The role of F0 contours in determining foot boundaries in Czech

Zdena Palková and Jan Volín
Institute of Phonetics, Charles University in Prague
E-mail: zdena.palkova@ff.cuni.cz

ABSTRACT

The definition of the stress-group or foot is often derived from the lexical stress prominence. However, in Czech no regular pattern in prosodic properties bound to the stressed syllable has been proven. The more realistic approach seems to be to base the unit delimitation on the linear course of the overall sound properties throughout the unit. We investigated the role of the F0 contour in this respect. Experiments were built on both natural and synthetic speech material in which division into feet determines the meanings. Alternative boundaries, and one-foot vs. two-foot solutions within identical syllable chains were examined. Analyses of F0 tracks and extensive perceptual experiments provide information about the degree of sound - meaning influence and confirm the relevance of the F0 contour for the inner cohesion of the foot.

1. INTRODUCTION

The indispensable step that has to be taken in connection with the description of the prosodic structure of a language is the choice of suitable descriptive units and their hierarchy. The definitions of these units are consequently based on the relationship between the segmentation of the speech continuum and sound prominences. This relationship is not necessarily identical for units at different levels and may vary for different languages.

The unit that is commonly used in intonation languages at the word level is the *stress-group* or *foot* while the prominence is referred to as *stress*. Traditionally, the unit is derived from the prominence, which is an adequate approach in many languages. The stress plays a pivotal role and its distribution and properties are studied in scientific analyses. The linear unit itself, the foot, is formally defined with varying degree of respect to the word boundary.

However, practical applications of this approach in the analysis of natural speech or in speech synthesis run into difficulties if the description is not based on sufficiently specified sound properties. As to the description of Czech, the problem arises when we want to specify the objective sound qualities which are supposed to signal the indispensable prominence.

In the sense of traditional structural description, the stress in Czech is fixed on the first syllable of a word. The stress contrast is not reflected on the segmental level: the components of unstressed syllables are reduced neither in their quality nor in their quantity. Word boundaries often coincide with foot boundaries. Co-occurrence of more words within one foot is caused by the existence of monosyllabic words, which hardly ever form feet on their own. The attempts to find any stable characteristics of the first syllable in the domain of F T I properties, whether in relation to the surrounding syllables or within the stressed syllable as such, have failed so far [1], [2]. Outcomes of various experiments have indicated that in case of Czech it is much more useful not to rely on recurrent prominences, but to base analyses on the linear course of the overall sound properties throughout the foot [3]. The adequacy of such approach has been confirmed through its successful application in the systems of TTS synthesis [4], [5]. The objectives of the current research are to detect cohesion supporting features on the one hand and sound qualities leading to the segmentation into feet (cohesion prohibiting features) on the other hand. The significance of the durational and F0 course seems to be unquestionable. Our paper presents major findings of several test series investigating the role of various types of F0 contours in segmentation of the speech flow into feet.

2. METHOD

2.1. Materials and testing

In all its parts, the experiment is based on the speech material in which segmentation into feet determines the meaning. We focused on 2, 3, and 4-syllable feet since in Czech these occur most frequently and as to the sound properties they are probably most variable. Series A of the tests contained material in which a given sequence of syllables could be divided into feet in two different ways (e.g. /svjetlo/vɲiːmajiː/ vs. /svjetlovɲiː/majiː/, meaning ‘*they perceive light*’ vs. ‘*they have light in it*’), while in the series B a given sequence of syllables forms either one or two feet (e.g. /proci/vɲejʃiːm/ vs. /procivɲejʃiːm/, meaning ‘*against external’* vs. ‘*more bothering*’). For both types we created tests based on natural (N) and synthetic speech (S).

Natural speech items (N-items) were extracted from longer texts, which were read by non-professional speakers (6 male and 8 female in 4 tests). N-items were always located in the middle part of a sentence outside the intonational nucleus. Synthetic speech items (S-items) were based on the hypotheses drawn from natural speech tests and the only manipulated feature was the fundamental frequency.

In perceptual tests, naïve listeners were asked to match the presented item with one of the two possible meanings. Tests A contained the ambiguous sequence of syllables on its own, tests B and S-item-based tests presented the ambiguous sequence in a short context suitable for both possible meanings. 50-100 listeners took part in each of the four tests. The groups of listeners were homogeneous as to their age, education and motivation.

2.2. Measurements

The analysis in the whole set of items was primarily focused at the course of the F0. However, the basic data concerning the amplitude and duration were registered as well. Part of the material was analyzed with the help of the MultiSpeech software, the whole material was later reanalyzed with the Praat package. The results gained by the two analysis tools were very close, which we consider a positive sign.

The F0 track was schematized, so that each syllable was represented by one F0 value. This value was chosen in a standard way from the central part of the syllabic nucleus in melodically stable syllables and from the second third of the syllabic nucleus with substantial F0 changes, provided the dynamics was not dropping considerably. These representative values were linked into stylized contours, which formed the basis for the assessment of the listener's judgements about parsing.

3. RESULTS

3.1. Listeners' agreement with speakers' intention

Results based on N-items provide some insight into the listeners' dependence on the sound in recovering the meaning in an extreme linguistic situation. Table 1 shows the agreement between the listeners' judgements about segmentation and the speakers' intentions. Altogether, there were 120 AN-items and 136 BN-items representing 10 different sequences with alternative boundary placement and 13 sequences with one-word vs. two-word option. As to the foot length, the material contained 218 2-syllable feet, 136 3-syllable feet, 74 4-syllable feet, 8 5-syllable and 8 6-syllable feet.

	1	**2**	**3**	**4**	**5**	**6**	**7**
AN1	10	3	55	19	74	10	16
AN2	10	3	32	29	61	11	29
BN1	8	4	61	21	82	6	13
BN2	9	4	34	42	76	11	13
$\bar{x}$			**46**	**28**	**73**	**10**	**18**

Table 1: The agreement between the listeners' judgements about segmentation and the speakers' intentions in four perception tests. Column 1 gives the number of tested sequences, col. 2 number of speakers. Col. 3 gives cases (%) in which agreement between listeners' judgements and speakers' intentions was 80-100%. In col. 4 it is 60-80%, in col. 5 it is 60-100%. Col. 6 gives cases (%) in which listeners decided against the intentions of the speakers. Column 7 represents cases where listeners did not reach common stand.

The results show that division of the speech continuum into words is equally dependent on the sound and the meaning. The sound itself helped to recover the speaker's intention in 46% cases on average (agreement 80-100%). The cases of disagreement between the listeners' judgements and speakers' intention occurred in each of the tests. They pinpoint the influence of the sound context that is longer than the modulation inside the given sequence. Closer analysis of the data revealed the expected differences in the clarity of the individual speakers' production.

3.2. Listeners' judgements and the sound modulation in N-tests

It is not surprising that in our tests we found no significant correlation between the recurrence of dynamic changes and the feet patterning. An explicit example is provided by the results in the B-tests, where two vs. one-word solutions are contrasted. Only cases with the listeners' agreement of 80% or more are taken into account. Tab. 2 shows cases (in %) in which the first syllable of the second part of the sequence was dynamically stronger (>), weaker (<), or equal (=) to the first syllable of the whole sequence if the sequence was perceived as one word or to the preceding syllable if the sequence was perceived as two words. (Changes of at least 1dB were considered. Obviously, certain changes may be caused by intrinsic qualities of individual phones.)

	>	=	<		>	=	<
one-word	52	26	22	two-word	27	37	37

Table 2: Cases (in %) in which the 1st syllable of the second part of the ambiguous sequence was dynamically stronger (>), weaker (<), or equal (=) to the 1st syllable of the whole sequence if the item was perceived as one word or to the preceding syllable if the item was perceived as two words.

The figures show clearly, that the division of a sequence into two words by means of the greater intensity at the beginning of the second part happened in only less than one third of the cases. Meanwhile, however, greater intensity inside longer words in one-word solutions occurs in more than one half of the cases. This is related to the crescendo tendency in the mid part of longer feet recently discovered in spoken Czech.

Studying listeners' judgements against the F0 contours is much more revealing and it brings relatively stable results. The most information value can be attributed to the items with high listeners' agreement, but also to the items that were perceived contrary to the intention of the speaker. We may hypothesize that sequences which the listeners parse differently are also different in their F0 course. We can distinguish melodic contours that are *cohesion supporting*, *cohesion neutral*, and *cohesion prohibiting*.

Examples of contours with the typically high degree of agreement between the listeners can be seen in fig. 1 and 3 for 3:2 division and fig. 2 and 4 for 2:3 division (tests A), fig. 5 and 7 for two-word solutions and fig. 6 and 8 for one-word solution, i.e. cohesion of the foot (tests B).

The strongest tendency found in our material is a negative one. The contour with clear F0 drop in its middle does not represent an acceptable form of a foot. This is supported by findings in both A and B-series of the tests. The syllable with the lowest F0 is usually the initial syllable of the

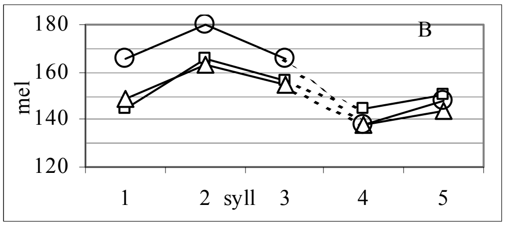

Fig. 1: AN, 3:2 ≥ 80, Speaker B

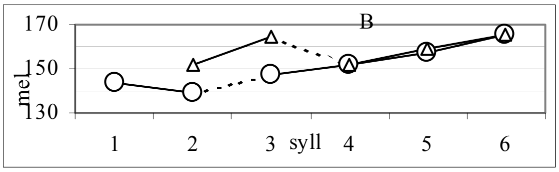

Fig. 2: AN 2:3,2:4 ≥ 80%, Speaker B

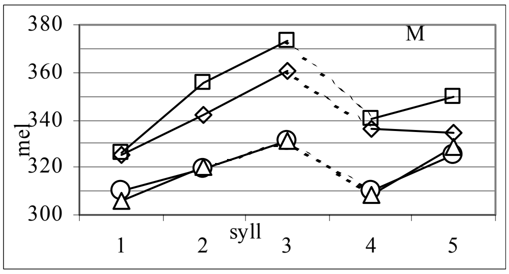

Fig. 3: AN, 3:2 ≥ 80%, Speaker M

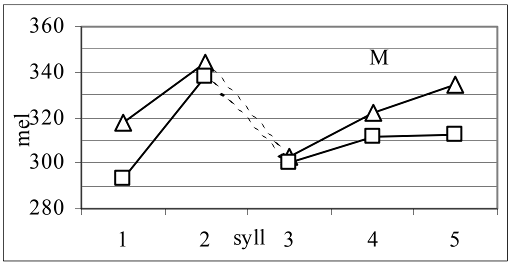

Fig. 4: AN, 2:3,2:4 ≥ 80%, Speaker M

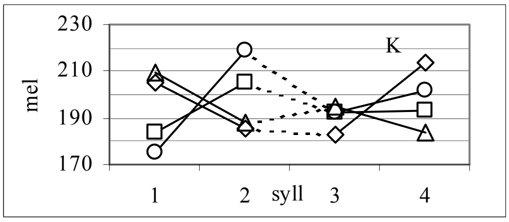

Fig. 5: BN, 2 words ≥ 90%, Speaker K

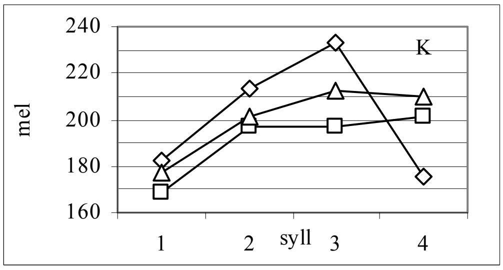

Fig. 6: BN, 1 word ≥ 80%, Speaker K

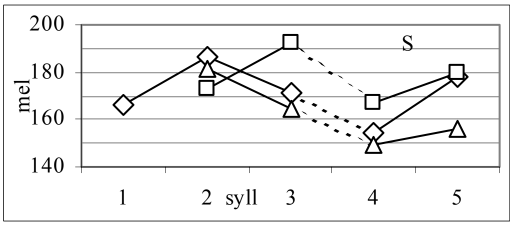

Fig. 7: BN, 2 words ≥ 90%, Speaker S

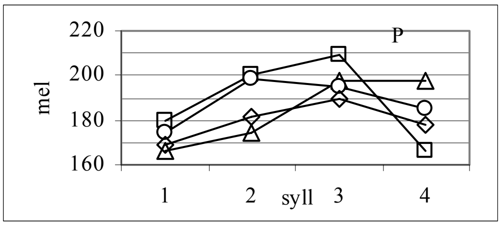

Fig. 8: BN, 1 word ≥ 80%, Speaker K

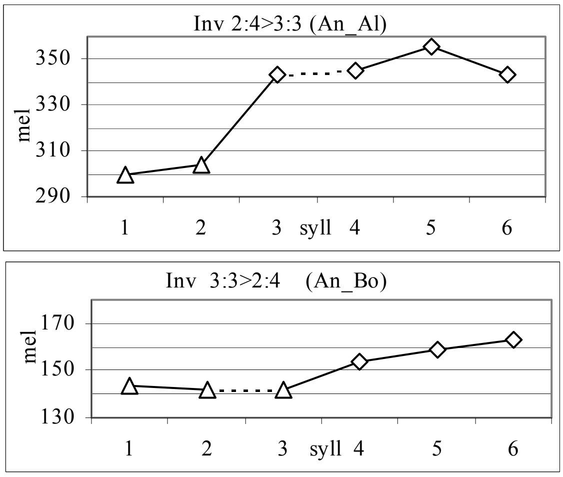

Fig. 9: AN, Inversions in evaluation

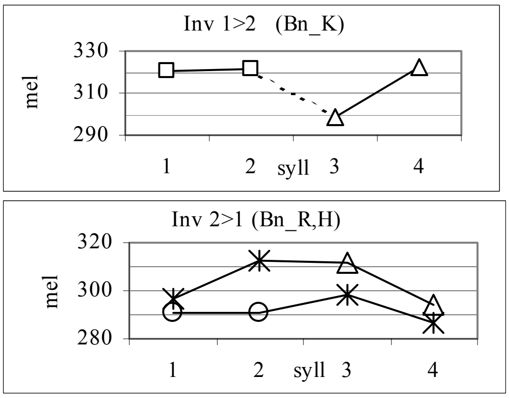

Fig. 10: BN, Inversions in evaluation

second foot (see fig. 1-5,7). Two subsequent rising contours further strengthen the tendency towards division (esp. fig.3,4). In the A-tests (i.e. those with the alternative boundary placement within the sequence), the cohesion-supporting contour in three-syllable feet was the rise-fall, especially for the first position in the sequence. When a three-syllable foot was positioned as the second in the sequence, linear rising or level contours were more significant (comp. fig.1,7 vs. 2,4,9). Cohesion supporting rising-falling contour was usually asymmetrical with less movement in its second portion. In the B-tests, the rising-falling contour or linear rising and linear falling contours supported the cohesion of the whole. This effect could be bolstered by small steps between neighbouring syllables with the exception of the last one (fig. 6,8).

The items that were perceived contrary to the intention of the speaker suggest that the listeners based their decisions on the relative acceptability of the whole sequence: either as a succession of two feet or as a solid unit (fig. 9,10). A more detailed analysis revealed characteristics of individual speakers. These involved preferences for using various types of height changes in the slopes of uni-directional contours or the slopes of the rising-falling contour. A simple cumulative graph provides example of such tendencies (fig.11).

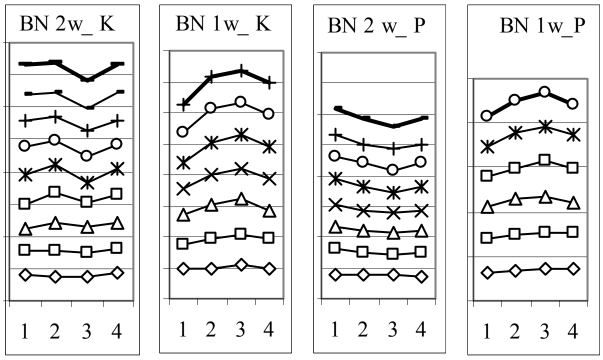

Fig. 11: Cumulative graphs revealing individual trends in F0

Tests based on synthesized speech (AS and BS tests) verified the negative influence of the F0 drop on the cohesion of the foot. We employed diphone synthesis which allows independent manipulation of the F0 values in syllabic peaks (For description see [4]). For the AS-test, four contours were chosen to represent the following directions of melodic changes: \\, \/, /\, //. Each type had four variants. These were used in the middle part of nine ambiguous sentences. Only small melodic steps were utilized, which resulted in rather low agreement in among listeners. The highest agreement between listeners was achieved in two contours which contained an F0 drop inside the potential foot. This solution was refused by the listeners. The effect grew stronger in the contours with more clearly defined shapes (fig.12).Tests of the BS-series investigated the influence of the direction of the changes in connection with the range of the melodic steps and the contour symmetry. There were 6 types tested (/, \, \\, \/, /\, //) always in two variants in 5 ambiguous sequences. The following graphs indicate that greater melodic steps weaken the cohesion of the foot and asymmetrical contours are more acceptable (fig.13).

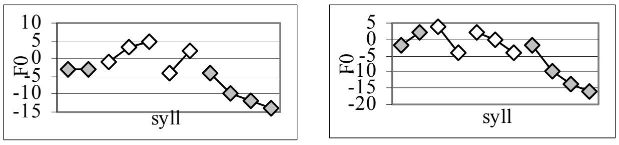

Fig. 12: AS, F0 patterns of type 4 >> 2 words

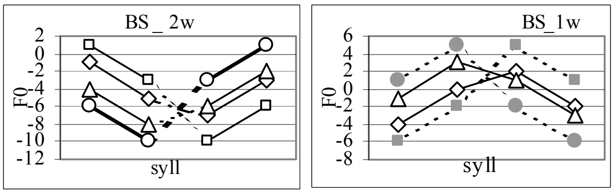

Fig. 13: BS, F0 patterns recognized as 2 or 1 words

4. CONCLUSIONS

Our experiments draw from the fact that the task imposed upon subjects does not require theoretical explanation of terms and is close to the real functioning of sound units in normal communication. It is evident that the task is within the capabilities of the subjects. The results are in good agreement with the preliminary hypothesis. For the future, it might be useful to enlarge the materials in terms of speakers and also, investigate different positions of ambiguous sequences in higher units. Additional types of contours might be sought as well.

The relationship between the foot as a sound unit and the word as the sign unit is essential for users of certain languages. Closer relationship makes the foot an existing reality, not just a descriptive concept. In fixed-stress languages like Czech, where the identity of a word is not dependent on the position of the prominence, the linear course of the sound properties throughout the unit plays probably a stronger role. For recognition, it is necessary to identify the unit as a whole. Stability of certain tendencies in the distribution of sound properties throughout the foot can become an effective means of foot cohesion, which in turn facilitates linear segmentation of texts.

ACKNOWLEDGEMENTS

Research supported by GACR 102/02/0124
GACR 405/02/0201.

REFERENCES

[1] P. Janota, “An experiment concerning the perception of stress by Czech listeners,” in *Phonetica Pragensia II*, pp. 45-68, Acta Universitatis Carolinae, Praha 1967.

[2] P. Janota and Z. Palková, “Auditory evaluation of stress under the influence of context,” in *Phonetica Pragensia IV*, pp. 29-60, AUC, Praha 1974.

[3] Z. Palková, “Einige Beziehungen zwischen prosodischen Merkmalen im Tschechischen,” in *Proceedings of the XIVth Internationalen Congress of Linguists*, Vol.I. pp. 507-510, Berlin1987.

[4] Z. Palková and M. Ptáček, ”Modelling prosody in TTS diphone synthesis in Czech“, in *Speech processing,* H.-W. Wodarz (ed.), pp. 59-77, Forum Phoneticum 63, Frankfurt am Main, 1997.

[5] P. Horák and J. Hanika, ”Dependences and independences of Text-to-Speech“, in *Papers in Phonetics and Speech Processing,* pp. 27-39, Z.Palková and H.-W. Wodarz, (eds.), Forum Phoneticum 70, Frankfurt am Main, 2000.