Implementing English Speech Interface to Jaguar Robot for SWAT Training

Matus Pleva, Jozef Juhar, and Anton Cizmar
Department of Electronics
and Multimedia Communications,
Faculty of Electrical Engineering and Informatics,
Technical university of Kosice,
Letna 9, 04120 Košice, Slovakia.
Email: matus.pleva@tuke.sk,
jozef.juhar@tuke.sk, anton.cizmar@tuke.sk

Christopher Hudson, Daniel W. Carruth, and Cindy L. Bethel
Center for Advanced Vehicular Systems,
High Performance Computing Collaboratory (HPC^2),
Bagley College of Engineering,
Mississippi State University,
200 Research Blvd., Starkville, MS 39759.
Email: chudson@cavs.msstate.edu,
dwc2@cavs.msstate.edu, cbethel@cse.msstate.edu

***Abstract*—This paper describes the development of a specialized application for voice command recognition for the Jaguar V4 robot in conjunction with the Starkville, MS, USA Special Weapons and Tactics (SWAT) team during training. This training took place at The Center for Advanced Vehicular Systems (CAVS), which provides a specialized environment for police SWAT training. This reconfigurable space, setup during this study as a two bedroom apartment, includes video monitoring of the space, sound playback and capturing, reconfigurable lighting, etc. This training environment is used for testing different kinds of human-robot interfaces in SWAT training operations. The results of the voice integration evaluation indicated that voice commands could be successfully used for controlling additional functions of the robot after a short introductory training session with a few of the police officers. These preliminary observations were encouraging and provides support for further investigation into the usefulness of this technology.**

***Keywords*—Human-Robot Interface, Automatic Speech Recognition, Robotics**

I. Introduction

The use of robots in army and police operations is a wide research and development topic, which includes a variety of robotic platforms, with the traditional focus around Unmanned Aerial Vehicles - UAVs [3] and drones [15]. The use of robots in SWAT operations has become more attractive with the increased availability and configurability of robots [8].

In law enforcement operations, there is a subset of situations that are common. These situations usually include finding a victim or suspect in a building, forcing a suspect to drop a weapon, and having a suspect exit a building with their hands in the air. Entering a building is always a risky operation for police officers and utilization of robots should lower the risk of injuries that may be associated with entering a building [6].

As this field advances, new interfaces need to be designed to allow for better integration of robots with SWAT teams. Using the reconfigurable testbed located at the Center for Advanced Vehicular Systems (CAVS), interface designs can be evaluated [1] and iterated on quickly and reliably [14]. The team from the Technical University of Košice during a visiting scholarship at CAVS, provided the knowledge of voice command integration [9], software development for this technology, and the development of possible use case scenarios for the integration of voice commands for robots used in SWAT trainings.

II. Jaguar V4 Robot Setup

The robot used for the integration of the voice control software was a Jaguar V4 by Dr.Robot[1] (Figure 3). The Jaguar v4 is a treaded robot with two sets of flippers for maneuverability. It is capable of operating for several hours off a single battery charge, which is helpful in law enforcement responses. The Jaguar V4 for this integration was configured with two cameras, a forward facing upward camera (Figure 1), and a forward facing drive camera located in the base of the robot (Figure 2).

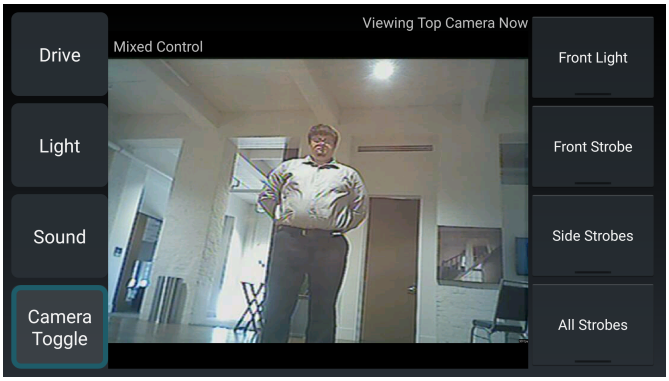

Fig. 1. Forward Facing Upward Camera View

The forward facing upward camera (Figure 1) is used to identify suspects, and any threats they might pose to the officers. This viewpoint, allows the officers to see the upper body and hands of a suspect to determine if they are carrying weapons or other potential threats. The forward facing drive camera (Figure 2) is used to help navigate through the environment. The Jaguar v4 was outfitted with distraction devices for use with the SWAT team, which included an LED forward

[1] http://jaguar.drrobot.com/specification_V4.asp

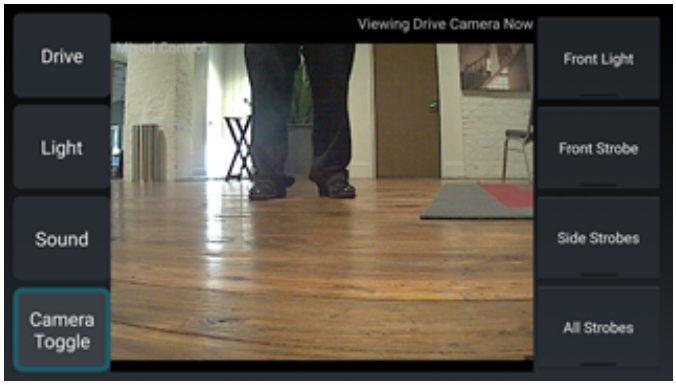

Fig. 2. Forward Facing Drive Camera View

facing light bar, two side facing blue/white high intensity LED strobe lights, and a loudspeaker (Figure 3). The loudspeaker provides the ability to play sirens, alarms, and other sounds to cover the movements of the officer and also to distract potential suspects.

Control of the Jaguar V4 is comprised of three core component systems. These three system components handle specific tasks in order to support a fully functional robot for SWAT team use. These systems are the base robot (Jaguar V4) for navigation and reconnaissance, the laptop system running the Robot Operating System (ROS) [12] for robot control, and the Arduino Nano [2] for the control of the distraction devices. These components work together to provide a robot capable of searching a building in conjunction with the SWAT team. Each component is described in detail below.

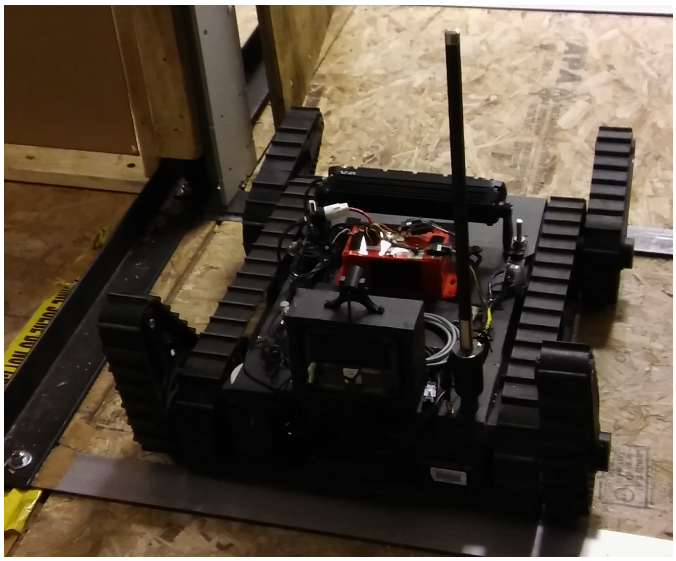

Fig. 3. Jaguar V4 robot with Arduino, lights, and loudspeaker upgrade approaching the breaching door into the CAVS testbed apartment

[2]https://www.arduino.cc/en/Main/ArduinoBoardNano

A. Jaguar V4 Robot Base Unit

The base unit of the robot is responsible for the execution of all movement commands received from the backpack control unit. It takes any message received and translates the commands to activate movements in the motors and flippers on the robot using ROS. Additionally, the base unit of the robot houses the cameras used for navigation and information gathering. The cameras located on the robot are Internet Protocol (IP)-based units, which are accessible by anyone on the network with the login information to receive the data. The video from the cameras has been designed to stream to an Android device that is wearable by the officers to provide them with critical information about what is happening in the view of the robot cameras prior to their entry into a potentially dangerous environment. By sending the robot in first, officers can gain an understanding of what they may encounter upon entry and helps to keep them safer.

Fig. 4. Nunchuck movement remote

B. Backpack Unit

The base robot only executes instructions it receives from the control unit. The control unit for the Jaguar V4 is a backpack unit that can be worn by the officers. The control hardware was put into a backpack to make it portable for use by the SWAT team. In the backpack is a laptop installed with an Ubuntu Linux distribution and ROS, which acts as the brains of the Jaguar V4 robot. ROS is a message passing protocol which defines a standard means for communication for robotic systems [12]. It uses a publish/subscribe protocol in which messages are place into a structure known as a topic. A topic distributes the message to any subscribing node, which wants that message type. This allows for multiple nodes to receive (subscribe to) a copy of a message that has been distributed (published) by the system. These standard messages can take many forms, including image messages, movement commands, Inertial Measurement Unit (IMU) data, Global Positioning System (GPS) data, and light detection and ranging (LIDAR) sensor data. Nodes that produce and publish these messages, receive the corresponding data from the sensor on the Jaguar base unit, and translate the data into the ROS message for that data type. These nodes, then take this message

which now conforms to the ROS protocol and pushes them into a topic. Any nodes that wish to consume this data, can subscribe to the topic. When a new message is published to a topic there is a callback function, which gives the node access to that data. A node can be both a publisher and subscriber. An example of this type of node is an autonomy node, which listens for LIDAR, GPS, and odometry data, then uses this data to make decisions about movements to be performed. Once a decision about a movement or set of movements is made from this data, these nodes can publish a message to a movement topic, to direct the robot to move in a specific way. The power in using ROS is that packages, which are written in ROS are generic for any type of robot, such that a robot can use any package it wants, so long as it provides the corresponding sensor input required.

Fig. 5. Typical accessories used during SWAT trainings: Wiimote/Nunchuck, Android mobile device, external antenna system & headphones with microphone.

C. Arduino Nano

Attached to the top of the Jaguar V4 Base Unit is an Arduino Nano inside of a protective container. This Arduino is used to power and control the light and sound systems attached to the base unit of the robot. The Arduino stores audio files (e.g., siren, alarm, dog bark, and footsteps) on an integrated microSD card for playback when activated by the officers through the Android mobile device. The Arduino Nano listens for HTTP requests to playback these files, or to activate the lights in either a strobe pattern or as a constant spotlight.

D. Accessories

Additional interfaces may be attached to the system. In the evaluation setup for the research presented in this paper, a headset (Figure 7) was attached to the backpack/laptop unit, several android devices were used, a Wiimote and Nunchuck system was used for teleoperation (Figure 4). These were connected to the backpack and Arduino units for the control and activation of the robot and the distraction devices. The system overview for this configuration can be seen in Figure 6. The Android interface (Figures 1 and 2) and Android mobile devices (Figure 5) are used for watching the camera streams from the robot and then executing commands such as playing a sound or turning on a light.

This research identified some challenges with this current approach for the control of the robot and the distraction devices, that may be addressed through the use of speech resources. One challenge is that using an Android mobile device puts officers in potential danger because the screen can “backlight” the officers making them identifiable targets. Another challenge is that it is difficult to use a touchscreen device due to the equipment and gloves the officers wear as part of their standard issue gear. The current approach may also prove to be distracting when in actual response conditions.

III. Speech Resources Development & Implementation

To accomplish the goal of integrating voice control software into the existing platform, it was important to understand how to best incorporate voice commands into the system [13]. Given the proposed integration, it was decided to operate the robot partially by voice to activate the distraction devices such as lights and sounds. The decision to use voice commands to activate lights and sounds was made in order to move toward a system in which officers would have minimal deviation from their trained protocols for clearing a building. By removing the need to interact physically with a device in order to activate a distraction device, the overall time the officers had to spend interacting with the robot was reduced.

Integration of the voice control software in this dynamic, unpredictable, and often noisy environment lends itself to unique problems, one of these problems being noise. Several steps where taken to reduce the overall noise effecting the system, the first of these steps was to use a keyword. Using a keyword allowed the system know that the officers where about to issue a command to the robot, and not simply talking to each other. It was important to select a keyword that is not frequently used by the officers during a police operation so that anytime it was said, it was understood by the system and the officers that whatever follows is directed toward the robot only. For this integration, the word ’Apple’ was used in the initial integration, other words evaluated for use where ’System’, Houston’, and ’Dragon’. Following the use of this keyword, a command could be issued for execution by the system to activate a light or sound.

A. Grammar

During the initial implementation which served as a proof of concept, nine commands where evaluated: Down, Dark, Suspect, Be, Flash, Alarm, Quiet, Siren and Lights. After a demonstration and discussion, these nine commands where further expanded, and a total of twenty commands where implemented, evaluated, and tested during a SWAT training session. The results of the evaluation from this training session revealed that it was more useful to have a small subset of commands, opposed to a larger set of commands. This discovery resulted from the officers’ difficulty in remembering the full list of commands available to them during the stress-filled training exercise meant to mimic a real response scenario.

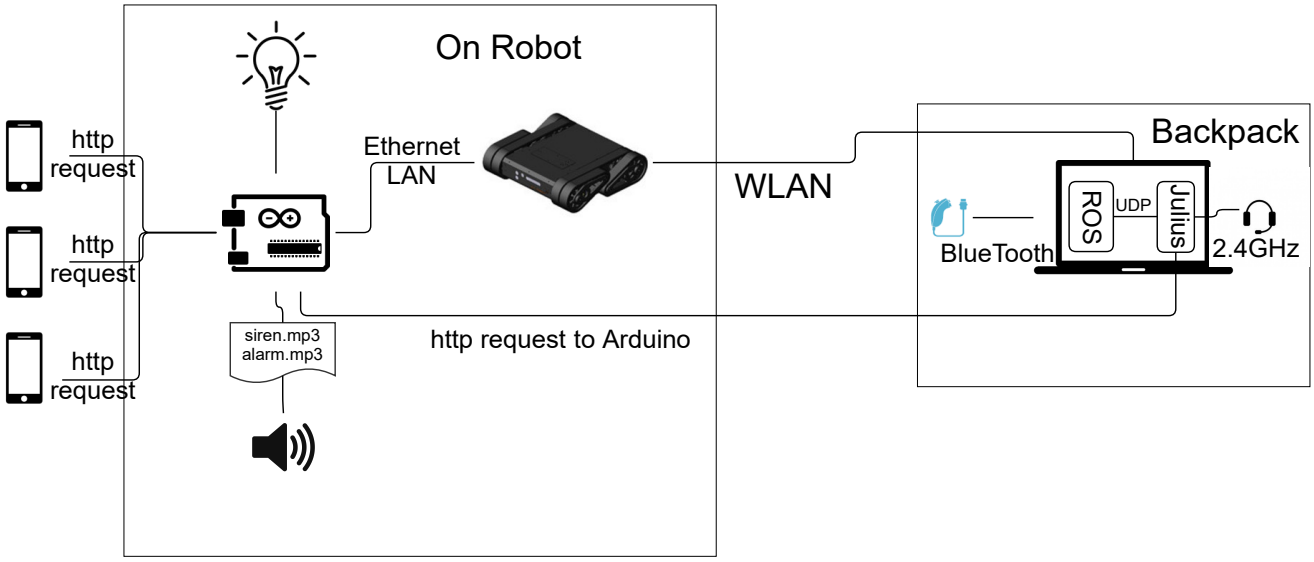

Fig. 6. System Overview

B. Acoustic Model

The acoustic model was chosen from the open source community as the first option. But the chosen VoxForge[3] acoustic model had poor quality and did not provide the needed tri-phones (combination of phonemes) for the word siren during the proof of concept testing. Next, the freely available WSJ+TIMIT baseline models [16] from Keith Vertanen were used.

However, it was noted during testing that when the robot's siren was playing in the background, the Julius recognition system had difficulty finishing the hypothesis and sent no results for several seconds. To address this issue, it was decided to use noise models from the TUKE (Technical University of Košice) acoustic models repository [11]. When the noise model was joined to the original acoustic models, the final combined model was able to operate better during noisy situations such as those encountered in SWAT operations.

C. Automatic Speech Recognition Engine

For the task of Automatic Speech Recognition (ASR) it is important that a stable engine is selected with the capability to send the recognized command to another application. For this purpose we selected the Open-Source Large Vocabulary CSR Engine Julius [4] which we has been previously used for the development of the SCORPIO robot speech interface in the Slovak language [10]. The recognition process was started simultaneously with the ROS control software in the operator command backpack. For interacting with the voice control software, a wireless (Figure 7) and/or wired headset was connected as a peripheral and the microphone input was activated.

[3]VoxForge - Open Source Free Speech Resources for Automatic Speech Recognition (Linux, Windows, and Mac), http://www.voxforge.org/

Fig. 7. Wireless headphones used for voice commands

IV. Automatic Speech Recognition Integration

The Julius open source Large Vocabulary Continuous Speech Recognition (LVCSR) [5] software was integrated with the Jaguar V4 robot through a multi-step process. First, we established direct communication between the Julius software and the micro-controller, an Arduino Nano, on the Jaguar V4 robot that controlled the activation of the lights and sounds. The Julius LVCSR software recognized specified keywords and commands and activated the lights and sounds using HTTP requests. For example, using very simple keywords and commands such as APPLE LIGHTS we were able to activate the strobe lights on the Jaguar V4. We were able to quickly demonstrate and test basic voice communication with the robot and onboard systems. A demonstration video is available for review at the following link: http://bit.ly/1sxaqcK.

To complete the integration of Julius with the Jaguar V4, it was necessary to extend beyond sending commands to activate lights and sounds through the onboard Arduino Nano. A full integration would mean that the command and keyword captured by Julius was sent to ROS, the control software. To accomplish this, a message containing four pieces of information was sent via UDP from Julius to ROS. This

message was formatted as follows: a keyword, a command, an options field, and a probability. A UDP connection was selected as the message transport system because Julius is a stand alone program, and by keeping the two pieces of software separate, it allowed for future reconfiguration of the system, should the Julius or ROS software be moved to a different computer. Once the message was received by ROS, it was parsed into a custom ROS message, which was placed in a topic so that any node which was written in the control software that wished to use it could access the voice inputs as they were parsed by Julius.

To confirm that the integration was working as intended, a series of nodes were written to use the commands parsed by the Julius system and received by ROS. The node written activated the lights or sounds, in the same way the direct connection with the Arduino Nano would accept them, with one small addition, it gave the operator the ability to override the choice by selecting a button on the controller. The successful implementation and use of this node demonstrated that the Julius software was fully integrated and operational with the Jaguar V4 robot.

V. Evaluation of the System During SWAT Training

The integration of Julius with the Jaguar V4 robot was sufficiently advanced to perform a demonstration the system during a SWAT training at CAVS. Additionally a new streaming video integration to Google Glass[4] smart glasses was implemented from the front camera of the robot. Together with voice interaction, the solution brought a new dimension to the SWAT training, because the operator did not need to watch or touch a display of another device (smart phone, PDA, etc.) and could be fully concentrated on the crime scene and his weapons. Next, the backlighting of the conventional mobile display was a important drawback because the SWAT team must be hidden in the environment and the back part of the smart glass displays could be covered to prevent that effect in low-to-no light conditions often observed in SWAT operations (Figure 8 shows the night vision testbed monitoring from different camera perspectives).

Fig. 8. Night vision monitoring of the testbed during actual SWAT training operations

[4]https://developers.google.com/glass/

Multiple SWAT officers were invited to interact directly with the system and provide feedback on the keywords, commands, and actions they wished to see implemented in the system. This training provided great insight into the potential issues that could arise when integrating such a system into a tactical team, especially when noise is a particular issue of concern. The results of this training informed the design process moving forward for the integration of the voice command system. Different kinds of microphones are being explored, as well as software and hardware solutions for limiting the amount of noise introduced into the system as a result of being in a dynamic and loud environment.

VI. Conclusion

The successful integration of the Julius software with the Jaguar V4 robot was accomplished and shown to be operational in training sessions at CAVS with the Starkville, MS, USA Police Department's SWAT team. The feedback and training sessions informed the design of the integration of the voice control system with the usage scenarios presented by the SWAT team. Deployment of this system in the training session revealed several areas for future work going forward. The use of voice commands may enhance human-robot interaction as it relates to the integration of robots with SWAT teams. The use of voice commands reduces the need for touchscreen displays, which have the inherent problem of "backlighting" officers and putting them at risk. Overall the integration of the Julius software was deemed a success, and provided a new means of interaction for the SWAT officers with the Jaguar V4 robot.

VII. Future Work

- Training Tool – CAVS team will develop a training tool to allow officers to work with the voice recognition software outside of training. The training tool will present a voice command cue and a description or visualization of the effect of the command. The user will speak the command. The tool will record the audio and present feedback to the user as to whether the command was successfully recognized.
- Additional Interface Designs – CAVS team will explore additional interface designs for use of the voice control system. In the prototype design, the robot was always listening, was prompted using the keyword + command combination, and no feedback was given by the ASR system. Alternative designs include using a button press to activate listening or, as previously used by TUKE, having the robot repeat back the command and having the user press a button to confirm the action [10].
- Collection of Audio Data for Improved Acoustic Models/Improved Command Recognition – Using the training tool and through recordings of future police trainings, Mississippi State will collect recordings of officers attempting voice commands and of ambient noise during trainings. These recordings will be made available to the TUKE team for consideration to improve models and command recognition.

- Identification of Opportunities for User Studies – Both teams will look for opportunities to present reports on current and future efforts resulting from his collaboration[5]. A user study is planned at CAVS under the supervision of their Human Factors and Ergonomics research group.
- Connection with Database – The speech interface could be used in a more valuable way if there was a connection to the Database with the possibility to simply query information or data using only the voice commands and developed NLP (Natural Language Processing) analysis [2] or Speaker Identification analysis from robot's microphone [7]. The initial idea was to authenticate a suspect using his voice entered ID number and provide the police officers with his name and image from a police or institution private database and then display this information to the robot operator.
- Noise Suppression and Microphone Selection – During the SWAT training exercises and testing, it was determined that the noise and voices in the background decreased the quality of the ASR output but also affected the ability of the recognition engine to finish the current hypothesis, which led sometimes to freezing of the interface until it received a less noisy input. A direct and most effective method was to use a dynamic microphone headset, which was much less sensitive to background noise because of its physical design. Next, a specialized noise suppression microphone headset will be tested and compared with common capacitive microphones in the headset available off-the-shelf.

Fig. 9. Fish eye optic image capture of the office room entrance

- Fish Eye Optics – During the SWAT training session, testing was performed with the use of an off-the-shelf fish eye optics that was attached to the robot camera. It demonstrated that using the forward upward facing camera with the fish-eye optics could be enough for navigating the robot and identifying a suspect. In case of using the fish eye optics, no switching of the cameras was needed. The distortion of the space was visible and as a wide angle lens, it provided from 120 - 360 degrees for field of view, but the human brain was very adaptible to this distortion (Figure 9). We plan to evaluate more types of wide angle optics and test the human experience after extended usage of these devices.

[5]Virtual Collaboration Arena, www.virca.hu

ACKNOWLEDGMENT

The research presented in this paper was supported by the Slovak Research and Development Agency under the research project APVV-15-0731 and by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the projects VEGA 1/0075/15 & KEGA 055TUKE-4/2016.

The authors would like to thank Dr. Bohumir Jelinek, former Ph.D. student of Prof. Čižmár from Technical University of Košice and CAVS Director Dr. Roger King, who provided the opportunity for this joint research project. Additionally, we would like to thank the Starkville, MS, USA Police Department and the SWAT team members for their assistance in this collaborative research effort.

REFERENCES

[1] C.L. Bethel, D. Carruth, T. Garrison. *Discoveries from integrating robots into SWAT team training exercises.* In Safety, Security, and Rescue Robotics (SSRR), 2012 IEEE International Symposium on, p. 8, 2012.

[2] D. Hladek, S. Ondas, J. Stas, *Online natural language processing of the Slovak language*, In proceedings: Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on, pp. 315–316, 2014.

[3] H.L. Jones, S.M. Rock , D. Burns, S. Morris, *Autonomous robots in swat applications: Research, design, and operations challenges*. In proceedings: Proceedings of the 2002 Symposium for the Association of Unmanned Vehicle Systems International (AUVSI '02), Orlando, Florida, p. 15, 2002.

[4] A. Lee, T. Kawahara, *Recent development of open-source speech recognition engine Julius*. In proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, Sapporo, Japan, pp. 131–137, 2009.

[5] A. Lee, T. Kawahara, K. Shikano, *Julius an open source realtime large vocabulary recognition engine*, In proceedings: EUROSPEECH, 7th European Conference on Speech Communication and Technology, September 3 - 7, 2001, Aalborg Congress and Culture Centre, Aalborg, Denmark, pp. 1691–1694, 2001.

[6] C. Lundberg, H. I. Christensen, *Assessment of Man-Portable Robots for Law Enforcement Agencies,* In proceedings: Performance Metrics and Intelligent Systems (PerMIS'07) Workshop, Gaithersburg, p. 8, 2007.

[7] L. Mackova, A. Cizmar and J. Juhar. *A Study of Acoustic Features for Emotional Speaker Recognition in I-Vector Representation*, Acta Electrotechnica et Informatica, 15 (2), pp. 15–20, 2015.

[8] R. R. Murphy, *Human-robot interaction in rescue robotics,* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34 (2), pp. 138–153, 2004.

[9] S. Ondas, et al. *Service robot SCORPIO with robust speech interface,* International Journal of Advanced Robotic Systems (vol. 10, no. 1), InTech, SAGE Publishing, p. 11, 2013.

[10] S. Ondas, et al., *Speech technologies for advanced applications in service robotics,* Acta Polytechnica Hungarica, 10 (5), pp.45–61, 2013.

[11] M. Pleva, J. Juhar, *TUKE-BNews-SK: Slovak Broadcast News Corpus Construction and Evaluation*, In proceedings: LREC 2014, Ninth International Conference on Language Resources and Evaluation, ELRA, Reykjavik, pp. 1709–1713, 2014.

[12] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, et al., *ROS: An Open-Source Robot Operating System,* In proceedings: ICRA Workshop on Open Source Software, Kobe, Japan, p. 5, 2009.

[13] M. S. Maucec, Z. Kacic, and A. Zgank. "Speech recognition for interaction with a robot in noisy environment." Przeglad Elektrotechniczny, 89 (5), pp. 232–236, 2013.

[14] P. Smolar, J. Tuharsky, Z. Fedor, M. Vircikova and P. Sincak, *Development of cognitive capabilities for robot Nao in Center for Intelligent Technologies in Kosice,* Cognitive Infocommunications (CogInfoCom), 2011 2nd International Conference on, Budapest, IEEE, p. 5, 2011.

[15] G. G. De la Torre, M. A. Ramallo, E. Cervantes, *Workload perception in drone flight training simulators*, Computers in Human Behavior, Volume 64, November 2016, pp. 449–454, 2016.

[16] K. Vertanen, *Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments*, Technical report, Cambridge, United Kingdom: Cavendish Laboratory, 2006.