MODELLING DANISH SENTENCE AND PHRASE INTONATION

Niels Reinholt Petersen
Institute of General and Applied Linguistics, University of Copenhagen, Denmark

ABSTRACT

The work reported in the paper explores the possibility of formulating (parts of) a model for Danish intonation in mathematical terms. It is hypothesised that—if the prosodic phrasing is known—F_0 of a stressed syllable can be described by a linear function with two independent variables, viz. its position in the sentence and in the prosodic phrase. Analysis of a material of read-aloud sentences shows the hypothesis to be tenable. Further analysis indicates that the mathematical model can be used for generating sentence and phrase intonation contours in a text-to-speech system.

1. INTRODUCTION

The work reported in the present paper is part of the development of the intonation rule module for a Danish text-to-speech system. Previous research on Danish intonation carried out primarily by Nina Grønnum (see e.g. [1] and [2] and references therein) has shown that Danish intonation can be described in terms of a hierarchical model where components of smaller temporal scope are superposed on components of larger temporal scope. So far the following components have been established: 1) a *text component*, 2) a *sentence component*, 3) a *prosodic phrase component*, 4) a *stress group component*, 5) a *'stød' component*, and 6) a *microprosodic component*.

The work described below focuses attention on the sentence and prosodic phrase components. In Danish, sentence function is signalled globally by the slope of the general fundamental frequency contour: in terminal declarative sentences the contour is steeply falling over the sentence, in continuative sentences and non-echo questions, the fall is less steep, and in echo-questions the contour is horizontal. In terminal declaratives, the slope of the sentence contour has been shown to vary with sentence length, being steeper in short than in longer sentences, i.e. the fundamental frequency range is independent of sentence length. Further, in sentences of more than 4 – 5 stress groups the sentence is divided into a number of prosodic phrases, each having its own declining F_0 contour starting with a (partial) resetting and superposed on the sentence contour [3, 4]. The prosodic phrasing of a sentence seems to be the result of a complex—and not very well understood—interaction between a number of factors factors, such as syntactic structure, semantic content, and a tendency to avoid long (more than 5 stress groups) prosodic phrases.

The fact that Danish intonation can be described by a hierarchical, superpositional model and that sentence function is signalled globally leads to the hypothesis that—if the prosodic phrasing of a sentence is known—the fundamental frequency of a stressed syllable can be described in mathematical terms as a *linear function* of two independent variables, namely its position in the sentence and its position in the prosodic phrase:

(1) $$F_0 = \alpha_S\, p_s + \alpha_S\, p_p + \beta$$

where α_s, α_p, and β are the sentence slope, the phrase slope, and the intercept, respectively, and p_s and p_p the position of the syllable in the sentence and in the prosodic phrase.

This basic hypothesis can be tested by means of multiple linear regression analysis of a spoken sentence material, and if the hypothesis turns out to be tenable (i.e. if the correlation coefficients obtained are sufficiently high) it will be meaningful to examine whether the constants of the linear regression equation can be derived from information independent of the fundamental frequency contour, such as sentence length and phrasing.

2. METHOD

2.1. Material and speakers

Measurements from two previously recorded materials were analysed. One sentence material (henceforth material A) consisted of 8 terminal declarative sentences spoken once each by 8 male speakers of Standard Danish. The sentences were syntactically relatively complex and varied in length between 7 and 12 stress groups (see further [5]). The other material (material B) consisted of two sets of terminal declarative sentences recorded and measured by Nina Grønnum [3, 4], who kindly made the measurements available for the present research. The sentences were syntactically relatively simple and varied in length between 2 and 8 stress groups. Data for two male speakers of Standard Danish (one was the author) was used here. One set of sentences was spoken 6 times by both speakers and the other set was spoken 6 times by one speaker and 3 times by the other speaker. Altogether the material comprised 211 sentences (a sentence meaning here and in the following one rendering of a sentence by one speaker), 64 sentences in material A and 147 in material B.

2.2. Regression analysis

In each sentence the measured fundamental frequencies were converted into semitones relative to the mean F_0 of the sentence in question.

The hypothesis to be tested (as formulated in equation 1) presupposes that the prosodic phrase boundaries are known. This is not the case; the lack of precise knowledge of the complex interaction between several factors makes it highly problematic to predict prosodic phrasing with any certainty from external (higher level) information, nor can the prosodic boundaries be determined unambiguously by visual inspection of fundamental frequency tracings. In the present investigation, therefore, each sentence was analysed using all prosodic phrasings possible for that sentence (including the possibility that the sentence comprises only one phrase), the only constraint being that a phrase must consist of at least two stress groups. The phrasing yielding the highest correlation coefficient was selected as *the* prosodic phrasing for the sentence in question, and this was the phrasing employed in the subsequent treatment of the data.

Multiple regression analysis was used for all possible phrasings of two or more phrases. In possible phrasings of one prosodic phrase, simple regression analysis was carried out, and the phrase slope was set to zero.

It not clear, *a priori*, whether the position of a stressed syllable should be defined as physical time relative to sentence and phrase onset or whether it should be defined as its ordinal number in the sentence and the phrase. In order to answer this question, material A was analysed using both definitions of syllable position. Material A was used because it showed a greater variation in stress group duration than did material B, and could, therefore, be expected to reveal any differences between the two definitions of syllable position. As it turned out, the multiple correlation coefficients were slightly higher for number than for time as the definition of syllable position. Although the mean difference was small (0.015), a t-test for paired observations showed it to be stastistically significant (p<0.001, one-tailed). On this basis it was decided to define the position of a syllable as its number in the sentence and the phrase.

3. RESULTS

3.1. The basic hypothesis

The results of the regression analyses are summarized in table 1. It is seen that the correlation coefficients obtained are generally very high, whether the entire material is considered or whether the sentences with only two stress groups (where the correlation coefficient will always be 1) are left out. In material A, which is the more heterogeneous, each sentence having been spoken only once by each speaker, the correlation coefficients are slightly smaller; but still the vast majority of the coefficients are higher than 0.9.

	median	R>0.9	R>0.95	R>0.98	N
All sentences	0.973	192 (91)	154 (73)	87 (41)	211
>2 stress groups	0.970	171 (90)	133 (70)	66 (35)	190
Material A	0.933	46 (72)	25 (39)	5 (8)	64

Table 1. Median multiple correlation coefficients and numbers and percentages (in brackets) of coefficients greater than 0.9, 0.95, and 0.98 obtained in the entire material (211 sentences), in sentences with more than two stress groups (190 sentences), and in material A (64 sentences).

Figure 1 displays estimated and observed fundamental frequencies (in semitones) over the entire material. The correlation between estimated and observed values is 0.952, which means that 90.5% (=100×r^2) of the total F$_0$ variation in the stressed syllables can be accounted for by their position in the sentence and the prosodic phrase. The amount of the total F$_0$ variation in material A accounted for is 86% (r=0.928).

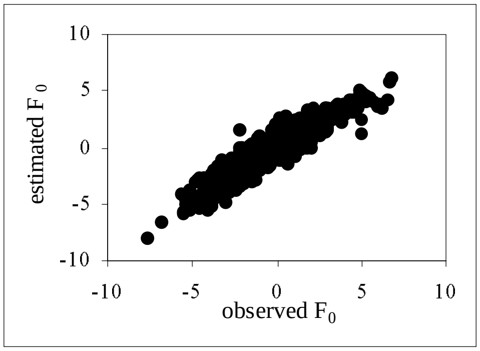

Figure 1. Estimated versus observed F$_0$ in semitones in the entire material, 211 sentences, 1374 data points.

These results present substantial evidence in support of the basic hypothesis—as formulated in equation 1 above—that Danish sentence and phrase intonation can be described quantitatively in terms of a linear regression model, and they also speak in support of the general model of Danish intonation as a hierarchical, superpositional system.

3.2. Derivation of regression constants

Apart from the correlation coefficient, the regression analysis supplies the linear equation constants (sentence slope, phrase slope, and intercept) determined by means of the method of least squares from the data analysed. In view of the high correlation coefficients obtained in the present material, it seems worthwhile to examine whether the constants describing the sentence and phrase intonation of a given sentence can be derived from information on its length and phrasing, i.e. from 'F$_0$-external' information. If the F$_0$ values estimated from derived constants compare reasonably well with those observed in the material, it seems justified to utilize the regression model as the basis for a set of rules generating sentence and phrase intonation in a text-to-speech system for Danish.

3.2.1. The sentence intonation slope. Since, as mentioned in section 1, the fundamental frequency range of a sentence is independent of its length, the sentence intonation slope, α_s, can be expected to vary as a hyperbolic function of sentence length. Figure 2 shows the observed sentence slopes together with the best fitting hyperbolic function, which is given by the equation

$$\alpha_S = \frac{-1}{0.2044(n_S - 1)} \tag{2}$$

where n_s is the number of stress groups in the sentence.

The correlation between observed sentence intonation slopes and slopes estimated by equation 2 is high, r=0.934, which means that 87% of the total observed variation of the sentence slopes can be explained by sentence length.

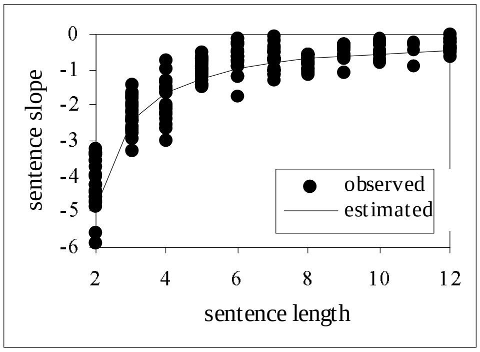

Figure 2. Observed sentence intonation slopes, α_s versus sentence length in number of stress groups. The line is the hyperbolic function (equation 2) fitted to the data points.

3.2.2. The phrase contour slope. The observed phrase slopes in sentences with more than one prosodic phrase (in one-phrase sentences the phrase slope was set to 0, as mentioned in section 2.2) varied between –0.02 and –3.218 semitones per stress group. In order to reveal any systematic pattern of variation of the phrase slopes, a number of correlation analyses were carried out using sentence length and number of phrases in the sentence as independent variables. As it turned out, the best fit to the observed phrase slopes in sentences with two phrases or more is given by linear function

(3) $$\alpha_p = 0.0025 n_s - 0.289 n_p - 0.117$$

where α_p is the phrase slope, n_s the number of stress groups in the sentence, and n_p the number of phrases in the sentence. The correlation between observed and estimated phrase slopes is moderate, r=0.444, i.e. slightly less than 20% of the total variation can be explained by the two independent variables.

With a phrase slope of 0 in one-phrase sentences and estimated elsewhere by equation 3, the correlation between observed and estimated slopes in the entire material becomes 0.714 (see figure 3).

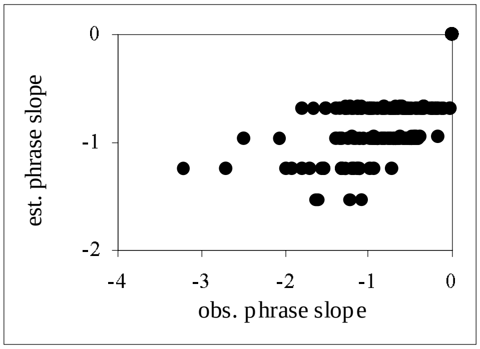

Figure 3. Oberserved and estimated prosodic phrase slopes in semitones per stress group for all sentences in the material.

3.2.3. The intercept. In a linear function the intercept will be negatively correlated with the slope, or slopes if more independent variables are involved. In the present material the correlation between observed sentence and phrase slopes and observed intercepts is certainly high, r=.84. But in a set of rules generating fundamental frequency, the intercept will have to be computed from sentence and phrase slopes derived on the basis of external information as described above. The best estimation of the observed intercepts from the slopes estimated by equations 2 and 3 is given by the equation

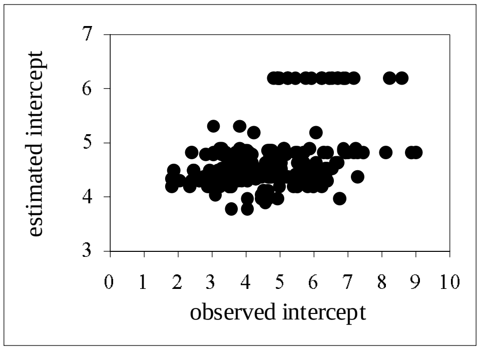

Figure 4. Observed and estimated intercepts in semitones for all sentences in the material.

(4) $$\beta = -0.739\alpha_s - 1.505\alpha_p + 2.568$$

where β is the intercept, α_s the sentence slope, and α_p the phrase slope.

Figure 4 displays the observed and estimated intercepts over all sentences in the material. As can be seen the correlation is moderate, r=.399; less than 16% of the total observed variation can be accounted for by the derived sentence and phrase slopes.

3.2.4. Comparison of observed and generated F_0. Figure 5 shows the observed fundamental frequencies versus those estimated from the derived regression constants. The correlation between observed and estimated F_0 over the entire material is high, r=0.883, and is only slightly lower than the correlation found by the regression analyses (cp. Section 3.1), which is 0.952. For material A the correlation coefficients are 0.847 and 0.928, respectively.

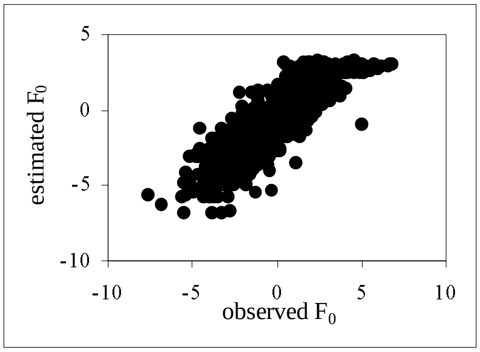

Figure 5. Observed F_0 and F_0 estimated from derived regression constants, 211 sentences, 1374 data points.

On this basis it seems fair to conclude that the fundamental frequency in sentences and prosodic phrases can be generated with reasonable accuracy on the basis of derived regression equation constants.

4. DISCUSSION

The results of the research reported in the present paper have theoretical as well as practical implications.

At the theoretical level, they contribute quantitative evidence in favour of a hierarchical, superpositional model of Danish intonation. In particular, the close relationship between sentence length and sentence intonation slope found in the present material speaks strongly in support of one of the crucial points of the model, namely that sentence function is signalled globally, implying a look-ahead mechanism to be involved in the production of fundamental frequency.

A second point of theoretical interest is the finding that an ordinal definition of (stressed) syllable position gives a more accurate (although slightly so) estimation of observed F_0 than does position defined as physical time. Although it needs further corroboration, this finding could be speculated to suggest that the planning of fundamental frequency takes place at an earlier stage in the speech production process than does the planning of timing.

The most evident practical implication of the present research is that it has been demonstrated that the regression model is able to generate realistic sentence and prosodic phrase intonation contours. On this basis it can be assumed that the model can be incorporated in an intonation rule system for Danish without essential modifications.

Another practical implication is that—in a language with an intonation system similar to that of Danish—regression analysis carried out on a number of possible prosodic phrasings of a given sentence, as described in section 2.2, can be utilized as a powerful tool for the locating of prosodic phrase boundaries from fundamental frequency measurements. The methods previously used for locating phrase boundaries (i.e. to determine where F_0 resettings occur) have been based on either visual inspection or on criteria based on the relation between neighbouring stressed syllables: if the slope between two such syllables is less steeply falling than the adjacent slopes [3, 4] or if the slope is level or rising [5], then a resetting is assumed to exist between the two stressed syllables. Both criteria are based on local F_0 relations and are, therefore, highly sensitive to random variation. In contradistinction to this, the regression analysis takes the entire sentence into account and the multiple correlation coefficient provides an objective criterion for choosing the best prosodic phrasing from among several possible phrasings, and at the same time it is a measure of the validity of the phrasing chosen.

A prerequisite for the implementation of the regression model as part of an intonation rule system is that prosodic boundaries can be reliably specified on the basis of external (higher level) information. In continuation of the research described in [5], work is in progress looking into the relationship between syntax and prosodic phrase boundaries using the best phrasings as determined by regression analyses. The preliminary results indicates that syntactic structure plays an essential role as a predictor of prosodic phrasing, but tends to predict too few prosodic phrase boundaries. Thus, other factors presumably involved in the control of prosodic phrasing, such as semantic contents and a tendency to avoid long phrases, will have to be taken into consideration. In the present material there are indications suggesting an interplay between sentence length and phrase length which—if it can be formalized—may account for some of the prosodic phrase boundaries not accounted for by syntactic structure.

REFERENCES

[1] Grønnum, N 1992. *The groundworks of Danish intonation*. Copenhagen, Museum Tusculanum Press.

[2] Grønnum, N. 1995. Superposition and subordination in intonation: a non-linear approach. In Elenius, K. and P. Branderud (eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences,* vol. 2. Stockholm.

[3] Thorsen, N. 1980. Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish. *Annual Report of the Institute of Phonetics, University of Copenhagen*, 14, 1-29.

[4] Thorsen, N. 1981. Intonation contours and stress group patterns in declarative sentences of varying length in ASC Danish – supplementary data. *Annual Report of the Institute of Phonetics, University of Copenhagen*, 15, 13-47.

[5] Reinholt Petersen, N. and P. Molbæk Hansen 1994. Fundamental frequency resettings, pauses, and syntactic boundaries in read-aloud Danish prose. *Acta Linguistica Hafniensia*, 27, 2, 383-401.