

Pitch Accents, Boundary Tones and Contours: Automatic Learning of Czech Intonation

Tomáš Duběda¹ and Jan Raab²

¹ Institute of Phonetics, Charles University in Prague

² Institute of Formal and Applied Linguistics, Charles University in Prague
dubeda@ff.cuni.cz; raab@ufal.mff.cuni.cz

Abstract. The present paper examines three methods of intonational stylization in the Czech language: a sequence of pitch accents, a sequence of boundary tones, and a sequence of contours. The efficiency of these methods was compared by means of a neural network which predicted the f_0 curve from each of the three types of input, with subsequent perceptual assessment. The results show that Czech intonation can be learned with about the same success rate in all three situations. This speaks in favour of a rehabilitation of contours as a traditional means of describing Czech intonation, as well as the use of boundary tones as another possible local approach.

Key words: Pitch accents, boundary tones, contours, Czech, prosody, automatic learning, neural networks.

1 Rationale

Systematic descriptions of Czech intonation available are based exclusively on contours (e. g. [2], [10], [8]), while no systematic tonal approach has been proposed thus far. Aside from this, the contour-based models are heavily biased towards the nuclear parts of intonation phrases (i.e. mostly final parts, with important phonological and paralinguistic functions), considering prenuclear material (i.e. parts preceding the nucleus) implicitly as uninteresting. The only extensive account of prenuclear intonation, oriented towards speech synthesis, can be found in [9].

From a theoretical viewpoint, the current lack of tone-based models might be more than an effect of historical inertia (most traditional approaches of intonation have been contour-based, while discrete targets are a relatively recent innovation): in [7], for instance, it is argued that Czech intonation is resistant to modeling by means of local events, and that the relevant building stone is the stress unit with its holistic contour. In addition to the pitch accent/stress unit dichotomy, a third theoretical possibility would be that intonation is based on boundary tones at either edge of the stress unit (cf. J. Vaissière's theory of 'English as a stress language' and 'French as a boundary language', [11]). It is precisely these three hypotheses that we test in the present study.

Czech is a non-tonal language with word stress located always on the first syllable of stressable words (cf. examples below). Content words may form stress units with following grammatical words. In some cases, stress units may start with one or more unstressed syllables due to preceding clitics. Czech accents are felt as weak by many foreign listeners, which may be due to low tones which frequently accompany the stressed syllable, as well as to the troublesome interference of segmental vowel length.


2 Goal, Material and Methodology

2.1 Goal

The question underlying this research is whether intonation is better anchored as a sequence of pitch accents, as a sequence of boundary tones, or as a sequence of contours. The tool for this comparison is a neural network which should learn how to predict the f_0 curve with the input provided. The output will then be assessed numerically and perceptually. A similar – though not identical – approach to intonational stylization can be found e. g. in [3].

We shall thus compare the target approach with the contour approach, and, within the target approach, compare the importance of pitch accents and boundary tones. We shall study non-nuclear intonation only. In this manner, there are three types of input to the neural network, responding to three theoretical assumptions:

1. Prenuclear intonation in Czech can be modeled as a sequence of pitch accents (PAs). We use bitonal, phonetic pitch accents which typically code intonational changes within a 3-syllable window centred around the stressed syllable, each tone expressing the change from one syllable to the other:



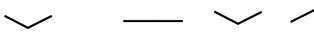
S	H		H	H		S	L	H	H	
'pru:	mpɛr	ne:	'vi:	nɔ	sɪ	'tʃɛs	kix	'vɪ	juʦ	...

Průměrné výnosy českých vinic...

‘Average yields of Czech vineyards...’

There are three relative tones: Higher, Same and Lower. The first tone of the intonation phrase is always (formally) S. Thick frames mark stress units.

2. Prenuclear intonation in Czech can be modeled as a sequence of boundary tones (BTs). We use two phonetic boundary tones in each stress unit: one at the beginning and one at the end. In this way, we typically code a 3-syllable window centered around the last syllable of each stress unit:



S		L	H		S	S	L	H	H	
'pru:	mpɛr	ne:	'vi:	nɔ	sɪ	'tʃɛs	kix	'vɪ	juʦ	...

3. Prenuclear intonation in Czech can be modeled as a sequence of contours (CONTs). We use one contour in each stress unit, only reflecting unit-internal intonational behaviour:

HL			H			L		H		
'pru:	mjɛr	ne:	'vi:	nɔ	st	'tʃɛs	kix	'vi	ju̯ts̄	...

The inventory of attested contours is: H (rising), HL (rising-falling), L (falling) and S (flat).

2.2 Material

The material was a set of manually-annotated recordings of a text read by 5 speakers of standard Czech (3 men and 2 women). The length of the text was ± 900 syllables, out of which ± 700 were pre-nuclear, i.e. relevant for our study. The total number of syllables processed was $\pm 3\,500$.

The tonal annotation, as exemplified above, was carried out by the first author on a perceptual basis with simultaneous inspection of the manually-checked f_0 curve. The labeling was perception-based and syllable-synchronized (each syllable to be annotated received exactly one tone). Three relative tones were used (Higher, Lower and Same). The terms ‘pitch accents’ and ‘boundary tones’ as used in this paper are not part of a phonological model, and should better read ‘phonetic PAs/BTs’ or ‘pre-theoretical PAs/BTs’.

2.3 Network Input

The task of the neural network is to predict intonation with each of the three inputs (PAs, BTs and CONTs). The prediction runs syllable by syllable, from left to right. At each stage, the network knows the real f_0 value of the preceding syllable.

We are dealing with three types of input, of which PAs and BTs are perfectly comparable, whereas CONTs differ in several respects, mainly because the contour spreads over several syllables. Instead of taking the stress unit for the basic unit of input, we decided to maintain the syllable-by-syllable approach, but to enrich it in the case of CONTs by rough positional information, so that the network can see the relative position of the syllable within the stress unit. This is in fact not an artificial solution because in the case of PAs and BTs, the network also gets some kind of positional information (tonal annotation is present on specific syllables only), and it also holds in terms of the quantity of information provided, which is a prerequisite for transparent comparison (cf. the bits approximation below).

Along with tonal and, in the case of CONTs, positional data, the input also contains scaling information, consisting of the average f_0 over the intonation phrase, and the f_0 measured on the first syllable of the nucleus. This latter value was necessary to guarantee a good matching between the end of the pre-nuclear

part, which was the object of prediction, and the beginning of the intonation nucleus, which was not modeled but simply copied from the original sentence. The proportion between the two scaling values also serves as a rough coding of the overall declination.

The input parameters for each syllable are summarized in Table 1.

Table 1. Input parameters for an individual syllable in the three training/testing situations. In tonal targets, “0” means “no target” (a syllable without a tone).

	PAs	BTs	CONTs
Tonal information	- tonal target (H; L; S; 0) of the processed syllable		- contour of the stress unit to which the processed syllable belongs (H; L; HL; S)
Positional information			- relative position of the syllable within the stress unit (syllable number/ unit length)
Scaling information	- mean f_0 of the intonation phrase - f_0 of the first nuclear syllable		
History	- f_0 of the preceding syllable (does not apply to phrase-initial syllables)		

The predicted parameter is the f_0 of the processed syllable, as measured in the centre of the syllable nucleus.

In terms of bits of information (cf. the novel view of prosody as a channel for information coding in [6]), the input for each syllable can be approximated as displayed in Table 2 (scaling information and history, being identical for all three situations, are omitted here):

Table 2. Bit approximation of input parameters.

Type of input	Information	Total
PAs and BTs	1 tone out of 4, i.e. 2 bits	2 bits
CONTs	1 contour out of 4, i. e. 2 bits, but identical for the whole stress unit (average length 3.1 syllables), thus $2 / 3.1 = 0.65$ bits 1 syllable position out of average 3.1, i. e. $\log_2 3.1 = 1.63$ bits	2.28 bits

This rough estimation makes it evident that adding positional information compensates for the tonal underspecification of CONTs, making the comparison of the three situations more realistic. Despite the fact that the network can react differently to positional information which is implicit, as in PAs and BTs, and to that which is explicit, as in CONTs, it was hard to imagine any other satisfactory

solution. As it is with the setting described, CONTs have a slightly richer input than PAs and BTs (a difference of 0.28 bits for the differing part).

Additionally, we used two more situations which should serve as a reference delimiting the quality bottom and ceiling of our prediction: the NULL version (no tonal information, only positional information, scaling and history), and the ORIG version (original sentence).

2.4 Network Architecture

For the training, we used the Matlab software. We tried many different configurations of the network. We chose such a configuration, which gave satisfactory results provided it was simple enough (therefore it should generalize). Here we present detailed description of our final configuration.

The network was trained in 300 iterations. We ran every training 5 times and averaged the output in order to reduce individual diversions of a single run.

For PAs and BTs, the network was composed of 5 input neurons, and for CONTs, of 6 input neurons. The first three neurons represented three possible tonal descriptions of the current syllable (0 or 1), the next two represented scaling values (see Table 1). An extra input neuron with a relative position in the stress unit was added to CONTs. We used one hidden layer composed of 50 neurons. There was one neuron at the output. All pitch values at the input were converted into semitones to make the pitch changes independent of the pitch range.

Since we had a relatively small set of data, we used so-called “cross-validation” for the training and application of the network. We divided the data into thirds. We trained the network on 2/3 of the data and applied it to the remaining 1/3. This process was repeated 3 times with different thirds for application. Thus we obtained predicted values for the whole data.

2.5 Perceptual Evaluation

The 10 selected sentences were re-synthesized with the f_0 values obtained as the average of 5 runs of the network, for each of the four inputs (PAs, BTs, CONTs and NULL). The technique used was PSOLA under Praat [1]. In this way, we obtained 40 predicted items (10 sentences x 4 inputs). Values for nuclear stress units were not predicted, but copied from the original. This set was completed by the 10 unmodified sentences (ORIG).

The perceptual assessment of these sentences was carried out by means of an interactive application where each question contained two versions of the same sentence. There were two main possibilities as to how to formulate the question: either in terms of similarity with the original, or in terms of naturalness, independent of the original. We decided to adopt the second approach because it is less demanding for the listeners: instead of choosing a point on a scale, they only have to compare the naturalness of two same-worded sentences. Also, the question of similarity with the original is partly answered by the acoustic measures (see below).

Since using all two-term combinations of 5 sets of 10 sentences would have led to $10 * 10 = 100$ questions, which would have made the test extremely lengthy, we only retained all two-term combinations of PAs, BTs and CONTs (30 questions), and added 10 combinations of NULL with a balanced selection of PAs, BTs and CONTs, and another 10 combinations of ORIG with a balanced selection of PAs, BTs and CONTs. This augmented the number of evaluated pairs of sentences to 50. The order of sentences, speakers and versions was randomized.

The instruction was: “In each of the pairs of sentences, decide whether sentence A sounds more natural, equally natural, or less natural than sentence B. Only intonation should be taken into account.” This three-term choice, which includes the “same” answer, prevents the listeners from being categorical when they hear no or almost no difference, and makes the data richer because the “same” answers can be filtered when a more categorical approach is needed.

There were 31 respondents, all BA-level students of linguistics or modern languages with no hearing impairments, speaking Czech as their only mother tongue.

3 Results

In an informal inspection of the output, it turned out that the described network meets well with our expectations: in most cases, it reacts correctly to categorical tonal input (e. g., for an H tone in input, it predicts a rise), it generates declination, and it ensures a rather good fit with the nuclear contour. The predicted prenuclear intonation seems clumsy at certain points, but generally, it leads to a clear demarcation of words. Some sentences are nearly perfect. Generally speaking, the predicted intonation is somewhat flatter compared to the original, which is a product of the network learning as well as of the averaging over five runs. As expected, the NULL version is very monotonous. The nuclear parts, copied from the original, contribute rather strongly to the overall quality impression.

3.1 Numeric Evaluation of the Output

We evaluated the difference between the original f_0 shape and the predicted one in terms of correlation coefficients. Only prenuclear parts were taken into account. The results are displayed in Table 3.

Table 3. Performance of the prediction in acoustic terms.

Prediction type	f_0 correlation with the original
PAs	0.48
BTs	0.50
CONTs	0.53
NULL	0.37

The overall correlation seems to be rather poor. PAs, BTs and CONTs cluster around 0.50, with only slight differences between them. NULL has the worst score, as expected. Inter-speaker variability is relatively small, except for BTs, where the highest correlation is 0.64 and the lowest 0.32. For ORIG, the correlation would be, of course, 1.00.

3.2 Perceptual Evaluation of the Output

The perceptual assessment of PAs, BTs and CONTs altogether against NULL (prediction with no tonal information) and ORIG (original sentences) is schematized in Figure 1. The vertical division line in each rectangle expresses the preference ratio: the further it is from a variant, the better this variant was assessed, e.g. there was a 54% preference of PAs/BTs/CONTs over NULL. The error lines correspond to inter-speaker standard deviation.



Fig. 1. Performance of the prediction in perceptual terms – PAs/BTs/CONTs vs. NULL and ORIG.

There was a 70% preference for ORIG over PAs/BTs/CONTs, which indicates that the listeners were able to separate original sentences from the predicted ones. On the other hand, the NULL versions were assessed only slightly worse than PAs/BTs/CONTs. This shows, among other things, that the positional information alone is sufficient to predict acceptable intonation. However, if we compare this fact with the acoustic results contained in Table 2, we can conclude that the listeners were rather tolerant towards versions with low-variability intonation, as predicted without tonal information.

The perceptual assessment of pairs of versions (PAs, BTs and CONTs between them) is displayed in Figure 2.

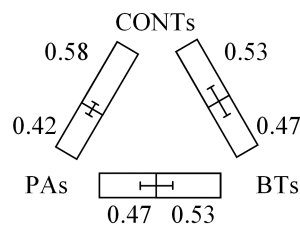


Fig. 2. Performance of the prediction in perceptual terms – PAs vs. BTs vs. CONTs.

The scheme shows that the only significant difference was found between CONTs and PAs, where CONTs were assessed better. The other pairs of versions show only insignificant preferences, which are however in line with the results contained in Table 2. To test the robustness of Figure 2, we calculated the same data a) without the “same” answers (i.e. we only counted answers “A is better” or “B is better”); b) without outlying listeners (exhibiting low correlation with average results); c) without listeners who were bad at telling the difference between ORIG and the other versions. In all three cases, the obtained values confirmed the setting displayed in Figures 1 and 2.

When examining speakers individually, one obtains the following orders of preferences:

- CONTs were assessed better than PAs in all speakers;
- CONTs were assessed better than BTs in 3 speakers out of 5;
- BTs were assessed better than PAs in 2 speakers, the same in 1 speaker, and worse in 2 speakers.

4 Conclusion

Before giving a general account of the results, we should point out possible sources of chaos and biases (some of which have been mentioned above):

- the most off-centre perceptual assessment is 0.70, expressing the preference of ORIG over PAs/BTs/CONTs; this indicates that the prediction led generally to rather acceptable results, which outranked the original sentences in 30% of the cases; the rest of the differences lies in the somewhat narrow interval 0.42–0.58;
- the listeners may have used different strategies to assess the output: among other things, it seems that they preferred flat intonation with tiny variations over rich excursions where the risk of sounding false is greater;
- the network processing may have been too gross to account for differences which are rather small;
- nuclear parts of intonation, identical with the original, increased considerably the quality of the output, and made the assessment less sharp;
- one should not forget that the input for CONTs was slightly overspecified (see section 2.3);
- generally speaking, the input was fairly rich (PAs and BTs: 0.64 tone per syllable on the average; CONTs: 0.39 tone per syllable on the average + positional information), which may have led to a ceiling effect.

Bearing all these facts in mind, one can summarize the results as follows:

- Surprisingly, within our methodology, Czech intonation can be learned with about the same success rate by means of PAs, BTs and CONTs. CONTs seem to be slightly more successful than the other two types of input, but this fact may be a product of a slightly richer input.

- With regard to the hypothesis that prenuclear intonation can be universally best described by means of PAs, our results speak in favour of a rehabilitation of CONTs as a traditional means of describing Czech intonation, as well as of the use of boundary tones as another possible local approach.
- However, we were not able to demonstrate within the given setting that either of the three approaches was really worse than the other two.
- Since our results may be affected by the method used, namely by the network function, we should test alternative methods, especially HMM, to validate or refine these preliminary conclusions.
- Our conclusions are relevant to the level of automatic learning. Implications regarding the processes of intonation coding and decoding in humans can only be indirect.

Acknowledgments. This research was carried out under the GAČR 405/07/0126 and the Information Society 1ET101120503 grants.

References

1. Boersma, P.: Praat, a system for doing phonetics by computer. *Glott International* 5 (9/10), 341–345 (2001)
2. Daneš, F.: *Intonace a věta ve spisovné češtině* [Intonation and the sentence in standard Czech]. Praha: ČSAV (1957)
3. Demenko, G., Wagner, A.: *The stylization of intonation contours*. *Speech Prosody* 2006. Dresden: TUD (2006)
4. Duběda, T.: Structural and quantitative properties of stress units in Czech and French. In: Braun, A., Masthoff, H. R. (eds.) *Festschrift for Jens-Peter Köster*. Stuttgart: Steiner, 338–350 (2002)
5. Hirst, D., Di Cristo, A. (eds.): *Intonation Systems. A Survey of Twenty Languages*. Cambridge University Press (1998)
6. Kochansky, G.: Prosody beyond fundamental frequency. In: Sudhoff, S. et al. (eds.) *Methods in Empirical Prosody Research*. Berlin – New York: Walter de Gruyter, 89–122 (2006)
7. Palková, Z.: Einige Beziehungen zwischen prosodischen Merkmalen im Tschechischen. *XIVth Congress of Linguists*, Vol. I. Berlin, 507–510 (1987)
8. Palková, Z.: *Fonetika a fonologie češtiny* [Phonetics and phonology of Czech]. Praha: Karolinum (1997)
9. Palková, Z.: The set of phonetic rules as a basis for the prosodic component of an automatic TTS synthesis in Czech. In: Palková, Z., Janíková, J. (eds.) *Phonetica Pragensia X*. Praha: Karolinum, 33–46 (2004)
10. Romportl, M.: *Studies in Phonetics*. Praha: Academia (1973)
11. Vaissière, J.: Language-Independent Prosodic Features. In: Cutler, A. (ed.) *Prosody. Models and Measurements*, 53–66 (1983)