ON BUILDING PHONETICALLY AND PROSODICALLY RICH SPEECH CORPUS FOR TEXT-TO-SPEECH SYNTHESIS

Jindřich Matoušek[1]
[1]Department of Cybernetics
University of West Bohemia
Plzeň, Czech Republic
email: jmatouse@kky.zcu.cz

Jan Romportl[1,2]
[2]Institute of Formal and Applied Linguistics
Charles University
Prague, Czech Republic
email: rompi@kky.zcu.cz

ABSTRACT
This paper proposes a way of preparing and recording a speech corpus for unit selection text-to-speech speech synthesis driven by symbolic prosody. The research is focused on a phonetically and prosodically rich sentence selection algorithm. Symbolic description on a deep prosody level is used to enrich the phonetic representation of sentences (by respecting the prosodeme types phones appear in). The resulting algorithm then selects sentences with respect to both phonetic and prosodic criteria. To cover supra-sentential prosody phenomena, paragraphs were selected at random and recorded as well. The new speech corpus can be utilised in unit selection speech synthesis and also for training a data-driven prosodic parser.

KEY WORDS
natural language processing, text-to-speech, speech synthesis, sentence selection, speech corpus, prosody

1 Introduction

In today's world, natural language processing and voice technologies play an important role. Text-to-speech synthesis (TTS), along with automatic speech recognition (ASR), ranks among the most important applications of natural language processing and voice technologies. Since they aim to make our lives more comfortable, we can encounter them more often in our everyday lives. For example, we can utilise an ASR system to dictate a letter or to summarise talks using keywords, or we can employ a TTS system to read SMS in handhelds, e-mails and other e-documents aloud. There are usually two criteria of evaluating TTS systems: firstly, synthetic speech must be *intelligible* (depends on segmental quality, i.e. phonetic segments); secondly, it should sound *natural* (depends on suprasegmental, i.e. prosodic characteristics). Although there are many TTS applications available today, the quality of synthetic speech is still somewhat limited, in particular from the point of view of the naturalness. After the quality is improved in that the naturalness of the resulting speech approaches the quality of real human speech, the massive usage of TTS technologies in real life applications could be expected.

Generally, the quality of synthetic speech produced by a *concatenation-based synthesis system* crucially depends on the quality of an *acoustic unit inventory*. Several factors contribute to the quality of the acoustic unit inventory, such as *speech corpus* from which the units are extracted, the type of the unit (i.e. phone, diphone, triphone etc.), labelling accuracy, the number of instances per each unit, prosodic richness of each unit etc. This paper proposes a way of preparing and recording a both phonetically and prosodically rich speech corpus for the use in concatenative corpus-based TTS synthesis applications.

The importance of a speech corpus is constantly increasing in modern *corpus-based speech synthesis systems*, where a trend is to extract relevant linguistic knowledge from the corpus. The relevant linguistic knowledge (including phonetic and prosodic richness) should be encoded in the corpus and exploited later during speech synthesis. Obviously, great attention should be paid to the speech corpus preparation and design processes.

Although the prosodic richness of a speech corpus is also very important (especially in the case of unit selection speech synthesis techniques where the amount of prosodic modifications is desired to keep at minimum, if any), prior works have focused mainly on the phonetic coverage. In our previous work [1] we tried to propose a sentence selection algorithm to select sentences with each triphone occurring "sufficiently enough". Although only 3 occurrences per each triphone were required, the desired distribution of triphones was not achieved mainly due to rare triphones together with the requirement of having a "reasonable" number of 5,000 sentences. François & Boëffard solved the problem of sentence selection as a set covering problem and proposed a method to select sentences with the uniform distribution of the most frequent triphones [2]. Rare triphones (defined as triphones with less than 10 tokens) were considered to be unimportant or even unwanted as they could cause some problems in the subsequent processing stages (e.g. they could be difficultly modelled using statistical methods), and so sentences containing those triphones were excluded from the selection. As a result, 12,217 sentences were selected. Having the ease of recording in mind, Kominek & Black in [3] proposed a method for selecting a compact set of about 1200 "nice" sentences containing at least one occurrence of every diphone. The mentioned methods of sentence selection did not take prosodic coverage into account (except for word stress which was used to distinguish stressed vowels from unstressed ones in [3]). In this paper, a sentence selection method respecting both phonetic and prosodic coverage is presented.

The paper is organised as follows. Section 2 describes text corpus related issues, including sentence collection and preprocessing. The basic sentence selection algorithm is also depicted in this section. Section 3 deals with extensions of the basic selection algorithm with respect to the prosodic events. In Section 4 a few remarks on recording the selected sentences are mentioned. Finally, Section 5 contains the conclusion and outlines our future work in this field.

2 Text corpus

Before the sentence selection is performed, one must provide a large number of sentences from which to select. Ideally, the sentences should cover the domain for which the synthesiser is supposed to be employed (so-called target domain) both in the sense of the vocabulary and the style of speaking. In general TTS systems it is hard to determine the target domain because the system should be ideally "good enough" for every text that could appear at the input of the system. We decided to use sentences from several Czech newspapers (and various domains like news, sport, culture, economy, etc.). Consequently, the synthesiser should cover various topics and deliver speech in news-like style.

2.1 Sentence preprocessing

In order to have all input sentences in a desired format (a plain text file with a single sentence on each line) and also to have the representation of these sentences in the form suitable for the selection (the phonetic or pronunciation form in the case of our basic algorithm), several auxiliary steps should be carried out:

1. **From HTML to plain text.** Since texts were obtained in HTML format (each HTML file containing one or more articles), the first task was to divide each HTML file into particular articles and to transform each article to a plain text in a sentence-by-sentence manner [4]. In this way, 1,114,747 sentences were collected.

2. **Excluding "non-standard" words.** To facilitate reading and to avoid reading errors, sentences with "non-standard" words were excluded from the set of collected sentences. The exclusion was done using a vocabulary created during the collection of the sentences. Each word from the sentences is present in the dictionary together with an information about its type (4 types are distinguished: regular words, exceptional words, foreign words (also including foreign names and geographical terms, figures, and abbreviations).

   As foreign words and abbreviations are prone to mispronunciation and figures could be pronounced in different ways, they were marked as non-standard and sentences containing these words were excluded. The number of sentences available for the next processing was then reduced to 638,678 in this step.

3. **Various filtering & normalisation.** In this step, another adjustments were performed to normalise the sentences:

   - replacing multiple space with a single space
   - removing sentences with "expressive" words
   - normalising punctuation (previous word should be followed by a punctuation mark with no space between and after this mark a single space should be present right in front of the next word)
   - removing duplicate sentences.

   All these adjustments should improve the readability of the sentences. The number of sentences was reduced to 524,472.

4. **Sentence structure parsing.** Rule-based punctuation- and conjunction-driven parsing was implemented to estimate the prosodic structure of sentences [5]. As a result, symbolic description on a deep prosody level (mainly boundaries between clauses, the major groups of words bounded by a pause or breath) of each sentence is obtained.

5. **Phonetization.** Since the sentence selection algorithm selects sentences based on some phonetic criteria (see Section 2.2), grapheme-to-phoneme rules [6] were employed to convert the textual sentences (i.e. the written form) to the phonetic (i.e. pronunciation) form. In addition, word stress and the corresponding prosodic words (PW, one or more words belonging to a word stress [5]) are estimated as well. Such a phonetic representation is enriched with prosodic features further in Section 3.1 and used then in the extended sentence selection algorithm.

2.2 Sentence selection

There is usually an effort to cover significant linguistic events in the selected sentences. Traditionally, phonetic criteria are taken into account and the aim is to select sentences that follow the desired distribution of phonetic units. In the basic algorithm described in this section, phones and diphones were taken into account. An extension of the algorithm, which copes with prosodic features of speech, is proposed in Section 3.1.

In speech recognition tasks, *naturally balanced sentences*, which contain phonetic events with respect to their frequency in natural speech (so-called *phonetically balanced sentences*), are often utilised [4]. On the other hand, different sentence selection techniques are usually employed for the purposes of speech synthesis [1, 2, 3] where sentences containing all phonetic events with as much uniform distribution as possible are to be selected (so-called *uniformly balanced sentences* or, more frequently, *phonetically rich sentences*) [7].

The question is how to deal with rare units. In fact, they could be modelled in the same way as the other units [1] or they could be excluded from the selection scheme [2]. We decided to count them in, because we believe synthesising even such rare units could be important for listeners when they evaluate the overall quality of a speech synthesis system. Unlike [3] we do not restrict the sentences strictly to be easily readable (except for foreign names, etc., see Section 2.1), because not easily readable sentences often contain rare units. The length of sentences was also chosen less restrictive: sentences that are not between 3 and 30 prosodic words were filtered out (sentences in this range were found to cover all relevant language phenomena while still being reasonably readable). However, such loose criteria could complicate the recording process because it could be hard to read such sentences consistently and correctly. Hence, special attention should be paid to the recording process (see Section 4).

The upper boundary of 30 prosodic words was determined according to the following criterion: let $p(i)$ be the relative frequency of a sentence with i prosodic words. The test function is

$$C(i) = \log \frac{|p(i) - p(i-1)|}{|p(i-1) - p(i-2)|}. \tag{1}$$

If the relative frequency of the sentences with the length i is comparable to those with the length $i-1$ (i.e. the difference is "small enough"), the absolute value of $C(i)$ is small, whereas bigger absolute values of $C(i)$ imply that the frequency of the sentences with i prosodic words does not follow the given tendency and thus their deployment in the data might not be "stable" enough. The mean is $E\left\{C(i)\right\} = -0.07$ and the standard deviation $\sigma\left\{C(i)\right\} = 0.4$. We have excluded the sentences with the length $i \geq m$ for the first $m \notin \langle E - \sigma; E + \sigma \rangle$. In our case $m = 30$.

The lower boundary of 3 prosodic words was determined experimentally: after thorough exploration of the source text we have found that 1- and 2-word "sentences" are often rather only sentence-like structures disallowing reasonable uttering (such as standalone names, various codes or notations, etc.).

The sentence selection algorithm is based on a modified version of *greedy algorithm* [1, 4]. In addition to the requirement of phonetically rich sentences, we also want all phonetic units in the list of the sentences selected so far to occur at least P-times. Phones as the phonetic units were used for this "preselection".

1. The sentence with the highest number of different phones which do not occur P-times in the sentences selected so far is moved from the input list of sentences to the list of up to now selected sentences.

2. If any sentence in the sentences selected so far could be removed so that the frequency of no phone falls below the determined threshold P, the sentence is moved from the list of sentences selected so far back to the list of the input sentences.

3. The steps 1 and 2 are repeated until all phones occur at least P-times in the selected sentences.

The algorithm then starts to balance the distribution of phonetic units. Unlike [1], where it was very hard to find a reasonable number of sentences with uniform distribution of triphone occurrences, less combinatoric units, diphones, are used here for balancing. Moreover, diphones are to be used as the basic units in our unit selection speech synthesis based on the speech corpus proposed in this paper. The algorithm of the selection of the phonetically rich sentences (with respect to diphones) has 4 steps:

4. For each phonetically transcribed sentence in the set of sentences selected so far (including the preselected sentences) a score S defined in equation (2) is computed. It reflects how well diphones contained in the sentence are represented in the sentences selected so far.

5. The sentence with the maximum score S is selected and moved to the list of sentences selected so far.

6. A recount is performed for the sentences selected so far and the sentence with the worst score is moved back to the list of the input sentences. However, if the removal causes the occurrence of any phone to fall below the specified threshold P, the sentence is not excluded and the algorithm goes to the step 7. Similarly, the exclusion takes effect only in the case that the sentence to be excluded differs from the last added sentence, i.e. from the sentence selected in the step 5 (the sentence would be selected again in the next iteration).

7. The steps 4 to 6 are repeated until the desired number of sentences is selected or until the desired score is achieved.

The score S is computed as the entropy of the inspected sentence and the sentences selected so far. The entropy reaches the maximum when all diphones occur equally in the selected sentences. The score is computed:

$$S = -\sum_{i=1}^{I} \frac{n_i + n'_i}{n} \, \log_2 \frac{n_i + n'_i}{n}, \tag{2}$$

where

$$n = \sum_{i=1}^{I} (n_i + n'_i), \tag{3}$$

where I is the number of different diphones that we wish to have in the selected sentences, n_i is the frequency of the i-th diphone in the up to now selected sentences, n'_i is the

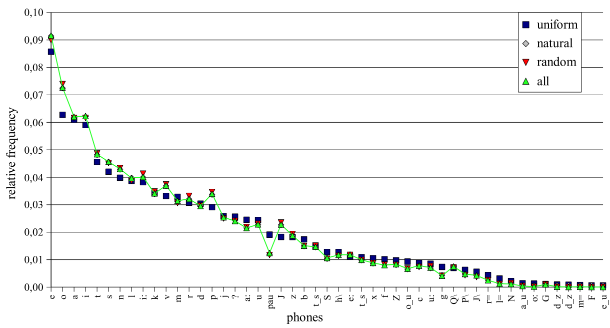

Figure 1. Distribution of phones in uniformly and naturally balanced sentences, randomly selected sentences and all sentences in the text corpus.

frequency of the i-th diphone in the inspected sentence and $0 \log_2 0 \equiv 0$. In our case, the sentence selection algorithm was stopped when a "reasonable" number of sentences N (in the sense of the feasibility of their recording, see Table 2) were selected.

As can be seen in Fig. 1 (where similar distributions in naturally and randomly selected sentences and also in all sentences available in the text corpus are highlighted by a line), the distribution of units in phonetically rich sentences cannot obviously be uniform because of language rules the sentences must obey (lexical, phonological, syntactical, etc.). Consequently, when a sentence with a rare unit is chosen, other more frequent units in this sentence are chosen too, thus also reflecting the natural distribution of units in real speech. As a result, the final distribution of unit occurrences resembles the distribution in phonetically balanced sentences with the exception that the rare units occur a little more frequently at the expense of the most frequent units being slightly less frequent. The characteristics of diphones in the selected sentences are shown in Table 1).

Table 1. *Characteristics of diphones in the selected sentences.*

diphone	text corpus	selected sentences		
		uniform	natural	random
coverage	100%	91%	84%	82%
avg. freq.	0.066%	0.072%	0.078%	0.080%
stdev. freq.	0.139%	0.126%	0.147%	0.150%
entropy	8.78	9.07	8.80	8.76

3 Extensions of the basic selection algorithm

In this section, two extensions of the basic sentence selection algorithm described in Section 2.2 are presented. Both of them incorporate prosodic features, the one in Subsection 3.1 proposes to include prosodic features directly into sentence selection algorithm and in Subsection 3.2 an effort to take supra-sentential prosodic phenomena into account is discussed.

3.1 Incorporating prosodic features

The building of a corpus for unit selection speech synthesis also requires having every phonetic unit included in the corpus in various prosodic contexts. Such a requirement, however, together with the standard phonetic coverage requirement would result in an enormous number of sentences to record. Maybe, that is why prior works usually ignored prosodic criteria and the sentence selection algorithm was driven only by the phonetic requirements without any possibility to affect the occurrence of units in different prosodic contexts. Since the richness of various prosodic contexts is very important in unit selection speech synthesis, especially in those approaches where the modifications of speech signals are avoided or at least kept to the minimum, our effort was to take prosodic features into account when selecting sentences to record. In our approach, instead of the most intuitive specification of explicit prosodic contours, a symbolic description using *prosodic phrase grammar* is used to parse every sentence on a deep prosody level [5]. The advantage is that the symbolic prosodic description is general enough not to complicate the sentence selection algorithm with many new requirements. Moreover, our unit selection speech synthesis system is based on the symbolic prosodic description – it does not employ any explicit prosodic contours when generating the resulting speech [8].

It is also important to note that many natural language processing techniques are tailored to English while omitting phenomena specific for other languages. When processing the Czech language (or actually all Slavic languages) we have to cope with very rich word flexion and partially free word order (which – together with prosody – influences the meaning of sentences). On the other hand the Czech language has almost regular stress placement which means the stress is less functionally involved than in English and thus we do not have to analyse for example its position within prosodic phrases when balancing the speech corpus.

Prosodemes, abstract prosodic units established in certain communication functions within the language system [5], were used to distinguish diphones. The prosodeme as a distinctive feature was preferred to word stress, because word stress is not as distinct, does not distinguish words and its usage is fairly regular in Czech, as it was already mentioned. For the purposes of the sentence selection we distinguish 6 types of prosodemes: declarative, "expressive" (imperative or optative), inquiring and supplementary interrogative (all of these being terminating prosodemes), non-terminating and "null" prosodemes [5]. Each word of a sentence belongs to a certain prosodeme according to the rules of the prosodic phrase grammar. Since there is a huge amount of declarative sentences in the source text corpus – more than 96 % of all sentences are the declarative ones (DEC), inquiring ("yes-no") questions (QIN) take approximately 3 %, the rest is formed by supplementary ("wh-") questions (QSU) and "expressive" sentences (EXP), we decided not to copy the distribution of the sentence types in the selected sentences because the prosodic richness of the sentences other than DEC cannot be satisfactorily underlaid by such a number while keeping also the segmental richness. Instead, we reinforced the number of non-declarative sentences to select and divided the sentences available in the text corpus into 4 groups according to the sentence types (i.e. according to terminating prosodemes). The sentence selection algorithm described in Section 2.2 was then run in parallel for each group. More detailed specifications can be found in Table 2. For QSU and EXP, where a small number of sentences were available in the source text corpus, no diphone balancing was carried out – just phone preselection was performed (marked by '*' in Table 2).

This way each diphone is differentiated into 6 types according to the prosodeme it appears in. The sentence selection algorithm then works with the text data represented by this extended set of diphones. Its effort to balance the diphone occurrences thus also implicitly leads to better prosodic balancing. Although the prosodeme placement itself is carried out on the text automatically by a rule-based algorithm (obviously, it must be done before the sentences are actually uttered) and, therefore, the final utterances often prosodically differ (due to the influences of the speaker) from what was expected during the balancing process, the sentence selection algorithm using the extended diphone set still selects diphone instances in as many prosodic situations (i.e. prosodemes) as possible.

Table 2. *Sentence selection for different types of sentences. The ad hoc selected numbers N and P correspond to those in Section 2.2.*

sentence type	text corpus sentences		selected sentences		
	number	coverage	N	coverage	P
DEC	506,090	96.50%	3,500	69.71%	50
QIN	14,697	2.80%	900	17.92%	12
QSU*	1,712	0.33%	310	6.17%	20
EXP*	1,973	0.38%	311	6.19%	20

3.2 Taking supra-sentential prosodic phenomena into account

Semantic, pragmatic and thus also prosodic forms of an utterance depend on a broader context of the utterance itself. A sentence uttered apart from its context is an artificial construct and the speaker has no cues for its proper prosodic articulation. Crossing the sentence boundary can be linguistically described in terms of *topic-focus articulation* – its relation to prosody is formally described by so-called *prosodic structures* in [5]. In our opinion, up-to-date TTS systems should reflect this because incompatibility between generated prosody of the synthetic speech and the context-given semantic configuration of the sentence can result in subtle, yet serious phenomena evaluated as unnatural and unfit.

Concerning the aforementioned, we have decided to record continuous portions (i.e. paragraphs) of speech crossing the sentence boundary so as to obtain suitable speech data for the prosodically natural unit selection synthesis as well as the material for further research of *prosodic homonymy* and the development of a prosodic structure parser [5].

Each paragraph consists of four subsequent sentences from the text data described above (contrary to the isolated sentence selection, no restrictions on sentence lengths were imposed and the sentences with figures, abbreviations and foreign words were not excluded; figures and abbreviations being transcribed in their full word forms). The goal is to select prosodically balanced and rich paragraphs as well (i.e. covering as many suprasegmental phenomena as possible) but unlike the isolated sentences, the paragraphs should be prosodically annotated according to the really uttered prosodic forms after being recorded (for the sake of the prosodic parser training). Since at the time of the paragraph selection process we have no prior knowledge of the possible prosodeme distribution (it is to be discovered from the data we will have acquired this way), we keep scientific methodological attitude of maximum entropy for situations where constraints are present in a system but are not explicitly known. The same is to be followed as to paragraph diphone balancing because due to the presence of foreign words and numerals we cannot estimate the error rate of the automatic transcription of these specific words. The text therefore includes an unknown (yet quite significant) number of wrong assessed diphones whose usage in diphone balancing would lead actually to random results. It means the best way to choose the paragraphs is to select them randomly in a uniform manner (i.e. no paragraphs were a priori preferred).

We have recorded 1,300 paragraphs giving the total number of 5,200 sentences. This data can be added to the isolated sentences to extend their number for the sake of the speech unit database. The length of the paragraph (4 sentences) was chosen according to the assumption that the sentence context potentially considered during the prosodic structure parsing can be limited to two preceding sentences. This way we have two subsequent sentences with their full considered context in each paragraph. Some statistics of the final speech corpus are shown in Table 3.

4 Recording

Sentences and paragraphs selected in the previous sections have been recorded in a soundproof studio located at our university. An AKG C 3000B large-diaphragm cardioid condenser microphone with a pop filter installed to reduce the force of air puffs emerging from bilabial plo-

Table 3. *Speech corpus statistics (P = phones, D = diphones).*

statistics		text corpus	selected sentences: sentences	paragraphs	total
# sentences		524,472	5,021	5,200	10,221
phones per sent.		77.5	57.6	83.4	70.62
PWs per sent.		10.0	7.3	10.8	9.04
P	types	49	49	49	49
	tokens	40,655,456	289,047	427,582	716,629
	avg. freq.	2.041%	2.041%	2.041%	2.041%
	stdev freq.	2.126%	1.934%	2.150%	2.060%
D	types	1,522	1,388	1,347	1,434
	coverage	100%	91%	89%	94%
	avg. freq.	0.066%	0.072%	0.074%	0.070%
	stdev freq.	0.139%	0.126%	0.147%	0.134%

sives and other strongly released stops was used. A high fidelity capture card capable of up to 96 kHz AD conversion (48 kHz AD conversion has been actually performed) was utilised. Glottal signal measured by an electroglottograph machine (suitable for the detection of glottal closure instants that are used for accurate pitch contours estimation, voiced/unvoiced signal detection, or smooth concatenation of speech segments in unit selection speech synthesis [8]) has been recorded along with the speech signal.

A female voice talent possessing a pleasant voice, good voice quality and professional recording experience was chosen to record the corpus. Being aware of the importance to keep the recordings consistent both in phonetic and prosodic (within the framework of symbolic prosody description) terms, an expert in acoustic phonetics and orthoepy supervised the recording; his job was to check the consistency of recordings and also the constancy of speaker's voice quality and pronunciation. The average duration of a recording session was about 4 hours which resulted in 25 recording sessions.

5 Conclusion & Future work

In this paper the issues related to the preparation and design of a speech corpus for the purposes of unit selection text-to-speech synthesis, an important application of natural language processing, were described. The focus was put on an algorithm for selecting phonetically rich sentences. Consequently, the basic algorithm was extended by respecting prosodic contexts of phonetic units. As for the prosodic features, symbolic description on a deep prosody level was adopted. The extended algorithm was then applied to uniform diphone balancing with the minimum phone-frequency preselection. In this way, 5,021 sentences were selected and recorded. To cover supra-sentential prosody phenomena in our speech corpus, 1,300 four-sentence paragraphs were selected at random and recorded as well. In the end, 10,221 sentences are available in the corpus.

At the time of writing, the recording of sentences and paragraphs is being finished. It is being prosodically annotated within the framework of our prosodic grammar and is going to be used to train a data-driven prosodic parser. Of course, it is going to be employed in our unit selection speech synthesis system driven by prosodic-grammar-based symbolic prosody. Obviously, various modifications of the sentence selection scheme presented here could be proposed, e.g. by defining different weights for phonetic and prosodic contexts, taking other prosodic phenomena into account, experimenting with different numbers of sentences, etc.

6 Acknowledgements

This research was supported by the Ministry of Education of the Czech Republic projects No. 2C06020 and LC536, and the EU 6th Framework Programme IST-034434.

References

[1] J. Matoušek, J. Psutka, J. Krůta, Design of speech corpus for text-to-speech synthesis, *Proc. Eurospeech*, Ålborg, Denmark, 2001, 2047–2050.

[2] H. François, O. Boëffard, Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem, *Proc. Eurospeech*, Ålborg, Denmark, 2001, 829–832.

[3] J. Kominek, A. Black, The CMU ARCTIC speech databases, *Proc. Speech Synthesis Workshop*, Pittsburgh, USA, 2004, 223–224.

[4] V. Radová, J. Psutka, UWB_S01 corpus – a Czech read-speech corpus, *Proc. ICSLP*, Beijing, China, 2000, 732–735.

[5] J. Romportl, Structural data-driven prosody model for TTS synthesis, *Proc. Speech Prosody*, Dresden, Germany, 2006, 549–552.

[6] J. Matoušek, J. Romportl, D. Tihelka, Z. Tychtl, Recent improvements on ARTIC: Czech text-to-speech system, *Proc. ICSLP/Interspeech*, Jeju, Korea, 2004, 1933–1936.

[7] D. Gibbon, R. Moore, R. Winski, *Handbook of standards and resources for spoken language systems* (Berlin: Mouton de Gruyter, 1997).

[8] D. Tihelka, Symbolic prosody driven unit selection for highly natural synthetic speech, *Proc. Eurospeech*, Lisbon, Portugal, 2005, 2525–2528.