Acoustic Analysis of Czech Stress: Intonation, Duration and Intensity Revisited

Tomáš Duběda[1] *& Jan Votrubec*[2]

[1]Institute of Phonetics, Charles University in Prague, Czech Republic
[2]Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic
dubeda@ff.cuni.cz, anna.kubatova@seznam.cz

10.21437/Interspeech.2005-509

Abstract

By examining acoustic marks of Czech stress, this paper attempts to provide an answer to the question of whether or not perceived accents in the Czech language have an objective existence. A neural network is used to predict the position of accents without lexical information. Three parameters (intonation, duration and intensity) are considered individually, in pairs and altogether. Fundamental frequency seems to be the best predictor of stress, both alone and combined with other parameters. The analysis of the individual prediction errors allows for a closer look at factors which are critical in accent prediction.

1. Introduction

The genesis of stress is generally motivated by a need to delimit words in continuous speech. Both intonational and rhythmic paths of accentogenesis have been considered, also, the iconic nature of stress has been pointed out [1], [2]. However, in some languages, stress as a phonological category seems to make less sense than in others, mostly because of its fixed position and weak acoustic correlates. This claim has been made e. g. for Amharic [3], French [4], [5], Hindi [6] and Telugu [7].

Czech is unanimously described as a language with word-initial stress [8]:

Český přízvuk má iniciální polohu.
[ˈt͡ʃɛskiː ˈpr̝̊iːzvʊkmaː ˈʔɪmɪt͡sɪjaːɫɲiː ˈpɔlɔɦʊ]
'Czech stress has an initial position.'

Yet, in current speech, Czech accents are often very weak and subtle. Among their potential suprasegmental correlates, the use of duration is largely inhibited by its phonological role in distinguishing short and long vowels (the length being completely independent of stress). Additionally, fundamental frequency demonstrates a much more complex behaviour than the tacitly expected peak on the stressed syllable, and, moreover, intensity does not rise on the stressed syllable. Segmental differences between stressed and unstressed syllables are negligible [8], [9]. In 1967, experiments carried out by Janota [10] seemed to demonstrate that all three prosodic parameters are likely to trigger the impression of accent in Czech, though in different proportions. However, this finding, based on observations of fully controlled, delexicalized synthetic speech, does not correspond with descriptions of accent in natural speech.

Despite claims against the very existence of stress in Czech, we can find positive, though indirect evidence which supports the existence of objective prominences in the language: (1) Stress can be easily demonstrated in emphatic speech, in scansion or in poetry, (2) stress is neither random nor arbitrary, and (3) Czech speakers are sensitive to wrong accent placement or its wrong realization in dialects and in non-native speech. Our pilot study [11], which, to the best of our knowledge, provided the first stochastic view of Czech stress, showed that the knowledge of normalized prosodic values (f_0, duration, intensity) over a 3-syllable frame is sufficient to predict the presence or absence of accent on the intermediate syllable with an accuracy of 80 % (agreement with a human listener). Left-only or right-only contexts provided much worse results than either-side contexts. The knowledge of a context larger than 3 syllables did not contribute significantly to the prediction score. Since accents are mostly located on the first syllable of a stress unit, the importance of the left context is evident. The syllable to the left of the accented one, being the last of the preceding stress unit, is responsible for most durational, dynamic and intonational contrast in the Czech language. This confirms the contextual hypothesis of Czech stress, which emphasizes global prosodic configurations of stress domains, as well as prosodic discontinuities between them [12].

2. Objective, corpus and methodology

2.1 Objective

Apart from verifying the results of the pilot study on a larger set of data, the present research also strives to delimit the role of individual prosodic parameters in accent prediction. Thus, the question underlying this article is twofold:

1. To what extent can Czech accents be predicted solely from prosodic parameters, without lexical information, by means of a neural network?
2. What are the respective roles of fundamental frequency (f_0), duration (d), and intensity (i) in accent prediction?

Considering the neural network as a reasonable copy of the human perceptual system, we test how the prosodic parameters provided in input are able to predict whether or not the syllable in question is accented. Stress assignment provided by a human listener is used as a reference. We assume that the neural network is able to detect any type of regularity (absolute or relative prosodic specification, positive or negative deviation, mono- or multiparametric characterization of stress, enhancement or compensation of the parameters, etc.).

2.2 Corpus

The recording analyzed was an informative text (970 syllables), read aloud by a male semi-professional speaker (in his twenties, standard pronunciation, average articulation rate 6,12 syll./s, average f_0 112 Hz). After a manual segmentation in sounds and syllables, the accents were perceptually assigned by the first author of this paper. 32 % of the syllables of the corpus were marked as accented.

2.3 Prosodic values and their normalization

The normalization of raw prosodic data was inevitable to grant a good performance of the network, given that it has no information about segmental structure. Since we only test prosodic parameters of stress, the exclusion of segmental variability is a methodological assumption here; in any case, segmental correlates of stress, as has been said above, are very weak in Czech. The basic unit of processing is the syllable. The measures and normalizations were carried out as follows:

1. duration: one value per syllable (ms), twofold normalization. i) duration of each segment weighted by its average duration throughout the corpus, so as to exclude intrinsic duration; ii) duration of each syllable weighted by the number of segments contained in it, so as to exclude variability in syllable structure.
2. intensity: one normalized value per syllable (dB). Average intensity of the nucleus, weighted by the average intensity of this segment, so as to exclude intrinsic intensity.
3. F_0: three values per syllable (Hz), taken in 20, 50 and 80 % of its duration ($f_0$1, $f_0$2, $f_0$3), no normalization.

In contrast to our pilot study, we used three intonation points per syllable instead of one, which gives access, within the limits of the present methodology, to finer intonation contours. Also, the duration normalization was more elaborate than in the pilot study.

To test the reliability of the normalization procedure, we calculated the correlation coefficients between the prediction score (cf. section 3) and the individual prosodic parameters in their raw form, as well as with the syllable structure:

ρ (prediction score; duration of the syllable) = 0,07
ρ (prediction score; mean intensity) = –0,03
ρ (prediction score; mean f0) = –0,04
ρ (prediction score; number of segments) = 0,01

Since all the coefficients are very close to zero, we can conclude that the network does not perform differently for different ranges of values, a result which validates the normalization method.

2.4 Training configurations and network architecture

The training input was:

1. prosodic description of the trained syllable
2. accent value of the trained syllable (0 – *unaccented* or 1 – *accented*)
3. prosodic description of the preceding syllable
4. prosodic description of the following syllable

By prosodic description, we mean the set of 5 values (d, i, $f_0$1, $f_0$2, $f_0$3) as described in the previous section. Pauses are formally taken for syllables, but with adjusted parameters (d = 250 ms; i = 0 dB; $f_0$1, $f_0$2, $f_0$3 = 0 Hz).

The training configurations cover all possible combinations of the three parameters used:

1. d, i, f_0
2. d, i
3. d, f_0
4. i, f_0
5. d
6. i
7. f_0

For each of the 7 situations, the parameter which is not mentioned was set to 0 in the input, so that it was excluded as a source of variability. For instance, in situation 1, we test the simultaneous effect of all three parameters on accent prediction, whereas in situation 5, we test the effect of duration variations only.

To amplify the amount of data, the input data were divided into three equally sized parts, out of which two were used for training and one for testing, in three possible combinations.

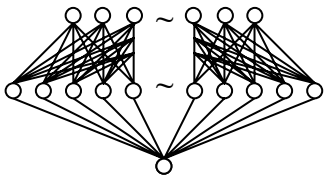

Fig. 1: The network architecture

As for the network structure, we used a back-propagation architecture with 15 input neurons (corresponding to 3 syllables, each described by 5 parameters), 30 neurons in the hidden layer, and 1 neuron in the output (accented +/–). The input values were weighted so as to fit into the interval 0…1. The output value, varying from –1 to +1, indicates the probability of accent as predicted by the network. For each of the training/testing configurations, we used 100 iterations with random initialization in 30 runs.

3. Prediction scores

The prediction score is defined as the proportion of matches between the rounded network prediction and the label provided by the human listener. If, for instance, the network predicts, for a given syllable, the value –0,3, we count this prediction as negative (*unaccented*). If this is in agreement with the decision of the human listener, the prediction score is 1, if not, then it is 0. The average value over the 30 runs of the network corresponds to the final prediction score. We did not display standard deviations or significance coefficients, assuming that a stochastic task in itself is, at least partly, a guarantee of systematic behaviour.

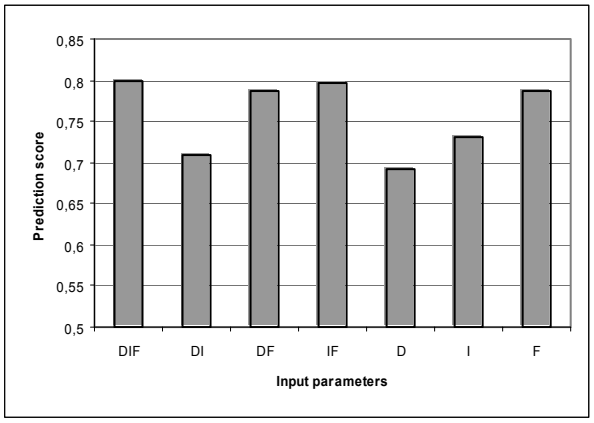

Fig. 2: Mean prediction score for seven different training/testing configurations (DIF – duration, intensity and fundamental frequency in input; DI – duration and intensity in input, fundamental frequency set to zero, etc.) The prediction score corresponds to the number of syllables whose accentual value was predicted in agreement with the human listener.

The results shown in Fig. 2 reflect the following facts:

1. The extremes of the prediction score are 69 % and 80 %. Since the default value for the network is *unstressed* (68 % of the syllables bear this mark), two of the seven values in the graph are highly fortuitous.
2. The four highest values correspond to situations where f_0 is present in input, whereas the three lowest scores were all attained without f_0.
3. F_0 alone is a good predictor of stress; the score does not change much when it is accompanied by either of the remaining parameters. The best score was achieved with all three parameters.
4. On the whole, the results confirm the logical behaviour of the network, and do not reveal any major problem. The fact that the two lowest scores were achieved with duration might indicate normalization problems, but the evidence in section 2.3 does not confirm this.

The pauses were predicted with a mean score of 80 %, which does not show a large difference from the overall mean, despite the fact that we expected better recognizability, according to "human" criteria. This point is probably not the only one where the network is not an exact copy of the human perceptual system.

4. Error analysis

The linguistic and acoustic properties of syllables in which we achieved the worse prediction scores may help explain what constitutes the problematic nature of Czech stress. The error analysis was performed for the three-parameter input, where the mean prediction score was 80 %.

Among the syllables which achieved a mean prediction score smaller that 50 %, there were 62 accent omissions (where the network did not see the accent), and 103 accent additions. These errors are rather uniformly distributed between accented and unaccented syllables (there are more unaccented syllables than accented, which means that the probability of error on an unaccented syllable, i. e. accent addition, is higher than the inverse).

Two contexts can be expected as particularly likely to perturb the network:

1. Adjacent accents, e. g. *dva tisíce* [ˈdva ˈcɪsiːt͡sɛ] 'two thousand'; 6 cases attested in the present corpus.
2. Anacruses, i. e. unstressed words preceding the accented syllable in a stress group after a pause; this situation diverges from the default pattern where every post-pausal syllable bears accent, e. g. *když odešel* [gdɪʃ ˈʔɔdɛʃɛl] 'when he left'; 16 cases attested in the present corpus.

In the first case, both syllables bearing adjacent accents were marked as specific; the average prediction score for these syllables is 84 %. In the second case, we marked all anacrusis syllables (mostly one, in some cases two, and in very few cases three) as well as the accented syllable; the mean prediction score was 88 %. In both situations, the network does not seem to have major problems with accent prediction. A closer auditory inspection of these contexts shows that the accents are marked with more emphasis, which might indicate a tendency of the speaker to anticipate such situations. Also, there is a universal tendency toward better prosodic specification after a pause. It is important to notice at this point that the prediction for each syllable is made independently of the previously assigned accent values, which is obviously not the case in speech perception by humans.

To obtain additional numerical indices of positions where the network encountered problems, we calculated correlation coefficients between the prediction score and a set of positional and structural descriptors. These descriptors are:

1. final syllable in the stress group; $\rho = 0{,}06$
2. first syllable in a stress group longer than 1 syllable (to avoid syncretism with point 1); $\rho = -0{,}05$
3. non-boundary syllable (other than initial of final, in stress groups longer than 2 syllable, to avoid syncretism with points 1 and 2); $\rho = -0{,}0001$

All the coefficients are very low, which does not allow us to determine, within the performance limits of the network used, critical positions for stress assignment.

A more detailed analysis of the 40 worse prediction scores, also taken from data obtained with a three-parameter input, shows the following distribution:

A) 15 accent omissions:

1. 10 in normal context. These cases indicate the weak acoustic salience of some Czech accents. These accent lapses would probably be at least partly eliminated if the network were able to make rhythmic predictions. A certain number of these cases turned out to be really weak when checked auditorily.
2. 4 after pause. This situation deviates from the unmarked character of post-pausal syllables. No auditory traces of abnormal realization were found.
3. 1 after adjacent accent

B) 25 accent additions:

1. 10 on a word-final syllable followed by a stressed syllable. 5 of these additions correspond to the word-final vowels [ɪ] or [iː], which may be due to normalization problems in duration or intensity. Most of these cases show no significant final lengthening or major intonation contours.
2. 6 on a word-interior syllable. When checked auditorily, these cases show weak though well detectable accent on the first syllable.
3. 4 on a word-initial syllable where the accent had not been heard. The accent prediction made by the network is not seen as totally false when checked auditorily.
4. 3 on a pause
5. 1 on a grammatical word
6. 1 on a word-final syllable before pause

5. Conclusions

The presented stochastic data and the subsequent error analysis has shown the following tendencies:

1. Despite more complex normalization and the use of three intonation values per syllable instead of one, the overall prediction score (with all three parameters in input) did not exceed the result of our preliminary study [...].
2. The prediction scores show the relative importance of the individual prosodic parameters in accent characterization. Fundamental frequency is the strongest predictor according to our data, both alone and in combination with other parameters.
3. The prediction errors, analyzed for the 40 worst prediction scores, can be roughly classified into those which are due to network problems (as in A1 or B4 above), those which are probably triggered by normalization inadequacy (as in B1) and those which are at least partly in agreement with human perception (as in A1 or B3).
4. Hypothesized critical contexts for accent prediction (i. e. anacruses and adjacent accents) did not generate worse than average results.
5. In accordance with the present data, obtained by the analysis of standard read Czech, the no-accent hypothesis has been rejected once again.

Possible areas of further research include comparative studies (comparison with other languages with "weak" stress), as well as studies focusing on technical optimization of the network.

6. Acknowledgements

This research was supported by the GAČR 405/04/P238 grant.

7. References

[1] Hyman, L. M., "On the nature of linguistic stress", in: *Studies in Stress and Accent, Southern California Occasional Papers in Linguistics 4*, ed. L. M. Hyman, Los Angeles, California, pp. 37–82, 1977.

[2] Duběda, T., *Jazyky a jejich zvuky*. Carolinum, Praha, in print.

[3] Bell, A. (1977) Accent placement and perception of prominence rhythmic structures, in: *Studies in Stress and Accent*, Southern California Occasional Papers in Linguistics 4, ed. L. M. Hyman, Los Angeles, California, pp. 1–14

[4] Vaissière, J. (1991b) Rhythm, accentuation and final lengthening in French, in: *Music, language, speech and brain*, eds. J. Sundberg – L. Nord – R. Carlson, pp. 108–120

[5] Rossi, M. (1979) Le français, langue sans accent?, in: *L'accent en français contemporain*, Didier, Ottawa, pp. 13–52

[6] Ohala, M., "Hindi", in : *Handbook of the IPA*, CUP, 1999.

[7] Lisker, L. – Krishnamurti, B., "Lexical stress in a 'stressless' language: judgements by Telugu- and English-speaking linguists", in: *Actes du XII^{e} Congrès international des sciences phonétiques*, Aix-en-Provence, Vol. 2, pp. 90–93, 1991

[8] Palková, Z., *Fonetika a fonologie češtiny*, Karolinum, Praha, 1994

[9] Duběda, T., "Structural and quantitative properties of stress units in Czech and French", in: *Festschrift for Jens-Peter Köster on the Occasion of his 60th Birthday, Phonetics and its Applications*. Angelika Braun/Herbert R. Masthoff (ed.), Stuttgart: Steiner, 2002.

[10] Janota, P., "An experiment concerning the perception of stress by Czech listeners, in: *Acta Universitatis Carolinae, Phonetica Pragensia*, Charles University, Prague, pp. 45–68, 1967.

[11] Duběda, T., Votrubec, J., "Acoustic Correlates of Stress in Czech: a Stochastic View", *14th Czech-German Workshop*, ed. R. Vích, AV ČR, Praha, pp. 36-43, 2004

[12] Palková, Z., Einige Beziehungen zwischen prosodischen Merkmalen im Tschechischen", in: *Proceedings of the XIVth Congress of Linguists*, Vol. I., Berlin, pp. 507–510, 1987.