# Speech Technologies for Advanced Applications in Service Robotics

**Stanislav Ondáš[1], Jozef Juhár[1], Matúš Pleva[1], Martin Lojka[1], Eva Kiktová [1], Martin Sulír[1], Anton Čižmár[1], Roland Holcer[2]**

[1] Department of Electronics and Multimedia Communications, FEI, Technical
University of Košice, Park Komenského 13, 041 20 Košice, Slovak Republic
E-mail: {stanislav.ondas, jozef.juhar, matus.pleva, martin.lojka, eva.kiktova,
martin.sulir, anton.cizmar}@tuke.sk

[2] ZŤS VVÚ KOŠICE a.s., Research, Development, Design & Supply Company,
Slovakia, Južná trieda 95, 041 24 Košice, Slovakia
E-mail: roland.holcer@ztsvvu.eu

*Abstract: The multimodal interface for controlling functions of the complex modular robotic system, which can be deployed in difficult conditions as are rescue works, natural disasters, fires, decontamination purposes was designed. Such interface involves several fundamental technologies such as speech recognition, speech synthesis and dialogue management. To enable human operator to cooperate with designed robotic system, the sophisticated architecture was designed and described technologies were implemented. The automatic speech recognition system is introduced, which is based on Hidden Markov models and enables to control functions of the system using a set of voice commands. The text-to-speech system was prepared for producing feedback to the operator and dialogue manager technology was adopted, which makes it possible to perform the information exchange between operator and robotic system. The system proposed is enriched with acoustic event detection system, which consists of a set of five microphones integrated on the robotic vehicle, the post-processing unit and detection unit.*

*Keywords: service robots; speech technologies; speech recognition; speech synthesis; multimodal interface*

# 1 Introduction

The field of service robotics is growing rapidly due to the increasing need of automation, safety and time saving. Robotic systems are mainly deployed in industry and nowadays we see robot applications s for healthcare or home usage too [1], [2]. A special effort is devoted to design and development of robotic systems for independent living of elderly people [3], [4].

It can be concluded, that in such "advanced" applications, successfulness of robotic devices depends critically on the reliable, user-friendly and intuitive interface, which enables "cooperation" between a user and a robotic system. The other important reason for paying close attention to designing a so-called "human-machine interfaces" (HMI) is, that so complex systems have a lot of various functionalities, which can be difficult to control by "old school" interfaces, as are keyboard, buttons, joystick etc. Usage of more natural form of an interface, which enables human-like interaction involving speech, vision and other channels, can provide more usable, intuitive and ergonomic interface.

Speech together with auditory perception can be identified as the most important channels to exchange ideas and thoughts among people. This dominance relates to the fact, that speech is an acoustic representation of language, which relates to the representation of the world in a human mind. The other important attribute of human speech is its capability to replace other modalities, when they are not available (e.g. in telephony interaction). These attributes make speech the most important modality in HMI. Therefore speech technologies become widely used in advanced robotic applications.

To enable human-robot speech communication, a several technologies, which can be labeled as "speech technologies", need to be involved. The automatic speech recognition together with text-to-speech synthesis is the most important. But mentioned technologies alone are not enough for successful human-robot interaction. Some dialogue logic need to be integrated which will be able to hold information about dialogue state and will decide about next step in the interaction (next dialogue state). Previous sentence is the definition of dialogue management technology, which creates a core of spoken dialogue systems [5]. To close the communication ring of human-machine interaction, natural language understanding and generation technologies should be mentioned too.

When we are talking about speech technologies, another technology can be mentioned, whose relation to speech technologies is based on analyzing the audio signal and on similar processing mechanisms – acoustic event detection. Acoustic event detection processes the input audio signal and classifies incoming sounds. In robotics, recognition of dangerous sounds and alerting the operator can be very useful due to the fact that robots often work in distant (or closed) environment. If the operator cannot see the robot directly, the video stream from robot's camera is the only medium for accessing information about environment, in which the robot moves. Extending robot with a set of microphones and acoustic event detection system can bring possibility of collecting additional information from this environment [6].

Although these technologies are being researched and implemented for a long time, their language dependency together with applying in new applications and situations creates a new space for their researching, designing and development. It is also true, that there are a lot of problems, which were not solved successful yet, e.g. robustness of speech recognition, naturalness of speech synthesis or the complexity of dialogue management. Therefore, the paper proposed introduces the work, which has been done in the area of speech technologies for Slovak language for deployment in robotics applications in *Laboratory of speech technologies in telecommunications* on the *Department of Electronics and Multimedia Communications FEI TU* in Košice.

In Section 2, the complex modular robotic system for service robotics development is presented on functional prototype described in the following section. Next, the description of technologies of speech recognition, speech synthesis, audio events detection, and multimodal interaction developed for the service robot prototype, is provided. Finally the future work is described and discussed.

## 2 Complex Modular Robotic System

The complex modular robotic system and particular modules have been designed and are being developed within a range of three projects. The first project, named "*The complex modular robotic system of middle category with highest intelligence*" (funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic MŠ SR- 3928/2010-11), continues in our earlier cooperation with ZŤS VVÚ Košice a.s. company, described in [7]. Next projects are being performed in cooperation with our partners from industry and universities in range of Research & Development Operational Program funded by the ERDF - *Research of modules for intelligent robotic systems* (ITMS project code 26220220141) and *Competence Center for Innovation Knowledge Technology of production systems in industry and services* (ITMS project code 26220220155).

Mentioned projects are focused on the design and development of the complex system of intelligent modules to be used for development of robotic systems. Such complex robotics systems will be deployed in difficult conditions such as rescue works, natural disasters, fires, decontamination purposes and so on.

The mobile robotic platform (3D model in Figure 1), developed by ZŤS VVÚ a.s. Košice, able to move autonomously in rugged terrain with the speed about 3-5 km/h and elevation of 45°, of usable capacity of 400 kg is intended to be the output of the work within the aforementioned projects. The robotic vehicle will bear several superstructures (extensions):

- the robotic arm with 6 degrees of freedom and nominal load 200 kg
- the extricate system with expansion force 20 kN
- the decontamination system for removing of toxic substances

The following systems for control and navigation will be included into the robotic platform:

- systems for intelligent teleoperator control
- systems for autonomous performance of tasks of advanced cognitivity
- systems for multi-sources navigation

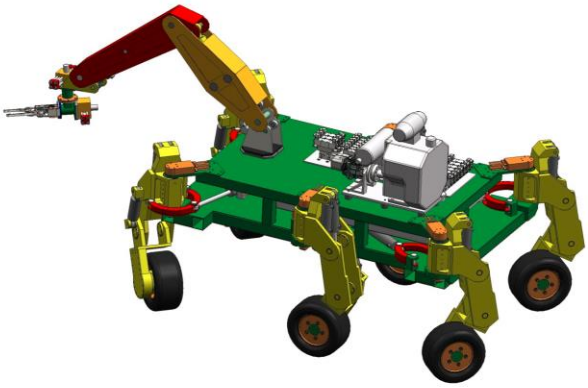


Figure 1
3D model of the complex modular robotic system

Although the introduced robotic system will be able to perform several tasks totally autonomously, a lot of other tasks need some cooperation with a human operator. The usage scenario of such systems is, that the mobile robotic platform (vehicle) has wireless connection with the control panel computer, which can be encapsulated into the robust briefcase (e.g. in [7]). Such control panel provides an interface between operator and robotic vehicle, which enables controlling the vehicle by joystick, keyboard or buttons, and the information from vehicle are depicted on display.

Adding the possibility of controlling selected functions by speech can significantly increase usability of the robotic system, due to the fact, that a large range of system's functions is difficult to control only by hands. The situation is very similar as for in-vehicle systems, because driver should pay attention to driving, what means, that his gaze watches the traffic and his hands are on the steering.

Therefore, modules, which enable cooperation with robotic system through speech interaction, need to be designed and integrated. The analysis of usage scenarios has shown the need of designing the more sophisticated multimodal interface (will be described in Section 3.1), because the interaction with robotic systems requires besides speech recognition and speech synthesis technologies also controlling the graphical output, transformation the speech commands into the system's actions, and some approach for acknowledgement of recognized commands (e.g. using "dead-man" button of the joystick) [8].

# 3 Technologies for the Robotic System Speech Control

## 3.1 The Robot Functions Multimodal Interface Architecture

For enabling usage of the speech modality for cooperation with robotic systems, a complex solution needs to be designed. Such solutions are usually not only unimodal and often present their output in graphical manner accompanied by speech output. Nowadays, a lot of systems enable also to control the system by touchscreens in combination with speech. Therefore user interfaces for cooperation with robots can be seen as a multimodal interactive system. The key components can be identified - input processing engines (automatic speech recognition, gesture recognition, touch recognition, etc.), multimodal fusion and fision components, dialogue manager, and output presentation engines (text-to-speech system, graphical output system). The proposed speech and multimodal technologies modules were realized and tested on the evaluation board provided, which is the computational center of the modular robotic system.

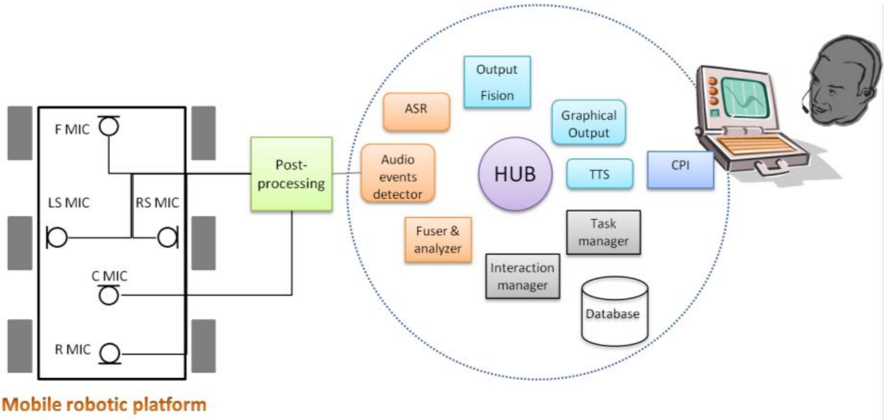


Figure 2
The robot speech interface

The newly designed robot multimodal interface (Figure 2) is de facto an asymmetric multimodal system, where speech is an input modality, and speech and graphical output are the output modalities. The system's input is extended with the acoustic event detection system, which carries the information about acoustic surrounding of the robot to the operator.

The Galaxy hub-server infrastructure [9] was adopted to construct the base structure of the system. The system designed consists of Galaxy hub process and nine servers:

- **Automatic speech recognition server** transforms speech signal into the corresponding text representation.
- **Audio events detector** serves for detecting of audio events as are gun shots or broken glass.
- **Fuser & analyzer** enable fusing several inputs into the corresponding events and perform analysis of incoming inputs. It transforms text representation of input into events, which represent voice commands for robotic system.
- **Output fision server** will serve for planning and fision of system outputs.
- **Text-to-speech server** performs speech synthesis, which transforms text output to speech signal.
- **Graphical output server** prepares text and graphical outputs, which should be presented to the user.
- **Interaction manager** is a part of the dialogue manager. It manages interaction by selecting appropriate tasks.
- **Task manager** is a part of the dialogue manager. It controls realization of simple tasks from which the interaction consists.
- **Control Panel Interface** (CPI) module was designed as an interface between the robot speech interface and control panel software.

To enable the control of the robot vehicle through voice commands, we decided to take an advantage of wireless close-talk headset to ensure as good SNR as possible. Usage of vehicle-integrated microphones directly for speech recognition purpose is not effective, because the acoustic surrounding of the robotic vehicle and also its own sound (from engine, movement, etc.) cause interferences to the input speech channel. Usage of wireless headset enables operator moving or standing close to the robot vehicle and expressing the voice commands as though directly to the robot. Of course, the signal from headset is still to be transmitted to the briefcase central control panel.

## 3.2 Automatic Speech Recognition

The Automatic speech recognition server is the most important part of the speech interface. Its performance (accuracy) determines successfulness of interaction. A lot of specific requirements can be identified on speech recognition technology for controlling the service robots, such as high accuracy, noise robustness, low hardware requirements, or the capability to run on specific hardware and software configurations. The ASR server was created by integrating an open-source solution into the Galaxy server. The advantage of using such approach is in the possibility to deliver audio stream both through Galaxy broker channel and directly through TCP/IP, also from remote locations.

For controlling robotic system by voice, rather than large vocabulary continues speech recognition system, the simpler recognition system being able to recognize isolated words and phrases should be used. Such systems are more accurate and also more robust, because of limited complexity and easier identification of start and end of commands.

The appropriate parameterization, acoustic and language models was designed and prepared for speech recognition process.

The most used parameterization, based on MFCC (Mel-frequency cepstral coefficients) was selected. The vector of parameters consist of twelve static MFCC coefficients, zero coefficient (0), delta (D), acceleration or acceleration coefficients (A) and subtraction of cepstral mean (Z) – (MFCC_D_A_Z_0) and its length is 39 values.

Acoustic models were trained on the telephone speech database SpeechDatE-SK [10] using training procedure from the COST-249 project [11]. Three states phoneme-based Hidden Markov models were prepared. We prefer phoneme models to triphone models for this task, because of their lower computation requirements. The accuracy of prepared models is higher than 95% (if SNR>20dB) for speech commands used for controlling SCORPIO service robot developed previously [7].

In case of the recognition system for recognizing a limited set of commands, the deterministic language model in form of context-independent grammar can be more effective and usable. The first step to building deterministic language model is performing analysis of system functions, which are considered to be controlled by voice. The analysis has shown that, there are six cooperation scenarios between robotic system and operator, which relate to:

- Movement of the robot
- Obtaining values from robot’s sensors
- Turning on/off robot’s devices and functions
- Manipulation with robotic arm
- Controlling of decontamination and extricate system
- Controlling of robot’s legs

A set of commands was designed for each scenario. Approximately 200 voice commands were constructed and the speech grammar was prepared, which defines a recognition network. When the operator wants to control a specific part of the system, they need firstly to say the name of that part, e.g. “arm” or “decontamination system”. Then they can start pronouncing commands for activated part, e.g.: “move right” or “lift the right front leg”.

The pronunciation dictionary was also prepared, which contains all words from recognition grammar and their phonetic transcription.

## 3.3 Acoustic Event Detection

Short audio segments which can represent a potential threat or other interesting events are considered as acoustic events. The possibility to monitor acoustic scene of robotic vehicle surrounding and to detect acoustic events significantly improve usefulness of the robotic system, especially in situation, when robotic vehicle is not visible (and not audible) for operator and it moves in remote (long distance) or closed environments. Usually the operator watches mainly the output of the camera in direction of vehicle movement and some events can occur wherever in environment. Almost every physical event has its own acoustic accompaniment, which can be detected from acoustic signal, e.g. breaking glass, explosions, crash, gunshot, calling for help or crying. Acoustic event detection systems (AEDS) enable to detect mentioned events and can be able to detect also direction to the source of incoming sounds.

The acoustic event detection system was designed for earlier described complex modular robotic system, which is intended to serve for monitoring of acoustic environment around the robotic vehicle and for delivering information about possibly interesting or dangerous events that occurred within this surrounding. AEDS consists of a set of microphones on the robotic vehicle; the post-processing unit and acoustic events detector (see Fig. 2).

The set of five microphones was designed, where four microphones are localized on every side of the vehicle and fifth microphone is placed in center of the vehicle and serves as reference microphone. This deployment enables to detect also the direction of incoming sound. The reference microphone will be used for compensation of the robot's own produced noise.

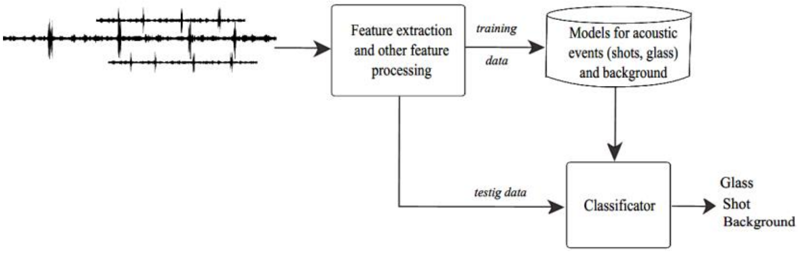

Feature extraction and other feature processing
training
data
Models for acoustic events (shots, glass) and background
testig data
Classificator
Glass
Shot
Background

Figure 3
Acoustic event detection

The post-processing unit performs a compensation algorithms for noise reduction and enables to estimate the direction of incoming sound. Robot's own noise is compensated in all channels using the central, reference microphone.

The third part of the AEDS is the detection system, which consists of the feature extraction and processing unit, and classification unit. Its principal scheme is depicted in the Fig. 3.

The feature extraction unit transforms input acoustic signal into the sequence of feature vectors, which are then send to the classification unit. The output of the system has a form of so-called "alerts", which inform the operator about recognition of specific acoustic event.

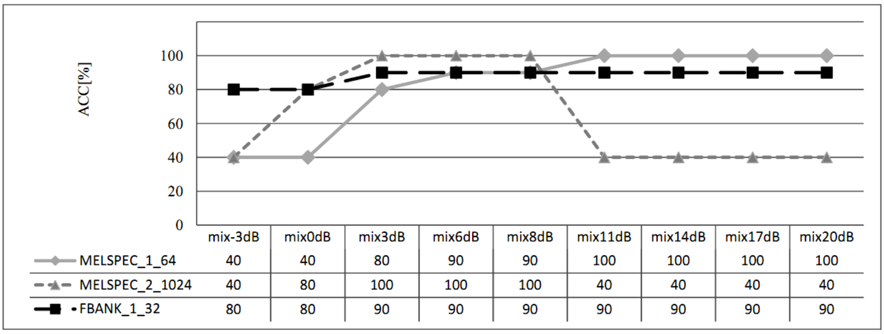
Different algorithms can be used to describe the nature of acoustic signals. Some of them are inspired by the descriptors defined in MPEG-7 standard or they are speech-based, such as MFCC (Mel-Frequency Cepstral Coefficients), or they are based on other temporal and spectral features. Deeper analysis was performed in [12], [13]. An exhaustive search for suitable parametric representation of selected set of acoustic events (gunshot and breaking glass) was performed. In our experiments we investigated different parametric representations of the selected acoustic events [14], [15], [16], [17]. Our research was concentrated mainly on the gunshots and breaking glass detection. Many tests were performed using various feature extraction algorithms and also with using of mutual information-based feature selection [16].

There are several approaches to audio event classification. Usage of Hidden Markov models (HMMs) [18], [19], Support Vector Machines (SVMs) [20] [21] or GMM binary trees [22] [23] are the most popular.

Due to our previous experiences, classification based on HMM (Hidden Markov Model) was designed and implemented. Basically the classificator is based on Hidden Markov Models (HMM) with modified Viterbi search algorithm that includes our segmentation function [24]. In more detail, our classificator uses HMMs with dictionary of audio events combined and converted into Weighted Finite-Sate Transducer (WFST) as a search network [25]. Classificator is receiving extracted input features vectors and building up possible hypotheses that are corresponding with input audio signal. When classification ends using our segmentation algorithm the most probable hypothesis is outputted. Then it continues with processing of the rest of the signal until next segmentation.

Two types of models were trained – model which represents background sounds (noise of the environment), and models which represent foreground events, that we need to detect. Ergodic HMM models for background, gunshot and sound of breaking glass were trained. Models in range of one to four states with 1 to 1024 PDFs (Probability Density Function) were prepared and tested.

A lot of experiments were performed with several types of parameterization, as are Mel-spectrum coefficients, Logarithmic Mel-spectrum coefficients, classical MFCC and other. Promising results were obtained with speech inspired features MELSPEC_DAZ (Mel-spectrum coefficients) and FBANK_DAZ (Logarithmic Mel-spectrum coefficients) with first (D) and second time derivation (A) and with cepstral mean normalization (Z). The perfect recognitions were achieved for SNR levels higher than 0dB [18]. The results achieved from the experiments are depicted in the following Fig. 4. The detection performance is limited by various factors, e.g. weather condition, SNR, sound similarities, overlapping events, etc. SNR impact was investigated in the works [15], [17].



| | mix-3dB | mix0dB | mix3dB | mix6dB | mix8dB | mix11dB | mix14dB | mix17dB | mix20dB |
|---|---|---|---|---|---|---|---|---|---|
| MELSPEC_1_64 | 40 | 40 | 80 | 90 | 90 | 100 | 100 | 100 | 100 |
| MELSPEC_2_1024 | 40 | 80 | 100 | 100 | 100 | 40 | 40 | 40 | 40 |
| FBANK_1_32 | 80 | 80 | 90 | 90 | 90 | 90 | 90 | 90 | 90 |

Figure 4

Accuracy of the audio events classification for different parameterizations and several SNR levels

To test newly-designed segmentation mechanism for classification process and to see if it interferes with detection accuracy, we conducted simple test on extended JDAE – TUKE database, described in [26]. Database contains (48 kHz, 16-bit per sample) recordings: 463 realizations of shots from commonly used weapons, 150 realizations of glass breaking and approximately 53 minutes of background sound. For testing purposes the database contains 13 glass breaking and 46 shots realizations. HMM models were trained using HTK environment [27] on MFCC feature vectors with zero, delta and acceleration coefficients resulting into 39 dimensional feature vectors. The segmentation algorithm was set to segment classification process and output results when the most probable hypothesis ending with background model and has lasted for 100ms time. For segmentation algorithm enabled and disabled we have achieved the same results 96.36% of accuracy. There was one missed alarm and 4 false alarms were produced [14].

## 3.4 Speech Synthesis

Text to speech (TTS) systems represent one of the most important parts of the speech interaction with robotic interface. Speech synthesizer enables to provide backchannel and information to the operator. It can be used also for indicating an occurrence of detected audio events.

The main objective of advanced speech synthesis nowadays is to use these artificial voices in various spheres of life without any limitation to their use in different situations. The essential requirements, which should be achieved in each TTS system, are the highest possible intelligibility, fluency and naturalness of speech at the output of this system.

The development of Slovak TTS systems for robotic interface is concentrated into two basic techniques of speech synthesis. The first voices, which were developed for this purpose, were based on the diphone concatenation method within the *Festival* TTS engine. In this case, it was necessary to create the voice data for

*Festival* (Slovak male and female diphone database) and additionally extend it using its embedded scheme scripting interface to incorporate the Slovak language support. The voices obtained by this approach can convert arbitrary text to speech with the quality corresponding to the diphone concatenation method [28]. Evaluation of voices using subjective and objective methods was conducted and from the obtained results it is clear that voices have acceptable intelligibility and fluency of speech at the output, but naturalness of speech significantly lags. The main advantages of this method are relatively small computing demands and small memory footprint, because the diphone concatenation method implemented in *Festival* uses LPC coefficients. Another advantage is the possibility to convert the obtained voices to the Festival-lite format, and we can use them with the *Flite* engine (cmuflite.org) which represents a small, fast run-time synthesis engine designed for small embedded machines.

Our recent research is focused on the creation of the new series of Slovak voices, which are based on the HMM-based speech synthesis method. This method uses the context-dependent HMM models, which are trained from speech corpus, as generative models for speech synthesis process [29]. We carried out some experiments with a relatively small database of Slovak language and the first female voices were built. The HMM-based TTS systems were developed by using several tools which are intended for this purpose. The quality of newly created voices was tested jointly with diphone voices and we can say that the HMM-based voices achieved much better results than diphone voices, even though the size of the input speech corpus in the case of HMM-based voices was not too large (speech corpus contained only 30 minutes of pure speech) [29]. Therefore, our work is now focused on building a bigger phonetically balanced speech corpus of male and female voice and we have to take into account that in case of development of the speech synthesis system for robotic interface, it is necessary to consider the fact, that we will probably synthesize short sentences or confirmation of commands, not long sentences nor a set of sentences. So the final speech database of both voices will consist of several variations of shorter sentences and confirmation.

## 3.5 Management of Interaction

The distributed dialogue manager, developed for communication with virtual agent [30], was adopted in simplified version also for management of interaction with the robotic system. The adopted solution is the distributed dialogue manager which consists of two agents – Interaction manager with user&history module included and Task manager. These agents cooperate over the common data space. Interaction is driven by events, which are represented by triggers and data objects. They could be caused by system, user or data. Triggers initiate small tasks. If user is passive, dialogue manager can invoke new tasks (system-initiative dialogue) by putting a new trigger into the trigger queue. The approach designed enables also building transition networks or state machine, which enables to create trees or networks, where tasks are nodes of such structure. Each transition holds also decision condition, which must be true, to use the concrete transition from one task to another.

**The interaction manager** is responsible for event loop mechanisms, which consist of initialization of the Interaction Pool, event-selection algorithm and destroying the pool. During the initialization phase initial set of triggers, which invokes a welcome task and the next tasks according user experience level are included to the trigger queue. Event-selection algorithm checks trigger queue and selects the next trigger, which will be handled by the Task Manager. Interaction manager has two algorithms for selection of triggers. The first one searches for the first unhandled trigger in the trigger queue. When the trigger handling is finished, the second algorithm passes the transition network and searches for the next node (task), which should be processed. The transition condition is checked to evaluate if the transition is possible.

**The task manager** is responsible for performing the particular tasks, which were selected by the interaction manager. The task handling mechanism has three fundamental algorithms – data object values collection, cooperation with external data and output concepts generation. Each task can have a general prompt, which introduces particular task. Tasks can require filling zero or more data objects, which are attribute-value pairs for holding information obtained from user, related to particular task. When specific (or all) data objects has filled their values, task manager can perform specific function, e.g. writing data to the database, consult database and obtain information for user, or it can perform transition to other task or simply nothing.

The approach proposed for management of interaction enables to control interaction in simple, effective and variable manner. The interaction can be controlled solely by the user (user-initiative interaction) by his utterances, from which triggers are extracted and conveyed into the trigger queue. Manager also enables strictly system initiative dialog, where the dialog flow is determined by transition network (system-initiative interaction). The mixed-initiative dialogues are allowed by combining previous scenarios too.

## 3.6 The Interaction

Four tasks (*ObtainCom, GiveBCh, AckCom, SendCom*) as well as transition network (Fig. 5) were prepared for controlling the interaction with the complex robotic system.

Each task performs specific small interaction. *ObtainCom* task serves for obtaining voice command from the operator. According to user’s experience level, it can prompt the user to say the command, or to leave him to be active and the system only waits for the command. *GiveBch* (Give backchannel) task repeats the recognized command to inform operator about what was recognized. Then, the interaction is moved to the AckCom (Acknowledge command) task, where the system waits (during the specified time interval) on acknowledgment of the command. If the command is confirmed, SendCom (Send command) task is invoked and command is sent to the central control panel software, which performs it.

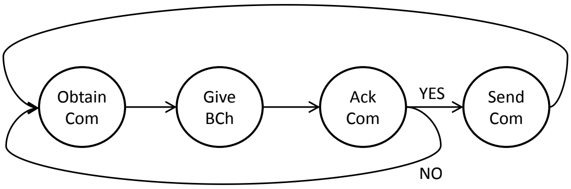


Figure 5
The transition network of tasks for controlling the robot using voice commands

Two examples of interaction are described below. The scenario no. 1 shows the base version, where the system firstly prompts the operator to say the command. The scenario no. 2 shows the scenario, when the operator is active and also the situation, when the command is not recognized correctly and the operator needs to repeat it.

**Scenario 1:**

S: Say your command. (*ObtainCom* task)

U: Robot, turn right. (*ObtainCom* task)

S: Robot, turn right. (*GiveBCh* task)

U: *(User pushes "dead man" button on the top of the joystick within a specified time interval)* (AckCom task)

System perform required task. (*SendCom*)

**Scenario 2:**

U: Robotic arm, base position. (*ObtainCom* task automatically invoked)

S: Arm, base position (*GiveBCh* task)

U: *(User pushes "dead man" button on the top of the joystick within a specified time interval)* (*AckCom* task)

System performs the required task. (*SendCom*)

U: Front camera (*ObtainCom* task)

S: Rear camera (*GiveBCh* task)

U: Front camera (*ObtainCom* task)

S: Front camera (*GiveBch* task)

….

As it is presented, the transition network enables to concatenate several tasks together, but the flow of the interaction is more flexible, because the user can in each state simply say the new command, and the first task (*ObtainCom*) is invoked.

Each task can be simply modified to change its prompts or data objects. The flexibility is assured by the possibility to access the database, where tasks as well as transition network are defined. The database can be modified also on the fly, which enables immediately react on situation.

## Conclusion

The modules for implementation of speech technologies enabling cooperation of the human operator and the complex modular robotic system were introduced and evaluated on evaluation board provided. Presented modules are organized in the multimodal interface for controlling functions of the robotic system. All modules work well in laboratory conditions. However, to evaluate usability of the interface and to find possible sources of problems, tests and evaluation in real outdoor environment will be performed after final prototype of the whole robotic system will be introduced.

The special attention will be paid to the robustness of the speech recognition system [31] as well as to cognitive multimodal interaction aspects [32], to ensure ergonomics and usability of designed interface. Our further work will be also focused on design and development of the interface application for touchscreens of modern mobile tablet environment.

## Acknowledgement

The research presented in this paper was supported partially (50%) by Research & Development Operational Program funded by the ERDF through project *Research of modules for intelligent robotic systems* (ITMS project code 26220220141) and partially (50%) by *Competence Center for Innovation Knowledge Technology of production systems in industry and services* (ITMS project code 26220220155).

## References

[1] Broadbent E., Stafford R., MacDonald B., “Acceptance of Healthcare Robots for the Older Population: Review and future directions”, International Journal of Social Robotics, 1 (4), pp. 319-330, 2009

[2] Boissy P., Corriveau H., Michaud F., Labonté D., Royer M.-P., “A Qualitative Study of In-Home Robotic Telepresence for Home Care of Community-Living Elderly Subjects”, Journal of Telemedicine and Telecare, 13 (2), pp. 79-84, 2007

[3] Shibata T., Wada K., “Robot Therapy: A New Approach for Mental Healthcare of the Elderly - A mini-review”, Gerontology, 57 (4), pp. 378-386, 2011

[4] McColl D., Louie W.-Y.G., Nejat G., “Brian 2.1: A Socially Assistive Robot for the Elderly and Cognitively Impaired” IEEE Robotics and Automation Magazine, 20 (1), art. no. 6476702, pp. 74-83, 2013

[5] McTear M. F., “Spoken Dialogue Technology: Enabling the Conversational User Interface”, ACM Computing Surveys, 34 (1), pp. 90-169, 2002

[6] Kotus J., Lopatka K., Czyzewski A., "Detection and Localization of Selected Acoustic Events in Acoustic Field for Smart Surveillance Applications", Multimedia Tools and Applications, p. 17, online first, 2012

[7] Ondas S. et al., “Service Robot SCORPIO with Robust Speech Interface”, International Journal of Advanced Robotic Systems. Vol. 10, art. no. 3, pp. 1-11, ISSN: 1729-8806, 2013

[8] Vaughan B., Han J. G., Gilmartin E., Campbell N., “Designing and Implementing a Platform for Collecting Multi-Modal Data of Human-Robot Interaction“, Acta Polytechnica Hungarica, Special Issue on Cognitive Infocommunications, Vol. 9, No. 1, pp. 7-17, 2012

[9] Seneff S., Hurley E., Lau R., Pao C., Schmid P., Zue V., “Galaxy-II: a Reference Architecture for Conversational System Development.”, In Proceedings of the 5th International Conference on Spoken Language Processing – of ICSLP’98, Sydney, Australia, pp. 931-934 ,1998

[10] Pollak P. et al., “SpeechDat(E) Eastern European Telephone Speech Databases” Proceedings of LREC Satellite workshop XLDB, Athens, Greece, pp. 20-25, 2000

[11] Johansen F. T. et al., “The COST 249 SpeechDat Multilingual Reference Recogniser” LREC - International Conference on Language Resources and Evaluation, Athens, Proceedings Vol. 3, pp. 1351-1355, 2000

[12] Kim H. G., Moreau N., Sikora T., “MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval,” New York: Wiley & Sons, p. 304, 2005

[13] Toh A. M., Togneri R., Nordhoolm S., “Investigation of Robust Features for Speech Recognition in Hostile Environments.” Asia-Pacific Conference on Communications, Perth, art. n. 1554204, pp. 956-960, 2005

[14] Vozarikova E., Juhár J., Čižmár A., “Acoustic Events Detection Using MFCC and MPEG-7 Descriptors,” Multimedia Communications, Services and Security, Springer, Vol. CCIS 149, pp. 191-197, 2011

[15] Vozarikova E., Pleva M., Vavrek J., Ondáš S., Juhár J., Čižmár A., “Detection and Classification of Audio Events in Noisy Environment,” Journal of Computer Science and Control Systems, 3(1), pp. 253-258, 2010

[16] Vozarikova E., Juhár J., Čižmár A., “Performance of Basic Spectral Descriptors and MRMR Algorithm to the Detection of Acoustic Event,” Multimedia Communications, Services and Security, Springer, Vol. CCIS 287, pp. 350-359, 2012

[17] Kiktova E., Lojka M., Pleva M., Juhar J., Cizmar A., “Comparison of Different Feature Types for Acoustic Event Detection System,” Multimedia Communications, Services and Security, Springer, Vol. CCIS 368, pp. 288-297, 2013

[18] Zhou X., Zhuang X., Liu M., Tang H., Hasegawa-Johnson M., Huang T., “HMM-based Acoustic Event Detection with AdaBoost Feature Selection,” Multimodal Technologies for Perception of Humans, Springer, Vol. LNCS 4625, pp. 345-353, 2008

[19] Ntalampiras S., Potamitis I., Fakotakis N., “On Acoustic Surveillance of Hazardous Situations,” ICASSP: IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, pp. 165-168, 2009

[20] Bach J.-H., Kayser H., Anemuller J., “Audio Classification and Localization for Incongruent Event Detection,” Detection and Identification of Rare Audiovisual Cues, Springer, Vol. SCI 384, pp. 39-46, 2012

[21] Giannakopoulos T., Kosmopoulos D., Aristidou A., Theodoridis S., “Violence Content Classification Using Audio Features” 4th Helenic Conference on AI, SETN, Heraklion, LNCS 3955, pp. 502-507, 2006

[22] Atrey P. K., Maddage N. C., Kankanhalli M. S., “Audio-based Event Detection for Multimedia Surveillance,” ICASSP: IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 5, pp. V813-V816, 2006

[23] Vavrek J., Vozáriková E., Pleva M., Juhár J., “Broadcast News Audio Classification Using SVM Binary Trees,” TSP: 35th International Conference on Telecommunications and Signal Processing, Prague, July 3-4, art. no. 6256338, pp. 469-473, 2012

[24] Pleva M., Lojka M., Juhar J., Vozarikova E., “Evaluating the Modified Viterbi Decoder for Long-Term Audio Events Monitoring Task” Elmar - International Symposium: Electronics in Marine, art. no. 6338500, pp. 179-182, 2012

[25] Lojka M., Juhár J., “Fast Construction of Speech Recognition Network for Slovak Language,” Journal of Electrical and Electronics Engineering, Vol. 3, No. 1, pp. 111-114, 2010

[26] Pleva M., Vozáriková E., Doboš L., Čižmár A., “The Joint Database of Audio Events and Backgrounds for Monitoring of Urban Areas,” Journal of Electrical and Electronics Engineering, Vol. 4 (1), pp. 185-188, 2011

[27] Young S., Kershaw D., Odell J., Ollason D., Valtchev V., Woodland P., The HTK Book Version 3.4, Cambridge University Press, 2006

[28] Sulír M., Juhár J., Ondáš S., “Speech Synthesis Evaluation for Robotic Interface”, Complex Control Systems, Vol. 11(1), BAS, pp. 64-69, 2012

[29] Zen H., Oura K., Nose T., Yamagishi J., Sako S., Toda T., Masuko T., Black A. W., Tokuda K., "Recent Development of the HMM-based Speech Synthesis System (HTS)" APSIPA ASC: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Sapporo, pp. 121-130, 2009

[30] Ondáš S., Juhár J., "Design and Development of the Slovak Multimodal Dialogue System with the BML Realizer Elckerlyc" CogInfoCom 2012: 3rd IEEE International Conference on Cognitive Infocommunications, December 2-5, Košice, pp. 427-432, 2012

[31] Staš J., Hládek D., "Recent Progress in Language Modeling and Continuous Speech Recognition of the Slovak Language", SCYR: Scientific Conference of Young Researchers of Faculty of Electrical Engineering and Informatics, FEI TU, Košice, pp. 339-342, 2011

[32] Baranyi P., Csapó A., "Definition and Synergies of Cognitive Infocommunications", Acta Polytechnica Hungarica, 9 (1), pp. 67-83, 2012