Speech Server based Lithuanian Voice Commands Recognition

G. Bartisiute[1], K. Ratkevicius[1]
[1]*Speech Research Laboratory, Kaunas University of Technology,*
Studentų St. 65-108, LT-51369, Kaunas, Lithuania, phone: +370 686 96736
kastytis.ratkevicius@ktu.lt

***Abstract*—Paper deals with application of *Microsoft Office Communications Server Speech Server* or *MSS'2007* for Lithuanian voice commands recognition. Voice servers together integrate telephony, speech and internet providing tools for developing applications that run over a telephone. Using of transcriptions of Lithuanian words so far is the only solution of voice servers' application for Lithuanian language. The results of investigation of Lithuanian digit names recognition by German, English, French and Spanish speech recognition engines implemented on MSS'2007 are presented. The best accuracy of Lithuanian digit names and voice commands recognition was achieved by Spanish recognizer. Achieved recognition accuracy is suitable for the real applications of speech server for Lithuanian language.**

***Index Terms*—Speech analysis, speech processing, speaker recognition, accuracy.**

I. Introduction

Speech technologies are in developing for so many years. We might ask a question: what is it for? It's a simple answer – the user. Probably you heard these commands by calling to customer service: "At this moment all operators are busy. If you want to hear X, push 1, if you want to hear Y, push 2, if you want to hear Z, push 3". Probably this form of information presentation is confusing, after hearing a long list of commands it's easy to forget which button you have to press, occurs frustration, dissatisfaction and so on. The use of speech technologies enables to create more natural, intuitive and preferable information service to a consumer at a lower price. The use of language interface means that the information will be available at all times, despite of operators working hours. Speech technologies could be divided into two groups: automatic speech recognition (*ASR*) and text to speech (*TTS*) technology. The quality of speech applications is characterized by the accuracy of speech recognition and how natural the information is presented by voice

Speech servers, such as *Microsoft Speech Server* or *IBM WebSphere Voice Server* provide ASR and TTS resources which are the basis of speech interface. A special program placed in server runs the dialog between a human and computer. So far the only voice server application for Lithuanian language approach is by using foreign language transcriptions for Lithuanian words.

There are examples in the world when a module created for one language could be applied to another [1]. For this purpose linguistic or acoustic experience is in use [2]. Accurate researches of English recognizer application for the recognition of Lithuanian last names and Lithuanian digits names are presented in sources [3], [4]. The use of other language recognition tools for Lithuanian language is based on transcribing Lithuanian words or phrases into another, for example, English language: by using IPA (*International Phonetic Alphabet*) transcriptions Lithuanian word *„nulis"* could be automatically transcribed into English: *"n uh l ih s"*. Now these symbols could be recognized by English recognizer.

The essential component of various telephony services is the necessity to identify the person. In many cases it is possible to organize pronunciation of the personal identification number using spelling of the digit numbers. In this situation we obtain limited set of ten digit names that could be used to carry out core voice information for the public service. The main requirement for such applications is to develop Lithuanian speaker independent spoken digit recognizer with more than 95% accuracy.

II. The basics of speech server application for Lithuanian language

Microsoft Office Communications Server Speech Server (*MSS'2007)* [5] was chosen for preparing of telephony services. It performs speech recognition, speech synthesis and telephony control operations. For creating of new programs *Microsoft Visual Studio'2005* pack is in use. Voice output can be performed from the synthesized text, from processed audio files or may be derived from the synthesized files and pre-prepared mixture audio files. MSS'2007 has four language recognizers to choose from. For research we chose German (*Microsoft Speech Recognizer 9.0 for MSS (German-Germany)*), English (*Microsoft Speech Recognizer 9.0 for MSS (English-US)*), French (*Microsoft Speech Recognizer 9.0 for MSS (French-Canada))* and Spanish *(Microsoft Speech Recognizer 9.0 for MSS (Spanish-US)*) language recognizers.

Manuscript received March 12, 2012; accepted May 12, 2012.

Parts of this work were done under research project No.:31V-33/12 funded by "High technology development programme for 2011-2013" project "Hybrid recognition technology for voice interface" (INFOBALSAS).

Very good Lithuanian digit names recognition accuracy was reached using *Microsoft English (U.S.) v6.1* recognizer (99.8% for female speaker) [6]. That led to conclusion that using of speech server for Lithuanian digit names recognition would reach the same high results. Unfortunately the results of recognition experiments have shown that IPA transcriptions are not suitable for speech server [6]. UPS (*Universal Phone Set*) type of transcriptions should be used for MSS'2007 speech server [7]. It complicates the choice of transcriptions, since we do not know yet how to automate the testing procedure of voice commands recognition by speech server.

In order to check the suitability of selected transcriptions for MSS'2007, the test for the measuring of the accuracy of the voice commands recognition was prepared on MSS'2007 (Fig. 1): speech dialog component *answerCallActivity1* answers an incoming call, *questionAnswerActivity1* – asks the question and gets the user's answer, *gotoActivity1* – jumps to another component, *disconnectCallActivity1* – disconnects an existing call. Such framework is suitable for testing of Lithuanian voice commands recognition by selected speech recognizer. The prompt, grammar and target properties of *questionAnswerActivity1* and *gotoActivity1* speech dialog components should be defined before the testing procedure.

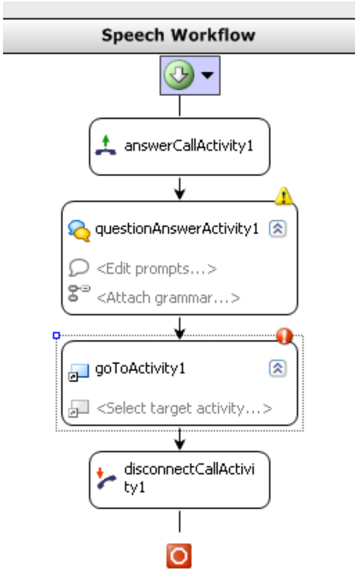
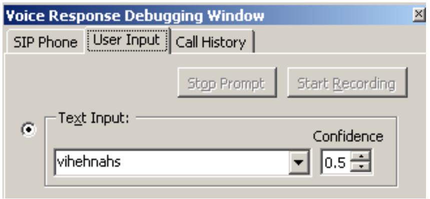


Fig. 1. The view of Dialog workflow designer window of prepared test.

Testing program at debugging mode presents the recognized word transcription and the confidence measure (Fig. 2): the word is recognized, if the confidence measure is above 0.2.



Text Input: vihehnahs

Confidence 0.5

Fig. 2. The view of Voice Response Debugging window after the recognition of digit „vienas".

More detailed information about the testing procedure of voice commands recognition using speech server was presented in [6].

III. EXPERIMENTAL EVALUATION OF SPEECH RECOGNITION ACCURACY BY TWO SPEAKERS

Digit recognition in many areas of life is essential, for example, in the bank operations, etc., therefore the names of Lithuanian digits *„nulis", „vienas", „du", „trys", „keturi", „penki", „šeši", „septyni", „aštuoni", „devyni"* have been chosen for the experiments with the already mentioned German, English, French and Spanish language recognizers. First of all Lithuanian digit names were rewritten to other language transcriptions using "synthesis", i.e., each Lithuanian digit name was synthesized with other language synthesizer and the most similar to Lithuanian pronunciation foreign transcriptions of digit names were selected The number of found transcriptions for each digit is unequal: for a short number, as *"du",* 7 transcriptions were enough, but for longer digits ("*septyni", "aštuoni")* - 10 transcriptions were selected.

At the transcription selection stage, a separate grammar and test were prepared for each digit and each recognizer (overall 40 grammars and tests). Two speakers (man and woman) took part in the experiments. Each digit was spoken 100 times through the microphone; the recognized transcriptions were selected as the winners. For example, the best transcriptions for digit *"nulis"*: *nuhlihs* (for German language), *nulis* (for English language), *nouluece* (for French language), *nuhlihs* (for Spanish language). If the same digit had more than one recognized transcription, all those transcriptions were used for further research and testing.

At the digit recognition stage, four tests were prepared for each recognizer. All winning transcriptions were used in the grammars. The accuracy of digit recognition with other language recognizers is presented in the Table I and the average confidence measure - in the Table II.

TABLE I. AVERAGE RECOGNITION ACCURACY OF LITHUANIAN DIGIT NAMES USING GERMAN, ENGLISH, FRENCH AND SPANISH RECOGNITION ENGINES.

Speaker	**German**	**English**	**French**	**Spanish**
KR, male	58.4	76.4	52.8	88.2
GB, female	51.8	59.0	76.8	98.8
Average	55.1	67.7	64.8	93.5

The best results were achieved with Spanish recognizer: the average accuracy of speaker-man is 88.2% and speaker-woman – 98.8%.

TABLE II. AVERAGE CONFIDENCE MEASURE OF LITHUANIAN DIGIT NAMES USING GERMAN, ENGLISH, FRENCH AND SPANISH RECOGNITION ENGINES.

Speaker	**German**	**English**	**French**	**Spanish**
KR, male	0.44	0.48	0.48	0.60
GB, female	0.42	0.37	0.57	0.77
Average	0.43	0.43	0.53	0.68

The results in Table II show that the highest confidence measure was gained also using Spanish language recognizer: the average confidence measure of speaker-man is 0.6, and speaker-woman – 0.77 (the confidence measure may vary from 0 to 1).

The recognition results of separate digit names using

German, English, French and Spanish recognizers are given in Fig. 3. Digit names "*vienas", "du", "penki"* and *"devyni"* were recognized without errors by Spanish recognizer.

IV. EXPERIMENTAL EVALUATION OF SPEECH RECOGNITION ACCURACY USING SPEECH CORPORA

Speech corpora – properly collected speech signal databases, which have the significance meaning to studies of speech signals and development of speech technologies. The characteristics of speech are very individual, different phonemes have very different characteristics. Phonetic units and acoustic characteristics and features of speech elements can be defined using speech corpora. It requires a lot of time and accuracy to process speech corpora: all incorrect recording must be deleted, because they could strongly influence the results of experiments.

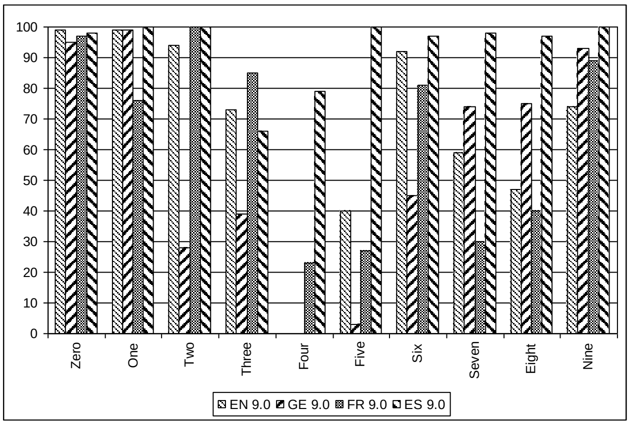


Fig. 3. The recognition accuracy of 10 Lithuanian digit names using English (EN 9.0), German (GE 9.0), French (FR 9.0) and Spanish (ES 9.0) speech engines.

For next experiments with Lithuanian digit names, speech corpora were collected, which contain utterances from 20 different speakers (5 speaker-men, 15 speaker-women). In this case each speaker pronounced 10 digit names 20 times. English and Spanish recognizers were selected for the comparison.

Speech server needs telephony format of speech input, so speech corpora were adopted by down-sampling the speech corpora from the original 16 kHz to 8 kHz sampling rate. We can mention other experiments concerning speech recognition at multiple sampling rates: the recognition performance was investigated on the speech signals sampled at different sampling rates (16, 11 and 8 kHz) with a HTK (Hidden Markov Model Toolkit) based recognizer [8]. It worked without a remarkable loss in recognition performance at above mentioned sampling rates. The efficiency of the recognition system using 11.025 kHz and 6 kHz sampled records was tested in [9]. The noticeable decrease of recognition accuracy was observed in the case of 6 kHz sampled records. Investigations have been carried out to determine the influence of speech coding on the performance of speech recognition systems in [10]. The deterioration of recognition performance was observed for all speech coders comparing with PCM-coded speech.

The averaged recognition accuracy of ten Lithuanian digit names using English and Spanish engines is shown in the Table III. Spanish speech engine (ES 9.0 I var.) enabled to achieve significantly higher recognition accuracy of Lithuanian digit names than English engine (EN 9.0): overall recognition accuracy increased from 77.0% for the English engine to the 97.0% for the Spanish engine.

For further research new transcriptions were created to achieve better results with Spanish recognizer. We started with words that had the lowest recognition score, for example *"trys"* and *"keturi"*. The grammar of the test contained new and old transcriptions. The average recognition accuracy of ten Lithuanian digit names reached 99.2% (ES 9.0 II var.).

In order to test if the selected transcriptions of digit names are suitable for other speech corpora, additional speech corpus was prepared from 10 new speakers. Similar results were achieved: the average recognition accuracy of ten Lithuanian digit names - 98.8% (ES 9.0 testing).

TABLE III. AVERAGE RECOGNITION ACCURACY OF TEN LITHUANIAN DIGIT NAMES USING ENGLISH AND SPANISH ENGINES.

Recognizer			
EN 9.0	ES 9.0 I var.	ES 9.0 II var.	ES 9.0 testing
77.0	97.0	99.2	98.8

The recognition accuracy of each digit using English and Spanish recognizers is shown in Fig. 4.

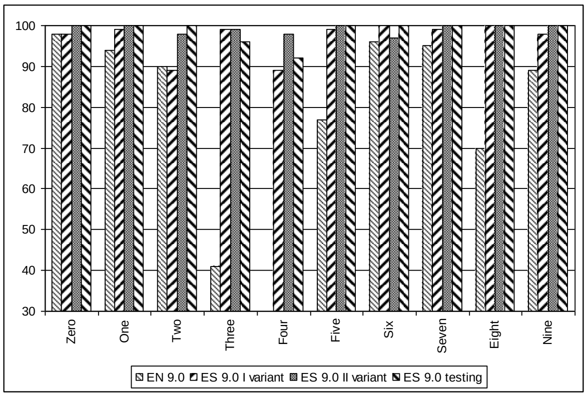


Fig. 4. The recognition accuracy of 10 Lithuanian digit names using English (EN 9.0) and Spanish (ES 9.0) speech engines and speech corpora.

Average recognition accuracy of ten Lithuanian digit names using Spanish engine significantly outperforms all earlier achieved results of similar experiments in Lithuania.

The above mentioned technique of proper transcriptions selection was used in the experiments of recognizing other Lithuanian digits names and some Lithuanian phrases. 9 Lithuanian phrases *("domestic accident", "accident on the way to the work", "donor", "epidemic", "nursing of the patient", "disease", "professional disease", "prosthesis", "observation of healthy child"*), which are used in the internet system of electronic documents management [11], presenting the cases of disability were chosen. Speech corpus consisted of 14 speakers with 60 utterances. In the case of Lithuanian digits names there were 26 digits *("dešimt", "vienuolika", "dvylika",…"milijonas", "milijonai", "milijonų"*) and 20 speakers with 20 utterances of each digit. 36 digits were got by adding the names of Lithuanian digits *„nulis", „vienas",*

„du“, „trys“, „keturi“, „penki“, „šeši“, „septyni“, „aštuoni“, „devyni“.

Excellent result of phrases recognition was achieved, but the predefined accuracy threshold of 95% wasn’t reached in the experiments of Lithuanian digits names recognition. Experimental results are presented in the Table IV.

TABLE IV. AVERAGE RECOGNITION ACCURACY OF SOME LITHUANIAN SPEECH CORPORA USING SPANISH ENGINE.

Speech corpus	9 phrases	26 digits	36 digits
Accuracy, %	100	93.4	91.8

The main reason of poor recognition of 26 or 36 Lithuanian digits names is the confusion of those digits names, which differ only by the ending part of digit name, for example, *“milijonas”, “milijonai”*. So big numbers should be separated into digits from 0 to 9 in the planned applications of speech server for Lithuanian language.

V. CONCLUSIONS

Different foreign language speech engines have different capabilities to recognize Lithuanian voice commands. Spanish speech engine enabled to achieve significantly higher recognition accuracy of Lithuanian digit names than English engine: overall recognition accuracy increased from 77.0% for the English engine to the 99.2% for the Spanish engine. Excellent result of nine Lithuanian phrases recognition by Spanish engine was achieved too. Such recognition accuracy is suitable for the real applications of speech server for Lithuanian language and significantly outperforms all earlier achieved results of similar experiments in Lithuania.

Big Lithuanian numbers should be separated into digits from 0 to 9 in the planned applications of speech server for Lithuanian language.

REFERENCES

[1] T. Schultz, A. Waibel, “Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition”, *Speech Communication,* vol. 35, no. 1–2, pp. 31–51, 2001. [Online]. Available: http://dx.doi.org/10.1016/S0167-6393(00)00094-7

[2] A. Zgank, et al., “The COST278 MASPER initiative – croslingual speech recognition with large telephone databases”, in *Proc. of the 4th International Conference on Language Resources and Evaluation (LREC’04),* 2004, pp. 2107–2110.

[3] P. Kasparaitis, “Lithuanian Speech Recognition Using the English recognizer”, *Informatica*, vol. 19, no. 4, pp. 505–516, 2008.

[4] R. Maskeliūnas, “Lithuanian Voice Commands Recognition Based on the Multiple Transcriptions”, Ph.D. dissertation, KTU, Kaunas: Technologija, 2009, p. 159.

[5] M. Dunn, *Pro Microsoft Speech Server 2007: Developing Speech Enabled Applications with .NET.* New York: Apress, 2007, p. 275.

[6] R. Maskeliūnas, A. Rudžionis, K. Ratkevičius, V. Rudžionis, “Investigation of Foreign Languages Models for Lithuanian Speech Recognition”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 3, pp. 37–42, 2009.

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

[8] H. G. Hirsch, K. Hellwig, S. Dobler, “Speech Recognition at Multiple Sampling Rates”, in *Proc. of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001),* Aalborg, Denmark, 2001, pp. 1837–1840.

[9] 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.

[10] H. G. Hirsch, “The Influence of Speech Coding on Recognition Performance in Telecommunication Networks”, in *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP2002),* Denver, USA, 2002, pp. 1877–1880.

[11] *The system of electronic documents management*. [Online]. Available: http://epts.sodra.lt/bendra-informacija.jsp