Prosodic Phrases and Semantic Accents in Speech Corpus for Czech TTS Synthesis*

Jan Romportl

[1] SpeechTech, s.r.o., Morseova 5, Plzeň, Czech Republic
[2] Department of Cybernetics, Faculty of Applied Sciences
University of West Bohemia, Univerzitní 8, Plzeň, Czech Republic
jan.romportl@speechtech.cz

Abstract. We describe a statistical method for assignment of prosodic phrases and semantic accents in read speech data. The method is based on statistical evaluation of listening test data by a maximum-likelihood approach with parameters estimated by an EM algorithm. We also present linguistically relevant quantitative results about the prosodic phrase and semantic accent distribution in 250 Czech sentences.

1 Introduction

The aim of this paper is to describe a method which is able to objectively assign (annotate) prosodic phrases and semantic accents in read speech data and present results of such an assignment in the Czech language. The concept of *prosodic phrase* – as understood here – basically corresponds to a traditional view or to what is meant by the term "phonemic clause" (or "discourse segment") in Czech literature [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 can also contain an "intonation peak". However, as [1] discusses, there is no empirical evidence supporting any stronger assumption about the intonation peak presence/absence or their number in a Czech utterance.

We therefore assume solely that a speaker may *emphasise any number* of words by acoustic means to express (perhaps even unintentionally) their prominence in comparison with other words. The acoustic prominence of a word can deliver various kinds of information – from this point of view we can observe the acoustic prominence even on words where a phrase end is acoustically realised. However, such a prominence has a different function: we want to find words whose acoustical prominence has an emphasising function in terms of semantics or pragmatics. We have decided to call such a phenomenon a *semantic accent*.

Prosodic annotation within the ToBI framework and its objectiveness in terms of quantitative inter-transcriber reliability (together with extensive references on the topic) is presented for example in [2]. Such annotation, however, is very time-consuming and requires highly trained annotators. Moreover, in conformity with [3] we prefer labelling

* Support for this work was provided by the Ministry of Education of the Czech Republic, projects 2C06020 and LC536.

of what an annotator hears rather than the shape of F0. Similar task has been described in [4] where the authors conclude with the statement that it is also reliable in some way to work with naive listeners.

As it will become evident further in our paper, phrases and semantic accents in Czech speech often lack any clear and reliable acoustic cues, therefore the number of naive listeners annotating our data must be relatively very high to generate a robust statistics. A similar strategy can be found in [5] where 74 students took part in listening tests on prominence and boundary perception, albeit our goal (as well as test protocol and nature of the speech data) here is different – we primarily want to find out *where* in our speech data the intonation phenomena *are*, whereas *how* they are *perceived* is secondary for us.

2 Prosodic phrases, semantic accents and speech synthesis

Performance of a unit selection TTS synthesis approach strongly depends on the way how the basic speech units (e.g diphones) in its database are parameterised in terms of their higher-level features (i.e. linguistical – as opposed to acoustical, which are obviously very important as well). Such a parameterisation should be compact and should try to cluster the speech units according to some repetitive suprasegmental patterns present in the speech data. This way usability of each unit can be as broad as possible while disallowing it to be used in such contexts where its specific properties would cause unwanted speech disruptions.

A discussion over the need of a *consistent* and reliable annotation of phrase boundaries together with problems in achieving it can be found in [6]. Moreover, the suprasegmental acoustic characteristics of strings of the basic speech units in emphasised words (i.e. the words bearing the semantic accent) are different from those which cover non-emphasised words. It is therefore important to have such units parameterised so that these differences are present in this parameterisation, hence the semantic accents must be designated in the speech corpus too.

The idea of the whole process is following: prosodic phrase boundaries and semantic accents 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 and semantic accents (their model respectively) obtained this way are considered to be the "real" ones in the sense of "objectiveness", no matter our subjective opinion. In the second phase (not covered by this paper) a machine classifier trained on these data can automatically extend the phrase boundary and semantic accent designation to the rest of the speech database, without being "confused" by inconsistencies in training data subjectively annotated by a single person.

3 Inter-subjective annotation process

The inter-subjective agreement on the phrase boundary and semantic accent annotation has been achieved by a statistical model applied on data acquired by two independent 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 [7] designed as the source dataset for our text-to-speech system ARTIC. A test layout description and technical issues together with information about the source corpus and test participants are in [6].

The first listening test is described in detail together with its evaluation also in [6] and we will not discuss it here. It is relevant for us now that its result are 100 sentences from the aforementioned corpus with labelled prosodic phrase boundaries. No semantic accents were labelled in this test and the participants had taken part in it actually with no knowledge about the phenomenon of semantic accent.

The second listening test (which was carried out 3 months after the first test and we will analyse it in the present paper) consisted of two parts (further in the text denoted as Part 1 and Part 2). Part 1 was aimed at finding the semantic accents in the sentences where the prosodic phrase deployment is already given: the same sentences as in the first listening test have been used again and the participants had been instructed to listen to these sentences very carefully and subsequently designate words where they perceived the semantic accent. The textual form of the sentences was displayed together with the a priori prosodic phrase deployment acquired from the first test. The participants had to accept this phrasing and bring into line their semantic accent assignment with it. This part has also served as a "tutorial" for Part 2 because the participants could learn this way what is statistically considered as phrase boundaries.

Part 2 was actually a combination of Part 1 and the first test: we have selected another 150 sentences from our corpus and the participants were again instructed to listen to the sentence recordings and designate the semantic accents. Moreover, in this part the task was also to designate words where the participants *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). It means that for every word in each of these 150 sentences the test comprises three options: a) this word is emphasised; b) after this word there certainly is the phrase boundary; c) after this word there might be the phrase boundary; the options *b* and *c* are obviously mutually exclusive.

We have eventually received correctly finished electronic answer sheets from 99 participants. It is worth mentioning that the first listening test described in [6] was finished by 103 participants from which 46 took part also in the second test and finished it (i.e. they have already had previous experience with phrase boundary assignment).

3.2 Statistical evaluation

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

Let X be a random process defined as $X = \{X_t : t \in T\}$, where $T = \{1, 2, \dots n\}$ is a set of time points respective to the 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. Exactly the same can be done for the semantic accents, such a random process is analogical to X and will be denoted as Y. We assume that the random processes X and Y are mutually independent.

Now let the test participants be numbered by the set $J = \{1, 2, \dots m\}$, i.e. the first participant has $j = 1$, the last one has $j = m$. We can define m random processes $O^{(1)}, \dots O^{(m)}$ representing the participants' responses (observations) such that $O^{(j)} = \{O_t^{(j)} \ : \ t \in T\}$, 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 $O^{(1)}, \dots O^{(m)}$ we want to estimate the hidden trajectory of the process X which best satisfies the given observations.

This can be analogically defined for the process Y with the only difference that the observations refer to the semantic accents and are based on the whole set of 250 sentences, whereas X describes only the subset of 150 sentences additionally chosen for the second listening test. For the sake of lucidity we will speak further in the text only about the process X assuming that everything which holds for it, holds analogically also for the process Y. It is supported by the fact that the two variants of the answers on the phrase boundary presence/absence (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 think 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. The difference between these two variants is utilised in the participants' agreement calculation (see section 4). We have decided not to use such two variants for the semantic accents because the semantic accent is defined more vaguely and a positive answer about its presence is actually most often an opinion like "this word might be emphasised".

The aforementioned goal can be transformed into the problem of finding the most likely model parameters given the observed data – a *maximum likelihood* approach. The relations between the unknown "real" boundary and a participant's assumption is expressed by the probabilities:

$$P(O_t^{(j)} = 1 | X_t = 1) = r_X^{(j)} \tag{1}$$
$$P(O_t^{(j)} = 0 | X_t = 1) = 1 - r_X^{(j)} \tag{2}$$
$$P(O_t^{(j)} = 0 | X_t = 0) = f_X^{(j)} \tag{3}$$
$$P(O_t^{(j)} = 1 | X_t = 0) = 1 - f_X^{(j)} \tag{4}$$

We further presuppose that X is a stationary process with the alternative probability distribution, thus:

$$X \sim A(p) \tag{5}$$

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 1–5 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

$$L(\Theta) = P(O|\Theta) \tag{6}$$

$$\Theta^* = \arg\max_{\Theta} L(\Theta) \tag{7}$$

There is not an analytical solution to the equation 7 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. However, it tends to 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 and equal for both the estimation of X and of Y : $p = 0.5$, $r_t^{(j)} = 0.7$ and $f_t^{(j)} = 0.9$ for all j and t. The parameters converged in both cases already after 10 iterations of the EM algorithm to Θ^*. The parameter of the alternative distribution converges to $p_X^* = 0.8470$, i.e. $\forall t : P(X_t = 0) = 0.8470$, and $p_Y^* = 0.9783$, i.e. $\forall t : P(Y_t = 0) = 0.9783$. The values $r^{*(j)}$ and $f^{*(j)}$ are generally different for all j but still their brief characterisation can be found in Table 1. We have used the 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 $L(\Theta)$.

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

$$P(X_t = 1|O_t) = \frac{\prod_{j \in J} P(O_t^{(j)}|X_t = 1) \cdot P(X_t = 1)}{P(O_t)} \tag{8}$$

and therefore we can formulate the decision criterion as

$$X_t = 1 \iff P(X_t = 1|O_t) > 1 - P(X_t = 1|O_t) \tag{9}$$

and since $P(O_t)$ is constant for the given t, we can omit it and compute only the numerator from the equation 8. The same criterion holds also for the process Y.

4 Results and conclusions

After having formally decided which words bear the phrase boundaries and semantic accents using the method described in the previous section, we can formulate some assertions about phrasing and emphasising in the Czech language. These assertions are based on the quantitative evaluation of the acquired data from various points of

Table 1. Probabilities of two kinds of errors the participants have done in placing the phrase boundaries (X) and semantic accents (Y), as estimated by the EM algorithm. These probabilities refer to the equations 2 and 4.

	$X : P(0\|1)$	$X : P(1\|0)$	$Y : P(0\|1)$	$Y : P(1\|0)$
average	0.125	0.054	0.380	0.159
st. dev.	0.099	0.049	0.191	0.096
min	0.000	0.004	0.036	0.013
max	0.421	0.351	0.864	0.400

view and have rather linguistically-theoretical value. On the contrary, the labelled data themselves have great practical value for our TTS system development.

Pause presence is one of the most important acoustic features signalling the phrase boundary, therefore it is useful to classify all phrase boundary occurrences into two classes: the boundary without a pause (B1) and the boundary with a pause (B2). Table 2 comprises an overview of phrase boundary type frequencies together with similar information on the semantic accents (SA). The frequencies do not include phrase breaks at the sentence ends – it means that only "intra-sentential" boundaries have been considered. This table also shows the inter-participant agreement for both boundary types, given as the relative number of the participants who have placed these boundaries. We can see that if the phrase boundary is followed by a pause, then in average 97 % of the participants agree on the phrase boundary assertion. Considering also the standard deviation it is clear that most of the cases with a pause are agreed on by more than 90 % of the participants. This is in contrast with 72 % average agreement on the non-paused phrase boundaries but even such a number is still a very good indicator that the phrase boundaries without a pause are well recognised too. The minimum agreement in the class B1 is less than 50 % – in such a case the EM algorithm has "decided" to place the boundary because the votes from the more "credible" participants (i.e. with higher $r_X^{(j)}$ and lower $1 - f_X^{(j)}$) have higher weight. There are 23 phrase boundaries with the agreement value less than 50 %. The average agreement on the semantic accents is much lower than on the phrase boundaries – this is in conformity with much higher error probabilities (see Table 1), which point out that the semantic accent is a less clear language phenomenon than the prosodic phrase (we will see that a similar conclusion implies also from the labelling agreement among the participants).

Table 2. Boundary type (B1 – without a pause, B2 – with a pause) and semantic accent (SA) frequencies together with agreement among the participants on phrase boundary and semantic accent placement.

		agreement			
	frequency	average	st. dev.	min	max
B1	201 (39.11 %)	72 %	17 %	43 %	100 %
B2	313 (60.89 %)	97 %	4 %	75 %	100 %
SA	227	49 %	10 %	83 %	36 %

Another important factor describing the phrase distribution is length of the phrases, viewed either as the number of phrases in a sentence or the number of words in a

Table 3. Distribution of the sentence lengths given as the number of phrases in a sentence and the phrase lengths given as the number of words in a phrase.

sentence length (in phrases)	frequency		
	whole test	part 1	part 2
1	4 (1.6 %)	0 (0.00 %)	4 (2.67 %)
2	89 (35.6 %)	22 (22.00 %)	67 (44.67 %)
3	78 (31.2 %)	35 (35.00 %)	43 (28.67 %)
4	56 (22.4 %)	27 (27.00 %)	29 (19.33 %)
5	14 (5.6 %)	11 (11.00 %)	3 (2.00 %)
6	9 (3.6 %)	5 (5.00 %)	4 (2.67 %)
phrase length (in words)	frequency		
	whole test	part 1	part 2
1	58 (7.59 %)	35 (10.23 %)	23 (5.45 %)
2	172 (22.51 %)	93 (27.19 %)	79 (18.72 %)
3	224 (29.32 %)	101 (29.53 %)	123 (29.15 %)
4	145 (18.98 %)	55 (16.08 %)	90 (21.33 %)
5	97 (12.7 %)	47 (13.74 %)	50 (11.85 %)
6	42 (5.5 %)	7 (2.05 %)	35 (8.29 %)
7	15 (1.96 %)	2 (0.58 %)	13 (3.08 %)
8	8 (1.05 %)	2 (0.58 %)	6 (1.42 %)
9	3 (0.39 %)	0 (0 %)	3 (0.71 %)

phrase. The distribution of these lengths is in Table 3. Since the phrase deployment in Part 1 of the test has been acquired without regards to the semantic accents, whereas the phrases in Part 2 have been deployed concurrently with the semantic accents, we give comparison of the phrase lengths in both parts separately as well. We can simply state that the fact whether the phrases are assessed with or without regard to the semantic accents *influences* the resulting phrase assessment. This is most apparent in case of short phrases from Part 1 – they are often replaced in Part 2 by longer phrases with the semantic accent elsewhere than on the last word.

The latter statement is supported by data shown in Table 4. According to this table there is in Part 2 a significant decrease of the number of phrases with the semantic accent on the last word in comparison with Part 1. We can make one more important conclusion from this table: it proves the hypothesis that there can be at the most one semantic accent in a phrase. The test participants had no prior information about expected or allowed numbers of the semantic accents in the phrases. Still the resulting statistically underlain deployment places no more than one semantic accent into any phrase, no matter whether the semantic accents have been assessed concurrently with the phrase boundaries or separately afterwards.

We can also present calculation of the overall inter-participant agreements A_X (for the phrase boundaries) and A_Y (for the semantic accents) given as the average of mutual agreement of all possible pairs of the participants. The mutual agreement of two participants is calculated as a quotient s/n where s is the number of word tokens where both participants had the same answer ("same answer" for the phrase boundary assignment means also the situation when one participant answers "maybe boundary" and the other can then answer anything) and n is the total number of tokens where they answered. We have obtained $A_X = 0.95$ and $A_Y = 0.76$. Another criterion representing the annotation reliability based on the inter-participant agreement is the Fleiss' kappa measure. Again, we have calculated it separately for the phrase boundary labels and the semantic accents. The overall Fleiss' kappa values for 99 % confidence level (i.e. $a = 0.01$)

are $\kappa_X = 0.6636$ and $\kappa_Y = 0.1283$. Especially the value for the phrases (κ_X), which means that the agreement is much above chance, is very similar to what [5] shows. Even in spite of κ_Y being much smaller ([5] also reports κ smaller for prominences) we can expect the EM algorithm to produce reasonable labelling of the semantic accents as well.

Considering these values we can conclude the paper with a statement that we have obtained a reliable phrase annotation which has not been (to our knowledge) available for Czech so far and the correctness can be supported by the fact that the inter-participant reliability measures are in conformity with results of other researchers.

Table 4. Distribution of the number of semantic accents (SA) in a phrase and the number of phrases with the semantic accent placed on the last word.

	frequency (number of phrases)		
number of SA in phrase	whole test	part 1	part 2
0	537 (70.29 %)	224 (65.50 %)	313 (74.17 %)
1	227 (29.71 %)	118 (34.50 %)	109 (25.83 %)
SA on last word	181 (23.69 %)	100 (29.24 %)	81 (19.19 %)

References

1. Palková, Z.: Rytmická výstavba prozaického textu (with English resume: The rhythmical potential of prose). Academia, Prague (1974).
2. Yoon, T-J.; Chavarría, S.; Cole, J.; Hasegawa-Johnson, M.: Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. Proc. Interspeech. Jeju, Korea (2004) 2729–2732.
3. Wightman, C. W.: ToBI or not ToBI. Proc. Speech Prosody. Aix-en-Provence, France (2002) 25–29.
4. Buhmann, J.; Caspers, J.; van Heuven, V. J.; Hoekstra, H.; Martens, J-P.; Swerts, M.: Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. Proc. LREC. Canary Islands, Spain (2002) 779–785.
5. Mo, Y.; Cole, J.; Lee, E-K.: Naïve listeners' prominence and boundary perception. Proc. Speech Prosody. Campinas, Brazil (2008) 735–738.
6. Romportl, J.: Statistical evaluation of prosodic phrases in the Czech language. Proc. Speech Prosody. Campinas, Brazil (2008), 755–758.
7. Matoušek, J.; Romportl, J.: Recording and annotation of speech corpus for Czech unit selection speech synthesis. Lecture Notes in Artificial Intelligence, vol. 4629. Springer, Berlin-Heidelberg (2007) 326–333.