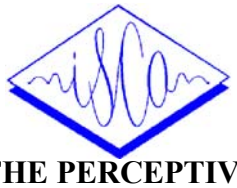


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

UNINTENDED PREFERENCES IN THE PERCEPTIVE EVALUATION OF RHYTHMICAL UNITS IN CZECH

Zdena Palková and *Jitka Janíková*
Institute of Phonetics, 116 38 Prague 1, Czech Republic
zdena.palkova@ff.cuni.cz jitka.janikova@ff.cuni.cz

ABSTRACT

The experience shows the evaluative judgements obtained through perceptive testing may contain unintended preferences made by listeners with no apparent motivation.An advantage for additional interpretation adjustments is if a general trend in such a distortion and its main direction can be detected. .In experiments devoted to finding out sound qualities that motivate the Czech listener to dividing a syllable chain into stress units ('phonetic words') in a given manner, some listeners tend to choose one of .the two possible variants with higher probability with no apparent motivation. Results from experiment presented in the contribution allow to assume that the tendency to prefer, in the 5-syllable sequence segmentation, a 3:2 chain of syllables over a 2:3 chain is possibly a prevalent listening preference in Czech.

10.21437/Eurospeech.1999-417

1. INTRODUCTION

The relationship between sound phenomena of particular spoken texts and generalised concepts instrumental in describing sound structure of language is in focus at any level of phonetic research. A relevant task is then the perceptive evaluation of speech signal by hearers in listening tests of several kinds. Monitoring tolerance of listeners for deflections from expected forms brings out remarkable information both on weight of individual sound parameters and on the share of vagueness in speech communication.

The experience, however, shows the evaluative judgements obtained through perceptive testing should be interpreted with regard to a possible presence of factors bringing a distortion to the evaluation results. An advantage for additional interpretation adjustments is if a general trend in such a distortion and its main direction can be detected. Cases which suggest that such a performance might be a general trend in a given language are particularly interesting. Results obtained in our current research of Czech prosody can demonstrate it.

2. RESEARCH IN STRESS UNIT PARSING IN CZECH

2.1 This research is partially devoted to finding out sound qualities characteristic of the *stress unit* in Czech, a rhythmical unit closely related to the word. What is looked for are the sound features that motivate the listener to dividing a syllable chain into stress units in a given manner. The research includes listening tests that provide sequences of syllables in which word juncture placement determines phrase meaning [1]. Hence the listeners, while deciding which way to segment the text, draw on their language experience, not on theoretic assumptions. (They decide whether e.g. „SVÌT -LO-VNÍ-MA-JÍ" should be segmented as SVÌTLO / VNÍMAJÍ, lit. „they perceive light," or rather SVÌTLO V NÍ / MAJÍ, „they have light in there", while the context admits both of them). The research employs both natural and synthesised speech samples.

The material to present here contains sets of 5-syllable two-stress unit chains, which can be parsed into sequences of either 2:3, or 3:2 syllables. The sound shape variability in such sequences is high in Czech and the tolerance of Czech listeners to both alternatives in parsing is high, too. Consequently, individual test items are considered by listeners with variable degree of agreement.

2.2 Generally, the results to be expected in parsing a 5-syllable chain are as follows:

60 to 80 % out of all the listeners' judgements in tests based on natural samples taken from non-professional speakers are in agreement with the original intention of the speaker. Concerning the response on individual tested items, the listeners show a significant interpersonal agreement in 65 to 85 % of the items; still, a few cases always keep emerging (about 10 %) of listeners parsing the syllable chain in a significant mutual agreement which is, however, inverse to the speaker's intention. These cases seem to point to the influence of a larger context in stress unit parsing in Czech. The rest are cases of inconsistent agreement in listeners' judgement.

The results in synthesised tests obviously depend on the way the prosodic features are utilised, especially the size of the changes in the contour and how they combine. E.g. using a short break or certain changes in the F0 contour provide a trivial way to produce a high degree of agreement in the listeners' judgement.

What seems significant for our topic is to pay close attention to the results in the tests that use just low-key changes in prosodic qualities that correspond to a non-expressive speech style. The samples we are going to deal with herein have been made by means of an automatic TTS diphone synthesis where only two

prosodic parameters are varied: the diphone duration related to the number of phones within the stress unit, and the F0 contour in the stress unit. The program can perform standard synthesis of long texts without requiring information on semantics and syntax while the result is well acceptable for a general listener (more on the concept in [2]. The sequences of two and three syllable stress units do not exceed in most cases the difference in duration of 10%. The F0 change bandwidth in the tests does not run over 10 % in a non-terminal position (±5 around 100 % which is 100 Hz).

The materials such as these obtain the judgement agreement between the expectation and the total volume of judgements in five syllable chain parsing from 55 to 65 %. Concerning the individual item evaluation agreement, 70 to 85 % of the cases are rather clearly determinate. Unlike the natural speech signal, however, the agreement in items evaluated satisfactorily in keeping with the original expectation is considerably lower, about 40 to 60 %. Inverse opinion cases can count as much as 30 %. The synthetic signal is also affected by the lexical composition of the items; with higher probability, some combinations get evaluated in preference of one way over the other. Probably, the sound shape of diphones may have its impact as natural speech samples having the same words display no such tendencies.

2.3 What seems interesting in view of our current topic, is that a close analysis of the results both as a whole and as subsets of test items, especially the items of lower listeners' agreement, reveal a certain assymetry in decisions. Some listeners seem to prefer one of the two variants for no obvious reason. The preferences apparently concern more often just one out of the two options: the Czech listeners seem to prefer the ratio. 3:2 over 2:3 in parsing five syllable chains.

3. EXPERIMENTS

Let us present results of 4 tests. The tests hereafter labelled as N1 and N2 contain items extracted out of natural speech, the tests labelled S1 and S2 contain samples made by synthesis.

3.1 The material for the tests

3.11 The items in the N1 and N2 tests are 8 five-syllable chains each. The chains can be parsed in Czech into two stress units in the ratio of 3:2 or 2:3, each with a different meaning.

The pairs have been extracted out of a connected text where they have been found in the middle of a sentence. The texts have been read by non-professional speakers and were long enough to prevent the readers from realising there were sequences with a double meaning. Each test uses either version performed by three different speakers.

3.12 The items in the S1 and S2 tests contain 9 five-syllable chains each, put in the middle of a four stress-unit sentence. The initial stress unit is two syllables, the final stress unit is five syllables. Their lexical composition is identical in a pair. Each chain in every test is included as four different F0 versions for the two parsing variants.

3.13 The listening groups for all the tests are rather homogenous in age, education as well as motivation, mostly students in pregraduate level their study who are 18 to 24. The number of listeners for the following results — N1: 75, N2: 80, S1 and S2: 46 each.

3.2 Selected results

3.21 The results based on the set of the total number of judgements

The Tab. 1 draws on the summary of all the judgements as to their agreement with the original intention (which is a speaker's intention in N1 and N2, and the type of intonation and diphone duration variant in S1 and S2). The columns show the agreement / disagreement with the original intention totalled for all the items of the type. The intention (type) is signified in capitals A or B, the listeners' judgements relate to the lower case *a* and *b*. The present contribution has always A variant for 3:2 parsing, B for 2:3 parsing.
(The tests did not have a consistent variant order so as to show if the position on the form affects the listeners' judgements. No such effect has been proved.)

Tab. 1 The agreement of the total judgement number with the expectation in %

	A as a	A as b	B as a	**B as b**	% of agreement
	3 : 2	3 : 2	2 : 3	**2 : 3**	**average**
N1	**81**	19	37	**63**	**72**
N2	**66**	34	41	**29**	**63**
S1	**64**	36	50	**50**	**57**
S2	**60**	40	48	**52**	**56**

The table shows the agreement with the intention is higher in all instances of the 3:2 parsing.

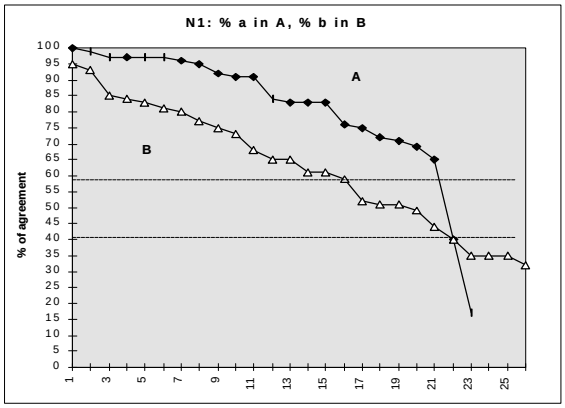
3.22 The results based on the judgement agreement concerning individual items

The graphs 1 to 4 show in N1, N2, S1 and S2 test a listeners' agreement in deciding on individual items and the solution related to the expectation. The items in each of the tests are separated into two sets corresponding to the types of A (symbol) and B (symbol). The *y*-axis placement of an item indicates the number of judgements corresponding to the expectation (in %). The market out middle range (40 to 60 %) covers the items indifferent to the listeners. The higher range contains instances with a majority decision (n >60 %) agreeing to the expectation, while the lower range

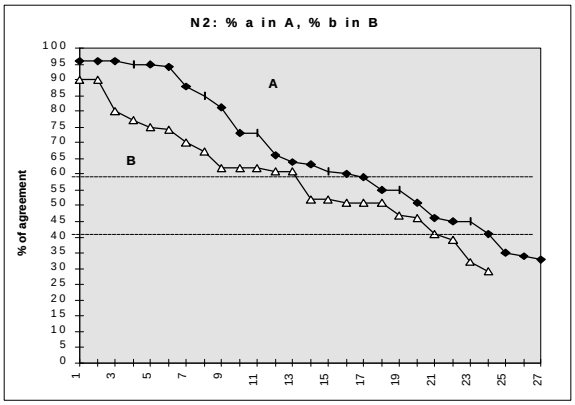
consists of the cases of a majority decision contrary to the expectation.

The evaluation of the data supporting the 3:2 parsing shows, across all the tests, a higher agreement of the listeners in cases decided in line with the expectation. The substitution of this type for the inverse happens less S1 tests especially). The 2:3 parsing rather reveals

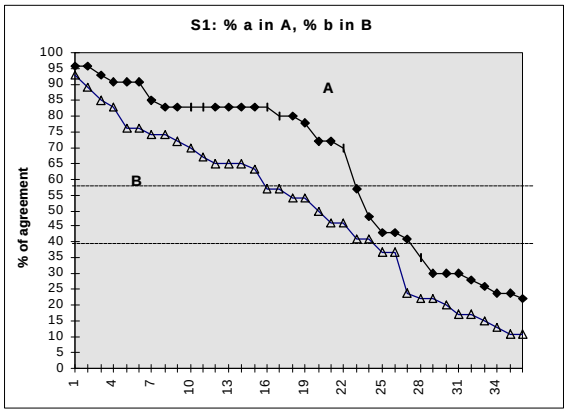
Graph 1



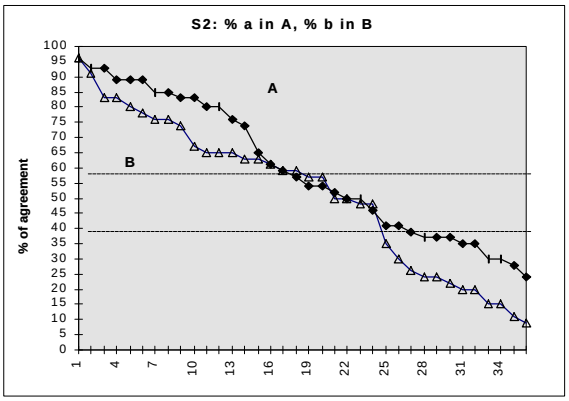
Graph 2



Graph 3



Graph 4



frequently than the reverse substitution (see the N1 and a higher percentage of agreement in the instances of inverse evaluation.

3.23 The results based on judgement analysis of individual listeners

The capacity available permits to present only a fraction of the material. The following examination aims at finding if a particular listener is open to „hear" either possibility equally as well, or whether they are prone to prefer one of them.

Please note that the way the test was presented lets one believe the semantic content of the pairs has only negligible impact on the choice. The N1 and N2 tests show no dependence on texts of particular types. The S1 and S2 tests demand to allow for an effect of the diphone quality in support of one or the other parsing. The tests have been therefore balanced to have the two variants in an equal number (3 indifferent sentences and two sets of 3 with a preference).

The graphs 5 to 8 show the number of a / b judgements in the N1, N2, S1 and S2 tests with the same eleven listeners. The order in the graph (*y*-axis placement) is also indicative if the evaluation is or is not in conformity to the expectation. The middle portions indicate agreement, the outer disagreement with the expectation. As the two variants have been almost or exactly balanced in their numbers, an outlying column in either direction suggests a preference for a type.

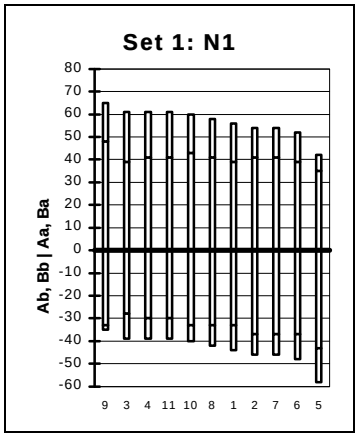
Graph 9 present results as to the application of the evaluation a / b for a larger group of 39 listeners. Let us take the N2 test for an example.

The presented data show that the preference for a type emerges in a conspicuous number of our tested persons. With a prevalence, it is the 3:2 preference. So e.g. in Graph 9, if the preference of one of the two variants is understood as the difference in a / b variant usage > 15%, then 26% of persons seems to be affected in the tests, including 20% in preference of the A variant.

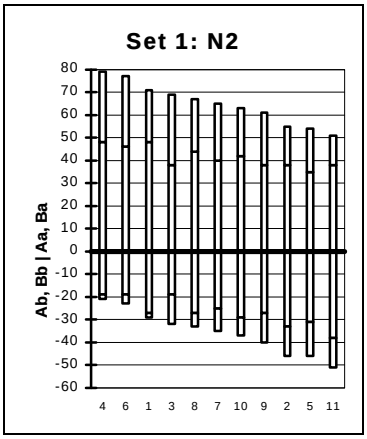
4. DISCUSSION

The same trend to parse a five-syllable chain in the ratio of 3:2 over 2:3 proves in all the tests processed. It is observed in a higher agreement in identifying the stress unit sequences which conform to such parsing, in a reduced readiness to substitute them as well as in an easier substitution for sequences of the inverse type.

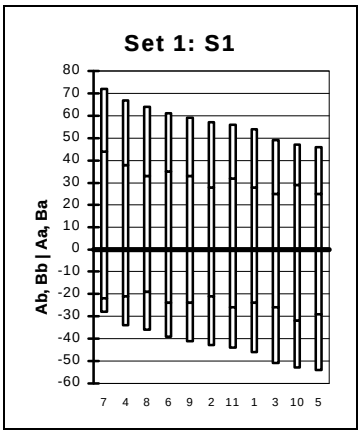
Graph 5



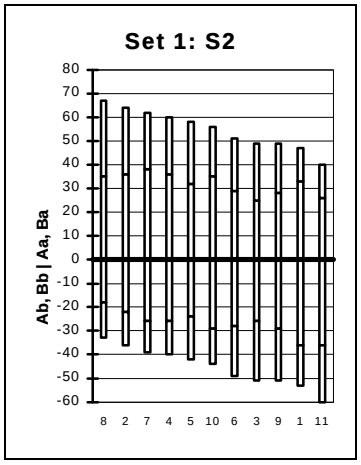
Graph 6



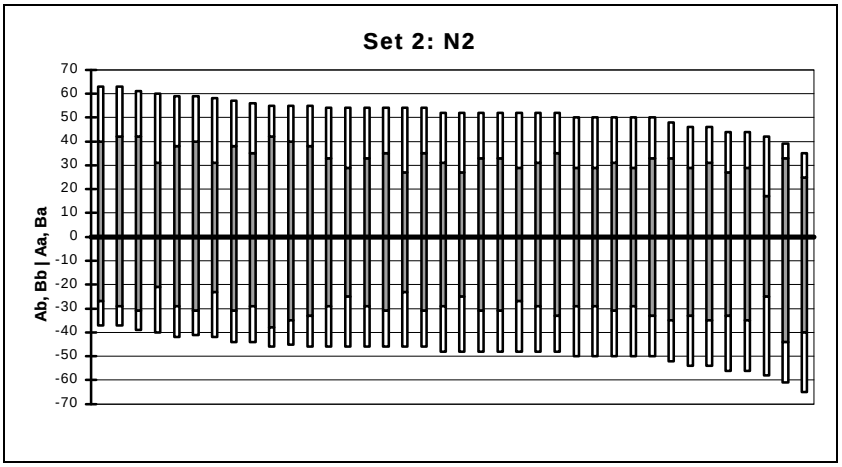
Graph 7



Graph 8



Graph 9



4.1 We feel no conclusion should be drawn concerning the source of the preference under the current condition of the research. The text versions leading to the initial three-syllable stress unit mostly represent two words in the stress unit and three words in the whole sequence which is a structure allowing a great variability of sound. The text leading to the 2:3 parsing is mostly made of two words, there is less sound variability, and that is why a more secure decision might be expected. The results to the contrary motivate us to believe the essential factor is rhythm rather than primarily phonology. Another explanation, however — that the effective background is the mostly descending rhythm of Czech, which is a feature of the language — also cannot be overlooked. Additional data are needed, e.g. based on material where the double meaning sequences would get into various places in longer utterances.

4.2 Recognition of unmotivated factors that affect perceptual judgements of sound phenomena helps us in interpreting results in tests of various kinds. It is significant especially where we meet with weaker tendencies and face the decision of what degree of evaluative agreement we should accept as still relevant.

Acknowledgement
The research is supported by the Grant Agency of the Czech Republic (research project GA CR 405/99/0172).

REFERENCES

[1] Palková, Z. (1996): Concurrent Context Intentions: A Relevant Factor in Connected Speech. In *Phonetica Pragensia IX,* AUC Philologica 1, Grague, pp.207-216

[2] Palková, Z. - Ptáèek, M. (1997): Modelling Prosody in TTS Diphone Synthesis in Czech. In: *Speech processing*, H.W.Wodarz (ed.), Forum Phoneticum 63, Frankfurt a.M., pp. 59-7