Intonation contours in Danish spontaneous speech

John Tøndering
Department of General and Applied Linguistics, University of Copenhagen, Denmark
johtnd@cphling.dk

ABSTRACT

The current model of Danish intonation can be formalized as a mathematical function, where F_0 of any stressed syllable is the result of a linear function with two independent variables: the position of the stressed syllable in the sentence and its position in the prosodic phrase. I have applied two similar functions to a small corpus of *spontaneous speech*. Whether F_0 in the stressed syllables can be accounted for by a mathematical function has been tested by means of (a) multiple linear regression analysis and (b) multiple polynomial regression analysis. Only 14.1% (a) or 34.1% (b) of the variation in the observed F_0 can be accounted for by these functions. The linear regression on read speech reaches 90.5%.

1. INTRODUCTION

Intonation in Danish spontaneous speech is almost unexplored. This paper will report on a study whose goal is to investigate whether parts of the intonation in Danish spontaneous speech could be described by Grønnum's (henceforth NG) model of intonation in Danish read aloud speech [1]. NG has shown that Danish intonation can be described as a layered, superpositional model where intonation contours of smaller temporal scope are superposed on contours of larger temporal scope (see Figure 1).

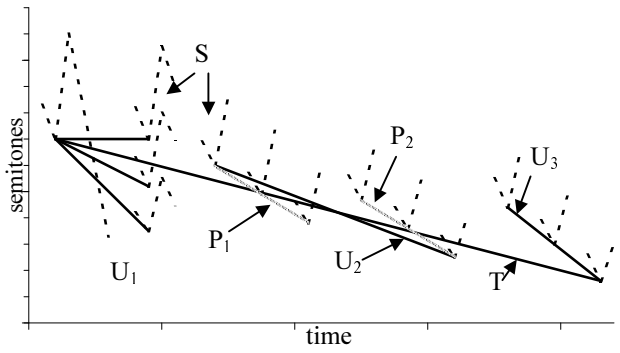

Figure 1: Grønnum's model of Danish intonation, adapted from [1]. *T* is the textual intonation contour, *U* is the utterance contour, *P* is the prosodic phrase contour, and *S* is a prosodic stress group pattern. See further the text.

The intonation contour of a short text, symbolized *T* (cf. Figure 1), is characterized by a gradual global descent. The onset is defined as the first stressed syllable in the first utterance, U_1, and the offset is defined as the last stressed syllable in the last utterance, U_3. The utterance contours are superposed on the textual contour. Utterances with more than 4-5 stressed syllables are divided into a number of prosodic phrases, each phrase contour, P_1 and P_2, being superposed on the utterance contour, U_2. Finally, each stress group pattern, *S*, is superposed on the phrase contour. The local minimum in this pattern coincides with the stressed vowel. In this manner the stressed vowels actually define the contours.

The utterance function in Standard Danish is signaled globally by the slope of the utterance contour. Terminal declarative sentences are typically accompanied by the most steeply falling contour and echo-questions are typically accompanied by a horizontal contour. Other questions and non-final clauses will have a slope somewhere between these two extremes (cf. the three alternative slopes of U_1 in Figure 1).

[2] and [3] showed that parts of NG's model can be formulated in mathematical terms. If the prosodic phrasing of a sentence is known, the fundamental frequency of a stressed syllable can be described as a linear function of two independent variables, namely the stressed syllable's position in the sentence and its position in the prosodic phrase. Reinholt Petersen tested his hypothesis by means of multiple regression analysis and found that the two variables could account for 90.5% of the total F_0 variation in the stressed syllables. His test material was read speech and the results, not surprisingly, confirm NG's model

In the present investigation it is hypothesized that the mathematical function can be expanded to cover spontaneously produced texts. The goal is to explore to what extent NG's model could be said to hold for intonation in spontaneous speech. But the steps from utterance to text and from read to spontaneous speech could not be taken without making certain assumptions.

2. METHOD

2.1 MATERIAL

The material used in the present investigation is part of a larger material recorded by NG. The investigated subpart are a couple of map task monologues. I have analyzed 13 of these route descriptions, chosen from 8 speakers. The stressed syllables were marked up by NG, and I extracted F_0 along with its time coordinate. There are from 27 to 108 stressed syllables in each text, totaling 681.

2.2 FROM UTTERANCE TO TEXT

Reinholt Petersen ([2]) investigated syntactically well

ISBN 1-876346-nn-n

formed read sentences. But spontaneous speech does not consist exclusively of syntactically well formed sentences. I circumvented the problems inherent in disentangling the syntax/prosody interplay by simply ignoring the sentence component: A route description is a text, which is decomposed into prosodic phrases. Of course this short cut will have an adverse effect on the final results.

2.3 FROM READ TO SPONTANEOUS SPEECH

NG's model is restricted to informal but distinct monologues and is based on context neutral utterances [4]. This is a fundamental constraint of the model and ensures that all stressed syllables have equal degrees of prominence. But this condition does not hold for spontaneous speech. Instead of trying to compensate for this variation the assumption is made that all stressed syllables are equally prominent. Again, this can have an adverse effect on the results.

2.4 THE HYPOTHESES

As mentioned earlier Reinholt Petersen [2] found that F_0 of a stressed syllable in a sentence could be described in mathematical terms as a linear function of two independent variables, namely the stressed syllable's position in the sentence and the stressed syllable's position in the prosodic phrase. In the present investigation the text level is involved, but the sentence level is ignored. This leads to the hypothesis that F_0 in a stressed syllable could be described as a linear function with two independent variables, that is the stressed syllable's position in the *text* and its position in the prosodic phrase:

$$F_0 = \alpha_t p_t + \alpha_p p_p + \beta \qquad (1)$$

α_t is the slope of the textual intonation contour, α_p is the prosodic phrase contour, and β is the intercept. p_t and p_p is the position of the stressed syllable in the text and in the prosodic phrase, respectively. The validity of (1) should be tested by means of multiple linear regression analysis.

If long text contours level out medially, as one would expect from [4], a third degree polynomial may be a better approximation:

$$F_0 = \alpha_{t^3} p_t^{\,3} + \alpha_{t^2} p_t^{\,2} + \alpha_t p_t + \alpha_{tp} p_t p_p + \alpha_p p_p + \beta \qquad (2)$$

p_t and p_p still refer to the position of the stressed syllable in the text and in the prosodic phrase.

Equations (1) and (2) presuppose that prosodic phrase boundaries be known. When Reinholt Petersen [2] tested his hypothesis he tested it on all the possible prosodic parsings of the sentences. But in the present investigation that is not practicable. If a prosodic phrase contains a minimum of 2 and a maximum of 8 stressed syllables, a text with 99 stress groups would have $51 * 10^{18}$ possible phrasings.

The problem of prosodic boundaries is discussed in [5]. Here the boundaries were established by listener judgments. But in the present investigation the boundaries were determined mechanically. Since, according to NG's model, a prosodic phrase will have an almost rectilinear slope, and a prosodic phrase boundary is associated with a reset (cf. *P* in Figure 1), this was accomplished by finding rectilinear phrase intonation contours separated by a reset. Accordingly, a linear regression line associated with the stressed syllables of a prosodic phrase would have a high correlation coefficient. This means that the prosodic phrasing could be established as the best fit among mechanically selected series of rectilinear regression lines. But it is also obvious that the potential phrasing of a text depends on the point of departure. If e.g. stressed syllables number 1 to 4 form a prosodic phrase then stressed syllables number 3 to 6 are not candidates – only number 5 and 6 are eligible. These considerations led to 3 different criteria for finding the prosodic phrases of a text:

a) The procedure starts in the first syllable of the text. The correlation coefficient of the regression line associated with stressed syllable number 1, 2 and 3 is calculated. If the correlation coefficient is below 0.9, syllable number 1 and 2 constitute a prosodic phrase and a new prosodic phrase starts in syllable number 3. If the correlation coefficient is equal to or above 0.9, the regression line associated with syllable number 1 to 4 is considered. And so on, to stressed syllable number 8.

b) This is the mirror image of (a) – that is the procedure takes the last syllable of the text as point of departure.

c) Identical to (b) but the demand on the correlation coefficient is lowered to 0.8.

To summarize: Two hypotheses are being tested: a) a linear function, and b) a polynomial function. Both hypotheses are based on two constraining assumptions, namely that the utterance level can be ignored, and that all the stressed syllables have equal degrees of prominence. Finally it should be noted that the position of a stressed syllable can be defined either as a time coordinate, relative to text and phrase onset, or it can be defined by its ordinal number in the text and the phrase. Both definitions were tested.

3. RESULTS

3.1 LINEAR REGRESSION ANALYSIS

The results of the multiple linear regression analysis are summarized in Table 1. As can be seen from Table 1 the correlation coefficients obtained are very low. The best correlation between observed and estimated F_0 is obtained with model (a) and "ordinal" stressed syllables. The correlation is 0.376, which means that 14.1% of the total F_0 variation in the stressed syllables can be accounted for by their ordinal position in the text and in the phrase. Reinholt Petersen [2] obtained 90.5% with this model.

3.2 POLYNOMIAL REGRESSION ANALYSIS

The results of the multiple polynomial regression analysis are shown in Table 2. They are better. First of all, more results are significant at the 5% level or better – actually only one text does not fulfill this demand under the ordinal

number and model (a) conditions. However, only 34.1% (0.584^2) of the total F_0 variation in the stressed syllables can be accounted for by their ordinal position in the text and in the phrase.

		R – ordinal number		
Phrasing model		a	b	c
Mean value		0.351	0.360	0.339
p<0.05		7	7	7
Observed	vs.	**0.376**	0.351	0.343
		R- physical time		
Phrasing model		a	b	c
Mean value		0.282	0.304	0.314
p<0.05		4	2	4
Observed	vs.	0.302	0.280	0.273

Table 1. Multiple linear regression analysis – Equation (1). The mean values summarize the results from all 13 analyzed texts. p<0.05 shows the number of significant results, and the correlation between the observed and the estimated F_0 values is also shown.

	R – ordinal number		
Phrasing model	a	b	c
Mean value	0.573	0.560	0.530
p<0.05	12	10	8
Observed vs. estimated	**0.584**	0.577	0.540
	R- physical time		
Phrasing model	a	b	c
Mean value	0.541	0.524	0.509
p<0.05	9	10	7
Observed vs. estimated	0.550	0.537	0.515

Table 2. Multiple polynomial regression analysis – Equation (2). See further the caption to Table 1 and the text.

This result is not impressive but NG's observation from read speech that textual contours level out medially may in fact carry over to spontaneous speech.

3.3 OTHER INTERACTIONS

The ensemble of constraining assumptions (equally distributed prominence and no autonomous sentence intonation domain) and hypotheses (textual intonation contours are either linear or polynomial functions) must be rejected. The assumption of equally distributed prominence is very likely highly responsible for the negative outcome. But equation (1) also presupposes that all prosodic phrases of a text have the same slope. That is explicitly not so in NG's model, but a consequence of the mathematical function.

As can be seen from Figure 2, equation (2) – the third degree polynomial – does not presuppose that all prosodic phrases of a text have the same slope, but the slopes of the prosodic phrases are deflected from the slope of the textual intonation contour. Though this seems to be more consistent with the data it still makes a claim about the slope of the prosodic phrases that is inconsistent with NG's

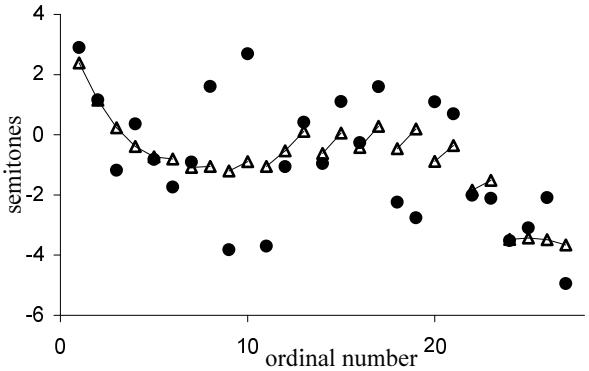

Figure 2. Test item 2. The observed F_0 values are shown as bullets, the estimated F_0 values are shown as triangles. The lines indicate the prosodic phrasing. The estimation is based upon equation (2).

model.

The text in test item 2, Figure 2, is (stressed vowels in bold face): *d**u** kører n**e**d til eller d**u** går n**e**d til kr**y**dset for **e**nden af V**e**stergade og s**å** drejer du til til v**e**nstre ad Stati**o**nsvej til du k**o**mmer til det n**æ**ste kr**y**ds og ved kr**y**dset ved det ved d**e**t kr**y**ds d**e**r går du **o**p ad Dronning D**a**gmars All**e** og **o**ppe for **e**nden af Dronning D**a**gmars All**e** til h**ø**jre der l**i**gger **a**lderdomshjemmet*

'you drive down to or you walks down to the crossing at the end of Vestergade and then you turn to to the left along Stationsvej until you reach to the next crossing and at the crossing at that crossing there you walk up along Dronning Dagmars Alle and at the end of Dronning Dagmars Alle to the right there is the retirement home'

Reinholt Petersen [2] did not examine sentences of more than 12 stressed syllables. Maybe the low correlation coefficients obtained in the present investigation are due to the long texts, which vary from 27 to 108 stressed syllables. Longer text must be inherently more difficult to model than shorter ones.

Figure 3 shows that there is correlation between text length and the obtained correlation coefficients, R^2=20.8%. But if the outliers (marked with arrows) are ignored, only 8.0% of *the variation* of the obtained correlation coefficients is due to text length. It is not possible to interpolate *from* the correlation between text length and correlation coefficients *to* the generally low correlation coefficients obtained. The same correlation would be obtained if all the results were equally higher or lower. The correlation between text length and correlation coefficient – although it is quite low – only tells us, that longer texts may be more difficult to model, and perhaps an intermediate utterance level ought to be introduced.

ISBN 1-876346-nn-n

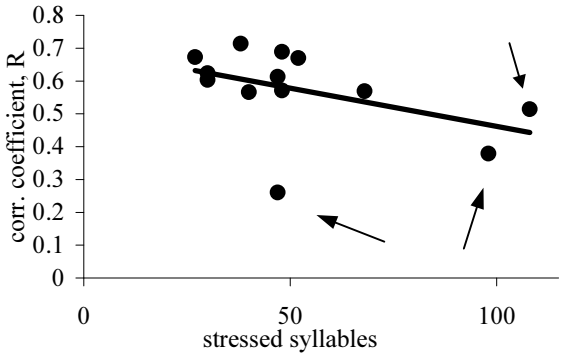

Figure 3. Correlation coefficients, R (obtained by multiple polynomial regression analysis of 13 texts) against the length of the text. The line shows the regression line associated with the results.

4. CONCLUSION

Even though the results of the investigation do not invalidate NG's model of Danish intonation, they call attention to some topics that should be considered when further research on intonation in Danish spontaneous speech is carried out. Firstly, it is highly important that the nature of prosodic boundaries be explored if we are to locate such boundaries confidently. Secondly, we need to take into account varying degrees of prominence and its effect on the F_0 pattern and the prosodic phrase contour. Although Grønnum has investigated this matter (cf. e.g. [6]), there is more to say. From the data analyzed in the present investigation it appears, e.g., that prominence can also entail a large F_0 excursion on the post-tonic syllables of a stress group. Thirdly, the data also show that intonation contour slopes can be positive, a feature not foreseen in NG's model. The obvious final caveat is that not enough is known about the interplay between intonation and other parts of the speech signal.

REFERENCES

[1] N. Grønnum, "Superposition and subordination in intonation – a non-linear approach," in *Proceedings of the XIIIth International Congress of Phonetic Sciences 1995*, vol. II, pp. 124–131, Stockholm 1995.

[2] N. Reinholt Petersen, "Modelling Danish sentence and phrase intonation," in *Proceedings of the XIVth International Congress of Phonetic Sciences 1999*, vol. II, pp. 925-928, San Francisco 1999.

[3] N. Reinholt Petersen, "Modelling fundamental frequency in first post-tonic syllables," in *Proceedings of Eurospeech 2001*, vol. II, pp. 939-942.

[4] N. Grønnum, *The groundworks of Danish intonaiton. An introduction*. Museum Tusculanum Press, Copenhagen 1992.

[5] M. Swerts, E. Strangert and M. Heldner, "F_0 declination in read-aloud and spontaneous speech," in *Proceedings of the International Conference on Spoken Language Processing, Philadelphia,* vol. 3, p. 1501-1504, Philadelphia, 1996.

[6] N. Thorsen, "Neutral stress, emphatic stress, and sentence intonation in advanced standard Copenhagen Danish," in *Annual Report of the Institute of Phonetics University of Copenhagen*, pp. 121-205, Copenhagen 1980.