Computing and Informatics, Vol. 35, 2016, 1467–1490

DEVELOPMENT OF THE SLOVAK HMM-BASED TTS SYSTEM AND EVALUATION OF VOICES IN RESPECT TO THE USED VOCODING TECHNIQUES

Martin Sulír, Jozef Juhár

Department of Electronics and Multimedia Communications
Technical University of Košice
Letná 9
042 00 Košice, Slovakia
e-mail: {martin.sulir, jozef.juhar}@tuke.sk

Milan Rusko

Institute of Informatics
Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava, Slovakia
e-mail: utrrrusk@savba.sk

Abstract. This paper describes the development of a Slovak text-to-speech system which applies a technique wherein speech is directly synthesized from hidden Markov models. Statistical models for Slovak speech units are trained by using the newly created female and male phonetically balanced speech corpora. In addition, contextual informations about phonemes, syllables, words, phrases, and utterances were determined, as well as questions for decision tree-based context clustering algorithms. In this paper, recent statistical parametric speech synthesis methods including the conventional, STRAIGHT and AHOcoder speech synthesis systems are implemented and evaluated. Objective evaluation methods (mel-cepstral distortion and fundamental frequency comparison) and subjective ones (mean opinion score and semantically unpredictable sentences test) are carried out to compare these systems with each other and evaluation of their overall quality. The result of this work is a set of text to speech systems for Slovak language which are characterized by very good intelligibility and quite good naturalness of utterances at the

output of these systems. In the subjective tests of intelligibility the STRAIGHT based female voice and AHOcoder based male voice reached the highest scores.

Keywords: Hidden Markov models, Slovak language, statistical parametric speech synthesis, text to speech

Mathematics Subject Classification 2010: 68-T35

1 INTRODUCTION

Text-to-speech (TTS) systems represent one of the most important part of the speech interaction with computer and the research in this area is carried out in many countries and companies around the world. Nowadays, the speech synthesis systems are represented by computer systems which can convert input text into output audio file – speech. The main task of these systems is making life easier, either to people with physical disabilities such as blind people or to ordinary people who use these systems to facilitate day to day activities. Research in this area is aimed to the point, when it will be possible to use these voices in various spheres of life, without any limitations or acting unnaturally. The use of speech synthesis in practice is often faced with the reluctance of users to communicate with device, whereas synthetic speech acts artificially and creates a barrier in communication [1]. That is why the research in the field of speech synthesis is focused on the development of new advanced methods and their improvement in order to make final speech output from these systems as close to the human interpretation as possible. However, the human speech production is a complex physiological process and this complexity is also included in the development of text-to-speech systems.

Traditional methods for speech synthesis are nowadays mainly corpus-based, so for developing an artificial voice for all languages it is necessary to have certain amount of data. The amount and quality of input data in these corpus-based systems determine the voice quality at the output [2, 3]. Statistical parametric synthesis based on Hidden Markov Models (HMMs) is one of the most effective and dynamically growing corpus-based method of speech synthesis in the last period [4]. This method uses statistical HMMs to model the spectral and prosodic parameters of the speech units. These models are trained using an input corpora which should contain phonetically balanced sentences that have the highest possible coverage of phonetic contextual units to provide a high quality speech synthesis, so the speech corpus derived from the mother text is a basic and important part in the development of corpus-based speech synthesis system.

Recently, the HMM-based speech synthesis technique has been reported for many languages, such as for large languages including the Mandarin Chinese [5], Spanish [6, 7], English [8], Portuguese [9] or Japanese [10], but the flexibility in development of those systems also enabled the integration of small languages, such

as Thai [11], Korean [12], Slovenian [13] or Greek [14]. In this paper, implementation and evaluation of newly created Slovak HMM-based speech synthesis system is described. The motivation for the development of comprehensive and applicable HMM-based speech synthesis system in the Slovak language was the absence of it as well as increased demand for the implementation of this type of speech technology in many newly created Slovak interactive applications. Moreover, the first author took the advantage of the fact, that he has been involved in the development of the Zure TTS multilingual speech synthesis system at the eNTERFACE'14 ISCA Training School, where he became familiar with the AHOcoder [26].

This paper is organized as follows: in Section 2 HMM-based speech synthesis system together with its vocoding approaches are described. Section 3 describes implementation of the Slovak HMM-based speech synthesis system from the input corpus design to the synthesis of a given utterance. In Section 4, experiments and results of objective and subjective evaluation are presented. The conclusions are listed in Section 5.

2 VOCODING APPROACHES IN HMM-BASED SPEECH SYNTHESIS SYSTEM

Statistical parametric synthesis based on HMM is the method of speech synthesis, which has been gaining in popularity in recent years, as evidenced by the research and development around the world [3]. Most of the research, development and implementation activities in this field are associated with HTS tools, which represent basic tools for HMMs training and partly also for speech generation from these models [15]. Together with this set of tools it is necessary to use text analysis module, which is not included in HTS tools. Language dependent text analysis module is responsible for input text processing and for these purposes, couple of complex software tools may be used, such as Festival Speech Synthesis System [16], DFKI MARY Text-to-Speech System [17] or Flite [18]. These software tools allow to use pre-built text analysis modules for languages, such as for example English, Japanese or German and they also provide set of tools for own text analysis module development. Figure 1 shows a block diagram of HMM-based speech synthesis system.

The system for speech synthesis is divided into two parts, namely the training part and the part of speech synthesis. The basic principle of this method is to use the context-dependent HMM models, which are trained from speech corpus, as generative models for speech synthesis process [4].

The main task of the training part is the extraction of spectral and excitation parameters from speech corpus as well as the implementation of HMMs training. In HMM-based speech synthesis system, each HMM correspond to left-to-right model with explicit state durations (Hidden Semi-Markov Models) where each output vector is composed of two components, namely it consists of the spectrum part represented by mel-cepstral coefficients and their delta and delta-delta coefficients;

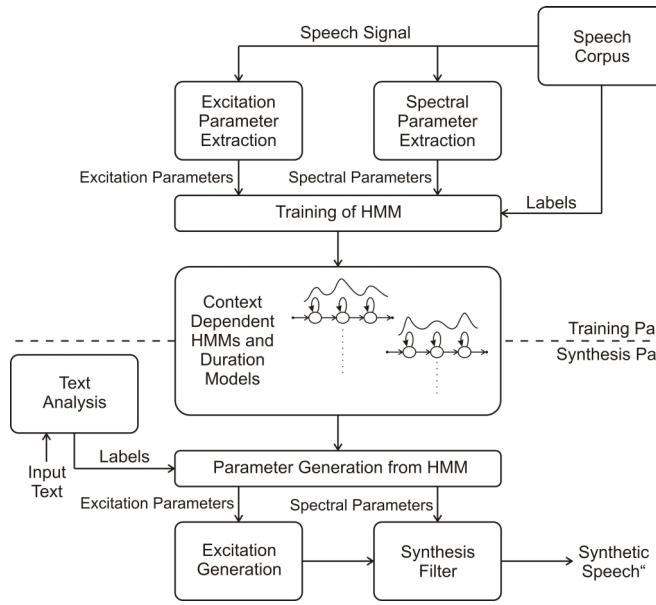

Figure 1. Block diagram of the HMM-based speech synthesis system

and the excitation part which is represented by excitation parameters and the corresponding delta and delta-delta dynamic features [19]. The HMMs include also density distributions for state duration for the purposes of reproduction of temporal structure of speech. The training of HMMs using fundamental frequency and mel-cepstrum simultaneously is enabled in a unified framework by using multi space probability distribution HMMs and multidimensional Gaussian distributions [20]. The simultaneous modeling of fundamental frequency information together with spectrum resulted in the set of context-dependent HMMs. Context-dependent clustering of Gaussian distributions is performed independently for spectrum, fundamental frequency information and duration because of the influence of the different clustering factor.

The synthesis part of HMM-based speech synthesis system consists of two main components. First component is represented by text analyser, which converts given text into contextual label sequence. Second component consists of several blocks which are responsible for parameter generation from context dependent HMMs and duration models; excitation generation based on generated excitation parameters and synthesis filter. This component composes an HMM sequence by concatenating context-dependent HMM according to input label sequence. Subsequently, state durations for the concatenate HMM sequence are determined in order to maximize the output probability of the state durations. From the obtained HMM sequence with the appropriate estimated state durations included, a sequence of mel-cepstral coefficients and logarithmical values of fundamental frequency (in-

cluding the voiced/unvoiced decisions) are generated using Case 2 of the algorithm which is presented in [21]. These vectors of mel-cepstral coefficients and logarithmic values of generated fundamental frequency values (logF0) represent input into speech synthesis filter and with its help the final speech waveform is formed.

Conventional TTS system based on HMMs works as mel-cepstral vocoder with a simple impulse train as the excitation signal, where a sequence of periodic pulses and white noise together with MLSA (Mel Log Spectrum Approximation) filter are used. Speech synthesis process is carried out by filtering of pulse train in case of voiced segments, and white noise filtering in case of unvoiced speech segments. Excitation is controlled by logarithmic values of generated fundamental frequency. The parameters of filter are adjusted according to the input mel-cepstral coefficients obtained in the parameters generation process. However, the use of such type of vocoder, which utilizes a simple model of excitation causes, that the speech acts too robotic.

To deal with the problems presented above, several high-quality vocoders with a more advanced excitation were implemented into HMM-based speech synthesis system. Such methods include e.g. MELP (Mixed Excitation Linear Prediction) method [22], HSM (Harmonic/Stochastic Model) model [23], excitation model based on modeling of residues [24], STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of Weighted Spectrum) [25] or AHOcoder [26]. In the next section, mixed excitation signal based on STRAIGHT and AHOcoder is described, because these two vocoders together with conventional MLSA filter system were adopted for the Slovak language.

STRAIGHT vocoding method represents an improvement of conventional HMM-based speech synthesis system, realized through three basic steps. The first step of STRAIGHT vocoding is fundamental frequency (F0) extraction, which is carried out by fixed-point analysis [25]. Subsequently, the periodicity of signal is removed in time domain with the help of F0-adaptive spectral analysis with surface reconstruction method and the methods for measuring aperiodicity of signal are applied [25]. Furthermore, a mixed excitation is applied by a weighted mixture of white noise and pulse train in the synthesis part. The additional modifications include phase manipulation on the excitation signal and use of aperiodicity coefficients in the weighting process.

AHOcoder vocoding method parametrizes speech waveforms into three different streams, namely it handles with the logarithmic values of F0, mel-cepstral coefficients and the so-called maximum voiced frequency, which stands for the local degree of harmonicity of the given signal [26]. Assuming that a speech frame is the sum of a harmonic component and a noise component which occupy the lower and the upper band of speech, respectively, then AHOcoder measures the logarithmic values of F0, the mel-cepstral representation of the spectral envelope, and the maximum voiced frequency defined as the boundary between harmonic and noisy bands.

3 IMPLEMENTATION OF HMM-BASED SLOVAK VOICES

The implementation of HMM-based speech synthesis system for the Slovak language follows the framework described in Figure 1. Its adaptation considers the particular characteristics of the Slovak language and mainly focuses on the analysis and the contextual modeling since contextual information is language dependent. However, the HMM framework provides a general setup for sufficient context modeling that can be adopted for many different languages. An important part of the implementation of the Slovak speech synthesis system was the design of the Slovak speech corpora which were designed especially for the Slovak HMM-based speech synthesis with respect to their phonetic coverage. Additionally, the module of text analysis with large pronunciation dictionary has been designed for the Slovak text processing. All the above mentioned parts will be described in more detail in the following sections.

3.1 Characteristics of the Slovak Language, Contextual Information and Its Context Clustering

The Slovak language belongs to the group of West Slavic languages. The Slovak alphabet consists of forty-nine letters, where seventeen of them represent vowels (including 4 diphthongs) and the other thirty-two letters represent consonants. The newly created speech synthesis system employs a set of 42 phonemes (including long and short pause models) as the basic acoustic units. A distinction between the number of letters in alphabets and the number of phonemes is caused by the same phonetic form of different letters (for example letter i and y have the same phonetic transcription and so on). Slovak SAMPA (Speech Assessment Methods Phonetic Alphabet) was used for phoneme representation [27]. Table 1 shows the Slovak vowels classification (in orthoepic SAMPA format and orthographic format) while Table 2 shows the Slovak consonants classification.

		Tongue Position							
		Front			Back				
		Short	Long		Short	Long			
Tongue Height	High	I [i,y]	I: [í,ý]		U [u]	U: [ú]	I_^U [iu]	Close lip rounding	Lip Rounding
	Mid	E [e]	E: [é]	I_^E [ie]	O [o]	O: [ó]	U_^O [ô]	Open lip rounding	
	Low	{ [ä]			a [a]	a: [á]	I_^a [ia]	Neutrally open	
		Monophthong		Diphthong	Monophthong		Diphthong		

Table 1. Slovak vowels classification

Contextual information extraction is one of the language dependent aspects of HMM-based speech synthesis system. Contextual labels, which represent input

V - voiced U - unvoiced		Place of Articulation												
		Bilabial		Labiodental		Alveolar		Postalveolar		Palatal		Velar		Glottal
		V	U	V	U	V	U	V	U	V	U	V	U	V
Manner of Articulation	Nasal	m [m]				n [n]				J [ň]				
	Plosive	b [b]	p [p]			d [d]	t [t]			J\ [ď]	c [ť]	g [g]	k [k]	
	Affricate					dz [dz]	ts [c]	dZ [dž]	tS [č]					
	Fricative			v [v]	f [f]	z [z]	s [s]	Z [ž]	S [š]				x [ch]	h\ [h]
	Trill					r, r=: [r, ŕ]								
	Approximant					l, l=: [l, ĺ]		L [ľ]		j [j]				

Table 2. Slovak consonants classification

to the speech synthesis system, are used to determine the corresponding HMMs in the set of models. Accordingly, contextual information which is represented in contextual labels was necessary to be considered in order to obtain the best possible quality of prosody representation at the output of the system. The following basic contextual information was considered for the first experiments with the Slovak HMM-based speech synthesis system:

- Level of phonemes
  - phoneme identity {before previous, after next} phonemes;
  - {previous, current, next} phonemes identity;
  - position of the current phoneme identity in the current syllable {forward, backward};
- Level of syllables
  - number of phonemes in {previous, current, next} syllables;
  - accent in {previous, current, next} syllables;
  - stress in {previous, current, next} syllables;
  - position of current syllable in the current word {forward, backward};
  - position of current syllable in the current sentence {forward, backward};
  - number of stressed syllables {before, after} the current syllable in the current sentence;
  - number of accented syllables {before, after} the current syllable in the current sentence;
  - number of syllables from the previous {stressed, accented} syllable to the current syllable;

  - number of syllables from the current syllable to the next {stressed, accented} syllable;

- Level of words
  - part-of-speech of the {previous, current, next} words;
  - number of syllables of the {previous, current, next} words;
  - position of current word in current sentence {forward, backward};

- Level of sentences
  - number of {phonemes, syllables, words} in the sentences;

In order to carry out decision tree-based context clustering, which is applied to generate the respective unseen model, when a given contextual label does not have a corresponding HMM in the set of trained models, a set of context clustering questions was proposed. The questions were proposed according to phonetic characteristics of vowels, diphthongs, and consonants of the Slovak language. The following phoneme categorisation was proposed:

- vowel – {monophthong, diphthong}, {front, back}, {long, short}, {high, mid, low}, {close lip-rounding, open lip-rounding, neutrally open}, {a-, e-, i-, o-, u-vowel};
- consonant – {voiced, unvoiced}, {nasal, plosive, affricate, fricative, trill, approximate}, {bilabial, labiodental, alveolar, post-alveolar, palatal, velar, glottal};
- silence and pause;

3.2 Phonetically Balanced Slovak Speech Corpus for Speech Synthesis

Two single speaker speech corpora have been carefully recorded under studio conditions for purposes of the Slovak HMM-based speech synthesis, where each consist of 4 526 phonetically balanced sentences [28]. In the following section of this subsection we describe all the steps that were necessary to obtain these corpora, namely the input text data filtering, the optimal text selection, the recording of obtained sentences and their processing.

Input text corpus for optimal text selection represents text data which were collected for the purposes of language modeling in speech recognition system [29]. The following data represents the Slovak sentences (over 10 million sentences) which are divided into separate lines and normalized. Therefore, these text data constitute a sufficient basis for building the phonetically balanced text corpus for speech synthesis, but it was necessary to further modify them. The first modification that had to be done was filtering of too long sentences because these sentences would be too difficult to record and in many cases they contain meaningless concatenated words and normalization tags. The same procedure was carried out with too short sentences that contain only one word. The next step was filtering of sentences that

contain words the pronunciation of which is not clearly defined in the Slovak language. In many cases those were abbreviations, non-Slovak words and misspelled words.

The output of these steps was the text corpus, which contains two and a half million sentences, so something more than 7 million of sentences were filtered out. The filtered text corpus was divided into five subcorpora, where each subcorpus contains approximately five hundred thousand of sentences and each of them were then processed separately. The separate processing of smaller subcorpora enables us to use the standalone or combined corpora in the speech synthesis for evaluation its quality depending on the size of the speech corpus or, for example, we can divide combined corpora for training and evaluation sets.

Each subcorpus was subsequently transferred to its phonetic form, because the optimal text selection is carried out at the level of phonemes. The optimal text selection algorithm based on Greedy algorithm was then applied on each transcribed subcorpus in level of diphones. The diphones have been selected in order to minimize the computational demands because they represent a compromise between their informative value and selection time requirement. This algorithm selected approximately 1 000 sentences from the input subcorpus which represented the phonetically balanced subcorpus. After this selection, it was necessary to manually check the individual selections and delete the sentences that still contain the inappropriate words for recording. The last step of the subcorpora processing was re-selection of the most appropriate sentences from the manually checked selected sentences of each subcorpora. The resulting phonetically balanced corpus was obtained by combining the subcorpora together. The average length of the selected sentences is equal to 9.748 words with standard deviation 2.693 words with the limited sentence length from 2 to 14 words. The results of optimal text selection with the proposed methodology are shown in Table 3.

Subcorpus	No. of Different Phonemes	No. of Selected Sentences	No. of Different Diphones	Total Number of Diphones in Selection
1	45	962	1 211	40 919
2	45	933	1 195	39 229
3	45	888	1 193	38 542
4	45	884	1 201	37 749
5	45	859	1 198	36 610
Total	45	4 526	1 260	193 049
Percentage coverage: 1 260/1 600 = 78.75 %				

Table 3. Optimal text selection results

As we can see, the final version of corpus contains 4526 sentences and the total number of different diphones is 1260. Considering that the speech synthesis in the Slovak language uses from 1 200 to 1 600 diphones, we can say that the percentage coverage of this selection is 78.75 % in the worst case (if we consider 1600 diphones). The following percentage coverage of the elements of language is a good result re-

gardless of complexity of the Slovak language as such (Slovak language belongs to the group of inflected languages). The final step in the creation of the phonetically balanced speech corpus was the recording and processing of the selected sentences. The recording took place in a professional recording studio with one male and one female speaker. The same set of 4526 phonetically balanced sentences was recorded by each of them. A Neumann TLM 103 cardioid condenser microphone with SPL Gold Mike Mk2 pre-amplifier and a hard disk recording system equipped with RME Fireface 400 audio interface was used during the sessions. A 48 kHz sampling frequency and 16 bit resolution were used. Because the recording was carried out in parts, where each part contained one hundred sentences, it was necessary to process them into isolated sentences. Also, the segments of silence had to be aligned at about 200 milliseconds at the beginning and end of the sentences [30]. A detailed specification of obtained subcorpora is shown in the Table 4.

	Male Voice 48 kHz	Female Voice 48 kHz
Number of sentences	4 526	4 526
Total duration	5 hrs 50 mins	6 hrs 51 mins
Total duration (without silences)	5 hrs 20 mins	6 hrs 14 mins

Table 4. Slovak speech corpus specification

3.3 Text Analysis Module

Text analysis module is one of the most important part of speech synthesis processes. It provides conversion of input raw text into sequence of contextual labels and instruction for speech synthesis filter in synthesis part of HMM-based speech synthesis system. The implementation of this module for the Slovak language was carried out using software tool Festival [16], which allows building individual block of this module as well as the final speech synthesis based on HMMs through integrating relevant modules and blocks. The first step in the raw text processing is text analysis that means analysis of input raw text into pronounceable words and sentences. In practical use of TTS systems, input text contains many of Non-Standard Words (NSW), such as for example numbers, abbreviations, acronyms, etc., and they need to be normalized. Then, it is necessary to convert the normalized text into orthoepic form by which the input processed text can be easily synthesized. Therefore, the above mentioned steps had to be implemented into Festival tool for the Slovak language, since the text processing is a language dependent task. The output of the sections below is a new language text analysis module for the Slovak language for Festival which can be used without limitations with supported methods of speech synthesis and it also allows building an arbitrary new voice.

3.3.1 Slovak Text Analysis and Normalization

The Festival new language implementation can be divided into several steps. At the beginning it was necessary to define phoneset together with the detailed description of each phone in term of its classification. The phoneset has been defined as indicated in Tables 1 and 2. The basic model of text analysis in Festival means that each token will be mapped as a list of words. In the Slovak language, this token based method was used for the unambiguous abbreviations, acronyms and for cardinal numbers. The lists of unambiguous acronyms and abbreviations were created where altogether more than a thousand of these tokens were defined. Subsequently, a list of punctuation and special characters which describes token to word rules for punctuation and special characters in their varied occurrences were also defined. This list also includes multiple variations for some entries to solve their ambiguity when they are used in different context and situations (for example symbol "-" may represent a minus symbol but also a hyphen). The text analysis module with pronunciation token to the word list is able to normalize mail addresses, percentage expressions, simple "one line" mathematical equations and so on. A problem with cardinal numbers tokens was solved algorithmically. The algorithm based on the numbers concatenation was implemented with the help of Festival Scheme programming language. In this case, the list includes token to word rules for the numerals from zero to nineteen and round numbers with irregular pronunciation. The other numbers are already algorithmically created by concatenation. This solution provides token to word conversion for the numbers up to trillions. The newly created NSW groups together with examples are shown in Table 5.

NSW Group	Written Format	Token to Word Format	SAMPA Transcription
Abbreviations	atď.	a tak ďalej	a tak J\alEj
Acronyms	MVSR	ministerstvo vnútra slovenskej republiky	mInIstErstvO vnU:tra slOvEnskEj rEpUblIkI
Punctuation & special characters	martin.sulir @tuke.sk	martin bodka sulir zavináč tuke bodka es ká	martIn bOdka sUlI:r zavIna:tS tUkE bOdkA Es ka:
Cardinal numbers	1988	tisíc deväťsto osemdesiat osem	cIsI:ts J\Ev{cstO OsEmJ\EsI_^at OsEm

Table 5. Slovak NSW groups with examples

3.3.2 Slovak Text Phonetic Analysis

Phonetic analysis for the Slovak language in Festival is primarily ensured by pronunciation dictionary, where the system is looking for direct transcription of a word to its phonetic form. If the input word is not listed in the dictionary letter-to-sound (LTS) rules were defined where the word is transcribed grapheme by grapheme to its phonetic form. The LTS rules were defined as a direct transcription from grapheme

to phoneme. The approach based on the LTS rules is very effective in case where orthographic form is equal to phonetic form (it is necessary only remap graphemes to phonemes) which represents the majority of cases in the Slovak language. The exceptions are represented by various specific phenomena such as softening of some syllables at the end of words or mutation of some phonemes and so on. Solution, where the text analysis module uses LTS rules for words with direct transcription and a pronunciation dictionary only for words where transcription is not straightforward was also adapted.

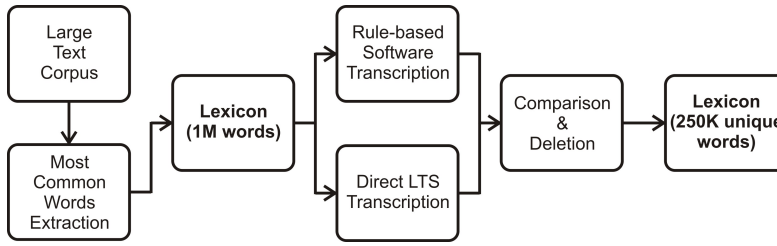

Figure 2. Block diagram of pronunciation lexicon creation process for Slovak language

Pronunciation dictionary was especially build for purposes of the Slovak HMM-based speech synthesis. It consists of 250 000 words which are listed together with their phonetic transcription and divided to syllables. All words in the dictionary have different grapheme and phonetic form. It is very effective solution, because searching for words in the dictionary is a computationally demanding task, so that we were able to decrease size of dictionary from 1 000 000 of most common words in the Slovak language to around 250 000. The pronunciation lexicon arose from a large text corpus from which one million of most common Slovak words were extracted. Subsequently, the two transcription methods were applied to this lexicon. The first one consisted of the application of a rule-based software tool specially built for Slovak grapheme to phoneme transcription. This tool was additionally modified to perform the transcription as well as a hyphenation into syllables what was necessary for marking syllables with the accent. The second approach was based on the application of the direct LTS transcription, where graphemes were directly rewritten to their phonetic form. Then they were compared and similar entries were deleted which resulted in the pronunciation dictionary which contains only the unique words with the transcription with exceptions.

3.4 Description of Speech Synthesis Systems for Evaluation

Together, six HMM-based speech synthesis systems were created and evaluated for the Slovak language speech synthesis. The HTS toolkit together with the newly created text analysis module implemented in Festival and appropriate speech parametrization technique was used. Each of these newly created systems were based on previously described speech corpora, where they were divided into training part

and evaluation part. The training part consisted of four subcorpora marked as Subcorpus 1–Subcorpus 4 (as indicated in Table 3) and evaluation part was formed by the remaining Subcorpus 5. It was necessary to divide corpora because Subcorpus 5 was used for objective evaluation with MCD (Mean Cepstral Distortion) method, where reference waveform is compared with its synthetic version. A default vector dimension was used and the fundamental frequency values were extracted with the help of AHOcoder extraction method for each vocoder. A more detailed description of systems is shown in Table 6.

System	Vocoder	Input Corpus	Parameters
Conventional female voice (FS1)	Mel – generalized cepstrum	Female corpus 3667 phonetically balanced sentences, 5 hrs 11 mins (w/o silence)	MGC: 34 + F0*: 1, Pulse plus noise excitation
STRAIGHT female voice (FS2)	STRAIGHT with critical band excitation		MGC: 49 + F0*: 1, Band Aperiodicity: 25, Multi-band Mixed Excitation
AHOcoder female voice (FS3)	AHOcoder HNM		MGC: 39 + F0*: 1, Multi-band Mixed Excitation
Conventional male voice (MS1)	Mel – generalized cepstrum	Male corpus 3667 phonetically balanced sentences, 4 hrs 19 mins (w/o silence)	MGC: 34 + F0*: 1, Pulse plus noise excitation
STRAIGHT male voice (MS2)	STRAIGHT with critical band excitation		MGC: 49 + F0*: 1, Band Aperiodicity: 25, Multi-band Mixed Excitation
AHOcoder male voice (MS3)	AHOcoder HNM		MGC: 39 + F0*: 1, Multi-band Mixed Excitation
* Same F0 coefficients for all male and all female systems extracted with AHOcoder F0 extraction algorithm, because it provides the most accurate extraction			

Table 6. Description of newly created Slovak HMM-based speech synthesis systems

4 EVALUATION AND RESULTS

The evaluation of newly created voices was performed by objective and subjective tests. An evaluation part of the Slovak speech corpora, which was created by separating of one fifth of the entire corpus, was used for objective evaluation. Taken together, a spectrum of generated utterances was evaluated by measuring of their Mean Mel-Cepstral Distortion and F0 contours were compared with F0 contour of reference database. Subjective evaluation was carried out using a web-based questionnaire. The intelligibility and naturalness of synthesized speech were evaluated with the help of MOS (Mean Opinion Score) and SUS (Semantically Unpredictable Sentences) tests. The detailed explanation of all experiments and their results will be given in the following subsections.

4.1 Objective Evaluation

As mentioned above, MCD evaluation method together with generated F0 values comparison were used for objective evaluation of the newly created Slovak HMM based TTS systems. MCD evaluation method represents a distance measure calculated between mel-cepstral coefficients of reference (or original) and evaluated speech samples [31]. Mel-cepstral distortion is defined by

$$MCD = \frac{10}{\ln(10)} \sqrt{2 \sum_{d=1}^{D} (mc_d^{orig} - mc_d^{synt})^2} \; [dB] \tag{1}$$

where mc_d^{orig} and mc_d^{synt} are the d^{th} mel-cepstral coefficient of the original and synthesized speech sample. The 0^{th} cepstral dimension is not considered in MCD computation because it describes the overall signal power, but we evaluated speech samples with constants speaker's loudness (studio recording etc.), so the distortion measure is not as much influenced by the signal power.

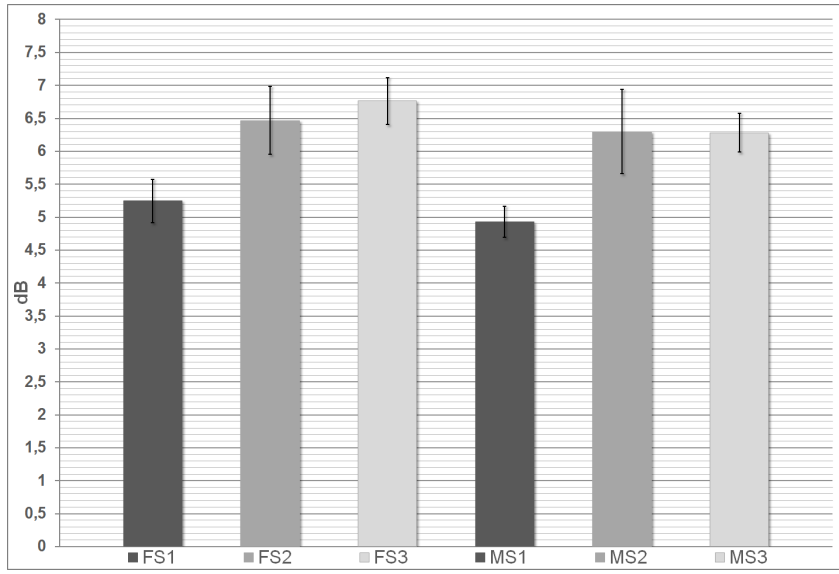

Figure 3. Mean MCD objective test results with SD for various HMM-based Slovak TTS systems

In these experimentation 859 recordings of one reference male and one reference female speaker are considered. We applied this metric to the mel-cepstral coefficients generated by six test systems, which are described in Subsection 5.4. These coefficients were compared with the reference mel-cepstral coefficients, which were extracted from Subcorpus 5 of male and female speaker. The content of generated

speech samples of all system and all utterances had the same content as the reference. Finally, the acquired results from all 859 comparisons were averaged to obtain the Mean Mel-Cepstral Distortion of each evaluated system. The results of objective evaluation of Slovak HMM-based speech synthesis systems with Mean MCD method are shown in Figure 3.

As we can see, the obtained result for all voices are in the range from 4.9 to 6.8 dB. The MS1 together with FS1 system, which represent male and female conventional HMM-based system obtain the best score. It is also evident, that the score of the remaining four systems is almost the same. The key difference is only in different standard deviations (SD), where STRAIGHT-based systems (FS2 and MS2) achieve much larger range of values in individual measurements. This may indicate ample variations in the quality of individual synthesized utterances. From the results it is also clear that the male voices seems to have higher quality than the female voices.

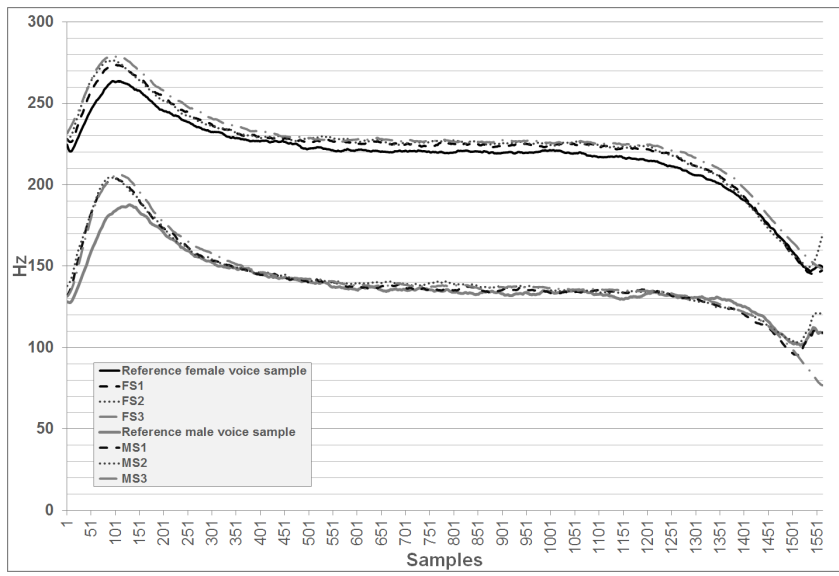

Figure 4. Comparison of aligned F0 values of Slovak HMM TTS

Objective evaluation was also made for the fundamental frequency of the generated utterances. In this case, the evaluation part of created corpora was used again for the assessment whether the generated F0 values are correct and appropriate. Evaluation procedure consisted of fundamental frequency extraction from male and female reference database together with the same extraction from generated utterances of all six newly created Slovak TTS systems. This extraction was performed on each of the 859 recordings of each system with AHOcoder F0 extraction tool and subsequently the alignment was performed with the help of interpolation to

achieve vector alignment to the same length for averaging and comparison. Alignment of each of the 859 vectors of each system was followed by averaging of the values in each sample of acquired vectors to get the average fundamental frequency contour of male and female generated utterances and reference samples. The comparison of aligned and averaged F0 values of all Slovak HMM-based speech synthesis systems together with the reference male and female contour are shown in Figure 4.

From the results it is apparent that almost all tested systems generate fundamental frequency of artificial speech with almost the same values as it is in the reference. The basic difference is evident at the beginning of each contour of TTS systems, where these contours reach higher values, but in general we can say, that the contours of each TTS system to a large extent follow the fundamental frequency contour of reference as in the case of the female and also the male voice.

4.2 Subjective Evaluation

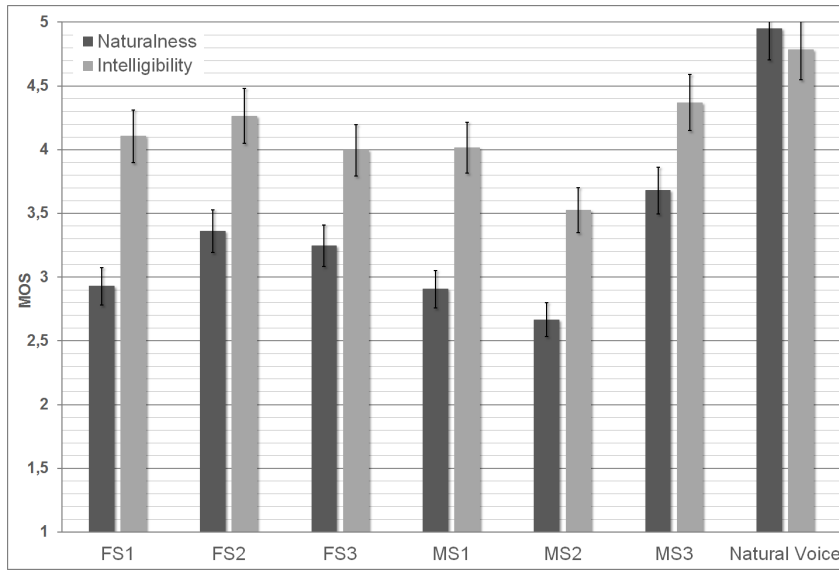

Figure 5. MOS naturalness and intelligibility evaluation results. Confidence interval of 95 % is shown

In case of subjective evaluation of newly created Slovak HMM-based speech synthesis systems, three experiments were employed. One large MOS (Mean Opinion Score) listening test was conducted to evaluate the intelligibility and naturalness of obtained speech. Subsequently, the SUS (Semantically Unpredictable Sentences) test was conducted, and finally the various aspects of speech were evaluated with the help of news articles extracted from internet. A total of 40 native Slovak speakers

participate in the experiments. The average age of participants was 27.251 years, with standard deviation of 6.967 years. There were 11 undergraduate participants, 16 with Master degree, 7 with Ph.D. degree and 1 Full Professor. Together, 3 participants have never listened to synthetic speech, 12 participants listen to it rarely, 10 participants yearly, 9 monthly, 4 weakly and 2 listen to synthetic speech daily. Only nine participants have indicated that they are speech experts.

The evaluation consisted of 20 sets of recordings where each of them was composed of 30 recordings (15 different sentences for male and female voice which consisted of 5 sentences of individual systems) for MOS naturalness and intelligibility evaluation, 30 recordings (the same layout as in the previous case) for SUS intelligibility evaluation of the same systems and 6 recordings (3 different articles for male and female voice which consisted of 1 article of individual systems) of synthesized news articles with the help of female and male synthetic voices. Together, 300 randomly selected sentences extracted from the large Slovak text corpus were subjected to MOS scale evaluation, 300 of previously prepared semantically unpredictable Slovak sentences were used in SUS intelligibility evaluation and 20 news articles were evaluated in the last part. Generated sentences of individual systems were arranged in random order in each subset, so as to be difficult to predict the used system and each of the systems were evaluated with the same subsets of recordings for the purposes of comparison of obtained results.

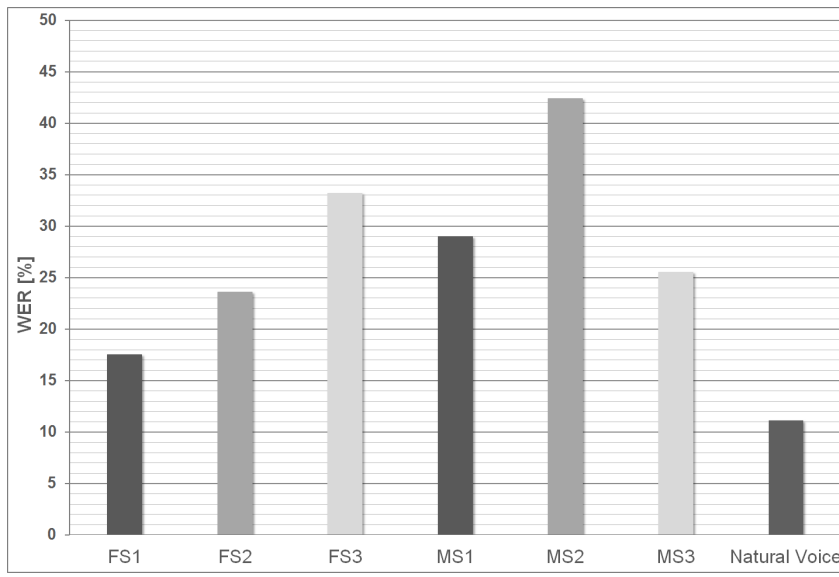

Figure 6. SUS intelligibility evaluation results

The first part of the assessment of all systems was the evaluation of basic speech parameters (intelligibility and naturalness) using MOS scale. This scale represents

the subjective rate for overall speech quality testing. This type of test is based on evaluation of the individual speech parameters using the scale 1 to 5 and subsequently calculating the arithmetical mean of obtained evaluations. The scale is composed of five values assigned to the grade so that the value 1 is the poor quality and the value of 5 is the excellent quality of the evaluated parameter. The results of subjective evaluations of naturalness and intelligibility with the help of MOS scale are shown in Figure 5.

It is evident, that intelligibility of individual systems reached a value between 3.5 and 4.3 what can be considered as a very satisfactory result. In case of female voice, the best results were achieved using the STRAIGHT-based system, but on the other hand, male voice based on AHOcoder was assessed as the most intelligible (by a considerable margin to compare the STRAIGHT-based one). In general, we can say that all newly created systems are quite intelligible. The evaluation of synthetic speech naturalness has shown that each of the newly created systems reached a value between 2.6 and 3.6 which is substantially lower than in the case of intelligibility of synthetic speech. These lower values were caused by using a filter at the output of the entire HMM-based speech synthesis system but it is evident that the naturalness evaluation results copies the intelligibility results in each of the systems what can together indicate their overall quality.

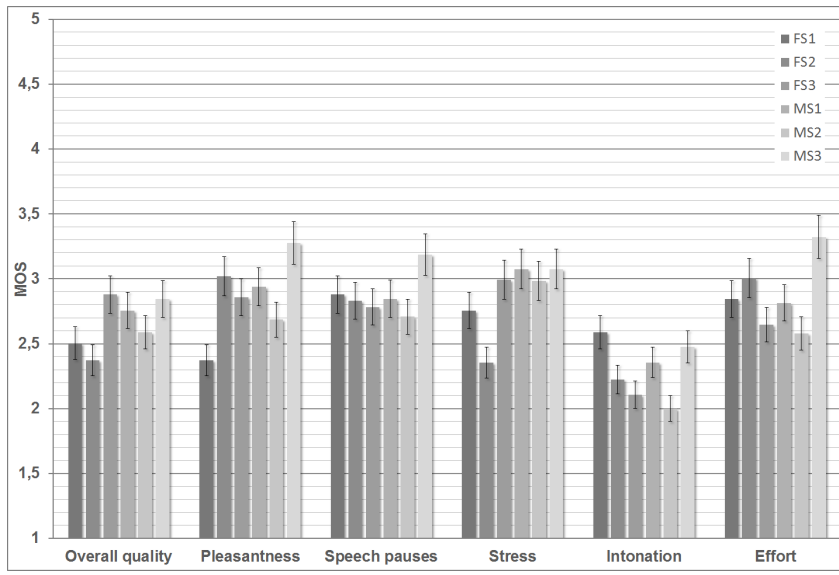

Figure 7. Speech synthesis quality evaluation by using of short news articles results. Confidence interval of 95 % is shown

The Semantically Unpredictable Sentences test was primarily developed for the purposes of speech intelligibility testing of TTS systems at the sentence level [32]. The basic idea of this test is to use the semantically unpredictable sentences –

sentences without meaning, but the essential requirement is that the sentences have to be syntactically correct. This structure of sentences allows users to understand every word in the sentence, but prevents estimating its contents from the initial words of the sentence. The participants must devote sufficient attention to each word in a sentence and this allows better evaluate speech intelligibility in case of TTS systems. The simplest way to score the results of evaluation is to take into account the correctly recognized words in sentences. In this case participants write only the words they capture. The score is then calculated as a percentage of incorrectly captured words from the set of all words in a sentence.

The newly created Slovak HMM-based TTS systems were evaluated with previously proposed set of semantically unpredictable sentences for Slovak language [33]. The essential feature of these newly proposed sentences is that they are phonetically balanced what allows to evaluate as many phonetic units as possible with the help of this evaluation method. Together 300 semantically unpredictable sentences with special syntactic structure were used in evaluation where these sentences were synthesized and the participants had to write all captured word of the sentence.

Each recording could be played only once in this test to obtain the most accurate evaluation of artificial speech intelligibility. The acquired assessments were subsequently shown as word error rate (WER), which represented a percentage share of incorrectly recognized words. The results of subjective evaluations of intelligibility with the help of SUS test are shown in Figure 6. From the results it is apparent that almost all tested systems got very good WER percentage score. The conventional female TTS labelled as FS1 even got WER score equal to 17.5 % what is a very good result especially when recordings can be played only once in the evaluation. On the other hand, STRAIGTH-based male TTS system gained the largest value of WER score (42.4 %) what is clearly related also with the lowest MOS score in the previous subjective evaluation.

The last part of the evaluation of the Slovak TTS systems was the assessment of artificial speech quality in short news articles, which represented a situation in which the speech synthesis is commonly found. 20 news articles extracted from online news websites were used in this type of evaluation. The average words in each article was about 53.619 words, with standard deviation of 17.716 words. In this case, the participants listened to recording of synthetic speech and subsequently they evaluated the overall artificial speech quality of presented recording, how pleasing the voice is, whether the pauses in speach are appropriate, they evaluated stress, intonation of speech and listening effort. This evaluation was carried out using the MOS scale in all cases. The results of speech synthesis quality evaluation by using of short news articles are shown in Figure 7.

The results show a slight decline in MOS score values in all evaluated parameters in comparison with values obtained in case of naturalness and intelligibility. A lower score in all evaluated parameters may result from the lower naturalness of the newly created voices and because these voices may be annoying in the speech synthesis of longer texts, what could also be reflected in this evaluation.

The MOS values for intonation are relatively lower than other measurements. This can indicate that the coverage of the intonation phenomena is insufficient in the databases and adding a special set of intonation rich sentences could help.

5 CONCLUSIONS AND FUTURE WORK

This paper presents the current development of a text-to-speech system based on HMM for the Slovak language. The performance of the systems have been evaluated through objective and subjective listening tests. Three vocoding techniques have been adopted for Slovak HMM-based speech synthesis with the help of specifically developed male and female speech corpora. Objective and subjective evaluation results showed that each of these systems provides very good intelligibility of artificial speech at the output. Naturalness of synthetic speech reaches maximum possible quality which is, however, limited by the current state of the used technology and the limited volume of the recorded speech data in databases. The results also points to the qualitative differences between the different vocoders. However, these results are encouraging since this is one of the first HMM speech synthesis system built for the Slovak language, and many improvements are possible. The results are also highlighted by the fact that the newly created Slovak speech synthesis systems have already been used in some practical applications and implementations. Firstly, these voices were adopted into multimodal interface for controlling functions of the modular robotic system which can be used in difficult conditions such as rescue works, natural disasters, fires or decontamination [34]. The Slovak HMM-based speech synthesis systems were implemented to produce feedback for the operator, in addition a dialogue manager technology was adopted, which allows to perform the information exchange between operator and robotic system. These voices were also used as part of the new version of ZureTTS system which is an initiative of Aholab Signal Processing Laboratory of University of the Basque Country to provide a personalized speech synthesizer to people with speech impairments, and also to those who completely lost their voices [35]. The new version of ZureTTS system was undertaken by an international team of researchers during eNTERFACE'14 ISCA Training School, covering up to 8 languages: English, Spanish, Basque, Catalan, Galician, Chinese, German and also Slovak.

Future work will focus on the adoption of a more sophisticated model for speech reconstruction and the inclusion of more prosodic properties in order to increase the naturalness of the produced speech. The text analysis module improvement will also form part of future work. The primary task will be in its extension by other functionalities and thereby to ensure increasing flexibility of the permissible input. Improvement of this module thus allows us to use this speech synthesis system in different applications. Speaker and emotional adaptation techniques for HMM-based speech synthesis will also be one of the areas of interest in the near future with regard to the processing of existing automatic speech recognition system corpora.

Acknowledgements

The research in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project VEGA 2/0197/15 and the Slovak Research and Development Agency under the project APVV-15-0731.

REFERENCES

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

[2] Dutoit, T.: Corpus-Based Speech Synthesis. Springer Handbook of Speech Processing. Springer Berlin Heidelberg, Berlin, 2008, pp. 437–456.

[3] King, S.: Measuring a Decade of Progress in Text-to-Speech. Loquens, Vol. 1, 2014, No. 1, Article No. e006.

[4] Tokuda, K.—Nankaku, Y.—Toda, T.—Zen, H.—Yamagishi, J.—Oura, K.: Speech Synthesis Based on Hidden Markov Models. Proceedings of the IEEE, Vol. 101, 2013, No. 5, pp. 1234–1252.

[5] Qian, Y.—Soong, F.—Chen, Y.—Chu, M.: An HMM-Based Mandarin Chinese Text-to-Speech System. Chinese Spoken Language Processing. Proceedings of the 5$^{\text{th}}$ International Symposium on Chinese Spoken Language Processing (ISCSLP 2006). Lecture Notes in Computer Science, Vol. 4274, 2006, pp. 223–232.

[6] Gonzalvo, X.—Iriondo, I.—Socoro, J.C.—Alias, F.—Monzo, C.: HMM-Based Spanish Speech Synthesis Using CBR as F0 Estimator. Proceedings of ISCA Tutorial and Research Workshop on Non Linear Speech Processing (NoLISP 2007), 2007, pp. 7–10.

[7] Camacho, A. H.—Ávila, F. R.: Development of a Mexican Spanish Synthetic Voice Using Synthesizer Modules of Festival Speech and HTS-Straight. International Journal of Computer and Electrical Engineering, Vol. 5, 2013, No. 1, pp. 36–39.

[8] Tokuda, K.—Zen, H.—Black, A. W.: An HMM-Based Speech Synthesis System Applied to English. Proceedings of 2002 IEEE Workshop on Speech Synthesis 2002 (WSS-02), 2002, pp. 227–230.

[9] Maia, R. S.—Zen, H.—Tokuda, K.—Kitamura, T.—Resende, F. G. V.: An HMM-Based Brazilian Portuguese Speech Synthesizer and Its Characteristics. Journal of Communication and Information Systems, Vol. 21, 2006, pp. 58–71.

[10] Tokuda, K.: Speech Parameter Generation Algorithm for HMM-Based Speech Synthesis. Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), Vol. 3, 2000, pp. 1314–1318.

[11] Chomphan, S.—Kobayashi, T.: Implementation and Evaluation of an HMM-Based Thai Speech Synthesis System. Proceedings of the 8$^{\text{th}}$ Annual Conference of the International Speech Communication Association, 2007, pp. 2849–2852.

[12] Kim, S. J.—Kim, J. J.—Hahn, M. S.: Implementation and Evaluation of an HMM-Based Korean Speech Synthesis System. IEICE – Transactions on Information and Systems, Vol. E89-D, 2006, No. 3, pp. 1116–1119.

[13] Vesnicer, B.—Mihelic, F.: Evaluation of the Slovenian HMM-Based Speech Synthesis System. Text, Speech and Dialogue. Proceedings of 7th International Conference on Text, Speech and Dialogue (TSD 2004). Lecture Notes in Computer Science, Vol. 3206, 2004, pp. 513–520.

[14] Karabetsos, S.—Tsiakoulis, P.—Chalamandaris, A: HMM-Based Speech Synthesis for the Greek Language. Text, Speech and Dialogue. Proceedings of 11th International Conference on Text, Speech and Dialogue (TSD 2008). Lecture Notes in Computer Science, Vol. 5246, 2008, pp. 349–356.

[15] Zen, H.—Nose, T.—Yamagishi, J.—Sako, S.—Masuko, T.—Black, A. W.—Tokuda, K.: The HMM-Based Speech Synthesis System Version 2.0. Proceedings of the 6th ISCA Workshop on Speech Synthesis (ISCA SSW6 2007), 2007, pp. 294–299.

[16] Taylor, P.—Black, A. W.—Caley, R.: The Architecture of the Festival Speech Synthesis System. Proceedings of the 3th ESCA Workshop in Speech Synthesis 1998, 1998, pp. 147–151.

[17] Schröder, M.—Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. International Journal of Speech Technology, Vol. 6, 2003, No. 4, pp. 365–377.

[18] Open JTalk. `http://open-jtalk.sourceforge.net/`. Accessed 12 December 2014.

[19] Zen, H.—Tokuda, K.—Masuko, T.—Kobayashi, T.—Kitamura, T.: Hidden Semi-Markov Model Based Speech Synthesis. Proceedings of the International Conference on Spoken Language Processing (ICSLP 2004), 2004, pp. 825–834.

[20] Tokuda, K.—Masuko, T.—Miyazaki, N.—Kobayashi, T.: Multi-Space Probability Distribution HMM. IEICE – Transactions on Information and Systems, Vol. E85-D, 2003, No. 3, 2003, pp. 229–232.

[21] Tokuda, K.—Yoshimura, T.—Masuko, T.—Kobayashi, T.—Kitamura, T.: Speech Synthesis Generation Algorithms for HMM-Based Speech Synthesis. Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000), 2000, pp. 1315–1318.

[22] Yoshimura, T.—Tokuda, K.—Masuko, T.—Kobayashi, T.—Kitamura, T.: Mixed Excitation for HMM-Based Speech Synthesis. Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), 2001, pp. 2259–2262.

[23] Erro, D.—Moreno, A.—Bonafonte, A.: Flexible Harmonic/Stochastic Speech Synthesis. Proceedings of the 6th ISCA Workshop on Speech Synthesis (ISCA SSW6 2007), 2007, pp. 194–199.

[24] Maia, R. S.—Toda, T.—Zen, H.—Nankaku, Y.—Tokuda, K.: An Excitation Model for HMM-Based Speech Synthesis Based on Residual Modeling. Proceedings of the 6th ISCA Workshop on Speech Synthesis (ISCA SSW6 2007), 2007, pp. 131–136.

[25] Kawahara, H.: Straight, Exploitation of the Other Aspect of Vocoder: Perceptually Isomorphic Decomposition of Speech Sounds. Acoustical Science and Technology, Vol. 27, 2006, No. 6, pp. 349–353.

[26] Erro, D.—Sainz, I.—Navas, E.—Hernaez, I.: Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. IEEE Journal of Selected Topics in Signal Processing, Vol. 8, 2014, No. 2, pp. 184–194.

[27] Ivanecký, J.—Nábělková, M.: SAMPA Transcription and the Slovak Language. Jazykovedný časopis, Vol. 53, 2002, No. 2, pp. 81–95 (in Slovak with English abstract).

[28] Sulír, M.—Juhár, J.: Design of an Optimal Male and Female Slovak Speech Database for HMM-Based Speech Synthesis. Proceedings of the 7th International Workshop on Multimedia and Signal Processing (Redžúr 2013), 2013, pp. 5–9.

[29] Hládek, D.—Staš, J.: Text Mining and Processing for Corpora Creation in Slovak Language. Journal of Computer Science and Control Systems, Vol. 3, 2010, No. 1, pp. 65–68.

[30] Kominek, J.—Black, A. W.: CMU ARCTIC Databases for Speech Synthesis. Carnegie Mellon University, Pittsburgh, 2003.

[31] Kubichek, R.: Mel-Cepstral Distance Measure for Objective Speech Quality Assessment. Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 1993, pp. 125–128.

[32] Benoi C.—Grice, M.—Hazan, V.: The SUS Test: A Method for the Assessment of Text-to-Speech Synthesis Intelligibility Using Semantically Unpredictable Sentences. Speech Communication, Vol. 18, 1996, No. 4, pp. 381–392.

[33] Sulír, M.—Staš, J.—Juhár, J.: Design of Phonetically Balanced SUS Test for Evaluation of Slovak TTS Systems. Proceedings of the 2014 56th International Symposium Electronics in Marine 2014 (ELMAR), 2014, pp. 35–38.

[34] Ondáš, S.—Juhár, J.—Pleva, M.—Čižmár, A.—Holcer, R.: Service Robot SCORPIO with Robust Speech Interface. International Journal of Advanced Robotic Systems, Vol. 10, 2013, No. 3, pp. 1–11.

[35] Erro, D.—Hernaez, I.—Alonso, A.—García-Lorenzo, D.—Navas, E.—Ye, J.—Arzelus, H.—Jauk, I.—Hy, N. Q.—Magariños, C.—Pérez-Ramón, R.—Sulír, M.—Tian, X.—Wang, X.: Personalized Synthetic Voices for Speaking Impaired: Website and App. Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2015), 2015, pp. 1251–1254.

Martin Sulír received his M.Sc. (Ing.) degree in the field of telecommunications in 2012 at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice. He is currently Ph.D. student at the Department of Electronics and Multimedia Communications at the Technical University of Košice. His research interests include text to speech synthesis systems.

Milan Rusko graduated from the Slovak Technical University, Bratislava in 1994, he received his Ph.D. degree from the Slovak Technical University, Košice in 2013. Since 1993 he has been the Head of the Department of Speech Analysis and Synthesis at the Institute of Informatics of the Slovak Academy of Sciences. His research interests include speech acoustics, speech corpora, digital speech and audio processing, speech recognition and speech synthesis. Responsible leader of several national and international projects. Head of the section "Physiological, Psychological and Musical Acoustics" in the Slovak Acoustical Society. Management implementation and application of scientific results into practice.

Jozef Juhár graduated from the Technical University of Košice in 1980. He received his Ph.D. degree in radioelectronics from the Technical University of Košice in 1991, where he works as Full Professor at the Department of Electronics and Multimedia Communications. He is author and co-author of more than 200 scientific papers. His research interests include digital speech and audio processing, speech/speaker identification, speech synthesis, development in spoken dialogue and speech recognition systems in telecommunication networks.