Statistical Evaluation of Prosodic Phrases in the Czech Language

Jan Romportl

Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Pilsen
rompi@kky.zcu.cz

Abstract

The present paper understands prosodic phrases as units which take part in constituting the rhythmical structure of speech. Due to very subjective and inconsistent criteria for the prosodic phrase perception there must be an objectively underlain method for prosodic phrase assignment. This has been achieved by extensive listening tests (103 participants) and statistical evaluation of the acquired data (e.g. maximum likelihood model).

1. Introduction

The rhythmical structure of an utterance is a very important prosodic feature not only for theoretical prosodic research, but also for applications such as text-to-speech (TTS) systems or speech synthesis systems generally. The aim of the present paper is to describe an approach for objectively underlain assignment of prosodic phrases (as one of the rhythmical constituents) in read speech data. The results of this effort also comprise quantitative characteristics of one aspect of rhythmical behaviour of the Czech language.

2. Prosodic phrases and speech synthesis

The concept of *prosodic phrase* – as understood in this paper – basically corresponds to what is meant by the term "discourse segment" (or "phonemic clause") in [1], i.e. such a phonetic unit which constitutes perception of the *rhythmical* qualities in language. A prosodic phrase is mainly delimited by acoustical features of its boundaries and it also usually contains an "intonation peak".

In our more abstract conception suitable for TTS synthesis purposes, a prosodic phrase is a suprasegmental unit which clusters segmental units (phones) into atomic sets with the same prosodic functions. Positional features of the segments used in TTS thus refer only to these clusters.

Obvious problems in the prosodic phrase definition arise mainly due to rather subjective nature of rhythm perception and often significant differences between the ways how phrasing is intended (and perceived) by a speaker and how it is perceived (and interpreted) by a listener. Our goals allow some simplification of this problem: we want to structure the speech data in such a way that a) this structuring is consistent throughout the whole dataset; b) the structuring obeys the principle of intersubjective agreement.

It is therefore very difficult (and under some circumstances almost impossible) for a single person to create *consistent* and reliable annotation of phrase boundaries in a speech corpus. The annotator is usually certain only in those cases where a phrase boundary is accompanied by a clear pause or possibly a *very* clear terminating pitch configuration; in other cases he often *feels* that placing a phrase boundary could be appropriate but according to our experience he oscillates throughout the annotation process in how clear the pitch configuration must be to pose a boundary – it means that very slight pitch shifts are often considered to be phrase boundaries while other more significant (even possibly accompanied by duration lengthening) are omitted.[1]

In other words: prosodic phrase boundaries are manually designated in a reasonable sub-part of the whole (presumably very large) real speech database so that there is agreement as high as possible among many independent listeners. The phrase boundaries (their model respectively) obtained this way are considered to be the "real" ones in the sense of "objectiveness", no matter our subjective opinion (i.e. in this case we settle for what can be explained as "vox populi, vox dei"). The definition of "prosodic phrase" is thus reduced to the form that the prosodic phrase is what emerges from this intersubjective agreement. In the second phase the relation between acoustical properties of speech and the "objective" phrase boundaries is implicitly captured in the form of a machine classifier trained on these data. Then this classifier can automatically extend the phrase boundary designation to the rest of the speech database. A subjective assessment of the automatically designated boundaries can be indeed questionable (the speaker or listeners might not agree with it in some cases) but the goal is actually not to find the prosodic boundaries in the phonetical sense *per se*, but to divide *consistently* the speech data into segments which can serve as phrase models (i.e. phone clusters) for speech synthesis.

3. Phrase boundary assignment

The intersubjective agreement on the phrase boundary assignment has been achieved by a statistical model applied on data acquired by extensive listening tests.

3.1. Listening tests

The listening tests were organised on the client-server basis using specially developed web application. We have used our speech corpus [2] designed as the source dataset for the text-to-speech system ARTIC. The corpus was very carefully recorded

Support for this work was provided by the Ministry of Education of the Czech Republic, project LC536, and the European Commission, EC grant number IST-FP6-034434 (project Companions). Great thanks go to our colleague Jan Zelinka for his valuable consultations.

[1] It is important to note that this obviously is not a mere experience with one annotator (possibly even badly trained) – this phenomenon seems to be a very general manifestation of prosody perception. Moreover, if there were a way to perfectly train an annotator, it would be infeasible for the purposes of speech synthesis corpora preparation and still it would lack needed "objectiveness".

in a studio by an experienced male speaker (the choice of the speaker had been consulted with two experts from the Institute of Phonetics, Charles University in Prague) who had been instructed to read isolated sentences naturally, yet avoiding any expressiveness. The speaker did not know that the recorded sentences would be used also for the phrasing analysis. The way how the corpus has been recorded (i.e. the type of recorded speech) obviously influences the scope of linguistically relevant findings of the research – therefore relevance of the quantitative results presented further in this paper is limited to the aforementioned speech domain; however, the methods we have used definitely are not limited to this data.

We have randomly selected 100 sentences from this corpus for the purposes of the listening tests and loaded them together with their orthographic transcriptions into the web application. Potential test participants were addressed among university students from all faculties (with special focus on students of linguistics) and finished listening tests were financially rewarded (so as to increase motivation of the students). The participants could do all the work from their homes without any personal contact with the test organisers – we have thus undertaken various measures to detect possible cheating, carelessness or misunderstandings.

The participants have been instructed to listen to the sentence recordings very carefully and subsequently designate words where they *are sure* there is a phrase boundary and words where they feel there *might be* a phrase boundary (i.e. these two cases were distinguished). Prior to the test itself the participants have been briefly familiarised with phonetic background of the problem and in this tutorial they listened to several training samples which showed possible phrasing demonstrations. It is, however, very important to note that we intentionally did not want to make almost any a priori assumptions about phrase boundary qualities or behaviour: we wanted to create "notion of prosodic phrase" in the participants and let them designate whatever subjectively fulfils this notion. Any stronger a priori assumption – not being rigorously and statistically underlain – could falsely influence results of the experiment, and even the existence of the language phenomenon of prosodic phrases is by itself quite a strong presupposition.

We have eventually received correctly finished tests from 103 participants (the total number of students who took part in these tests was 174, some of the students have not finished their tests, some of them have not even started, and there were also several apparent cheating attempts) which already gives us quite a robust observation set for further evaluation. Several interesting yet less important facts about the tests are in Table 1.

Table 1: *Several quantitative facts about the listening tests.*

finished tests	103
participants with phonetic education	25
average time spent on one test	92 min
avg. num. of sentence replays	2.33
avg. num. of sessions per user	3.10
total number of sentences	100
total number of word tokens	1063
total length of speech	$\approx$ 508 s

3.2. Statistical evaluation

The goal of the listening tests was to find places in the given sentences where we can make intersubjective agreement on phrase boundary occurrences. The resulting phrase deployment is then to be treated as an objective basis for any further research. We can transform the problem of such a phrase deployment based on many independent observations into more abstract and formal level:

Let X be a random process defined as

$$X = \{X_t \,:\, t \in T\} \tag{1}$$

where $T = \{1, 2, \ldots n\}$ is a set of time points respective to ordinal numbering of words in the test sentences (i.e. the first word in the first sentence has $t = 1$, the second word in the first sentence has $t = 2$, and so on), and X_t are random variables which hold $X_t = 1$ iff the t-th word finishes a prosodic phrase, and $X_t = 0$ otherwise.

Now let the test participants be numbered by the set $J = \{1, 2, \ldots m\}$, i.e. the first participant has $j = 1$, the last one has $j = m$. We can define m random processes $O^{(1)}, \ldots O^{(m)}$ representing the participants' responses (observations) such that

$$O^{(j)} = \{O_t^{(j)} \,:\, t \in T\} \tag{2}$$

where t has the same meaning as for the process X, and $O_t^{(j)}$ are random variables which hold $O_t^{(j)} = 1$ iff the j-th participant asserts that the t-th word finishes a prosodic phrase, and $O_t^{(j)} = 0$ iff the j-th participant *does not* assert that the t-th word finishes a prosodic phrase.

Our goal can now be re-formulated as follows: knowing the observations $O^{(1)}, \ldots O^{(m)}$ we want to estimate the hidden trajectory of the process X which best satisfies the given observations.

We have applied two approaches towards the hidden trajectory estimation – the principle of simple majority and the principle of maximum likelihood – and further in this paper we compare their results. In both approaches the two variants of the test answers (i.e. "boundary for sure" and "boundary maybe") were treated equally – this was based on the assumption that if the "statistically relevant" number of participants thinks that there *might be* the phrase boundary at the given place, it *really is* there. The reason for allowing two levels of certainty from the participants' side was mainly due to the experience that if a listener is really not sure, he answers randomly – and this can be avoided by the "maybe" variant.

3.2.1. Simple majority

The approach based on the principle of simple majority is quite an easy and intuitive solution to the aforementioned problem. It also best models the intersubjective agreement understood as a voting process in which the participants vote for each word whether it should bear a phrase boundary or not. The basic idea is that a phrase boundary occurs at the given position in case at least 50 % of the participants vote for it. This can be possibly enhanced by weighting each participant's vote according to some confidence criterion. The threshold of 50 % can be indeed changed to a different value but if there is no clear reason for doing it, it would be methodologically incorrect.

Formally we can express this approach by a random variable given as a weighted average

$$R_t = \frac{\sum_{j \in J} w_j \cdot O_t^{(j)}}{\sum_{j \in J} w_j} \tag{3}$$

and the decision criterion is then

$$X_t = 1 \iff R_t > r \tag{4}$$

where $r = 0.5$ is the threshold corresponding to the simple majority. Further in this paper we will show the listening tests results for the most obvious case of the equal weights (i.e. $\forall j, l \in J : w_j = w_l$).

3.2.2. Maximum likelihood

The intuitive basis and simplicity of the previous approach are, however, also its main drawbacks. How can we be sure that the 50 % threshold is the right one or that a listener is really reliable although he for example always says for any word that there is a boundary?

We can now really make use of the benefits of the random process formalisation of our task and transform it into the problem of finding the most likely model parameters given the observed data. The relations between the unknown "real" boundary and a participant's assumption is expressed by the probabilities:

$$P(O_t^{(j)} = 1|X_t = 1) = r^{(j)} \tag{5}$$

$$P(O_t^{(j)} = 0|X_t = 1) = 1 - r^{(j)} \tag{6}$$

$$P(O_t^{(j)} = 0|X_t = 0) = f^{(j)} \tag{7}$$

$$P(O_t^{(j)} = 1|X_t = 0) = 1 - f^{(j)} \tag{8}$$

The equations 5 and 7 express the probabilities that the j-th participant correctly identifies the boundary presence or absence for a word, the equations 6 and 8 express the probabilities of two kinds of errors.

As it has been already justified in the previous paragraphs, we do not want to make any strong a priori assumptions about the random process X (i.e. phrase boundary deployment), therefore we can obey the principle of Occam's razor and presuppose that X is a stationary process with the alternative probability distribution, thus:

$$X \sim A(p) \tag{9}$$

where $\forall h, i \in T : p_h = p_i = p$. In this point we really intentionally pretend that we do not know anything about phrasing behaviour so that all words have equal probability of bearing a phrase boundary (phrase lengths, lexical, syntactical, semantical or any other factors are excluded on account of the methodological constraints).

Through the equations 5–9 we have postulated the structure of the probabilistic model of our problem and now we can see that it has the unknown parameters $r^{(j)}$, $f^{(j)}$ and p which we will further collectively denote as Θ.

The goal is to find the most likely parameters Θ^* given the observation $O = [O^{(1)}, \dots O^{(m)}]$, i.e. maximise the likelihood function

$$L(\Theta) = P(O|\Theta) \tag{10}$$

$$\Theta^* = \arg\max_{\Theta} L(\Theta) \tag{11}$$

There is not an analytical solution to the equation 11 and therefore we have decided to estimate the parameters by an expectation-maximisation (EM) algorithm. The EM algorithm is proved not to decrease the likelihood function in any iteration but it will converge to a local maximum, hence the initial parameters must be chosen reasonably and perturbed in more experiments.

We have set the initial parameters Θ_0 heuristically: $p = 0.5$, $r_t^{(j)} = 0.7$ and $f_t^{(j)} = 0.9$ for all j and t. These initial conditions (as well as their various perturbations) converged already after 10 iterations of the EM algorithm[2] to a saddle point. The parameters in this saddle point are considered to be Θ^*: $r^{*(j)}$ and $f^{*(j)}$ are obviously different for all j, the constant parameter of the alternative distribution converges to $p^* = 0.8509$, i.e. $\forall t : P(X_t = 0) = 0.8509$.

The probability that the t-th word bears a phrase boundary given the observations $O_t = [O_t^{(1)}, \dots O_t^{(m)}]$ is

$$P(X_t = 1|O_t) = \frac{\prod_{j \in J} P(O_t^{(j)}|X_t = 1) \cdot P(X_t = 1)}{P(O_t)} \tag{12}$$

and therefore we can formulate the decision criterion as

$$X_t = 1 \iff P(X_t = 1|O_t) > P(X_t = 0|O_t) \tag{13}$$

where

$$P(X_t = 0|O_t) = 1 - P(X_t = 1|O_t) \tag{14}$$

and since $P(O_t)$ is constant for the given t, we can omit it and compute only the numerator from the equation 12.

4. Data evaluation

The prosodic phrase boundaries have been deployed independently by both methods and the results have been compared. The EM algorithm has placed the boundaries on all the words as the simple majority approach but in addition to this it has designated 8 words with $R_t <= 0.5$ as bearing the phrase boundary (i.e. the agreement on these words was lower than 50 % but still they were more likely to have the boundary than not).

The lowest observed agreement which received the boundary by the EM algorithm was $R_t = 0.48$ but there were also cases with the same or higher (0.48 and 0.49) agreement which were not considered as the boundaries. It is quite a coincidence that this even occurred in a single sentence:

"... existenciální motivy, */(0.99) pravděpodobně *(0.48) v mnohém (0.48) autobiografické." ("... existencial motives, probably in many respects autobiographical.")

The numbers in the brackets mean the agreement among the participants on possible boundary placement, the slash indicates that the boundary was assigned by the simple majority approach and the asterisks designate the boundaries assigned by the EM algorithm.

4.1. Phrase boundary types and lengths

For the purposes of this paper we have not explicitly analysed acoustical properties of the designated boundaries – we have only distinguished two cases: the boundary without a pause (B1) and the boundary with a pause (B2). Although the explicit distinction and analysis of prosodic forms contributing on phrase delimitation (similar to [1]) is very important from the phonetical point of view, with this task we rely rather on implicit automatic analysis, as it has been outlined in the previous sections.

Table 2 comprises an overview of phrase boundary type frequencies, as assigned by the simple majority (SM) and maximum likelihood (ML) approaches. The frequencies do not include phrase breaks at the sentence ends – it means that only "intra-sentential" boundaries have been considered. It can be

[2] To be more precise, it was a Baum-Welch algorithm simplified to suit the needs of this problem. Instead of explicit maximisation of $L(\Theta)$ the algorithm maximises $P(X|O)$ by iterative gradient changes of the parameters Θ – this process ensures growth of $L(\Theta)$.

seen that almost two thirds of all the phrase boundaries are accompanied by a pause and that the difference between the assignment approaches is mostly in the cases without a pause (because vast majority of boundaries with pause had significantly higher agreement than 50 %).

Table 2: *Boundary type frequencies.*

	SM	ML
B1	91 (39.1 %)	98 (40.7 %)
B2	142 (60.9 %)	143 (59.3 %)
total	233	241

Information about an average phrase length is in Table 3. Unlike [1] we have measured the lengths in lexical words instead of prosodic words (phonetic words respectively). The reason is mainly that our tested data do not comprise prosodic word annotation and even if they did, it would make the whole task harder to statistically evaluate due to significant uncertainty in the prosodic word assignment itself. The results presented here can be partially comparable to [1] by assuming the average length of a prosodic word (according to [3] a prosodic word is in average 1.34 lexical words long) but still there must be a tolerance because this average length is not based on our data. It is clear that the listeners tend to perceive quite short phrases and our results correspond to findings of [1].

Table 3: *The average and maximum phrase length (in lexical words).*

	SM	ML
avg. len.	3.19	3.12
std. dev.	1.41	1.36
max. len.	9	8

4.2. Overall agreement

Another factor describing properties of prosodic phrase perception is a measure of agreement among the test participants. We have calculated specifically the agreement between each couple of the participants on placing the phrase boundaries. The overall agreement measure was then calculated as an average of these values.

We have chosen two approaches of computing the agreement between two participants: the first approach counts all the words where both participants assumed the same results, the second approach omits the words where none of the participants assumed a boundary. Formally, the agreement $A_1(i,j)$ between the participants i and j in the sense of the first approach is given as

$$A_1(i,j) = \frac{\sum_{t \in T} f_{ij}(t)}{n} \tag{15}$$

whereas

$$f_{ij}(t) = \begin{cases} 1 \Leftrightarrow (\varrho(O_t^{(i)}) = \varrho(O_t^{(j)})) \\ 0 \Leftrightarrow (\varrho(O_t^{(i)}) \neq \varrho(O_t^{(j)})) \end{cases} \tag{16}$$

where $\varrho(x)$ is integer rounding of x (its purpose will be explained later). The second approach defines the agreement $A_2(i,j)$ as

$$A_2(i,j) = \frac{\sum_{t \in T}(f_{ij}(t) \cdot c_{ij}(t))}{\sum_{t \in T} c_{ij}(t)} \tag{17}$$

where

$$c_{ij}(t) = \begin{cases} 1 \Leftrightarrow (O_t^{(i)} = 1 \vee O_t^{(j)} = 1) \\ 0 \Leftrightarrow (O_t^{(i)} = 0 \wedge O_t^{(j)} = 0) \end{cases} \tag{18}$$

The overall agreement is given as

$$A_1 = \frac{\sum_{i,j \in J; j>i} A(i,j)}{\frac{1}{2}m^2 - m} \tag{19}$$

and for A_2 analogically. In these equations the "sure" and "maybe" variants of the answers of the participants are treated equally (i.e. $O_t^{(i)} = 1$ in both cases: the i-th participant assumed that the t-th word bears the boundary or assumed that maybe bears) and the values A_1 and A_2 are in Table 4 placed in the column M1. Another possibility is to disregard all the "maybe" variants (i.e. $O_t^{(i)} = 0$ for all "maybe" answers) – the results for it are under the designation M2. The most interesting possibility is, however, the one where the "sure" and "maybe" variants are treated differently: in this case $O_t^{(i)} = 1$ in the "sure" variant and $O_t^{(i)} = 0.6$ in the "maybe" variant (M3 in Table 4). The rounding ϱ from the equation 16 is applied here and the equation 18 has a slightly different form:

$$c_{ij}(t) = \begin{cases} 1 \Leftrightarrow (O_t^{(i)} = 1 \vee O_t^{(j)} = 1) \\ 1 \Leftrightarrow (O_t^{(i)} = 0.6 \wedge O_t^{(j)} = 0.6) \\ 0 \; otherwise \end{cases} \tag{20}$$

Table 4: *Overall agreement among the participants.*

	M1		M2		M3	
	A_1	A_2	A_1	A_2	A_1	A_2
A_x	0.81	0.41	0.86	0.41	0.81	**0.56**
$stdev(A_x(i,j))$	0.04	0.06	0.04	0.06	0.04	0.09
$max(A_x(i,j))$	0.87	0.52	0.91	0.56	0.87	0.74
$min(A_x(i,j))$	0.62	0.14	0.71	0.12	0.62	0.19

5. Conclusions

The quantitative results presented here describe one aspect of rhythm perception in the Czech language. They are conclusive and underlain by objective methods and therefore they do not depend on subjective opinions. This way we have acquired a model of a virtual listener who is "always right" in prosodic phrase judgement. Of course the actual values and parameters of the phrase deployment strongly depend on the speech material, but the methods we have used are repeatable and would reproduce the same results on the same data.

Moreover, we have reached the results quantitatively comparable to those in classical studies of the Czech phonetics, such as [1], and this can be also understood as a kind of verification of our work. From the bold number in Table 4 we can see that 103 listeners have agreed on 56 % of all phrase boundaries, no matter their type or distinctiveness. It is also interesting to note that almost the same value has been reached when considering only the results of 25 participants with phonetic education.

6. References

[1] Palková, Z., 1974. Rytmická výstavba prozaického textu (with English resume: The rhythmical potential of prose). Prague: Academia.

[2] Matoušek, J.; Romportl, J., 2007. Recording and annotation of speech corpus for Czech unit selection speech synthesis. In *Lecture Notes in Artificial Intelligence, vol. 4629*. Berlin-Heidelberg: Springer, 326-333.

[3] Psutka, J.; Müller, L.; Matoušek, J.; Radová, V., 2006. Mluvíme s počítačem česky (Talking with Computer in Czech). Prague: Academia.