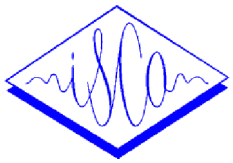


Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

Global and local evaluation of prosody: Discrete target and just noticeable differences

Tomáš Duběda

Institute of Phonetics, Charles University in Prague
dubeda@ff.cuni.cz

Abstract

We describe an experiment based on introspective assessment of inter-syllabic prosodic changes in two different conditions (global – continuous sentences, and local – two-syllable portions of speech), by means of discrete targets (*Higher, Same* and *Lower*). The results show that global evaluation is very problematic for the listeners and that f_0 is the best assessed parameter, followed by intensity and duration. They also provide a realistic, continuous-speech-based view of just noticeable differences, at least for f_0 and intensity.

1. Introduction and rationale

In the present experiment, we test the way listeners *describe* the evolution of three prosodic parameters (f_0, duration and intensity) in continuous speech, with the aim to learn more about how they *perceive* it. In a task where short Czech sentences were to be annotated with respect to height, length and loudness changes from one syllable to the other, we study the relations between such a description and the underlying acoustic data; this also includes the problem of just noticeable differences (JNDs).

The use of discrete targets in intonational studies has been a common method for a few decades [5, 6]. For duration and intensity, this approach has not yet been really exploited, mostly because these two parameters are more embedded in the segments and show less autosegmental behaviour. Their perception is less sharp, and their variations more subtle. In [3], an attempt to stylize f_0, duration and intensity within stress units by means of *Higher, Same* and *Lower* targets is described, based on JNDs found in literature.

The present experiment aims at estimating how realistic such a stylization is by comparing the listeners' annotation with the measured data in two different testing conditions.

In normal speech perception, we evaluate prosody:

(i) as a bundle of interacting parameters;

(ii) within a larger context;

(iii) with respect to linguistic levels of analysis.

In the formal analysis of prosody, however, we first have to move down to a more atomistic level, where the mentioned perceptual aspects (i)–(iii) should be neutralized. Stimuli may be then transformed e. g. in the following way:

(i) can be resolved e. g. by normalizing some parameters while leaving the studied one untouched. This method involves some artefacts [13], but should be reasonably safe for our purposes.

(ii) can be dealt with e. g. by reducing the available context. This solution is problematic because results obtained for short stretches of speech demonstrably do not apply to continuous speech, as far as prosody is concerned [13]. Therefore, we decided to study both the local and global evaluation of prosodic changes.

(iii) can be overcome e. g. by using delexicalized speech. This method seems to be adequate for intonation, but not for duration and intensity, where changing the underlying segments would imply fine – and risky – calculations of the target values. We thus decided to use natural speech, and to guide the listeners towards formal rather than linguistic judgments.

The language tested is Czech, a Western Slavonic language with a fixed stress (though rather weak) on the first syllable of stressable words, with no lexical tones, with phonological vowel length distinction and no pronounced reduction of unstressed syllables.

In sum, our perceptual tests should show how the listeners evaluate local changes of f_0, duration and intensity by means of a uniform set of discrete targets. This objective is not very ambitious for intonation, but could provide interesting evidence for the latter two parameters, despite the fact that the task is introspective. Unlike many psychoacoustic experiments on prosodic discrimination where stimuli are short and well-controlled [1], our test setting remains much closer to real speech.

2. Method

We used six different Czech sentences whose average length is 23 syllables. Three male speakers were involved, each pronouncing two sentences.

To isolate each of the three parameters (i. e. to resolve the obstacle (i), as defined above), we neutralized the other two in this way:

- evaluation of f_0: duration and intensity neutralized;
- evaluation of duration: f_0 and intensity neutralized;
- evaluation of intensity: f_0 and duration neutralized.

The neutralization was carried out in Praat [2], by means of re-synthesis:

- f_0: setting pitch to the average over the sentence;
- duration: setting the duration of each segment to its average intrinsic duration over a large sample of data measured for the same speaker;
- intensity: setting the intensity of each segment to its average intrinsic intensity.

In this manner, we obtained three different versions for each sentence.

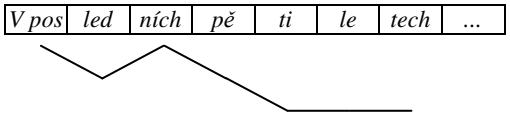
In a pilot test session, these sentences were transcribed by 6 listeners using *High, Low* and *Same* targets. A brief analysis of the results showed that there was very poor matching with the acoustic data, and that there were often technical rather than perceptual uncertainties about the use of the symbols. Therefore, we decided to switch to a graphical analysis which should be closer to the percepts. There were two different tests, one for global prosodic assessment and one for local assessment.

10.21437/SpeechProsody.2008-161

2.1. Test 1

12 listeners, students of phonetics or linguistics (6 post-graduate and 6 pre-graduate), all of them having previous experience in analytic listening, were asked to transcribe the evolution of each prosodic parameter (f_0, duration and intensity) in modified sentences by means of a broken line. The instruction was to evaluate each pair of syllables, and to use either of the elements ／, ＼ or ＿ to express whether the parameter is rising/falling/ staying the same from one syllable to the other, irrespective of the magnitude of this change. Within-syllable prosodic behaviour was to be ignored (theoretical accounts of Czech intonation seem to indicate that there is rarely more than one intonation target in a syllable; for duration and intensity, supposing more than one target would be nonsense).

In an interactive application, the listeners were first shown how to transcribe each of the parameters. Then they were presented with the 18 sentences (in three blocks – intensity evaluation first, then duration, then f_0). Each sentence could only be listened to as a whole, but as many times as needed. For duration, the instruction stated that no segmental length (short *vs.* long vowels), but only prosodic duration was to be analyzed. Also, the listeners were asked not to rely on any theoretical knowledge about Czech prosody, because the sentences had been modified in a specific way (which they actually were). In the form, each sentence was broken into boxes corresponding to syllables. A transcription example (*V posledních pěti letech...* [ˈfpɔslɛdɲiːx ˈpjɛcɪ ˈlɛtɛx] "In the last five years..."):



The resulting line should not be seen as a stylization of the pitch course, but as the sum of elementary pitch movements.

These graphical elements were then converted into *H, S* and *L* targets (or, numerically, into 1, 0 and –1 values):

–	L	H	L	L	S	S	...
–	–1	1	–1	–1	0	0	...

2.2. Test 2

In the second test, designed for the local evaluation of prosody, 12 listeners (4 post- and 8 pre-graduate; 3 having participated in Test 1) assessed two-syllable stretches extracted from the transformed sentences (with initial and final amplitude damping to prevent perceptual effects of the cutting point), assigning one of the values *rise/level/fall* (1/0/–1) to each syllable pair. Only half of the data (one sentence per speaker instead of two) was used in this "gating" test, to prevent it from being too long for the listeners. The recommended number of repetitions for each stimulus was two. The order of the stimuli remained the same as in the source sentence. The other aspects of the test (application, the order of items, linguistic cautions) were the same as in Test 1.

3. Results and discussion

3.1. Listeners' performance

To assess the matching between acoustic differences and the transcripts obtained, we compared each syllable-to-syllable difference with its mean rating (the average of the collected 1/0/–1 judgements). The acoustic data were established on syllable nuclei in the following way:

– f_0: average f_0 over the nucleus [Hz]; difference in %;
– duration: duration of the nucleus weighed by its intrinsic duration, so as to capture the prosodic effect of duration only; difference in %;
– intensity: intensity of the nucleus weighed by its intrinsic intensity, so as to capture the prosodic effect of intensity only; difference in dB.

Note that all acoustic descriptors are based on the nucleus only. For intonation, this solution does not seem to be unrealistic [7]; for intensity, vocalic nuclei have been shown to exhibit regular patterns in Czech stress groups [4]; for duration, informal tests have shown that normalized nucleus duration is a slightly better predictor than normalized syllable duration.

Table 1 provides an example of this analysis.

Table 1: *Example of the analysis (intensity).*

Syllable	Difference from the preceding syllable [dB]	Average rating	Standard deviation of the rating
V pos	N. A.	N. A.	N. A.
led	–4.5	–0.40	0.49
ních	1.8	–0.40	0.49
pě	–1.5	0.00	0.63
ti	–2.3	0.00	0.00
le	–0.9	–0.20	0.40
tech	–0.8	–0.20	0.40
Correlation between the second and the third column for this example: $\rho = -0.12$			

The correlation coefficient (as shown in Table 1) seems to be a good instrument for quantifying the matching between acoustics and perception; it is important, however, first to see what its variations in the present case may be. We calculated this coefficient for an ideal case where the listeners would assign a 1/–1 value to any difference which is beyond the putative threshold as used in [3] (i. e. 2% for f_0, 10% for duration and 1 dB for intensity), and a 0 value for differences which are below this threshold. This ideal correlation is 0.83 for pitch, 0.85 for duration, and 0.83 for intensity. All correlations obtained should be interpreted with respect to this topline, i. e. considered as effectively higher.

The overall correlation between the acoustic differences and the corresponding perceptual ratings was found to be rather poor. This means that the task was difficult, even with respect to a looser topline, and the results do not permit a relevant analysis. One possible way of reducing chaos in the data, unless we change the very definition of the task, is selecting listeners with good performance only: this strategy is sometimes used in phonetic research (e. g. asking trained phoneticians to distinguish sounds, or professional actors to mimic emotional prosody), and – hopefully – does not imply that "prosody is not for everyone". We decided to evaluate

both the mean correlation of each listener's rating with the acoustic data, and the number of correct identifications of the 15 greatest positive and 15 greatest negative acoustic changes, where the listeners' behaviour should be especially systematic; the five listeners who met these conditions best were selected for further analysis. The listeners' performance in the two test conditions is displayed in Table 2.

Table 2: *Listeners' performance (correlation between acoustic differences and corresponding mean perceptual ratings).*

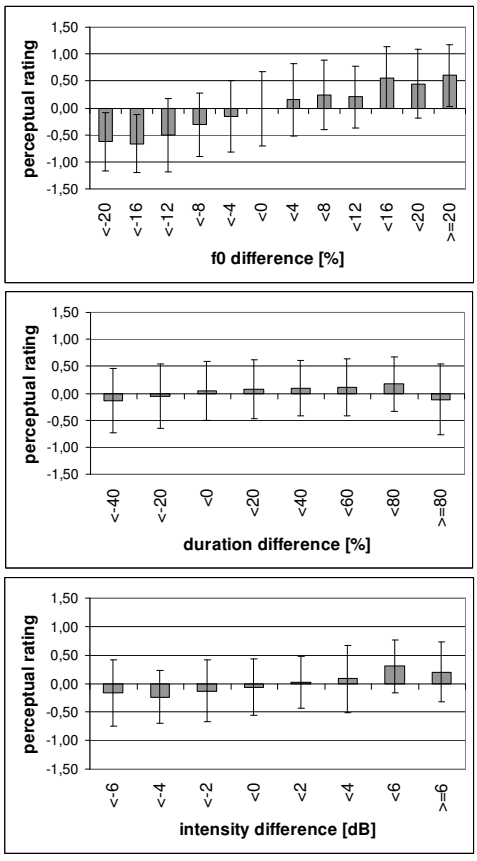
	Test 1		Test 2	
Parameter	All listeners	Selected listeners	All listeners	Selected listeners
f_0	0.63	0.67	0.82	0.77
duration	0.21	0.19	0.32	0.30
intensity	0.21	0.40	0.39	0.34

Test 2 led to a clearly better overall performance than Test 1, which was expected. The performance for selected listeners in Test 1 is slightly better for f_0, slightly worse for duration, and clearly better for intensity. Selected listeners in Test 2 performed slightly worse than all listeners, which can be explained by the way they were selected (apart from the overall score, their score for the greatest differences was also taken into account). Since we believe that perceptual reliability should include a good assessment of extreme differences, we shall keep the selected group despite their slightly worse results. Among the three parameters, f_0 was rated the best, with a correlation close to the topline in Test 2. Next comes intensity with a correlation which is approximately twice as small, followed by duration.

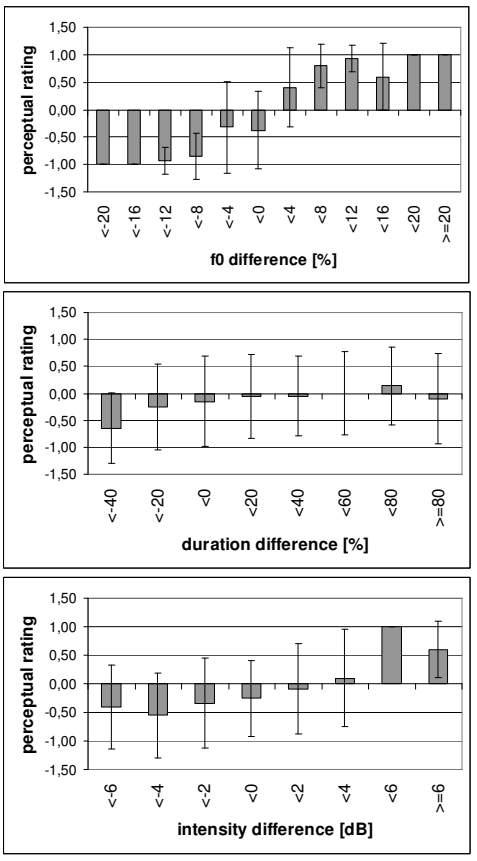
All following analyses are based on the selected listeners only.

3.2. Test 1

Figures 1–3 show how the selected listeners evaluated acoustic differences of different magnitude in continuous speech. In all three graphs, the ratings grow generally from left to right, as expected, but in the case of duration and intensity, the low scores for all differences (including large ones), as well as huge standard deviations indicate that it was virtually impossible for the listeners to assess these parameters (as measured on vocalic nuclei and normalized) in the given situation. For f_0, we again see high standard deviations, but listeners seemed at least to be unanimous about a non-rise for falls over 12%, and about a non-fall for rises over 12%. Hardly anything can be said about the JNDs in this situation.



Figures 1–3: *Perceptual rating of acoustic differences in continuous speech (Test 1) – means and standard deviations. On the absciss, negative values correspond to falls, and positive values correspond to rises. On the ordinate, "1" corresponds to full "rise" agreement, and "–1" corresponds to full "fall" agreement. Total N for each graph: 125 stimuli x 5 ratings.*



Figures 4–6: *Perceptual rating of acoustic differences in two-syllable stretches of speech (Test 2) – means and standard deviations. Total N for each graph: 62 stimuli x 5 ratings. See legend of Figures 1–3.*

3.3. Test 2

Figures 4–6 show that in isolated pairs of syllables, the results achieved – compared to Test 1 – were similar for duration, better for intensity, and much better for f_0. Let us repeat that all results put forward for Test 2 are based on half the data used in Test 1.

It is almost impossible to deduce anything from the duration graph (Figure 5), except for the fact that the listeners may have heard the second syllable as shorter than it was, because they expected a final lengthening [13]. This can be actually seen in the negative rating of positive changes up to 40%.

In the case of intensity (Figure 4), it seems that there was certainty about a non-fall when the positive deviation was greater than 4 dB; the certainty threshold for non-rises is less clear, and may lie around 4 dB as well. These perceptual boundaries could be declared JNDs if we were certain enough about the representativeness of the acoustic descriptor used. The roughness of the histogram also plays a role. Interestingly, no perceptual effect consisting in underestimating the intensity decrease on the second syllable [13] can be found in our data; on the contrary, negative perceptual ratings continue right from the 0% intensity change point.

F_0 changes were assessed with much unanimity if they were greater than 8% when falling, and greater than 4% when rising. These two points seem to mark category boundaries on the rating curve, and can be considered JNDs for the given test setting. They converge with 't Hart's results on JNDs [10] in continuous speech, and their asymmetry corresponds to the findings by the same author about the effect of pitch change direction on the JNDs [11]. Hence, they give evidence for the unmarkedness of the falls [12].

4. Conclusion

Our experiment confirmed the fact that global perception of prosody is more resistant to introspection than is its local assessment. The incapacity of describing our own prosodic percepts when we listen to continuous speech does not mean, however, that no perception occurred: prosodic parameters were only integrated into patterns and embedded into linguistic units. The difference between the Test 1 and Test 2 conditions may be compared to that between a landscape seen from a tower and from the ground. From the ground, we distinguish more details, but need more time to explore them, and may miss larger-sized patterns. This metaphor also holds for J. Vaissière's claim that *the well-established experimental methods developed in psychoacoustics do not actually apply in the field of intonation* [13, p. 240]. To understand speech perception, we should preferably construct tests from the "tower" perspective; many questions, however, cannot be formulated like this. The "ground" perspective may therefore be used, provided that it is interpreted adequately.

Going back to the motivation of this experiment – determining perceptual thresholds for parallel stylization of the three parameters – the data obtained in Test 2 are especially useful with respect to f_0, including the confirmation of the asymmetry between positive and negative deviations. For intensity, a largely neglected variable, our data are reasonably promising: it should not be impossible to make a perceptually founded stylization of this parameter. As for duration, the results are disappointing, and might indicate problems with the descriptor selected (normalized nucleus duration). In all three cases, the tests confirm that JNDs established for stationary signals (e. g. [1, 8]) cannot be directly applied in real speech prosody.

Possible refinements of this experiment include testing perception in a semi-global context (e. g. that of a stress unit), or comparing sensitivity to prosodic changes in different positions (initial, final, stressed, nuclear).

5. Acknowledgement

This research was supported by the GAČR 405/07/0126 grant.

6. References

[1] Beckman, M. E., 1986. *Stress and Non-Stress Accent,* Netherlands Phonetic Archives No. 7, Foris Publication.

[2] Boersma, P.; Weenink, D., 2006. *Praat: Doing phonetics by computer* (Version 4.4.04).

[3] Duběda, T., 2006. Investigating stylized prosodic contours in Czech. *16th Czech-German Workshop.* R. Vích (ed.). Prague, 41–47.

[4] Duběda, T., 2006b. Intensity as a macroprosodic variable in Czech. *Speech Prosody 2006*, Dresden, 185–188.

[5] Hirst, D.; Di Cristo, A. (eds), 1998. *Intonation Systems. A Survey of Twenty Languages*. CUP.

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

[7] Mertens, P., 2004. The Prosogram: semi-automatic transcription of prosody based on a tonal perception model. *Speech Prosody 2004*, Nara.

[8] Pols, L. C. W., 1999. Flexible, robust, and efficient human speech processing versus present-day speech technology. *Proceedings of the 14th ICPhS,* San Francisco, 9–16.

[9] Rosen, S. M.; Fourcin, A. J., 1986. Frequency Selectivity and the Perception of Speech. *Frequency selectivity in hearing*. B. C. J. Moore (ed.). Academic Press, 373–487.

[10] 't Hart, J., 1974. Discriminability of the size of pitch movements in speech. *IPO Annual Progress Report 9,* 56–63.

[11] 't Hart, J., 1981. Differential sensitivity to pitch distance, particularly in speech. *JASA 69(3),* 811–821.

[12] Vaissière, J., 1983. Language-Independent Prosodic Features. *Prosody. Models and Measurements*. A. Cutler (ed.), 53–66.

[13] Vaissière, J., 2005. Perception of intonation. *The Handbook of Speech Perception.* D. B. Pisoni; R. E. Remez (eds). Blackwell Publishing.