http://dx.doi.org/10.5755/j01.eee.20.9.8713

Web Services Based Hybrid Recognizer of Lithuanian Voice Commands

V. Rudzionis[1], G. Raskinis[2], R. Maskeliunas[3], A. Rudzionis[3], K. Ratkevicius[3], G. Bartisiute[3]
[1]Department of Informatics, Kaunas Faculty, Vilnius University,
Muitines St. 8, Kaunas, Lithuania
[2]Faculty of Informatics, Vytautas Magnus University,
Vileikos St. 8, LT-51367, Kaunas, Lithuania
[3]Centre of Information Systems Design Technologies, Kaunas University of Technology,
Studentu St. 50, LT-51368, Kaunas, Lithuania
vytautas.rudzionis@khf.vu.lt

***Abstract*—This paper presents the recently developed medical-pharmaceutical informative system with voice user interface. This is the first computerized system oriented towards healthcare services and industry where Lithuanian voice commands are used as a primary mean for control. Another essential property of the developed system is its hybrid nature: two different recognizers - an adapted commercial Spanish speech recognizer available from Microsoft and a locally developed HMM speech recognizer based on Lithuanian acoustic models – are operating in parallel. The recognition hypotheses produced by those recognizers are joined together using logical rules obtained using decision rules induction algorithms such as Ripper. All these measures and approaches allowed achieve very high speaker independent voice commands recognition accuracy acceptable for the system implementation in practice. The best achieved recognition was 98.9 % for 1000 Lithuanian voice commands. The paper presents optimization issues related with the development of the system.**

***Index Terms*—Speech recognition, speech analysis, human computer interaction, hybrid systems.**

I. INTRODUCTION

During last 15 years speech technologies (recognition, synthesis, speaker identification) became integral part of industrial applications of information technologies in various areas of human activities. It is widely accepted that one of the areas where speech recognition are gaining the strongest positions is healthcare industry and services. The main rationale for the application of voice processing in the healthcare is the desire to save the work time of highest qualification medical personal which is spent to routine operations of documentation as well as the desire to speed up and to easy up the information search and presentation and sometimes to allow healthcare practitioners to concentrate their attention to more important and urgent tasks. But implementation of speech recognition to the healthcare services has far more reaching consequences and has bigger benefits than just simply aiding the healthcare practitioners to document information or to find the necessary information. It is acknowledged that speech recognition has 5 main benefits for the healthcare industry: reusability, flexibility, less medical errors, productivity and so-called “one and done” effect [1]. Reusability means that speech recognition forces to produce structured text which is better suited for further processing in other applications and often is immediately available e.g. for billing. Flexibility means that voice recognition isn’t linked to a single device and is well suited to the cloud implementation which means that physician may use the same recognition profile from any device or any place. From a reimbursement standpoint, the closer the organization is between delivery-of-care and recording-of-care the more accurate the information is. Less medical errors means that speech recognition systems provide structured text and often strictly predefined vocabulary which is easier transferred and understood exactly with higher probability to another healthcare institution (historically medical documentation started off when the need to transfer patient from doctor to another occurred). Productivity means that using EMR (electronic medical records) physicians are able to serve more patients (some studies showed, that up to one third more patients [1]). “One and done” principle means that using speech recognition physician is able to perform some tasks independently which otherwise requires the help of other people (transcriptionists, secretaries, etc.): physicians can dictate and then see what they said, make small corrections, sign the documentation, then send it off in this way avoiding multiple steps and people involved in them.

Looking at the practice worldwide it is evaluated that about half a million health care practitioners have the possibilities to use speech technologies in their daily activities [2]. Among well-known examples of speech recognition applications in medical practice is one of the biggest healthcare services providers in Northern America – Providence Health and Services – installed voice controlled electronic health records management system Epic in 27 hospitals and more than 250 clinics. It is stated that this system is used by more than 8000 practitioner physicians daily. The financial benefits of voice recognition is well illustrated by the example that Dragon Dictate software allowed to save the Norman Regional Health System –

Manuscript received February 24, 2014; accepted June 10, 2014.

single hospital centre based in Oklahoma City – $1.8 million in transcription services only in 2013 [3]. According to a recent UN study [4] to many healthcare stakeholders the number of people with access to cell phones – around 6 billion, the study estimates – constitutes a major opportunity to increase access to healthcare worldwide. It would be difficult to improve access to healthcare services in many places of the world otherwise.

In any case the main prerequisite to implement such systems in Lithuania is the development of speech engine with proper reliability for the recognition of Lithuanian medical terms and other voice commands and phrases. It is obvious that from the user perspective, the more universal system is, and the more flexible vocabulary is used, the better. [5]. But evaluating the possibilities to achieve practically implementable system that could be useful for healthcare practitioners some vocabulary restrictions should be applied. The developed Lithuanian medical information system is able to recognize the names of most often met disease names, the most popular drug names and most often met in medical practice complaints. The total number of voice commands implemented in the system, is about 1000, which is probably the biggest number of Lithuanian voice commands implemented in practically oriented Lithuanian speech recognition system so far: the lack of available speech resources causes that still many Lithuanian speech recognition studies are devoted to issues of speaker-dependent recognition including some recent ones (e.g. [6]).

To achieve the accuracy goals, there was made a decision to develop hybrid speech recognition system for Lithuanian medical and pharmaceutical terms. Under the term hybrid system we assume the exploitation of several different speech recognizers and the combination of the hypotheses produced by them to derive a final decision [7].

Our earlier experience suggested the use of Microsoft Spanish speech engine as the foreign language speech engine for the adaptation purposes as one of the recognizers [8]. The CD-HMM based speech recognizer was used as the preferred selection for proprietary Lithuanian speech recognizer since it has shown that such model is the most efficient one in wide variety of applications. In [9] we showed that such approach could have the potential to reach the goals since the errors produced by different recognizers aren’t correlated and different approaches may be used to improve overall performance. This paper is devoted to various system optimization issues and the final design.

Further this paper is organized as follows. The Chapter II presents the methodology used to design and evaluate speech recognition system for Lithuanian healthcare institutions. Chapter III presents the speech corpora used to train and optimize the performance of the system. The Chapters IV and V present some experimental evaluation results and system optimization issues which enabled to achieve the highest recognition accuracy. Finally some conclusions are presented.

II. ALGORITHMS AND APPROACHES

This chapter introduces methods and principles used when developing recognition system. As was mentioned earlier hybrid system uses two different recognizers: adapted to recognize Lithuanian voice commands Spanish commercial speech recognizer and CD-HMM based proprietary Lithuanian recognizer. Adaptation of foreign language recognizer is based on principles of multilingual recognition [10]. These approaches could be summarized as an expectation that phonetic properties of one language (usually less widely used) could be largely described by the acoustic-phonetic models of other language. The aim of the training is to find the best compliance among the phonetic models in different languages. This adaptation procedure is usually called mapping and was investigated in many studies [11]–[13]. The research showed that in many cases this method allows to achieve high enough recognition accuracy. The main problem of adaptation, in our case, is to find the best phonetic sequences (transcriptions) to write Lithuanian voice commands. Our previous activities showed that the use of Spanish synthesizer or automatic generation of a wide set of possible transcription sequences and experimental evaluation of their efficiency may be useful.

Proprietary Lithuanian speech recognizer was built using continuous density hidden Markov models (CD-HMM) [14]. This model proved to be the most reliable and widely used approach today when designing speech recognition engines for a very wide set of applications. The essential problem is to have necessary speech corpora to train CD-HMM models. The absence of large enough amounts of speech data is the main obstacle when developing recognizers for such language as Lithuanian. Naturally proprietary Lithuanian recognizer has the higher potential to achieve highest recognition accuracy, if enough training data is available.

III. SPEECH CORPORA

To train and evaluate medical-pharmaceutical Lithuanian voice command recognition system special speech corpora MEDIC has been developed [9]. In the first stage recordings of 631 voice commands were collected. Among them were 217 names of most widely used diseases, 208 of most frequent complaints and 206 most often used drugs. This command set was formed from 1114 separate words, among them 777 words were unique. Later recordings of 100 more drug names were collected too. In this way the total number of drug names was 306 and the total number of recorded voice commands was 731. Each voice command was recorded 20 times by 12 different speakers (5 females and 7 males). The total duration of recordings was about 100 hours. Finally the list of voice commands has been supplemented with the names of 83 diseases, 92 complaints and 94 drug names without new recordings. In this way the total list of voice commands that speech recognizer is trained and is able to recognize is 1000.

IV. OPTIMIZATION OF SPANISH RECOGNIZER

In [9] we presented the basic results showing that adapted Spanish recognizer and proprietary Lithuanian recognizer provided uncorrelated results. The recognition accuracy of adapted Spanish recognizer wasn’t high. This chapter presents some issues on optimization of Spanish recognizer. Several transcription selection criteria were evaluated: a) synthesis of transcribed Lithuanian voice commands with Spanish synthesizer and selecting transcription providing the best sounding realization; b) empirical selection of best transcriptions; c) intuitive selection. In the case a) investigator writes Lithuanian command using Spanish grammar rules and then feeds the transcribed command to the synthesizer. If the synthesized command sounds naturally enough, the transcription is used for the recognition. The empirical selection is based on the fact that foreign recognizers have lexicon design tools. E.g. for Lithuanian command "viduriavimas" the tool provided following set of possible Spanish transcriptions (called PRON transcriptions):

```
B I D U DX J A B I M A S
B I D U DX J A B J M A S
B J D U DX J A B I M A S
B I D W DX J A B I M A S
B I D U DX I A B I M A S
```

Intuitive transcription selection method is based on analogies from other commands or other heuristic methods. The optimal transcriptions selection process is very time consuming process. For example we are giving the procedure of optimization of command "nemiga". Initial transcription enabled to achieve only the 28.7 % accuracy rate. Lexicon design tool generated 6 transcription candidates (N E M J G A, N E M I G A, N E M I K A, N A M I K A, N A M I G A, N I A M I K A), while using intuitive approach 6 more transcriptions were obtained too ((nemiga, niamiga, niamige, namiga, naamiga, namika, namijka, niamika)). Analysing all 14 transcriptions was found template which enabled to achieve the highest recognition accuracy of 79.6 % for this command. Table I shows the recognition accuracy of some commands.

TABLE I. RECOGNITION ACCURACY OF SEVERAL VOICE COMMANDS USING ADAPTED SPANISH RECOGNIZER BEFORE AND AFTER OPTIMIZATION.

Command	Accuracy, %	
	Prior optimization	After optimization
zemas kraujo spaudimas	0.0	10.0
gelta	18.7	80.8
roZine	10.4	71.2
tolura	49.2	50.4
plaviksas	11.2	20.0

These results allows to make conclusion that for the middle size vocabulary (more than 500 commands) recognition tasks there are no universal method to find the best transcriptions since there is necessary to evaluate many possible candidates and good result will not be guaranteed. But properly evaluating lexical and grammatical constraints recognition of a limited vocabulary could provide to relatively good accuracy level.

V. Training and Evaluation of Proprietary Lithuanian Recognizer and Realization of Recognition System

Proprietary Lithuanian speech recognizer is based on CD-HMM model. Its basic version uses triphones as a basic speech element to model acoustic events occurring in speech utterance. Gaussian mixtures are used to model probabilities of particular acoustic events. Acoustic properties were described using MFCC features. Viterbi search algorithm was used as a basis for the decoding procedure to find the most likely sequences of acoustic events.

First acoustic models were obtained using earlier collected speech corpora (about 35 hours of recordings) which content wasn't related with the medical-pharmaceutical topics. Later models were retrained using the specially designed speech corpora described above. Data from the MEDIC corpora was used to evaluate the efficiency of recognizer too but seeking to model speaker – independent mode the recordings of the tested speaker weren't used during evaluation. Table II shows the recognition accuracy of the 731 commands from speech corpora averaged per all speakers.

TABLE II. RECOGNITION ACCURACY OF 731 VOICE COMMANDS FROM MEDICAL-PHARMACEUTICAL CORPORA.

Corpora, (commands)	Correct, %	Misrecognitions, %
Diseases (217)	99.65	0.35
Complaints (208)	99.64	0.36
Drugs (306)	98.24	1.76
All (731)	98.92	1.08

Similar experiments were carried on with 1000 command set including those voice commands that weren't specially included to the corpora. In this case overall recognition accuracy for 1000 commands was 98.83 %. This is minor decrease comparing with the case when only 731 commands were used. It is also important that similar slight accuracy decrease has been observed for each speaker used to test the accuracy but there were no situations when some particular speaker saw significant performance degradation.

Finally both recognition approaches were combined into the single hybrid recognizer. To make the final decision rule induction algorithm Ripper was applied [15]–[16]. The detailed description of combination approach and the rules used to combine the output of two recognizers could be found in [17]. Each object in the training set (set of recognizer output parameters) has been described using seventy features Among those features are such parameters as confidence of the result provided by SP recognizer, average log probability of the LT recognizer hypothesis, proportion and likelihood of all sounds present in the hypothesis produced by both recognizers and some other parameters (such as gender probability, silence probability at the start and the end of the utterance, etc.). Hybrid decision rule efficiency was evaluated standard cross checking procedure: data from 11 speakers were used to derive the rule while 12th speaker was used to check efficiency. Later the results of all speakers were averaged. Experiments evaluation showed that Ripper rule operates with 97.85 % ± 2.30 % accuracy. Since the logical rule is invoked only when outputs of recognizers differs we can conclude that overall accuracy of hybrid recognizer should be 98,92 %. Such accuracy for 1000 commands speaker independent task could be treated as a very high and has been never achieved for Lithuanian speaker-independent recognition tasks using the vocabularies of similar size.

Further we will provide some insights into practical realization of hybrid recognition system.

We have developed a voice controlled web service based prototype, mainly targeted at servicing medical personnel, though in principle compatible with other business institutions working in Pharmacy and health industries. Our application model can be described as a set of internet objects and functions to access the remote based service operations. All calculations are done on a server side and our users are presented with an HTML5 based frontend compatible with any modern browsers and devices, including Android based phones and tablets. This client-server principle allows us full control, support and improvement of speech recognition processes, as well as reducing the calculations on a client device, as speech processing is very computationally intensive. The web service itself was developed using NET4 WCF libraries and is compatible with industry standard applications.

The main recognition process can be explained in three steps:

1. A user pronounces a voice prompt using his device of choice, thus a sound recording is produced and sent to our server for further processing. To achieve high availability we are able to do this even using any HTML5 compatible browser.
2. As soon as the server receives the recorded audio file, signal processing components are activated and further passed to our proprietary speech recognizer, which then continues the recognition process and produces the possible semantic meaning and probability of a recognized answer. To improve the recognition quality we use a hybrid approach by running two parallel speech engines [17], a commercial, based on foreign acoustic models and a proprietary build specifically for the Lithuanian language, each with its own advantages and disadvantages.
3. If the produced probability of a recognized utterance is high enough, a response is generated and sent back via encrypted string back to the client device (application or web script) and is then further shown on screen, passed for further application steps, or even pronounced using a proprietary Lithuanian TTS.

VI. CONCLUSIONS

The medical information system using voice command recognition is presented. One of the basic properties of the system is the implementation of a hybrid approach: both proprietary Lithuanian recognizer and adapted foreign language recognizer are used in the system and the final decision is made combining hypotheses from both recognizers. The combination is done using a set of logical rules derived with the help of automatic logical rules construction algorithm Ripper. All these measures together allowed achieve very high speaker-independent recognition accuracy of 1000 medical-pharmaceutical voice commands (98.9 %). The achieved recognition accuracy completely satisfies the needs and developed system forms good basis for its integration into medical information systems used in practice.

REFERENCES

[1] Speech Recognition 2013: Going from Back to Front and Beyond. *KLAS Report*. [Online]. Available: http://www.klasresearch.com/ News/ PressRoom/2013/speechrec

[2] S. Deschennes, “5 Financial Benefits of Voice Recognition Technology”, *Healthcare Finance News*, 2012.

[3] D. Carr, “Voice recognition speeds HER use for Oklahoma hospital”, *Information Week Healthcare*, 2014. [Online]. Available: http://www.informationweek.com/healthcare/mobile-and-wireless/voice-recognition-speeds-ehr-use-for-oklahoma-hospital/d/d-id/1113302

[4] J. Rowe, “3 reasons you’ll need voice recognition for better patient engagement”, *Healthcare IT News*, 2013. [Online]. Available: http://www.healthcareitnews.com/news/3-reasons-youll-need-voice-recognition-better-patient-engagement

[5] K. Ananthkrishnan, A. Hashemi, A. Barnes, S. Bailes, “Performance of speaker-independent speech recognition of Australian English”, in *Proc. 11th Australian Int. Conf. on Speech Science & Technologies*, Auckland, 2006, pp. 494–499.

[6] T. Sledevic, G. Tamulevicius, D. Navakauskas, “Upgrading FPGA implementation of isolated word recognition system for a real-time operation”, *Elektronika ir Elektrotechnika*, vol. 19, no 10, pp. 123–128, 2013.

[7] L. Deng, “Front-end, back-end, and hybrid techniques for noise-robust speech recognition”, in *Robust Speech Recognition of Uncertain or Missing Data*, Springer, 2011, pp. 67–99. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21317-5_4

[8] R. Maskeliunas, A. Rudzionis, V. Rudzionis, “Advances on the use of the foreign language recognizer”, in *Proc. Second (COST 2012) Workshop, Dublin, Lecture Notes on Computer Sciences*, 2010, pp. 217–224.

[9] V. Rudzionis, G. Raskinis, R. Maskeliunas, A. Rudzionis, K. Ratkevicius, “Comparative analysis of adapted foreign language and native Lithuanian speech recognizers for voice user interface”, *Elektronika ir Elektrotechnika*, vol. 19, no. 7, pp. 90–93, 2013.

[10] H. Lin, Li Deng, D. Yu, Y. Gong, A. Acero, C.-H. Lee, “A Study on multilingual acoustic modeling for large vocabulary ASR”, in *Proc. IEEE (ICASSP 2009)*, Taipei, 2009, pp. 4333–4336.

[11] L. Burget *et al.*, “Multilingual acoustic modelling for speech recognition based on subspace Gaussian mixture models”, in *Proc. IEEE (ICASSP 2010), Dalas*, pp. 4334–4337, 2010.

[12] S Thomas, S. Ganapathy, H. Hermansky. “Multiligual MLP features for low-resource LVCSR systems”, in *Proc. IEEE (ICASSP 2012), Kyoto*, pp. 4269–4272, 2012.

[13] J. Huang, J. Li, D. Yu, L. Deng, Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers”, in *IEEE Proc. (ICASSP 2013), Vancouver*, pp. 7304–7308, 2013.

[14] L. Rabiner, “A tutorial on hidden Markov models on selected applications in speech recognition”, in *Proc. IEEE*, vol. 77, no 2, 1989, pp. 257–286. [Online]. Available: http://dx.doi.org/ 10.1109/5.18626

[15] W. Cohen, “Fast effective rule induction”, in *Proc. Twelfth Int. Conf. Machine Learning*, 1995, pp. 115–123.

[16] G. Pappa, A. Freitas, “Automatically evolving rule induction algorithms”, in *Machine Learning (ECML 2006), Lecture Notes in Computer Science*, vol. 212, pp. 341–352, 2006.

[17] V. Rudzionis, K. Ratkevicius, A. Rudzionis, G. Raskinis, R. Maskeliunas, “Recognition of voice commands using hybrid approach”, in *Proc. 19th Int. Conf. ICIST 2013*, Kaunas, Lithuania, 2013, pp. 186–197.