RS by special rules for different vowel and consonant categories and location within a syllable frame.

The upper part of Figure 4 shows a co-variation of perceptually graded RS values together with RS data predicted from syllable duration. Except for a few failures, e.g. in the middle of the compound accent 2 region, where the graded F0 is higher than the predicted value, there is a fair match. Duration alone apparently has a predictive power.

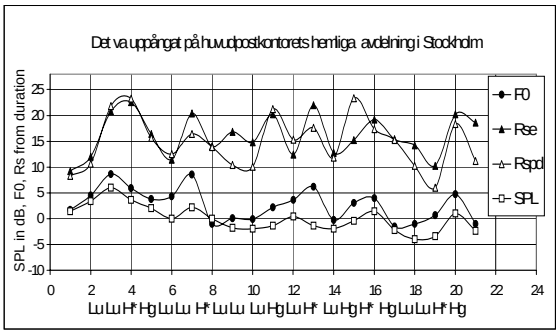

Figure 4. Co-variation of F0 and SPL with perceived RS and RS predicted from duration. Mean of five subjects.

4.3. Sub-glottal pressure and F0

We have a unique set of data from true sub-glottal pressure measurements in a part of our prose-reading corpus. It is documented in [6] and in greater detail in [9]. Here follow some brief comments only.

The main course of Psub within a declarative neutral sentence is a start of about 3.5 cm H_2O followed by a rapid increase to an initial level of about 7 cm H_2O and an overall decay towards the end of the sentence to a final value of about 2 cm H_2O, during which we can expect a simultaneous fall of SPL of the order of 5 dB and of accent peaks by about 5 semitones.

Sentence duration will influence the extent of the decay. Psub is elevated in emphasized regions, but there is an interesting finding that Psub increases well ahead of a stressed syllable and then shows a decay starting at the syllable boundary [6,11]. If the vowel is long, the F0 peak occurs in a region of falling Psub and SPL. At higher RS levels, the F0 contour attains a typical hat pattern [14], more precisely the shape of a symmetrical Gaussian prototype with a base of about 250 ms and a height of the order of 6-9 semitones depending on prominence.

Moreover, a prosodic boundary after a stressed syllable ending with a long vowel may be initiated by an early Psub fall in which the F0 peak occurs in a final interval of decaying intensity [11].

5. Text-to-speech modelling.

Our speech analysis data reveal structures that can be modeled by rules for text-to-speech synthesis. The first step is to divide a sentence into a number of major prosodic groups with associated boundaries to be marked by F0 reset, associated pauses and final lengthening. This is achieved by a gross syntactic analysis supported by a bottom-up survey of data and experience from analysis-by-synthesis procedures.

A sentence may thus be allotted one or more prosodic groups, each with a base curve of prescribed rise and fall characteristics upon which F0 accent modulations of the type shown in Figure 2 are superimposed. An example is shown in Figure 5.

Accent modulations are determined by prominence RS, and also by position within a prosodic group. The FK system is target oriented, with a decision for every syllable, including prosodic carry-over effects. RS may be given default values according to word class.

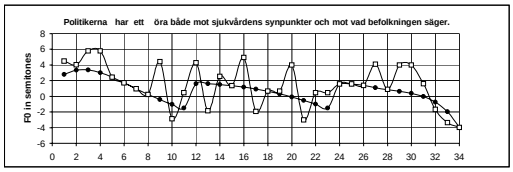

Figure 5. Illustrating the superposition of accent modulations on three successive prosodic base curves. Text translation: "Politicians have an ear, both with respect to health-care needs and to what the population says."

Focal increase and pre- and post-focal reductions have been treated in [4-5]. Duration is modeled by special rules according to phoneme class and RS .

Our data of pause durations from the prose reading material [7] show rather consistent values of the order of 1 second between complete sentences, but there are large individual variations of sentence internal pauses [1, 7]. We tentatively use three levels, 450, 175 and 40 ms. The rules also cover pre-pause lengthening and additional pre-boundary lengthening, which often occur without a pause.

Our results from text-to speech Mbrola synthesis are promising. We achieve a quite natural prosody. Predicted F0 patterns match group averages from our reference data with a mean error of the order of 1.5 semitones [3, 5. 10]. Errors in predicted syllable duration are of the order of 25 ms.

A main conclusion is that the FK system is a powerful tool for experimental variation of prosodic structures and speaking style.

5. References

1. Fant G, Nord L & Kruckenberg A (1986). Individual variations in text reading. A databank pilot study. *STL-QPSR* 4/1986, 1-17.
2. Fant G, Kruckenberg A (1989). Preliminaries to the study of Swedish prose reading and reading style, *STL-QPSR, KTH*, 2/1989, 1-83.
3. 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, KTH*, 2002: 161-164.
4. Fant G, Kruckenberg A & Liljencrants J (2002). Acoustic-phonetic analysis of prosody in Swedish Intonation. In: *Analysis, Modelling and Technology*, Botinis A (ed.). Kluwer, Academic Publishers, 55-86.
5. Fant G & Kruckenberg A (2004). The FK prosody model. Analysis and synthesis. *Proc. Nordic Prosody, IX*, Lund.
6. Fant G & Kruckenberg A (2004). An integrated view of Swedish prosody. In: Fant G (ed) *Speech Acoustics and Phonetics. Selected Writings*. Kluwer Academic Publishers/Springer, 249-300.
7. Fant G, Kruckenberg A, Barbosa-Ferreira (2003). Individual variations in pausing. A study of read speech. *Fonetik 2003, Phonum 9*, Umeå Univ, 193-196.
8. Fant G, Kruckenberg A & Liljencrants J (2000). The Source-filter frame of prominence. *Phonetica* 57: 113-127.
9. Fant G, Kruckenberg A, Liljencrants J & Hertegård S (2000). Acoustic phonetic studies of prominence in Swedish. *TMH-QPSR, KTH*, 2/3: 1-52.
10. Fant G & Kruckenberg A (2004). Intonation analysis and synthesis with reference to Swedish. *International Symposium on Tonal Aspects of Language, TAL 2004*, Beijing, 57-60.
11. Fant G & Kruckenberg A (2005). Covariation of sub-glottal pressure, F0 and intensity, *Eurospeech 2005.*
12. Bruce G (1977). *Swedish Word Accents in Sentence Perspective* . Lund: Gleerup.
13. 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: Botinis A (ed). *Intonation. Analysis Modelling and Technology*, Kluwer Academic Publishers, 291-320.
14. Collier R (1991). Multi-language intonation synthesis. *Journal of Phonetics*, 19: 61-73.
15. Portele P & Heuft B (1997). Toward a prominence based synthesis system. *Speech Communication*, 21: 61-72.
16. Jakobson R, Fant G & Halle M (1952). Preliminaries to speech analysis. The distinctive features and their correlates. MIT press, 7th ed, 1967.