ANALYSIS OF MORPH-BASED LANGUAGE MODELING AND SPEECH RECOGNITION IN SLOVAK

Jan STAS[1], *Daniel HLADEK*[1], *Jozef JUHAR*[1], *Daniel ZLACKY*[1]

[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University of Kosice, Park Komenskeho 13, 042 00 Kosice, Slovak Republic

jan.stas@tuke.sk, daniel.hladek@tuke.sk, jozef.juhar@tuke.sk, daniel.zlacky@tuke.sk

Abstract. *The inflection of the Slovak language causes a large number of unique word forms, which produces not only a large vocabulary, but also a number of out-of-vocabulary words. Morph-based language models solve this problem by decomposition of inflected word forms into small sub-word units and resolve the general problem of sparsity the training data. In this paper, we present several rule-based and data-driven approaches to the automatic segmentation of words into morphs. These data are later used in the modeling of the Slovak language for large vocabulary continuous speech recognition. Preliminary results show a significant decrease in the number of out-of-vocabulary words and reduction of resultant language model perplexity.*

Keywords

Automatic word segmentation, language modeling, morphological analysis, speech recognition.

1. Introduction

In highly-inflectional or agglutinative languages, the rich morphology and relatively free order of words in sentences causes several problems of using standard *n*-gram models in the statistical language modeling for *large vocabulary continuous speech recognition* (LVCSR). Growing vocabulary causes great increasing the number of *n*-grams and the size of a resultant language model (LM) as well, which has a significant impact on time and memory requirements in the process of decoding word sequence pronounced by user. Another problem is the sparsity of training data, which causes the fact that the estimation of the conditional probabilities for the infrequent *n*-grams is not reliable enough. Therefore, several methods have been proposed to eliminate the mentioned disadvantages of the standard LMs with respect to the morphology of the given language. One of the possible ways how to eliminate the data sparsity problem is to use the morph-based language modeling in large vocabulary continuous speech recognition.

Morph-based models resolve the data sparsity problem by decomposing words into small sub-word units - morphs in order to reduce the resultant size of vocabulary and number of *out-of-vocabulary* (OOV) words, because the number of all possible morphs is always smaller than the number of all possible words in the given language and the average occurrence of these sub-word units in a training corpus is always larger than the average occurrence of the words in the same corpus. In general, morphs can be represented by the *syllables*, *pre-fixes*, *stems*, *roots*, *suffixes*, *endings*, or by other meaning-ful sub-word units. Than the selection of proper sub-word units in the statistical language modeling depends on the several parameters, for example on the average length of words in a language, the bounds of the word-inflection, the manner of word-forming, morph pronunciation, or their inter- or intra-word coverage.

Contemporary language modeling distinguishes three main categories of the representation morphs in LMs:

- *syllable-based models*, successfully applied in the modeling of the Polish language [1],
- *models based on grammatical or statistical morphs*, mainly used in the modeling agglutinative languages, such as Finnish or Estonian [2],
- *models with using stems and endings*, which decompose words only into two word segments and are often used in modeling inflective languages, such as Czech [3], Slovenian [4], or Russian [5].

By detailed study of the inflected languages, we came to the knowledge that the inflection in the Slovak is concentrated on the border of the stem and ending of a word. In this way, we can simply increase the predictive ability of a LM in comparison to the previous approaches, where the history of morphs within one word for *n*-gram model is generally never known and that may generate many grammatically incorrect words by their chaining. Therefore, we have focused on the analysis and utilization the morph-based LMs using stems and endings of the words. This approach is a certain compromise among the very short word segments and whole words.

This article is organized as follows. The next section introduces an overview about the selected approaches and principles used in automatic word segmentation into sub-word units by using rule-based and data-driven methods and principles for segmentation of Slovak words that have been proposed untill now. Our proposed algorithm for word segmentation into stems and endings in Slovak will also be presented in this section. Further section describes the speech recognition setup used in experiments, about results of which the fifth section gives a short overview. Finally, at the end of this paper, main contributions and future directions in the Slovak language modeling will be mentioned.

2. Automatic Word Segmentation

An In general, methods for automatic word segmentation to the sub-word units can be divided into three main categories:

- *rule-based approaches*, which decompose words by using predefined rules for word segmentation in a given language,
- *data-driven approaches*, which are based on the statistical techniques,
- *hybrid segmentation approaches*, which combine both mentioned approaches.

One of the first data-driven approaches is the *semantically oriented segmentation of words*, proposed for the inflected languages that are based on the principle of the *latent semantic analysis*, while boundaries for word segmentation are given by the branching factor in a tree structure for a group of words with similar properties (affix candidates) [6].

Among the most well-known word, segmentation methods is the *letter successor variety* (LSV) *segmentation* [7]. This *knowledge-free morphology segmentation* is the base for the other different data-driven methods. Algorithm is based on the computing frequency of distribution of the character variants after (or before) the group of characters (respectively) in a given word. The segmentation boundaries are specified in the places after (or before) maximal (or minimal) value of the occurrence variety of characters on the given position. Bordag later extended LSV segmentation by a combination of the different evaluation metrics involved in the surrounding context. Even though this method considers several different metrics for determining the optimal word segmentation, we have observed approximately equal results in the segmentation of Slovak words in comparison to the standard LSV segmentation [8].

Similar approach, called the *minimum description length* (MDL) *principle* was proposed for Finnish [9]. This relatively complex unsupervised data-driven algorithm considers a model that would be able to describe examined language, the morphological regularities in a language or entire model by using the probability distribution of morphs in a group of words with similar properties in the simplest and shortest way. MDL algorithm was later successfully applied in the system called Morfessor for the modeling of the Finnish language using statistical morphs [10], later in modeling Slovenian [4], agglutinative Estonian, inflective Turkish or conversational Arabic using morph-based LMs [11].

Among the other approaches, the segmentation with using morphological analyzer by application of *two-level morpho-analysis* in the modeling of the Czech language using segmentation into stems and endings [3] or statistical approach for determining syllabic boundaries using Sylseg based on the *hidden Markov models* (HMM), which has been trained on the text data segmented by simple rules for syllabic segmentation of Slovak words [12] may be mentioned.

3. Word Segmentation in Slovak

Selected approaches for automatic word segmentation to the sub-word units mentioned in the previous section have been applied to the Slovak language in order to choose the most suitable segmentation needed for language modeling.

In the following experiments described in [13] we have observed that using Harris or Bordag segmentation method, based on the LSV computing, a correct segmentation in a few number of cases has been reached only. Only the word prefixes were correctly segmented. On the contrary, a large amount of words contained in dictionary, remained non-segmented, which limits the applicability of these methods in the Slovak language modeling.

As it was presented in the introduction of this paper, segmentation into two sub-word units - stems and endings, appears the most suitable in the case of the inflective languages. This approach has also been applied in the modeling of the Slovak language using classes of words, derived from the word suffixes [14]. As the proposed word segmentation methods consider statistically-selected morphemic boundaries only without considering correct phonetic morph pronunciation, it was necessary to propose segmentation rules regarding morphemic and phonetic boundaries within words.

These rules can be summarized in follows:

- morphological segmentation into stems and endings satisfies rules for automatic word segmen-tation to the syllables in Slovak [12], where each suffix agrees with last syllable of a words,
- only words than have more than two syllables or 7 characters are segmented,

- each suffix has length between 2 and 4 characters,
- remaining words with length less than two syllables or 7 characters are not segmented,
- segmentation was constrained by the lexicon, which contains about 1 million of grammatically correct words from the Slovak National Corpus database [15].

Another type of segmentation was implemented by modification of the segmentation using Morfessor [10], by satisfying the rules described above. In this case, if some words were segmented at the end into several morphs (infixes or affixes), then they were merged into one sub-word unit.

Syllable segmentation using Sylseg was later extended by a novel training set, with which the memory requirements of the original algorithm increased. Therefore, it was necessary to modify the existing algorithm by application of a binomial tree structure, which reduced the time requirements for localization of appropriate morphs in the model for segmentation. These experimental results for automatic syllable segmentation of Slovak words using modified Sylseg are published in [13].

4. Speech Recognition Setup

Experiments have been performed with bi-, tri- and quadrigram morph-based LMs, created by using SRILM Toolkit [16] and vocabulary sizes from 25*k* up to 125*k* (with step 25*k*) of the most frequent words, stems or endings in a training corpus (see Tab. 1). All morph-based models were smoothed by the Witten-Bell back-off algorithm. They have been trained on a newspaper text corpus size of about 180 million of tokens, gathered from the newspaper web-pages written in Slovak, from 2007 to 2011 year, by our system for text gathering and processsing called webAgent [17].

As the acoustic model (AM), the triphone context-dependent model based on the HMMs has been used, where each state of the HMM has been modeled by 32 Gaussian mixtures. The model has been generated from feature vectors containing 39 mel-frequency cepstral (MFC) coefficients and created using about 60 hours of readings by professionally trained speakers obtained from the Slovak Broadcast News (BN) database, recorded from 2007 to 2009 year. The acoustic database is characterized by gender-balanced speakers and contains read and spontaneous speech [18]. Rare triphones have been modeled by the effective triphone mapping algorithm [19].

In the decoding process, we have used the LVCSR engine Julius based on two-pass strategy, where the input data are processed in the first pass with bigram LM and the final search is performed with trigram LM using the results of the first pass to narrow the search space [20].

The test data were represented by about 4 hours of randomly selected speech recordings from the Slovak BN acoustic database that were not used in training of the AM and contain 40 656 words in 4 343 sentences.

For evaluation, two standard measures based on the model perplexity (PPL) and the word error rate (WER) have been used. Perplexity is defined as the reciprocal of geometric probability assigned by the LM to each word in the test set and WER is computed by comparing the reference text read by a speaker against the recognized results. It takes into the account substitution, insertion and deletion errors and evaluates the overall performance of the LVCSR system.

Tab.1: Statistics of morphs in dictionaries.

ID	Number of			Size of	
	Syllables	Stems	Endings	Words	Vocab
Reference					
25*k*	-	-	-	25 025	25 025
50*k*	-	-	-	50 107	50 107
75*k*	-	-	-	75 271	75 271
100*k*	-	-	-	100 558	100 558
125*k*	-	-	-	125 930	125 930
Proposed stem-ending segmentation					
25*k*	-	13 506	1 449	10 053	25 007
50*k*	-	28 395	2 062	19 777	50 234
75*k*	-	43 759	2 523	29 224	75 506
100*k*	-	59 344	2 883	37 905	100 132
116*k*	-	70 315	3 013	43 410	116 738
Proposed stem-ending segmentation – modified vocabulary					
25*k*	-	12 556	3 013	9 432	25 011
50*k*	-	27 714	3 013	19 343	50 070
75*k*	-	43 263	3 013	28 920	75 196
100*k*	-	59 344	3 013	37 905	100 262
116*k*	-	70 315	3 013	43 410	116 738
Stem-ending segmentation using Morfessor					
25*k*	-	5 255	1 295	18 459	25 006
50*k*	-	10 191	2 179	37 795	50 162
75*k*	-	14 802	3 097	57 137	75 033
100*k*	-	18 971	4 032	77 352	100 352
125*k*	-	22 857	4 944	99 171	126 969
Morfessor and proposed stem-ending segmentation					
25*k*	-	12 689	2 316	10 008	25 011
50*k*	-	25 538	3 844	20 759	50 139
75*k*	-	36 875	5 350	32 810	75 033
100*k*	-	47 217	6 757	47 649	101 621
125*k*	-	54 921	7 848	62 816	125 583
Syllable segmentation using Sylseg					
25*k*	4 685	-	-	20 405	25 090
50*k*	5 341	-	-	45 418	50 759
62*k*	5 452	-	-	56 936	62 388

5. Experimental Results

Regarding to the results of the analysis of morph-based LMs, experiments have been oriented on the evaluation of the number of OOV words, the model perplexity and the word error rate computed on the test data, in order to discover the impact of the selected techniques of word segmentation and statistical language modeling on the overall precision of the LVCSR system in the task of the Slovak Broadcast News transcription. The proposed technique for word segmentation into stems and ending was compared to the modified segmentation using

Morfessor and the combination of this two mentioned approaches, and with syllable segmentation using Sylseg, in the task of the statistical language modeling using bi-, tri- and quadrigrams morph-based models of the Slovak language with respect to the standard word-based LMs. Statistics about the number of the most frequent morphs and words obtained by the counting in a training corpus can be seen in the Tab. 1, and the number of OOV words, PPL and the WER of the LVCSR system is in the Tab. 2.

Tab.2: Evaluation of the quality of *n*-gram morph-based models of the Slovak language for different types of word segmentation.

ID	Size of	OOV	Bigrams		Trigrams		Quadrigrams	
	Vocab	[%]	PPL	WER [%]	Vocab	WER [%]	PPL	WER [%]
Reference								
25*k*	25 025	12,12	472,52	29,80	368,14	28,48	362,49	28,48
50*k*	50 107	6,85	566,23	20,75	435,74	19,60	429,67	19,59
75*k*	75 271	4,82	615,25	17,64	471,37	16,51	464,83	16,58
100*k*	100 558	3,74	644,95	16,06	493,63	14,83	486,74	14,88
125*k*	125 930	3,05	669,48	15,05	512,10	13,94	505,00	13,94
Proposed stem-ending segmentation								
25*k*	25 007	4,43	199,10	27,04	119,83	24,56	112,85	24,33
50*k*	50 234	2,36	211,67	22,95	124,99	20,48	117,60	20,41
75*k*	75 506	1,73	217,35	21,87	127,81	19,45	120,26	19,24
100*k*	100 132	1,47	220,24	21,28	129,51	18,73	121,90	18,60
125*k*	116 738	1,38	221,28	21,17	130,11	18,90	122,46	18,61
Proposed stem-ending segmentation – modified vocabulary								
25*k*	25 011	4,64	198,95	27,58	120,79	25,29	113,25	25,09
50*k*	50 070	2,40	211,60	23,04	125,08	20,55	117,68	20,42
75*k*	75 196	1,73	217,45	21,79	127,90	19,36	120,34	19,10
100*k*	100 262	1,47	220,25	21,34	129,51	18,82	121,89	18,66
116*k*	116 738	1,38	221,28	21,17	130,11	18,90	122,46	18,61
Stem-ending segmentation using Morfessor								
25*k*	25 006	7,49	323,25	27,04	219,08	25,24	211,97	25,16
50*k*	50 162	3,87	365,35	21,10	244,20	19,16	236,04	19,27
75*k*	75 033	2,54	389,92	19,31	255,47	17,47	246,84	17,45
100*k*	100 352	1,88	396,99	18,45	263,69	16,56	254,82	16,55
125*k*	126 969	1,47	404,47	17,81	268,42	16,01	259,38	16,06
Morfessor and proposed stem-ending segmentation								
25*k*	25 011	4,28	198,13	27,67	120,40	24,94	113,46	24,77
50*k*	50 139	2,10	210,04	23,71	125,67	21,11	118,26	21,01
75*k*	75 033	1,40	215,60	22,80	128,49	20,15	120,89	19,92
100*k*	101 621	1,02	218,96	22,57	130,38	19,93	122,60	19,72
125*k*	125 583	0,83	221,19	22,76	131,51	20,01	123,75	19,86
Syllable segmentation using Sylseg								
25*k*	25 090	4,02	97,49	42,23	50,22	36,33	42,04	35,26
50*k*	50 759	3,78	99,44	41,96	51,23	36,29	42,89	35,15
62*k*	62 388	3,75	99,80	42,17	51,42	36,49	43,06	35,23

As we can see, from the experimental results of the word segmentation into sub-word units, the best results were achieved by using our proposed word segmentation technique, despite the fact that by the unsupervised segmentation using Morfessor have produced slightly larger number of unique endings, which were the major part of a sub-word unit in around one third of cases have been overlapped. This word segmentation approach also produces the highest number of *n*-grams in LMs. On the contrary, by using the proposed technique in combination with unsupervised segmentation using Morfessor we have achieved a significant reduction appro-ximately 70 % in the number of OOV words, also in the case of the use smaller vocabularies. This fact resulted in the lowest values of PPL, practically in all cases about 68 %, relatively. A moderate degradation was only observed in WER, approximately increased by 3,92 % in the case of bigram, 1,75 % for trigram and 1,54 % for quadrigram LMs, absolutely. As we can see in the Tab. 2, no improvements were achieved by modification of vocabulary by introducing all possible endings into vocabulary. The best results of the speech recognition with using morph-based LMs have been observed with modified segmentation using Morfessor. In all these cases, only for the size of vocabulary 25*k* significant decreasing of WER within the range from 8,20 % to 12,79 %, relatively was reached.

On the contrary, even at the lowest PPL values of syllable-based LMs, word segmen-tation using Sylseg produced the highest WER and it can be concluded that this type of word segmentation is not very suitable for the Slovak language modeling.

At the end of this article, it can be noted that using stems and endings the trigram morph-based models are sufficient for modeling of the Slovak language. Only small shift in evaluation values was observed by using higher-order *n*-grams. The most suitable type of segmentation is our proposed algorithm or modified version of the unsupervised segmentation using Morfessor, whose efficiency can be improved in the future by introducing *model-based word segmentation*. Currently we cannot say exactly what size of vocabulary is the most appropriate to design the effective morph-based models of the Slovak language in a specific domain-oriented task of the LVCSR system.

6. Conclusion

This paper was oriented on the analysis of methods for automatic word segmentation and morph-based statistical modeling of the Slovak language. Since the Slovak language belongs to the group of highly-inflective languages, where the flection is concentrated on the last syllable of a word, we have decided to direct our research to the problem of word segmentation into stems and endings. With respect to the fact that already proposed techniques for segmentation of Slovak words do not consider the segmentation on the morpho-phonetic boundaries, we have decided to propose a new algorithm for word segmentation into stems and endings that achieved the best results in the number of OOV words and model perplexity with a slight decrease in word error rate of our LVCSR system in the task of the Slovak Broadcast News transcription in comparison to the other examined approaches. Further research will be oriented on the improvement of the data-driven algorithms for an unsupervised word segmentation, for example by optimization of the segmentation using Morfessor, in order to increase the quality and decrease memory requirements of the Slovak morph-based language models in LVCSR systems.

Acknowledgements

The research presented in this paper was supported by the Ministry of Education under research project VEGA 1/0386/12 (20 %), MS SR 3928/2010-11 (30 %) and Research and Development Operational Program funded by the ERDF under the project IMTS-26220220141 (50 %).

References

[1] MAJEWSKI, Piotr. Syllable Based Language Model for Large Vocabulary Continuous Speech Recognition of Polish. In: *Proc. of the 11th International. Conference on Text, Speech and Dialogue, TSD'2008*. Berlin: Springer Verlag, 2008, vol. 5246, pp. 397-401. ISSN 0302-9743. ISBN 978-3-540-87390-7. DOI: 10.1007/978-3-540-87391-4_51.

[2] CREUTZ, M., T. HIRSIMAKI, M. KURIMO, A. PUURULA, J. PYLKKONEN, V. SIIVOLA, M. VARJOKALLIO, E. ARISOY, M. SARACLAR and A. STOLCKE. Analysis of Morph-Based Speech Recognition and the Modeling Out-of-Vocabulary Words across Languages. In: *Proc. of NAACL-HLT'2007*. Rochester: ACM, 2007, pp. 380-387. Avaible at: http://www.speech.sri.com/cgi-bin/run-distill?/pubs/papers/Creu0704:Morph//document.ps.gz.

[3] BYRNE, W. J., J. HAJIC, P. IRCING, P. KRBEC and J. PSUTKA. Morpheme-Based Language Models for Speech Recognition of Czech. In: *Proc. of the 3rd International Workshop on Text, Speech and Dialogue, TSD'2000*. Berlin: Springer Verlag, 2000, pp. 211-216. ISBN 978-3-540-41042-3. DOI: 10.1007/3-540-45323-7_36.

[4] ROTOVNIK, T., M. S. MAUCEC and Z. KACIC. Large Vocabulary Continuous Speech Recognition of an Inflected Language using Stems and Endings. *Speech Communications*. 2011, vol. 49, iss. 6, pp. 437-452. ISSN 0167-6393. DOI: 10.1016/j.specom.2007.02.010.

[5] KARPOV, A., I. S. KIPYATKOVA and A. RONZHIN. Very large Vocabulary ASR for Spoken Russian with Syntactic and Morpheme Analysis. In: *Proceedings of INTERSPEECH'2011*. Florence: ISCA, 2011, pp. 3161-3164. ISBN 978-1-61839-270-1.

[6] SCHONE, P. and D. JURAFSKY. Knowledge-Free Induction of Morphology using Latent Semantic Analysis. In: *Proc. of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning*. Lisbon: Association for Computational Linguistics, 2000, pp. 67-72. DOI: 10.3115/1117601.1117615.

[7] HARRIS, Zellig S. From Morpheme to Phoneme. *Language*. 1955, vol. 31, no. 2, pp. 190-222. ISSN 0097-8507.

[8] BORDAG, Stefan. Unsupervised Knowledge-Free Morpheme Boundary Detection. In: *Proc. of International Conference on Recent Advances in Natural Language Processing, RANLP'2005*. Borovets: INCOMA, 2005, pp. 1-7. ISBN 954-91743-3-6. DOI: 10.1.1.109.2571.

[9] BARRON, A., J. RISSANEN and B. YU. The Minimum Description Length Principle in Coding and Modeling. *IEEE Transaction on Information Theory*. 1998, vol. 44, iss. 6, pp. 2743-2760. ISSN 0018-9448. DOI: 10.1109/18.720554

[10] CREUTZ, M. and K. LAGUS. Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora using Morfessor 1.0. *Computer and Information Science*. 2005, vol. Tech. report A81, pp 27. ISSN 1913-8997.

[11] HIRSIMAKI, T., J. PYLKKONEN and M. KURIMO. Importance of Higher-Order N-Gram Models in Morph-Based Speech Recognition. *IEEE Transaction on Audio, Speech and Language Processing*. 2009, vol. 17, iss. 4, pp. 724-732. ISSN 1558-7916. DOI: 10.1109/TASL.2008.2012323.

[12] IVANECKY, Jozef. Analysis of Rule-Based Phonetic Transcription Technique applied to Slovak Language. In: *Proc. of the 3rd International Seminar SLOVKO'2005: Computer Treatment of Slavic and East Europe Languages*. Bratislava: VEDA, 2005, pp. 130-136. ISBN 80-224-0895-6.

[13] MIRILOVIC, M. and J. JUHAR. Morphological Segmentation of Word Units for Large Vocabulary Automatic Speech Recognition in Slovak. In: *Proc. of the 3rd Baltic Conference on Human Language Technologies, Baltic HLT'2007*. Kaunas: Association for Computational Linguistics, 2007, pp. 189-196. ISBN 978-9955704539.

[14] HLADEK, D., J. STAS and J. JUHAR. Word Clustering for Slovak Class-based Language Model. *Journal of Electrical and Electronics Engineering*. 2012, vol. 5, no. 1, pp. 85-88. ISSN 2067-2128.

[15] r-mak-2.0 – Rucne morfologicky anotovany korpus (manually morphologically annotated corpus. L. STUR INSTITUTE OF LINGUISTICS, Slovak Academy of Science. *Slovak National Corpus* [online]. Bratislava, 2007. Available at: http://korpus.juls.savba.sk/stats.html.

[16] STOLCKE, Andreas. SRILM – An Extensible Language Modeling Toolkit. In: *Proc. of ICSLP'2002*, Denver: International Computer Science Institute, vol. 2, 2002, pp. 901-904. ISBN 1 876346 43 4.

[17] HLADEK, D. and J. STAS. Text Mining and Processing for Corpora Creation in Slovak Language. *Journal of Computer Science and Control Systems*. 2010, vol. 3, no. 1, pp. 65-68, ISSN 1844-6043.

[18] PLEVA, M., J. JUHAR and A. CIZMAR. Slovak Broadcast News Speech Corpus for Automatic Speech Recognition. In: *Proc. of the 8th International Conference on Research in Telecommunication Technology*, *RTT'2007*. Slovak Republic, 2007, pp. 334-337. ISBN 978-80-8070-735-4.

[19] DARJAA, S., M. CERNAK, M. TRNKA and M. RUSKO. Effective Triphone Mapping for Acoustic Modeling in Speech Recognition. In: *Proc. of INTERSPEECH'2011*. Florence: ISCA, 2011, pp. 1717-1720. ISBN 978-1-61839-270-1.

[20] LEE, A., T. KAWAHARA and K. SHIKANO. Julius – An Open Source Real-Time Large Vocabulary Recognition Engine. *Proc. of EUROSPEECH'2001*. Aalborg: ISCA, 2001, pp. 1691-1694. ISBN 87-90834-09-7.

About Authors

Jan STAS was born in Bardejov, Slovakia in 1984. In 2007 he graduated M.Sc. (Ing.) at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. He received his Ph.D. degree in 2011 at the same department in the field of Telecommunications. He is currently working as a post-doctoral researcher at the Department of Electronics and Multimedia Communications at the Technical University of Kosice. His research interests include natural language processing, computational linguistics and statistical language modeling in large vocabulary continuous speech recognition (LVCSR) systems.

Daniel HLADEK was born in Kosice, Slovakia in 1982. In 2006 he graduated M.Sc. (Ing.) at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Kosice. He has obtained Ph.D. degree in 2009 at the same department in the field of Computational Intelligence. He is currently working as a post-doctoral researcher at the Department of Electronics and Multimedia Communications at the Technical University of Kosice with focus on programming, natural language processing, statistical language modeling in large vocabulary continuous speech recognition (LVCSR) systems.

Jozef JUHAR was born in Poproc, Slovakia in 1956. He graduated from the Technical University of Kosice in 1980. He received Ph.D. degree in Radioelectronics from Technical University of Kosice in 1991, where he works as an Associate 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.

Daniel ZLACKY was born in Poprad, Slovakia in 1988. He 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 Kosice. His research interests include automatic word segmentation and language modeling in large vocabulary continuous speech recognition (LVCSR) systems.