2. Semenkin E., Semenkina M. 2012. *Spacecrafts' Control Systems Effective Variants Choice with Self-Configuring Genetic Algorithm.* In: Ferrier, J.-L., Bernard, A., Gusikhin, O. and Madani, K. (eds), Proceedings of the 9th International Conference on Informatics in Control, Automation and Robotics, vol. 1, p. 84–93.

3. Davendra D. (ed.) 2010. Traveling Salesman Problem, Theory and Applications. InTech.

4. Papadimitriou C. H., Steiglitz K. 1982. Combinatorial Optimization. Algorithms and Complexity. Englewood Cliffs, NJ: Prentice-Hall.

5. Lin S., Kernigan B.W. 1973. *An Effective Heuristic Algorithm for the Traveling-Salesman Problem*. Operations Res. 21, p. 498–516.

6. Shah-Hosseini, H. 2007. Problem solving by intelligent water drops. *Proc. of IEEE Congresson Evolutionary Computation*, Swissotel The Stamford, Singapore, p. 3226–3231.

7. Dorigo M., Gambardella L. M. 1997. *Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem.* IEEE Transactions on Evolutionary Computation, p. 53–66.

8. Eiben A. E., Smith J. E. 2003. Introduction to evolutionary computing. Springer-Verlag, Berlin, Heidelberg.

9. Schaefer R., Cotta C., Kołodziej J., Rudolph G. 2010. *Parallel Problem Solving from Nature*. PPSN XI 11th International Conference, Kraków, Poland, p. 11–15.

УДК 004.93

EMOTION RECOGNITION AND SPEAKER IDENTIFICATION FROM SPEECH

M. Yu. Sidorov[1], S. G. Zablotskiy[1], W. Minker[1], E. C. Semenkin[2]

[1]Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany
E-mail: maxim.sidorov@uni-ulm.de, sergey.zablotskiy@uni-ulm.de, wolfgang.minker@uni-ulm.de
[2] Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: eugenesemenkin@yandex.ru

The performance of spoken dialogue systems (SDS) is not perfect yet, especially for some languages. Emotion recognition from speech (ER) is a technique which can improve the SDS behavior by finding critical points in the human-machine interaction and changing a dialogue strategy. Inclusion of the speaker specific information, by conducting the speaker identification procedure (SI) at the set up of ER task could also be used in order to improve the dialogue quality. Choosing of both appropriate speech signal features and machine learning algorithms for the ER and SI remain a complex and challenging problem. More than 50 machine learning algorithms were applied in the study for ER and SI tasks, using 9 multi-language corpora (Russian, English, German, and Japanese) of both acted and non-acted emotional utterance recordings. The study provides the results of evaluation as well as their analysis and future directions.

Keywords: emotion recognition from speech, speaker identification from speech, machine learning algorithms, speaker adaptive emotion recognition from speech.

РАСПОЗНАВАНИЕ ЭМОЦИЙ И ИДЕНТИФИКАЦИЯ СПИКЕРА ПО РЕЧЕВЫМ СИГНАЛАМ

М. Ю. Сидоров[1], С. Г. Заблоцкий[1], В. Минкер[1], Е. С. Семенкин[2]

[1]Университет города Ульма
Германия, 89081, Ульм, Аллея Альберта Эйнштейна 43. E-mail: maxim.sidorov@uni-ulm.de, sergey.zablotskiy@uni-ulm.de, wolfgang.minker@uni-ulm.de
[2]Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, Красноярск, просп. им. газ. «Красноярский рабочий» 31
E-mail: eugenesemenkin@yandex.ru

Производительность диалоговых систем, основанных на естественном языке, по-прежнему находится на достаточно низком уровне, особенно для некоторых языков. Распознавание эмоций на основе речевого сигнала представляет собой подход, способный улучшить качество работы таких систем посредством определения критических точек в диалоге между человеком и компьютером и последующей адаптации диалога. Использование процедуры идентификации улучшает качество распознавания эмоций на основе речевого сигнала пользователя, так как становится возможным построение моделей эмоций конкретного человека. Выбор подходящих параметров речевых сигналов и алгоритма моделирования для задач идентификации говорящего и распо-

знавания эмоций остаются важными проблемами. Более 50 алгоритмов машинного обучения были применены к задачам на 9 речевых корпусах различных языков (русского, английского, немецкого и японского). В используемых корпусах содержатся как реальные эмоции, так и постановочные диалоги актеров. Приведены результаты применения алгоритмов для обеих задач и их анализ.

Ключевые слова: распознавание эмоций и идентификация говорящего по речевым сигналам, алгоритмы машинного обучения, адаптивная процедура распознавания эмоций по речевым сигналам.

Nowadays, SDSs are included into car navigation systems, mobile devices and personal assistants and, thus, getting more and more popular. However, there are some problems which prevent the widespread using of such technologies. Firstly, one of the most important parts of SDSs is a speech recognition component, which provides the mapping between the speech signal and texts on the natural language, is not able to provide the ideal recognition accuracy. Secondly, some ambiguity is provided by the dialogue manager (DM) component. Therefore, the end-users are often disappointed or even angry while using such SDSs.

We have suggested here to use the additional information about the dialogue to improve its quality. Speaker specific information, through the speaker identification procedure, as well as gender specific information, through the gender identification from speech, and information about emotional state of a user, through emotion recognition from speech, could improve the performance of SDSs.

The solid line blocks in Figure 1 shows the baseline cycle of SDSs execution. In each turn the user communicates with the application. Recognized speech from a user comes to the DM block. The response of the system is sent back to user by a speech synthesis block The proposed techniques are demonstrated with the dash lines in Figure 1. Emotion specific information comes to dialogue manager, which makes a decision about the user satisfaction and adapts a dialogue strategy if needed.

We have focused here on the speaker identification and the emotion recognition procedures. The solution to such problems depends completely on the machine learning algorithms used for the modeling. We have applied more than 50 algorithms for solving these problems in order to figure out which algorithms should be used in real world applications. All evaluations were conducted on 9 different speech corpora to obtain representative speech samples and more objective results.

This paper is organized as follows: the used corpora and speech signal features are described in section 2; section 3 briefly describes the used machine learning algorithms; evaluation results as well as their analysis are shown in section 4; finally, there are some conclusion and direction for the future work in the 5th section.

Corpora Description and Feature Extraction. All evaluations were conducted using several speech databases. Here are their brief description and statistical characteristics.

Emotion recognition databases. *Berlin emotional database* [1] was recorded at the Technical University of Berlin and consists of labeled emotional German utterances which were spoken by 10 actors (5 f). Each utterance has one of the following emotional labels: neutral, anger, fear, joy, sadness, boredom and disgust.

Let's Go emotional database [2–4] comprises non-acted English (American) utterances which were extracted from the SDS based bus-stop navigational system. The utterances are requests to the system spoken by real users of this system. Each utterance has one of the following emotional labels: angry, slightly angry, very angry, neutral, friendly and non-speech – critical noisy recordings or just silence.

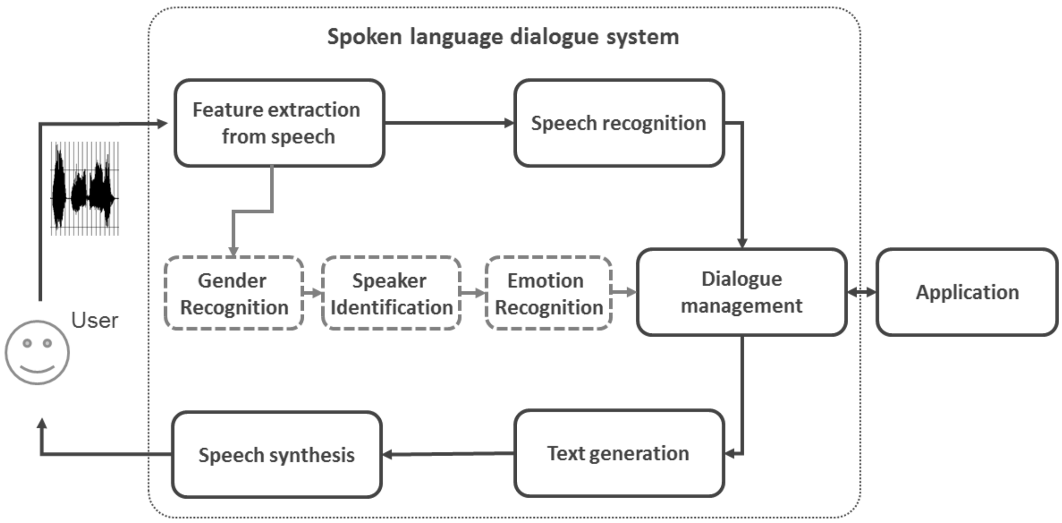

Fig. 1. SDSs execution cycle

SAVEE (Surrey Audio-Visual Expressed Emotion) corpus [5] was recorded as a part of an investigation into audio-visual emotion classification, from four native English male speakers. Emotional label for each utterance is one of the standard set of emotions (anger, disgust, fear, happiness, sadness, surprise and neutral).

UUDB (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database [6] consists of spontaneous Japanese speech through task-oriented dialogue which was produced by 7 pairs of speakers (12 f), 4737 utterances in total. Emotional labels for each utterance were created by 3 annotators on the 5-dimensional emotional basis (interest, credibility, dominance, arousal and pleasantness). To produce the labels for classification task we have used just pleasantness (or evaluation) and arousal axis. The corresponding quadrant (counterclockwise, starting in positive quadrant, assuming arousal as abscissa) can also be assigned emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene [7].

VAM-Audio database [8] was created at the Karlsruhe University and consists of utterances extracted from the popular German talk-show "Vera am Mittag" (Vera at afternoon). The emotional labels of the first part of the corpus (speakers 1-19) were given by 17 human evaluators and the rest of the utterances (speakers 20–47) were labeled by 6 annotators on the 3-dimensional emotional basis (valence, activation and dominance). The emotional labeling was done in a similar way to the UUDB corpora, using valence (or evaluation) and arousal axis.

Emotions itself and their evaluations have subjective nature. That is why it is important to have at least several evaluators of emotional labels. Even for humans it is not always evident to make a decision about an emotional label. Each study, which proposed an emotional database, provides also an evaluators confusion matrix and statistical description of their decisions.

Speaker identification databases. *Domian database.* Originally, it is a German radio talk-show [9] where people talk to a moderator about their private troubles. We have prepared a database based on the utterance extraction from these talk-show recordings. The collection of the data is still ongoing and by now it contains the utterances of 59 speakers.

The ISABASE-2 corpus [10] used in our work is one of the largest high-quality speech database of Russian and is normally used for Russian speech recognition [11] but we have used it to evaluate the speaker identification models as well. It was created by the Institute of System Analysis of the Russian Academy of Science with the support of the Russian Foundation of Fundamental Research in collaboration with a speech group of the Philological Faculty of Moscow State University and consists of more than 34 hours of clear, high-quality utterances spoken by 110 speakers (55 f).

The recording of the *PDA Speech Database* [12] was done at the Carnegie Mellon University using a PDA device. Each of 16 native speakers of American English reads about 50 sentences.

VAM-Video Database is a part of VAM-Corpus [8] has no emotional labels but still can be used to evaluate a speaker identification approaches. The number of speakers is 98.

The statistical description of the databases is in the tables 1 and 2.

Note, that the emotional databases were used for both ER and SI problems.

Feature extraction. The choice of the appropriate speech signal features for both problems is still an open question [13], nevertheless in this study the most popular ones have been chosen.

Table 1

Speaker identification corpora

Database	anguage	Full length (min.)	Number of speakers	File level Duration		Speaker level Duration	
				Mean (sec.)	Std. (sec.)	Mean (sec.)	Std. (sec.)
Berlin	German	24,7	10	2,7	1,02	148,7	40,5
Domian	German	235,6	59	6,1	5,1	239,6	80,9
Isabase	Russian	2053,6	110	4,8	1,06	1120,1	278,3
Let's Go	English	118,2	291	1,6	1,4	24,3	33,04
PDA	English	98,8	16	7,09	2,4	370,6	50,7
SAVEE	English	30,7	4	3,8	1,07	460,7	42,2
UUDB	Japanese	113,4	14	1,4	1,7	486,3	281,3
VAM-Audio	German	47,8	47	3,02	2,1	61,03	33,03
VAM-Video	German	75,7	98	3,1	2,2	46,3	35,6

Table 2

Emotion recognition corpora

Database	Number of emotions	Emotion level Duration		Notes
		Mean (sec.)	Std. (sec.)	
Berlin	7	212,4	64,8	Acted
Let's Go	5	1419,5	2124,6	Non-acted
SAVEE	7	263,2	76,3	Acted
UUDB	4	1702,3	3219,7	Non-acted
VAM-Audio	4	717,1	726,3	Non-acted

Average values of the following speech signal features were included into the feature vector: power, mean, root mean square, jitter, shimmer, 12 MFCCs and 5 formants. Mean, minimum, maximum, range and deviation of the following features have also been used: pitch, intensity and harmonicity (37-dimentional feature vector for one speech signal file, in total). The Praat [14] system has been used in order to extract speech signal features from wave files.

We have applied each algorithm in a static mode, i. e. each speech signal was parameterized by a single 37-dimensional feature vector consisting of corresponding average values.

Machine Learning Algorithms. A number of different algorithms were applied for both tasks in order to figure out which ones should be used to produce appropriate results in real world applications. This section provides short description of the used algorithms.

One may group the used algorithms into the following clusters: tree based modeling, artificial neural networks, Bayesian modeling, Gaussian modeling, instance based algorithms, rule based approaches, models based on the different fitting functions, support vector modeling and fuzzy rules.

Decision tree based algorithms. Such kind of models is based on a tree-like graph structure. The main advantages of such models are that they could be understandable for people and properly explained by Boolean logic. The majority types of decision trees based on recursive procedure, where on each iteration the entropy of every attribute using the data set is calculated; the data set is split into subsets using the attribute for which entropy is minimum; a decision tree node containing that attribute is created and recourse on subsets using remaining attributes. The standard ID3, C4.5 and M5 algorithms for decision tree building as well as the tree structure with logistic regression (LMT) and naïve Bayes classifiers (NBTree) at its leaves, in addition a random tree and a forest of random trees model were applied for ER and SI problems.

Rule based algorithms based on transformation from decision trees to rules. The most of such models grow the decision tree and produce logic rule from the best leaf. The baseline RIPPER algorithm for the growing rules, the hybrid algorithm of the decision table and Naïve Bayes classifier (DTNB) as well as the C4.5 and M5 rule growing algorithms were used in the study.

Artificial Neural Networks is a class of algorithms based on structural and functional modeling of human brain. Such algorithms are capable to solve difficult tasks of modeling, prediction and recognition. The state-of-the-art multi-layer perceptron (MLP) and neural networks designed by evolutionary algorithms (AutoMLP) were applied for the classification tasks.

Bayesian modeling algorithms based on the Bayesian theorem. The simple Naïve Bayes classifier and Naïve Bayes classifier with a kernel function as well as Bayes Network were applied to the problems.

Support Vector Machine (SVM) is a supervised learning algorithm based on a construction of a hyper plane or set of hyper planes in a high- or infinite- dimensional space. These models can be used for a classification, regression and other tasks.

Function fitting is a class of algorithms assumes that a model has some structure and the main task is to figure out the appropriate parameters of that structure. For instance, the linear regression model assumes that a data set is linearly separable in the feature space. A multinomial logistic regression is also based on the logistic function and generalizes a simple logistic regression by allowing more than two discrete outcomes. These algorithms as well as the Pace regression were applied for the modeling. The PLS classifier is a wrapper classifier based on the PLS filters which is able to perform predictions.

Lazy (or instance based) algorithms use only instances from a training set to create a class hypothesis of unknown instances. Basically they use different types of distance metrics between already known and unknown samples to produce a class hypothesis. The well-known k-Nearest Neighbors (kNN) algorithm uses the Euclidian metric, whereas the K-Star algorithm uses the Entropic based metric.

Fuzzy rule algorithms based on fuzzy logic and linguistic variables. This approach has a number of advantages, because they could deal with uncertainty, noisy and subjective data. It also could take a subjective experience into account. In this study Mixed Fuzzy Rule Formation algorithm was used for numeric data labeling.

Note, that some of the used algorithms can deal only with binary labels or could provide just regression procedure. Well-known one-against-all approach has been applied in the first case and classification by regression procedure (max of corresponding output) in the second one.

In order to evaluate the performance of the described algorithms the following systems were exploited: Weka [15], RapidMiner [16] and KNIME. Some additional algorithms were implemented in C++ and MATLAB programming languages from scratch.

Evaluation Results. This section demonstrates evaluation results. All data from each corpus had been parameterized before they were split into training and test partitions (0,7 vs. 0,3 correspondingly). The best algorithm for the ER task was the Gaussian Process (the highest average recognition accuracy over all corpora) which slightly outperformed decision tree with logistic regression at its leaves. The logistic regression, PLS classifier, linear regression and multi-layer perceptron also achieved a high value of recognition accuracy (see fig. 2).

The five best algorithms for the speaker identification task (see fig. 3) were multi-layer perceptron (the highest average identification accuracy over all corpora), decision trees with logistic regression at its leaves, functional trees, neural networks designed by evolutionary algorithm and k-nearest-neighbors algorithm.

Conclusions and future work. The study has revealed the most appropriate algorithms for emotion recognition and speaker identification tasks from speech. Evaluations have been conducted using cross-corporal and multi-language approach so the results can be assumed to be representative.

It is evident that the classification accuracy strongly depends on the amount of speech data for each class. Therefore, a high level of accuracy was achieved for the PDA and SAVEE corpuses and the low one for the VAM-Video and Let's Go databases (see *Number of classes* and *Class level duration* columns for the corresponding corpora in table 1 and table 2).

In the study we have used one average feature vector for each speech signal (machine learning algorithm applications in the *static* mode). Such approach has some advantages and the main one is the execution time of a feature extraction procedure. Using this approach, such procedures like the ER and the SI can be deployed in real time.

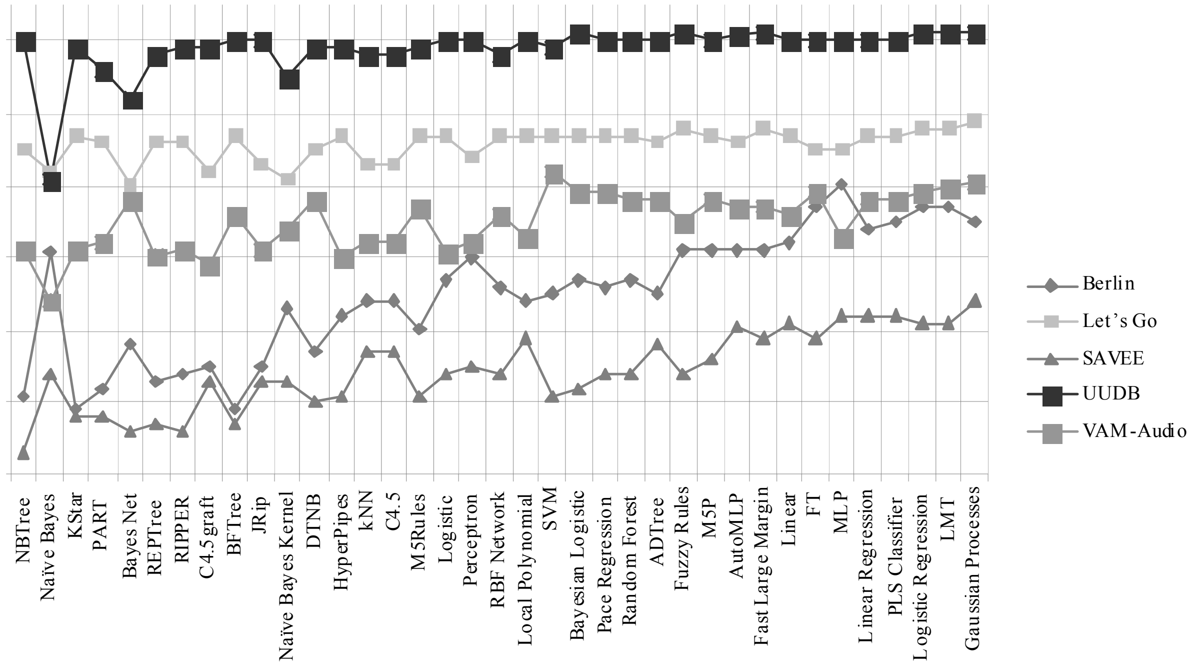

Fig. 2. Emotion recognition accuracy over all corpora

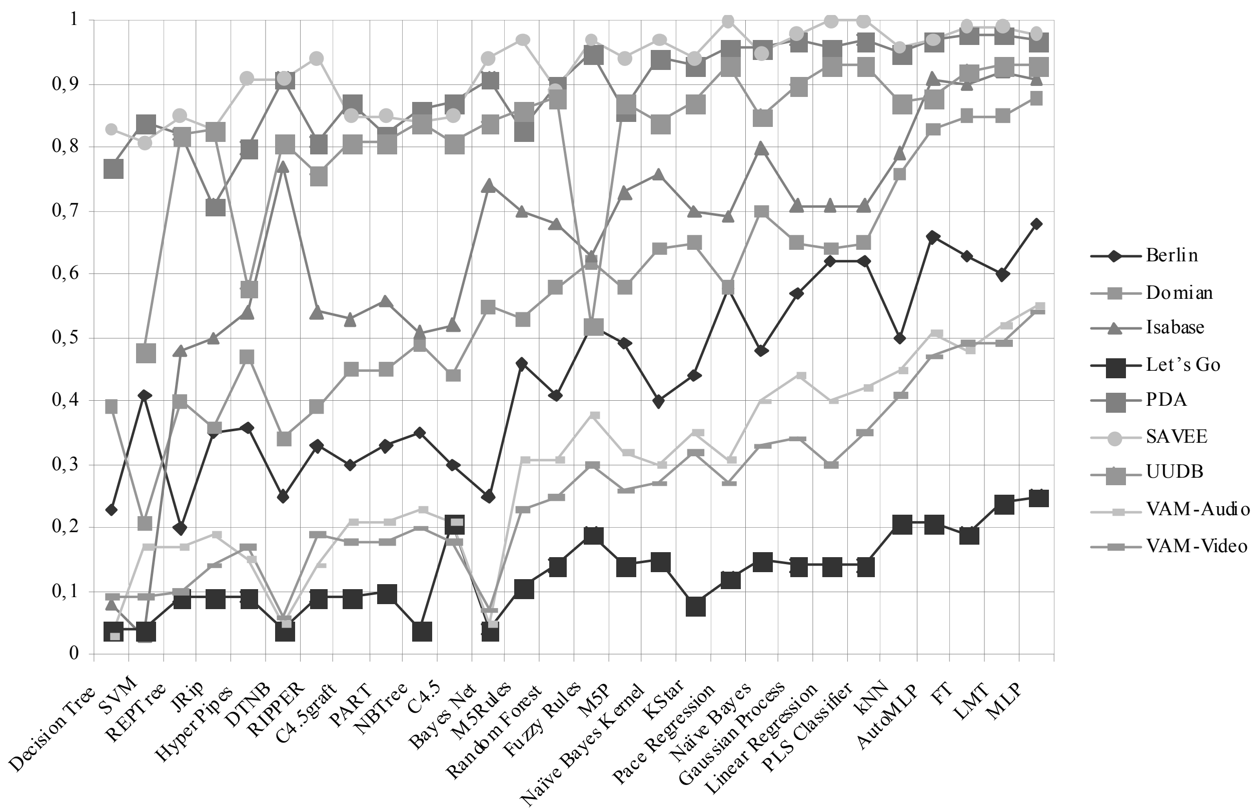

Fig. 3. Speaker identification accuracy over all corpora

Our future direction is the investigation of the machine learning algorithm applications in the *dynamic* mode. In this case the feature vectors are extracted consequently every short period of time (for example each 0,01 sec.). Moreover, speaker specific and gender specific information should be used in order to improve the emotion recognition accuracy from speech. The emotion recognition accuracy (as well as a SDS's performance in general) might be significantly improved by training of the speaker specific emotional models and using gender specific information as well. The next step is the exploitment of the best algorithms for emotion recognition and speaker identification from speech in order to build a speaker dependent emotion recognition systems.

References

1. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., & Weiss B. (2005, September). *A database of German emotional speech*. In Proc. Interspeech (Vol. 2005).
2. Maxine Eskenazi, Alan W Black, Antoine Raux, and Brian Langner. *Let's Go Lab: a platform for evaluation of spoken dialog systems with real world use* In Proceedings of Interspeech 2008 Conference, Brisbane, Australia.
3. Alexander Schmitt, Benjamin Schatz and Wolfgang Minker. *Modeling and predicting quality in spoken human-computer interaction.* In Proceedings of the SIGDIAL 2011 Conference, Association for Computational Linguistics, 2011.
4. Schmitt A., Heinroth T. and Liscombe J. *On NoMatchs, NoInputs and BargeIns: Do Non-Acoustic Features Support Anger Detection?* Proceedings of the SIGDIAL 2009. Conference, Association for Computational Linguistics, London, UK, 2009, p. 128–131.
5. Haq S. and Jackson P. J. B. *Multimodal Emotion Recognition*. In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, chapter 17, 2010, p. 398–423.
6. Available at: http://uudb.speech-lab.org/.
7. Schuller B., Vlasenko B., Eyben F., Rigoll G., & Wendemuth A. (2009, November). *Acoustic emotion recognition: A benchmark comparison of performances.* In Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, p. 552–557.
8. Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. *The Vera am Mittag German Audio-Visual Emotional Speech Database*. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hannover, Germany, 2008.
9. Available at: http://www.einslive.de/sendungen/domian/.
10. Bogdanov D. S., Bruhtiy A. V., Krivnova O. F., Podrabinovich A. Ya. and Strokin G. S., 2003. *Organizational Control and Artificial Intelligence, chapter Technology of Speech Databases Development (in Russian)*, p. 448. Editorial URSS.
11. Zablotskiy S., Shvets A., Sidorov M., Semenkin E. and Minker W. *Speech and Language Resources for LVCSR of Russian.* In Proceedings of the LREC 2012, Istanbul.
12. Available at: http://www.speech.cs.cmu.edu/databases/pda/README.html.
13. Sidorov M., Schmitt A., Zablotskiy S. and Minker W. *Survey of Automated Speaker Identification Methods.* In Proceedings of Intellegent Environment 2012, Athens, Greece. In Press.
14. Boersma Paul & Weenink David (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.50, retrieved 21 May 2013 from http://www.praat.org/.
15. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, vol. 11, Issue 1.
16. Mierswa, Ingo and Wurst, Michael and Klinkenberg, Ralf and Scholz, Martin and Euler, Timm. *YALE: Rapid Prototyping for Complex Data Mining Tasks*. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

УДК 519.6+004.9

TWO-STEPS SYSTEM IN SEARCHING SIMILAR WORDS FOR FAST AND RELIABLE AUTOMATIC CONCATENATION OF RUSSIAN SUB-WORD UNITS

A. Spirina[1], S. G. Zablotskiy[2], M. Yu. Sidorov[2]

[1]Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation. E-mail: s_nastia@mail.ru
[2]Ulm University
43, Albert-Einstein-Allee, Ulm, 89081, Germany.
E-mail: sergey.zablotskiy@uni-ulm.de, maxim.sidorov@uni-ulm.de

In this paper we describe and investigate the two-steps system sorting out inappropriate words in searching of similar words in the lexicon for automatic concatenation of Russian sub-word units. This two-steps system consists of com-