RESCORING N-BEST LISTS FOR RUSSIAN SPEECH RECOGNITION USING FACTORED LANGUAGE MODELS

Irina Kipyatkova[1], *Vasilisa Verkhodanova*[1], *Alexey Karpov*[1,2]

[1]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[2]University ITMO, St. Petersburg, Russia
{kipyatkova, verkhodanova, karpov}@iias.spb.su

ABSTRACT

In this paper, we present a research of factored language model (FLM) for rescoring N-best lists for Russian speech recognition task. As a baseline language model we used a 3-gram language model. Both baseline and factored language models were trained on a text corpus collected from recent news texts on Internet sites of online newspapers; total size of the corpus is about 350 million words (2.4 GB data). For FLMs creation, we used five factors: word, its lemma, stem, part-of-speech, and morphological tag. We investigate the influence of factor set on language model perplexity and word error rate (WER). Experiments on large vocabulary continuous Russian speech recognition showed that FLM can reduce WER.

Index Terms— factored language model (FLM), automatic speech recognition (ASR), N-best lists, Russian language processing

1. INTRODUCTION

Russian is morphologically rich inflective language. Words in Russian can inflect for a number of syntactic features: case, number, gender etc., this leads to a large number of possible word forms and consequent problems connected with sparseness of data. Vast majority of lexical items (except adverbs, prepositions etc.) modify its basic form (lemma) according to grammatical, morphological, and contextual relations. This is a common characteristic for all Slavic (or Slavonic) languages [1]. Thus, it is possible to cite as example a comparison of word *nice* and its inflected equivalents in Russian: one word compared to 24 words.

Rich morphology of Russian as well as of many other Slavic languages results in extremely large vocabulary. New words with similar meaning can be created by adding single or multiple prefixes, suffixes and endings to a stem, or also by modifying a stem itself [2]. Even grading of adjectives and adverbs is done by adding specific suffixes and prefixes. Slavic morphology is primarily fusional, that is a given affix frequently combines the expression of a number of grammatical categories [3]. Thus, automatic speech recognition (ASR) vocabulary for Slavic languages requires millions of words, that is 10 or 20 times larger than ASR systems for English language where the inventory of 50 thousands most frequent words yields the coverage rate about 99% [2].

Another feature that is characteristic for Russian language as well as other Slavic languages is a relatively free word order: for example, the subject-verb-object triple in Russian is possible in all 6 surface orders. This became possible due to rich morphology, because the role of the word in the sentence is determined by its inflected form. But in contrast to free word order there is a complex grammatical agreement systems in Slavic languages [3].

The appearance of these features results in the increasing of vocabulary size and the number of out-vocabulary (OOV) words. In terms of OOV rates, Russian is comparable to some other morphologically rich European languages, such as Finnish, Czech, Hungarian, Lithuanian or Turkish [4, 5, 6].

In recent years a number of approaches dealing with mentioned issues were widely introduced and tested for speech recognition systems for the Russian language. The survey of Russian ASR systems is given in [7], while here we present several new works.

In [8], authors deal with a Russian speech recognition system developed within the Quaero program. The system uses two different acoustic front-ends in order to train the acoustic models. 4-gram case sensitive language models (LMs) with vocabulary of 500K were trained on broadcast news, web data, books, and audio transcripts. Experiments showed word error rate of about 20% on the official Quaero 2010 evaluation set. The carried out analysis of recognition errors showed that many recognition errors were caused by inflections and Yo-homonyms.

In [9], a maximum entropy language model for Russian with features specifically designed to deal with the inflections in Russian language is described. This model combined with subword based language model was used for N-best list rescoring. This led to reduction of word error rate by 1.2%.

A large vocabulary continuous speech recognizer that uses syllable-based LM is presented in [10]. A method for recognized syllables concatenation and error correction is proposed. The syllable lexicon has about 12K entries. The final sentence is constructed from the recognized syllables by the designed co-evolutionary asymptotic probabilistic genetic algorithm (CAPGA).

In [11], authors deal with method of syntactic links accounting in language model. They used such processing stages as part-of-speech (POS) tagging, dependency parsing and factored language models for hypotheses rescoring. Experiments were performed on parts of Russian National Corpora and have shown that only one of mixed models showed slightly better results than the simple 3-gram model. The best accuracy was 91.77%, that is 1.26% better than results obtained with the baseline model.

A continuous Russian speech recognition with deep belief networks in conjunction with HMM is presented in [12]. The first of two recognition stages was the use of deep belief networks to calculate the phoneme state probability for feature vectors. At the second stage Viterbi decoder used these probabilities for generating resulting sequence of words. The experiments were performed on the collected by FSSI Research Institute corpus of telephone speech, with 25 hours used for training, 1 hour for validation and 1 hour for test. Additionally 17 hours of unlabeled speech were used for pretraining of deep neural networks. Experiments showed best results with deep neural networks in the case of 5 layers with 1000 elements for one layer. In that case accuracy was 45%.

In [7], syntactico-statistical language model is proposed to take into account long-distance syntactic dependencies between word pairs. This model was created by joint application of statistic and syntactic analysis of training text data. Application of the model to large vocabulary speech recognition task allowed to decrease WER from 30.5% to 26.9%

Yandex SpeechKit [13] provides an ASR search application for Russian language. For improvement of acoustical modeling authors used deep neural networks. At the moment application allows searching general information and geo information (streets, places). Authors claim that accuracy is 84% for general information and 94% for geo information.

Finally, for automatic voice search in the Internet, Google Inc. has developed the on-line Voice Search service [14], which uses speech recognition technology. This service allows users to find necessary information in the Internet pronouncing a word or a phrase. For the LM creation, written queries to Google search engine were used. This technology is also applied to other Google services, for example, Google maps, where it is possible to perform voice request for searching a place on the map. For short and common sentences it works pretty well, but it fails for conversational Russian speech.

2. FACTORED LANGUAGE MODELING

Alternative to N-gram language models is factored language models (FLM) that for the first time was introduced in order to deal with morphologically rich Arabic language [15]. Then it has been used for many other morphologically rich languages. This model incorporates various morphological features (factors) and it can be used for inflective languages. So, a word is viewed as a vector of k factors: $w_i = (f_i^1, f_i^2, ..., f_i^k)$. Factors of a given word can be word, morphological classes, stems, roots, and other grammatical features. Probabilistic language model is constructed over sets of the factors.

There are two main issues in FLM developing [16]:

1. choosing an appropriate set of factor definitions by using data-driven techniques or linguistic knowledge;
2. finding the best statistical model over these factors.

In FLM, there is no obvious way of backing-off path [15]. In word N-gram modeling backing-off is performed by dropping first the most distant word, followed by the second most distant word, and so on until the unigram language model is used. This process is illustrated in Figure 1 (a). In FLM any factor can be dropped at each step of backing-off, and it is not obvious which factor to drop first. In this case, several backoff paths are possible, what results in a backoff graph. An example of backoff graph is presented on Figure 1(b). The graph shows all possible single step backoff paths, where exactly one variable is dropped per backoff step.

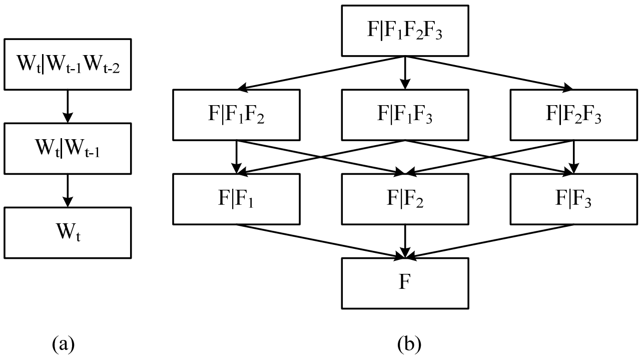

Figure 1. N-gram and FLM backoff trees: (a) backoff path for a 3-gram language model over words; (b) backoff graph for with three parent variables F_1, F_2, F_3

In [17], factored language model is incorporated at different stages of the speech recognition system: at the stage of N-best list rescoring and at recognition stage. Because the use of FLM at the recognition stage is problematic, for speech decoding a word-based language model rescored with FLM was used. Recognition results showed an improved WER with the FLM used for N-best rescoring task by 0.8-1.3% depending on the test speech corpus, and usage of FLM at speech recognition gave additional improving of WER by 0.5%.

FLM is applied for lattice rescoring in [18]. For speech recognition HTK decoder was used. The decoder generated a lattice of 100 best alternatives for each test sentence using a word-based bigram language model with 5K vocabulary. Then the lattice was rescored with various morpheme-based and factored language models. Word recognition accuracy obtained with baseline model was 91.60%. Usage of FLM increased word recognition accuracy up to 92.92%.

In [19], morpheme-based trigram language mode for Estonian was used for N-best list generating. Vocabulary of the language model consisted of 60K particles. Then the obtained morpheme sequences were reconstructed to word sequences. Factored language model which used words and their part-of-speech tags was applied to rescore N-best hypotheses. A relative WER improvement of 7.3% was obtained on a large vocabulary speaker independent recognition task.

In [20], FLM was combined with recurrent neural network for Code-Switching Language Modeling task. The combined language model gave a relative improvement of 32.7% comparing to the baseline 3-gram model.

An application of the FLM for Russian speech recognition is described in [21, 22]. FLM was trained on the text corpus containing 10M words with vocabulary size of about 100K words. FLMs were created using the next factors: word, lemma, morphological tag part-of-speech, and gender-number-person factor. TreeTagger [23] tool was used for obtaining the factors. Investigation of influence of different factors and backoff path on perplexity and WER was carried out. FLM was used for rescoring of 500-best list. Evaluation experiments showed that FLM allows to achieve 4.0% WER relative reduction, and 6.9% relative reduction was obtained when FLM was interpolated with baseline 3-gram model.

3. THE BASELINE SPEECH RECOGNITION SYSTEM

3.1. Architecture of the baseline speech recognition system

An architecture of the software of automatic analysis, recognition and diarization of Russian speech (PARAD-R) is presented on Figure 2. PARAD-R software is built on the basis of a three-level processing (client, server, program-mathematical core) [24]. The client and server can be located either on the same computer or on different computers and can communicate over a computer network. The exchange of information between client and server is implemented using protocols MRCPv2 (Media Resource Control Protocol) and RTSP (Real-Time Streaming Protocol).

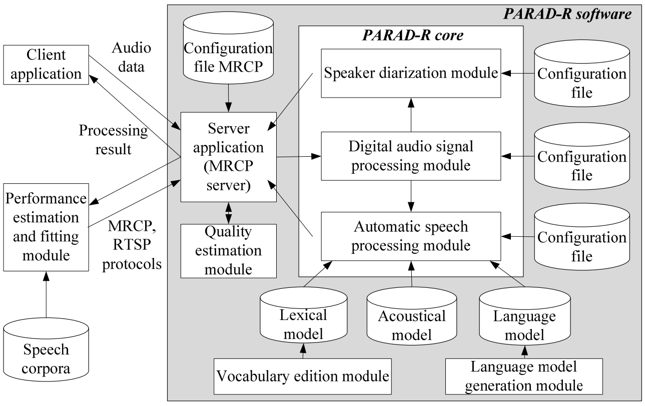
PARAD-R software
PARAD-R core
Client application
Audio data
Configuration file MRCP
Speaker diarization module
Configuration file
Processing result
Server application (MRCP server)
Digital audio signal processing module
Configuration file
Performance estimation and fitting module
MRCP, RTSP protocols
Quality estimation module
Automatic speech processing module
Configuration file
Lexical model
Acoustical model
Language model
Speech corpora
Vocabulary edition module
Language model generation module

Figure 2. Architecture of the PARAD-R speech analysis software

The server consists of the following software modules: a server application - MRCP server, the modules of vocabulary editor, language model generator and quality estimator. Each of these modules, except the last, is implemented as an executable file running OS MS Windows XP/Vista/7. In addition to these software modules, the server is also linked to the core of mathematical software, which includes: digital audio processing, speaker diarization, automatic speech recognition [25-30]. Each of these modules is implemented as a static library to be connected to the server application.

3.2. Acoustic modeling

Training of acoustic models of speech units is carried out with the use of a Russian speech corpus. In this research, we have used our own corpus of spoken Russian speech Euronounce-SPIIRAS, created in 2008-2009 in the framework of the Euro-Nounce project [31]. The speech data were collected in clean acoustic conditions, with 44.1 kHz sampling rate, 16-bit audio quality. A signal-to-noise ratio (SNR) at least 35-40 dB was provided. The database consists of 16,350 utterances pronounced by 50 Russian native speakers (25 male and 25 female). Each speaker reads 327 phonetically-balanced and meaningful sentences carefully, but fluently one time only. Total duration of speech data is about 21 hours.

Hidden Markov Models (HMM) are used for acoustic modeling, and each phoneme (speech sound) is modeled by one continuous density HMM. A phoneme model has three states: the first state describes phoneme's start, the second state presents a middle part, and a third state is phoneme's end. HMM of a word is obtained by connection of phoneme's models. Similarly the models of words are connected with each other, generating the models of phrases. The aim of training of the acoustic models based on HMM is to determine such model's parameters that would lead to maximum value of probability of appearance of this sequence by training sequence of observations [32].

3.3. Baseline Language Modeling

For the language model creation, we collected and automatically processed a new Russian text corpus of on-line newspapers. This corpus was collected from recent news published on freely available Internet sites of on-line Russian newspapers (www.ng.ru, www.smi.ru, www.lenta.ru, www.gazeta.ru, www.interfax.ru, ria.ru) for the years 2006-2013. The procedure of preliminary text processing and normalization is described in [7]. The size of the corpus after text normalization and deletion of doubling or short (<5 words) sentences is over 350M words, and it has above 1M unique word-forms.

For the statistical text analysis we used the SRI Language Modeling Toolkit (SRILM) [33]. We created 3-gram language models with different vocabulary sizes, and the best speech recognition results were obtained with 150K vocabulary [34]. Perplexity of the baseline model is 553. So this vocabulary was chosen for further experiments with N-best list rescoring.

4. FACTORED LANGUAGE MODEL CREATION

The software "VisualSynan" from the AOT project [35] was used for obtaining morphological word features. We used five factors: the word, its lemma, stem, part-of-speech (POS), and morphological tag.

The training text corpus was processed to replace words with their factors. For example, the word 'схеме' ("scheme") is replaced with the vector {W-схеме: L-схема: S-схем: P-сущ: G-bc}, where W is a word, L is a lemma, S is a stem, P is POS, G is a morphological tag that means noun POS, feminine gender, singular, dative case. We created 4 models with the word plus one of the other factors using Witten-Bell discounting method.

We have tried 2 fixed backoff paths:

1. The first drop was of the most distant word and factor, then – of the less distant ones.
2. The first drop was of the words in time-distance order, the drop of the factors in the same order.

Figure 3 shows an example of these backoff paths for a model W+L.

Table 1 presents perplexity of the obtained models calculated on text data consisting of phrases (33M words in total) from another online newspaper "Фонтанка.ru" (www.fontanka.ru). Perplexity is given with two different normalizations: counting all input tokens (PPL) and excluding end-of-sentence tags (PPL1).

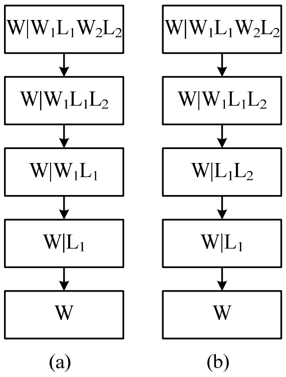

Figure 3. Backoff paths for a model W+L: (a) backoff path 1; (b) backoff path 2

Table 1. Perplexity of different FLMs with different backoff paths

Factors	Backoff path 1		Backoff path 2	
	PPL	PPL1	PPL	PPL1
W	-	-	553	878
W+L	826	1405	1007	1739
W+S	1637	2937	1834	3320
W+G	750	1264	900	1539
W+P	725	1219	727	1223

The models built with backoff path 1 have smaller perplexity. The largest value of perplexity has the model with word and stem factors.

5. EXPERIMENTS

To test the speech recognition system we used a speech corpus that contains 500 phrases pronounced by 5 speakers. The phrases were taken from the materials of the on-line newspaper «Фонтанка.ru» (www.fontanka.ru).

For speech recognition we used decoder Julius ver. 4.2 [36]. WER obtained with the baseline 3-gram language model was 26.54%. The OOV rate for the test set was 1.1%. RTF was 2.5 for the speech decoder installed on a desktop PC with multi-core Intel Core i7-3770K 3.5 GHz processor.

We produced several N-best lists with different number of hypotheses and carried out rescoring of N-best lists using created FLMs. The results are presented in Table 2.

Table 2. WER obtained after rescoring of N-best lists with FLMs with different backoff paths

Models	50-best		20-best		10-best	
	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2
W+L	28.05	29.06	27.83	28.39	**26.95**	27.77
W+S	29.33	30.30	29.01	29.46	27.90	28.63
W+G	**27.88**	**28.39**	**27.30**	**27.58**	27.88	**27.15**
W+P	28.75	29.40	27.72	28.48	27.32	27.60

Table 2 shows that in most cases model with W and G factors gave the better results, but WER was worse than WER obtained before N-best list rescoring. Then we carried out linearly interpolation of FLMs with baseline 3-gram model. Performance of obtained models in terms of WER is presented in Table 3.

Table 3. WER obtained after rescoring of N-best lists with FLMs interpolated with 3-gram model

Models	50-best		20-best		10-best	
	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2
3gram and W+L	26.10	26.16	25.69	25.84	25.90	26.05
3gram and W+S	26.46	26.27	26.01	26.03	26.20	26.09
3gram and W+G	26.01	25.45	25.71	**25.28**	25.79	25.45
3gram and W+P	25.88	26.03	**25.51**	25.75	25.75	25.84

Some better results were obtained after rescoring of 20-best lists than of 50-best and 10-best lists. The lowest WER (25.28%) was obtained by means of the baseline model interpolated with FLM, in which W and G factors were used (backoff path 2). The second best result was obtained when the FLM with W and P factors (backoff path 1) was used for interpolation with 3-gram model. So, we created the model that is the linearly interpolation of 3-gram model, W+G model (backoff path 2), and W+P model (backoff path 1). In this case WER was equal to 25.19%. So we obtained a relative WER reduction of 5% comparing to baseline the system.

On Figure 4 the distribution graph of mean values of best hypothesis number for different N-best sizes is presented. It shows that when N-best list is increasing in number more than 30, the increase of mean number of best hypothesis slows down.

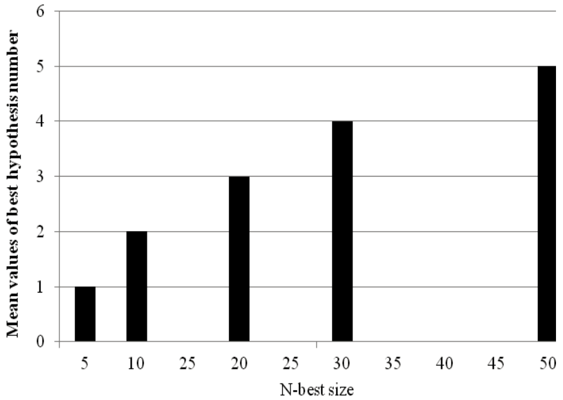

Figure 4. Distribution graph of mean values of best hypothesis number for different N-best sizes

Our results are consistent with those obtained in [22]. But comparing to [22] we used another morphological parser - AOT [35] while authors in [22] used TreeTagger [23]. For our experiments we used training set of 350 million words that is 35 times larger set than in [22]. In the end our results are better and support the hypothesis of [22] that FLM improve recognition accuracy.

6. CONCLUSION

Rich morphology of Russian complicates the creation of language models. FLMs can help to include additional information in language model and thereby to improve Russian speech recognition system.

In the paper we have investigated an application of FLM for N-best lists rescoring for Russian speech recognition. We made a comparison of influence of factor set on speech recognition results. We obtained relative WER reduction of 5% comparing to the baseline system.

In further research we plan to investigate FLMs with more than two factors and try generalized backoff in which multiple different paths are chosen dynamically at a run time.

7. ACKNOWLEDGEMENTS

This research is supported by the Russian Foundation for Basic Research (project No. 12-08-01265), by the Russian Humanitarian Scientific Foundation (project No. 12-04-12062) and by the Government of Russian Federation (Grant 074-U01).

8. REFERENCES

[1] L. Ryazanova-Clarke, T. Wade, “The Russian Language Today Routledge”, Language Arts & Disciplines, 2002.

[2] J. Nouza, J. Zdansky, P. Cerva, J. Silovsky, “Challenges in speech processing of Slavic languages (Case studies in speech recognition of Czech and Slovak)”, A. Esposito et al. (Eds.): Development of Multimodal Interface: Active Learning and Synchrony, LNCS 5967, Springer-Verlag, Heidelberg, pp. 225–241, 2010.

[3] B. Comrie, G. G. Corbet. The Slavonic Languages. London and New York: Routledge Press, 1993.

[4] P. Ircing, J. Hoidekr, J. Psutka, “Exploiting linguistic knowledge in language modeling of Czech spontaneous speech”, in Proceedings of Int. Conf. on Language Resources and Evaluation LREC’2006, Genoa, Italy, pp. 2600–2603, 2006.

[5] M. Kurimo, et al., “Unlimited vocabulary speech recognition for agglutinative languages”, in Proceedings of Human Language Technology Conference of the North American Chapter of the ACL, New York, USA, pp. 487–494, 2006.

[6] A. Vaičiūnas, “Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition”, PhD thesis, Vytautas Magnus University, Kaunas, 2006.

[7] A. Karpov, K. Markov, I. Kipyatkova, D. Vazhenina, A. Ronzhin, "Large vocabulary Russian speech recognition using syntactico-statistical language modeling", Speech Communication. 2014, Vol. 56, January, pp. 213-228.

[8] Y. Titov, K. Kilgour, S. Stüker, and A. Waibel, "The 2011 kit quaero speech-to-text system for the Russian language," in Proceedings of the 14th International Conference "Speech and Computer" (SPECOM'2011), pp. 136-143, 2011.

[9] E. Shin, S. Stüker, K. Kilgour, C. Fügen, A. Waibel, "Maximum Entropy Language Modeling for Russian ASR", in Proc. of the International Workshop for Spoken Language Translation (IWSLT 2013), Heidelberg, December 5-6, 2013.

[10] S. Zablotskiy, A. Shvets, M. Sidorov, E. Semenkin, W. Minker, "Speech and Language Recources for LVCSR of Russia", in Proceedings of LREC'2012, Istanbul, Turkey, pp. 3374–3377, 2012.

[11] M. Zulkarneev, P. Satunovsky, N. Shamraev, "The use of d-gram language model for speech recognition in Russian", SPECOM 2013, Springer LNAI 8113, pp. 362-366, 2013.

[12] M. Zulkarneev, R. Grigoryan, N. Shamraev, "Acoustic modeling with deep belief networks for Russian Speech", SPECOM 2013, Springer LNAI 8113, pp. 17-23, 2013.

[13] SpeechKit API, http://api.yandex.ru/speechkit/.

[14] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen , M. Kamvar, B. Strope, "Google Search by Voice: A Case Study". Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, pp. 61-90, 2010.

[15] J. A. Bilmes, K. Kirchhoff, "Factored language models and generalized parallel backoff", in Proc. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 2, Stroudsburg, PA, USA, pp. 4–6, 2003.

[16] K. Kirchhoff, J. Bilmes, and K. Duh, "Factored Language Models Tutorial", Tech. Report UWEETR-2007-0003, Dept. of EE, U. Washington, June 2007.

[17] D. Vergyri, K. Kirchhoff, K. Duh, A. Stolcke, "Morphology-Based Language Modeling for Arabic Speech Recognition", in Proc. ICSLP 2004, pp. 2245-2248, 2004.

[18] M.Y. Tachbelie, S. Teferra Abate, W. Menzel, "Morpheme-based language modeling for amharic speech recognition", in Proc. of the 4th Language and Technology Conference, LTC-2009, Posnan, Poland, pp. 114-118, 2009.

[19] T. Alumae, "Sentence-adapted factored language model for transcribing Estonian speech". in Proc. of ICASSP 2006. Toulouse, France, pp. 429–432, 2006.

[20] H. Adel, N. T. Vu, T. Schultz, "Combination of Recurrent Neural Networks and Factored Language Models for Code-Switching Language Modeling", in Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria, 2013.

[21] D. Vazhenina, K. Markov, "Evaluation of advanced language modelling techniques for Russian LVCSR", SPECOM 2013, Springer LNAI 8113, pp. 124-131, 2013.

[22] D.Vazhenina, K.Markov, "Factored Language Modeling for Russian LVCSR", In Proc. International Joint Conference on Awareness Science and Technology & Ubi-Media Computing, Nov. 2013.

[23] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in Proc. the International Conference on New Methods of Language Processing, Manchester, UK, 1994, pp. 44–49.

[24] A. Ronzhin, V. Budkov, I. Kipyatkova, "PARAD-R: Speech Analysis Software for Meeting Support", in Proc. of the 9th International Conference on Information, Communications and Signal Processing ICICS-2013, Tainan, Taiwan, 2013.

[25] Al.L. Ronzhin, M.V. Prischepa, A.A. Karpov, "A Video Monitoring Model with a Distributed Camera System for the Smart Space", Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2010, LNCS 6294, 2010, pp. 102–110.

[26] R.M. Yusupov, A.L. Ronzhin, "From Smart Devices to Smart Space", Herald of the Russian Academy of Sciences, MAIK Nauka, Vol. 80, Number 1, 2010, pp. 63–68.

[27] I. Kipyatkova, A. Karpov, V. Verkhodanova, M. Zelezny, "Modeling of Pronunciation, Language and Nonverbal Units at Conversational Russian Speech Recognition", International Journal of Computer Science and Applications. – 2013. – Vol. 10, N 1. – pp. 11-30.

[28] A.L. Ronzhin, V.Yu. Budkov, "Multimodal Interaction with Intelligent Meeting Room Facilities from Inside and Outside", Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2009, LNCS 5764, 2009, pp. 77–88.

[29] V. Budkov, Al. Ronzhin, S. Glazkov, An. Ronzhin, "Event-Driven Content Management System for Smart Meeting Room", Springer-Verlag Berlin Heidelberg, S. Balandin et al. (Eds.): NEW2AN/ruSMART 2011, LNCS 6869, 2011, pp. 550–560.

[30] A. Ronzhin, V. Budkov. "Speaker Turn Detection Based on Multimodal Situation Analysis", SPECOM 2013, Springer LNAI 8113 LNAI 8113, 2013, pp. 302–309.

[31] O. Jokisch, A. Wagner, R. Sabo, R. Jaeckel, Cylwik N., M. Rusko, A. Ronzhin, R. Hoffmann, "Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system", in Proc. of SPECOM'2009, St. Petersburg, Russia, pp. 515–520, 2009.

[32] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition. Prentice Hall, 1995.

[33] A. Stolcke, J. Zheng, W. Wang, V. Abrash, "SRILM at Sixteen: Update and Outlook", in Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop ASRU'2011. Waikoloa, Hawaii, USA, 2011.

[34] I. Kipyatkova, A. Karpov, "Lexicon Size and Language Model Order Optimization for Russian LVCSR", SPECOM 2013, Springer LNAI 8113, pp. 219-226, 2013.

[35] A. Sokirko, "Morphological modules on the website www.aot.ru", in Proc. "Dialogue-2004", Protvino, Russia, pp. 559–564, 2004 (in Rus.).

[36] A. Lee, T. Kawahara, "Recent Development of Open-Source Speech Recognition Engine Julius", in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009), Sapporo, Japan, pp.131–137, 2009.