Clause position within a sentence: human vs. machine recognition

Zdena Palková & Jan Volín

Institute of Phonetics
Charles University, Prague
zdena.palkova@ff.cuni.cz

Abstract

The paper presents a combined experiment in which recognition of a prosodic phrase position within a larger syntactic structure by human listeners is confronted with recognition by artificial neural networks. Apart from the success rate we are predominantly interested in similarities in the error pattern of the two recognition modes. The results suggest that the automatic recognition could help to determine which of the selected parameters are relevant for human listeners, since it provides linguistically interpretable outcome.

1. Introduction

1.1. Theoretical background

A substantial portion of the current interest of linguists in general properties of verbal communication and in discourse analysis is devoted to the aspects of sound patterns in speech and their systematic exploitation in the process of information transfer. In this respect, it is predominantly the intonation structure that is in the centre of attention (see, e.g., [1], [2]). Recent developments in intonology have indicated that rather than description of isolated sentences there is nowadays a critical need of analyses of larger spoken texts.

The fundamental frequency changes are utilized in the phonological structure of languages on the post-lexical level as a functional device in two basic ways. First, they complement or modify messages coded in the text by adding further communicative (discoursal) meanings. Such modifications can be fully grammaticalized (e.g., syntactic/ pragmatic notions as finality, interrogation, or even attitudinal modality, etc.), but they can also belong to a considerably more diverse choice of expressive and emotional means (e.g., [3], [4]).

Second, they guide the listener through the linearity of the spoken text. In this sense the pitch changes break the text into cognitively manageable units and hint at the relationship among the units by either delimitative means or by cohesion signals ([3] esp. Chapter 5, or [5]: p. 28 - Contextual effects; p. 31 - Focalization and emphasis, etc.).

Intonation in Czech exploits all the above-mentioned possibilities. So far, attention has been paid mainly to its role in attitudinal modality marking, in anaphoric focalization, and in linear segmentation of spoken texts. Systematic research has indicated that intonation can also facilitate orientation of the listener in longer continuous texts either by signalling boundaries of various strengths or by cues that refer to the position of a unit within another, hierarchically higher unit (e.g., the position of a stress-unit within a tone-unit or a tone-unit within an utterance [6]). This function of intonation assists the listener to follow the development of the text, whether it is a course of a narrative or a line of reasoning.

Differences among intonation contours which are dependent on their position can be traced within units of various lengths, i.e., even above the level of the utterance [7]. Besides the traditional interest in unit-final contours, which in Czech, like in many other languages, can serve as carriers of syntactic/pragmatic functions, there is also a need to analyse characteristics of intonation contours in non-final positions. Therefore, we set out to investigate whether a clause extracted from its context carries enough acoustic cues so as to allow its specification as sentence-initial, sentence-medial, or sentence-final.

1.2. Methodological approach

The analysis of longer spoken texts requires specialized research methods. In order to ascertain the presence or the absence of differences in characteristics of intonation contours with respect to their position in a larger context, and to verify their perceptual validity for the percipient, one has to perform analyses which are demanding not only with respect of the amount of the data required, but also with respect of the number of parameters that may affect the perceptual process. When analysing longer texts, the number of parameters that have to be taken into account increases while the certainty in determining their effect can actually decrease.

To solve this methodological problem we suggest considering a comparison of results of perceptual test collected from human language users with results provided by statistically based analyses of carefully selected speech signal parameters. It is specifically artificial neural networks which allow for stipulating and verifying hypotheses with a large number of input variables.

Similarity or differences in results from both sources might provide relevant information about the adequacy and appropriateness of the selected parameters, especially with regard to the percipient's point of view. The most revealing aspect might be the analysis of the error patterns: even though humans and machines might easily use completely different strategies and still arrive at quite similar success rate, it is quite unlikely that different strategies will lead to the same error pattern.

2. Experiments

2.1. Material

The core of the material comprised 16 syntactically well-formed clauses, each of which occurred in a larger complex sentence in an initial, medial, or final position. Three parallel narratives with analogous content were created to nest always just one of the variants (i.e., initial, medial, final) of the clause at a time. Each of the sixteen clauses consisted of four stress-units. The size of the stress-units varied but the clause, realized usually as one tone-unit, was always 10 syllables long. For example, the clause …*protože otce výborně znala*…

(…*because she knew her father very well*…) occurred sentence-initially: *Protože otce výborně znala, nebála se, že by ji opravdu vyhnal* (*Because she knew her father very well, she was not afraid that he might really expel her*.), sentence-medially: *Princezna se ale nebála, protože otce výborně znala a uměla ho vždycky obměkčit* (*But the princess was not afraid, because she knew her father very well, and was always able to talk him in*.), and sentence-finally: *Princezna se jen smála a nebrala hrozbu vážně, protože otce výborně znala* (*The princess just laughed and did not take the threat seriously, because she knew her father very well*).

The narratives were read by three non-professional speakers (two women and one man) in a sound-treated booth. The procedure produced 144 items: 16 clauses in 3 positions read by 3 speakers.

2.2. Method

The above-mentioned 144 items were extracted from the recordings of the texts and they were used to compile two parallel, virtually identical perceptual tests. The reason for compiling two rather than just one test was to avoid the respondents' fatigue. The tests were administered to small homogeneous groups of native Czech listeners. In total, 117 respondents took part: 59 were presented version A of the test, 58 listened to version B. The respondents were presented individual items in a sound-treated room and were asked to decide whether the phrase they heard was an initial, medial of final part of a longer utterance. They marked their decision in an answer sheet. The information about the extent of the material and the number of subjects is summarized in Table 1.

Object	*Quantity*
clauses explored	16
positions in a sentence	3
speakers	3
test items	144
respondents	117
respondents' judgements	8420

Table 1: *Basic characteristics of the material.*

As to the artificial neural network (ANN) predictions, we based our analysis on the properties of individual stress-units. The idea that the F0 course in a stress-unit might be the fundamental element of the intonation description of the Czech tone-unit is discussed in [8]. Therefore, at this stage of the research, the parameterisation of the melody in the clauses is built on the properties of individual stress-groups disregarding their mutual relationship. In other words, the relationship between two consecutive stress-units is currently not taken into account. However, the variable which we called *register* captures the difference between the mean F0 within the given stress-unit and the speaker's overall mean F0.

The other input variables were *range* and *contour*. The *range* was based on the standard deviation of F0 data points in the given stress-unit from the mean. In order to make the range variable more intuitive, we multiplied the standard deviation by 2 × 1.96. This transformation ensures that 95 % of the F0 values are included in the parameter. The *contour* was a nominal variable. It was based on the F0 measurements of the centres of syllabic nuclei. The stress-units consisted of two to four syllables. The increase in F0 values by more than 2 percent from one syllabic nucleus to another was considered a rise, the opposite was a fall. The changes that were smaller than 2 percent were labelled as level contours. Thus, each of the stress-units was assigned a label from the following invetory: fall, rise, level, rise-fall, and fall-rise. Since there were four stress-units in each of the clauses, the input vector comprised 12 elements: 4 x *register*, 4 x *range*, and 4 x *contour*. We used perceptron network with the architecture of 12:24-13-3:1 (Random ½ of the data training set, ¼ testing set, ¼ validation set, for final validation reshuffled.)

3. Results

Table 2 shows the correct recognition of individual clause positions by human listeners. Interestingly, there is quite a high success rate not only for the sentence-final positions, where the correct recognition was expected, but also for the initial and medial positions. Actually, the success rate for the difference between the two non-final types was considerably higher than we anticipated.

%	*I obs*	*M obs*	*F obs*
I pred	77.1	27.1	0.0
M pred	22.9	54.2	0.0
F pred	0.0	18.8	100.0

Table 2: *Confusion matrix for recognition by human listeners. The values are in percentages. I - initial, M - medial, F - final position; obs - observed, pred - predicted.*

Table 3 presents the results of the artificial neural network (ANN) predictions. The success rate, which is based on three selected variables for each of the four stress-units in a clause (i.e., 12 information entries per case), is comparable, though not identical with the predictions made by human listeners.

%	*I obs*	*M obs*	*F obs*
I pred	77.1	22.9	0.0
M pred	14.6	52.1	12.5
F pred	8.3	25.0	87.5

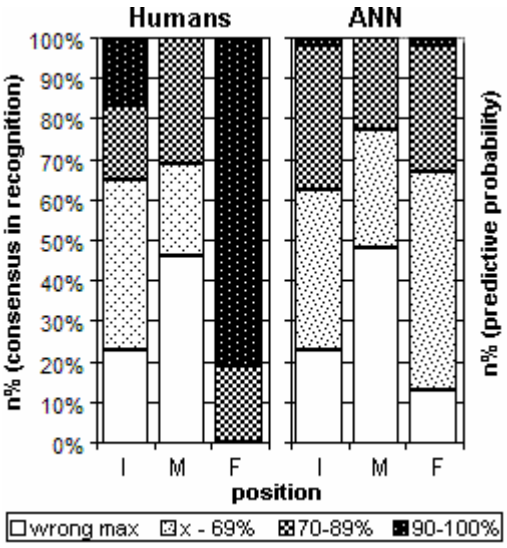
Table 3: *Confusion matrix for prediction by ANN. The values are in percentages. Abbreviations are the same as in Table 2.*

First of all, human listeners are clearly superior in recognizing the final position of clauses. This fact is actually also responsible for the difference in the overall (total) success rate: 77.1 % for humans and 72.2 % for the ANN. Second, neither the machine nor humans confuse the final clauses for the initial ones. Human listeners are in additon generally better at not confusing the initial clauses for the final ones, and the final ones for the medial ones.

It has to be remembered, however, that the confusion matrices in Tables 2 and 3 are based on winning candidates: the recognition of an item is based on the probability ratings for the three positions with the highest probability determining the outcome. Although this method is undoubtedly informative and is commonly used, it masks certain aspects of the recognition task. For example, the one-hundred-percent success rate for final position in human recognition hides the fact that 150 judgements (5.4 %) concerning final clauses were wrong. Some people actually thought they were listening to a sentence-medial or even sentence-initial clause

when they were presented a sentence-final one. In other words, Tables 2 and 3 do not show the confidence with which the ANN or humans as a group opted for the given solution. This problem is dealt with in the following paragraphs.

Graph 1 compares the increasing consensus of listeners in correct recognition of an item with the ANN probabilities of correct predictions. It provides an even more explicit evidence than Tables 1 and 2 that the final position is considerably easier to recognize for humans than for the ANN. On the other hand, there is practically no difference in recognition of medial positions and only a minor difference in determining initial positions.



Graph 1: *Comparison of the listeners' increasing consensus in correct recognition with the ANN predictive probabilities.*

From the human point of view, the worst recognition of final positions still had a consensus of more than 70 % (specifically 74 %). Most of the final items were recognized with more than 90% confidence. In the case of ANN, such high probability is limited to only a very narrow band that does not differ from the situation in initial positions. Human recognition of the initial position again displays the highest degree of confidence for nearly 20 % of the items. However, if we ignore the division into the two top probability/confidence bands, we can see that the performance of human listeners is very similar to that of the ANN. This also holds for the medial position where both modes of recognition lack the band of items recognized with more than 90% confidence/probability. Yet the human listeners are by about 9 % more successful in the 70-89% confidence/probability band.

Table 4 shows how individual variables contributed to the performance of our ANN. The most important variable was the contour in the fourth stress-unit (Cont4th). The first four (one third) most important variables also included contours in the first and third stress-units, and the mean F0 relative to the effective voice range of the speaker in the third stress-unit (Regist3rd). On the other hand, our current model did not make much use of the information concerning the intonation contour in the second stress-unit (Cont 2nd), and pitch range descriptors in the first and fourth stress-units.

As to the incorrectly recognized items, we found 17 cases where human listeners erred in the same direction as our ANN recognizer. Knowing that human listeners recognized wrongly 33 out of 144 items, we can see that one half of their errors were identical with the errors made by the ANN recognizer.

More specifically, both humans and the ANN labelled two initial clauses as medial, six medial clauses as initial, and nine medial clauses as sentence-final. These 17 item can provide valuable information about properties that are relevant to the two modes of recognition.

Variable	*Importance*	*Sensitivity*
Cont 1st	3	1.10
Cont 2nd	12	0.77
Cont 3rd	4	1.10
Cont 4th	1	1.20
Range 1st	10	0.98
Range 2nd	6	1.03
Range 3rd	8	1.00
Range 4th	11	0.98
Regist 1st	9	0.98
Regist 2nd	7	1.03
Regist 3rd	2	1.15
Regist 4th	5	1.06

Table 4: *Sensitivity analysis results for the trained ANN. Importance of individual variables is ranked. Higher sensitivity values signal greater contribution to the performance. (Variables: Cont1st - contour in the first stress-unit ... Regist4th - register in the fourth stress-unit.)*

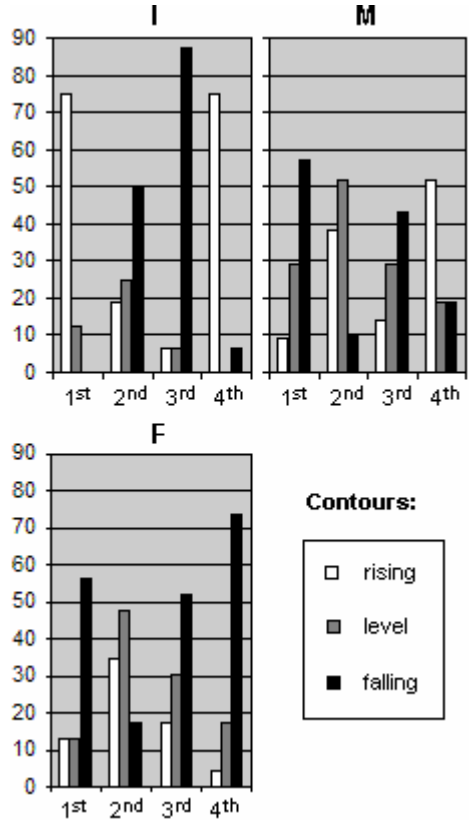
4. Discussion

The sensitivity analysis suggested that the shapes of contours in stress-units 1, 3, and 4 played an important role in the recognition process. We believe that the role of the contour in stress-unit 2 is also important, but its importance is relational: the contour is not important per se, but in relation to the first and the third contours. Relational approach will be our task in the next stages of our research.

It is still interesting to take a closer look at the occurrence of the individual types of contours in the items with high consensus of the listeners about the correct recognition and in items that were recognized wrongly. That is facilitated in Graphs 2 and 3. (For the sake of clarity, we present only rising, falling, and level contours. There was also marginal presence of rising-falling (8 %) and falling-rising (2 %) contours. This, however, was rare and was limited only to the position of a nuclear tone.)

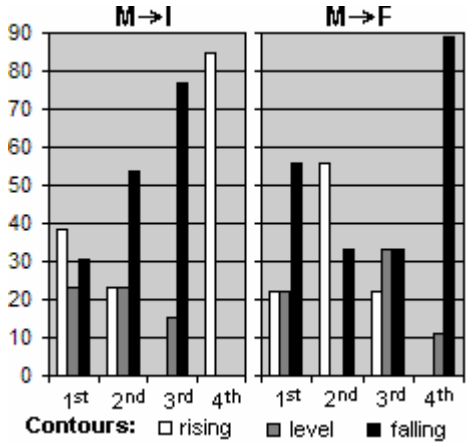
Graph 2 shows frequency of occurrence of the three most common types of contours in stress-units of clauses in initial (I), medial (M), and final (F) positions. The graph is based on contours that were recognized with relatively high confidence. The word 'relatively' refers to the consensus typical for the given position (see above, Graph 1). Thus for the initial position it was the band of 75-100 %, for the medial position it was the band of 60 % and higher, and for the final position it was the band of 97-100 %.

Graph 2 indicates that position I displays a typical rise in the 1st and 4th stress-units, while the 2nd and especially the 3rd stress-units fall. Positions M and F are similar to one another in composition of the first three stress-units, but there is a difference in the 4th stress-unit: position M ends with a rise, whereas position F ends, as expected, with a fall.



Graph 2: *Relative occurrences (in %) of the three most common types of contours in items with high consensus in correct recognition. The values on the ordinate give percentages of occurrence; the abscissa marks the individual stress-units.*

Graph 3 shows the relative occurrence of contours in medial items that were wrongly recognized as either initial or final both by human listeners and ANN. The left part of the graph (M→I, i.e., M recognized as I) indicates that these items had a greater number of rising contours in the 1st stress-unit, and falling contours in the 2nd and 3rd stress-unit (Compare with part I and M of Graph 2).



Graph 3: *Relative occurrences (in %) of the three most common contours in erroneously recognized items. The ordinate gives percentages of occurrence; abscissa refers to individual stress-units.*

The right part of Graph 3 (M→F) shows that the medial items that were erroneously recognized as final had a considerably high number of falling contours in the 4th stress-unit (Compare with part M and F of Graph 2). Another difference is the prevalence of the rising contours in the 2nd stress-unit. Whether that is a relevant finding remains to be seen in the future research.

Graphs 2 and 3 are in agreement with the ANN sensitivity analysis as to the importance of the shapes of intonation contours in the 1st, 3rd, and 4th stress-units. However, the importance of the 2nd stress-unit seems to be also significant especially for the initial position recognition.

5. Conclusions

The results of our experiments confirmed the existence of intonation characteristics which support orientation of listeners in the linear design of a longer structured spoken text. Human recognition finds sufficient amount of prosodic cues to determine the position of clauses within a higher syntactic unit even if these clauses are extracted from their contexts.

ANN predictions exhibit a satisfactory success rate for all the three positions while using quite a limited set of variables. It seems that the selected intonation characteristics are relevant for the given recognition task. They also support the hypothesis that F0 course in a stress-unit provides a convenient base for the intonation description of the Czech clause.

There are still substantial differences between the human perception and the ANN performance, especially with regard to the degree of certainty in recognition of the three positions (I, M, and F). It has been confirmed that the stochastic modelling is difficult to interpret without comparisons with the result from perception tests. Particularly, the differences in error patterns might hint at further directions of the research.

The future research will require a substantial extension of the number of variables to see if a closer match between human evaluation and computational recognition is possible.

6. Acknowledgement

This research was supported by grants GACR 405/05/0436 and VZ MSM0021620825

7. References

[1] Coulthard, M., 1992. The significance of intonation in discourse. In *Advances in spoken discourse analysis,* M. Coulthard, (ed.), Routledge, London - New York, 35-49.

[2] Brazil, D., 1997. *The communicative value of intonation in English*. Cambridge: Cambridge Univ. Press.

[3] Ladd, D.R., 1996. *Intonational phonology*. Cambridge: Cambridge Univ. Press.

[4] Gussenhoven C., 2004. *The phonology of tone and intonation*. Cambridge: Cambridge Univ. Press, p. 22

[5] Hirst, D.J. and Di Cristo, A., 1998. *Intonation systems*. Cambridge: Cambridge Univ. Press.

[6] Palková, Z., 2001. Positional relevancy within tone units. In *Proceedings of LP'2000*, B.Palek, O.Fujimura, (eds.) Praha: The Karolinum Press, 131-146

[7] Wichmann, A., 2000. *Intonation in Text and Discourse*. Harlow: Longman, Pearson Education Limited.

[8] Palková, Z. and Janíková, J., 1999. Positionally determined differences in F0 patterns validity in Czech. In *Proc. of XIV ICPhS*. San Francisco: Congress Org., 1545-1548.