Comparative Analysis of Adapted Foreign Language and Native Lithuanian Speech Recognizers for Voice User Interface

V. Rudzionis[1], G. Raskinis[2], R. Maskeliunas[3], A. Rudzionis[3], K. Ratkevicius[3]
[1]*Department of Informatics, Vilnius University Kaunas Faculty of Humanities, Muitines St. 8, Kaunas, Lithuania*
[2]*Faculty of Informatics, Vytautas Magnus University, Vileikos St. 8, Kaunas, Lithuania*
[3]*Faculty of Informatics, Kaunas University of Technology, Studentu St. 65-108, LT-51369 Kaunas, Lithuania, kastytis.ratkevicius@ktu.lt*

***Abstract*—Paper presents research results obtained when building a speaker independent hybrid speech recognizer. This recognizer will be integrated as a phrase recognizer in a medical-pharmaceutical information system. The hybrid speech recognizer consists of two recognition components: an adapted commercial Microsoft Spanish speech recognizer and a locally developed hidden Markov models based recognizer implementing Lithuanian acoustic models. Efficiency of both recognition components was evaluated on multiple speaker independent speech recognition tasks. The average accuracy of Lithuanian recognizer was higher reaching 0.6% phrase error rate for user requests in medical-pharmaceutical domain. The adapted commercial Spanish speech recognizer showed the ability to improve the accuracy of Lithuanian recognizer in the worst recognition scenarios. These results proved the hypothesis formulated when proposing the basic idea of hybrid recognition approach: recognition errors from different recognizers built using various techniques are not strongly correlated. This fact could be exploited for improved overall speech recognition accuracy.**

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

I. INTRODUCTION

During last 15 years speech technologies (recognition, synthesis and identification) became an integral part of industrial applications of information technologies in various areas of human activities [1]. The variety of the areas where speech processing technologies are applied grows every year. One of the most perspective areas for the application of speech technologies is the healthcare industry. The main rationale for the application of voice processing in the healthcare is the desire to save the time of highly qualified medical personnel that is routinely spent on operations of documentation as well as the desire to speed up and to ease the information search and presentation. In this way, time spent on documentation and other trivial tasks could be allocated for the tasks requiring higher qualification. The biggest savings are achieved by using speech recognition technology to fill-in medical documents or parts of documents by voice. If the speech recognition is combined with modern means of communication (computer, internet and telephony), entirely new possibilities to perform medical documentation anytime and anywhere may occur.

It is estimated that about half a million of health care practitioners world wide have the possibility of using speech technologies in their daily activities [2]. Among well known examples of applications in medicine is a voice controlled electronic health records management system Epic. It is provided by one of the biggest healthcare service providers in Northern America – Providence Health and Services – and installed in 27 hospitals and more than 250 clinics. One of the leading companies in speech processing technologies – Nuance Communications – began to distribute their product called Dragon Medical Mobile Dictation in 2012. This product is aimed at transcribing remarks, comments, and notes about patient conditions dictated by voice and using mobile devices. Another example illustrating the potential of speech processing applications in healthcare industry is another product of Nuance called Dragon Medical 360, which is intended for further processing and proper management (placing to the proper location and files) of medical comments recorded by voice.

Modern voice user interface implementing systems are complex systems incorporating achievements in many areas of research and technology: telecommunications, signal processing, microelectronics and information technologies. Most of these technologies could be easily moved to Lithuania but some not: main prerequisite to implement such systems in Lithuania is the development of speech recognition engine of sufficient reliability for the recognition of Lithuanian medical terms and other related voice commands and phrases. Looking from the user's perspective, the system is better if it is universal and has a

Manuscript received December 22, 2013; accepted April 17, 2013.
This research was funded by a research project "Hybrid recognition technology for voice interface" (INFOBALSAS) funded by "High technology development program 2011 - 2013" of ASIT.

flexible vocabulary. However, in order to evaluate the feasibility of making a system that could be useful for healthcare practitioners in Lithuania, some vocabulary restrictions should be applied.

To achieve this goal the decision was made to develop a hybrid speech recognition system for Lithuanian medical and pharmaceutical terms. We understand the term "hybrid system" as the combination of two speech recognition engines: an adapted foreign language recognition engine and a proprietary Lithuanian speech recognition engine. It should be noted that many state-of-the-art speech recognition systems implements several speech recognizers [3], [4]. But in most of these cases are used basically the same recognizers trained on different features or having slightly different structure. Here we are using two entirely different approaches to supplement each other. The inclusion of adapted foreign language speech engine is aimed at exploiting the elements of well developed acoustic models of foreign languages thus easing and speeding up the development of a voice recognition system of Lithuanian voice commands. The development and inclusion of a proprietary Lithuanian speech recognition engine is aimed at improving the recognition accuracy of voice commands that may be poorly recognized by an adapted foreign language engine. Our earlier investigations suggested the use of Microsoft's Spanish speech recognition engine as the foreign engine for the adaptation purposes. The continuous density hidden Markov models (CD-HMM) based recognizer was used as the preferred option for the proprietary Lithuanian speech recognizer.

The rest of the paper is organized as follows. Chapter 2 presents the speech corpora used in the development of Lithuanian medical information system. Chapter 3 presents some experimental results investigating the adapted foreign language recognizer. In Chapter 4, we present some experimental results in developing and investigating the proprietary Lithuanian speech engine.

II. SPEECH CORPORA

One of the key factors in developing a robust and efficient speech recognition system is the employment of proper speech corpus. The main purpose of the corpus is to provide acoustic-phonetic material (recordings) for the training process, i.e. for finding the parameters of acoustical models. Speech corpus should comply with a wide range of requirements. Few of them we will mention explicitly:

- corpus should be as good as possible in representing acoustic-phonetic content which will be used by the system;
- corpus should be as good as possible in covering the variety of speakers which will use the system.

In practice, compromises should be made. For the development of our system, it was decided to collect the specialized corpus containing the names of diseases, patient complaints which are most often met in medical practice, and the names of the most frequent prescription drugs. On the other hand, it was decided to use speech resources already at our possession, i.e. earlier collected speech corpora (containing about 50 hours of speech) to train acoustic models. Since available speech corpora were presented in other papers [5], we will present details about specialized medical and pharmaceutical speech corpus here.

The large part of medical terms used in practice by healthcare professionals is contained in the official list of diseases and disorders that is approved by the Ministry of Health. This list contains 14179 diseases and disorders. It is composed of more than 88000 lexical tokens (not all are of medical origin) and has 10955 unique lexical types. It may be surprising that specific medical terms are used much less often than general terms (e.g. switches). 5991 lexical types cover 75 percent of the whole list. This analysis showed that it is unfeasible to develop a full scale medical Lithuanian recognition system in short time. At the same time, not all diseases or drugs are used equally often. A big part of the daily voice requests could be successfully handled using a relatively small number of voice commands. In collaboration with industrial partners who have the expertise in developing IS for medical professionals we selected the 631 diseases names, complaints and drug names containing 777 lexical tokens. This list represents the most frequently used medical terms in Lithuania. Each voice command has been recorded by 7 male speakers and 5 female speakers in laboratory conditions. Every speaker pronounced each voice command 20 times. Full semi-automatic and manual validation with respect to the completeness and correctness has been performed. The size of the medical speech corpus is about 100 hours.

III. RECOGNITION EXPERIMENTS USING ADAPTED FOREIGN LANGUAGE ENGINE

The rationale behind the adaptation of a foreign language speech recognition engine for the recognition of Lithuanian voice commands is the assumption that acoustic-phonetic properties of one language are partially present in another language. In this case, the aim of training is to find the best mapping between Lithuanian phonetic units and acoustic models of phonetic units of a foreign language. It has been shown that such mapping allows the achievement of acceptable recognition accuracy for a large group of applications. The Spanish speech recognition engine was used since our earlier investigations showed that Spanish has the closest acoustic similarity to Lithuanian among commercially available recognizers [6].

The first task in the adaptation procedure is to find the proper method of mapping between the phonetic units in one language and acoustic models in another. The initial comparison and evaluation of various mapping methods has been done using Lithuanian corpus of family names. These experiments showed that an average recognition error rate of 6.67% is achieved when simplest transcriptions of Lithuanian words (just slightly differing from their orthographic representation) are used. Optimization of the mapping of the worst recognized words has decreased the error rate to about 5%.

More complicated mapping methods were investigated using the larger corpus of Lithuanian digit names. In this case, 5 different mapping methods were applied: M5 – the intuitive mapping method used in the experiments presented above, and the M1-M4 mapping methods that were set up

during several experimental studies. Table I shows the recognition word error rate. Small corpus CORP11 (utterances of 11 speakers) was used to optimize the transcriptions for different methodologies. The corpus DIGITS10 (utterances of 100 speakers) was used to get more robust results. The results showed that small corpus was unable to get optimized results and led to improvement of the straightforward method M5.

TABLE I. THE RECOGNITION ACCURACY (WORD ERROR RATE IN PERCENT) FOR DIFFERENT MAPPING METHODS.

Corpora	Mapping method				
	M1	M2	M3	M4	M5
CORP11 (10 speakers)	6.45	5.30	6.10	5.35	8.55
DIGITS10 (100 speakers)	8.28	9.40	9.53	12.7	17.9

The simplest method of transcription mapping was also used for the evaluation of recognition accuracy of medical and pharmaceutical terms using the part DISEASES (names of 217 diseases only) of the corpus MEDIC (631 possible commands). Here, the recordings of 11 speakers were used. This means that 4340 utterances were recognized for each speaker. The system was allowed to select its response among 631 possible commands of MEDIC.

TABLE II. THE RECOGNITION ACCURACY (WORD ERROR RATE IN PERCENT) USING THE NAMES OF DISEASES.

	Speakers			
	Average	Best speaker	Worst speaker	Average of 3 worst speakers
WER, %	14.5	8.53	30.2	24.0

It could be seen that the average error rate was 14.5%. But more careful analysis showed that utterances of three speakers resulted in a significantly higher error rate in comparison to the rest. If the system was restrained to select its response among the alternatives in the DISEASES group (217 options) then the word error rate dropped down to 8.59%. It should be noted that about 50% of all commands were recognized with the recognition accuracy over 90% and only a small number of voice commands were recognized very poorly. Table III lists 13 the most poorly recognized voice commands (the number of times when command was correctly recognized among 220 utterances).

TABLE III. LIST OF WORST RECOGNIZED COMMANDS FROM THE DISEASE NAMES SET.

Command	Number of correct answers
rožė	2
rožinė	27
rožiniai spuogai	51
vidurinės ausies uždegimas	59
rožinė dedervinė	77
egzema	81
venų varikozė	96
hemofilija	101
gingivitas	108
šlapimo nelaikymas	108
rinitas	109
juostinė pūslelinė	110
nugaros skausmas	110

The results presented above could be interpreted in various ways. The overall recognition accuracy does not meet the requirements usually set for the practical applications using speech recognition. On the other hand, these results weren‘t optimized and were achieved using intuitive mapping. The experience in developing other applications suggests that it is possible to increase the recognition accuracy by at least 30% through the optimization and selection of transcriptions. On these assumptions we could expect that the overall recognition accuracy using foreign language speech recognition engine with adapted transcriptions alone could achieve more than 90% and become close to the general requirements for the implementation of voice user interface in practice. Another observation is that some commands and some speakers make too many mistakes to be expected to achieve appropriate performance. This shows that the application of the proprietary Lithuanian recognizer and the implementation of hybrid approach are necessary: usually proprietary speech recognizers perform better than the adapted or retrained using foreign language recognizers. The first step in building hybrid recognizer is to evaluate the performance of a proprietary Lithuanian recognizer for such task and to find if it‘s performance may be considered as the supplementary to the performance of an adapted foreign language speech recognizer (recognition errors aren‘t completely correlated).

IV. RECOGNITION EXPERIMENTS USING LITHUANIAN SPEECH RECOGNIZER

For the development of purely Lithuanian speech recognizer as a component of hybrid recognition technology well known continuous density hidden Markov models (CD-HMM) approach has been used. The selection of the CD-HMM approach was caused by the fact that this technology is most often used in speech recognition systems worldwide. Despite some of the well known drawbacks it still provides the best recognition results with respect to other approaches. The core of the Lithuanian recognizer is HMM models of Lithuanian phonetic units which are trained in offline mode. Experiments described in this section are based on the acoustic models that were obtained/trained on a previously collected general purpose corpus of read speech (recordings of about 50 hours). Parts of the medical and pharmaceutical corpus MEDIC were used for the evaluation purposes, though MEDIC is to be used for training and improving initial acoustic models later in our research.

Thus, the first version of Lithuanian recognizer was obtained using recordings of 50 speakers (of them 25 male and 25 female speakers). Each speaker read Lithuanian text for about 1 hour. The speech was fully annotated, i.e. the sequence of phonetic units that matched the read text was known. This sequence also included the knowledge of word stresses and some non speech phenomena (pauses, noise).

The training process could be briefly summarized as follows. First of all, speech recordings were transformed to the feature vectors. The recordings were sampled at 16 kHz, split into 20 ms duration frames using 10 ms displacement. Then 12 mel-frequency cepstrum coefficients (MFCC) were obtained out of 26 outputs of triangular filters linearly spaced in the nonlinear mel-frequency scale using discrete

cosine transform. In addition to 12 cepstrum coefficients also 12 first and second order cepstrum change rate coefficients (derivatives) were calculated. These coefficients together with energy, first and second order energy derivatives were packed to a single vector of 39 coefficients for each 20ms signal frame. For the description of acoustic units the set of 141 monophones were used [7], e.g. diphthongs, affricates, palatalized and non-palatalized consonants, accented and non-accented vowels were modeled as distinct phonetic units. Lithuanian HMM speech recognizer was trained following the sequence of processing steps described in other references. To model acoustic changes HMM models had 1-5 states. Most of the HMM prototypes had the topology generally known as left-to-right without skips. Near 75000 triphone models represented by about 40000 physical different HMMs were obtained as a result of training.

For recognition purposes, each voice command was described as a sequence of triphones. For instance, Lithuanian command/word „aštuoni“ (eight) was represented by two possible pronunciations (corresponding to different accentuation patterns) each having 6 triphones:

```
aštuoni  sil-a+S a-S+t S-t+"uo t-"uo+n' "uo-n'+i
n'-i+sil
aštuoni  sil-a+S a-S+t S-t+uo t-uo+n' uo-n'+"i n'-
"i+sil
```

Then Lithuanian recognizer was optimized and tested for the tasks that are characteristic for the medical information systems using speech data from other corpora. The results are summarized in Table IV. It is important to note that the total amount of the test recordings was about 35 hours and none of the test recordings were used for training.

TABLE IV. THE RECOGNITION ACCURACY (WORD ERROR RATE IN PERCENT).

The title and a brief description of an experiment	Word error rate, %
DIGIT10: recognition of digit names 0-9 (among 10 possible options).	1.08
DIGIT100: recognition of digit names 0-9 (among 10 digit names and 90 the most frequent Lithuanian words)	18.22
DIGIT1000: recognition of digit names 0-9 (among 10 digit names and 990 the most frequent Lithuanian words).	38.26
SYLLAB: recognition of 66 open syllables from the list of 66 possible syllables	57.61
NAMES1: recognition of name-surname pairs from 100 possible name-surname pairs	0.10
NAMES2: recognition of name-surname pairs when names and surnames are selected from two separate lists of names and surnames	3.94
MEDIC: recognition of voice commands (names of diseases, disorders, drugs, patient complaints) from the list of 631 commands	0.60

The results above suggest several remarks. It could be seen that short voice commands are difficult to discriminate even if they are not numerous (experiment SYLLAB) while long voice commands are recognized with better accuracy. When the number of possible voice commands grows up the recognition accuracy goes down since more and more confusable pairs appears in the vocabulary (experiments DIGITx): digit name “devyni” (nine) often confused with the word “tėvynė” (fatherland) in DIGIT1000 experiment.

Hybrid approach may be useful if at least one of the recognizers provides the correct answer and it is possible to associate the correct answer with a higher degree of confidence. Since the recognition accuracy of Lithuanian recognizer was higher in average than that of the adapted Spanish recognizer, there was an analysis made to see if the errors of Lithuanian recognizer could be corrected using the adapted Spanish recognizer. A few observations confirmed this intuition. In the experiment DIGIT10 was observed that out of 244 total errors in recognizing digit names 106 errors were related to the digit “trys”. All these misrecognitions were given to the adapted Spanish recognizer and 93 out of 106 utterances were recognized correctly. In the second case (corpora MEDIC) Lithuanian recognizer made 236 errors (out of 44560 utterances). 8 worst recognized commands (135 falsely recognized utterances) were presented to the Spanish recognizer which reduced error rate nearly twice (to 70 errors). These observations allow us to conclude that hybrid approach may be useful in developing VUI based systems. Various methods to select the recognizer could be used, e.g. acoustic confusability measure based on Levenshtein distance could be applied.

V. CONCLUSIONS

The medical information system using voice command recognition is presented. The key element of the system is the implementation of a hybrid approach: both proprietary Lithuanian recognizer and adapted foreign language recognizer will be used in the system. The experiments showed that Lithuanian recognizer outperformed an adapted foreign language recognizer even if it was trained on a general purpose speech corpus. On the other hand, error analysis showed that errors made by different recognizers aren’t exactly correlated and the performance of one recognizer may be used for the improvement of performance of the other recognizer. This shows that the hybrid approach could be a promising technique in developing practical voice user interface based systems.

REFERENCES

[1] D. Suendermann, R. Pieraccini, *Spoken Language Understanding*, John Wiley & Sons, Ltd., pp. 171–194, 2011. [Online]. Available: http://dx.doi.org/10.1002/9781119992691.ch7

[2] Nuance Healthcare Inks Deal with Providence Health & Services. *Speech Technology Magazine*. [Online]. Available: http://www.speechtechmag.com/Articles/PrintArticle.aspx?ArticleID=83918

[3] G. Tur, A. Stolcke, *et al.*, “The CALO Meeting Speech Recognition and Understanding System”, in *Proc. of IEEE Spoken Language Technology Workshop*, 2008, pp. 69–72.

[4] G. Saon, J. T. Chien, “Large-Vocabulary Continuous Speech Recognition Systems: A Look at Some Recent Advances”, *IEEE Signal Processing Magazine*, vol. 29, no. 6, 2012, pp. 18–33 [Online]. Available: http://dx.doi.org/10.1109/MSP.2012.2197156

[5] A. Vaičiūnas. “Statistical Language Models of Lithuanian and their Application to Very Large Vocabulary Continuous Speech Recognition”, Ph.D. dissertation, Vytautas Magnus University, Kaunas, 2006

[6] R. Maskeliūnas, K. Ratkevičius, V. Rudžionis, “Voice-based Human-Machine Interaction Modeling for Automated Information Services”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)*, no. 4, pp. 109–115, 2011.

[7] A. Raškinis, G. Raškinis, A. Kazlauskienė, “SAMPA (Speech Assessment Methods Phonetic Alphabet) for Encoding Transcriptions of Lithuanian Speech Corpora”, *Information Technology and Control,* vol. 29, no. 4, pp. 50–56, 2003.