# Correction of prosodic phrases in large speech corpora

No Author Given

No Institute Given

**Abstract.** Nowadays, in many speech processing tasks, such as speech recognition and synthesis, really large speech corpora are utilized. These speech corpora usually contain several hours of speech or even more. To achieve possibly best results, an appropriate annotation of the recorded utterances is often necessary. This paper is focused on problems related to the prosodic annotation of the Czech speech corpora. In the Czech language, the utterances are supposed to be split by pauses into so-called prosodic clauses containing one or more prosodic phrases. The types of particular phrases are linked to their last prosodic words corresponding to various functionally involved prosodemes. The clause/phrase structure is substantially determined by the sentence composition. However, in real speech data, different prosodeme type or even phrase/clause borders can be present. This paper deals with 2 basic problems: the correction of the improper prosodeme/phrase type and the detection of new phrase borders. For both tasks, we proposed new procedures utilizing hidden Markov models. Experiments were performed on 4 large speech corpora recorded by professional speakers for the purpose of speech synthesis. These experiments were limited to the declarative sentences. The results were successfully verified by listening tests.

**Key words:** speech corpora, prosody, annotation

## 1 Introduction

Nowadays, in many speech processing tasks, such as speech recognition and synthesis, really large speech corpora are utilized. These speech corpora usually contain several hours of speech or even more. To achieve possibly best results, an appropriate annotation of the recorded utterances is often necessary.

In connection with using the large speech corpora, the automatic phonetic and prosodic annotation of speech [1, 2] became an important task. This paper deals with 2 basic problems: the correction of the improper prosodeme/phrase type and the detection of new phrase borders.

### 1.1 Prosody model

For our purposes, we used the formal prosody model proposed by Romportl [3]. On the basis of this model, an utterance is divided into prosodic clauses separated by short pauses. Each prosodic clause includes one or more prosodic phrases containing certain continuous intonation schemes. Furthermore, phrases are composed of prosodic words.

The communication function the speaker intends the phrase to have (or shortly the type of phrase) is linked with the last prosodic word in the phrase. For this purposes,

so called prosodemes are defined. The last prosodic word is linked with a functionally involved prosodeme, other words with null prosodemes. For the Czech language, the following basic classes of functionally involved prosodemes were defined [3]:

P1 – prosodemes terminating satisfactorily (in declarative sentences)
P2 – prosodemes terminating unsatisfactorily (in questions)
P3 – prosodemes non-terminating (in non-terminal phrases of compound sentences)

Since this research is limited to the declarative sentences and neutral speech (i.e. without emphasis, expressions etc.), prosodemes P0, P1.1 and P3.1 were applied. According to the theoretical assumption, all the compound sentences consist of several phrases, where the last one is terminated with prosodeme P1.1 and the other phrases ends with P3.1; see a simple example in Figure1.

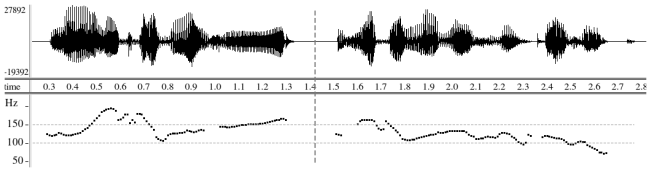


**Fig. 1.** Declarative compound sentence *"Málokdo věří, že by mohl zvítězit."* (in English *"Few believe that he could win."*). This prosodeme combination corresponds to the prosody model: the first phrase ends with P3.1 prosodeme and the last one with P1.1.

Particular prosodemes are linked with specific speech features: P1.1 is characteristic with a pitch decrease within its last syllable and a pitch increase is typical for P3.1. Beside the pitch shape (which is the most relevant), spectral features, duration and energy can be different for particular prosodemes. Naturally, particular types of phrases do not vary solely within their last prosodic words. Some specific prosodic differences can be present throughout the whole phrase. However, those differencies are often rather content-related (e.g. emphasis on some key words) and a more complex prosody model would be required. The utilized prosody model based seems to be sufficiently descriptive for the phrase type classification task [4, 5].

### 1.2 Problems in real speech

In real speech data, a different prosodeme than expected could be present. The most frequent case of this inconsistency is a compound sentence that can be split into several independent sentences. Within the compound sentence, all phrases (except the last one) should be terminated with the prosodeme P3.1. However, when the link between particular sentences is rather weak, the utterance can be split into independent sentences which are naturally terminated by the prosodeme P1.1. This is illustrated in Figure 2.

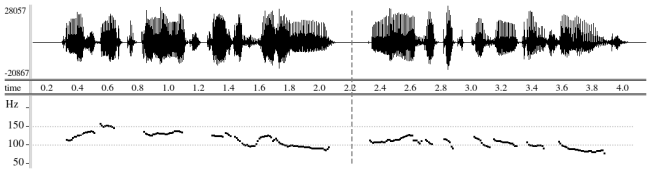


**Fig. 2.** Declarative compound sentence *"My jsme ekonomické oddělení, ne detektivní kancelář."* (in English *"We are the economic department, not a detective agency."*). The first phrase is terminated by an evident prosodeme P1.1.

In the Czech text, particular phrases are supposed to be separated by punctuation marks, usually commas[1]. Corresponding segments of speech are supposed to be prosodic phrases ended by functionally involved prosodemes. However, this theoretical assumption is not always fulfilled:

- Pauses can appear inside text phrases, especially when they are long.
- More text phrases can be uttered together without indication of any functionally involved prosodeme. However, the pause absence does not always lead to the absence of a functionally involved prosodeme; please compare Figures 3 and 4.

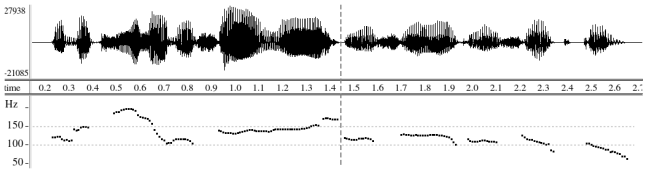


**Fig. 3.** Declarative compound sentence *"Aby cíle dosáhl, musí mít výsledky."* (in English *"To achieve the goal, the results are necessary."*). Though there is no pause, the prosodeme P3.1 terminating the first part is obvious.

Badly annotated speech corpora can be a source of various troubles. In speech synthesis (specifically, in unit selection method), prosodeme labels are important attributes for selecting sequence of optimal speech units for building resulting speech [6]. Using units from an inappropriate prosodeme or mixing units from different types of prosodemes can cause a degradation of the overall speech quality.

## 2 Proposed approach

To model the prosodic properties of speech we employed a similar HMM framework as it is specific for the HMM-based speech synthesis [7]. Speech was described by

[1] This is in contrast with English, where using commas has more complex rules. However, some copulative conjunctions in Czech are also used without a comma, e.g. *"a", "nebo", "ani"*, etc. (in English *"and", "or", "nor"*, respectively).

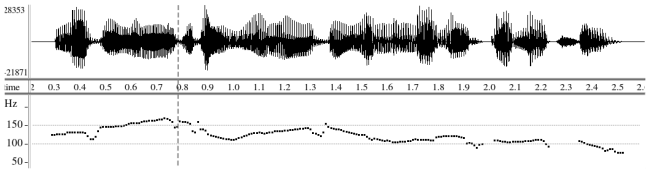


**Fig. 4.** Declarative compound sentence *"Nevím, kdo jiný by jim mohl pomoci."* (in English *"I don't know who else could help them."*). The punctuation in text has no evident impact on prosody realization; neither pause nor functional prosodeme are present.

a sequence of parameter vectors containing 40 mel cepstral coefficients obtained by STRAIGHT analysis method [8] and the fundamental frequency (log $F_0$) extracted by using the PRAAT software[2] [9]. The speech parameter vectors were modelled by a set of multi-stream context dependent HMMs by using the HTS toolkit[3].

In the HMM-based speech synthesis framework, the phonetic, prosodic and linguistic context are taken into account, i.e. a speech unit is given as a phone with its phonetic, prosodic and linguistic context information. In this manner, the language prosody is modelled implicitly – in various contexts different units/models can be used. Within our experiments, a context-depended unit is represented by a string

$$\mathbf{a}_\ell\texttt{-}\mathbf{a_c}\texttt{+}\mathbf{a_r}\texttt{@P:}\mathbf{p_f}\texttt{\_}\mathbf{p_b}\texttt{@S:}\mathbf{s_{f1}}\texttt{|}\mathbf{s_{f2}}\texttt{\_}\mathbf{s_{b1}}\texttt{|}\mathbf{s_{b2}}\texttt{@W:}\mathbf{w_f}\texttt{\_}\mathbf{w_b}\sim\mathbf{p_x}$$

where all subscripted bold letters are contextual factors defined as

| | | |
|---|---|---|
| $\mathbf{a}_\ell$, $\mathbf{a_c}$, $\mathbf{a_r}$ | ... | left context, current phoneme and right context |
| $\mathbf{p_f}$, $\mathbf{p_b}$ | ... | forward and backward position of phone in prosodic word |
| $\mathbf{s_{f1}}$, $\mathbf{s_{b1}}$ | ... | forward and backward position of syllable in prosodic word |
| $\mathbf{s_{f2}}$, $\mathbf{s_{b2}}$ | ... | forward and backward position of syllable in phrase |
| $\mathbf{w_f}$, $\mathbf{w_b}$ | ... | forward and backward position of prosodic word in phrase |
| $\mathbf{p_x}$ | ... | prosodeme type |

### 2.1 Training stage

For our experiments, we used 4 large speech corpora (described hereinafter) recorded for the purposes of speech synthesis. At the beginning, all utterances were segmented to phrases only by detected pauses, i.e. all phrases correspond to clauses. This manner of phonetic annotation is also used in our unit selection TTS system [**?**], since functionally involved prosodemes are ensured at the end of all phrases.

**Model training** Model parameters were estimated from the speech data by using maximum likelihood criterion. 3-state left-to-right MSD-HSMM with single Gaussian output distributions were used. For a more robust model parameter estimation, the context clustering based on the MDL (Minimum Description Length) criterion was performed. In this stage, the default prosodic annotation of particular phrases is used.

[2] Praat: doing phonetics by computer, `www.praat.org`

[3] HMM-based Speech Synthesis System (HTS), `http://hts.sp.nitech.ac.jp`

**Prosodeme correction** This procedure is a modified version of a more general method described in [5]. First, each individual phrase terminated by the prosodeme P3.1 is transcribed by using the prosodeme P1.1, i.e. the default and new transcriptions differ only in the prosodeme contextual factors within the last prosodic word; it is analogous to the example in Table 1, but simpler. Then corresponding speech features are successively forced-aligned with both transcriptions and the transcription with the best value of alignment score is selected for the given phrase.

When a new corrected transcription of all utterances is available, the whole process can be run iteratively. This correction procedure works on the assumption that most utterances correspond to the theoretical prosody model with some rare exceptions. Then, the trained speech HMMs are correct and can be used to reveal and correct those exceptions. However, problems can occur in the case of less consistent prosody in speech, since inconsistencies can cumulate, a part of models can be badly trained and some prosodemes can be changed incorrectly.

To cope with that, an additional step is performed at the end of each iteration. In this step, the prosodeme correction procedure is performed by using HMMs from another speaker. Only corrections performed by both its own and other speaker's models are kept, other changes are annulled, therefore this step is referred to as the *annulling step*.

**Splitting phrases by punctuation** First, each individual phrase containing a comma is split into particular phrases terminated by prosodemes P3.1 (excluding the last phrase, naturally). A simple example is presented in Table 1. When more commas are present in the phrase, all possible split combinations have to be taken into account. Again, the corresponding speech features are successively forced-aligned with all transcriptions and the transcription with the best value of alignment score is selected.

**Table 1.** An example of splitting utterances by the punctuation into phrases: *"Řekl, že přijde."* (in English *"He said that he will come."*). Its phonetic transcription "`$|RekL|Ze|pQijde|$`" including word separator "`|`" and formal border pauses "`$`". Changed contextual factors are bold.

| phns | default (one phrase) | split phrases |
|---|---|---|
| $ | $ | $ |
| R | $-R+e@P:1_4@S:0\|0_2\|**5**@W:1_**2**∼**0** | $-R+e@P:1_4@S:0\|0_2\|**2**@W:1_**1**∼**31** |
| e | R-e+k@P:2_3@S:1\|1_2\|**5**@W:1_**2**∼**0** | R-e+k@P:2_3@S:1\|1_2\|**2**@W:1_**1**∼**31** |
| k | e-k+L@P:3_2@S:1\|1_1\|**4**@W:1_**2**∼**0** | e-k+L@P:3_2@S:1\|1_1\|**1**@W:1_**1**∼**31** |
| L | k-L+Z@P:4_1@S:2\|2_1\|**4**@W:1_**2**∼**0** | k-L+Z@P:4_1@S:2\|2_1\|**1**@W:1_**1**∼**31** |
| Z | L-Z+e@P:1_8@S:0\|**2**_3\|3@W:**2**_1∼11 | L-Z+e@P:1_8@S:0\|**0**_3\|3@W:**1**_1∼11 |
| e | Z-e+p@P:2_7@S:1\|**3**_3\|3@W:**2**_1∼11 | Z-e+p@P:2_7@S:1\|**1**_3\|3@W:**1**_1∼11 |
| p | e-p+Q@P:3_6@S:1\|**3**_2\|2@W:**2**_1∼11 | e-p+Q@P:3_6@S:1\|**1**_2\|2@W:**1**_1∼11 |
| Q | p-Q+i@P:4_5@S:1\|**3**_2\|2@W:**2**_1∼11 | p-Q+i@P:4_5@S:1\|**1**_2\|2@W:**1**_1∼11 |
| i | Q-i+j@P:5_4@S:2\|**4**_2\|2@W:**2**_1∼11 | Q-i+j@P:5_4@S:2\|**2**_2\|2@W:**1**_1∼11 |
| j | i-j+d@P:6_3@S:2\|**4**_1\|1@W:**2**_1∼11 | i-j+d@P:6_3@S:2\|**2**_1\|1@W:**1**_1∼11 |
| d | j-d+e@P:7_2@S:2\|**4**_1\|1@W:**2**_1∼11 | j-d+e@P:7_2@S:2\|**2**_1\|1@W:**1**_1∼11 |
| e | d-e+$@P:8_1@S:3\|**5**_1\|1@W:**2**_1∼11 | d-e+$@P:8_1@S:3\|**3**_1\|1@W:**1**_1∼11 |
| $ | $ | $ |

## 3 Evaluation and results

For our experiments, we used 4 large speech corpora recorded for the purposes of speech synthesis: 2 male voices (denoted as AJ and JS) and 2 female voices (denoted as KI and MR). Each corpus contained about 10,000 declarative sentences. The detailed description of experimental data is present in Table 2.

**Table 2.** Description of experimental data. Please note that the total number of phrases is given as phrases ended by a comma + phrases without any end punctuation + phrases ended by a dot (equal to the number of utterances).

| speaker | | AJ | JS | KI | MR |
|---|---|---|---|---|---|
| utterances | | 9,996 | 9,846 | 9.896 | 9,878 |
| commas | total | 11,400 | 10,851 | 10,841 | 11,249 |
| | inside phrases | 1,998 | 1,001 | 8,503 | 5,013 |
| phrases | total | 22,971 | 20,097 | 13,166 | 18,236 |
| | ended by comma | 9,381 | 9,847 | 2,332 | 6,217 |
| | without end punctuation | 3,594 | 404 | 938 | 2,141 |

Although the sets of sentences are almost the same, some statistics are very different. This indicates various speaking styles of particular speakers. For example, the number of commas inside phrases corresponds how often speakers join text segments separated by a comma into one phrase. By contrast, the number of phrases without any end punctuation tells how often speakers make pauses inside continuous text segment.

To illustrate the prosody consistency of particular speakers, we performed one iteration of the proposed correction procedure without the annulling step. The higher number of changes is, the lower the consistency is supposed to be – see Table 3.

**Table 3.** The initial number of prosodemes and the number of P3.1 $\rightarrow$ P1.1 changes.

| speaker | AJ | JS | KI | MR |
|---|---|---|---|---|
| # P1.1 prosodemes | 9,996 | 9,846 | 9,896 | 9,878 |
| # P3.1 prosodemes | 12,974 | 10,250 | 3,269 | 8,358 |
| # changes | 114 | 60 | 21 | 452 |

The iterative correction procedure with annulling step was tested only on voices AJ (male) and MR (female). The annulling step was performed by using models from JS and KI, since these voices seem to be more consistent and therefore their models are expected to be more robust. Results are presented in Table 4.

Splitting phrases by punctuation was performed for all speakers, results are presented in Table 5. Since this splitting procedure is presented as fully new, we did not perform iterations, nor the annulling step in our experiments.

**Table 4.** Changing prosodemes P3.1 → P1.1: the initial number of prosodemes and the number of changes in particular iterations. Please remember that the corrections are always performed on the default corpora (the correction procedure is not cumulative).

| speaker | # P1.1 | # P3.1 | # changes | | |
|---|---|---|---|---|---|
| | | | iter.1 | iter.2 | iter.3 |
| AJ | 9,996 | 12,974 | 49 | 56 | 59 |
| MR | 9,878 | 8,358 | 223 | 257 | 273 |

**Table 5.** Splitting utterances by the punctuation: the initial number of phrases and P3.1 prosodemes and the number changes. The number of changes affects equally both phrases and prosodemes since each splitting produces a new phrases ended with the P3.1 prosodeme.

| speaker | | AJ | JS | KI | MR |
|---|---|---|---|---|---|
| # phrases | | 9,996 | 9,846 | 9,896 | 9,878 |
| # P3.1 prosodemes | | 12,974 | 10,250 | 3,269 | 8,358 |
| # changes | annulled | 154 | 47 | 412 | 813 |
| | performed | 245 | 116 | 524 | 714 |

### 3.1 Listening tests

The suitability of the performed corrections was verified by one overall listening test. It contained 120 individual utterances with one selected prosodic word. Listeners should select a proper prosodeme linked to this word; 5 choices were available

- it is definitely a P1.1 prosodeme
- it is probably a P1.1 prosodeme
- it is definitely a P3.1 prosodeme
- it is probably a P3.1 prosodeme
- it is a null prosodeme (or an indecisive case)

The test contained 40 utterances (20 from speakers AJ and MR) for the evaluation of the prosodeme changing procedure

- $2 \times 10$ utterances where a P3.1 to 1.1 correction was performed
- $2 \times 10$ utterances where that correction was annulled in the second stage

The remaining 80 utterances (20 from each speaker) were intent for the evaluation of the splitting procedure

- $4 \times 10$ utterances that were additionally split by a comma (split utterances)
- $4 \times 10$ utterances that contain a comma, but the splitting was not performed (non-split utterances)

All the utterances were mixed together, i.e. speakers and issues took turns randomly. Sentences were selected to be short and simple like the examples in Figures 1-4. Five participants took part in this test, all of them were speech processing experts capable to distinguish various prosodeme types.

**Changed prosodemes** The distribution of listeners' choices is present in Figure 5 and Table 7. The most relevant entries are the percentages of changed prosodemes that were marked as P3.1: 90% and 76% for AJ and MR, respectively. The other 6% and 20% were marked as P1.1 and the remaining 4% (equally for both speakers) were indecisive cases. Since only 3 iterations of correction procedure were performed and it wasn't the final state, further improvement could be expected.

As was explained in Section 2, the purpose of annulled changes is to increase the robustness within several initial iterations of the correction procedure. All those changes can be still applied in the following stage without the annulling step. Anyway, the more annulled cases really does not match the desired prosodeme the more beneficial the annulling step is. In our case, this rate is about 82% and 62% (all non-P1.1 cases).

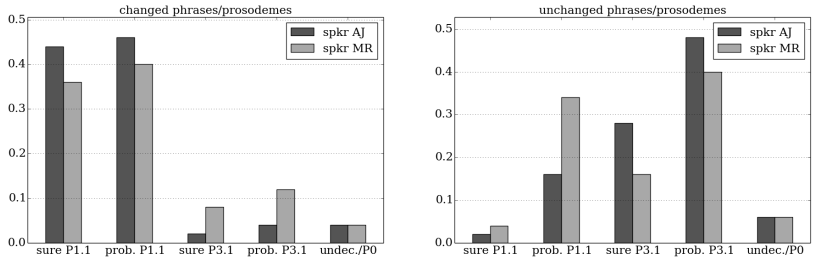


**Fig. 5.** Results of listening test on changing prosodemes P3.1 → P1.1.

**Table 6.** Results of listening test on changing prosodemes P3.1 → P1.1: percentage of particular listeners' choices. The agreement between human listeners and the proposed correction procedure is expressed mainly by the bold values.

| phrases | speaker | prosodeme P1.1 | | | prosodeme P3.1 | | | P0 |
|---|---|---|---|---|---|---|---|---|
| | | sure | probably | total | sure | probably | total | |
| changed | AJ | 44.0 | 46.0 | **90.0** | 2.0 | 4.0 | 6.0 | 4.0 |
| | MR | 36.0 | 40.0 | **76.0** | 8.0 | 12.0 | 20.0 | 4.0 |
| | all | 40.0 | 43.0 | **83.0** | 5.0 | 8.0 | 13.0 | 4.0 |
| unchanged | AJ | 2.0 | 16.0 | 18.0 | 28.0 | 48.0 | 76.0 | 6.0 |
| | MR | 4.0 | 34.0 | 38.0 | 16.0 | 40.0 | 56.0 | 6.0 |
| | all | 3.0 | 25.0 | 28.0 | 22.0 | 44.0 | 66.0 | 6.0 |

**Split phrases** Results of listening test are presented in Figure 6 and Table 7. A high consistency between listeners and the proposed procedure is evident: prosodemes in split utterances were annotated as definitely or probably P3.1 in about 88% cases for all speakers (ranged between 84% for KI and 91% for AJ). Surprisingly, an appreciable

amount of P1.1 prosodemes appeared in listeners' selections. Actually, it is in accordance with the experiment on changing prosodemes and P1.1s could be expected here, too.

The actual benefit of the splitting procedure should be also apparent by a comparison of results for the split and non-split utterances. Above all, significantly less P3.1s and more null prosodemes should be present in non-split sentences. This is true, nevertheless the number of P3.1 prosodemes in non-split utterances is higher than expected, especially 72% for JS. The reason could be also the influence of the sentence structure on the listeners' decision; moreover it evidently depends on the actual speaker, too.

The splitting procedure could be also simply modified to work iteratively (similarly as the procedure for changing the type of prosodeme). Then a lower number of prosodemes P3.1 could be expected in non-split utterances.

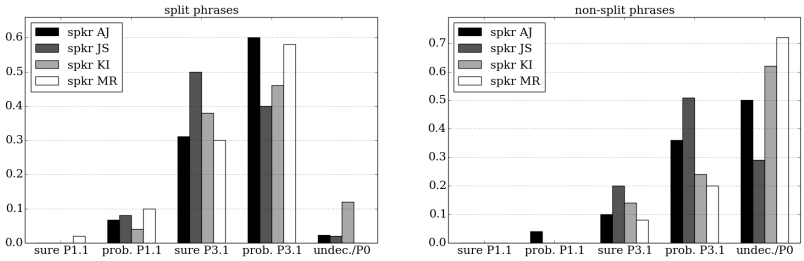


**Fig. 6.** Results of listening test: splitting utterances into phrases by punctuation.

**Table 7.** Results of listening test on splitting phrases by punctuation: percentage of particular listeners' choices. The agreement between human listeners and the proposed splitting procedure is expressed mainly by the bold values.

| phrases | speaker | prosodeme P1.1 | | | prosodeme P3.1 | | | P0 |
|---|---|---|---|---|---|---|---|---|
| | | sure | probably | total | sure | probably | total | |
| split | AJ | 0.0 | 7.0 | 7.0 | 31.0 | 60.0 | **91.0** | 2.0 |
| | JS | 0.0 | 8.0 | 8.0 | 50.0 | 40.0 | **90.0** | 2.0 |
| | KI | 0.0 | 4.0 | 4.0 | 38.0 | 46.0 | **84.0** | 12.0 |
| | MR | 2.0 | 10.0 | 12.0 | 30.0 | 58.0 | **88.0** | 0.0 |
| | all | 0.5 | 7.3 | 7.8 | 37.3 | 51.0 | **88.3** | 4.0 |
| non-split | AJ | 0.0 | 4.0 | 4.0 | 10.0 | 36.0 | 46.0 | 50.0 |
| | JS | 0.0 | 0.0 | 0.0 | 22.0 | 51.0 | 73.0 | 29.0 |
| | KI | 0.0 | 0.0 | 0.0 | 14.0 | 24.0 | 38.0 | 62.0 |
| | MR | 0.0 | 0.0 | 0.0 | 8.0 | 20.0 | 28.0 | 72.0 |
| | all | 0.0 | 1.0 | 1.0 | 13.5 | 32.8 | 46.3 | 53.3 |

## 4 Conclusion

This paper presented 2 procedures for the correction of the type and borders of prosodic phrases in large speech corpora. Experiments were performed on 4 corpora. Although all contained almost equal sentences and were recorded by professional speakers, the prosody structure of recorded utterances, its consistency and the corresponding number of performed corrections varied for particular speakers.

The results have been verified in a listening test. The agreement between the test participants and the proposed procedures was about 83% for changing the prosodeme type and 88% for splitting utterances into phrases by the punctuation.

In our future work, more experiments on prosodeme classification will be performed. Both proposed procedures should be joint together into one iterative correction process. The annulling step (or maybe the whole procedure) can be improved by employing speaker-independent models and their adaptation. Other types of phrases (e.g. various types of questions) will be included, too. A big challenge is the automatic prosody annotation of speech data, especially of non-professional speakers whose prosody could be problematic due to its bad consistency.

## References

1. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing **2** (1994) 469–481
2. Toledano, D., Gómez, L., Grande, L.: Automatic phonetic segmentation. IEEE Transactions on Speech and Audio Processing **11** (2003) 617–625
3. Romportl, J., Matoušek, J., Tihelka, D.: Advanced prosody modelling. In: Proceedings of the 7th International Conference TSD 2004. Volume 3206 of Lecture Notes in Artificial Intelligence. (2004) 441–447
4. Hanzlíček, Z., Grůber, M.: Initial experiments on automatic correction of prosodic annotation of large speech corpora. In: Text, Speech and Dialogue. Volume 8655 of Lecture Notes in Computer Science. Springer International Publishing (2014) 481–488
5. Hanzlíček, Z.: Classification of prosodic phrases by using hmms. In: Text, Speech and Dialogue. Volume 9302 of Lecture Notes in Computer Science. Springer International Publishing (2015) 497–505
6. Tihelka, D., Matoušek, J.: Unit selection and its relation to symbolic prosody: A new approach. In: Proceedings of Interspeech '06. (2006) 2042–2045
7. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication **51**(11) (2009) 1039–1064
8. Kawahara, H., Masuda-Katsuse, I., de Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication **27** (1999) 187–207
9. Boersma, P., van Heuven, V.: Praat, a system for doing phonetics by computer. Glot International **5**(9/10) (2001) 341–345