Some Aspects of Voice User Interfaces Development for Internet and Computer Control Applications

R. Maskeliunas[1], K. Ratkevicius[1], V. Rudzionis[2]
[1]*Speech Research Laboratory, Kaunas University of Technology,*
Studentų St. 65, LT–51369, Kaunas, Lithuania, phone: +370 37 300600
[2]*Department of Informatics, Kaunas Humanities Faculty of Vilnius University,*
Muitines St. 8, LT–44280 Kaunas, Lithuania, phone: +370 37 422566
rytis.maskeliunas@ktu.lt

***Abstract*—This paper deals with some aspects of development voice user interfaces for several applications. The realistic scenario when developing VUI for such languages as Lithuanian is adaptation of foreign language speech recognizer. But not all voice commands could be recognized successfully enough using adapted recognizer. Hence the importance to implement hybrid recognizer arises. Several practical applications using Lithuanian voice command recognitions are presented. The experimental investigation showed that 90 percent recognition accuracy was achieved in average using adaptation of foreign language speech engine. Detailed analysis showed that most commands are recognized with very high accuracy. For those commands that aren't recognized accurately enough hybrid recognition principle may be applied.**

***Index Terms*—Speech processing, speech analysis, audio user interfaces, lithuanian language applications.**

I. INTRODUCTION

Voice user interfaces has a series of advantages for the control of various applications. The speech is often the most natural, easiest to use and the most convenient way of human-machine interaction. The advantages of voice user interfaces were described and shown in various papers such as [1].

It became some sort of mantra to talk that voice controlled user interfaces more and more often are implemented in various applications. The one big problem with such statements is that being largely true they dealt mainly with the implementations that were oriented to the special areas of application and were virtually unknown for the wider audience. But one application must be mentioned in particular. Despite that it was presented only several months ago it became so widely talked about and showed so much promises for the so large number of potential customers that we could even say that that this application gave new boost for the entire speech recognition solutions implementing industry. This is personal digital assistant Siri introduced by Apple as an integral part of iPhone4 series. As was stated by some independent business and market analysts Siri is already the best voice recognition application in history [2]. The success of Siri is caused by three main components: improved speech recognizer, improved natural language modeller and as semantic analyser. These three components enabled the implementation of voice user interface with the flexibility never seen before in practical applications available to the general public. It should be emphasized that speech recognition accuracy is the first and basic element of such types of systems since only good enough voice recognition may to allow implement semantic analysers.

The history of the development of such flexible voice recognition using systems as Siri shows how complicated task are facing the researchers in this field: Siri was the outcome of the very big and very expensive research project which was carried for more than five years, grants were provided by US DoD DARPA agency, project cost was about 200 million USD, it involved more than 300 researchers from about 25 leading universities and private research laboratories.

The extent of these projects raised the necessity to rethink the principles and strategy when developing voice user interfaces were Lithuanian language is used as a primary means of communication. Very important characteristic of voice based interfaces is the dependability of the phonetic, syntactic and lexical properties of the language spoken by the user. This means that it is impossible to move technologies developed for the recognition of one language to the recognition of another automatically. At least some sort of adaptation would be necessary. It is obvious that projects of such scale can't be carried on in Lithuania so we need different strategy to achieve the use of Lithuanian in modern information systems with VUI. Two possible solutions could be proposed in Lithuanian case:

1) To use carefully designed guided human-machine dialogues to avoid or to minimize impact of semantic variability;
2) To adapt foreign language speech recognition engines to speed up development of applications and to

Manuscript received March 12, 2012; accepted May 13, 2012.
Parts of this work were done under research project No.: 20101216-90 funded by EU SF project "Postdoctoral Fellowship Implementation in Lithuania" (VP1-3.1-ŠMM-01).

reduce the costs of the development.

One of the possible solutions for some class of applications is the adaptation of foreign language based speech engines via the selection of proper phonetic transcriptions. In our previous studies the advantages of such method and its possible uses were established [3], [4]. But this approach has some limitations too: not all words needed for the application and not all phonetic units could be adapted successfully enough and they require another solution. In this case hybrid recognition principle could be used which will be discussed below.

II. Importance of Hybrid Recognition Approach

Voice command recognition is the crucial and most important step in the development of voice user interfaces. Despite the significant progress in the field still the best speech recognition systems can't achieve human recognition accuracy.

Seeking to improve the voice command recognition accuracy state-of-the-art recognizers implements various techniques. At the front-end these techniques include vocal tract length normalization, cepstral feature normalization, linear discriminant feature transforms, and non-linear discriminant transforms affected by multilayer perceptrons. Hidden Markov model acoustic models based on clustered Gaussian mixtures are trained using discriminative criteria such as minimum phone error and a related feature-level transform. Feature transforms are also used to bridge differences in signal bandwidth between the background and target data [5]. Particularly important is that all state-of-the-art systems works in the batch mode, recognizing speech or voice commands multiple times for the purpose of unsupervised acoustic adaptation (e.g. using maximum likelihood linear regression) and also for the purpose of combining multiple hypothesis streams, often based on subsystems that differ in the features or models used so as to generate complementary information [6]. For example, a system might recognize speech based on both Mel cepstral coefficients and perceptual linear prediction cepstrum, and combine the result. In other words all state-of-the-art speech recognition systems are using hybrid recognition approach: implementing several recognizers operating in parallel and combining the results of several recognizers to get the final integral decision. So it is hard to expect achieve the appropriate recognition results using single recognizer as experience of various researchers in the field suggests.

When using adaptation of foreign language commercial speech recognizer to recognize Lithuanian voice commands hybrid recognition will be slightly different from those approaches described above. Typical drawbacks of ASR engines are very clear in noisy environments and in the case of disturbances and variations not defined in a model used when training the recognizer: the most problematic factor remains the discrimination of acoustically similarly sounding words. We expect to overcome some of these limitations by utilizing a hybrid system capable of combining "the best of two worlds". Our method consists of the adaptation of traditional commercial CD-HMM as a basis engine for the recognition of most voice commands and a proprietary engine for the discrimination of relatively poorly recognized voice commands, capable of utilizing an appropriate algorithms. .Hybrid recognition methods of such type are still not utilized in the current applications with VUI, though the potential and benefits are quite obvious. These benefits are clear also while developing a recognizer from a scratch and especially while integrating and disseminating the results into an already used standardized IT systems. The methodic allowing the selection of the more poorly recognized voice commands from a provided dictionary still does not exist and the future research should provide insights into it. From other point of view hybrid methods are well suited for the integration into parallel, cloud and specialized hardware using voice recognition systems [7].

III. Some Perspective Lithuanian Speech Recognition Application Examples

Analysis of the development of various voice user interfaces using applications in other countries and some examples of the good practice in other countries we are developing several applications for the call centre automation, control of Internet browser by voice and filling of Internet forms by voice. The selection of these areas of application was done after analysing the possibilities to achieve the stage of implementation (it means the feasibility of the application) and the potential of user benefits that proposed system may achieve.

The main requirements for the applications under development are:

1) To develop Lithuanian speaker independent spoken digit recognition with more than 95% accuracy;

2) To develop proper names recognition system for 500 distinct names and family names;

3) To develop computer and internet control tools implementing voice user interface which achieves commands recognition accuracy of at least 95%.

Below we will briefly describe the basic ideas of the projects under development.

Automated call centre. This e-service is targeted at various commercial and governmental institutions. A user of this application can call a provided phone number(s) and say by voice (or by entering text using a smartphone, or by DTMF using a regular phone) the name and surname of the person he is looking for and (if necessary) the office (title) or the department. An automated system recognizes the query, checks the databases and responds to the user the following ways: by pronouncing a phone number using a natural speech, by sending a phone number using SMS, by a direct transfer to a person he was looking for. The e-service also has a capability to renew the vocabulary (adding the new names and surnames of the new workers) thus maintaining future functionality.

Control of computer and the internet browser by voice. This e-service is targeted at the internet browsing by voice. It will enable to open or to close the websites, to control the browser by voice commands and to read the selected text by voice. The essential features of the e-service: the possibility to work with any installed in the system speech recognition engine, the possibility to make and to change voice commands and to change reactions to voice commands. The accuracy of 100 Lithuanian voice commands recognition should be more than 95%. The list of voice commands consist of about 50 commands for internet browsing (up, down, home, end, right, left, select all, read etc.) and about 50 commands for internet websites opening. This e-service could be applied for visually impaired people.

Filling of internet forms by voice. This e-service is targeted at the internet forms filling by voice. It will enable to open or to close the websites and to fill by voice the fields in the internet websites, used for searching of information, e-shopping and so on. It has the possibility to input the text by spelling, to input the digits or various marks by voice commands. The accuracy of 100 Lithuanian voice commands recognition should be more than 95%. The list of voice commands consist of about 60 commands for input of symbols (letters, digits, marks), about 30 commands for internet websites opening and about 20 commands for input of text fragments. This e-service should be useful for visually impaired people.

IV. EXPERIMENTAL EVALUATION OF SPEECH RECOGNITION ACCURACY

All the applications mentioned above are different in many aspects (vocabulary, technical realization, etc.). But at least one thing is common: the necessity to recognize digit names as accurately as possible. In our previous experiments we showed that proper selection of phonetic transcriptions enables to achieve high enough recognition accuracy of Lithuanian voice commands using foreign language speech engine.

This study is the continuation of our earlier research. In [8] we showed that Microsoft SAPI could be applied to recognize Lithuanian voice commands. In [4] method to select transcriptions to recognize Lithuanian commands using foreign language speech engine was shown. In [3] very high accuracy of recognition of ten Lithuanian digit names was demonstrated.

The main aim of these experiments was to establish the limits of possibilities to improve the recognition accuracy of Lithuanian voice commands using recognition when whole word is used as a main unit and digits recognition when words are composed from subunits (phonemic units). In the first group of experiments utterances of ten digit names (0-9) were used which were pronounced by 20 speakers and 20 times each. Together the possibilities to adapt English and Spanish recognition engines were investigated in this context. Table I shows the obtained results (EN 7.0 means English engine with ARPAbet-based transcriptions [9], ES 8.0: Microsoft Spanish 8.0 engine with UPS-based transcriptions and comparison with the results obtained using Microsoft Speech Server English and Spanish 9.0 engines with UPS transcriptions obtained earlier [10]).

TABLE I. AVERAGE RECOGNITION ACCURACY OF LITHUANIAN DIGITS USING ENGLISH AND SPANISH ENGINES AND DIFFERENT SETS OF TRANSCRIPTIONS.

Transcription	Recognizer			
	EN 7.0	ES 8.0	EN 9.0	ES 9.0
Words	90.0	84.8	-	-
Subunits	90.2	87.9	77.0	97.0

It could be seen that the accuracy achieved with Microsoft Speech server Spanish engine 9.0 wasn't achieved with the computers using Microsoft Windows with speech engines set for this operating systems. Further optimization is necessary since predefined accuracy threshold of 95% wasn't achieved in these experiments.

Fig. 1 shows the recognition results for each of the digit name used in the experiments.

Another group of experiments was carried on investigating the accuracy of recognition of commands for internet browser, text editor and media player control. In the case of internet browser there were 18 commands and 17 speakers with 20 utterances of each command, in text editor case there were 24 commands while in media player case there were 13 commands. In both cases 16 speakers pronounced each command 20 times. Table II shows the results of experiment.

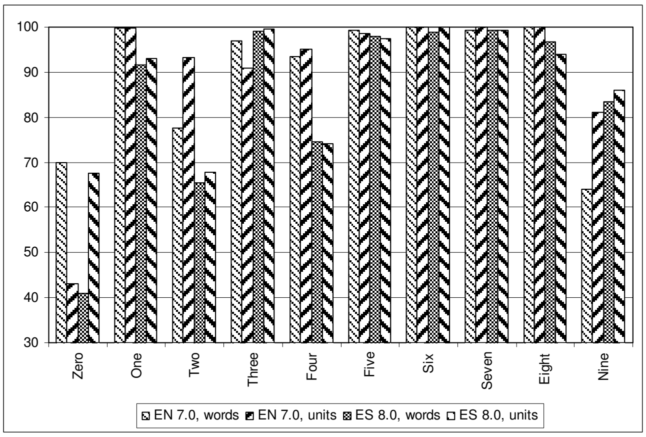

Fig. 1. The recognition accuracy of 10 Lithuanian voice commands using English and Spanish speech engines.

TABLE II. THE RECOGNITION ACCURACY OF LITHUANIAN CONTROL COMMANDS USING ENGLISH AND SPANISH SPEECH ENGINES AND DIFFERENT SETS OF TRANSCRIPTIONS.

Controlled program	EN 7.0		ES 8.0	
	Words	Units	Words	Units
Internet browser	95.1	95.7	96.5	97.4
Microsoft Office Word	92.3	91.7	90.0	92.3
Audio player	86.4	85.2	86.4	87.5
Average	91.3	90.9	91.0	92.4

It could be seen that average recognition accuracy for all control applications was slightly above 90 percent. That is close but still below the defined target (the target has been achieved only for internet browser control by voice applications). The worst performance was obtained for the audio player voice control commands were average recognition accuracy was only about 85-87%. From other point of view these initial results provide the basis for further improvements and make believe to achieve defined accuracy level reasonable. Another interesting observation that both adapted to recognize Lithuanian commands English and Spanish engines performed nearly at the same level of accuracy. This observation goes into contradiction with our earlier studies using Spanish engine from Microsoft Speech server. In this study Spanish engine had clear advantage over English engine. Further investigation is necessary to find the reasons of these phenomena.

Fig. 2 shows the recognition accuracy for each command used in the internet browser control application. Analyzing them you can see that most of the commands were recognized with very high accuracy level (97 -100 %) while the overall average result was degraded by several commands recognized poorly. The recognition of those commands should be investigated further and if necessary proprietary Lithuanian recognizer should be used for them. In this way hybrid recognizer will be implemented in the framework of these applications.

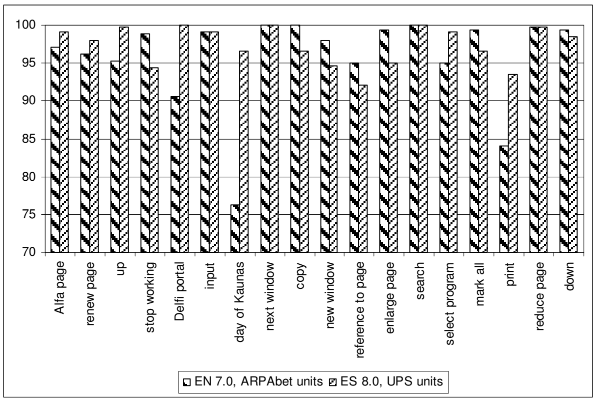

Fig. 2. The recognition accuracy of internet browsing Lithuanian commands using English and Spanish speech engines and different sets of transcriptions.

V. Conclusions

Several practical applications under development using recognition of Lithuanian voice commands were presented. Recognition accuracy of Lithuanian voice commands aiming to implement in these applications was investigated. Adaptation of Microsoft SAPI English and Spanish recognizers allowed achieve recognition accuracy of about 90%. More detailed analysis showed that most commands were recognized with the very high recognition accuracy (98-100%) while several commands were recognized poorly. There are two possible strategies to improve recognition of those commands: to optimize further adaptation or to train proprietary recognizer and apply hybrid approach.

References

[1] A. Rudžionis, K. Ratkevičius, V. Rudžionis, *Voice interactive systems. In The engineering handbook of smart technology for aging, disability and independence*, pp. 281–297.

[2] E. Jackson, “Why Siri is Google Killer”, *Forbes Magazine,* 2011. [Online]. Available: http://www.forbes.com/sites/ericjackson/2011/10/28/why-siri-is-a-google-killer/

[3] R. Maskeliunas, K. Ratkevicius, V. Rudzionis, “Voice-based Human-machine Interaction Modelling for Automated Information Services”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering*), no. 4, pp. 109–112, 2011.

[4] R. Maskeliunas, A. Rudzionis, K. Ratkevicius, V., Rudzionis, “Modelling of Call Services for Public Sector”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering*), no. 4, pp. 85–86, 2010.

[5] T. Hain, et al., “The AMI system for the transcription of speech in meetings”, in *Proc. of the ICASSP,* 2007, pp. 357–360.

[6] G. Tur, et. al., “The CALO Meeting Assistant System”, *IEEE Transactions on Audio, Speech and Language Processing,* 2010, vol. 18, no. 6, pp. 1601–1611. [Online]. Available: http://dx.doi.org/10.1109/TASL.2009.2038810

[7] G. Tamulevičius, V. Arminas, E. Ivanovas, D. Navakauskas, “Hardware Accelerated FPGA Implementation of Lithuanian Isolated Word Recognition System”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering*), no. 3, pp. 57–62, 2010.

[8] A. Rudzionis, K. Ratkevicius, T. Dumbliauskas, V. Rudzionis, “Control of Computer and Electrical Devices by Voice. on the Use of the Foreign Language Recognizer”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering*), no. 6. pp. 16–21, 2008.

[9] D. Jurafsky, J. Martin, *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition,* 2nd ed., Prentice-Hall, 2009, p. 1024.

[10] *Universal Phone Set (UPS).* [Online]. Available: http://msdn.microsoft.com/en-us/library/hh361647.aspx