Proceedings of the Institute of Acoustics

SOME PHONETIC CORRELATES OF STYLISATION IN THE STEP-DOWN CONTOUR

Mike Johnson and Martine Grice

University College London

ABSTRACT

In this paper a particular type of stylised intonation contour is examined: a stepping down sequence of two pitch plateaux. This is often referred to as the calling contour, as it may, amongst other things, be used to attract a person's attention.

It has been suggested by previous commentators that the minor third predominates as the pitch interval between the two plateaux in such contours. At the same time, the favoured end pitch of the second plateau is hypothesised to be in mid-range.

This paper presents the results of a pilot perceptual study designed to investigate the favoured interval, preferred end pitch and preferred starting pitch in stylised step-down contours. Both intonationally experienced and naive subjects were presented with step-down contours of varying pitch separation and asked to rate them on a seven point scale. Results indicated that intervals in the range minor third to fourth (3-5 semitones) were preferred over other ones, with a non-significant preference for the major third being displayed. In addition, for a selected subset of the subjects, there was an increasing preference for the contour as it shifted up in the pitch range.

1. INTRODUCTION

1.1 Stylisation

The notion of stylisation was introduced by Ladd [1], [2], to account for intonation contours which comprise one or more pitch plateaux, separated in pitch from their neighbours. Where stylisation is understood to mean 'simplification of a phenomenon to its bare essentials', a direct iconic link can be observed between the form and function of these contours. Their stylised form, involving pitch jumps and plateaux as opposed to glides, is said to express a stylised function: that of routine or stereotype, where no real information transfer takes place.

1.2 The step-down contour

The contour examined in this paper involves a stepping down sequence, such as is found in the following examples (notation as in Bolinger, [3] and elsewhere):

Ro bert Daddy fell down sta irs

Many analysts (for example Liberman [4], Leben [5], Gibbon [6]) have noted that this step-down contour is a form appropriate to calling out to someone at a (sometimes metaphorical) distance. The contour is thus often referred to as the 'calling contour'. The function implied here of securing uptake or establishing contact is supplemented, in certain circumstances, by the connotations of routineness, boredom or ritual.

2. THREE HYPOTHESES REGARDING THE STEP-DOWN CONTOUR

There are a number of features which characterise the step-down contour. Here we concentrate on three hypotheses regarding its location on the pitch axis.

2.1 Hypothesis 1: There is a standard pitch interval between the plateaux

It has been observed that the step down contour often carries a pitch interval of a minor third ([4], [6], [2]). Gibbon [4] and Ladd [2]both point out, however, that the interval can be more or less than the minor third. This may depend on the degree of chant, where chant can be viewed as a gradient function, manifested at its extreme by rigorous isochrony, prolongation of syllables bearing the pitch plateaux, and distinct voice quality, in addition to a fixed pitch interval.

2.2 Hypothesis 2: The pitch of the second plateau is in mid-range

Most commentators treat the step-down contour as an arrested fall or a raised fall-rise with a reduced rise. In either case, the pitch of the second plateau would be in mid range.

2.3 Hypothesis 3: The pitch of the first plateau is high in the range

Although the step-down contour is downward moving, there is no a priori reason why it should not start in mid range and end low. It may need to start high because of its (possibly conventionalised) calling function.

2.4 Relationship between the factors in the respective hypotheses

It would appear that the factors in these hypotheses are not independent. For instance, if the pitch of the first plateau is high in the range and that of the second is in mid-range, then the pitch interval between them might seem to be fixed. By similar considerations, each factor could be seen to depend on the other two. However, this observation depends on the requirement that 'high in the range' and 'in mid range' are discrete pitch levels, as in classical American intonation analyses. This is another thing to be learned from experimentation with the step-down contour. In any case, regardless of dependency between factors, it could be the case that one or two are more characteristic of step-down contours than the other(s).

In this paper, which forms the starting point for a series of experiments, the factors in the hypotheses are considered to be components in a model of intonation perception; they may also correspond to factors in a phonological model of intonation but may, to some extent, be determined by productive physiological factors.

2.5 The position of phonological models in respect of the three hypotheses

Hypothesis 1 could receive support in two cases:
A) The step-down contour is analysed as comprising two tones with a fixed pitch interval in an operative 'register' (cf. Ladd 1987). Step-down contours of fixed pitch separation would vary in height within the speaker's overall speaking range as the register shifted up and down within that range.
B) If a musical system of tones were used (such as the well-tempered scale of 12 pitch levels), then a fixed pitch interval within that system could be used anywhere within a speaker's overall range. In this case, the interval would not be calculated as a fraction of the speaker's range, but be independent of it.

Hypothesis 2 appears to be favoured by current phonological treatments. In the phonological systems described by Ladd [2] and [7], Gussenhoven [8] and Pierrehumbert [9], the hypothesis which receives the most support is the second: that the pitch of the second plateau in the step-

down is in mid range. Ladd and Pierrehumbert do not, however, refer directly to a mid level. Rather, they claim that the second plateau is above the bottom of the range: Gussenhoven refers to the level of the second plateau as 'lowered' mid, a reduced form of an L tone, the product of a process called 'fusion'; Pierrehumbert represents it as a downstepped high tone followed by a raised low boundary tone; Ladd [7] describes it as a low tone modified by the feature [-floor]. We have argued [10] that a more satisfactory account might be obtained by allowing stylised contours to occur within an augmented tonal system where mid is no longer derived, but is an integral part of a three or four tone system. These levels would still necessarily be relative to a given range.

Hypothesis 3 would seem to receive no support from any phonological model independent of the fact that step-down contours are represented as having an H tone (or raised mid, in the case of Gussenhoven) on their first element.

2.6 **Productive physiological considerations in respect of the three hypotheses**
It is not clear that there is strong physiological support for hypothesis 1. The only constraint on pitch interval might be that it should not be too large: all other things being equal, it is likely to be easier to achieve monotone after a reasonably small jump in pitch than after a large one which, again, might lead to unavoidable onglides. In this case there would be no physiological argument in favour of an interval of any fixed size.

Hypothesis 2 appears to have some productive physiological support. Recent work by the authors [10] has led them to suggest that stylisation should be seen to occur only in strictly monotone sequences of pitch where no emotional colouring such as suppressed pique or contradiction vitiates the quality of the voice[1] . In this context, it is hypothesised that pitch plateaux in the mid range are approached with fewer problems of laryngeal adjustment; that, all other things being equal, a monotone sequence is more easily produced if it is in the middle of a speaker's range than if it is high or low, where rising and falling onglides are often present.

Hypothesis 3 does not receive any direct support in the physiological domain.

3. THE EXPERIMENT

The experiment was designed to investigate the effect of changing step size and placement in speaker pitch range of the step-down contour. To encourage listeners to use a specific reference pitch range, it was decided that step-down stimuli should be presented with a prior aural context in the form of a dialogue in which the speaker whose voice was used in the synthesis was the predominant participator. The dependent variable in the experiment was chosen to be listeners' evaluation of the appropriateness of the step-down contour in that context,because questions regarding the formal naturalness of the stimuli thus became indirect, reducing the effect of judgements of other independent factors, such as quality of resynthesis, gradient of pitch glide through the intervocalic consonant, etc.

---

[1] Boredom is not treated here as a voice quality modifier; it is generally suppression of emotion which affects voice quality in level stretches of pitch. Examples of high level contours of the type: '$_{I}$ wouldn't be so sure of that!' (suppressed contradiction), 'You $_{bastard!}$' (suppressed pique) and '$_{Ro}$ bert!' (suppressed admonition).

SOME PHONETIC CORRELATES OF STYLISATION

3.1 Subjects

23 subjects, 12 female and 11 male, both intonationally experienced and naive, and with no reported hearing loss, were recruited from staff and students at UCL. Subjects were paid no fee for their participation in the test.

3.2 Choice of context and stimuli

The aim of finding a semantically neutral token in a context for which securing of uptake or routine are possible interpretations led to the choice of a name which was presented as the final part of a dialogue, the full text of which is as follows:

M: OK, we've covered syllabus - Irma's going to sort that out; Irma's going to look into the operational ramifications of the new education bill. And then there's the question of a statement of policy in the prospectus; Irma's going to ask John about that. Er...what else?

F: Well, I think that we have to make sure that we attract more postgrads.

M: Erm, yes, well um, they'll need a place to work. We really need to check how many rooms we have available. Oh dear! That's Irma's domain too - who it seems is busy scribbling notes on all the other things she has to do ... Irma

The name 'Irma' was chosen in order to reduce segmental effects on F0, it having central vowels and a nasal consonant. In the calling context, in a non-rhotic variety of British English, it is pronounced ['ɜ:mɜ:].

3.3 Method

3.3.1 Preparation of stimuli Seventy eight test stimuli were prepared using as the base token a natural monotone rendering of the name 'Irma' with a fundamental frequency of 137 Hz. The base token was sampled at 12.8 KHz on a Masscomp 6000 series 32-bit minicomputer and downsampled to 10KHz; synthetic stimuli were then generated by LPC analysis[1] (using the ILS

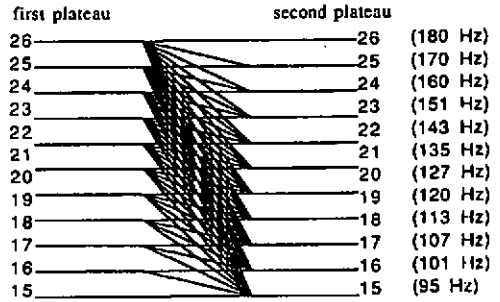


Figure 1. Schema of F0 contours synthesised. Values are in semitones relative to 40 Hz.

[1] It is expected that future experiments will use the PSOLA method of analysis and resynthesis (see [12]) in order to enhance the naturalness of the synthesis throughout the range and to make a closer match between context and stimuli. Classical LPC was chosen in this case because of its resynthesis schema using pulse excitation - consistent across the pitch range, though rather synthetic-sounding.

SOME PHONETIC CORRELATES OF STYLISATION

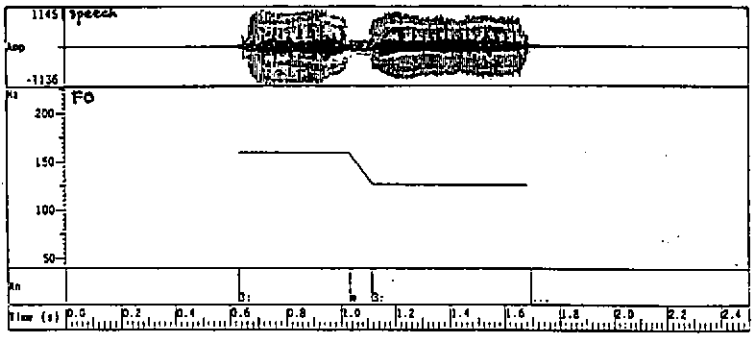


Figure 2. An example step-down token: from 24 to 20 semitones (relative to 40 Hz).

API analysis program) and resynthesis (using the ILS SNS synthesis program). The F0 contours used in resynthesis comprised the non-rising permutations of 12 pitch levels according to the schema in figure 1, where the plateau values are in semitones relative to 40Hz. The plateaux extended for the whole duration of each vowel. The pitch through the nasal was a linear interpolation between the plateau values. Contours were either level or falling, F0 intervals ranging from 0 to 11 semitones. The range of the stimuli is equivalent to that of the speaker, calculated over the dialogue passage forming the context of the test. An example token (a step-down of 24 to 20 semitones) appears in figure 2.

3.3.2 **Presentation of stimuli** The dialogue context was presented to the subjects aurally via an audio-cassette tape-recorder, and on a printed sheet. In addition, a summary of the context was displayed on the computer screen throughout the test. Speech tokens forming the stimulus set were stored on the Masscomp and were replayed via a D to A converter and a 4.5 KHz low pass filter. After listening to 5 sample tokens, subjects were presented with 3 blocks of 78 randomised tokens. They listened to the dialogue context before each block. Both context and stimuli were presented in free field conditions. The duration of the test was circa 30 minutes.

3.4 **Collection, storage and analysis of results**
Results were entered by subjects at the keyboard of a standard BBC Micro acting as a front end to the Masscomp computer. They used arrow keys in order to select one of seven numbered boxes in a screen display of the following format:

```
1     2     3     4     5     6     7
|-----|-----|-----|-----|-----|-----|
entirely                          entirely
inappropriate <-------------------------------------------> appropriate
```

Tokens were accessed after each response, thus enabling the test to be carried out at the subject's own pace. Ratings were decremented by 1, so that the range of rating values in the results is 0-6. Result files were then sorted and statistical analysis was performed using the SAS package.

After being tested, subjects were asked for their general impressions of the test and for the specific strategy they may have adopted in their assessment of the appropriateness of the tokens.

4. RESULTS AND DISCUSSION

This pilot test was designed to give some indications about which range of values for each of the factors was appropriate for study in later experiments, and which subjects could be satisfactorily used in those experiments. For this task-simplification exercise, it was not considered necessary to construct a balanced experiment. Indeed, because all non-rising permutations of step-size in the defined pitch range required inclusion, a balanced test was not possible. Within the stimulus set, smaller steps are prevalent. Similarly, there are more contours ending low than high. The restricted statistical analysis performed reflects these facts.

By inspecting the plots of individual subjects and taking into account information gleaned from subjects immediately after the test, it was clear that a number of subjects had interpreted the task in a way which had not been foreseen in the design. Certain subjects expected to hear a rising pitch in the given context and therefore rated level stimuli highly. Other subjects expected the speaker to be irritated by Irma's lack of attention and therefore preferred steps of eleven semitones (which can be used to mark irritation, although typically when accompanied by modification of voice quality[1]). A number of inexperienced subjects placed all of the tokens in the lower half of the rating range and stated that voice quality considerations influenced their decisions. All these subjects were treated as a separate group, and the results from the remaining 11 subjects (henceforth referred to as the select group) were pooled for the purposes of studying the variation of the data.

4.2 Variation of rating against the three hypothesis factors

Figure 3 displays plots of mean rating against step size, pitch of first plateau (P1) and pitch of second plateau (P2) for the select group. For this group, it appears that step-size is the most significant factor in the variation of the data - steps of between 3 and 5 semitones appear to group separately from other steps; in addition, however, the rating function of P2 shows a clear bias towards the middle of the range (20 and 21 semitones are the preferred values). Also, upper pitch values are preferred for P1. Prima facie, there is some corroboration for all three hypotheses.

4.3 Statistical Analysis

The large size of the imbalance in the data means that it is not possible to perform rigorous statistical analysis on the three hypothesis factors. If only balanced variables are considered, the pilot is a two-factor mixed model experiment, with fixed effect **token** and random effect **subject**. A two-way ANOVA was therefore performed on the data, in order to determine that the factor **token** is significant. Also, a Tukey HSD test was performed on the same factor, for comparison of mean ratings of tokens.The factor **token** was found to be a highly significant source of variation in the data ($F(77,1694) = 25.89$; $p < 0.0001$). The Tukey HSD test comparing mean ratings showed that the overall preferred token was 26-22 (mean rating = 4.536, N=69), a step down of four semitones from the top of the range to upper-mid range. Furthermore, the ratings for the eighteen most preferred tokens were not significantly different from each other, and all contained step sizes of between 3 and 5 semitones.

[1] The proposed use of PSOLA is expected to reduce these problems; enhanced naturalness is expected to lead to more consistent ratings in respect of voice quality.

The same analysis performed on the select group of eleven subjects revealed a similar result: (token: $F(77,770) = 29.71$; $p<0.0001$. Tukey HSD: Preferred token = 26-22; mean rating = 5.212, N=33; top 18 tokens not significantly different from each other, all of between 3 and 5 semitone intervals.)

4.4 Discussion

For the select group of subjects, it seems very likely that, on the whole, **step-size** is a more significant effect in the experiment than either **P1** or **P2**. Of the latter two, it appears that **P2** might be slightly more significant than **P1**.

The covariation between hypothesis factors is summarised in the three 3-D plots in Fig. 5, which show the rating profile of covariation between the three hypothesis factors. These plots show that the favoured step size is 4 semitones, and that there is a strong effect of increasing pitch height of **P1** and **P2** on Appropriateness Rating. Thus, there may be evidence for a fixed pitch interval, supporting Hypothesis 1 above, and some support for hypotheses 2 and 3.

5. CONCLUSION

This pilot experiment suggests that the favoured perceptual form of a step-down contour is one of a major third starting high in the speaker's speaking range. The results suggest that all three hypotheses have some support, but that the factor of fixed step size is most fundamental (although the minor third is less favoured than the major). For a near-octave range, the preference for a major third starting at the top of the range suggests a division of the range into quartiles for stylisation purposes. The pilot study indicates that a balanced test restricted to step sizes ranging between 2 and 6, and to the upper half of the speaking range of a number of speakers, would be the most profitable next step.

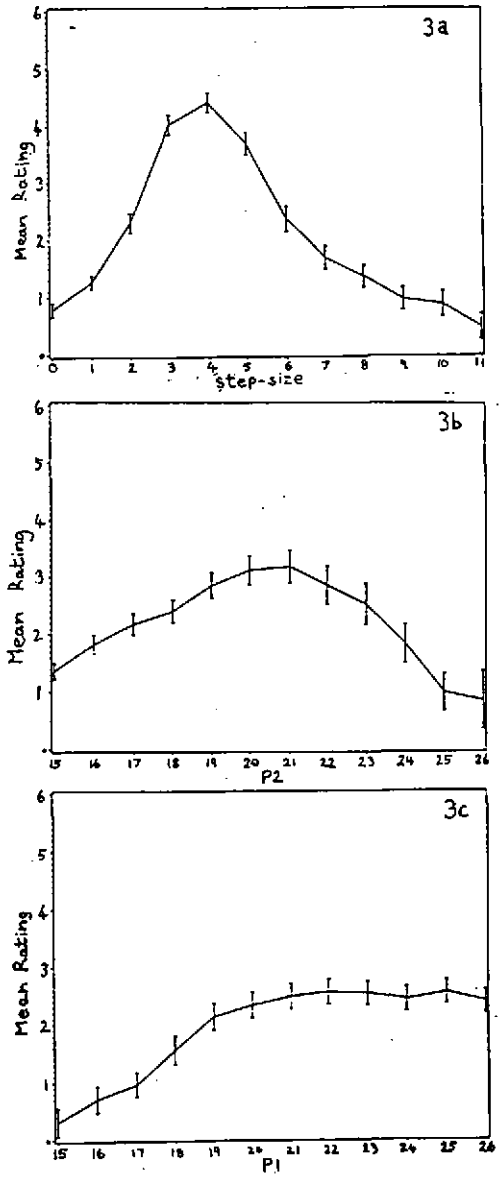
6. REFERENCES

[1] **Ladd**, D. Robert (1978) Stylised intonation, Language 54, 517-41.
[2] **Ladd**, D. Robert (1980) The structure of intonational meaning: Evidence from English, Bloomington: IUP.
[3] **Bolinger**, Dwight (1986) Intonation and its parts, Edward Arnold.
[4] **Liberman**, Mark (1975) "The intonational system of English", MIT dissertation [New york:Garland (1980)].
[5] **Leben**, William (1976) The tones in English intonation, Linguistic Analysis 2, 69-107.
[6] **Gibbon**, Dafydd (1976) Perspectives of intonation analysis, Bern: Lang.
[7] **Ladd**, D. Robert (1983) Phonological features of intonational peaks, Language 59, 721-759.
[8] **Gussenhoven**, Carlos (1983) A three dimensional scaling of nine English tones, J Semantics 2,183-203
[9] **Pierrehumbert**, J. (1980) "The phonology and phonetics of English intonation", MIT dissertation.
[10] **Johnson**, M and M **Grice** (1990) The phonological status of stylised intonation contours, Speech Hearing and Language: Work in progress, UCL, Vol 4, pp 227-256.
[11] **Ladd**, D Robert, (1977) A model of intonational phonology for use in speech synthesis by rule, In Laver and Jack: Proc. European Conference on Speech Technology. Edinburgh.
[12] **Hamon**, C, E **Moulines** and F **Charpentier** (1989) A diphone synthesis system based on time-domain modifications of speech, Proc. Int. Conf. ASSP, Glasgow.

Acknowledgements

Thanks are due to Andy Faulkner and Bill Barry, and mistakes are due to Johnson and Grice. This work was funded by the MRC (Johnson) and Esprit SAM Project no. 2589 (Grice).

SOME PHONETIC CORRELATES OF STYLISATION



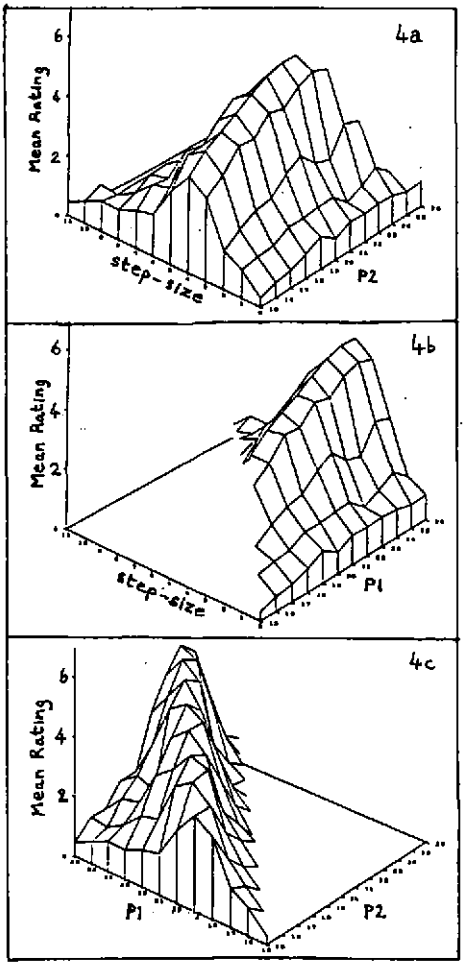


Fig. 3 Mean appropriateness ratings against a) step-size, b) P2 and c) P1 for the select group with 95% confidence limit bars.

Fig. 4 Rating profiles of the three hypothesis factors (select group):
a) step-size against P2
b) step-size against P1
c) P1 against P2