Journal of Computing and Information Technology - CIT 14, 2006, 4, 307–313
doi:10.2498/cit.2006.04.06

Croatian HMM-based Speech Synthesis

S. Martincic-Ipsic and I. Ipsic

Department of Informatics, Faculty of Philosophy, University of Rijeka, Croatia

The paper describes the development of a trainable speech synthesis system, based on hidden Markov models. An approach to speech signal generation using a source-filter model is presented. Inputs into the synthesis system are speech utterances and their phone level transcriptions. A method using context-dependent acoustic models and Croatian phonetic rules for speech synthesis is proposed. Croatian HMM-based speech synthesis experiments are presented and generated speech results are discussed.

Keywords: corpus-based speech synthesis, hidden Markov models, context-dependent acoustic modelling, Croatian speech corpora.

1. Introduction

One of the challenges put in front of researchers is to build natural conversational interfaces. When we speak about speech-based interfaces for computer system, we refer to two basic technologies: speech recognition and speech synthesis. The goal of synthesis systems is to provide the spoken output to the users by generating speech from text. Speech synthesis is used in: spoken dialog systems, applications for blind and visually-impaired persons, applications in telecommunication, eyes and hands free applications.

Speech synthesis methods can be grouped into three categories: articulatory synthesis, formant synthesis and concatenative synthesis. Articulatory synthesis is based on physical models of the human speech production system. Main reasons why articulatory synthesis has not led to quality speech synthesis is the lack of knowledge of the complex human articulation organs [5].

Formant speech synthesis is based on the rules which describe the resonant frequencies of the vocal tract. The formant method uses the source-filter model of speech production, where speech is modelled by parameters of the filter model. Rule-based formant synthesis can produce quality speech which sounds unnatural, since it is difficult to estimate the vocal tract model and source parameters [5].

More natural sound speech can be produced using concatenative synthesis. These techniques use stored basic speech units (segments), which are concatenated to the word sequences according to a pronunciation dictionary [4]. Special signal processing techniques to smooth the unit transitions and to model the intonation are used. Such systems can produce quality speech which often lacks naturalness, since the concatenation methods cannot efficiently model the prosodic characteristics of speech.

Methods which are able to produce more natural speech are a generalization of the concatenative synthesis based on dynamic selection of basic speech units from a large speech corpus. This method is also known as the corpus synthesis [4].

These methods consist of mainly two parts: procedures for selection and training of basic synthesis units and the part for the synthesis, where the phonetic and prosodic information is used for speech signal generation. One of the most promising methods is the use of context-dependent phone models, which are modelled with hidden Markov model (HMM) [14].

Recently we can find trainable synthesis systems for Japanese [14], English [1, 6, 13] and for a few other languages [3, 16]. For Croatian speech synthesis, so far only experiments using diphone concatenation synthesis have been reported [2].

The paper is organised as follows: in Section two we present the system architecture, the third part describes the source filter model of speech production, the fourth part describes the HMM methodology used for speech signal features generation and finally the Croatian speech corpus used for the experiments is described. The seventh part is dedicated to speech synthesis experiments. At the end, we conclude with Discussion and Conclusion.

2. Trainable Speech Synthesis

In speech recognition the goal is to find the spoken words in the speech signal. From the feature vectors using the Viterbi algorithm the most probable path through HMMs is used in order to find the spoken words [11].

In speech synthesis, the same procedure is used to find the most probable path through HMMs that can generate the speech signal feature vectors. The speech signal can then be synthesized from so generated feature vectors. The generated feature vectors from HMMs describe mel-cepstrum, pitch and duration of context-dependent phones.

Figure 1 gives the overview of the HMM-based speech synthesis systems. The inputs to the system training are speech utterances and their phone level transcriptions.

The training part includes spectral parameters and excitation parameters extraction. The feature vectors of extracted mel-cepstrum and fundamental frequency F_0 parameters, together with their dynamic features, are concatenated and used for HMMs training of context-independent and context-dependent acoustic models. The training of phone HMMs using pitch and mel-cepstrum simultaneously is enabled in a unified framework by using multi-space probability distribution HMMs and multi-dimensional Gaussian distributions [12]. The simultaneous modeling of pitch and spectrum resulted in the set of context-dependent HMMs. Context-dependent clustering of Gaussian distributions was performed independently for spectrum, pitch and duration because of the different clustering factor influence.

In the synthesis part, from the set of context-dependent HMMs according to the symbols in the entry text, the speech parameters are generated. The generated excitation parameters and mel-cepstrum parameters are used to generate the speech signal using the source-filter model.

The advantage of this approach is in capturing the acoustical features of context-dependent phones using the speech corpora. Synthesized voiced characteristics can also be changed easily by altering the HMM parameters and the system can be easily ported to a new language.

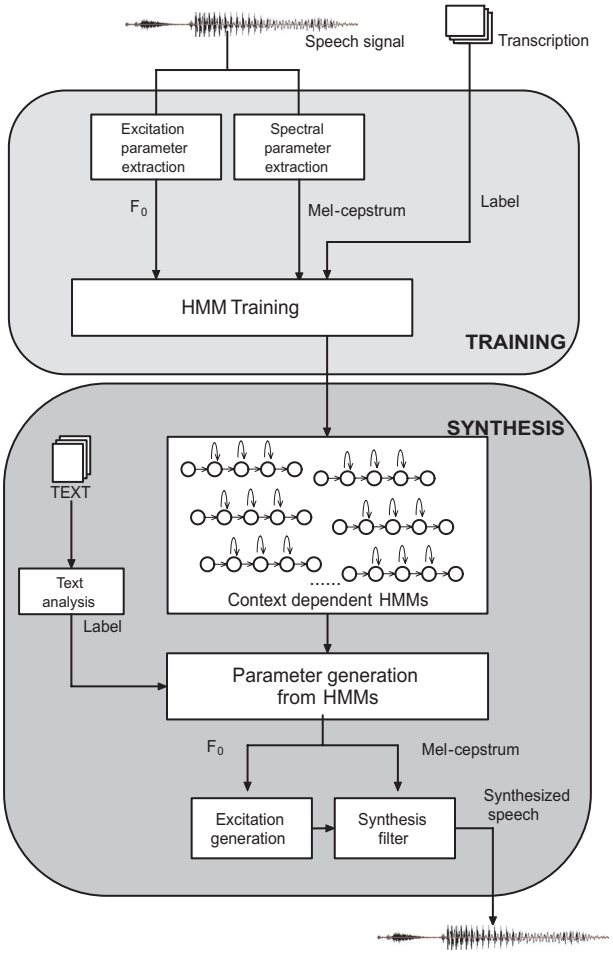

Fig. 1. Overview of the HMM speech synthesis.

3. The Source-Filter Speech Model

The source-filter model of human speech production is the basis of many speech synthesis approaches. Speech can be viewed as the output of a linear filter excited by a sound source. Typically, the sound source has a voiced sound component and an unvoiced sound component.

The filter simulates the frequency response of the vocal tract and shapes the spectrum of the signal generated by the source.

Figure 2 shows a source filter speech synthesis model where an impulse train for voiced sounds or white noise for unvoiced sounds are used as input to a time varying filter. For voiced fricatives mixed model is used by combining impulse train and white noise at the same time.

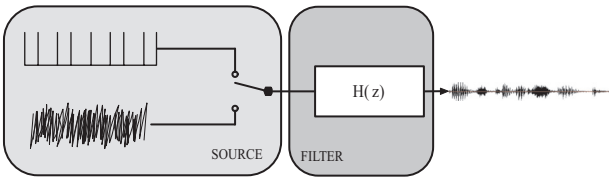

Fig. 2. The source-filter model.

The z-transform of the speech signal $S(z)$ can be modelled using

$$S(z) = U(z)\, H(z) \tag{1}$$

where $U(z)$ is the excitation model and $H(z)$ the transfer function of the filter model representing the vocal tract response.

Linear predictive coding (LPC) source-filter model is widely used because it is a fast, simple and effective way of estimating the main parameters of the speech signal. LPC model predicts the current speech sample from the linear combination of its past p samples

$$H(z) = \frac{X(z)}{E(z)} = \left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)^{-1} = \frac{1}{A(z)} \tag{2}$$

where p is the order of LPC analysis.

Another source-filter model approach is the Mel Log Spectral Approximation (MLSA) filter based on mel-frequency cepstral coefficients (MFCC) [7]

$$H(z) = \exp \sum_{m=0}^{M} c(m) \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{3}$$

where $c(m)$ are the mel-cepstral coefficients and α is the frequency compression parameter, which is used to compress mel-scale in order to approximate the human sensitivity to the frequencies of the speech signal.

4. Speech Signal Features Generation Using HMMs

A hidden Markov model $\lambda(A, B, \pi)$ is defined by its parameters: A – state transition probability, B – output probability and π – initial state probability.

Let us have the HMM λ that contains concatenated elementary triphone or monophone HMMs that correspond to the symbols in the word w, which has to be synthesized.

The aim of the speech synthesis is to find the most probable sequence of states features vectors $\hat{x}$ from the HMM λ. Figure 3 shows the model in state q_i at time t_i.

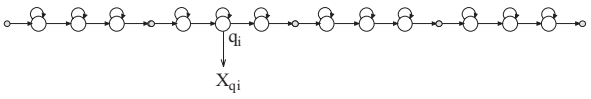

Fig. 3. Concatenated HMM chain.

X_{qi} is the M-dimensional generated feature vector at the state q_i of the model λ

$$X_{q_i} = (x_1^{(q_i)}, x_2^{(q_i)}, .., x_M^{(q_i)})^T \tag{4}$$

From model λ we want to generate a sequence of features vectors $\hat{x} = x_{q_1} x_{q_2} .. x_{q_L}$ of length L maximizing the overall likelihood $P(x\,|\lambda)$ of a HMM

$$\begin{aligned} \hat{x} &= \arg\max_x \{P(x|\lambda)\} \\ &= \arg\max_x \left\{ \sum_Q P(x|q,\lambda) P(q|\lambda) \right\}, \end{aligned} \tag{5}$$

where the $Q = q_1, q_2, .., q_L$ is the path through the states of the model λ. The overall likelihood of the model $P(x\,|\lambda)$ is computed by adding the product of joint output probability $P(x\,|q, \lambda)$ and state sequence probability $P(q\,|\lambda)$ over all possible paths Q [11].

Practically, we use Viterbi approximation of (5), because, theoretically, we have to search for all possible paths through the model which is too time consuming

$$\hat{x} \approx \arg\max_x \{P(x|q,\lambda,L) P(q|\lambda,L)\} \tag{6}$$

The state sequence $\hat{q}$ of the model λ can be maximized independently of $\hat{x}$

$$\hat{q} = \arg\max_{q} \{P(q|\lambda, L)\} \tag{7}$$

Let's assume that the output probability distribution of each state q_i is one Gaussian density function with a mean vector μ_i and covariance matrix Σ_i. The HMM model λ is a set of all means and covariance matrices for all N states:

$$\lambda = (\mu_1, \Sigma_1, \mu_2, \Sigma_2, \ldots, \mu_N, \Sigma_N). \tag{8}$$

Then the log-likelihood of (6) is given by

$$\ln P(x|q,\lambda) = -\frac{LM}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{L}\ln\left|\Sigma_{q_t}\right| - \frac{1}{2}\sum_{t=1}^{T}(x_t - \mu_{q_t})^T\Sigma_{q_t}^{-1}(x_t - \mu_{q_t}) \tag{9}$$

Maximizing x in (9) leads to the trivial solution $\hat{x} = (\mu_{q_1}\mu_{q_2}, \ldots, \mu_{q_L})$, where the sequence is equal to the means of the corresponding states. Such a solution does not represent the speech well because of the discontinuities at the state boundary. This can be solved by extending the feature vectors with first and second differentials

$$X_{q_i} = ((x_{q_i})^T, (\Delta x_{q_i})^T, (\Delta^2 x_{q_i})^T)^T. \tag{10}$$

In [8] a fast algorithm is given for the solution of equation (9).

5. Croatian Speech Corpus

The speech corpus is the essential part of all spoken technologies systems. The quality and volume of speech data in the corpus directly influences the performance of the system. Enough speech data is essential in all statistical approaches to speech modelling such as HMMs in order to estimate all the parameters of the models. The training of the HMM models for speech synthesis is based on speech data utterances and their transcriptions. Croatian speech corpora used for speech synthesis training contains speech of one professional speaker of the national radio. The speech was recorded using a PC with an additional Hauppauge WINTV/radio card.

The radio-broadcasted speech signals were sampled with 16 KHz and stored in a 16-bit PCM encoded waveform format. Mel-cepstrum and fundamental frequency F_0 was calculated for each utterance using the SPTK tool [17].

The speech is organized in 1111 spoken utterances lasting 85 minutes and containing 12265 words where 3840 are different. Language perplexity of the bigram model is 23.59. The phonetic dictionary contains accented words. Our system differentiates vowels with (marked by a *:*) and without accent, including the occurrence of *r* as a vowel. The number of seen cross-word triphones in the speech corpora is 8290 which is about 14% of the number of all possible triphones (60521).

The speech utterances are transcribed on the word level, so before training we had to perform initial phone level segmentation. Phone level segmentation was achieved using automatic alignment of speech signals and word transcriptions, based on monophone HMMs [9]. Automatic segmentation is performed by using the forced alignment of the spoken utterance and the corresponding transcription by using the monophone speech recognizer.

The forced alignment assumes that all phonemes in the utterance initially have equal segmentation distribution. The monophone recognizer was trained on the same data that was automatically segmented. The monophone HMMs contain 5 states. Ten iterations of Baum-Welch training were used to estimate the monophone models parameters. Finally the number of mixtures output Gaussian probability density functions per state was increased to 20. Further the Viterbi algorithm was used to find the most likely sequence of HMM states. The results of the Viterbi algorithm are automatically segmented monophones, which are used as the input for the speech synthesis training procedures.

The overall phone correctness of the used 20 mix monophone recognizer is 78.62%. Figure 4 shows the distribution of all phonemes in the speech database and correctness of automatic phone-level segmentation. Phoneme correctness was determined as a percentage of correctly determined phones, taking into account substitution and deletion errors [9].

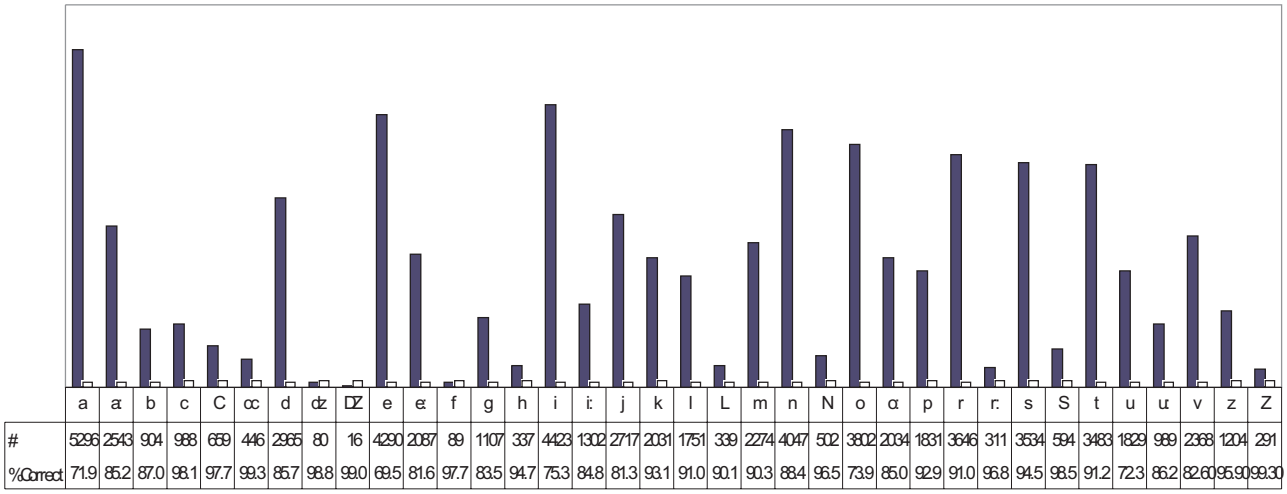

Fig. 4. Croatian phonemes distribution in the speech database.

Additionally, the automatically segmented phones were inspected and corrected. Human experts audio-visually adjusted automatically segmented phones using the signal spectrogram and signal transcription. Phone boundaries were specially adjusted for phonemes with low-occurrence.

6. Speech Synthesis Experiments

Training of the models was performed using HTS (HMM-based Speech Synthesis System) [18] which is an extension to the HTK Toolkit [15].

The speech signals were windowed using a 25 ms Blackman window and 5 ms frame shift. The feature vector consisted of spectral and excitation (pitch) parameters. The spectral feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient and its delta and acceleration coefficients. The pitch feature vector consisted of log F_0 and its dynamic parameters (delta and acceleration). The HMMs were embedded-trained on the features vectors consisting of spectrum, pitch and their dynamic features.

We used 7 states left to right HMMs with no skip, where the first and the last states were no-emitting states. First we trained 41 monophone models, 36 for the Croatian phone set including accented and not accented vowels and 5 for special events like breathing, pause, noise etc.

The triphone models were made out of monophone models and trained. Then the state tying procedure based on 216 Croatian phonetic rules was performed, and tied triphone models were reestimated [10]. The phonetic rules describe the class of the alophones according to their articulatory and acoustic characteristics.

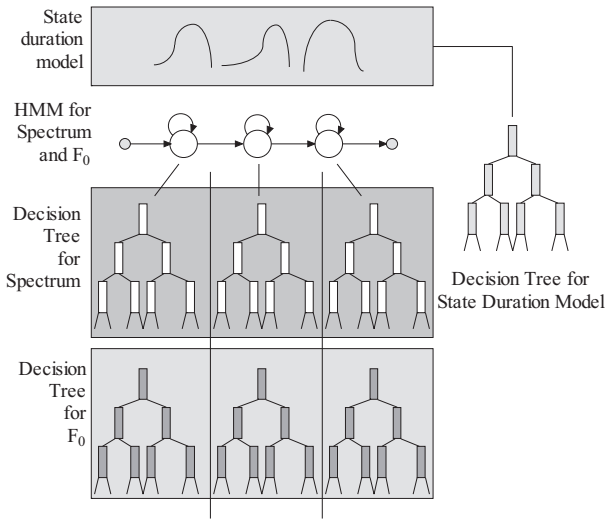

Fig. 5. Decision trees for spectrum, pitch and duration.

According to the state clustering of the models the F_0 and duration clustering trees were built as shown in Figure 5. The clustering trees were built separately because different context clustering factors are relevant for spectral part clustering, pitch clustering and duration clustering.

State duration densities were modeled by multivariate Gaussian distribution. The dimensionality of state duration density is equal to the number of states of corresponding HMM. The last step in the training part is parameter generation for unseen triphones according to the Croatian phonetic decision trees.

The synthesis part used prepared context-dependent HMMs, and state duration and pitch trees for generating the sequence of feature vectors for the test text. According to the phoneme sequence in text labels the context-dependent HMMs were concatenated. State durations of the sentence are determined by maximizing the likelihood of state duration densities. According to the obtained state the sequence of mel-cepstral coefficients and $\log F_0$ values including voiced/unvoiced decisions are determined by maximizing the output probability of HMM [12]. Finally, the speech is synthesized from generated mel-cepstral feature vectors and pitch values using the MLSA filter.

7. Speech Synthesis Results

The text-to-speech test included 41 Croatian sentences. The text labels were transformed into triphone format. For each sentence the speech in raw format, pitch and duration were generated. Figure 6 presents the result of generated speech for the sentences:

"*Vjetar u unutrašnjosti većinom slab, na Jadranu umjerena i jaka bura. <uzdah> Najviša dnevna temperatura od minus jedan do plus tri stupnja na Jadranu od deset do petnaest.*"

From the top the pitch, spectrogram and raw signal are shown.

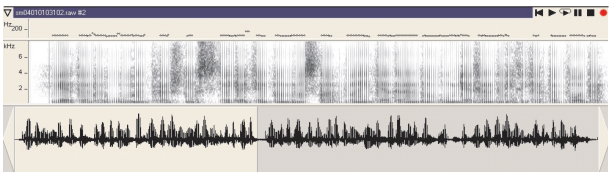

Fig. 6. Pitch and spectrogram of generated speech signal for utterance sm04010103102.

Figure 7 shows the pitch and spectrogram of the corresponding part of original signal.

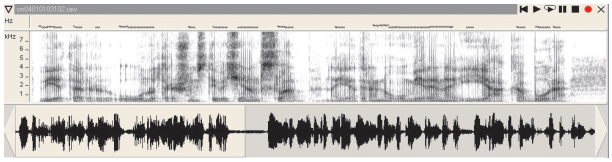

Fig. 7. Pitch and spectrogram of original signal for utterance sm04010103102.

8. Discussion and Conclusion

In this paper we presented the HMM-based Croatian speech synthesis system. The text-to-speech system was trained on 85 minutes of Croatian speech organized in 1111 spoken utterances. The used HMM approach is very effective in rapid development of the TTS system for new language. Although the quality of generated speech is "vocoded" buzzy speech, it can be understood. This speech synthesis will be incorporated in Croatian weather information spoken dialog system.

Further work on evaluation of intelligibility and naturalness of synthetic speech will be done. The human experts and users will evaluate the system. The rate for intelligibility, overall quality, naturalness and functionality will be collected. In order to improve the context-dependent phone models used for synthesis, more Croatian speech material will be recorded and annotated.

References

[1] A. ACERO, Formant Analysis and Synthesis Using Hidden Markov Models. *EUROSPEECH*, Budapest, Hungary, p.p. 1047-1050. 1999.

[2] J. BAKRAN, N. LAZIC, Fonetski problemi difonske sinteze hrvatskog govora. *Govor XV*,1998., Vol. 2, pp. 103–116.

[3] M. J. BARROS, ET AL., HMM-based European Portuguese TTS System, *INTERSPEECH '05*, Lisbon, Portugal,205., pp. 2581–2584.

[4] N. CAMPBELL, AND A. BLACK, Prosody and the selection of source units for concatenative synthesis, *Progress in Speech Synthesis*, ed. van Santen, J. Sproat, R., Olive, J., Hirsberg J., Springer, New York, 1997., pp. 663–666.

[5] D. H. KLATT, Review of Text to Speech Conversion for English, *Journal of the Acoustic Society of America*, 1987., Vol. 82, pp. 737–793.

[6] R. E. DONOVAN, P. C. WOODLAND, A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, 1999., Vol. 00, pp. 1–19.

[7] T. FUKADA, K. TOKUDA, T. KOBAYASHI, AND S. IMAI, An adaptive algorithm for melcepstral analysis of speech, *Proc. of ICASSP*, 1992., Vol. 1, pp. 137–140.

[8] K. TOKUDA, ET AL., An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. *EUROSPEECH 95*, 1995., 1: pp. 757–760.

[9] S. MARTINCIC-IPSIC, I. IPSIC, Recognition of Croatian Broadcast Speech. L. Budin (ed.), S. Ribarić (ed.).*XXVII. MIPRO 2004*, Opatija, 2004., Vol. CTS + CIS, pp. 111–114.

[10] S. MARTINCIC-IPSIC, I. IPSIC, Croatian Telephone Speech Recognition. L. Budin (ed.), S. Ribaric (ed.). *To appear in XXIX. MIPRO 2006*, Opatija, 2006., Vol. CTS + CIS.

[11] L. R. RABINER, A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 1989., Vol. 77, No. 2, pp. 257–286.

[12] K. TOKUDA, ET AL., Multi-Space Probability Distribution HMM. *IEICE Trans. Inf. & System.*, Vol. E85-D, No. 3, March, 2003.

[13] K. TOKUDA, H. ZEN, A. BLACK, An HMM-Based Speech Synthesis System Applied to English. *IEEE TTS Workshop 2002*. Santa Monica. California, USA. 2002.

[14] K. TOKUDA, ET AL., Speech Parameter Generation Algorithm for HMM-Based Speech Synthesis. HMM, *Proc. ICASSP.*, 2000., Vol. 3. pp. 1314–1318.

[15] S. YOUNG, ET AL., *The HTK Book* (for HTK Version 3.2). Cambridge University Engineering Department, Cambridge, Great Britain. 2002.

[16] B. VESNICER, F. MIHELIC, Sinteza slovenskega govora z uporabo prikritih Markovovih modelov. *Elektrotehniski vestnik*, 2004., Vol. 71, No. 4, pp. 223–228.

[17] Department of Computer Science, Nagoya Institute of Technology, *Speech Signal Processing Toolkit*, SPTK 3.0. Reference manual, `http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/`, Japan, December, 2003. [09.2005.].

[18] Department of Computer Science, Nagoya Institute of Technology, *HTS HMM Based Speech Synthesis System 1.0.* `http://hts.ics.nitech.ac.jp/`, Japan, 2004. [09.2005.]

Received: June, 2006
Accepted: September, 2006

Contact addresses:
Sandra Martincic-Ipsic
Department of Informatics
Faculty of Philosophy
University of Rijeka
Omladinska 14, 51000 Rijeka
Croatia
Phone: (385) 51–345 046 Fax: (385) 51345 207
e-mail:`smarti@ffri.hr`

Ivo Ipsic
Department of Informatics
Faculty of Philosophy
University of Rijeka
Omladinska 14, 51000 Rijeka
Croatia
Phone: (385) 51–345 046 Fax: (385) 51345 207
e-mail:`ivoi@ffri.hr`

SANDA MARTINCIC-IPSIC was born in 1970 in Rijeka, Croatia. She received B.Sc. degree in computer science from the Faculty of Computer Science and Informatics, University of Ljubljana in 1995 and M. Sc. degree in Informatics from the Faculty of Economy, University of Ljubljana in 1999. Currently she works as an assistant at the Department of Informatics, Faculty of Philosophy, University of Rijeka. In 2002 she started a PhD postgraduate study at the Faculty of Electrical Engineering and Computing, University of Zagreb. Her research interests are in speech recognition, speech synthesis, speech corpora development and spoken dialog systems, with special focus on Croatian language.

IVO IPSIC was born in 1963. He received his B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana, Slovenia in 1988, 1991 and 1996, respectively. From 1988–1998 he was a staff member of the Laboratory for Artificial Perception at the Faculty of Electrical Engineering, University of Ljubljana. In the academic year 93/94 Ivo Ipsic was a guest researcher at the Friedrich Alexander University of Erlangen-Nürnberg, Germany, at the Department of Computer Science. His work concerned acoustic modelling for continuous speech recognition. Since 1998 Ivo Ipsic has been an assistant professor of computer science at the University of Rijeka, Croatia. His current research interests belong to the field of pattern recognition, digital signal processing and artificial intelligence. Ivo Ipsic is author of more than 50 papers presented at international conferences or published in international journals.