Slovak Text-to-Speech Synthesis in ARTIC System*

Jindřich Matoušek and Daniel Tihelka

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
jmatouse@kky.zcu.cz, dtihelka@kky.zcu.cz

Abstract. This paper presents a brand-new Slovak text-to-speech system. It was developed within the framework of ARTIC system (primarily designed to synthesize Czech speech) with respect to the knowledge of Slovak language. Thus, statistical approach (using hidden Markov models) was employed to build an acoustic unit inventory of Slovak language in a fully automatic way. Both phonetic transcription and prosodic rules were proposed to convert an input text to its phonetic form and to estimate its suprasegmental features. As a result, a fully working text-to-speech system that converts an arbitrary Slovak text to the corresponding output speech was designed. The informal listening tests show the system is capable of producing speech of a high quality (with high level of intelligibility and good naturalness).

1 Introduction

Text-to-speech (TTS) synthesis is one of the most important tasks of computer speech processing. Nowadays, concatenative 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. In the case of such large acoustic unit inventories the automation of the inventory creation process is necessary. Thanks to the automation, different inventories can be created very quickly. Thus, new voices and languages can be developed within a framework of a single TTS system. In modern integrating world (especially in view of the expanding European Union) multilingual TTS systems become more and more important and enjoy bigger and bigger popularity.

In [1, 2], ARTIC, a modern TTS system was developed to synthesize Czech speech. Having been created on the principles mentioned above, it is capable of using different automatically built acoustic unit inventories. An important step towards multilinguality was achieved in [3] where a German voice was successfully designed within the ARTIC system. In this paper another language, Slovak, is modeled within the framework of ARTIC.

* This work was supported by the Ministry of Education of Czech Republic, project No. MSM235200004, and the firm SpeechTech.

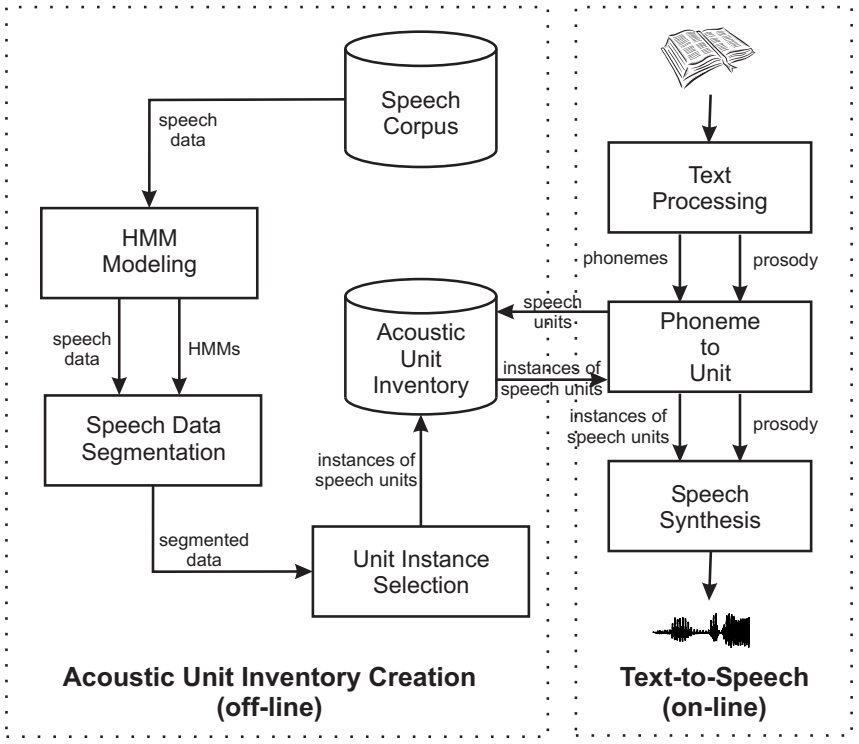

Fig. 1. A simplified scheme of ARTIC TTS System.

The paper is organized as follows. Section 2 briefly introduces TTS system ARTIC. In Section 3 steps necessary to create a Slovak voice within the ARTIC system are described. Finally, Section 4 concludes the paper by summarizing the main findings and outlines our future work.

2 Text-to-Speech System ARTIC

A brief introduction of the TTS System ARTIC is given in this Section. ARTIC (Artificial Talker in Czech) is a TTS system primarily designed to synthesize Czech speech. It is a corpus-based system that employs most widely used concatenative approach to speech synthesis [1]. The process of designing the synthesizer can be divided into two parts: the first part concerns the creation of the acoustic unit inventory, the second one takes care of text-to-speech conversion itself. The block diagram of ARTIC is shown in Figure 1.

Nowadays, two Czech voices (female and male) are available. In [3] a German TTS module was successfully created within the ARTIC system. It should be emphasized that all voices (or language modules) were realized fully automatically in the same way as described further in the paper (with respect to the their language-dependent properties).

2.1 Acoustic Unit Inventory

Statistical approach to acoustic unit inventory construction was applied – a carefully designed large speech corpus collected from a single speaker [2] is used to create the inventory in a fully automatic way. A set of three-state left-to-right single-density crossword-triphone hidden Markov models (HMMs) was employed to model context-dependent phone-sized units (*triphones*) on the basis of the large speech corpus [1, 3–5]. The speech corpus consisted of both linguistic and speech data. As for linguistics, both orthographic and phonetic transcriptions of each sentence were used. Speech signals were represented by their waveforms and their spectral properties were described by vectors of Mel Frequency Cepstral Coefficients (MFCCs) calculated using 20 ms windowed speech signal with 4 ms shift. In the current system 12 MFCCs plus normalized energy together with the corresponding first, second and third differential coefficients (52 coefficients in total) were used. Glottal signals were also recorded along with the speech using an electroglottograph.

HMMs were initialized using both *flat-start* (all HMMs start with the same parameters) and *bootstrap* (some pre-segmented speech data are available to initialize each HMM individually) technique [4]. To make more robust models and to enable modeling triphones not present in the speech corpus, a clustering procedure was employed to tie similar triphones. This is very important for TTS synthesis since clustering ensures that an arbitrary triphone, i.e. arbitrary text, could be synthesized. Clustering can be performed at two different levels: model (phone) level or state (sub-phone) level. Listening tests revealed the superiority of the model-level clustering over the state-level one [3]. Viterbi algorithm is then used to align speech waveforms of each sentence from the speech corpus with a sequence of corresponding tied-triphone HMMs. As a result, triphone-level segmentation of the speech corpus is produced [4, 5]. As the last step of the acoustic unit inventory creation process an off-line single instance selection was implemented to choose the "most representative" instance of every triphone. These instances are stored in the acoustic unit inventory and used later in on-line speech synthesis.

2.2 Text-to-speech

Text-to-speech module processes an arbitrary input text, converts it to its phonetic form (including prosodic feature estimation), and produces the output speech. Phonetic transcription of the input Czech text is done by rules [8, 1]. Recently, prosody generation module was designed for ARTIC [6]. It is able to carry out suprasegmental modulations of speech melody (i.e. fundamental frequency – $F0$ contour), intensity (i.e. volume) and timing (i.e. phone duration). A set of 21 rules is used to control various prosodic characteristics (e.g. a baseline F0 contour, a slope of an overall declining melody tendency, shapes of all declining/ascending cadences, intensity modulations, duration changes, influence of word/sentence stress, etc.).

As for the low-level synthesis, concatenative speech synthesis techniques can be employed to join speech units stored in the inventory. In fact, all standard concatenative techniques can be used in the system. OLA-like [1] and harmonic/noise-based (HNM) [7] synthesis methods have been implemented in the system so far. These methods use so-called *pitch-marks*, glottal closure instances (detected from glottal signals [2]), to change the prosodic characteristics of synthesized speech.

3 Modeling the Slovak Voice

In this section we describe the problems of creating a voice in a new language – Slovak. We will deal with spoken form of Slovak (and mainly phonetics and phonology) and focus on the differences between Czech and Slovak, because Czech is the main language the TTS system ARTIC has been designed for. Furthermore, the process of the automatic Slovak acoustic inventory creation and text-to-speech synthesis (including phonetic transcription and prosody generation) will be detailed.

3.1 The Slovak Language

In this subsection we will describe the properties of Slovak language very briefly. We will limit ourselves to the phonetic (or phonologic) and prosodic features – the most important properties from text-to-speech synthesis point of view.

First of all, it should be said that being a Slavic language, Slovak is very similar to Czech in all linguistic aspects (unlike e.g. German [3]). Phonetic forms of both languages are very similar to their orthographic forms (in fact, this feature is common to all Slavic languages). As a result, relatively simple phonetic transcription rules can be utilized to convert orthographic form (i.e. letters) to phonetic form (i.e. phones). Slovak phonetic transcription rules will be described in Section 3.3.

Despite the similarity between Czech and Slovak, there are some differences both in orthographic and phonetic forms. These differences should be taken into account when building the TTS system. As for orthography, there are some Slovak letters which are not used in written Czech (namely *ä, ô, ľ, ŕ, ĺ*).

Phonetic alphabet used in our Slovak TTS system consists of 54 phones, and is shown in Table 1. It is based on unofficial Slovak SAMPA [11, 12] and Slovak phonetic publications (e.g. [10]). Here is a comparison between Slovak and Czech [9] 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 diphtongs [i_^a, i_^e, i_^u, u_^o] occur in Slovak. None of them exists in Czech.

Table 1. The Slovak phonetic inventory used in our system.

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	ľad	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

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.

Prosodic features of Slovak are very similar to Czech as well. Stress is always on the first syllable (with the exception of non-syllabic prepositions and some monosyllabic words) in both languages. It means that it does not have a phonological-distinctive function, though it can help distinguish words in continuous speech. Intonational and temporal characteristics are also almost identical in Czech and Slovak.

3.2 Inventory of Slovak Acoustic Units

When modeling or synthesizing speech, the first step usually consists of defining the basic phonetic inventory of a language in focus. We use 54 Slovak phones (see Section 3.1) for our text-to-speech purposes.

Concatenative speech synthesis techniques employ acoustic unit inventories. Nowadays, these inventories are very large and are usually designed automatically on the basis of a large speech corpus. This is the case of the TTS system described in Section 2. The Slovak speech corpus has basically the same structure as the "general" corpus presented in Section 2. Here, 7.012 Slovak sentences were collected. All the sentences were pronounced by a single female speaker. Both speech and glottal signals were recorded. The characteristics of the corpus, compared to the corpora of other languages, are shown in Table 2.

Table 2. Speech corpora and acoustic unit inventories characteristics. CZ1 = Czech female, CZ2 = Czech male, DE = German Male, SK = Slovak Female Voice.

	CZ1	CZ2	DE	SK
Number of sentences	5,000	10,004	5,255	7,012
Amount of speech data [hours]	12.9	19.3	13	16.5
Number of phones	44	44	46	54
Number of occurrences per phone	6,300	8,396	4,804	4,715
Number of (clustered) triphones	7,106	7,016	4,687	6,621
Number of occurrences per triphone	32	44	35	32

The speech corpus is used as a basis for speech unit modeling. Slovak speech units were modeled in the same way as in Section 2.1, i.e. three-state left-to-right single-density crossword-triphone HMMs were employed to model Slovak acoustic units. Since no pre-segmented speech data were available, so called flat-start initialization was adopted. The clustering procedure was tuned to respect the features of spoken Slovak language. No experiments have been carried out to get optimal clustering results (i.e. minimum number of clustered units while maintaining the quality of synthetic speech) so far. Then, the automatic segmentation was performed to identify individual instances of each triphone in the

speech corpus. Finally, the same simple instance selection procedure as described in Section 2.1 was implemented.

3.3 Slovak Text-to-Speech

Once again, the Slovak text-to-speech process fundamentally copies the general one described in Section 2.2. Of course, phonetic transcription rules (in the form of [8]) specially designed for Slovak language were proposed. In the following text we will show just a couple of examples (in fact, more than 100 rules were defined in our system). For example written *t, d, n, l* are pronounced as alveopalatal [c, J\, J, L] in front of [i, i:, e]:

$$\mathrm{TDNL} \rightarrow \mathrm{ALVPAL} \,/\, \_ \; \langle \mathrm{i}, \mathrm{i}:, \mathrm{e} \rangle \,, \tag{1}$$

where $TDNL = \langle \mathrm{t}, \mathrm{d}, \mathrm{n}, \mathrm{l} \rangle$ and $ALVPAL = \langle \mathrm{c}, \mathrm{J}\backslash, \mathrm{J}, \mathrm{L} \rangle$. The symbols $<$ and $>$ define a set of phones or letters. Many exceptions to this rule exist, especially in words of foreign origin. But there are also some domestic words (e.g. *jeden*, *žiadni*) in which the rule must not be applied. Such words should be stored in a phonetic exception dictionary.

In continuous speech, groups of consonants are subject of so-called voice assimilation – simply said, all consonants in a group are either voiced or unvoiced according to the last consonant in the group. The basic rules for voice assimilation have the form:

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

$$\mathrm{UPC} \rightarrow \mathrm{VPC} \,/\, \_ \; \langle \mathrm{VPC}, |\mathrm{VPC}, |\mathrm{SON}, |\mathrm{VOW}, |\mathrm{v} \rangle \,, \tag{3}$$

where $\mathrm{VPC} = \langle \mathrm{b}, \mathrm{w}, \mathrm{d}, \mathrm{z}, \mathrm{d\_z}, \mathrm{Z}, \mathrm{d\_Z}, \mathrm{J}\backslash, \mathrm{g}, \mathrm{h}\backslash \rangle$ denotes voiced paired consonants, $\mathrm{UPC} = \langle \mathrm{p}, \mathrm{f}, \mathrm{t}, \mathrm{s}, \mathrm{t\_s}, \mathrm{z}, \mathrm{t\_S}, \mathrm{c}, \mathrm{k}, \mathrm{x} \rangle$ unvoiced paired consonants, PAU is a symbol of a pause, $SON = \langle \mathrm{m}, \mathrm{F}, \mathrm{n}, \mathrm{N}, \mathrm{J}, \mathrm{N}\backslash, \mathrm{l}, \mathrm{l}=, \mathrm{l}=:, \mathrm{L}, \mathrm{r}, \mathrm{r}=, \mathrm{r}=:, \mathrm{j} \rangle$ are (unpaired) sonorant consonants and $VOW = \langle \mathrm{a}, \mathrm{e}, \mathrm{i}, \mathrm{o}, \mathrm{u}, \{, \mathrm{a}:, \mathrm{e}:, \mathrm{i}:, \mathrm{o}:, \mathrm{u}: \rangle$ vowels. Symbol [|] denotes word boundaries (not pauses).

The same prosodic rules as for Czech were applied to Slovak. Moreover, the values of coefficients in these rules were left unchanged. Although the resulting synthetic speech sounds good, better results (i.e. more natural speech) may be obtained by adjusting the coefficients.

As for low-level speech synthesis methods, the same techniques as mentioned in Section 2.2 were employed to synthesize Slovak.

4 Conclusion & Future Work

In this paper a new Slovak language module designed for the ARTIC TTS system was presented. When creating the module, we took advantage of the experience with other languages, especially Czech. The system uses an automatically built acoustic unit inventory and a set of both phonetic and prosodic rules to convert an input text to the corresponding speech. As a fully working TTS system was

implemented, an *arbitrary* Slovak text can appear at the input of the system and the corresponding speech is produced. Although no comprehensive listening tests were carried out in the time of writing this paper, our simple informal listening tests showed a high level of intelligibility and a good naturalness of the synthetic speech. After Czech, German, and Slovak voices had been implemented within the ARTIC TTS system, the automatic HMM-based acoustic unit inventory construction process was definitely shown to be language-independent. Moreover, ARTIC can be called multilingual TTS system from now.

Since the first version of Slovak synthesis system has been designed so far, there are many parts which could be improved. There is no doubt synthetic speech could be even better. More detailed text processing (e.g. text normalization, proper text analysis avoiding ambiguous phonetization) should be worked out in the future. Prosodic rules should be tuned up to generate optimal suprasegmental characteristics of spoken Slovak. A data-based prosody model is also under construction now. Some improvements to the acoustic unit inventory construction process can be proposed as well by examining the influence of individual parameters of this process (e.g. unit-dependent HMM topology or speech parametrization) on the speech segmentation accuracy. Works on an algorithm for a dynamic on-line speech unit instance selection are also in progress now. Our future work will also comprise creating modules for other languages.

References

1. Matoušek, J., Psutka, J.: ARTIC: a New Czech Text-to-Speech System Using Statistical Approach to Speech Segment Database Construction. Proceedings of ICSLP 2000, vol. IV. Beijing (2000) 612–615.
2. Matoušek, J., Psutka, J., Krůta, J.: On Building Speech Corpus for Concatenation-Based Speech Synthesis. Proceedings of Eurospeech2001, vol 3. Ålborg (2001) 2047–2050.
3. Matoušek, J., Tihelka, D., Psutka, J., Hesová: German and Czech Speech Synthesis Using HMM-Based Speech Segment Database. Proceedings of TSD 2002. Springer-Verlag (2002) 173–180.
4. Matoušek, J., Tihelka, D., Psutka, J: Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-Specific Correction. Proceedings of Eurospeech 2003. Geneva (2003) 301–304.
5. Matoušek, J., Tihelka, D., Psutka, J: Experiments with Automatic Segmentation for Czech Speech Synthesis. Proceedings of TSD 2003. Springer-Verlag (2003) 287–294.
6. Romportl, J., Matoušek, J., Tihelka, D.: Prosody Model and its Application to Czech TTS System. Proceedings of UKROBRAZ 2002. Kyjiv, Ukraine (2002) 93–96.
7. Tychtl, Z., Matouš, K.: The Phase Substitutions in Czech Harmonic Concatenative Speech Synthesis. Proceedings of TSD 2003. Springer-Verlag (2003) 333–340.
8. Psutka, J.: Communication with Computer by Speech (in Czech). Academia, Prague (1995).
9. Czech SAMPA. http://www.phon.ucl.ac.uk/home/sampa/czech-uni.htm.
10. Kráľ, A.: Rules of Slovak Pronunciation (in Slovak). SPN, Bratislava (1996).
11. Slovak SAMPA. http://www.ui.savba.sk/speech/sampa_sk.htm.
12. Ivanecký, J., Nábělová, M.: Phonetic Transcription SAMPA and Slovak Language (in Slovak). Jazykovedny časopis, 53 (2002) 81–95.