Intonational equivalence: an experimental evaluation of pitch scales

Francis Nolan
Department of Linguistics, University of Cambridge, UK

E-mail: fjn1@cus.cam.ac.uk

ABSTRACT

Intonationally equivalent utterances by, for instance, a man and a woman result in different pitch 'spans' when plotted on a Hertz scale. Alternative, psycho-acoustic, scales are available, such as semitones, mels, Bark and ERB-rate. Previous experiments have found hearers' linguistic pitch-related perception to be well modelled by one or another of these scales or even by Hertz, but there is no consensus. In this experiment subjects were asked to replicate equivalent 'template' intonation patterns produced by a male and a female speaker. The utterances, by design, were in three pitch spans. The 'goodness of fit' of the subjects' imitations was evaluated when pitch was represented in each of the above scales. Results for both female and male subjects showed that semitones and ERB-rate best reflect subjects' intuitions about equivalence, with semitones marginally the better. The reasons for this result, particularly its relation to previous work using prominence judgments, are discussed.

1. INTRODUCTION

Two speakers can produce intonationally equivalent utterances even though they have different 'tessituras' or personal pitch ranges. Intonationally equivalent utterances will share the same pitch accents and boundary tone(s) – to use the analytic constructs of autosegmental-metrical theory – but if they are completely equivalent they will also share the same pitch span, that is, the size of the excursion between high and low points in the pitch contour [1,2]. Thus a good imitation of a relatively monotonous, depressed-sounding man's utterance by a woman will also sound relatively monotonous and depressed. Her span, however, will be much larger if expressed in Hertz because her tessitura is higher and, as is well known, the Hertz scale is linear whereas our perception and production of pitch generally are not.

The problem which this paper tackles, then, is how to transform a fundamental frequency contour expressed in Hertz into a representation which will correctly capture speakers' intuitions about equivalence of intonational span.

On the face of it, the solution is obvious: transform the data to a psycho-acoustic pitch scale. The problem is that there are a number of such pitch scales, and there seems to be no consensus on which should be used for intonation. The question is not easy to resolve in principle, partly because the perception of the pitch of complex waves such as (voiced) speech does not depend uniquely or even primarily on the fundamental harmonic, but is contributed to by higher harmonics [3]. The experiment reported here will compare Hertz, semitones, mels, Bark and ERB-rate.

2. PSYCHO-ACOUSTIC SCALES

The purpose of a psycho-acoustic scale is to provide steps which correspond to equal perceptual intervals. In the case of pitch perception, the best known such scale is the musical semitone scale. Each musical octave is divided into twelve semitones, easily visualised as the white and black notes on the piano. The semitone scale is a logarithmic transformation of the physical Hertz scale, as can be seen clearly in Figure 1.

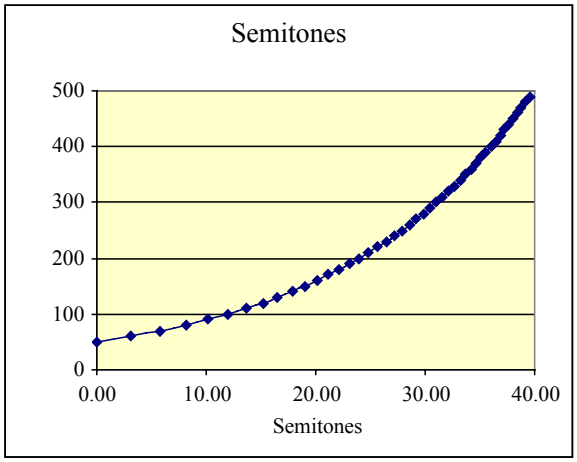

Figure 1: The relation between Hertz and semitones in the range 50 to 500 Hz

The Bark scale is derived from experiments which establish the maximum bandwidth within which masking noise will affect the perception of a tone. This 'critical band' becomes larger at higher frequencies, and the Bark scale is approximately logarithmic; for frequencies below 500 Hz, however, which is the main region for the fundamental frequency of the speech signal and those harmonics likely to play a role in pitch perception [3], is it nearly linear (see Figure 2 overleaf).

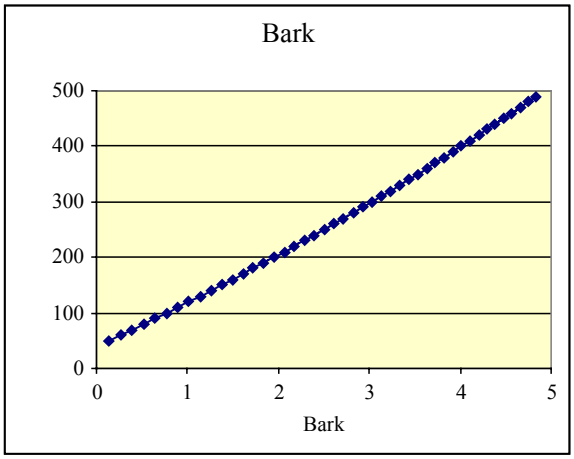

Figure 2: The relation between Hertz and Bark in the range 50 to 500 Hz.

The mel scale, which is derived from experiments in which subjects adjust one tone to be twice as high or half as high as a reference tone, arrives at a similar linear relationship below 500 Hz and logarithmic above. This is not shown.

The ERB-rate scale (Equivalent Rectangular Bandwith) is derived in a way conceptually similar to the Bark scale, but using a different technique. Here, the masking noise is presented with a rectangular 'notch' around a tone, and the minimum bandwith of notch established above which perception of the tone is not affected. Again, the resultant scale is logarithmic at higher frequencies, but, as shown in Figure 3, below 500 Hz it is between linear and logarithmic.

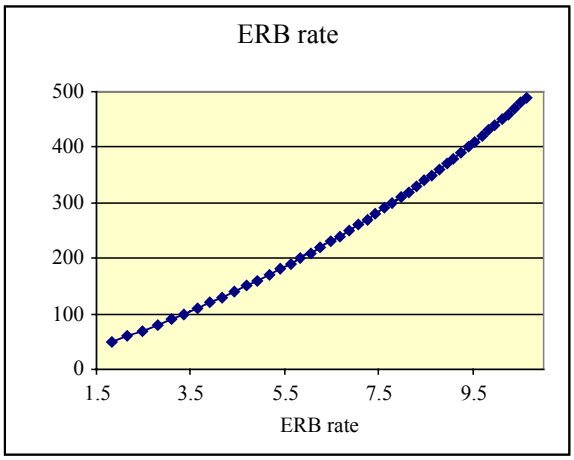

Figure 3: The relation between Hertz and ERB-rate in the range 50 to 500 Hz.

3. PREVIOUS WORK

In [4] subjects were asked to judge the relative strength of pairs of accents synthesised with different peak fundamental frequencies. The relevant outcome was that the judgments correlated better with excursion size when the excursion was expressed in Hertz than in semitones. This appears to speak against the use of psycho-acoustic scales. On the other hand, for instance, [5] reports an experiment in which subjects adjusted the height of synthetic stimuli to match the prominence of a reference in a lower or higher pitch register, and found that subjects' behaviour was better modelled by the ERB-rate scale than by either Hertz or semitones.

The conflicting results of previous experiments, and the fact that they have focussed on judgments of prominence rather than melodic equivalence, motivated a new experimental approach. It would also deal explicitly with cross-gender pitch equivalence.

4. EXPERIMENT

The approach used here aimed directly to tap speakers' intuitions about intonational equivalence by getting them to imitate utterances as closely as possible in their own voice. If this turned out to be a feasible task (it did), it was assumed that the best pitch scale would be the one which yielded the smallest 'error' between the pitch span of each template and that of its replication.

A conversational dyad was constructed which consisted of only sonorant sounds as follows, together with a specific intonation pattern (represented here in ToBI transcription):

[A] We were relying on a milliner.

H+L* H* L- L%

[B] A milliner

H* L- H%

In the first stage, template utterances were created by two phoneticians, one male (the author) and one female (RS). FN would produce a token of [A] and, after a gap, [B], in one of three impressionistically defined pitch spans: neutral, compressed, and expanded. In the gap immediately after [A], RS would attempt an exact replication of the utterance in her own tessitura; likewise immediately after [B]. The same procedure was followed for the other pitch spans, and the whole process repeated a further two times yielding three complete sets of utterances. From these, the auditorily most accurate male-female replication pairs, one for each pitch span, were subsequently chosen as the 'templates' for the subjects.

Once the twelve template utterances (3 [A] and 3 [B] utterances each for FN and for RS) had been selected, they were arranged in [A]-[B] pairs in which the [A] and [B] were always by different speakers (therefore always a male voice responding to a female voice, or vice versa). The pitch span of [A] and [B] within a pair was chosen randomly, and so did not match except by chance. This gave six pairs. The randomisation was done twice more, so that subjects faced the task of imitating 18 [A]-[B] pairs, in each of which the template speaker and the pitch span varied randomly. Each template utterance was presented

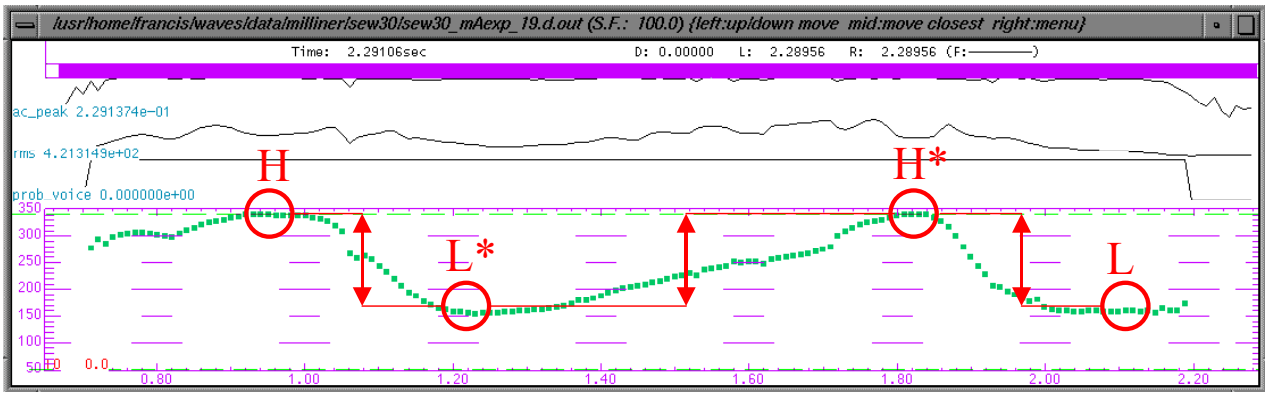

Figure 4: Illustration of three span measurements in the F0 trace of 'We were relying on a milliner'.

over headphones from a unix workstation, and the subject's replication recorded direct to disk.

The subjects were eight male and eight female students aged between 19 and 30 who had been trained in intonation analysis within the 'British' approach. The point of using intonationally trained subjects was to ensure accurate replication of the desired intonation patterns. All subjects spoke a variety approximating standard southern British English, though with traces of regional accents. All had had some musical training, but in no case to an advanced level.

Subjects were recorded one by one in the sound treated booth. They were given an instruction sheet which included each utterance and a transcription of its intonation in a system they had been taught. They were told that in the pause after each [A] or [B] utterance they should repeat it, aiming to 'produce an intonationally equivalent utterance in your own voice'. The intention of this instruction was that subjects should imitate the span, but not the tessitura, of the templates; and to a large extent subjects seemed to be able to achieve this, accurately replicating the intonational phonology of the utterances. This provided a dataset which directly tapped the speakers' intuitions about intonational equivalence – at least within the tolerances of their productive and perceptual accuracy (we must allow that performance errors will make replications less than perfect).

5. ANALYSIS

Analysis was carried out using Silicon Graphics Unix workstations running ESPS (Entropics Signal Processing Software) and Xwaves+. By design the intonation patterns of both utterances had easily definable high and low targets, as shown for a token of the [A] utterance whose fundamental frequency trace is shown in Figure 4. Span measurements were calculated as the difference between each successive high and low or low and high; three spans for [A] and two for [B].

The input to the comparison of pitch scales consisted in the 'error' between each span in each replication and the equivalent span of the template of which it was a replication. The idea is illustrated in Figure 5. In each scale the span of a replication was compared with that of the relevant template, and the difference calculated. Since it would not be possible to compare raw differences across the scales since each scale is numerically different, the difference was expressed as the absolute value of the ratio of the difference to the span of the template, expressed as a percentage. The error is thus expressed in a form which could be compared directly across the scales. The error was averaged over the three repetitions of a template by a subject, giving six error values per subject (one for each male and female template in three spans).

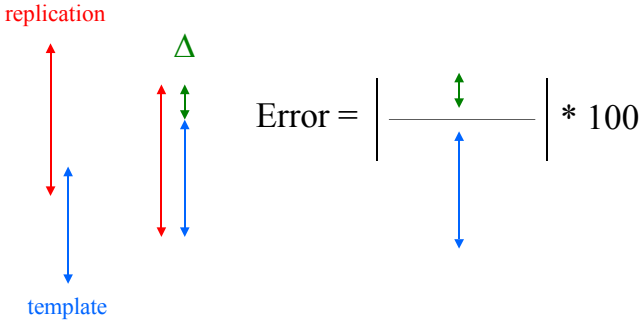

Figure 5: Schematic representation of the calculation of the 'error' between replication and template spans. The replication error is expressed as a percentage of the template span, in absolute value.

6. RESULTS AND DISCUSSION

The 'replication error' values were subject to a repeated measures ANOVA with replication error as the dependent variable and pitch scale and subject-sex as the independent variables. Pitch scale (with five levels, Hz, ST, ERB, mel, and bark) was shown to have a significant effect on error magnitude ($F(1.03, 7.21) = 63.12$, $p<0.001$, Greenhouse-Geisser correction), while subject-sex was not significant. There was a significant interaction between scale and subject-sex ($F(1.10, 7.70) = 23.71$, $p<0.001$, Greenhouse-Geisser correction). Replication error magnitudes are shown in Figure 6 separately for male and female subjects.

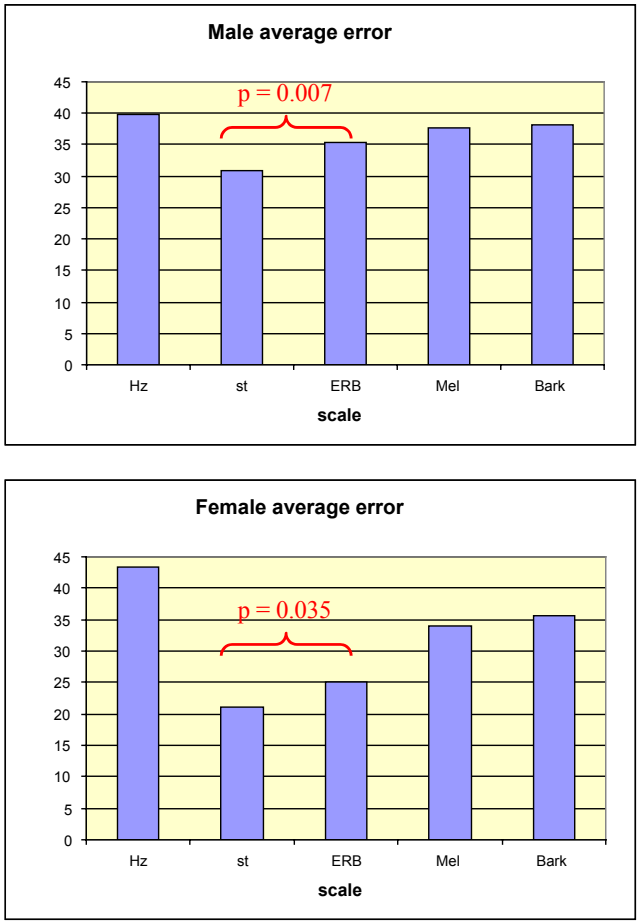

Figure 6: Error expressed as a percentage of the template; male subjects (above) and .female (below).

It can be seen that for both male and female speakers semitones result in the smallest replication error, with ERB-rate a close second. Post hoc paired-samples t-tests revealed that most between-scale differences were significant at the $p<0.01$ level, with the exception of the difference between semitones and ERB-rate for female speakers which was significant only at $p<0.05$. The experiment suggests that a logarithmic (st) or near logarithmic (ERB) scale best models intonational equivalence. The results point more in the direction of a logarithmic scale than those of [4], which supported a linear (Hz) scale, or those of [5], which found the ERB-rate scale better than semitones.

The explanation may lie in the nature of the task. Both those earlier experiments used judgments of accentual prominence, whereas the present experiment focussed on intonational pitch span. Speculatively, this may be a more purely 'musical' task than those involving accentual prominence, where the role of pitch may be confounded with that of the co-varying factor of amplitude. If so, it is not surprising that the semitone scale performed well. The practical motivation for this experiment was the question of what scale to plot intonation contours on, particularly when men and women are to be compared, and the results indicate that the semitone scale is a well-motivated choice.

Further analysis will be required to check that the method of assessing goodness of fit between replication and template is optimal, and to tease out whether the scales which come out best (semitones and ERB-rate) are superior in all permutations of the task, e.g. male subjects replicating the expanded-span female template and female subjects replicating the compressed-span male template.

4. CONCLUSIONS

The experiment reported here has demonstrated the viability of an experimental method which directly taps speakers' intuitions about equivalence of intonational span across speakers. Provisional analysis of the results indicates that these intuitions are best modelled by a psycho-acoustic pitch scale which is logarithmic (semitones) or near-logarithmic (ERB-rate) in the frequency range of interest.

ACKNOWLEDGMENTS

The research reported in this paper was carried in Cambridge University Phonetics Laboratory with the collaboration of Eva Liina Asu, Margit Aufterbeck, Rachael Knight, and Brechtje Post. We are grateful to Rachel Smith for producing the female templates, and to Geoffrey Potter for technical support.

REFERENCES

[1] D.R. Ladd, *Intonational Phonology*, Cambridge: Cambridge University Press, 1996.

[2] D. Patterson and D.R. Ladd, "Pitch range modelling: linguistic dimensions of variation", in *Proceedings of the 13th International Congress of Phonetic Sciences, San Francisco*, pp. 1169-1172, 1999.

[3] B.C.J. Moore, "Aspects of auditory processing related to speech perception", in *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver, Eds, pp. 539–565. Oxford: Blackwell, 1997.

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

[5] D.J. Hermes & J.C. van Gestel, "The frequency scale of speech intonation", *Journal of the Acoustical Society of America*, 90, pp. 97–102, 1991.