Slovak Unit-Selection Speech Synthesis: Creating a New Slovak Voice within a Czech TTS System ARTIC

Jindřich Matoušek, Daniel Tihelka, Jan Romportl, and Josef Psutka

***Abstract*—ARTIC (Artificial Talker in Czech) is a corpus-based text-to-speech (TTS) system that enables to synthesise an arbitrary text, mainly for the Czech language. Basically, two versions of ARTIC are available—a single unit instance system (also known as fixed-inventory synthesis) with the quality of resulting speech limited by the fixed inventory, and multiple unit instance system with the quality profitting from employing a unit-selection algorithm to select the longest suitable sequence of phonetic units from many units available. In this paper, a process of building a new Slovak voice for the unit-selection version of ARTIC is presented. All steps in the design, from the preparation of a suitable speech corpus to the creation of an acoustic unit inventory of the new Slovak voice and its use in the ARTIC system will be described. Text processing module, including automatic phonetic transcription and symbolic prosodic description of an arbitrary Slovak text, will be detailed. Finally, speech production module based on the unit selection algorithm will be mentioned as well.**

***Index Terms*—text-to-speech, corpus-based speech synthesis, unit selection, Slovak language.**

I. Introduction

TEXT-TO-SPEECH (TTS) synthesis is one of the most important tasks of computer speech processing. Nowadays, corpus-based synthesis is the most widely used approach to speech synthesis. The current trend in this approach is to use large speech corpora and acoustic unit inventories to catch as many speech phenomena (i.e. spectral variations, prosodic variations, etc.) in segments of speech as possible (see e.g. [1] or [2]). In the case of such large acoustic unit inventories, an automation of the inventory creation process is very helpful, especially for multilingual and/or multi-voice TTS systems. Thanks to the automation, different inventories/voices can be created very quickly within a framework of a single TTS system.

In our previous work, ARTIC, a modern TTS system was developed to synthesize primarily Czech speech [3]. Two corpus-based approaches to speech synthesis are currently supported in the system, resulting in two versions of TTS systems: a *single unit instance* (SUI) system (also known as *fixed-inventory synthesis* or *diphone-based synthesis* [1]) with the quality of resulting speech limited by the fixed inventory, and *multiple unit instance* (MUI) system with the quality profitting from selecting (employing *unit-selection algorithm* [1], [4], [5]) from many unit instances available or from modelling properties of speech statistically (using Hidden Markov models, HMMs, in *HMM-based speech synthesis* [6], [7]). ARTIC was primarily designed to synthesise Czech speech; two Czech voices for SUI and four high-quality Czech voices for MUI are currently available [8]. While focused on the Czech language, some experiments to synthesise speech in other languages, German [9] and Slovak [10], were also carried out. However, as previous-generation SUI version of the system was employed in these experiments, the quality of synthetic speech in these languages was limited and appears now to be insufficient in modern applications. Therefore, a new Slovak MUI system with a new unit-selection compatible voice was built from a new large speech corpus within the framework of ARTIC as described in [11] and further, in more detail, in this paper.

Since corpus-based speech synthesis (both unit selection and HMM based) is very popular today and was shown to be able to produce synthetic speech of a high quality, new corpora in many languages (of course, in major languages like English [14] or French [15] but also in minor languages like Czech [8], [16], [17] or Slovak [18]) have been intensively designed. Thanks to data-driven approaches in various phases of the preparation of a new corpus, a new voice derived from this corpus can be designed relatively quickly and easily. Following the principles on which the Czech corpora were built (and which also proved to be useful for automatic recognition of the Slovak language [19], [20]), a new Slovak corpus for unit-selection speech synthesis was

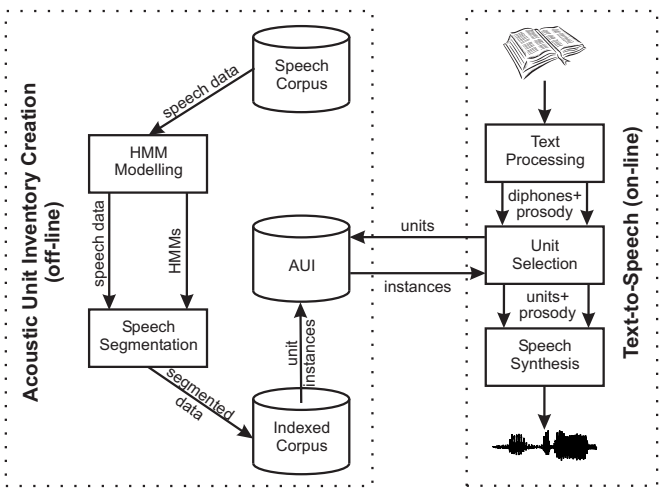

Fig. 1. Simplified block diagram of ARTIC TTS system.

Manuscript received January 25, 2012. This work was supported by the Ministry of Industry and Trade of the Czech Republic, project No. MPO FR-TI1/518, and by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

J. Matoušek, D. Tihelka, and J. Romportl are with the Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic, e-mail: {jmatouse, dtihelka, rompi}@kky.zcu.cz.

J. Psutka is with the University of West Bohemia, Pilsen, Czech Republic, e-mail: {psutka}@kky.zcu.cz

TABLE I
SLOVAK PHONETIC INVENTORY WITH 54 PHONES USED IN ARTIC (SYMBOLS ARE IN SAMPA NOTATION [12], [13]).

Phone	Word	Trans.	Phone	Word	Trans.	Phone	Word	Trans.
a	mama	mama	d	dom	dom	Z	žena	Zena
e	pes	pes	c	ťava	cava	x	chata	xata
i	pivo	pivo	J\	háďa	h\a:J\a	h\	had	h\at
o	bok	bok	k	oko	oko	G\	nechže	JeG\Ze
u	bubon	bubon	g	guma	guma	r	rak	rak
a:	páv	pa:f	?	áno	?a:no	r=	vrch	vr=x
e:	želé	Zele:	m	mama	mama	r=:	vŕba	vr=:ba
i:	víno	vi:no	F	amfiteáter	aFfitea:ter	l	loď	loc
o:	katalóg	katalo:k	n	nos	nos	l=	vlk	vl=k
u:	múr	mu:r	N	banka	baNka	l=:	vĺča	vl=:t_Sa
{	päť	p{c	J	vaňa	vaJa	L	ťad	Lat
i_^a	piatok	pi_^atok	N\	Slovensko	sloveN\sko	j	jama	jama
i_^e	mier	mi_^er	f	figa	figa	u_^	pravda	prau_^da
i_^u	paniu	paJi_^u	w	vdova	wdova	i_^	kraj	krai_^
u_^o	kôň	ku_^oJ	v	vlak	vlak	t_s	cena	t_sena
p	prak	prak	s	osa	osa	t_S	oči	ot_Si
b	bod	bot	z	zima	zima	d_z	medza	med_za
t	vata	vata	S	šek	Sek	d_Z	džungľa	d_ZuNgLa

designed and is described in this paper.

All steps in the design of a new Slovak unit-selection speech synthesis system are described in the paper. Section II briefly introduces the ARTIC text-to-speech system. In Section III, the process of the preparation of a new Slovak speech corpus and the creation of a new Slovak acoustic unit inventory, including the description of a Slovak phonetic alphabet and its differences from Czech, are introduced. Text-processing issues, including examples of Slovak phonetic transcription rules, are shown in Section IV. Speech production utilising a unit-selection algorithm is depicted in Section V. Finally, Section VI concludes the paper and outlines our future work in synthesis of Slovak speech.

II. ARTIC TTS SYSTEM

ARTIC (Artificial Talker in Czech) is a Czech text-to-speech system developed since 1997 [3]. It is a corpus-based system which is based on a large carefully designed speech corpus (annotated on orthographic, phonetic and prosodic levels [21], [8]). From the very beginning, automatic HMM-based approach to acoustic unit inventory creation was used. Two speech synthesis methods, fixed-inventory synthesis (a SUI approach) and unit-selection synthesis (a MUI approach), are currently implemented [3]. Experiments with HMM-based speech synthesis method (which could be, in its basic version, viewed as a sort of a SUI approach) have been carried out recently as well [7]. The block diagram of the ARTIC TTS system is shown in Figure 1.

The basic speech units used in the SUI system are *triphones*. Since only one instance of each triphone is employed in the system, the representative instances are selected off-line. Due to the fixed inventory, compact inventories can be utilized (tens of MBs), but the quality of output speech is limited. As a result, the SUI version of ARTIC is suitable for low-resource devices.

On the other hand, many instances of each speech unit (*diphones* in this case) are employed in the MUI version, and the optimal instances are selected on-line using a unit-selection algorithm. As a result, high-quality speech can be produced in this way at the expense of employing large acoustic unit inventories (hundreds of MBs). In this paper, a process of building of a new Slovak voice for the MUI approach with unit selection is described.

In the following sections, a process of building of a new voice in a new language, Slovak, within the ARTIC TTS system framework will be described. Unlike [10], focus will be given to the MUI version of the system (incorporating the unit-selection algorithm). Spoken form of Slovak (and mainly phonetics and phonology) will be dealt with too. As Czech is the main language the TTS system ARTIC has been designed for, we will also mention some differences between Czech and Slovak. Furthermore, all steps in the design, including the preparation of a suitable acoustic unit inventory of the new Slovak voice, text processing and speech production itself, will be detailed.

III. SLOVAK ACOUSTIC UNIT INVENTORY

We used the same *phonetic inventory* as described in [10] and shown in Table I. To create an *acoustic unit inventory* (AUI) for a unit-selection system, a new large speech corpus of a Slovak voice was designed first. It comprised 12,070 utterances (approx. 19 hours of speech, see Table II). The texts of utterances were downloaded from Internet news portals, selected to comprise all phones in sufficient number of occurrences and recorded by a semi-professional voice talent in an anechoic chamber using high-quality recording devices. The recorded utterances were then carefully annotated following the principles described in [21], and phonetic transcripts were obtained automatically by applying rules described in Section IV. Beside speech waveforms, glottal signals were also recorded using an electroglottograph and used as input signals to glottal pulses (pitch-marks) detection algorithm [22], [23], [24]. Pitch-marks are used for an accurate computation of fundamental frequency (F0) in

TABLE II
OVERVIEW OF UTTERANCES IN THE NEW SLOVAK CORPUS.

Type of utterances	**Number**	**Length**
Declarative	10,000	∼17 hours
"Yes/No"-questions	1,013	∼55 mins
"Wh"-questions	1,057	∼63 mins
Total	**12,070**	**∼19 hours**

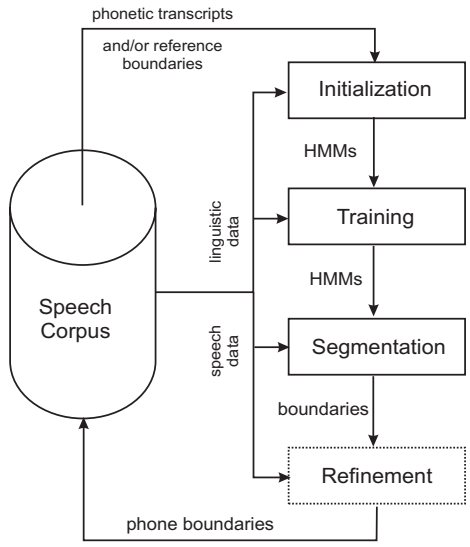

Fig. 2. Scheme of automatic acoustic inventory creation.

join cost (see Section V) and also as consistent concatenation points during speech synthesis.

When comparing Czech and Slovak, it should be said that, being Slavic languages, both languages are very similar in all linguistic aspects (unlike e.g. German [9]). However, despite the similarity, there are some differences both in orthographic and phonetic forms. These differences should be taken into account when building a Slovak voice in the ARTIC TTS system. As for orthography, there are some Slovak letters which are not used in written Czech (namely *ä, ô, ľ, ŕ, Ĺ*). Acoustic unit inventories also do differ. Standard Slovak phonetic alphabet (denoted using SAMPA [12], [13]) consists of 54 phones (see Table I) while 48 phones are usually used for Czech [25]. Here is a comparison between Slovak and Czech phonetic inventories:

Vowels. There are almost no distinctions between Czech and Slovak vowel systems—basically there are 5 short [a, e, i, o, u] and 5 long [a:, e:, i:, o:, u:] vowels in both languages. The only exception is an "additional" Slovak short vowel [{] which can rarely appear in spoken Slovak (often is pronounced as [e]).

Diphthongs. 4 diphthongs [i_^a, i_^e, i_^u, u_^o] occur in Slovak. None of them exists in Czech. On the contrary, there are three Czech diphthongs [o_u, a_u, e_u] that do not occur in Slovak.

Plosives. There are no differences between 9 Slovak and Czech plosives: [p, b, t, d, c, J\, k, g, ?]. [?] stands for glottal stop.

Affricates. 4 Slovak affricates are the same as the Czech ones: [t_s, t_S, d_z, d_Z].

Nasals. There are 5 "basic" nasals [m, F, n, N, J] in both Slovak and Czech. Moreover, another nasal [N\] can be pronounced in some contexts in Slovak.

Fricatives. There are 9 fricatives in "basic" Slovak [f, w, v, s, z, S, Z, x, h\]. They are the same as the Czech ones with the exception of [w] being an important variant of [v]. Moreover, due to voice assimilation "voiced *ch*" [G\] can be pronounced alternately with [h\] in both languages.

Liquids. In fact, 3 liquids occur in Slovak [r, l, L]. But there are also their significant allophones which express the syllabicity [r=, r=:, l=, l=:]. Symbol [=] denotes the syllabicity, [:] stands for "long" duration. "Long" syllabic phones [r=:, l=:] (written as *ŕ*, *ĺ*) and "soft" [L] (written as *ľ*) do not exist in Czech.

Glides. There are 3 glides in Slovak [j, u_^, i_^]. Just [j] occurs in Czech.

The process of the automatic acoustic unit inventory creation is illustrated in Figure 2. Hidden Markov models (HMMs) and ASR-based training procedures were employed to align HMMs with speech data, producing boundaries between phones in each source utterance. Optionally, the boundaries can be refined as described e.g. in [26], [27]. As diphones are used as basic speech units in ARTIC unit-selection system, diphone boundaries were then derived from the phone boundaries. More details about the automatic segmentation and acoustic unit inventory creation process can be found e.g. in [28], [29].

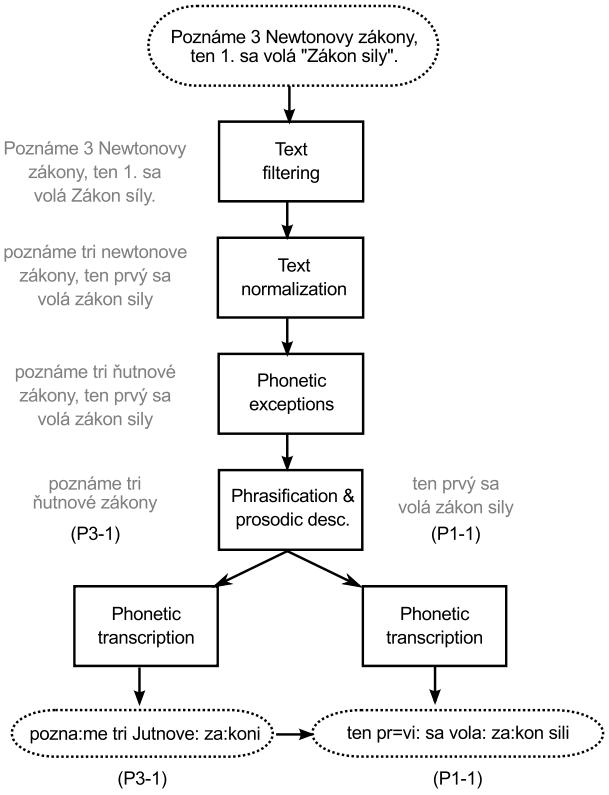

Fig. 3. Schematic view of text processing in ARTIC TTS system. For clarity, a single sentence is shown in the figure, phones-to-diphones conversion is not shown, and prosodic description is limited to prosodemes. Prosodeme P3-1 occurs in a non-terminating part of a sentence, while prosodeme P1-1 occurs in a declarative sentence (see Section IV-B). In the real system more detailed prosodic description is used [30], [31] or [32].

IV. Text Processing

Being a Slavic language, Slovak is similar to Czech in all linguistic aspects. Hence, very similar text-processing techniques as for Czech were carried out also for Slovak with the aim to reveal phonetic and prosodic aspects of an input text. Due to a complexity of such a task, current text processing in the ARTIC system is somewhat simplified to several main steps (see Figure 3):

- **Sentence boundary detection.** In this step, the decision whether a punctuation mark in the written text is or is not the end of a sentence. Neural networks or a heuristic classifier are used for the detection [33].

TABLE III
NOTATION USED IN THE EXAMPLES OF PHONETIC TRANSCRIPTION RULES.

Abbreviation	Description	Phones
*	any phone	*arbitrary phone symbol*
—	word boundary	*word boundary symbol*
-	syllable boundary	*syllable boundary symbol*
VPC	voiced paired consonant	b w d z d_z Z d_Z J\ g G\
UPC	unvoiced paired consonant	p f t s t_s S t_S c k x
TDNL	consonants "t d n l"	t d n l
ALV	alveopalatal consonants	c J\ J L
VOC	vocals and diphthongs	a a: e e: i i: o o: u u: { i_^a i_^e i_^u u_^o
SON	sonorant consonants	m F n N J N\ l l= l=: L r r= r=: j
CON	consonants	VPC + UPC + SON

- **Text normalisation.** This step consists in transcribing "non-standard" words (digits, abbreviations, acronyms, etc.) into their correct "full word forms". This is a significant problem in Slavic languages as morphological and semantic information is necessary to make the conversion unambiguous. Various tagging and lemmatization techniques are used [34], [35], [36].
- **Phonetic transcription.** In order to have a pronunciation form of the utterances in the written text, the normalised texts are phonetically transcribed. Detailed rule-based phonetic transcription including pronunciation dictionary of "exceptional" words (mostly foreign words) [37] is used in our system—more details are given in Section IV-A.
- **Phones-to-diphones conversion.** This is a simple step in which phonetic units (phones) are converted to more "technical units" the synthesizer works with. In our case, diphones (i.e., units which start in the middle of a phone and end in the middle of the following phone; thus, covering the co-articulation information between phones) are used.
- **Prosodic description.** In addition to phonetic transcription, each utterance is also transcribed in terms of prosodic symbols (prosodic clauses, phrases, prosodemes, etc.) using prosodic phrase grammar. More details are given in Section IV-B.

A. Phonetic Transcription

Similarly as for Czech, phonetic form of Slovak is similar to orthographic form. Thus, relatively simple phonetic transcription rules can be utilised to convert Slovak letters to phones. Based on phonetic rules defined by Slovak phoneticians in [38], approximately 150 rules in a form [37]

$$A \rightarrow B / C\_ \; D \tag{1}$$

were defined in our system. The rule can be read as follows: letter sequence A with both left context C and right context D is transcribed as phone sequence B. Some examples of these rules are given here (notation is explained in Table III):

$$\text{VPC} \rightarrow \text{UPC} \;\; / \; * \_ \; \langle \text{UPC}, |\text{UPC}, |\text{PAU}, |? \rangle \tag{2}$$

$$\text{UPC} \rightarrow \text{VPC} \;\; / \; * \_ \; \langle \text{VPC}, |\text{VPC}, |\text{SON}, |\text{VOC}, |\text{v} \rangle \tag{3}$$

$$\text{TDNL} \rightarrow \text{ALV} \;\; / \; * \_ \; \langle \text{i}, \text{i} :, \text{e} \rangle \tag{4}$$

$$\text{r} \rightarrow \text{r} = \;\; / \; \text{CON}\_ \; \langle \text{CON}, | \rangle \tag{5}$$

$$\acute{\text{I}} \rightarrow \text{l} =: \;\; / \; * \_ * \tag{6}$$

$$\text{ia} \rightarrow \text{i\_\^{}a} \;\; / \; * \_ * \tag{7}$$

$$\text{iu} \rightarrow \text{i\_\^{}u} \;\; / \; * \_ * \tag{8}$$

$$\hat{\text{o}} \rightarrow \text{u\_\^{}o} \;\; / \; * \_ * \tag{9}$$

$$\text{stsk} \rightarrow \text{sk} \;\; / \; * \_ * \tag{10}$$

$$\text{ts} \rightarrow \text{t\_s} \;\; / \; * \_ \; \langle \text{k}, \text{t} \rangle \tag{11}$$

$$\text{t\_SS} \rightarrow \text{t\_S} \;\; / \; * \_ \; \text{CON} \tag{12}$$

$$\text{VOC} \rightarrow \text{?VOC} \;\; / \; \text{PAU}\_ * \tag{13}$$

$$\text{v} \rightarrow \text{u\_\^{}} \;\; / \; * \_ \text{-} \tag{14}$$

$$\text{v} \rightarrow \text{f} \;\; / \; \text{-}\_ \; \langle \text{UPC}, |\text{UPC}, |\text{PAU} \rangle \tag{15}$$

$$\text{v} \rightarrow \text{w} \;\; / \; \text{-}\_ \; \langle \text{VPC}, |\text{VPC} \rangle \tag{16}$$

$$\text{v} \rightarrow \text{v} \;\; / \; \text{-}\_ \; \langle \text{n}, \text{J}, |\text{n}, |\text{J} \rangle \tag{17}$$

$$\text{j} \rightarrow \text{i\_\^{}} \;\; / \; * \_ \text{-} \tag{18}$$

The examples include special inter- and cross-word voice assimilation rules (2, 3), rules for "softening" (or palatalisation) of consonants (4), rules for transcribing short (5) and long (6) syllabic consonants, and a rule for transcribing glottal stop (13). Rules for transcribing diphthongs are shown in (7-9). Examples of simplifying consonantal groups are given by the rules (10-12). Special rules for transcribing written *v* as [v], [f], [w], or [u_^] are shown in (14-17) and written *j* as [u_^] in (18).

In these examples, for instance, the rule (4) can be read as "written *t, d, n, l* are pronounced as alveopalatal [c, J\, J, L] in front of [i, i:, e]". Multiple transcriptions of ambiguous Slovak contexts are supported as well. Pronunciation dictionary of "exceptional" words (foreign words but also some Slovak native words not obeying the transcription rules) is also employed.

B. Prosodic Description

In addition to the *phonetic description* (on a *segmental level*), *prosodic description* (on a *suprasegmental level*) plays a key role in synthesising speech with a high degree of naturalness. This kind of description is not directly linked to phones. It rather relates to larger phonetic units like syllables, words, parts of sentences or even to whole sentences or longer utterances.

In Slavic languages (and also in other Indo-European languages), prosody can be viewed to supplement the phonetic information by other linguistic aspects, such as sentence modality (in sense of declarative sentences vs. yes/no questions, terminating vs. non-terminating utterances, etc.), emotions and styles, or generally expressiveness and speaker's attitude during communication. As such, prosody helps listeners understand the meaning of the transmitted message. Prosody also helps in the division of longer utterances into sentences, sentences into shorter segments (clauses or phrases) and phrases to words.

In our system, prosodic analysis includes heuristic punctuation-driven sentence clause detection, rule-based word stress detection and symbolic prosodic description adopted from the Czech language [31]. The symbolic prosodic description is based on a *prosodic phrase grammar*, in which the following *prosodic structures* are distinguished:

- **Prosodic sentence (PS).** Prosodic sentence is actually a prosodic manifestation of a sentence (e.g. an utterance) as a syntactically consistent unit, yet it can also be unfinished or grammatically incorrect.
- **Prosodic clause (PC).** Prosodic clause is such a linear unit of a prosodic sentence which is delimited by pauses.
- **Prosodic phrase (PP).** Prosodic phrase is such a segment of speech where a certain intonation scheme is realized continuously. A single prosodic clause often contains more prosodic phrases.
- **Prosodeme (P0), (Px).** Prosodeme is an abstract unit (sort of a "suprasegmental phoneme") established in a certain communication function within the language system. We have postulated that any single prosodic phrase consists of two prosodemes: so called "null prosodeme" (P0) and "functionally involved prosodeme" (where (Px) stands for a type of the prosodeme[1], depending on the communication function the speaker intends the sentence to have).
- **Prosodic word (PW).** Prosodic word (sometimes also called phonemic word) is a group of words subordinated to one word accent (stress).
- **Semantic Accent.** By this term we call such a prosodic word attribute, which indicates the word is emphasised (using acoustic means) by a speaker.

So, the prosodic structures formally describe the linguistic functions of certain prosodic phenomena and can be parsed using the prosodic phrase grammar

$$
\begin{aligned}
(PS) &\longrightarrow (PC)\{1+\}\,\$\{1\}\\
(PC) &\longrightarrow (PP)\{1+\}\,\#\{1\}\\
(PP) &\longrightarrow (P0)\{1\}\,(Px)\{1\}\\
(P0) &\longrightarrow \emptyset\\
(P0) &\longrightarrow (PW)\{1+\}\\
(Px) &\longrightarrow (PW)\{1\}\\
(Px) &\longrightarrow (SA)\{1\}\,(PW)\{1+\}\\
(PW) &\longrightarrow w\{1\}.
\end{aligned}
$$

The terminal symbols \$ and # represent two types of pauses (\$ is a pause between sentences, # is a pause between prosodic clauses), $\emptyset$ is an empty symbol, and the symbol w is a variable that can represent any word of the language.

Using the prosodic phrase grammar, each sentence can be represented by a derivation hierarchical tree. An example of a hierarchical tree is shown in Figure 4.

To formally define the relation of the prosodic grammar with speech units used to create the synthetic speech (or, in general units used to model the prosody), let P_{NS}^{l} be a set

[1](P1) is a prosodeme at the end of declarative sentences, (P2) is a prosodeme at the end of question sentences, (P3) is a prosodeme at the end of non-terminating sentence parts. Each prosodeme can occur in several variants—for more details see [30], [31] or [32].

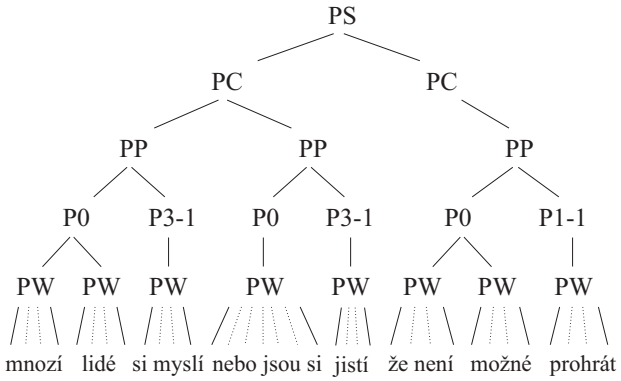

Fig. 4. An example of a prosodic-phrase-grammar-based hierarchical tree.

of all nodes in prosodic structure tree, and l be the index in the hierarchy of the structure ($1 = PS, \ldots, 5 = PW$). Since the level of prosodic words may not be enough for the purposes of speech units concatenation—not only prosodic, but also phonetic information affect the naturalness of a selected sequence of units—let us extend P_{NS} by the set of all phonetic symbols of the utterance underlying the prosodic structure, staying on level $l = 6$.

During speech synthesis, each speech unit is assigned with a set of symbolic prosodic features related to the prosodic structures. Taking the hierarchic nature of the grammar into account, the features (and their mutual relations) can be defined as:

$$t_l = \left(\mathcal{F}_l(P_{NS}^{l}, P_{NS}^{l-1}), t_{l-1}\right), \qquad l = 2, \ldots, 6 \quad (19)$$

$$t_1 = \emptyset \quad (20)$$

where $\mathcal{F}$ is a function defining the relation between levels l and $l-1$ in the prosodic structure (e.g. the relation of units to the prosodic word) which can differ for individual levels. The recursion allows us to fully describe the whole hierarchy.

V. Speech Production

In MUI version of ARTIC, resulting speech is generated by a unit-selection algorithm (see e.g. [5], [39]). Its principle is to smoothly concatenate (according to *join cost*) speech segments taken from speech unit inventories according to phonetic and prosodic criteria (*target cost*) imposed by the *target specification* given by the synthesised utterance [4]; the unit inventories are filled with diphone speech segments extracted from natural utterances using the automatically segmented boundaries. As there are many instances of each speech segment for most of them, there is a need to dynamically select the optimal (with respect to both target and join costs) instances during synthesis run-time (using a unit selection technique). A scheme of unit-selection approach to speech synthesis is shown in Figure 5. Although the unit-selection framework is, to some extend, language-dependent (at least its target-related part), the same setup as for Czech was employed, with slight adjustments respecting the phonetic and prosodic nature of the Slovak language in the target specification features.

To calculate the target cost, a prosodic structure of the to-be-synthesised utterance is described by means of *prosodic phrase grammar* symbols, and each diphone is assigned with the appropriate *P0, Px* and, optionally, *SA* labels, plus

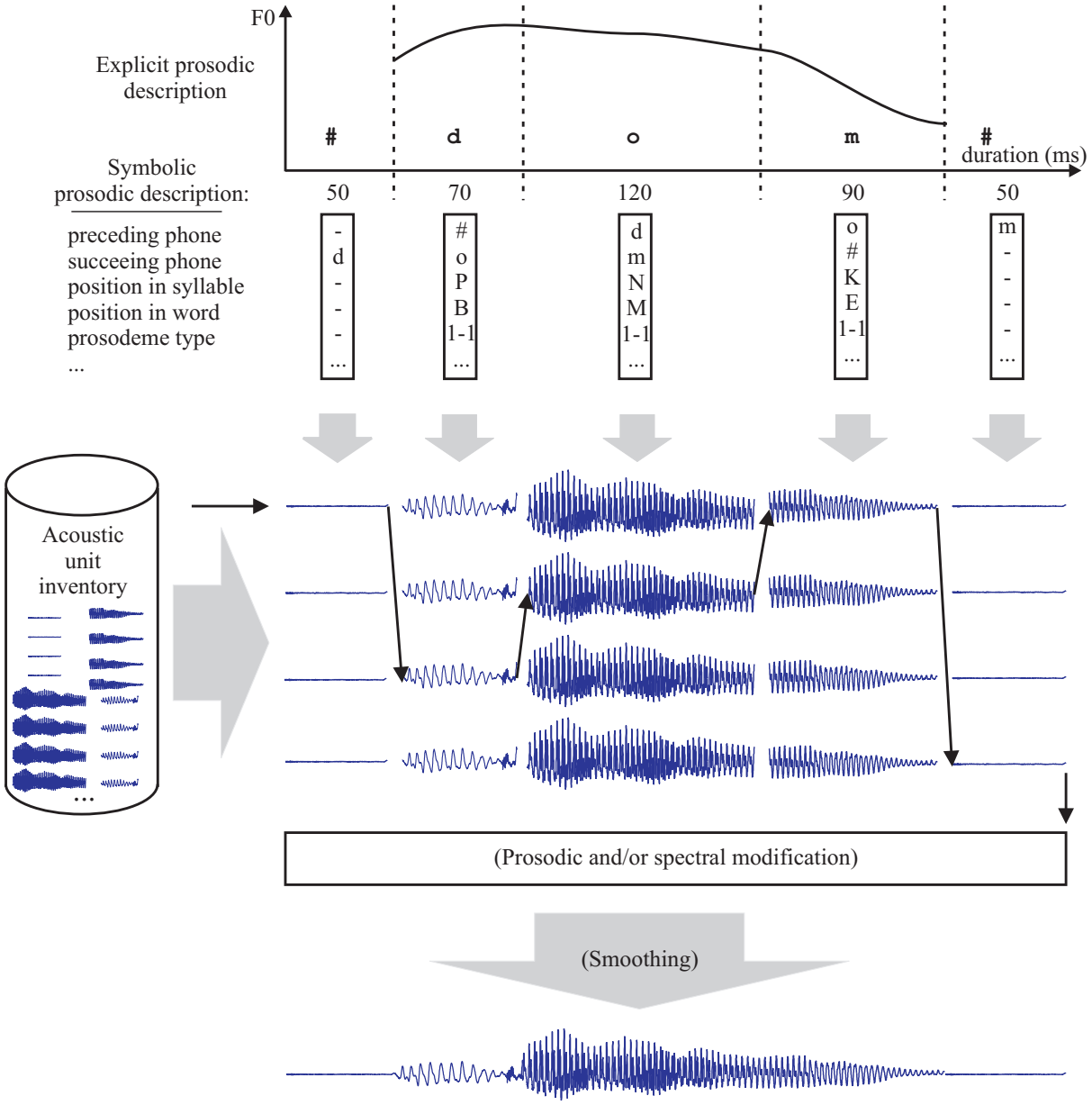

Fig. 5. A schematic view of unit-selection speech synthesis driven by symbolic prosody in the ARTIC TTS system (adapted from [1]). Position in syllables are denoted as (N)ucleus, (P)raetura, and (K)oda. Position in word is (B)eginning, (M)iddle, and (E)nd. Prosodeme 1-1 occurs in a declarative sentence (for list of all prosodemes, see e.g. [31]).

with additional lower-level features like phonetic context (immediate left and right phones to the diphone) and (non-discrete) "suitability" to the given position in the underlying prosodic word (*PW*) [5].

The join cost is, currently, evaluated as a mean of distances consisting of spectral difference (mel-frequency cepstral coefficients, MFCCs, are currently used), and the differences of F0 and energy measured around the concatenation points of the units to concatenate. Our experience and knowledge shows, however, that those features (especially the MFCC) do not reflect very much how concatenation smoothness is perceived by humans [40], [41].

After selecting the optimal sequence of (diphone) speech segments, neither prosodic nor spectral modifications are made by default in the ARTIC system except for simple smoothing at concatenation points, except when the change of tempo (duration) of synthetic speech is required. In this case, special WSOLA-based algorithm is employed to modify the speech signal of individual units [42], achieving high-quality speech speed-up/slow-down even for larger duration modifications factors. To preserve the naturalness, the MUI system does not allow to change or explicitly model the pitch (F0 contour), since we do not yet have a method achieving high-quality pitch modifications. To cope with high CPU power and memory cost typical for unit-selection systems, a computational optimisation, as proposed in [39], is used.

A. *Prosodic Synonymy and Homonymy in Unit Selection*

Regarding the target features, we intend to move from simple match/mismatch features comparison (usually used in today's unit selection frameworks) forward to what we call *prosodic synonymy/homonymy* of units [43]. In the "classic" concept, the target features and their individual prominences are set heuristically (sometimes the prominences are adjusted automatically [4], but still not changing the essence of the

concept), while it is hoped that the features describe the prosodic properties of the units. In fact, since *prosody* is suprasegmental feature, we cannot consider speech units (of phone-size, which is also our case) as holders of prosody. Instead, such features (position in *sth*, type of *sth*, left/right *sth*) try to fix the unit into a particular context in which it has been recorded in the source speech corpus, and the selection algorithm strives to put the unit into the same context (evaluating the features match/mismatch). The required/expected prosody (as the copy of prosody from the corpus) emerges only if a sequence of units, large enough, is successfully reconstructed.

The key point in *prosodic synonymy/homonymy* is to reach the natural prosody expressing the required communication function, while not necessarily copying the prosody from corpus, by building sequences of units in context different than those in which they were originally recorded—the units are synonymous for the context even when they originated from different contexts. Or from the point of view of prosody, homonymous prosody (in term of expressed communication function) is achieved with units originating in prosodic contexts different from that created.

To define it more formally, let $\mathbf{C} = \{c_1, c_2, \ldots, c_N\}$ be the set containing all candidates of a unit—this set is *sharply determined* by the particular candidates in it. Be further $t(m)$ a particular target requirement—one symbol (leaf) t_6 in the prosodic grammar. In the classic concept, the relation between the target and the candidates matching the target is

$$S : t(m) \rightarrow \mathring{C}(m) \subset \mathbf{C} \qquad \forall m = 1, \ldots, M \tag{21}$$

which assigns sharply delimited finite set of units to the given target (i.e. the units are described by the same grammar structure). Note that relation 21 is valid in general, not only for prosodic description using the grammar in Section IV-B. Taking into account all possible target descriptions $t = \bigcup t(m)$, $\forall m$, then

$$\bigcap \mathring{C}(m) = \emptyset \tag{22}$$

and even for large, although still limited, corpus there is a non-empty set $\bar{t} \subset t$ for which there are no units matching the prosodic description from the set, i.e.

$$\bigcup_{\bar{t}} \mathring{C} = \emptyset \tag{23}$$

On the other hand, when synonymy is taken into account the relation 21 becomes

$$S : t(m) \rightarrow \mathcal{C}(m) \subseteq \mathbf{C} \qquad \forall m \tag{24}$$

assigning to $t(m)$ a *not sharply determined* sub-set $\mathcal{C}$ of candidates which express the communication function given by $t(m)$. For all values of t it is also valid that

$$\bigcap C(m) \subseteq \mathbf{C} \tag{25}$$

As a result of relation 25, the set $\bar{t}$ is here much smaller, or even empty.

The classic understanding of what does "feature" mean is not very suitable here, since a unit description may, in general, vary depending in the context into which the unit is tried to be inserted—some classic features may be important for one context (their difference to what is required reflects what will be perceived), while in another context they may become insignificant. In general, there is also no need for target cost to be 0 for all synonymic candidates (although it will usually be met when a phrase occurring in the corpus is synthesized). Therefore, let target cost be more generally defined as a similarity function $\mathcal{G}$ between target features $t(c)$ of a candidate c and target specification t required for the candidate:

$$TC(c, t) = \mathcal{G}(t(c), t) \tag{26}$$

and for the correct function of the selection, the following equation must be met:

$$\forall c \in \mathcal{C} \text{ and } \forall d \in (C - \mathcal{C}) : TC(c, t) < TC(d, t) \tag{27}$$

In this concept, it is not possible to "make up" or define a set of target features; instead, they must be revealed, or trained, according to given speech corpus. Therefore, it seems that instead of target specification describing *what we want to achieve*, it will be more suitable to prescribe *what we must avoid* — in the sense of perception, the first is not a complement of the second. The very first experiment, described in [5], replaced the classic discrete positional features (e.g. index of unit in syllable and/or word) by continuous *suitability* for the given position. Although it does not obey the synonymy/homonymy concept (in fact, it is still rather closer to classic features handling), it showed that it is more feasible to think about selection features in terms of "suitable for" instead of "required for".

Similarly to the target cost, the concatenation features should also obey the principles of synonymy/homonymy, only from a perspective of the smoothness of the generated speech.

VI. Conclusion and Future Work

In this paper, a new Slovak voice suitable for a multiple-unit-instance version (with unit-selection algorithm) of the ARTIC TTS system was presented. The new Slovak voice was built on the same principles as the existing Czech voices within the framework of the ARTIC system. All steps in the design including the preparation of an appropriate acoustic unit inventory of the new Slovak voice, text processing and speech production modules were mentioned in the paper. According to informal listening tests the new MUI (unit-selection) version of Slovak clearly outperformed the SUI version. In fact, speech synthesised by the MUI version was never rated as worse than speech synthesised by the SUI version.

In our future work, we plan to focus mainly on text processing issues, especially on advanced text pre-processing, fine-tuning of phonetic transcription rules and updating of the pronunciation dictionary of exceptional words. Application of HMM-based speech synthesis method for Slovak language is also under consideration.

References

[1] T. Dutoit, "Corpus-based speech synthesis," in *Springer Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. Dordrecht: Springer, 2008, pp. 437–455.

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge: Cambridge University Press, 2009.

[3] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text-to-speech system ARTIC," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2006, vol. 4188, pp. 439–446.

[4] A. Hunt and A. Black, "Unit selection in concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, USA, 1996, pp. 373–376.

[5] D. Tihelka and J. Matoušek, "Unit selection and its relation to symbolic prosody: a new approach," in *Proc. INTERSPEECH*, Pittsburgh, USA, 2006, pp. 2042–2045.

[6] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[7] Z. Hanzlíček, "Czech HMM-based speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Springer, 2010, vol. 6231, pp. 291–298.

[8] J. Matoušek, D. Tihelka, and J. Romportl, "Building of a speech corpus optimised for unit selection TTS synthesis," in *Proc. LREC*, Marrakech, Morocco, 2008.

[9] J. Matoušek, D. Tihelka, J. Psutka, and J. Hesová, "German and Czech speech synthesis using HMM-based speech segment database," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, vol. 2448, pp. 173–180.

[10] J. Matoušek and D. Tihelka, "Slovak text-to-speech synthesis in ARTIC system," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, vol. 3206, pp. 155–162.

[11] J. Matoušek, D. Tihelka, and J. Psutka, "New Slovak unit-selection speech synthesis in ARTIC TTS system," in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2011, WCECS 2011*, San Francisco, USA, 2011, pp. 485–490.

[12] "Slovak SAMPA," http://www.ui.savba.sk/speech/sampa_sk.htm.

[13] J. Ivanecký and M. Nábělková, "Fonetická transkripcia SAMPA a Slovenčina (SAMPA transcription and the slovak language)," *Jazykovedný časopis*, vol. 53, no. 2, pp. 81–95, 2002, (in Slovak).

[14] J. Kominek and A. Black, "CMU ARCTIC speech databases, the," in *Proc. SSW*, Pittsburgh, USA, 2004, pp. 223–224.

[15] H. Francois and O. Boeffard, "Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem," in *Proc. INTERSPEECH*, Aalborg, Denmark, 2001, pp. 829–832.

[16] V. Radová and J. Psutka, "Recording and annotation of the Czech speech corpus," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2000, vol. 1902, pp. 319–323.

[17] J. Matoušek and J. Romportl, "On building phonetically and prosodically rich speech corpus for text-to-speech synthesis," in *Proc. CI*, San Francisco, USA, 2006, pp. 442–447.

[18] M. Rusko, M. Trnka, S. Daržágín, and M. Cerňak, "Slovak speech database for experiments and application building in unit-selection speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, vol. 3206, pp. 457–464.

[19] J. Psutka, J. Hajič, and W. Byrne, "The development of ASR for Slavic languages in the MALACH project," in *Proc. ICASSP*, Montreal, Canada, 2004, pp. 749–752.

[20] J. Psutka, P. Ircing, J. V. Psutka, J. Hajič, W. Byrne, and J. Mírovský, "Automatic transcription of Czech, Russian, and Slovak spontaneous speech in the MALACH project," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 1349–1352.

[21] J. Matoušek and J. Romportl, "Recording and annotation of speech corpus for Czech unit selection speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin: Springer, 2007, vol. 4629, pp. 326–333.

[22] M. Legát, J. Matoušek, and D. Tihelka, "A robust multi-phase pitch-mark detection algorithm," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 1641–1644.

[23] M. Legát, D. Tihelka, and J. Matoušek, "Pitch marks at peaks or valleys?" in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, vol. 4629, pp. 502–507.

[24] M. Legát, J. Matoušek, and D. Tihelka, "On the detection of pitch marks using a robust multi-phase algorithm," *Speech Commun.*, vol. 53, no. 4, pp. 552–566, April 2011.

[25] "Czech SAMPA," http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm.

[26] S. S. Park and N. S. Kim, "On using multiple models for automatic speech segmentation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2202–2212, 2007.

[27] J. Matoušek and J. Romportl, "Automatic pitch-synchronous phonetic segmentation," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008.

[28] J. Matoušek, D. Tihelka, and J. Psutka, "Experiments with automatic segmentation for Czech speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2003, vol. 2807, pp. 287–294.

[29] J. Matoušek, "Automatic pitch-synchronous phonetic segmentation with context-independent HMMs," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Springer, 2009, vol. 5729, pp. 178–185.

[30] J. Romportl, J. Matoušek, and D. Tihelka, "Advanced prosody modelling," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, vol. 3206, pp. 441–447.

[31] J. Romportl and J. Matoušek, "Formal prosodic structures and their application in NLP," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, vol. 3658, pp. 371–378.

[32] J. Romportl, "Prosodic phrases and semantic accents in speech corpus for Czech TTS synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2008, vol. 5246, pp. 493–500.

[33] J. Romportl, D. Tihelka, and J. Matoušek, "Sentence boundary detection in Czech TTS system using neural networks," in *Proc. ISSPA*, Paris, France, 2003, pp. 247–250.

[34] J. Kanis and L. Müller, "Using the lemmatization technique for phonetic transcription in text-to-speech system," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, vol. 3206, pp. 355–361.

[35] ——, "Automatic lemmatizer construction with focus on OOV words lemmatization," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, vol. 3658, pp. 132–139.

[36] J. Zelinka, J. Kanis, and L. Müller, "Automatic transcription of numerals in inflectional languages," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, vol. 3658, pp. 326–333.

[37] J. Psutka, L. Müller, J. Matoušek, and V. Radová, *Mluvíme s počítačem Česky (Talking with Computer in Czech)*. Prague: Academia, 2006, (in Czech).

[38] Ábel Kráľ, *Pravidlá slovenskej výslovnosti (Rules of Slovak Pronunciation)*, 2nd ed. Matica slovenská, 2009.

[39] D. Tihelka, J. Kala, and J. Matoušek, "Enhancements of Viterbi search for fast unit selection synthesis," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 174–177.

[40] M. Legát and J. Matoušek, "Identifying concatenation discontinuities by hierarchical divisive clustering of pitch contours," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, vol. 6836, pp. 171–178.

[41] ——, "Analysis of data collected in listening tests for the purpose of evaluation of concatenation cost functions," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, vol. 6836, pp. 33–40.

[42] D. Tihelka and M. Méner, "Generalized non-uniform time scaling distribution method for natural-sounding speech rate change," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, vol. 6836, pp. 147–154.

[43] D. Tihelka and J. Romportl, "Exploring automatic similarity measures for unit selection tuning," in *Proc. INTERSPEECH*, Brighton, Great Britain, 2009, pp. 736–739.