Lund University, Dept. of Linguistics
Working Papers 49 (2001), 38–41

A novel system for F0 analysis and prediction

Gunnar Fant and Anita Kruckenberg
Dept. of Speech Music and Hearing, KTH
gunnar@speech.kth.se

Abstract

This in an updated version of earlier work on quantification and normalisation of F0 data in a frame of the Swedish accent system. It serves two functions. One is descriptive, suitable for cross-subject comparisons and for studies of average intonation contours. The other is directed to rules for predicting intonation patterns of relevance for text-to-speech synthesis. Our continuously scaled prominence parameter, Rs, has an important role in the quantification of F0 data.

1 Introduction

This is a continuation of our earlier work on descriptive analysis, quantification and prediction of F0 patterns in Swedish (Fant and Kruckenberg 1999, 2001). We have clear evidence of the superiority of expressing F0 data on a semitone scale in stead of a linear Hz scale which is the standard in most speech processing and display systems. A second step in normalisation is to correct for the mean F0 of a particular subject. In addition we perform a temporal normalisation by referring sampled F0 data to successive syllable slots instead of a continuous time scale. As a result we can derive representative mean F0-patterns of a group of mixed male and female subjects or we may compare different subjects' realisations.

Our analysis is largely adapted to conform with the Bruce, 1977, classical system for description of accent 1 and accent 2 patterns, but our sampling and labelling routines are quite specific. We assign two samples per accented syllable and one sample only for all other syllables.

The perceived prominence of a syllable or a word, our Rs parameter, is determined from listening tests. It is continuously scaled from 0 to 30, with typical values of 10 for unstressed syllables and 20 for accented syllables. Studies of the co-variation of Rs with syllable type, accent 1 or 2, F0 and position of the syllable in utterances of various types have provided a base for prediction rules. These also include statistical data for predicting Rs from word class, position and some additional higher level considerations. Prosodic boundary rules are under way. It is now possible to incorporate our F0 predictions in text-to-speech systems.

The accuracy of prediction, as judged with respect to measured data, is high but there remain tests to be made comparing our approach with that of established systems.

Our original studies (Fant and Kruckenberg 1999, 2001) were based on the recordings of three males and two females reading a passage of one minutes length, corpus 1. We now have a material from a following part of the novel of the same length, corpus 2. Rules derived from the earlier recording can now be tested against a fresh material.

2 Analysis

2.1 Experimental techniques

F0 traces on a log scale were printed out in synchrony with oscillogram, spectrogram and intensity curves (Fant et al. 2000). Our calibration standard was 2 mm per semitone (st), i.e. 24 mm per octave. Measurements were made within 0.5 semitone. All F0 values were initially expressed in an absolute scale of semitones (st) relative 100 Hz.

A normalisation based on each speaker's average F0 in unstressed syllables in the entire reading was introduced. Accordingly, a correction of -7 st respectively -9.5 st were applied to the female data and -1 st respectively +1 st and 0 st for the male data. As a result the female data were effectively lowered to match the male data.

Similar individual corrections were derived from the second recording. Moreover, it was found that the average of F0 measures from all sampled data points of a speaker exactly matched those from unstressed syllables which reveals an interesting symmetry.

Assessments of Rs ratings for all five subjects have been limited to accentuated words. Nouns, adjectives and numerals scored average values of 21, verbs 19 and adverbs 18,5. No significant difference was found between accent 1 and 2.

2.2 Data sampling and display

Our F0 data sampling and labelling of accent parameters largely conforms with the canonical description and notations of Bruce (1977). A few minor additions and some specific interpretations of data labels have been introduced.

Our present routine is confined to the sampling at two positions within the major syllable of an accented word, L* and Ha for accent 1 and H* and L for accent 2. All other syllables, i. e. those denoted H, Hg or Lu are given one sample point only. H pertains to the pre-accent 1 syllable, Hg the secondary syllable of accent 2, and Lu all unaccented syllables. Thus the word "margarinlåda" would be denoted: Lu Lu H* L Hg Lu.

Lu measurements are referred to the middle of the vowel. In a weak accent 1 syllable the F0 contour may show a continuing fall instead of an L* Ha rise. As a consequence, our routine in these instances of sampling L* at the beginning and Ha at the end of the vowel produces a negative (Ha-L*). At higher degrees of prominence Ha refers to the peak of an F0 maximum in the middle of the vowel or of the voiced part of the syllable. Ha and Hg, potentially carrying "sentence accent", display an approximately equal rate of increase or decrease with prominence.

2.3 F0 data display

In order to visualise a connected F0 contour from the sampled data of a sentence we performed a smoothed continuous record of successive sample points based on Excel routines. The result is a time and frequency normalised intonation contour in which unvoiced portions are overbridged. The time scale is substituted by a sequence of data slots and the frequency scale is in semitones relative 100 Hz. Individual variations in timing and tempo are thus excluded and the data for any speaker is corrected for his or her average F0.

This system has obvious advantages for studies of the intonation of several subjects, and it provides efficient means of calculating average contours and individual variations. We thus found that the spread of the five subjects around their mean, expressed as a standard deviation, usually varied in the range of 1 to 2 semitones. An example comparing one of the females, AÖ, with the mean of the five subjects is illustrated in Figure 1.

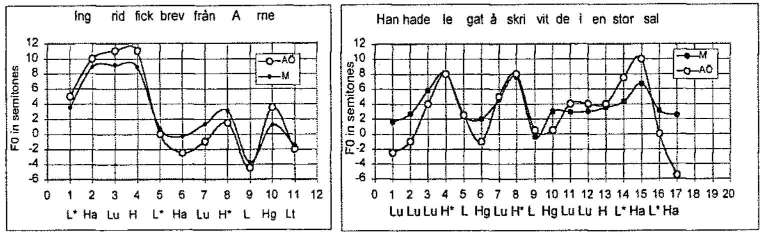

Figure 1. Normalized F0-contours. M is the mean of five subjects and AÖ is one of the female subjects.

3 Modelling and prediction

As described in earlier reports (Fant and Kruckenberg 1999, 2001) we have developed a model in which a normalised contour is derived from local accent modulations superimposed on a base contour defining F0 of unstressed syllables, denoted Lu, which has been found to coincide with the mean of the local modulation range. The base curve has a prescribed initial rise, declination rate, and in some contexts a specific final fall. At present we have two alternatives only, one for the initial clause of a new sentence and one for following clauses, typically a sentence final clause, which ends with a relative large pause. A more detailed modelling of the base curve, especially at its terminations, could take into account our finding that the size of the F0 reset at a major prosodic boundary varies in proportion to the pause duration (Fant and Kruckenberg 1999, 2001) and thus to the level of syntactic hierarchy.

The extent of F0 declination is approximately the same in short and long prosodic groups of the order of a breathgroup. We accordingly specify the influence of position in terms of a relative scale from 0 to 1.

The local modulations defined by the H L* Ha of accent 1 and the H*LHg of accent 2 contours are derived from an assumed Rs of the word as a single unit and relative positions of the data points on the base curve. Calculations involve highly non-linear regression analysis which takes into account the type of clause, primary or secondary.

4 Results

The results from a prediction of intonation contours performed on corpus number one have been quite promising (Fant and Kruckenberg 1999, 2001). Predictions based on assessed prominence of accented words usually agree within 2 semitones with the group average of the five speakers. The best fit occurs for H* turning points of accent 2 where the prediction error is of the order of one semitone only.

How effective is the modelling in a text-to-speech application, where Rs values have to be determined from lexical class and grammatical structure? An indication is given in Figure 2 which pertains to one primary and two secondary clauses from corpus 2 predicted from equations derived for corpus 1. The overall agreement is substantial.

5 Conclusions

We have developed a novel system of documenting F0 data based on frequency and time domain normalisation, which is well suited for comparing individual performances and deriving average data for a group of subjects (Fant and Kruckenberg 1999, 2001).

In addition we have developed the main body of programs for predicting F0 contours from text. Our prominence parameter Rs ensures a continuous scaling of accentual realisations thus avoiding limitations imposed by discrete phonological classifications such as focal accent, but on the other hand providing a frame for the phonological system.

Compared to established systems (Carlson and Granström 1973, Bruce et al. 2000) our modelling is more detailed which should have some advantages, but requires an additional set of rules for avoiding local irregularities. As in any other prediction scheme we need rules for phonetically definable assimilation and rules for juncture realisations, global gestures (Gårding 1989), and reorganisations imposed by focal prominence. Perceptually oriented system comparisons are needed.

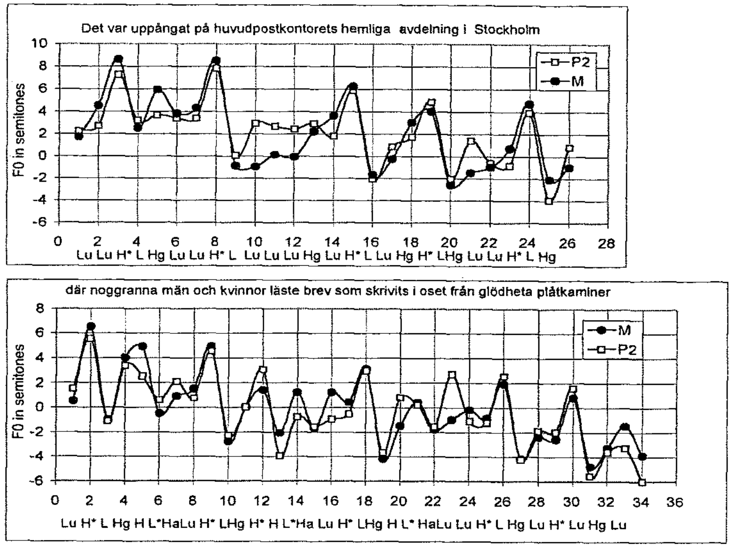

Figure 2. A test of corpus 1 rules applied to a corpus 2 sentence with Rs predicted from word class and position within the clause.

References

Bruce, G. 1977. Swedish Word Accents in Sentence Perspective. Gleerup, Lund.

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

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

Fant, G. & A. Kruckenberg. 1994. 'Notes on stress and word accent in Swedish'. *Proceedings of the International Symposium on Prosody, Sept 18 1994, Yokohama*. Also published in *STL-QPSR* 2-3/1994, 125-144.

Fant, G. & A. Kruckenberg. 1999. 'F0-patterns in text reading'. In J. Allwood (ed), *Proc of Fonetik 99*. Gothenburg papers in theoretical linguistics. Göteborg University, 53-56.

Fant G. & A. Kruckenberg. 2001. 'F0 analysis and prediction in Swedish prose reading'. In N. Grønnum and J. Rischel (eds.) *To honour Eli Fischer-Jørgensen*. Travaux du Circle Linguistique de Copenhague. Reitzel Copenhagen, 124-147.

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

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