Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Perceptual Judgments of Pitch Range

Rolf Carlson[1], Kjell Elenius[1] and Marc Swerts[2]*

[1]KTH, Sweden, and [2]Tilburg University,The Netherlands *Names in alphabetic order
rolf@speech.kth.se, kjell@speech.kth.se, m.g.j.swerts@uvt.nl

Abstract

This paper reports on a study that explores to what extent listeners are able to judge where a particular utterance fragment is located in a speaker's pitch range. The research consists of a perception study that makes use of 100 stimuli, selected from 50 different speakers whose speech was originally collected for a multi-speaker database of Swedish speech materials. The fragments are presented to subjects whom are asked to estimate whether the fragment is located in the lower or higher part of that speaker's range. Results reveal that listeners' judgments are dependent on the gender of the speaker, but that within a gender they tend to hear differences in range.

1. Introduction

One of the most controversial topics in intonation research has been the issue of pitch range. There is already confusion about the exact definition of the term, as it has been operationalized in quite different ways in the literature. Following Ladd [6], we assume that it covers two partially independent dimensions of pitch variation, i.e. level and span. The former refers to the "overall" key in which an utterance is produced, which can for instance be relatively high or low; the latter has to do with the tonal space a speaker exploits, in particular the distance between a speaker's upper and lower extreme in fundamental frequency (F0). Defined as such, it is clear that there exist inter-speaker differences in pitch range, both regarding level and span. For instance, a female speaker with a soprano voice will - on average - speak much higher than a male speaker with a bass voice. At the same time, some speakers may use a rather narrow span of frequencies, whereas for others the distance between lower and higher extremes is larger. Apart from such speaker-related differences, it is likely that –within speakers - variation in pitch range is dependent on factors such as speaking style or emotional content, or on the language that is spoken (see discussion of British English versus Dutch [5]).

While the notion of pitch range may be intuitively clear, it has been a highly debated research problem. One fundamental drawback is that there is as yet no consensus on how pitch range should be determined acoustically. For instance, measures vary between a rather sophisticated use of larger-scale declination lines (baselines and toplines), and a relatively simple annotation in the ToBI framework of the highest F0 value (HighF0) at the energy peak in an accented syllable. Presumably, the latter choice to look at HighF0 as a correlate of pitch range is motivated by the fact that (utterance-final) low targets are often claimed to be relatively stable, so that variation in pitch range can be modeled as fluctuation in the higher part of a speakers' frequency usage, whereas the low values can be seen as stable reference points (see however [6], [9]). Along the same lines, there is discussion as to what the best scale is to adequately represent pitch range (e.g. linear versus logarithmic). This choice of scale partly depends on the fact of whether or not the scale should reflect a listener's perception of range. From a listener's perspective, pitch range can be viewed as a frame of reference against which he or she "calibrates" the local pitch events of a speaker. Currently, there is no complete answer to the question as to whether listeners are indeed able to estimate a speaker's pitch range.

Our own rationale to start looking at pitch range phenomena stems from an earlier study which explored to what extent listeners are able to predict the occurrence of an upcoming break on the basis of prosodic properties of a particular utterance fragment [3]. It was found that listeners are indeed able to tell beforehand whether or not a break is coming up. Acoustic analyses revealed that listeners may have based these judgments partly on variation in pitch range: the estimated boundary strength of a break appeared to correlate highly with F0 values in the last 100 ms of the fragments (r=.62) (Figure 1). Since this perception experiment was based on stimuli coming from only one female speaker, it was not entirely clear from the test to what extent the listeners had been able to learn the pitch range properties of that speaker in the course of the experiment, or already knew them since the speaker is a famous Swedish politician. In particular, the exposure of many samples of that single speaker allowed listeners - in theory at least - to base their scores on within-speaker comparisons of pitch ranges in different stimuli.

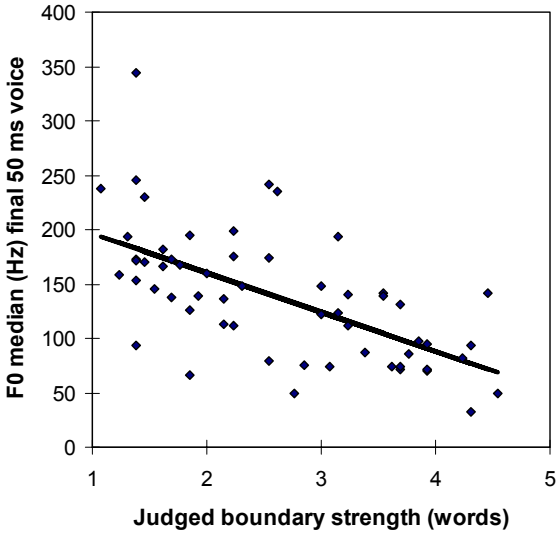

Figure 1: *Correlation between boundary strength and utterance-final F0 values (last 50 ms)*

This leaves us with the question to what extent listeners would be able to make more absolute judgments of pitch range. Therefore, we have conducted a project in which we use a larger database, consisting of utterances from many speakers in order to learn to what extent listeners are able to

10.21437/SpeechProsody.2004-159

tell apart high pitch range from low pitch range utterances, when they are not able to compare within speakers. The study differs from most previous work in that the experiment uses natural speech samples as stimulus materials rather than elicited or synthetic utterances with carefully controlled intonation contours (see also [8]).

2. Corpus study of F0 distribution

The research described here consists of a perception test which uses speech fragments from a multi-speaker database a stimuli. In order to get an estimate of how F0 varies for different speakers we made a corpus study of the F0 distribution in the Swedish SpeeCon database collected by KTH. SpeeCon focused on collecting linguistic data especially for speech recogniser training and testing. It was funded as a project under Human Language Technologies (HLT), part of the EU IST Programme. The fundamental frequency traces for 498 speakers were analyzed using the wavesurfer/ESPS Waves software. A cumulative distribution of the F0 measurments for each speaker was calculated based on 314 prompted utterances corresponding to about 30000 F0 observations (every 10 ms) per speaker (Figure 1). The median was chosen as a good speaker characteristic feature as has been pointed out by for example van Bezooijen [1] and also discussed by Rietveld [10]. The median, the 25 % and 75 % points in the distribution describe each speaker's typical range.

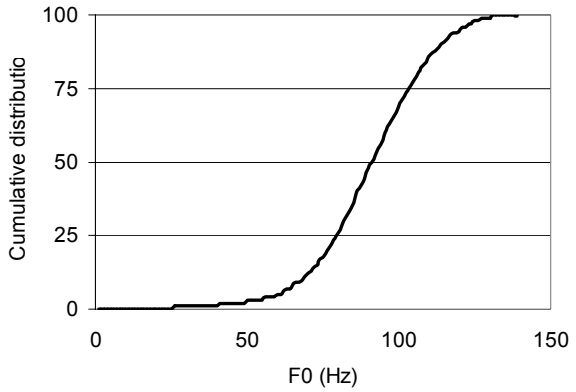

Figure 2. *The cumulative distribution of F0 for one subject, whose 25%, 50% and 75% levels are marked by the gridlines.*

All 498 speakers were arranged according to their F0 median, together with the speaker-specific 25 % and 75 % values (Figure 3). As can be seen the distance between the 25% and 75% values increases when the median gets a higher value, suggesting that pitch span is not completely independent from pitch level. A better way to describe the data and to make it more homogeneous is to use the semitone scale. This is clear from Figure 4 which shows that the lines representing 25% and 75% points run parallel with the median, as opposed to the diverging lines in Figure 3. In other words: the range becomes frequency independent (Figure 5). The advantage of using such a non-linear scale has been discussed in detail by e.g. [4] and [7].

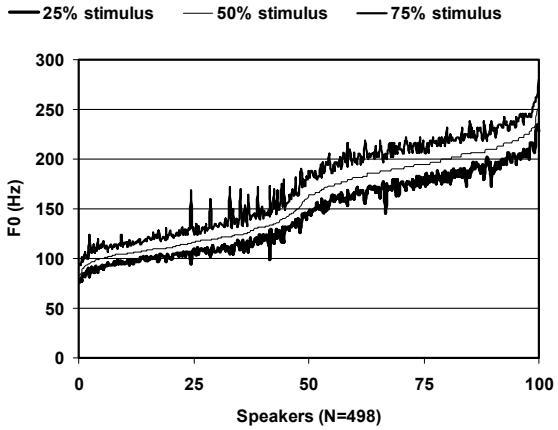

Figure 3*: Distribution of media, lower 25% and upper 75% levels for speakers of SpeeCon database (linear scale)*

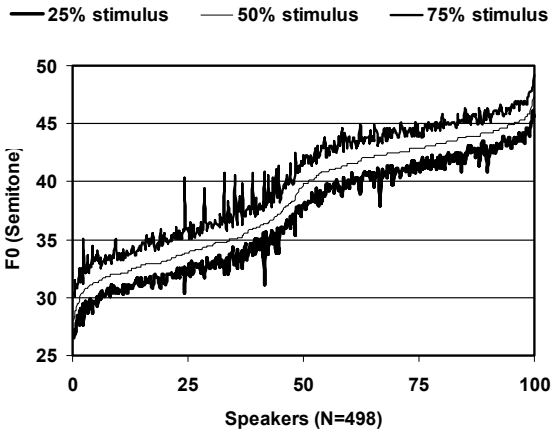

Figure 4*: Distribution of media, lower 25% and upper 75% levels for speakers of SpeeCon database (semitone scale)*

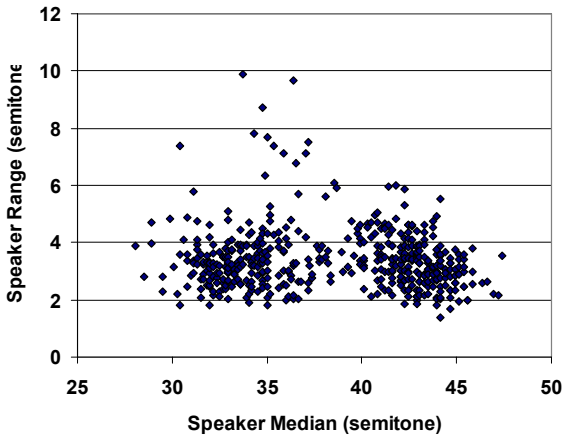

Figure 5. *Range distribution for all speakers (n=498) for different speaker medians (expressed in semitones)*

3. Experimental design

3.1. Stimuli

The stimulus selection for the perception experiment was carried out in three steps. Utterances with a median close to the 25% or 75% points were automatically chosen for all speakers (in order to avoid selecting outliers) and sorted according to stimulus duration. The top 100 speakers were selected having stimulus durations of about 1 second long. Finally, the number of speakers was reduced to 50, selected in such a way that the smaller group had an F0 median distribution which was representative of the distribution in the database as a whole, which was ten times larger.

3.2. Subjects

Subjects consisted of 13 speakers of Swedish, namely 4 colleagues from KTH and 9 students in logopedics from Umeå university, Sweden. They all participated as listeners in the current test on a voluntary basis.

3.3. Perceptual experiment

The 100 different stimuli (low and high pitch range utterance fragments from 50 different speakers) were mixed and presented sequentially to our listeners via a specifically designed interface, which allows to run perception experiments through the internet using a standard web browser with audio facilities. To minimize possible learning effects, each subject was presented with a differently randomized list of stimuli. In order to make sure that people made absolute judgments per speaker rather than compare pitch range levels within speakers, the stimulus set was split in two halves, such that the high and low version of one speaker did not occur in the same half of the test. The subjects' task was to rate, for each stimulus, on a 5-point scale whether they felt that the fragment was spoken in a relatively low range (1), a relatively high range (5), or at a range in between these two extremes (2-4). The actual test was preceded by a short introduction which briefly explained a few concepts (such as pitch range) and the actual task. No feedback was given on the "correctness" of their responses, and there was no interaction with the experimenters. During the test, subjects could listen as many times as needed to a given stimulus before giving an answer, but they could not return to a previous stimulus after a response had been entered. The task lasted between 10 and 15 minutes.

4. Hypotheses

Before we embark on the actual results of the experiment, let us first specify the different hypotheses that one could formulate. If we take as our zero hypothesis that listeners are not able to tell the difference between low and high range stimuli at all and therefore produce completely random results, there are at least three different ways in which that zero hypothesis could be rejected (see also Figure 6), namely:

Hypothesis H1: Listeners can make an estimate of a speaker's range and where an utterance is positioned in this range
Hypothesis H2: Listeners can not make an estimate of a speaker's range and make an absolute judgment of an utterance F0 irrespective of speaker characteristics.
Hypothesis H3: Listeners can estimate the speaker's gender and make an estimate where an utterance is positioned in the gender range

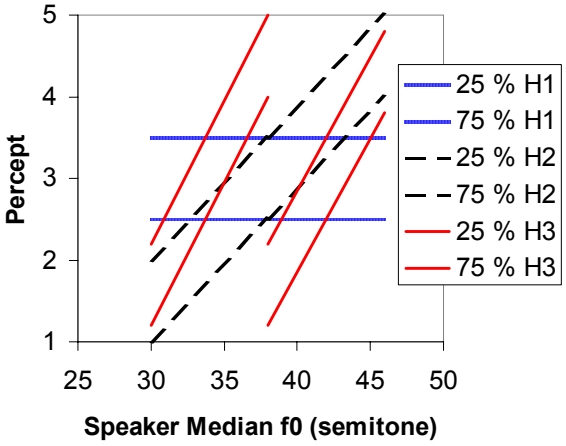

Figure 6*: Visualization of predicted outcomes for different hypotheses of the pitch range experiment*

5. Results

Figure 7 gives the overall difference in perception for stimuli with a relatively low and high pitch range, respectively. A paired t-test reveals that this difference is significant, both when comparing averages calculated per listener ($t=-6.83$, $df=12$, $p<0.001$), as when comparing averages calculated per speaker ($t=-4.65$, $df=49$, $p<0.001$). While this suggests that low pitch range stimuli on the whole can be distinguished reliably from high pitch range stimuli, it is obvious from looking at the averages that this task was a very difficult one. Also, the overall means do not allow us to decide which of the three hypotheses described above is the most probably one.

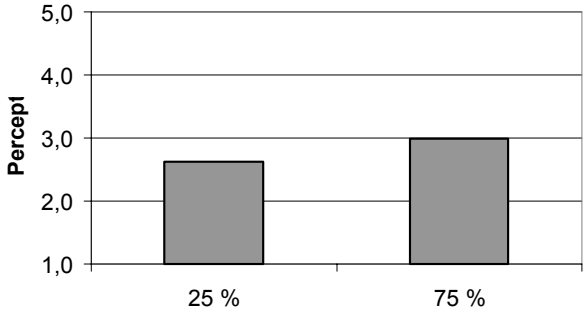

Figure 7*: Average range judgments for 25% and 75% stimuli*

Figure 8 shows the judgments for 25% and 75% stimuli for each selected speaker separately. Visual inspection of this figure suggests that our results are most compatible with hypothesis 3, as we can see that the plot can roughly be divided into two sections with the area between 35 and 40 semitones as a dividing point. Each section consists of two gradually increasing lines in which the 75% cases are generally judged to be higher than the 25% cases. In any case, the plot is different from what one would predict on the basis of hypothesis 1 (two parallel straight lines) or hypothesis 2 (one rather than two gradually increasing lines). Note also that the judgments appear to be speaker-specific: while in a majority of the cases the 75% values are judged higher than the 25% cases, this does not appear to be true for all speakers. Also, there is some variation in the distances between the two estimated values.

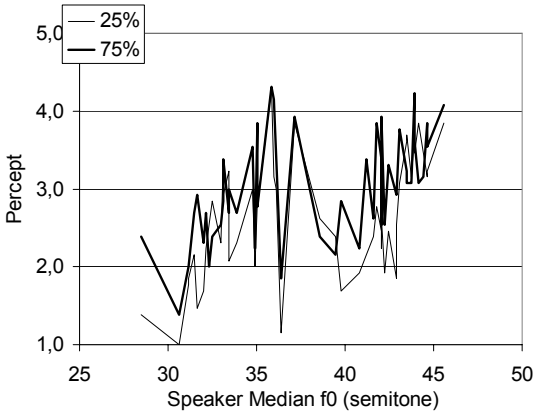

Figure 8*: Judgments of pitch range for 25% and 75% stimuli arranged per speaker*

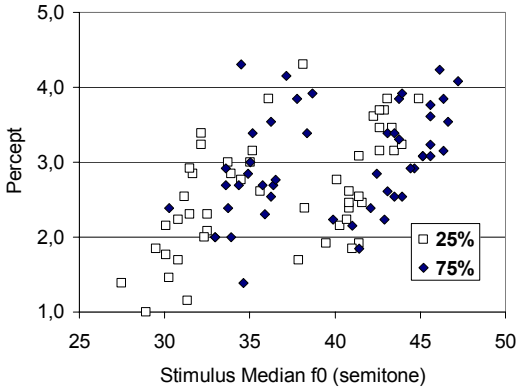

Figure 9*: Judgments of pitch range for each stimuli arranged according to stimulus median.*

Figure 9 presents the judgments by the listeners for different stimulus medians, **not** speaker medians. We can observe that the judgments are correlated with the stimulus median supporting the hypothesis that the listeners are unable to make a detailed model of the speakers range (besides the gender difference) when they are forced to make an absolute judgment on the basis of a single utterance.

6. Discussion and Conclusion

This paper has reported on a study that explores to what extent listeners are able to judge where a particular utterance fragment is located in that speaker's pitch range. The research consists of a perception study that made use of 100 stimuli, selected from 50 different speakers. Results reveal that listeners' judgments are gender-specific, but that within a gender they tend to hear differences in range. However, it is also clear from our data that the task was a very difficult one. We could therefore imagine follow-up studies, such as a within-speaker paired-comparison test to check whether such a paradigm will improve the results, or a test in which we take more extreme values than 25% and 75% as stimulus materials. It remains to be seen on what basis listeners were able to estimate the speakers' pitch ranges. It seems unlikely that they have based their judgments purely on utterance-final low targets, which are sometimes modeled as reference points for pitch range. As Ladd [6] already remarked, more recent findings suggest that these final lows are not as stable as traditionally assumed. Our own impression was that the higher range was somewhat more dynamic than the more monotonous lower pitch range values, which may therefore have served as a cue to listeners. Another possibility is that listeners have based their judgments on variation in voice quality. It is known that stretches of speech produced in a lower pitch range are sometimes characterized with particular unstabilities in pitch, such as jitter and shimmer, that are due to limitations of the vocal apparatus. Similarly, in our earlier study on break prediction [3], we found that relatively low pitch regions right before stronger breaks were accompanied by creaky voice. Finally, it has been shown that some aspects of the vocal source, such as open quotient, which has an impact on the perceived timbre of a person's voice, may also covary with pitch level, e.g. [11]. An adequate modeling of pitch range will improve our general understanding of intonation structure, since it is known that pitch range is exploited by speakers and listeners, for instance as a cue to prominence, phrasing and emotional connotations. For practical purposes, a good model of pitch range may be helpful to define an appropriate pitch scale, and could be beneficial for speech synthesis to adequately generate pitch variation.

7. Acknowledgments

Marc Swerts is also affiliated with the Fund for Scientific Research – Flanders (FWO - Flanders), with the Dutch NSF and with Antwerp University, Belgium. We would like to thank Theo Veenker for help with setting up the experimental environment. This work has been carried out within the Swedish project "Boundaries and groupings - the structuring of speech in different communicative situations" (GROG), a project whose overall goals is to model the structuring of Swedish speech in terms of prosodic breaks and groupings [2].

8. References

[1] Bezooijen R.A.M.G. van, 1984. *The characteristics and the recognizability of vocal expression of emotion*, Foris, Dordrecht, The Netherlands.

[2] Carlson, R,; Granström, B,; Heldner, M,; House, D,; Megyesi, B,; Strangert, E,; Swerts, M., 2002. Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. *Fonetik 2002*, TMH-QPSR, 44.

[3] Carlson, R,; Swerts, M., 2003. Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials, *ICPhS 03*.

[4] Fant, G,; Kruckenberg, A,; Gustafson, K,; Liljencrants, J, 2002. A new approach to intonation analysis and synthesis of Swedish, *Speech Prosody*, Aix-en-Provence, France.

[5] 't Hart, H,; Collier, R,; Cohen, A., 1990. A perceptual study of intonation. *CUP*.

[6] Ladd, D., 1996. Intonational Phonology. *CUP*.

[7] Nolan, F., 2003. Intonational Equivalence: An Experimental Evaluation of Pitch Scales, Proc. ICPhS 03.

[8] Portes, C,; Di Cristo, A., 2003. Pitch Range in spontaneous speech:semi-automatic approach versus subjective judgement , *ICPhS 03.*

[9] Rietveld A.C.M., 2003. (in progress).

[10] Rietveld A.C.M,; Gussenhoven C., 1985. On the relation between pitch excursion size and prominence, *Journal of Phonetics, 13*, 299–308.

[11] Swerts, M,; Veldhuis, R., 2001. The effect of speech melody on voice quality. *Speech com. 33:4*, 297-303