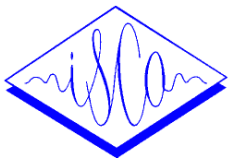

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

INTONATION CONTOURS AND THE PROMINENCE OF F0 PEAKS

Carlos Gussenhoven and *Toni Rietveld*

Department of English and Department of Language and Speech
University of Nijmegen, NL-6525 HT Nijmegen, the Netherlands

ABSTRACT

This paper reports the results of an experiment that sought to establish if the parameters in an existing F0 implementation model uniquely correspond to perceptual dimensions like 'emphasis' and 'liveliness'. These results are negative. A discussion of the rationale behind different kinds of perceptual experiments into perceived prominence leads to the conclusion that such experiments may reveal the effects of contour-internal relations like downstep and declination, or may more generally reveal the existence of various aspects of canonical contour shapes, but will not yield a calculus for determining the perceived prominence of pitch accents on the basis of anchor points in the surrounding unaccented parts of the contour.

I. INTRODUCTION

The general aim of experiments that seek to establish the perceived prominence of intonational peaks is the development of a theory of intonational scaling. As is to be expected, peaks that represent realisations of the same phonological pitch accent will vary across speakers with different mean speaking ranges, as well as across different locations in the contour. A successful theory would thus predict the F0 of peaks with equal perceived prominence occurring in different locations in a given contour, or in the same location in a given contour spoken in two different speaker-specific registers. Equally, it would predict what contiguous F0 intervals, for a given peak, would represent equidistant steps in perceived prominence.

The research questions can thus be divided into three types:

(i) those that try to establish relations between two peaks in the same contour, either relations that result from intonational processes like downstep, declination and final lowering, or relations that have to do with the degree to which the peaks belong together in a single contour;

(ii) those that try to establish relations between peaks in different contours, by finding some way of calculating the prominence of a peak from the unaccented minima around it;

(iii) those that try to establish what physical unit best expresses the perceptually equidistant steps between different peaks.

II. GENERAL RESULTS

2.1 Relations between peaks in the same contour

If we ignore the third of the above research questions (e.g. [1]), it is fair to say that, so far, coherent results have only been obtained in experiments dealing question (i). First, Pierrehumbert established that peaks later in the contour needed to have lower F0 for English listners than earlier peaks, which result was explained as the listener's adaptation to the fundamental frequency declination she expects the speaker to produce [2]. Second, Ladd et al. found that in contours with two peaks, P2, the later peak, is perceived as less prominent if the F0 of P1 is lower, provided the general F0 range of the two peaks is relatively low [3]. For higher values of F0, the effect reverses: lowering the F0 of P1 causes the perceived prominence of P2 to increase. Ladd et al. explain the first effect, a replication of an earlier fortuitous research result with Dutch listeners [4], as the effect of a prominence setting for the whole contour, as assumed by the listener. Since a lower P1 will induce a lower prominence perception for the contour encompassing both P1 and P2, P2 will likewise be perceived as less prominent. The reversal of the effect at higher values of F0 is explained by the abandonment of this assumption of contour integrity. Rather, the peaks are now taken to be separately emphasised, and a lower F0 for P1 will now have the effect of throwing P2 into relief, thus increasing the perceived prominence. Third, Gussenhoven & Rietveld argue that their research results show that Dutch listeners not only make allowance for declination when assessing the prominence of a final peak, but also for final lowering (Liberman & Pierrehumbert [5]), if the assumption is made that declination is linear across the contour [4].

2.2 Interpreting the perceived prominence of peaks on the basis of some model using anchor points in the pitch contour

In contrast to experiments investigating contour-internal dependencies between F0 peaks, experiments that have sought to predict perceived prominence of F0 peaks on the basis of other features of the contour have not led to easily interpretable results. It would at first sight seem reasonable to suppose, for instance, that the prominence of a peak can be predicted on the basis of the F0 of surrounding unaccented parts of the contour. In general, the expectation has been that the distance between the peak and some reference line will determine the perceived prominence of the peak. Ladd draws a distinction between theories that assume an overt baseline, i.e. one that can be drawn through observable F0 values, as in the IPO tradition, and theories that assume an abstract reference line, i.e. one given by some model of F0 scaling [6]. He shows that the assumption of an overt baseline, determined by the actual baseline on which the peaks are superimposed in synthetic stimuli, fails to account for the perceived prominence differences for the peaks Terken used in his stimuli [7]. He does this by showing that the distance between the peak and the overt baseline produces a poorer fit of the prominence data than the distance between the peaks and an abstract reference line, in this case an arbitrarily chosen line which runs through the F0 minimum before the second peak and which declines at the standard rate as used in the IPO speech synthesis model.

While the overt baseline would thus not appear to be used in any evident way to calculate peak prominences, Repp et al. do find an effect of the overt declining baseline on the perception of the prominence of P2 in a two-peak contour. Their results are not in accordance with the view that the distance between the unaccented low point before the peak determines the prominence of the peak [8]. Neither do they tally with Terken's earlier results, which showed, for stimuli in which the F0 of the peak covaried with the slope of the baseline, that the 'declination effect' found by Pierrehumbert is larger when an overt declining baseline is present [7]. This is because the results obtained by Repp at al. show that the 'declination effect' decreases with increasing steepness of the overt declining baseline. And inasmuch as there is an effect of the overt baseline at all, their results go against the hypothesis that an abstract reference line is strictly computed on the basis of peak values alone.

Since the final low point of falling intonation contours has been shown to be fairly invariant against changes in overall pitch [5], and since the perceptual salience of the utterance end is comparatively high, it seems promising to look for an anchor point in the F0 of the end of the utterance. However, Terken's suggestion that the final low pitch of the contour is used by the listener to determine a reference for the prominence of the peak(s) is refuted in our own unpublished results with two-peak contours [9]. While peak height and position (P1 or P2) were found to determine the perceived prominence in the expected fashion, there was no effect whatever of the F0 of the final low point.

2.3 Towards a different interpretation strategy?

The research briefly summarized in the preceding section has been conducted under a general working hypothesis to the effect that some combination of observable F0 values provide the listener, after a suitable transformation, with a prominence scale against which the peaks of the contour can be measured. It is of course conceivable that this assumption is incorrect. A more 'global' view of the problem would be that listeners apply a complex of criteria that are based on their experience with intonation contours. A voice quality that suggests anger or excitement might lead to greater perceived prominence, other things being equal, and voice quality differences suggesting different sexes or ages may well bias the listener towards greater of lesser prominence of F0 peaks, depending on the expected pitch range of the type of speaker that is perceived. Such a more global theory implies that the intonation contour is not sampled at specific critical locations, which values are subsequently used to model a prominence scale, but rather is evaluated more generally in terms of expected contour shapes. This would mean that every contour is evaluated by the listener in terms of the multitude of contours with which it associated, each of which will have its specific biassing effect on the listener. This explanatory strategy would force us to interpret the results of rating experiments of this type in terms of the reminiscences of actual contour types that the stimulus contours create in the rater's mind.

Let us take the results of Repp et al. as an example [8]. A steep overt declining baseline connecting two peaks may well suggest that the peaks belong to the same phrase, whereas a more level, low baseline may suggest they belong to different phrases, as suggested by the analysis in [10]. Listeners may associate sequences of phrases with 'phrasal downstep', while interpreting two highish peaks in the same phrase with the absence of accentual downstep. As a result, the perceived prominence of P2 in the virtual 'two-phrase' condition will be affected more by the compensatory strategy applied by the listener, causing it to sound more prominent than the P2 in the virtual 'one-phrase' condition.

This more global - or perhaps 'incidental' - interpretation of the results obtained in experiments of type (ii) above suggests that the perception of the speaker's sex may well cause the perceived prominence of the peaks to change. Thus, if we were to manipulate our stimuli by means of spectral transformations which create different speaker types, by changing a male into a virtual female voice or vice versa, while keeping the F0 the same, we would predict that the perceived prominence

would decrease when going from a (virtual) male to a female voice. This is because the listener will associate female voices with higher overall F0 than male voices, and a given prominence level therefore normally corresponds with a higher F0 peak in female speech than in male speech.

III. TESTING AN IMPLEMENTATION MODEL

3.1 Introduction

There is another approach to the problem of peak scaling that seems worth pursuing: we could test F0-scaling models that have been developed for the synthesis of pitch contours. In such models, we typically find parameters that are intended to model prominence differences among intonational contours. The model presented by Ladd is particularly interesting from this point of view because it includes several parameters that have the effect of boosting pitch excursions, N, for *Normalisation*, and W for *Width* [11]. A mathematical description is presented in [11,10]; a graphical representation is given in Figure 2. The accentual and phrasal downstep factors *da* and *dp* are not relevant to our topic.

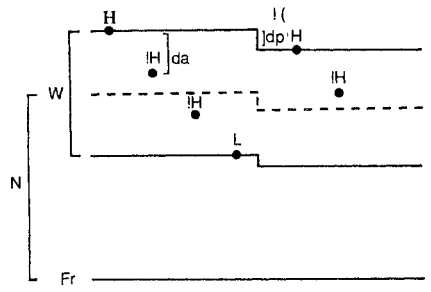

Figure 2. Ladd's implementation model for F0 targets [11] (with modification for phrasal and accentual downstep by van den Berg et al. [10])

The model defines a high and a low 'grid'-lines; increasing N will cause both the high and the low grid-lines to be raised by some multiplicative factor, which in absolute terms will have a much greater effect on the high grid-line than on the low grid-line. Increases in Fr (*Reference Frequency*) will do the same, but will cause increases or decreases of comparable magnitude for both grid-lines. Increasing W will raise the high and lower the low grid-line. The model, modified in [10] so as to be able to deal with separate factors for accentual and phrasal downstep, is used in the rule-based synthesis of pitch contours from autosegmental tonal representations. Without provision for the downstep factors it amounts to (1), which gives the value for the high grid-line. Replacing W with 1/W gives the low grid-line.

(1) $F0 = Fr * N^{W}$

The question arises if the different parameters Fr, N, and W uniquely correspond with perceptual dimensions, in the sense that N is related to the perception of liveliness, while W is related to the perception of emphasis, or vice versa. This question can be investigated by asking raters to judge a number of synthetic stimuli in which the same phonological representation is realised with a number of different settings for these parameters.

In Experiment I, we used two sentences as carriers which had similar segments in two accentable positions, and which had durations of 2245 ms and 2885 ms.

(2) a. Ik ga NOOIT meer met LEna naar school
'I will never go to school with Lena again'
b. Zo zou NOOIT meer naar LEningrad worden verwezen
'They'd never refer to Leningrad again,for instance'

The sentences were provided with two phonological contours each, a 'flat hat' (linked H*L H*L contour) and two 'pointed hats' (unlinked H*L H*L contour). For each sentence, we crossed these two contours with the vales for N and W given in (3). Fr was held constant at 50. For each sentence we thus produced 3 N's x 3 W's x 2 contours = 18 versions.

(3) N = 1.9, 2.1, 2.3
W = 1.4, 1.6, 1.8

We presented the stimuli in two random orders to 21 raters, whose task was to rate each stimulus on three 7-point semantic differentials. These were as given in (4). A additional scale 'High-pitched - Low-pitched' was used as a check on raters' behaviour.

(4) Natural - Not natural
Emphatic - Not emphatic
Agitated - Not agitated

In Experiment II we used a text of 17 secs, which was about environmental pollution. It consisted of three sentences, which were provided with phonological intonation contours judged to be natural by the experimenters. This time, we adjusted the N values downward relative to Experiment I, and added one step. Also, we used two values for Fr, corresponding with a deep and a highish male voice. The steps are given in (5).

(5) N = 1.5, 1.7, 1.9, 2.1
W = 1.4, 1.6, 1.8
Fr = 50, 70

In addition to the scales in (4), we had the text rated on a scale 'Confident-Not confident'. The 24 versions of the text were randomised, low-passed filtered to avoid undesirable artefacts of the allophone synthesis program used (3000 Hz), and presented to 32 listeners.

3.2 Results

Experiment 1. For every scale a Reliability coefficient (α) was calculated. These turned out to be satisfactory (Natural 0.77) to very satisfactory (Emphasis .95; Agitated 0.96). The data were pooled across sentence and contour. The results show that an increase of N leads to increases in perceived emphasis as well as in perceived liveliness, and that the same can be said about increases in W. These model parameters do not, therefore, seem to correspond to separate perceptual dimensions. This is also apparent from the high correlation between the scores on these two scales: .89.

Table 1. Explained variance (η^2's) of scores on scales 'Emphasis' and 'Liveliness', if significant at the 5% level. Single-sentence stimuli.

	Emphasis	Liveliness
N	.85	.91
W	.90	.85

Experiment II. Instead of single-sentence stimuli, in Experiment II we used different versions of a text, which raters were asked to rate on four scales. Reliability coefficients varied from reasonable ('Emphasis'.54) to satisfactory ('Confident' .81, 'Natural' .83) and very satisfactory (Lively .93). Analyses of variance were performed on the scores for each scale, with Fr (2 levels), N (4 levels), W (3 levels) and Subjects (32 levels) as factors (repeated measurements). P-values for factors of more than two levels were Huynh-Feldt corrected.

Table 2. Explained variance (η^2's) of scores on scales 'Emphasis' and 'Liveliness', if significant at the 5% level. Text stimuli.

	Emphasis	Liveliness
Fr	-	.59
N	.16	.53
W	.13	.15
Fr x N	-	.14

We again find that both N and W affect both perceptual dimensions: increases in either of these model parameters increase the perceived emphasis as well as the perceived liveliness. However, the effect sizes are very much smaller than for the single-sentence stimuli. In particular 'Emphasis' has only a small effect. The latter result becomes interpretable when we realise that emphasis, or prominence, is a more locally defined percept than liveliness. Prominence is typically assigned to an accent or a phrase, whereas liveliness is an attribute of larger speech segments, like paragraphs, and can even be a characteristic of a someone's speech in general.

Raising the Fr from 50 Hz to 70 Hz has an effect on 'Liveliness', but not on 'Emphasis'. Even more so than N, Fr determines the degree of raising and lowering of the pitch contour as a whole, without greatly altering the excursion sizes. While W can be used to create greater relief in the pitch contour, of the sort we may associate with both emphasis and liveliness, carrying out the same intonation contour in a higher register does not make the speaker sound more emphatic.

The inclusion of the scale 'Natural' (both experiments) enables us to say that, for Fr=50, an N of 1.9 and a W of 1.4 or 1.6, or an N of 1.7 a W of 1.8, give the most natural pitch range for Dutch. This corresponds with a grid with F0 values of around 70 Hz and 130 Hz, or approx. 10 semitones. This is also the general contour range which makes the speaker sound most confident (Experiment 2).

CONCLUSION

The Width and Range parameters in Ladd's model for F0 implementation do not uniquely correspond to perceptual dimensions like emphasis and liveliness. Evaluation of the results of experiments measuring perceived prominence of F0 peaks suggests that such experiments can be used to track down aspects of canonical contours of the language, but will not yield a model of prominence perception that takes the F0 of presumed critical locations as input.

References

[1] D. Hermes and J. van Gestel (1991) "The frequency scale of speech intonation", JASA 90, pp. 97-102.

[2] J.B. Pierrehumbert (1989) "The perception of fundamental frequency declination", JASA 66, pp. 363-369.

[3] D.R. Ladd, J. Verhoeven and K. Jacobs (1994) "Influence of adjacent pitch accents on each other's perceived prominence: Two contradictory effects". Journal of Phonetics 22, pp. 87-99.

[4] C. Gussenhoven and T. Rietveld (1988) "Fundamental frequency declination in Dutch: Testing three hypotheses", Journal of Phonetics 16, pp. 355-369.

[5] M.Y. Liberman and J.B. Pierrehumbert (1984) "Intonational invariance under changes in pitch range and length". In M. Aronoff and R.T. Oehrle (eds.) Language and Sound Structure: Studies in Phonology Presented to Morris Halle. Cambridge MA: MIT Press. pp. 157-233.

[6] D.R. Ladd (1993) "On the theoretical status of 'The Baseline' in modelling intonation", Language and Speech 36, pp. 435-451.

[7] J.M.B. Terken (1991) "Fundamental frequency and perceived prominence of accented syllables". JASA 89, pp. 1768-1776.

[8] B.H. Repp, H.H. Rump and J.M.B. Terken, "Relative perceptual prominence of fundamental freuqency peaks in the presence of declination", IPO Annual Progress Report 28, pp.59-62.

[9] J.M.B. Terken (1993) "Baselines Revisited: Reply to Ladd", Language and Speech 36, pp. 453-459.

[10] R. van den Berg, C. Gussenhoven and T. Rietveld (1992). "Downstep in Dutch: Implications for a model". In G.J. Docherty and D.R Ladd (eds.) Papers in Laboratory Phonology II: Gesture, Segment, Prosody, pp. 335-358. Cambridge University Press.

[11] D.R. Ladd (1991) "Metrical representation of pitch register". In Papers in Laboratory Phonology I (J. Kingston and M.E. Beckman eds.), pp. 35-57. Cambridge University Press.