Mirjam Sepesy MAUČEC, Zdravko KAČIČ and Andrej ŽGANK

Faculty for Electrical Engineering and Computer Science, University of Maribor

Speech recognition for interaction with a robot in noisy environment

Abstract. *One of the main problems with speech recognition for robots is noise. In this paper we propose two methods to enhance the robustness of continuous speech recognition in noisy environment. We show that the accuracy of recognition can be improved by better weighting the language model in the decision process. The second proposed method is based on language model adaptation. The experiments showed that both proposed techniques improve speech recognition accuracy by approximately 2% .*

Streszczenie. *W artykule przedstawiono dwie metody zwiększenia odporności na zakłócenia i skuteczności rozpoznawania mowy w zaszumionym otoczeniu. Wykazano, że odpowiednie dobranie współczynników wagowych w procesie decyzyjnym dla modelu języka zwiększa precyzję rozpoznawania dźwięków. Druga metoda opiera się na adaptacji modelu języka. Badania eksperymentalne wykazały, że obydwie metody zwiększają skuteczność rozpoznania mowy o około 2%. (**Rozpoznawanie mowy w interakcji z robotem w zaszumionym otoczeniu**).*

Keywords: speech recognition, large vocabulary continuous speech, noise robustness.
Słowa kluczowe: rozpoznawanie mowy, bogate słownictwo, mowa ciągła, odporność na zakłócenia.

Introduction

In the near future a robot should have capability of social interaction with people. Much of efforts in humanoid robot technology have been focused on robot locomotion aiming at safe walk and behaviours. Research affords were mainly devoted to robot vision. Since we are talking about humanoid robots, natural interaction between a human and a robot is also expected, which includes also the auditory function of robot [1, 2, 3]. Little have been done in technological development of hearing function of robot. Human-robot interaction through voice channel is one of key perceptive functions of a robot. Communication in natural language rather than in an artificial programming language is an advantage, which enables the communication with the novice users without a proper training and overcome many difficulties with localization.

Effective communication with a mobile robot using speech is a difficult problem. Information for speech recognition engine comes from the microphone array, which is exposed to noises in real environment. A lot of sound sources harm the speech that should be reliable recognized. Motor noises are inevitably generated. While a robot is moving and performing gestures it makes noise itself and is expected to recognize human speech. Noise is captured with strong power by the robot's microphones, because the noise sources are closer to the microphones than the speech source. Ambient noise caused by room acoustics usually additionally degrades quality and reliability of speech recognition.

The aim of this paper is to analyze noise robustness of our speech recognition system and to propose approaches to enhance the robustness. The paper is organized as follows. Next section reviews the literature. We describe large vocabulary continuous speech recognition in noisy environment. In this section the theoretical background of our idea is given. A detailed description of developed recognition system follows. We give the specifications of text and speech databases. The experiments in clean and noisy environment are described. Last section concludes the paper.

Related work

One of the main problems with automatic speech recognition (ASR) for robots is noise. To cope with such a noisy speech signal [4], noise adaptation techniques such as multi-condition training [5] and Maximum-Likelihood Linear Regression (MLLR) [6, 7] are commonly used. One of the big differences between robot noise and environmental noise it that robot noise can easily be estimated in advance. Each kind of robot motion or gesture performs almost the same noise every time it is performed. The noise can be recorded. Having recorded noise, many techniques exist to subtract the noise from the input signal. In [8] a method is proposed, that is based on three techniques, multi-condition training, maximum-likelihood linear regression, and missing feature theory.

In [9] the robot audition system is described that recognizes speech that is contaminated by simultaneous speech. The system is based on two key ideas, pre-processing of ASR and missing feature-theory based integration of pre-processing and ASR. Theoretically, more than three speakers can be supported, but the performance becomes worse. In case of three speakers it was around 60%. When assuming that the acoustic environment does not change much, ASR with multi-condition trained acoustic model tends to work well. As already mentioned, robots interact with multiple people under dynamically changing environments. Therefore, ASR should work with a single acoustic model by adapting it to a current environment. In [10] missing feature theory is adopted in a system that consist of geometric source separation, post-filtering, computation of missing feature mask and missing feature theory based ASR.

In our research we analyze noise robustness from the perspectives of acoustic and language models.

Speech recognition in noise environment

Speech recognition process consists of two main modules, speech pre-processing and speech decoding. First, input speech is processed by voice activity detection module. It determines the presence or absence of speech in input signal. If it fails in noisy environment it is difficult to recognize input speech accurately. After a speech segment is detected, the speech signal is analyzed to extract the useful information for speech recognition. For noisy speech recognition mel-scale frequency cepstrum coefficients (MFCC) and perceptual linear prediction method (PLP) were proved to be effective. Analyzed speech is than recognized in decoding process. A search for most probable word sequence is carried out using the information from acoustic and language models. Acoustic models are mostly trained on data, recorded in clean condition. Consecutively there is a mismatch between models and input features that come from noisy environment.

Language model assigns a probability to a sequence of words. The assigned probability is calculated based on counting short sequences in training data being in most cases clean, literal text. Speech in noisy environment differs to a great extent from the literal text. In addition to lexical words it contains hesitations and disfluencies that carry no linguistic information and harm the prediction power of language model.

Many efforts have been made in developing methods for noise robust voice activity detection and voice robust speech analysis [11]. Much less research was devoted to noise robust decoding process. Two types of approaches for decoding in noisy environment are analyzed in this paper. The first approach is based on noisy channel model. The second approach introduces adapted language model.

The noisy channel model is a framework used by search algorithm to find the intended word sequence given the acoustic evidence of input speech. It picks the most likely word sequence given the observed acoustic evidence.

$$\hat{W} = arg\ \max_{w} P(W|A) \tag{1}$$

$P(W|A)$ denotes the probability that words were spoken, given that the acoustic evidence was observed. The well known Bayes' formula allows us to rewrite the probability to take into account the acoustic and linguistic knowledge:

$$P(W|A) = \frac{P(A|W) \cdot P(W)}{P(A)} \tag{2}$$

Since the maximization is carried out with the variable A fixed, we can disregard the dominator and use the simple product of acoustic probability $P(A|W)$ and language model probability $P(W)$. In real systems, it is widely known that balancing between acoustic probability and language probability is needed to optimize the system performance. The typical form of combining the two knowledge sources is:

$$\hat{W} = arg\ \max_{w} \left(log\ P(A|W) + \alpha\ log\ P(W) - nQ \right) \tag{3}$$

where α is known as language model weight, Q is word insertion penalty and n is the number of words in the sequence W.

Both parameters are constants, and optimal values for them are not known. Obviously they depend on utilized acoustic and language models. We claim that there are also some dependence between those two parameters and the speech environment. It is natural to think that we need to somehow compensate for noisy input speech, i.e. noisy acoustic evidence. We see the probability estimate given by language model as more reliable estimate than the probability given by acoustic model in noisy environment. In the experiments we will confirm the correctness of the idea by systematic optimization of language model weight.

Language model assigns the probability to a sequence of words using the chain rule. In case of a trigram model it is:

$$P(W) = \prod_{i=1}^{n} P(w_i | w_{i-2}, w_{i-1}) \tag{4}$$

Basic trigram probabilities are estimated by trigram frequencies, calculated in very large corpora of training text. Since many possible trigrams are never actually encountered even in very large corpora, some smoothing method is used. Using this approach we get the model of general language.

It is quite common that speech recognition system is used in an environment, which can be characterized by language in use. In communication with a robot we can observe that the language is even more limited (for example to instructions given to it). Having characteristic text from the target environment, the model of general language can be adapted to the target language by interpolation:

$$P(W) = \lambda P_G(W) + (1 - \lambda) P_T(W); \tag{5}$$

where $P_G(W)$ is a probability estimate given by the model of general language, and $P_T(W)$ is a probability estimate given by the model of target language, and λ is the interpolation coefficient. To build the model of target language, characteristic text should be of a reasonable size. It can be collected in a WoZ experiment [12]. In this paper some additional experiments will be devoted to adapted language models. The relation between adapted language model and language model weight in noisy environment will be analyzed.

Speech recognition system

The speech recognition system used in these experiments is based on Hidden Markov Models (HMM) for acoustic modelling and statistical n-grams for language modelling. Such system belongs to the more complex ones in the area of speech recognition [13]. The basic description of speech recognition system is given in this section, for further details see [14].

The role of feature extraction is to convert the input speech signal into sequence of feature vectors, which can be used for speech decoding. The defined window size was 25 ms with 10 ms time shift. Mel-scale frequency cepstrum coefficients were used as feature extraction method, as they prove good performance in case of noisy environment. Energy was added to the basic set of 12 mel-cepstral coefficients, resulting in 13 features. Speech signal contains short-term variations [15], which are modelled by delta and delta-delta features, calculated over several adjacent frames. The final size of feature vector was 39. The robustness of feature extraction procedure to different channel conditions was improved with cepstral mean normalization.

The speech recognition usage scenario usually also determines the type of acoustic models. The speaker independent, context-dependent grapheme based acoustic models were best suitable for our task. Three state left-right hidden Markov models topology with weighted sum of continuous Gaussian probability density functions (PDF) was selected for acoustic models. The speech database transcriptions, which are needed for acoustic modeling [16], are created in a manual way. Therefore we applied acoustic modeling training procedure in three steps, consecutively improving the quality of transcriptions with forced-realignment method.

The initial parameters of monophone acoustic models were estimated as global values, calculated on randomized subset of training data. The initialized acoustic models were trained on baseline speech transcriptions using the Baum-Welch re-estimation in a stepwise manner [17]. The monophone acoustic models with 1 mixture of Gaussian PDF were then used for forced-realigning the speech databases transcriptions.

The transcriptions resulting from forced-realigning were then used for training the acoustic models, with stepwise increasing the number of Gaussian mixtures to 32. The

second set of acoustic models was again used for forced-realigning procedure, which resulted in improved speech transcriptions.

A new set of monophone acoustic models was initialized in the third step, using the improved transcriptions from the second step. The initialization was performed with locale values for each monophone acoustic model [18]. The acoustic models were then trained with Baum-Welch re-estimation. The cross word context-dependent acoustic models (triphones) were introduced to the set, to successfully model the effect of coarticulation in continuous speech. The main weakness of triphone acoustic models is the large number of free models' parameters, which must be estimated during the acoustic training procedure. Phonetic decision tree based clustering was applied to triphone acoustic models to control the number of free parameters in the proportion with the amount of available spoken training material. The decision trees were induced on phonetic broad classes, generated in a data-driven way. The number of Gaussian PDFs' per state was then again increased in a stepwise manner. The final triphone acoustic models used for evaluation had 16 Gaussian PDF per state and were speaker and gender independent.

Vocabulary contained the 64,000 words. All words from BNSI training set were included. Most frequent words from corpus FidaPLUS [19] were added.

Baseline language model was back-off model based on trigrams. Singletons bigrams and trigrams were excluded. N-grams with frequency, greater than 7, got maximum likelihood estimates. N-grams with lower counts were discounted under Good-Turing. Final baseline model contained 17M bigrams and 33M trigrams. Its perplexity on test set was 177.

Baseline language model was adapted to the target domain using the linear interpolation given in equation (5). Interpolation weight was set to 95%. The perplexity of adapted language model was 143.

Decoder is the essential module of an automatic speech recognition system. It needs three different data sources for its operation: acoustic models, language model and lexicon. The decoder included in the experiments was a dynamic one-pass Viterbi decoder with enabled beam pruning and limited number of active models to control the speech recognition speed.

The lexicon included in the decoder had 64,000 different words. The base unit in the lexicon was grapheme, which makes additional grapheme-to-phoneme conversion obsolete. The out-of-vocabulary rate for the test set was 3.12%, which is comparable with other highly inflectional languages.

Text and speech databases

The experiments presented in this paper were carried out for Slovenian language, but all proposed methods are language independent, and can be used for any other language with available language resources. Text database was used to train language models. It contained four components. The first component was BNSI speech training corpus. Its size was 573k words. The second component was BNSI-text. It is the collection of TV scenarios. The size of this component was 11M words. Both corpora contained samples of spoken language. Third and fourth component contained samples of written language. The third component was corpus Večer, which contained newspaper articles. Its size was 95M words. The fourth component was Slovenian national corpus FidaPLUS [19]. It is the largest database, containing 621M words.

Speech database, with large amount of annotated and transcribed spoken material, presents an important aspect of an automatic speech recognition system. The Slovenian BNSI Broadcast News speech database [20] was used in the experiments. The BNSI speech database comprises speech in several different acoustic conditions, which are presented in Table 1.

Table 1. Ratio of various acoustic conditions in the BNSI speech database.

Acoustic condition in database		**Ratio(%)**
F0	studio/read	36.56
F1	studio/spontaneous	16.23
F2	telephone	1.65
F3	music	6.02
F4	background	37.63
F5	nonnative	0.05
FX	other	1.86

Such variety of acoustic conditions in the BNSI database is appropriate for the acoustic modeling training task as it guarantees the channel robustness of acoustic models. The complete BNSI database consists of 36 hours of spoken material; 30 hours are used for acoustic training, 3 hours are devoted to as development set, and 3 hours are standard evaluation set. There are 1565 different speaker, 1069 of them male and 477 female. The gender of remaining 19 speakers was unknown.

One of the key factors influencing the robustness of speech recognition is noise. Four dedicated test sets with 633 sentences each, were created (see Table 2), to be able to completely control the influence of noise on speech recognition accuracy.

Table 2. Signal-to-noise ratio for different test sets.

Test set	**SNR(dB)**
Baseline	47.52
Low noise	39.45
Medium noise	34.22
High noise	28.80

The first test set (Baseline) contains the original continuous speech in clean acoustic environment. The signal-to-noise ratio (SNR) for this test case was 47.52 dB. White noise with different level of energy was added to the baseline test set recorded in clean acoustic environment. As a result, three additional test sets with various noise levels (low, medium, high) were created. In such a way the SNR of test sets was reduced, simulating usage scenarios in adverse acoustic conditions.

Experimental results

The analysis of the speech recognition system for noisy environment was carried out in several steps. The speech recognition results are presented as word accuracy, which is:

$$Accuracy(\%) = \frac{H - I}{N} \cdot 100 \tag{6}$$

where H denotes the number of correct words in the recognized set, I is the number of insertions and N denotes the number of all words in the test set.

First, the optimal language model weight for the baseline scenario has to be evaluated. As test set, continuous speech in clean acoustic condition with baseline interpolated trigram language model was used in this scope. The results of this analysis are shown in Figure 1.

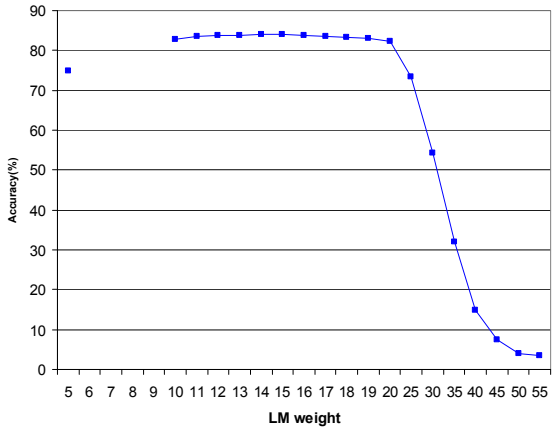

Fig.1. Speech recognition accuracy for baseline test set.

The best speech recognition accuracy of 84.12% was achieved with the language model weight 15. The speech recognition accuracy is near optimal between language model weights 10 and 20. When the weight is increased beyond 20, the accuracy drastically reduces. The achieved results for clean audio condition are comparable with other Slovenian continuous speech recognition systems of similar complexity. The overall accuracy for Slovenian continuous speech recognition systems is lower than for some other major languages (i.e. English, Spanish), as Slovenian belongs to the group of highly inflectional languages.

The next step of analysis added to the experiments the test set with low and medium noise acoustic conditions. The preliminary tests with high noise acoustic conditions showed drastic reduction of speech recognition accuracy (approx. 25%), consequently we decided to exclude this test set from the following experiments. The speech recognition results are presented in Figure 2.

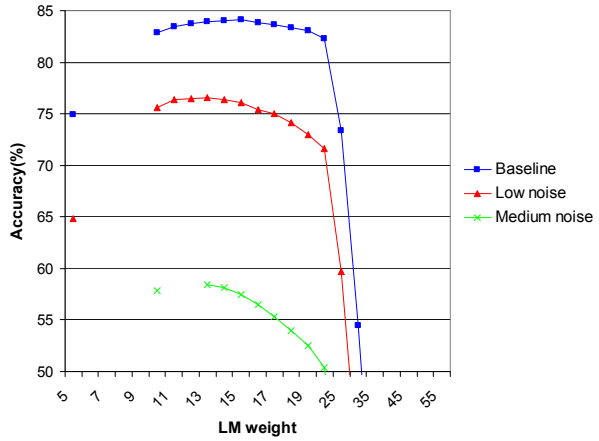

Fig.2. Speech recognition accuracy for test set in various acoustic conditions.

The analysis first shows the general influence of noise on speech recognition accuracy. The overall accuracy was reduced by approximately 10% when low noise was added to the test set, and by approximately 25% when medium noise was added. The analysis of best speech recognition accuracies shows the change of language model weight, which was needed to obtain optimal accuracy in adverse acoustic conditions. The best accuracy of 76.58% was achieved on low noise test set, when the language model weight was set to 13. The best accuracy of 58.45% on the medium noise test set was also obtained at the language model weight 13. This shift of language model weight clearly shows that also language model influences the robustness of speech recognition in noisy environment. The detailed analysis of type of speech recognition errors showed that the results achieved in low noise conditions could still be used in robotic environment, despite the lower overall accuracy. On the other site are the results for medium (and high) noise condition to worse to be usable in real-life environment.

Adverse acoustic conditions also influence the speed of speech recognition and with it the time-to-command delay. This is mainly caused by the increased search space in the speech recognition decoder, which results in the reduced accuracy. The continuous speech recognition speed is usually measured with real-time factor xRT. The analysis of speech recognition speed was carried out on a high performance computer server, such as are used in distributed speech recognition environments found in state-of-the-art client/server robot platforms [21, 22]. The results of increase of real-time factor are given in Table 3.

Table 3. Increase of real-time factor in various acoustic conditions.

Acoustic conditions	**Increase of xRT**
Baseline	1.00
Low noise	2.97
Medium noise	5.15
High noise	7.05

The speed of speech recognition with the low noise added to the test set already dramatically decreased. The real-time factory was increased by the factor of 2.97. Even worse results were observed for the medium (5.15 increase) and high (7.05 increase) noise conditions.

The last step of analysis evaluated the performance of adapted language models. The speech recognition results are given in Table 4.

Table 4. Speech recognition accuracy with adapted language model in various acoustic conditions

Acoustic conditions	Accuracy(%)
Low noise	78.48
Medium noise	60.33
High noise	25.86

The adapted language model improved the speech recognition accuracy for low noise acoustic conditions from 76.58% to 78.48%. The statistically significant improvement of speech recognition accuracy confirms the importance of adaptation procedures used during the development process.

Conclusion

The paper presented the research on speech recognition in noisy environment. Improvements are proposed that are based on two key ideas - optimizing language model weight and adapting the language model. We showed the effectiveness of the system through several experiments of speech recognition in noisy environments.

The work was partially funded by Slovenian Research Agency, under contract number P2-0069, Research Programme "Advanced methods of interaction in telecommunication".

REFERENCES

[1] ROJC, Matej. Web-based architecture RES based on finite-state machines for distributed evaluation and development of speech synthesis systems. *Int j. comput. linguist. res.* (Print)/, Mar. 2011, vol. 2, no.1, pp. 1-12.

[2] Drungilas D, Grisius G, Recognition of Human Emotions in Reasoning Algorithms of Wheelchair Type Robots. *Informatica*, 2010,Vol. 21, No. 4, pp. 521-532.

[3] Sato M, Iwasawa T, Sugiyama A, Nishizawa T, Takano Y, A Single-Chip Speech Dialogue Module and Its Evaluation on a Personal Robot, PaPeRo-Mini. IEICE *TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences* 2010, Vol.E93-A No.1 pp.261-271.
[4] Janusz Dulas: Automatic digits recognition for Polish – noisy phonemes identification, *Electrical Review,* 01/2011, pp. 280-283.
[5] R. P. Lippmann et al., Multi-style training for robust isolated word speech recognition, *Proc. of ICASSP, 1987*, 705–708.
[6] Liao, Y. F., Fang, H. H., Hsu, C. H., Eigen-MLLR Environment/Speaker Compensation for Robust Speech Recognition. *Proceeding Interspeech, 2008*, Brisbane, Australia, pp. 1249–1252.
[7] Donglin Wang, Leung, H., Keun-Chang Kwak, Hosub Yoon, Enhanced Speech Recognition with Blind Equalization For Robot "WEVER-R2".The 16th *IEEE International Symposium on Robot and Human interactive Communication*, 2007, pp. 684-688.
[8] Yoshitaka Nishimura, Mikio Nakano, Kazuhiro Nakadai, Hiroshi Tsujino and Mitsuru Ishizuka, Speech Recognition for a Robot under its Motor Noises by Selective Application of Missing Feature Theory and MLLR, *Proc. of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition* (SAPA, 2006), Pittsburgh, pp.53-58.
[9] Yamamoto, S., Nakadai, K., Nakano, M., Tsujino, H., Valin, J.-M., Komatani, K., Ogata, T., Okuno, H.G., Design And Implementation Of A Robot Audition System For Automatic Speech Recognition Of Simultaneous Speech, *IEEE Workshop on Automatic Speech Recognition & Understanding*, ASRU, 2007, pp. 11-116.
[10] Yamamoto, S., Valin, J.-M., Nakadai, K., Rouat, J., Michaud, F., Ogata, T., Okuno, H.G., Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory, Proceedings of the 2005 IEEE International Conference on Robotics and Automation, pp. 1477 -1482.
[11] Jianjun Huang, Yafei Zhang, Xiongwei Zhang, Tao Zhu: A Data Field method for speech enhancement incorporating Binary Time-Frequency Masking, *Electrical Review*, 2011, No 7, pp.225 - 229.
[12] Kelley, J.F., An empirical methodology for writing user-friendly natural language computer application. *Proceeding of ACM SIG-CHI, 1983 Human Factors in Computing Systems*. New York: ACM, pp. 193 - 196.
[13] Lee CH, On Automatic Speech Recognition at the Dawn of the 21st Century. *IEICE Trans. Inf. Syst*., vol. E86-D, 2003, No. 3, pp. 377-396.
[14] Maučec, M. S., Žgank, A., Speech recognition system of Slovenian broadcast news. Speech technologies. Rijeka: InTech. 2011, pp. 221-236.
[15] Lipeika A, Optimization of Formant Feature Based Speech Recognition. *Informatica*, 2010, Vol. 21, No. 3, pp. 361-374.
[16] Maskeliunas R, Rudzionis A, Rudzionis V., Advances on the Use of the Foreign Language Recognizer. Development of Multimodal Interfaces: Active Listening and Synchrony, *Lecture Notes in Computer Science*, 2010, Springer Verlag, vol. 5967, pp. 217-224.
[17] Cho Y, Yook D., Maximum Likelihood Training and Adaptation of Embedded Speech Recognizers for Mobile Environments. *ETRI Journal*, 2010, vol.32, no.1, pp.160-162.
[18] Pyz G, Simonyte V, Slivinskas V., Modelling of Lithuanian Speech Diphthongs. *Informatica*, 2011,Vol. 22, No. 3, pp. 411-434.
[19] Arhar, Š., Gorjanc, V., Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 2007, 52/2., pp. 95-110.
[20] Žgank, A., Verdonik, D., Markus, A. Z., Kačič, Z., BNSI Slovenian broadcast news database - speech and text corpus, *In INTERSPEECH, 2005*, pp. 1537-1540.
[21] Jang C, Lee S, Jung S, Song B, Kim R, Kim R, Lee CH, OPRoS: A New Component-Based Robot Software Platform. *ETRI Journal*, 2010, vol.32, no.5, pp.646-656.
[22] Rambow M, Rohrmüller F, Kouraks O, Brščivcić D, Wollherr D, Hirche S, Buss M (2010) A Framework for Information Distribution, Task Execution and Decision Making in Multi-Robot Systems. *IEICE TRANSACTIONS on Information and Systems,* 2010, Vol.E93-D No.6 pp.1352-1360.

***Authors**: prof. dr Mirjam Sepesy Maučec, prof. dr. Zdravko Kačič , prof. dr. Andrej Žgank, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia, e-mail: {mirjam.sepesy, kacic, andrej.zgank} @uni-mb.si.*