About Development and Innovation of the Slovak Spoken Language Dialogue System

Jozef Juhár[*], Stanislav Ondáš[*], Anton Čižmár[*] and Milan Rusko[**]
[*]Department of Electronics and Multimedia Communications,
Technical University of Košice, Faculty of Electrical Engineering and Informatics,
Letná 9, 042 00 Košice, Slovakia,
E-Mail: Jozef.Juhar@tuke.sk, Anton.Cizmar@tuke.sk, Stanislav.Ondas@tuke.sk

[**]Department of Speech Analysis and Synthesis,
Slovak Academy of Sciences, Institute of Informatics,
Dúbravská cesta 9, 845 07 Bratislava, Slovakia,
E-Mail: Milan Rusko@savba.sk

Abstract ***–The research and development of the Slovak spoken language dialogue system (SLDS) is described in the paper. The dialogue system is based on the DARPA Communicator architecture and was developed in the period from July 2003 to June 2006. It consists of the Galaxy hub and telephony, automatic speech recognition, text-to-speech, backend, transport and VoiceXML dialogue management and automatic evaluation modules. The dialogue system is demonstrated and tested via two pilot applications, „Weather Forecast“ and „Public Transport Timetables“. The required information is retrieved from Internet resources in multi-user mode through PSTN, ISDN, GSM and/or VoIP network. Some innovation development has been performed since 2006 which is also described in the paper***

Keywords: ***Dialogue System, Speech Recognition, VoiceXML, Text-To-Speech, Dialogue Management.***

I. INTRODUCTION

Due to the progress in the technology of speech recognition and understanding, the spoken language dialogue systems have begun to emerge as a practical alternative for a conversational computer interface. They are more effective than IVR systems since they allow a more free and natural interaction and can be combined with the input modalities and visual output.

In this paper we describe the development of the first SLDS which is able to interact in Slovak language. The system has been developed in the period from July 2003 to June 2006 and was supported by the National program for R&D “Building of the information society”. The main goal of the project was the research and development of a SLDS for information retrieval using voice interaction between humans and computers. The SLDS had to enable multi-user interaction in the Slovak language through telecommunication networks and to find information distributed in computer data networks such as the Internet. The SLDS is also a tool for continuing research in the area of spoken language technologies in Slovakia.

Contractors of the project were the Ministry of Education of the Slovak Republic and the Technical University of Košice. Collaborative organizations were the Institute of Informatics, the Slovak Academy of Sciences Bratislava, the Slovak University of Technology in Bratislava and the University of Žilina.

The choice of a solution had come from contemporary free resources, state-of-the-art in the topic and the experiences of the partners involved in the project. As described further the solution is based on the DARPA Communicator architecture based on the Galaxy hub, a software router developed by the Spoken Language Systems group at MIT, subsequently released as an open source package in collaboration with the MITRE Corporation, and now available on SourceForge. The proposed system consists of the Galaxy hub and six modules (servers). Functionality of the SDS is demonstrated and tested via two pilot applications – „Weather forecast for Slovakia“ and „Timetable of Slovak Railways“ retrieving the required information from internet resources in multi-user mode through telephone [1].

Since 2006 the SLDS is continuously improved at Technical University of Košice with collaboration of Slovak Academy of Science Bratislava. The innovation steps are described in the paper.

II. SYSTEM ARCHITECTURE

The architecture of the developed system is based on the DARPA Communicator. The DARPA Communicator systems use a ‘hub-and-spoke’ architecture: each module seeks services from and

provides services to the other modules by communicating with them through a central software router, the Galaxy hub. Java, along with C and C++ are supported in the API to the Galaxy hub. The substantial development based on the Communicator architecture has been already undertaken at Carnegie Mellon University and the University of Colorado.

Our system consists of a hub and seven system modules: telephony module, automatic speech recognition (ASR) module, text-to-speech (TTS) module, transport server, back/end module, module of dialogue management and the new evaluation module. The relationship between the dialogue manager, the Galaxy hub, and the other system modules is represented schematically in Fig. 1.

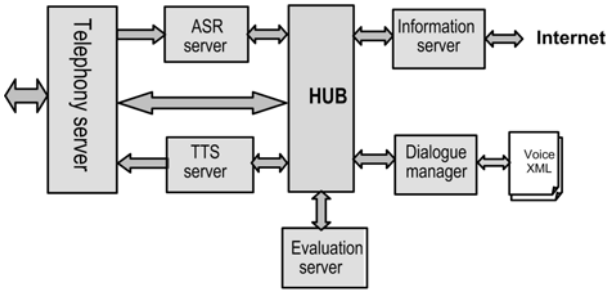

Fig. 1. Architecture of the Galaxy/VoiceXML based spoken Slovak dialogue system

The telephony module connects the whole system to a telecommunication network. It opens and closes telephone calls and through the/a Broker Channel transmits speech data to/from the ASR/TTS modules. The server of automatic speech recognition (ASR) performs the conversion of incoming speech to a corresponding text. Context dependent HMM acoustic models trained on SpeechDat-Sk and MobilDat-Sk speech databases and ATK/HTK based speech recognition engine were used in this task. The dialogue manager controls the dialogue of the system with the user and performs other specified tasks. The heart of the dialogue manger is as the interpreter of VoiceXML mark-up language. The information server connects the system to information sources and retrieves information requiered by the user. The server of text-to-speech (TTS) synthesis converts outgoing information in text form to speech, which is more user friendly. The evaluation server tracks communication among servers and computes a set of interaction parameters.

The communicator supports „Windows-only" as well as a mixed Windows/Linux platform solution. In this case a Transport Server, managing file transmissions between platforms is active.

III. MOBILDAT-SK SPEECH DATABASE

The MobilDat-SK is a speech database containing recordings of 1100 speakers recorded over a mobile (GSM) telephone network. It serves as an extension to the SpeechDat-E Slovak database and so was designed to follow the SpeechDat specification as closely as possible. It is balanced according to the age, regional accent, and sex of the speakers. Every speaker pronounced 50 files (either prompted or spontaneous) containing numbers, names, dates, money amounts, embedded command words, geographical names, phonetically balanced words, phonetically balanced sentences, Yes/No answers and one longer non-mandatory spontaneous utterance.

Every speaker called only once; from one acoustic environment. The required number of the calls from different environments were specified as a minimum 10% of the database for each environment. The home, office, public building, street and vehicle acoustic environments were chosen.

We decided to use the database content adopted from SpeechDat-E database, however according to assumed practical applications some new items were added:

- Sentence expressing a query on departure or arrival of a train, including names of two train stations from a set of 500.
- Name of the town or tourist area from a set of 500 names.
- Web domain or e-mail address from a set of 150 web domains and 150 e-mail addresses
- One non-mandatory item. A longer spontaneous utterance was added at the end of the recording. The caller had to answer a simple question from a set of 25 such as: "How do you get from your house to the post-office?". This item should considerably extend the spontaneous content of the database.

Except of the SpeechDat-E and MobilDat-SK databases speech utterances from the real interactions with Slovak SDS were collected, annotated and used for acoustic models adaptation (improvement) [2], [15].

A considerable time was dedicated to testing and comparison of trained acoustic models [12], [25], [26].

IV. AUTOMATIC SPEECH RECOGNITION

The basic idea of stochastic system for automatic speech recognition (ASR) is to use a statistical information about structure of sounds, words and sentences in particular language.

The ASR system can be broken into two main parts. First part is extracting observation vectors from input speech. In that way the data rate to the next, the most time-consuming part, can be lowered. The second part is processing all provided statistical information about language and the acoustic observation vectors in order to "decode" input speech. We use word "decode" because the user, as he speaks, is "encoding" the words from his mind into sounds coming from his mouth. On the side of ASR system we are then performing a reverse operation, the "decoding".

The development of reliable and fast speech recognizer is not an easy task. Fortunately there are several speech recognizers available for nonprofit

research. We have adapted two well-known speech recognizers as the ASR module for our system. The first is ATK, on-line version of HTK [3]. The ATK based ASR module was adapted for our SDS running on a Windows-only platform and on a mixed Windows/Linux platform as well. In the second case the ASR module runs on a separate PC with Linux OS. The second speech recognizer we adapted for our system is Sphinx-4 written in Java [10]. Both ASR modules provide similar results.

SpeechDat-SK and MobildDat-SK [7] databases were used for traning HMM's. Context dependent (triphone) acoustic models were trained in a training procedure compatible with "refrec" procedure [1], [4]. Dynamic speech recognition grammars and lexicons are used in the speech recognizers.

We trained also a filler model, which serves as a garbage in the word-spotting mode of ASR system. Fillers were built in to the speech grammars, which we use in the "How may I help you?" questions in the Slovak weather forecast service [16].

Modern voice applications bring a need of a large vocabulary continuous speech recognition functionality. Therefore there is a significant effort to build such speech recognition system [13], [14], [23].

The recognition network is based on finite-state acceptor (FSA) and/or finite-state transducer (FST) are being under development [6].

V. TEXT-TO-SPEECH SYNTHESIS

Two TTS modules have been designed using two different approaches – diphone and corpus based synthesis.

A. The Diphone Concatenative Synthesizer

This speech synthesizer is based on concatenation of small elements of a pre-recorded speech signal, mainly diphones. An original algorithm similar to the Time Domain Pitch Synchronous Overlap and Add (TD-PSOLA) was used for concatenation. The pronunciation is controlled by a block of orthographic-to-orthoepic (grapheme to phoneme) conversion based on a sophisticated set of rules supplemented by a pronunciation vocabulary and a list of exceptions. This elaborated unit has proven to be more reliable than our similar data driven system based on CART trees [8].

B. Corpus synthesizer

The second method is based on corpus synthesis i.e. concatenation of prepared acoustic units. The advantage of corpus-based method is in minimizing the number of concatenations in synthesized speech and thus reducing the need for speech processing causing artificiality.

The most critical phase of corpus synthesis is the selection of appropriate units. In our recent approach to synthesis we apply a phonetic unit pre-selection which reduces the universality and openness of the classical unit selection approach, but it excludes the most significant concatenative problems in advance, before the calculation of concatenative and unit costs has even started [8]. Modelling acoustic parameters of prosody in Slovak using classification and regression trees was implemented [27]

VI. DIALOGUE MANAGER

There are many approaches of how to solve dialogue manager unit and also many languages for writing a code for it. Voice Extensible Markup Language (VoiceXML) is a markup language for creating voice user interfaces that use automatic speech recognition (ASR) and text-to-speech synthesis (TTS). Many commercial VoiceXML-based speech applications have been deployed across a diverse set of industries, including financial services, government, insurance, retail, telecommunications, transportation, travel and hospitality. VoiceXML simplifies speech application development, enables distributed application design and accelerates the development of interactive voice response (IVR) environments. For these reasons, VoiceXML has been widely adopted within the speech industry, and for these reasons we decided that the dialogue manager unit be based on VoiceXML interpretation.

Fig. 2 shows the structure of the dialogue manager [11]. Its fundamental components are VoiceXML interpreter, XML Parser and ECMAScript unit. The dialogue manager is written in C++ and in its actual state the interpreter performs all fundamental algorithms of VoiceXML language and service functions for all VoiceXML commands, i.e. Form Interpretation, Event Handling, Grammar Activation and Resource Fetching Algorithms. It supports the full range of VoiceXML 1.0 and a large part of VoiceXML v2.0 functions.

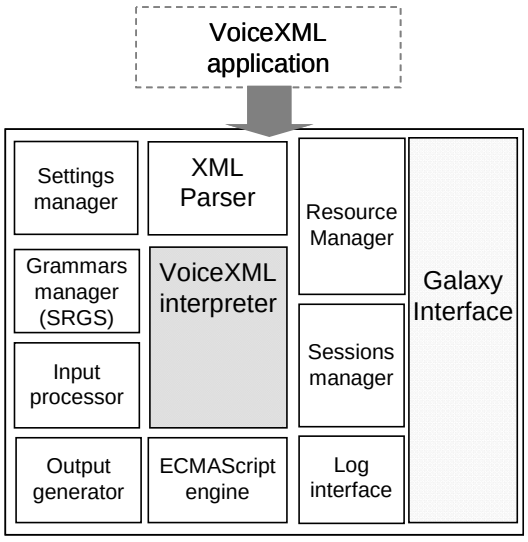

Fig.2 Basic components of VoiceXML based dialogue management unit

Developed dialogue manager module integrates three approaches for obtaining information from web. The standard way is using of the submit element, which sends an http request to the web server. Then the dialogue manager takes generated VoiceXML document and interprets its. The second way, mostly used for local information sources, uses an object element, which can run local scripts and executables. The third and most robust way is cooperation with backend module of the system, which takes responsibility for obtaining data.

VII. TELEPHONY MODULE

The telephony (audio) server connects the whole system to a telecommunication network. A direct (broker) connection between audio server and ASR or TTS server is established to transmit speech data and this way reduces the network load. The telephony server is written in C++ and supports telephone hardware (Dialogic D120/41JCT-LSEuro). The server fully supports DTMF communication and call-routing and the ongoing work will be concentrated on the support of barge-in. Concept of VoIP-based telephony module is under development [5].

VIII. BACKEND MODULE

After an analysis of various existing approaches of information retrieval from the web, and of the task to be carried out by the information server in our pilot applications, we came to the decision that no complicated data retrieval system is needed. On the contrary – a rule based ad-hoc application searching only several predefined web-servers with a relatively well known structure of pages will do a much better job. As the number of web-servers giving detailed weather forecast for Slovakia, as well as the number of web-servers providing information on train and bus connections in Slovakia is very limited, we had been checking several selected servers for stability and information reliability for about a month, and then we chose several candidate servers, from which the information is to be retrieved.

The information server (backend server) is capable of retrieving the information contained on the suitable web-pages according to the Dialogue Manager (DM) requests, extracting the needed data, analyzing it and if they are taken for valid, returning the data in the XML format to the DM. If the backend server fails to get valid data from one web source, it switches to a second wrapper retrieving the information from a different web-server.

The information server communicates with the HUB via the GALAXY interface. This module accomplishes its own communication with HUB, receives input requests, processes them and makes decisions to which web-wrapper (WW) the request should be sent, receives the answer and sends it back to the HUB.

The web-wrapper is responsible for the navigation through the web-server, data extraction from the web-pages and their mapping on to a structured format (XML), convenient for further processing. The wrapper is specially designed for one source of data; thus to combine data from different sources, several wrappers must be designed.

Wrappers are designed to be as robust as possible against changes in the web-pages structure . Nevertheless, in the case of substantial changes in the web-page design, the adaptation of the wrapper would probably be inevitable.

To speed up the system (to eliminate the influence of long reaction times of the www-pages) and to assure drop-out resistance while simultaneously keeping the information as current as possible, automatic periodic download and cashing of the web-pages content were introduced.

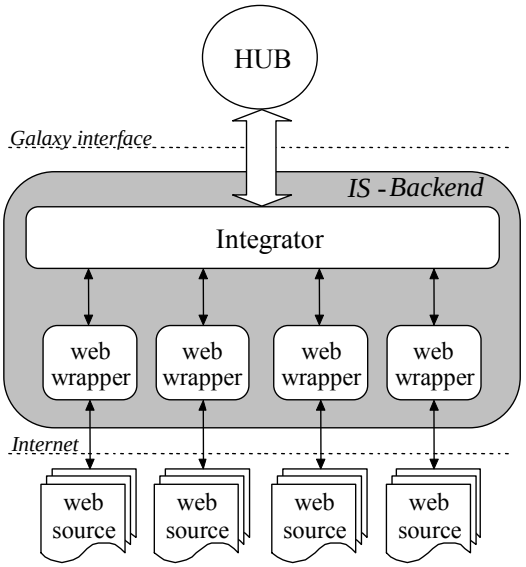

Fig.3 Progress of performance for consequently trained models

The server is open for future applications by the possibility of creating web-wrappers for new services and adding them to the existing wrappers. The information on the currently accessible wrappers is stored in the system's configuration file.

IX. EVALUATION MODULE

There is a lot of views, methods and approaches to the assessment and evaluation of the SLDS's quality. The used technique depends on the purpose of the evaluation. Our intention was to design such technique, which will be able to extract information directly and automatically during the system-user interaction.

According to [17] the evaluation in the Slovak SLDS is performance and quality evaluation realized in "field" environment. The object of the evaluation is the overall system and the type of used method is instrumental and

expert-based measurement of system and interaction parameters.

The developed evaluation server is a stand-alone galaxy server programmed in C++ language, which tracks message communication between servers of SDS and extracts interaction parameters according these messages.

Hub programs were also rebuilt for the purpose of evaluation. We added in to the relevant programs a new rule, for routing particular interaction to the evaluation module.

There was a need to specify a set of interaction parameters, which should be computed by this module. The specific of this task was that there were never been established any set of interaction parameters for SDS supporting W3C Voice standards. After the analysis of several evaluation methods, we took as a ground the work of Sebastian Möller [17]. Designed set of interaction parameters consists of 17 parameters (fifteen modified Möller's parameters, two new) [18].

Information about computed parameters is systematically logged in to three types of files. The main information file holds the basic information about all incomming calls on all telephony lines (channels). The evaluation server generates also the channel information files for all connected telephony lines, which contains detailed information about all calls on the given channel. The third file is the channel statistic file. It provides summarized information about calls on given channel.

The results of the evaluation on 309 calls are presented in our earlier paper [19].

The complementary type of SDS evaluation is the subjective evaluation, which obtains information from questionnaires fulfilled by users. The evaluation experiment was carried on the 26 users, whose made 52 interactions with the system [22]. The results show the relation of both evaluation experiments.

X. SERVICES

Since January 2006 the developed system is providing weather forecast service and train and bus timetable information services in Slovak language, which are available publicly through PSTN and GSM telephony network and through VoIP (Skype). In Fig. 4 the sub-dialogue structure of the services is shown. In March 2009 the new sub-service was started, which provides information about city buses in Slovak town Košice. Both services were designed with the consideration for elderly and disabled people [21].

The services have been designed as system-directed or with mixed-initiative. In the weather forecast service a "how-may-I-help-you" approach is used. Services have the open structure based on sub-dialogues. This structure enables replacing of arbitrary part of the service without the need of rebuilding the other parts. All timetable services together would create the complex timetable service for Slovakia.

Until now more than 5000 interactions were done by users and the system has more than 400 Skype's users.

Evaluation of the dialogue system based on ITU-T recommendations has been proposed and performed [18], [22].

Based on experience with development of service dialogues a corpus of reusable dialogue components for voice dialogues design in Slovak was built [24].

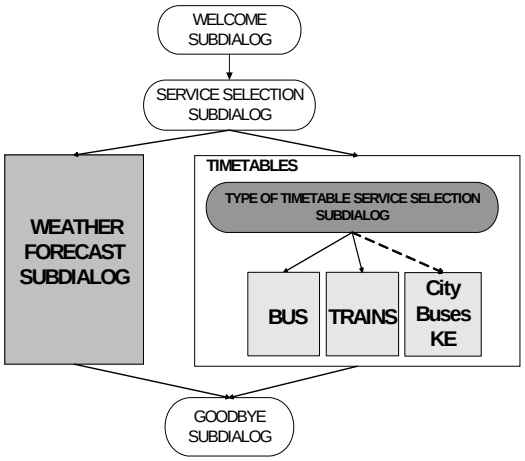

Fig. 4. The structure of services

XI. CONCLUSIONS

In this paper we described the development of the first Slovak spoken language dialogue system. Our main goal was to develop a dialogue system that will serve as a starting platform for further research in the area of spoken Slovak engineering. We successfully combined up to date free resources with our own research into functional system that enables in multi-user mode an interaction through telephone in the Slovak language to retrieve required information from Internet resources. The functionality of the SLDS is demonstrated and tested via two pilot applications, „Weather forecast for Slovakia" and „Slovak Timetables Service" [9]. Applying new findings we are continuing in further developments and improvements to the system.

ACKNOWLEDGEMENTS

The work presented in this paper was supported by the Ministry of education of Slovak Republic under research projects AV 4/0006/07 and AV 4/2016/08 and Slovak Research and Development Agency under research project APVV-0369-07

REFERENCES

[1] Juhár, J., Ondáš, S., Čižmár, A., Jarina, R., Rusko, M., Rozinaj, G., "Development of Slovak GALAXY / VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet", In Proc.

Interspeech 2006, Pittsburg, USA, Sept. 17-21, 2006, p. 485-488. ISSN 1990-9772.

[2] Pleva, M., Juhár, J., Čižmár, A., "Building of the annotated speech utterances database from the Slovak spoken dialog system interactions", In RTT 2008 - Research in Telecommunication Technology, 9th International Conference: Vyhne, Slovak Republic, September 10-12, 2008, 260-262. ISBN 978-80-227-2939-0.

[3] Young, S., "ATK: An application Toolkit for HTK, version 1.3", Cambridge University, January 2004.

[4] Lihan, S., Juhár, J., Čižmár, A., "Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases", In Proc. Interspeech 2005, Lisabon, Portugal, September 2005, pp. 225 – 228.

[5] Pleva, M., Juhár, J., Čižmár, A., "Concept of spoken dialog system based on voice over IP telephony", In RTT 2008 - Research in Telecommunication Technology, 9th International Conference: Vyhne, Slovak Republic, September 10-12, 2008, pp 263-265. ISBN 978-80-227-2939-0.

[6] Lojka, M., Juhár, J., "Finite-state machines for continuous speech recognition - an overview", In Digital Technologies – International Workshop (CD Proceedings), Žilina, Slovakia, November 20-21, 2008, ISBN 978-80-8070-953-2

[7] Rusko, M., Trnka, M., Darjaa S., "MobilDat-SK - A Mobile Telephone Extension to the SpeechDat-E SK Telephone Speech Database in Slovak", SPEECOM 2006, Saint-Peterburg, Russia, June 2006, pp.449-454.

[8] Darjaa, S., Rusko, M., Trnka M., "Three Generations of Speech Synthesis Systems in Slovakia", SPEECOM 2006, Saint-Peterburg, Russia, July 2006, pp. 297-302.

[9] Gladišová, I., Doboš, Ľ., Juhár, J., Ondáš, S., "Dialog Design for Telephone Based Meteorological Information System", in Proc. DSP-MCOM 2005, Košice, Slovakia, Sept., 2005, pp. 151-154.

[10] Mirilovič, M., Lihan, S., Juhár, J., Čižmár,A., "Slovak speech recognition based on Sphinx-4 and SpeechDat-SK", In Proc. DSP-MCOM 2005, Košice, Slovakia, Sept. 2005, pp. 76-79.

[11] Ondáš, S., Juhár, J., "Dialogue manager based on the VoiceXML interpreter", in Proc. DSP-MCOM 2005, Košice, Slovakia, Sept. 2005, pp.80-83.

[12] Mirilovič, M., Juhár, J. and Čižmár, A., "Comparison of grapheme and phoneme based acoustic modeling in LVCSR task in Slovak", In A. Esposito et al. (Eds.), Multimodal Signals: Cognitive and Algorithmic Issues, LNAI 5398, Springer-Verlag, pp. 242-247, 2009.

[13] Mirilovič, M., Juhár, J. and Čižmár, A., "Large Vocabulary Continuous Speech Recognition in Slovak", In AEI'08 - Applied Electrical Engineering and Informatics, International Conference: Athens, Greece, September 8-11, 2008, pp.73-77.

[14] Katrák, M., Juhár, J., "The classification of English phones using neural network", In RTT 2008 - Research in Telecommunication Technology, 9th International Conference: Vyhne, Slovak Republic, September 10-12, 2008, pp 100-103.

[15] Papco, M., Juhár, J., "Acoustic models trained on SpeechDat database extended to utterances recorded from users interaction with IRKR system", In RTT 2008 - Research in Telecommunication Technology, 9th International Conference: Vyhne, Slovak Republic, September 10-12, 2008, pp 247-250.

[16] Gladišová, I., Doboš, Ľ., Juhár, J., Ondáš, S., "Dialog design of the telephone based meteorological information system", In Proc. DSP-MCOM 2005, Košice, Slovakia, September 2005, pp. 151-154.

[17] Möller, S., "Quality of Telephone-Based Spoken Dialogue Systems", Springer Science + Business Media, Inc., 2005, eBook ISBN 0-387-23186-2.

[18] Ondáš, S., "Design and evaluation of the spoken dialogue systems and services", PhD. Thesis (in Slovak), September 2008, Technical University of Košice, Slovakia

[19] Ondáš, S., Juhár, J., "Automatic evaluation of Slovak spoken language dialogue system", In ECMS 2007 & Doctoral School, Liberec, May 21-23, 2007, p.120.

[20] Papco, M., Juhár, J., "Spectral-Subtractive Algorithm in Speech Recognition with MobilDat Corpus", In Digital Technologies – International Workshop (CD Proceedings), Žilina, Slovakia, November 20-21, 2008.

[21] Ondáš, S., Juhár, J., "Voice services for elderly and disabled people in MonAMI project in Slovakia", In Digital Technologies – International Workshop (CD Proceedings), Žilina, Slovakia, November 20-21, 2008.

[22] Ondáš, S., Juhár, J., Čižmár, A., "Evaluation of the Slovak Spoken Dialogue System based on ITU-T", In: P. Sojka, et.al. (Eds): Text, Speech and Dialogue, LNAI 5246, pp. 633 – 640, Springer-Verlag Berlin Heidelberg 2008.

[23] Mirilovič, M., Juhár, J., "Morphological Segmentation of Word Units for Large Vocabulary Automatic Speech Recognition in Slovak", In HLT 2007 – Proceedings of the Third Baltic Conference on Human Language Technologies, October 4-5, 2007, Kaunas, Lithuania, pp. 189-196.

[24] Ondáš, S., Juhár, J., "Building of Reusable Dialogue Corpus for Voice Dialogue Design in Slovak", In HLT 2007 – Proceedings of the Third Baltic Conference on Human Language Technologies, October 4-5, 2007 Kaunas, Lithuania, pp. 227-235.

[25] Lihan, S., Juhár, J., "Comparison of two Slovak speech databases in speech recognition tests", In ACOUSTICS High Tatras 06 : 33rd International Acoustical Conference - EAA Symposium : Štrbské Pleso, Slovakia, October 4 - 6, 2006. s. 130-133.

[26] Lihan, S., Juhár, J., Čižmár,A., "Comparison of Slovak and Czech Speech Recognition Based on Grapheme and Phoneme Acoustic Models", In: Proc. Interspeech 2006, Pittsburg, USA, Sept. 17-21, 2006, p. 149-152.

[27] Rusko M., Trnka M., Darjaa S., Kováč R., "Modelling acoustic parameters of prosody in Slovak using Classification and Regression Trees", In Human Language Technologies as a Challenge for Computer Science and Linguistics - Proceedings. Poznań, Poland, 2007. ISBN 978-83-7177-407-2, pp. 231-235.