MODELLING INTRA- AND INTER-SPEAKER PITCH RANGE VARIATION

D. Robert Ladd and Jacques Terken***
**University of Edinburgh, Scotland, **Institute for Perception Research, Eindhoven, The Netherlands*

ABSTRACT

This paper reports preliminary findings from a large-scale study of pitch-range variation within and across speakers. The purpose of the study as a whole is to understand better the ways in which pitch range may vary with different speaking modes and in relation to linguistic structure, within and across speakers. An important empirical goal of the study is the collection of a large amount of quantitative data. We report results of comparisons between normal speaking mode, raised voice and local emphasis, and discuss some preliminary conclusions.

INTRODUCTION

Beginning with Bruce [1], numerous instrumental studies (e.g. [2,3]) have pointed to the existence of relatively invariant pitch targets in intonation contours, normally local maxima or minima. This finding is the basis of autosegmental descriptions of intonation in terms of H(igh) and L(ow) tones, such as the influential description of English intonation proposed by Pierrehumbert [4]. Likewise, in descriptions of intonation based on pitch movements rather than pitch targets (e.g. [5]), it is assumed that the beginning and ending levels of movements can be accounted for by a quantitative model. We assume without further comment the validity of the idea that targets have a reality of some sort, and an important role to play in the quantitative modelling of intonation.

At the same time, however, it is clear that the phonological description of tonal targets is still controversial, in part because there is no general agreement about the nature of tonal targets even among proponents of autosegmental analyses. A significant source of disagreement in these debates is the treatment of pitch range variation. It is obvious that a given speaker's voice may be higher or lower than another's, and it is similarly obvious that individual speakers may raise or lower their voices for a variety of reasons. What is not well understood, often due to the lack of empirical data, are the relations between such global within- and across-speaker differences, and their relation to other, more linguistic sources of variation in the "scaling" of individual pitch targets.

For example, it is known that F0 is generally higher at the beginning of paragraphs or other large discourse chunks, and that the F0 on individual accented words can be locally raised to convey emphasis. Is the raising of F0 in these types of cases quantitatively identical to "raising the voice"? Or again, if one speaker has a low monotonous voice and another a high animated voice, are these voice types quantitatively comparable to the differences within a single speaker between speaking monotonously and speaking animatedly? For that matter, what is the relation between "low" and "monotonous" or "high" and "animated": what would be the quantitative characterisation of a "high monotonous voice"?

This paper reports preliminary findings from a large-scale study focusing on these and related questions about pitch range variation. The purpose of the study as a whole is to understand better (a) the ways in which pitch range may vary within speakers; (b) the extent to which target scaling is invariant both within and across speakers; (c) what sort of scale (linear, logarithmic, or other) is most appropriate for characterising any invariance involved. An important empirical goal of the study is the collection of a large amount of quantitative data. This paper reports conclusions based on the analysis of some of the data from speech materials read aloud. Further analysis is in progress and fuller reports will be published in due course.

METHOD

Speech Materials

Our basic approach was to measure F0 at specific pre-selected points (e.g. utterance-initial unstressed syllable, first accent peak, utterance-final low in statements, low accent valleys in questions, etc.), in multiple repetitions of utterances with comparable contours. By doing this we hoped to establish stable mean pitch values for certain putative targets, creating a kind of "map" of the relative pitch of these targets which could then be compared between speakers or between different pitch-range settings of the same speaker.

For the study as a whole, we designed several sets of sentences intended to elicit specific intonation patterns which we expected would have consistent and identifiable peaks and valleys at well-defined points. These included ordinary statements of varying lengths, short questions, statements with explicit contrasts (of the sort "X not Y"and "Not X but Y"), and a news bulletin containing 18 short paragraphs. The recording session also included a short section of spontaneous speech (a description of the speaker's route to work). Only a small portion of the data is discussed and analysed here.

The materials discussed here fall into three main groups. Group 1 sentences contain two noun phrases with a total of four accented words; there are two subtypes, one in which both noun phrases have two accented words ("2-2") and one in which the first has three and the second only one ("3-1"). For example (accented words are written in capitals):

2-2: *Je moet de MOOIE ROZEN in een GELE VAAS doen* (You should put the pretty roses in a yellow vase)
3-1: *Je moet de MOOIE GELE ROZEN in een VAAS doen* (You should put the pretty yellow roses in a vase)

Altogether there were 8 such pairs, and each sentence was read twice, for a total of 32 sentences in group 1. Targets to be studied in this group are initial pitch, peak and valley on auxiliary verb, the four accentual peaks, medial valley between the two noun phrases, and the final low.

Group 2 sentences were all of the form

We zouden wel eens naar [X] kunnen gaan (We really ought to be able to go to [X] sometime)

in which X was one of four place names. Each of the 4 versions was read 4 times, for a total of 16 utterances. We intended that these should be accented only on the place name, though in the event many speakers put a weak accent on the auxiliary *zouden* as well. Targets to be studied in this group are initial pitch, weak accent on auxiliary, valley immediately preceding accent, accent peak, and final low.

Group 3 sentences were all of the form

Ik zei niet [X], maar [Y] (I didn't say [X], but [Y])

where X and Y were similar-sounding words that might plausibly be confused in a real situation, e.g. *mannetjes/lammetjes* (`little men / little lambs'). There were 4 pairs of words, presented in both possible orders, with 2 repetitions of each sentence, for a total of 16 utterances. Targets to be studied in this group are initial pitch, valleys preceding accents, accent peaks, final low, and both valley and peak of medial continuation rise.

In constructing the sentence materials we balanced prosodic, pragmatic, and segmental phonetic considerations.In particular, we avoided words with high vowels (to minimise intrinsic F0 effects) and obstruents (to minimise segmental perturbations of F0). Further details are beyond the scope of this report.

Recording and analysis procedures

For the recording sessions, all the materials were organised into 8 blocks, and the session lasted typically 75 minutes with a short break in the middle. The sentences discussed here were included in three of the blocks: one (Block 2) in which speakers read normally without any special instructions, another (Block 4) in which the speakers were told to raise their voice as if talking on a bad overseas telephone connection

(the situation was made more realistic by exposing speakers to rather loud (over 90 dB) non-steady noise over headphones), and a third (Block 7) in which individual words were capitalised and the speakers were told to emphasise those words. Of the materials discussed here, Block 7 included only the group 3 sentences ('not X but Y').

Speakers were 16 native speakers of Standard Dutch, 8 males and 8 females, all students or employees at the Institute for Perception Research (IPO), Eindhoven. This report is based on results from only 8 speakers, 4 males and 4 females.

The recordings were made in a quiet recording studio at IPO, using professional equipment. The sentences to be read were presented one at a time on a computer screen placed on a table in front of the speaker. The experimenter controlled the presentation from a neighbouring control room.

The recordings (on DAT tape) were transferred to the computer system at IPO and separate speech files were made for each sentence. F0 extraction was done by means of an algorithm based on subharmonic summation ([6]) with tracking. F0 values for each target were determined on the basis of time-aligned displays of the waveform and the F0 trace, obtained by means of an interactive wave form processing package developed at IPO. Details of the measurement criteria are beyond the scope of this limited presentation.

The maximum number of utterances per speaker in the portion of the study reported here was as follows:

	Normal	Raised voice	Local_ Emph.
Group 1 2-2	16	16	--
Group 1 3-1	16	16	--
Group 2	16	16	--
Group 3	16	16	16

In many cases one or more utterances had to be discarded because of disfluencies, etc.

RESULTS AND DISCUSSION

Sample data (for Speaker RS's Group 3 and Speaker RW's Group 1 sentences) are shown in Figs. 1 and 2. For all the speakers whose data we have analysed so far, the contours for each sentence group, and the patterns of modification for the different conditions, are strikingly similar. Quantitative modelling of the similarities is still only at a very preliminary stage. However, several findings are common to most or all of the speakers and must presumably be incorporated into any quantitative model. These are summarised here:

(1) There is a clear distinction between overall raising and local emphasis. The former raises both peaks and valleys, whereas the latter affects only peaks. This is seen in Fig. 1. This could be incorporated into a quantitative model by distinguishing two aspects of what is often loosely called "pitch range", namely the overall *level* and the *width* of the space in which tonal targets (or tonal movements) are realised. In overall raising of the voice, it is primarily level that is affected. In local emphasis, level is unaffected, but the width of the tonal space is expanded.

(2) Whereas many earlier reports (e.g. [2]) suggest that final F0 low is very stable for individual speakers, it appears that overall rasing also slightly raises the speaker's final F0 low. This is seen in both Figs. 1 and 2.

(3) For all targets, the effect of raising overall pitch range is extremely constant. For all speakers the correlation between targets in normal range and corresponding targets in raised range is extremely high (on the order of r = .90).

(4) It appears that the most invariant characterisation of the F0 relations in our data is achieved using an ERB scale [7], which at typical speech F0 levels is intermediate between the linear Hz scale and a logarithmic scale. This is reflected in the fact that range modifications look most similar across speakers when expressed in ERB: on a Hz scale overall raising is generally greater for men than for women, while on a log scale it is generally greater for women than for men. On the ERB scale the amounts by which males and females raise their voices are most comparable. Further discussion of quantitative details is beyond the scope of this paper.

REFERENCES

[1] Bruce, G. (1977), *Swedish word accents in sentence perspective*, Lund: CWK Gleerup.

[2] Liberman, M. and Pierrehumbert, J. (1984), "Intonational invariance under changes in pitch range and length", In Aronoff, M. and Oehrle, R., editors, *Language Sound Structure*, Cambridge: MIT Press,.

[3] Van den Berg, R., Gussenhoven, C., and Rietveld, A. (1992), "Downstep in Dutch: implications for a model", In Docherty, G. and Ladd, D., editors, *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, pp. 335-359. Cambridge, U.K: Cambridge University Press,.

[4] Pierrehumbert, J.B. (1980), *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, Massachussets Institute of Technology, Cambridge MA (Bloomington, IN: Indiana University Linguistics Club).

[5] 't Hart, J., Collier, R. and Cohen, A. (1990), *A perceptual study of intonation*, Cambridge: Cambridge University Press.

[6] Hermes, D.J. (1988), "Measurement of pitch by subharmonic summation", *J. Acoust. Soc. Am.* Vol. 83, pp. 257-264.

[7] Hermes, D.J. and Van Gestel, J. (1991), "The frequency scale of speech intonation", *J. Acoust. Soc. Am.*, Vol. 90, pp. 97-102.

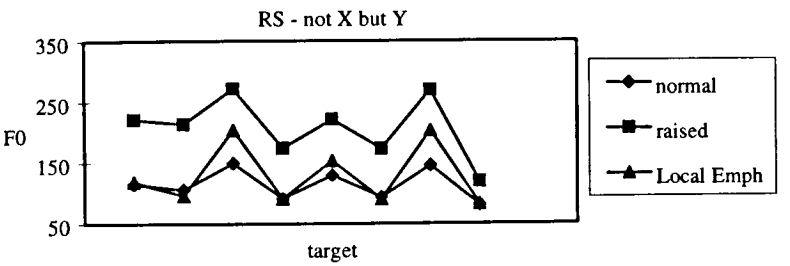

Figure 1. Average target values for successive targets (see text) in "not [X] but [Y]" utterances for speaker RS, in normal speaking mode, with raised voice, and with local emphasis on accented words.

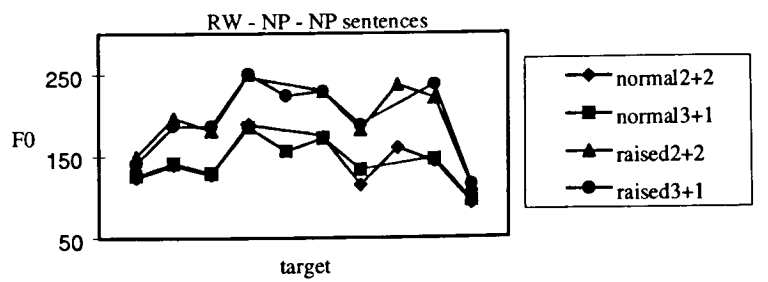

Figure 2. Average target values for successive targets in NP-NP sentences with 2+2 and 3+1 structure (see text) for speaker RW, in normal speaking mode and with raised voice.