INTECH
open science | open minds

OPEN ACCESS ARTICLE

International Journal of Advanced Robotic Systems

Service Robot SCORPIO with Robust Speech Interface

Regular Paper

Stanislav Ondas[1,*], Jozef Juhar[1], Matus Pleva[1], Anton Cizmar[1] and Roland Holcer[2]

1 Technical University of Kosice, Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics, Kosice, Slovakia
2 ZTS VVU KOSICE a.s., Research, Development, Design & Supply Company, Kosice, Slovakia
* Corresponding author E-mail: Stanislav.Ondas@tuke.sk

Received 31 May 2012; Accepted 7 Nov 2012

DOI: 10.5772/54934

Abstract The SCORPIO is a small-size mini-teleoperator mobile service robot for booby-trap disposal. It can be manually controlled by an operator through a portable briefcase remote control device using joystick, keyboard and buttons. In this paper, the speech interface is described. As an auxiliary function, the remote interface allows a human operator to concentrate sight and/or hands on other operation activities that are more important. The developed speech interface is based on HMM-based acoustic models trained using the SpeechDatE-SK database, a small-vocabulary language model based on fixed connected words, grammar, and the speech recognition setup adapted for low-resource devices. To improve the robustness of the speech interface in an outdoor environment, which is the working area of the SCORPIO service robot, a speech enhancement based on the spectral subtraction method, as well as a unique combination of an iterative approach and a modified LIMA framework, were researched, developed and tested on simulated and real outdoor recordings.

Keywords Service Robot, Automatic Speech Recognition, Speech Enhancement, Spectral Subtraction, Human-Robot Interaction

1. Introduction

Service robotics represents a special subset of robotic systems. There are several definitions of service robots, proposed by the International Federation of Robotics. According to [1], a service robot is *a mobile device carrying out services either partially or fully automatically.* In [2], it is *a robot that operates partially or fully autonomously to perform services useful to the well being. They are mobile or manipulative or combination of both.*

To help us describe such devices we can look at the primary application domains of service robots, which are services like manipulation, security monitoring, short distance shuttle transports, automatic cleaning, robot assistance or fire fighting, rescue and pyrotechnic assistance.

The level of robots' autonomy determines their controlling mechanisms. Most robots are not fully autonomous and often work in teleoperator mode. They are controlled remotely using a wired control panel (computer) or wirelessly using PDA or control panel hardware. These devices can also display a video stream from cameras on robots, as well as other important values measured by the robotic system. They can also present the current state of the robot's vehicle subcomponents (e.g. lights, cameras, …) .

Communication with robots is one of the key research fields in robotics. To facilitate natural interaction, robots should be able to perceive and understand several modalities used by humans during face-to-face interaction. Besides speech, probably the most prominent modality used by humans, these modalities also include pointing gestures, facial expressions, head poses, gaze, eye-contact, body language, etc. [3]. Multimodal interfaces in robotic systems also use a combination of different inputs, including head nodding, pointing, field of vision cooperating with dialogue management, semantics, context [4], and, with human-robot teamwork, also sharing knowledge using non-verbal communication [5] [6].

Using speech as an important modality of the communication interface is becoming more and more popular. Especially in teleoperation, where an operator needs to control large numbers of devices (using keyboard, buttons and joysticks), speech may significantly help in successfully controlling secondary functionalities of the service robot.

Integration of speech recognition capability into a service robotic system has its own specifics. One of the most important facts is that such systems operate mainly in outdoor, noisy environments [7], which has a negative impact on the robustness of the speech recognition process.

The *robustness* can be generally defined as the capability to deal with adverse conditions, or to adapt to such conditions. In our work here we focus on one aspect of speech recognition robustness – "*environment robustness*", which means the capability to achieve acceptable recognition accuracy in a noisy, outdoor environment.

There are several ways to increase the robustness of the speech recognition process [8]:

- using a more sophisticated voice-activity detection algorithm (cepstral-based, GMM-based),
- using speech enhancement techniques,
- using robust features for speech parameterization,
- training robust acoustic models.

The application of speech enhancement is the most frequently used solution for increasing the robustness of the speech interface. A lot of work has been done in this area with very promising results (e.g., [9]-12]). Speech enhancement techniques play an important role in the final functionality and usability of the system [9].

Not all enhancement techniques can be used for improving speech recognition. Enhancement also produces distortion in the speech. This kind of distortion is often acceptable for a human listener, but can negatively impact the accuracy of speech recognition. The spectral subtraction enhancement method is one of the most appropriate methods for use with speech recognition. It also develops some distortion (also in the form of musical noise), but there are several techniques to suppress this effect (see [11-14]).

Building on previous work described in [15], we continue to research and develop the speech interface for the mobile service robot SCORPIO. This is a small service robot working in teleoperator mode that can serve several purposes, especially in the support of monitoring, manipulation and movement in dangerous areas. Its working area is an outdoor environment, such as a street, or industrial buildings. The robot vehicle is operated by the human operator from a wirelessly connected control briefcase using joystick, keyboard, buttons and speech interface. The environment robustness of the integrated speech interface is an important requirement in such systems.

First, the acoustic conditions of the robot's working area and acceptable recognition accuracy were examined. The acoustic conditions and the accuracy of speech recognition can be numerically expressed by the signal-to-noise ratio (SNR) and word error rate (WER), respectively. Sound recordings captured on the street [25] showed the average SNR at around 15dB with dispersion in the interval ±5dB. Based on this, we can conclude that the proposed speech interface should operate in an environment with at least a 10dB SNR level and a WER of lower than 10%, which is the value that is also acceptable in an acoustically clean environment. The spectral subtraction methods were studied and applied for this purpose. The unique combination of a modified LIMA framework and an iterative spectral subtraction method is presented.

The paper is organized as follows. First, the service robotic system SCORPIO (the robot and the portable briefcase remote control device) is described. Then, the design and development of the speech interface implemented in the remote control device is introduced, including hardware design, recognition engine, communication with the control panel, acoustic and language models, utilization and start-up tests, and final implementation in the remote control device of SCORPIO. The second part of the paper introduces the spectral subtraction enhancement techniques and their implementation into the service robot speech interface. Testing to improve robustness in laboratory conditions, simulated outdoor conditions, and real outdoor environments is described in Section 6. The last part of the paper discusses the results and introduces some proposals for future work.

2. Description of the service robot SCORPIO

The SCORPIO is a small-size mini-teleoperator (or mobile service robot) for booby-trap disposal, especially underneath vehicles (height 130 mm) [16], which was developed by the ZTS VVU Kosice. It is able to serve several purposes, especially the support of monitoring, manipulation and movement in dangerous areas, pyrotechnical and chemical reconnaissance, etc. It is, for example, capable of carrying a water cannon able to destroy improvised explosive devices. The basic system consists of two parts – the mobile robot vehicle (Figure 1) and the control unit (panel) in the form of a portable briefcase remote control device (containing a low-resource embedded PC) (Figure 2), which enables the remote control of the mobile vehicle.

The SCORPIO robot vehicle operates five monochromatic BW cameras (two front, two rear, one top wide-angle), one colour camera for direction finding (all cameras are analogue), two laser pointers for direction targeting, three rangefinders (front, rear, cannon) and seven lights (two front, two rear, two top, one direction finding). The robot vehicle also contains an internal embedded PC, described below, and a digital RF module for connection to the controlling briefcase. For the transmission of the chosen analogue camera composite signal, a second RF module with common analogue modulation is used.

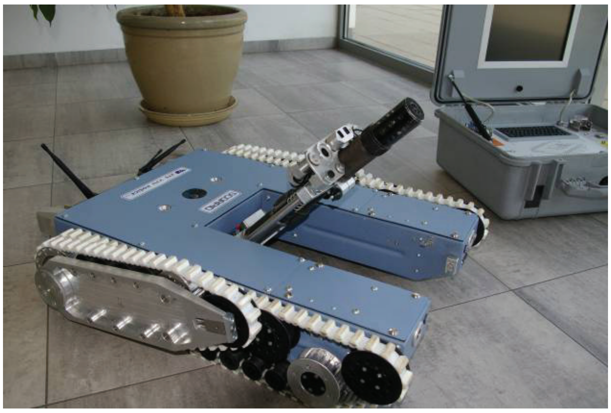

Figure 1. The mobile service robot (vehicle) SCORPIO

The main controlling hardware of the robot SCORPIO consists of two independent ultra-low-power Pentium-III class computers, one in the robot and one in the wirelessly connected briefcase, both running on batteries. The robot has no local controlling mechanisms, and it is not able to work without the portable briefcase remote control device connection.

The portable briefcase remote control device unit consists of the embedded computer with analogue video capturing card (AVC2000), a TFT 12" display and control panel with control buttons, a keyboard and a joystick with operator-presence control, or "dead man's switch" (which was not used during the operation).

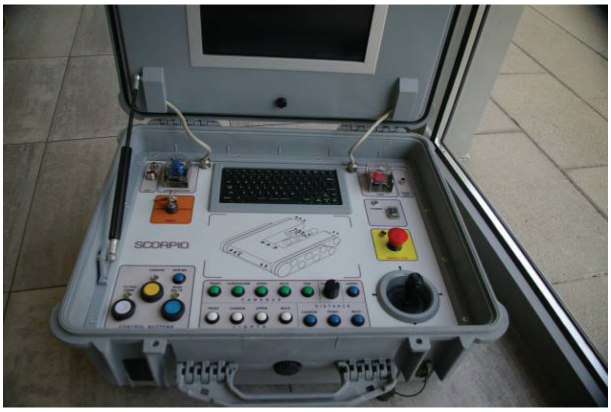

Figure 2. The wirelessly connected portable briefcase remote control device with control panel of the service robot SCORPIO

The main functionalities of the control panel GUI are displaying the video from the chosen robot's camera, displaying values measured by the vehicle (distance to nearest detected objects, temperatures, battery, communication information, etc.), controlling the movement of the robot vehicle using the joystick, selecting the chosen camera for display, and enabling/disabling the lasers and lights of the robot with a set of buttons. When communication with the mobile robot vehicle is not functional, the GUI informs the operator using a sound alarm and by displaying informational messages; the robot then stops all activity immediately.

3. SCORPIO speech interface

The service robot's wirelessly connected portable briefcase control device enables the operator to manually operate the robot just by using the joystick, keyboard and pre-installed buttons (see Figure 2). The limitation of such an interface is that there are a lot of devices (functionalities), which are difficult to control simultaneously. The operator needs his hands to control the movement of the robot and some other functionalities, such as strap position. The second problem is that he or she needs to continuously watch the screen with the output of robot's cameras, and is not able to concentrate his sight on the buttons of the control panel. Therefore, using speech as an input-output modality seems to have advantages. The next important limitation is that adding a new functionality requires a reconstruction of the portable briefcase remote control device (adding new buttons): without modifying the portable briefcase remote control device, controlling the new functions becomes very complicated.

The SCORPIO speech interface (Figure 3) has a simple structure. It consists of two modules – a control panel interface (CPI) and an automatic speech recognition engine (ASR engine), which is described further in this section.

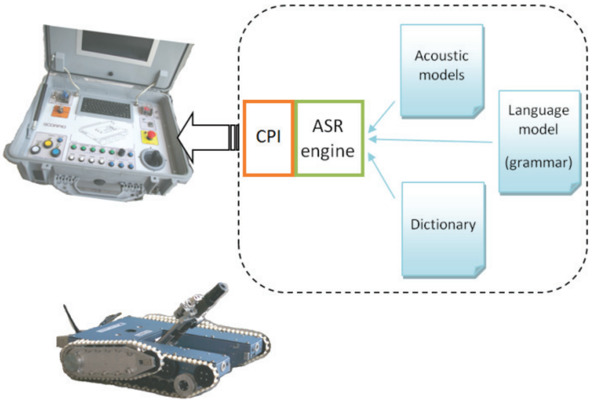

Figure 3. SCORPIO speech interface including ASR module and module for communication with the SCORPIO portable briefcase remote control device software

Developing the reliable, fast and easy-to-use ASR engine for noisy environments and low-resource hardware devices is not an easy task. Hardware limitations do not allow us to use large acoustic and language resources or complicated algorithms. However, there are some frameworks available able to fulfil our requirements and run on specified hardware configuration. We have adopted one of them [17] for building a specialized ASR module.

The development process of the SCORPIO speech interface consists of selecting the parameterization as well as the types of acoustic model, training the acoustic models, preparing language models, adapting and integrating the ASR engine and developing the CPI module. The last step was the implementation of the speech interface in the control briefcase.

3.1 Parameterization and acoustic model

Three-state left-to-right phoneme-based Hidden Markov models (32 probability density functions – PDF mixtures on state) with MFCC parameterization were selected as the most appropriate models, based on previous work done in our laboratory [20]. The parameter vector consists of 12 static MFCC coefficients, zero coefficient (0), delta (D), acceleration or acceleration coefficients (A) and with subtraction of cepstral mean (Z) – (HTK configuration sample: MFCC_D_A_Z_0). The vector consists of 39 values.

3.2 Acoustic model training

The acoustic models were trained on landline telephone speech database SpeechDatE-SK [21] using the reference recognizer training procedure from the COST-249 project [22]. The phoneme-based acoustic model training had two phases. The first phase focused on the best alignment of models using the *HTK-based flat-start method* [23] consisting of initialization of HMM's parameters with global means and variances, the Baum-Welch embedded re-estimation of parameters, re-estimation of added inter-word silence (sp – short pause) model, and alignment of training data using these models.

The second phase focused on *training final models* with estimation of phoneme models based on previous Viterbi forced alignment and re-estimation with the Forward-Backward algorithm. Finally, the number of PDF mixtures were doubled and then re-estimated in an iterative process.

3.3 Language models

In the case of speech interfaces, which enable users to control functionalities of the system through a limited set of commands, the deterministic language model is effective enough. Such a model is usually in the form of context-independent grammar. First, the set of intended commands was defined according the analysis of interaction with the robotic system SCORPIO. The analysis has shown that it could be helpful for the operator to use speech for controlling cameras, lights, rangefinders and track positions, because their hands must control the movements of the robot by joystick and their sight has to follow the screen where the output of the camera is visible.

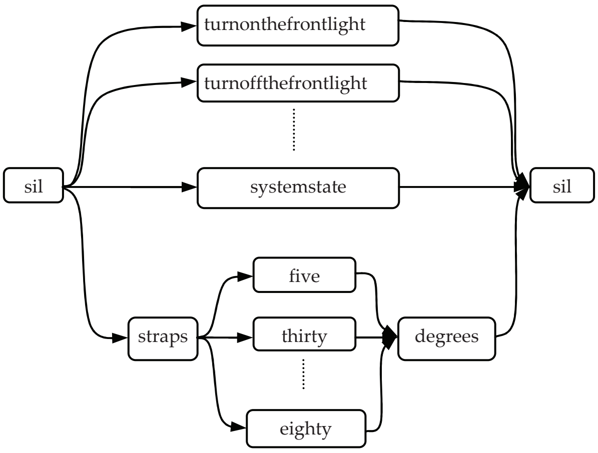

Figure 4. The recognition grammar network

Speech commands were distributed according to the relevant devices. 62 voice commands were defined, such as to turn on and off devices or set the position of tracks. Commands were structured into the parallel network (Figure 4). Commands consisting of more than one word were merged into one big word, e.g., command "zapnúť predné svetlo" ("turn on the front light") was merged into the command "zapnúťprednésvetlo" ("turnonthefrontlight"). This way the recognition network was simplified.

In addition to recognition grammar, the pronunciation dictionary was also prepared in two steps. In the first step pronunciation was added automatically with a simple transcription tool. After that, the pronunciation on word borders was manually corrected.

3.4 CPI communication module

After creating the ASR engine, a wrapper interface, or the Control Panel Interface (CPI) communication module, was designed. The module was designed to communicate with the robot's main control software (embedded in the portable briefcase remote control device) by sending UDP frames with specific structures, which contain information about recognized commands and the state of the speech interface. The module uses a TCP/IP connection to the ASR engine. It is responsible also for filtering the recognized commands by comparing the recognition confidence level with a specified threshold.

This communication structure helped us during the development phase because the GUI of the portable briefcase remote control device software was not functional on the development board. This was due to the fact that the video capturing card presents the panel exits with an error code. Therefore, we used a VPN connection to complete the robot's portable briefcase remote control device connection to the ZTS VVU intranet to test the command execution and transmission.

This architecture will enable the future development of recognition engines running on portable embedded devices outside the portable briefcase remote control device. Current mobile technology provides more computing power than the embedded PC integrated in this robotic system.

3.5 Hardware implementation

The setup was built on an existing embedded computer environment with an embedded Intel x586 compatible Tiny886ULP8-800/128-L-X computer [18] with CPU – 1GHz (TM5800), OS: Linux Debian 6, 500MB of memory, 1GB of storage for OS and applications (SD card) and USB interface.

After adapting the decoding core to a low-resource device [19], the first challenge was to prepare a universal audio interface for a microphone and loudspeaker connection, because the embedded computer used in the SCORPIO vehicle and controlling briefcase has no audio input or output capabilities. After testing seven different external USB audio cards, only two of them were able to be connected to a recognition engine using an Alsa interface [24]. It was finally discovered that the USB bandwidth control experimental kernel option switch had to be turned off and the USB modules recompiled. Ultimately, all connected USB audio devices were working properly.
Sound tests were then carried out for subjective hearing tests of the sound recorded from the microphone using different USB soundcards, USB headphones and finally from eyewear iWear™ VR920 with microphone and gyro sensor included. In the future it is planned to use the gyro sensor of the VR920 for controlling the cameras of the robot.

During the subjective hearing tests only two USB sound cards, one USB headphone and VR920 were found to be reliable in speech recognition applications. The others had poor sound quality with artefacts, or some kind of gate functions causing the first part of any speech to be lost (the first phonemes were destroyed).

Furthermore, a serious problem was encountered when the system booted up with connected USB devices. One USB soundcard and the VR920 caused a POST (power-on self-test) to hang, and the boot-up process was stopped. The only solution found was not to plug these devices before booting the embedded PC, which should need an HW modification for permanent speech interface setup.

After the software development, the computational resources tests were done – the portable briefcase remote control device runs on batteries, and also the overloaded CPU could cause communication failures with the SCORPIO robot or overheating. After using monophone models (16 PDF mixtures on state), only 16% of processing time and 4.5MB of RAM was used by the ASR and CPI communication module.

Finally, during the HW implementation process (the kernel recompilation and recognition engine development) the main storage capacity was upgraded by implementing a fourfold greater capacity (4GB Flash card), thus preparing for the development of a more robust system.

3.6 Scenario of interaction

After booting up the remote control briefcase and controlling robot software, the speech recognition engine (ASR and CPI communication module) starts to operate. If there is any soundcard connected, the CPI waits for the specific command "aktivuj hlasové povely" (activate the voice commands). When this occurs, the CPI module starts to listen to the next commands – it turns itself to the *active state*.

When the speech interface is in active state, an operator can use it simultaneously with joystick, keyboard and buttons. When a voice command is recognized with a sufficient confidence level, the CPI uses simple word TTS (Text to Speech) synthesis to replay the recognized command to the operator (they can also see the recognized command on the display). During the synthesized speech replay the operator could push the operator presence button on the joystick (which was not used before for any other function of the robot) and the command is executed by the robot. After that, the speech interface listens continuously to new commands.

When the operator does not want to use the speech interface, it can be set to *passive state* with the voice command "vypni hlasové povely" (turn off the voice commands).

The interaction with the SCORPIO robot through the speech interface can be seen here:
http://speetis.fei.tuke.sk/video/scor2012.wmv.

4. Initial evaluation of SCORPIO speech interface

4.1. The reference offline tests

The offline tests for evaluation of acoustic models were done to obtain the reference performance of such models. The obtained values show us the best accuracy that can be achieved with the acoustic models used. The tests were performed with application words, isolated digits, proper names and phonetically rich words. Data on 200 speakers from the SpeechDatE-Sk database have been used [21]. Word Error Rates (WERs) were calculated by the common formula defined in [23].

$$\mathrm{WER} = \frac{S + D + I}{N}, \tag{1}$$

S is the number of substitutions, *D* is the number of the deletions, *I* is the number of the insertions and *N* is the number of words in the reference file.

WER (%) / mixtures	Own names	**App. words**	**Isol. digits**	Phon. rich Words
16	21.09	6.00	1.08	20.66
32	17.53	4.63	1.08	18.75
64	12.41	2.06	0.54	15.31
128	9.21	1.63	0.54	14.41
256	8.68	1.46	0.00	15.56

Table 1. The results of offline tests with recordings from SpeechDatE-Sk database

The results (Table 1) show that our phonemes acoustic models can be powerful enough for the application words and isolated digits in good acoustic conditions.

4.2 Robot's workspace and definition of robustness

The SCORPIO service robot is mainly designed to work in outdoor, noisy environments. The acoustic conditions can be expressed by the signal-to-noise ratio (SNR).The average SNR in recordings made on the street [25] is around 15dB, with dispersion in the interval from 10 to 20 dB. The robustness of the robot's speech interface, which we have set as our goal, means that the system will be able to reach a word error rate (WER) of lower than 10% when the SNR level is in the mentioned interval.

4.3 Simulated outdoor environment tests (without enhancement)

Since no relevant results were available for real outdoor environment conditions, it was decided to prepare for the experiment with a simulated outdoor (street) environment. For the simulation of the real conditions we took a recording of a noisy street (with noises of cars, buses and trams) from the JDAE-TUKE database (Joint database of acoustic events and backgrounds), which was created by our laboratory [25]. In the next step, we prepared a group of testing recordings with eight participants. Recordings were recorded in a relatively quiet room with a standard headset microphone. Each recording contains all commands for controlling the SCORPIO robot. The overall length of all recordings is about 16 minutes.

The software tool FaNT (Filtering and Noise Adding Tool, described in [26]) has been used for creating a mix of the clear test recordings and the street noise. Six types of recordings have been prepared with specific signal-to-noise ratio (SNR) values of 35dB, 30dB, 20dB, 10dB, 5dB and 0dB. Recordings were manually annotated for the WER computation.

All recordings were tested using the speech recognition system of the SCORPIO speech interface, and the WER values were logged. In the case of the most disturbed recordings (10dB, 5dB and 0dB), the threshold of the voice activity detector must be experimentally set. The base threshold for all recordings was 2000. The threshold level was increased with increasing noise. The impact of these changes can be seen in Table 2. Tests were done with acoustic models with 16, 32, 64, 128 and 256 PDF mixtures.

WER (%)	Number of mixtures				
SNR (dB)	**16**	**32**	**64**	**128**	**256**
30	4.40	4.11	4.02	3.81	**3.52**
20	5.87	5.87	5.87	5.87	**4.99**
10(thresh.=4000)	34.90	32.26	31.73	27.27	**26.39**
0 (thresh.=8000)	65.59	61.88	60.97	57.48	56.30

Table 2. The results of offline tests with noisy recordings

The results in Table 2 show that the WER significantly increase in the 20 to 10dB interval. When the SNR is 10dB, the performance of the speech interface is not sufficient (26.39%) to fill the robustness criterion (WER lower than 10%). So, some enhancement is required. The noise reduction methods based on spectral subtraction were studied, developed and tested.

5. Spectral subtraction methods for increasing speech interface robustness

There are several ways to increase the robustness of speech recognition in noisy environments. The first is to use more a sophisticated voice activity detection algorithm. The described speech interface uses energy-based VAD, which is not sufficient. Cepstral-based or GMM-based detectors could be more reliable in noisy environments. The second way to increase robustness is to use special noise-adapted acoustic models, as well as robust features for the parameterization of noisy speech.

Using some noise-reduction algorithms is one possible and the most popular approach to increasing the robustness of speech interface. A lot of work has been done in this area [8]. There are several noise-reduction algorithms (spectral subtraction, Wiener filtering, MMSE). All of them reduce noise, but not all give better accuracy in the speech recognition process. Our earlier experiments with Wiener filtering and MMSE caused lowering of recognition accuracy. When the SNR was 10dB, and no enhancement was used, the WER was about 16%. The application of Wiener filter increases the WER to 35.29%, and the application of MMSE results in WER 36.4%.

First, we implemented a spectral subtraction speech enhancement algorithm in the SCORPIO speech interface to increase its robustness. Then, several experiments were done with the spectral subtraction (SS) algorithm and its modifications.

5.1 Theory of spectral subtraction

The basic assumption is that the noisy speech signal *y(n)* consists of a speech signal *s(n)* and an additive noise signal *d(n)* [8] as follows:

$$\mathrm{y(n)} = \mathrm{s(n)} + \mathrm{d(n)} \tag{2}$$

In the frequency domain, equation (2) is expressed as

$$\mathrm{Y}(\omega) = \mathrm{S}(\omega) + \mathrm{D}(\omega)\,, \tag{3}$$

where *Y(ω)*, *S(ω)* and *D(ω)* are spectra of signals *y(n)*, *s(n)* and *d(n)*.

Y(ω) can be expressed in exponential form as

$$\mathrm{Y}(\omega) = \left|\mathrm{Y}(\omega)\right| e^{j\varphi_{y}(\omega)}\,. \tag{4}$$

If the noise spectrum $\hat{\mathrm{D}}(\omega)$ can be estimated, then an approximation of speech spectrum $\hat{\mathrm{S}}(\omega)$ can be computed from final signal spectrum *Y(ω):*

$$\left|\hat{\mathrm{S}}(\omega)\right|^{\mathrm{p}} = \left|\hat{\mathrm{Y}}(\omega)\right|^{\mathrm{p}} - \left|\hat{\mathrm{D}}(\omega)\right|^{\mathrm{p}}, \tag{5}$$

where *p* is the power exponent.

The equation (5) represents the general algorithm of spectral subtraction. If *p is 1* then it is the basic version of spectral subtraction of magnitude spectra. If *p* is 2 then it is the algorithm of power spectral subtraction [9]. Sometimes the power exponent is marked as γ (see [11]).

After applying spectral subtraction the enhanced spectrum could contain some negative values, which is not allowed. Such situations can occur when the estimated noise spectrum is greater than the enhanced signal spectrum. Several solutions have been proposed. The simplest one was proposed by Boll [13]. He suggested simply substituting negative values of the spectrum with zero values, which is expressed by the following formula:

$$\left|\hat{\mathrm{S}}(\omega)\right|^{2} = \begin{cases} \left|\mathrm{Y}(\omega)\right|^{2} - \left|\hat{\mathrm{D}}(\omega)\right|^{2} & \text{if} \left|\mathrm{Y}(\omega)\right|^{2} > \left|\hat{\mathrm{D}}(\omega)\right|^{2} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

A different approach was proposed by Berouti et al. [10], based on using the oversubtraction factor *α* and the flooring factor *β*. Their method consists of subtracting an overestimate of the noise power spectrum while preventing the resultant spectral components from going below a preset minimum value (spectral floor) [14]. The realization follows the next equation:

$$\left|\hat{\mathrm{S}}(\omega)\right|^{2} = \begin{cases} \left|\mathrm{Y}(\omega)\right|^{2} - \alpha\left|\hat{\mathrm{D}}(\omega)\right|^{2} & \text{if} \left|\mathrm{Y}(\omega)\right|^{2} > (\alpha+\beta)\left|\hat{\mathrm{D}}(\omega)\right|^{2} \\ \beta\left|\hat{\mathrm{D}}(\omega)\right|^{2} & \text{else,} \end{cases} \tag{7}$$

where α is the oversubtraction factor (usually $\alpha \geq 1$), and $\beta\,(0 < \beta << 1)$ is the spectral floor parameter.

As mentioned in [8], these parameters enable a great amount of flexibility in the spectral subtraction algorithm and can be adjusted to obtain the best enhancement of speech. The parameter *α* determines the amount of subtracted noise and affects the amount of speech spectral distortion caused by the subtraction. The parameter *β* controls the amount of remaining residual noise and the amount of perceived musical noise. Optimization of speech recognition accuracy can be reached by varying these parameters.

5.2 Modified LIMA framework for spectral subtraction

Kleinschmidt [11] presents a modified LIMA (likelihood-maximizing) enhancement technique for spectral subtraction, where the values of power exponent *γ* and *β* floor parameter are optimized to best fit the instantaneous relationship between clean speech and noise signals. The proposed modification removes the need to access the state models and the state sequence information, as is necessary in classic LIMA framework methodology. Only access to full utterance likelihoods (accuracy) and word sequences is required. This approach is highly suitable for use with stand-alone or third-party speech recognition engines [11].

As a criterion for maximization, the word recognition accuracy (ACC) was taken. The results of the experiments proposed in [11] show the possibility to blindly optimize spectral subtraction parameters using only utterance level scores (ACC, WER).

The second important fact proposed in [11] is that there is the potential to achieve better performance when the values of γ and β are not constrained to their traditional values. For example, some improvement was achieved when γ was 1.5, which responds neither to magnitude spectral subtraction nor to power spectral subtraction. The same situation was presented for the floor parameter β. The theory of SS defines β as a value very close to zero, e.g., 0.002, but in [11] Kleinschmidt describes some improvements also when β was 0.5. This value seems to be more appropriate for use in speech recognition.

5.3 Iterative spectral subtraction

The main disadvantages of the spectral subtraction speech enhancement algorithms are that they develop musical noise in an enhanced signal and also cause distortion in speech. Speech distortion becomes severe when the degree of noise reduction is larger. It can be reduced by several modifications of the basic subtraction algorithm. Improvement can be achieved using the approach proposed in [10] and by adjusting oversubtraction factor α, floor factor β or power exponent γ (LIMA or modified LIMA approaches).

The next promising approach is using an iterative spectral subtraction technique, as proposed in several articles (e.g., [12] [14] [27] [28]). As Li wrote in [14], the principle of iterative spectral subtraction consists in the fact that the enhanced speech becomes the input signal, so music noise is seen as input noise to be reduced again. The results published in the mentioned papers show improvement potential, especially in second iteration. According to [29], a lower amount of musical noise is observed after using iterative "weak" spectral subtraction, when rather less noise is subtracted in particular iterations.

6. Experiments with spectral subtraction

6.1 Setup of the experiment with simulated outdoor conditions

A new, larger set of test recordings with 60 participants was prepared. Recordings were made with video eyewear iWear™ VR920, intended for use by the operator of the service robot. This device contains an integrated USB audio system with a built-in microphone (in the eyewear frame). Recordings were made in a room with office background quality (SNR was around 25dB). Each recording contains all commands for controlling the SCORPIO robot. The overall length of the recordings is about 80 minutes. All recordings were annotated in a two-stage process. In the first stage, recordings were recognized with the automatic speech recognition engine using the same acoustic and language models used by the SCORPIO speech interface. After that, the obtained annotations were manually corrected.

As in our earlier experiment for the simulation of the outdoor conditions, a recording of a noisy street from the JDAE-TUKE database was mixed with clear test recordings (mixing was done with tool FaNT [26]). Two groups of recordings were prepared with specific signal-to-noise ratio (SNR) values of 10dB and 0dB. Prepared recordings were firstly tested without using speech enhancement to obtain the reference values (Table 3).

All recordings were tested using the SCORPIO speech interface, and the overall WERs (Word Error Rate) were calculated. Tests were done with the phoneme-based acoustic models described above, trained on the SpeechDatE-Sk database [21] with 256 PDF mixtures. The context-free speech grammar presented in section 3.3 was used as a language model.

6.2 Reference test

First, the reference values of WER for clear (SNR ≈ 25dB) and noisy recordings (10 and 0dB) without enhancement were obtained in an offline test (see Table 3).

SNR (dB)	**25**	**10**	**0**
WER (%)	2.76	16.78	61.81

Table 3. Reference results (without enhancement)

Recordings with SNR=10dB result in WER about 14% higher in comparison with clear recordings. When SNR is 0dB, the speech recognition system is rather unusable (WER is 61.81%). This reference test confirmed early results obtained with the smaller group of recordings (about 16 minutes, eight participants) presented in Table 2. When SNR is around 10dB and lower, the robustness criterion is not fulfilled (WER = 16.78%).

6.3 Experiments with spectral subtraction based on modified LIMA framework

The modified LIMA framework presented in [11] makes it possible to optimize parameters γ and β of spectral subtraction according to the overall word or sentence error rate. We assume that, together with power exponent and floor parameter, also oversubtraction factor α can be adjusted to bring more robustness in speech recognition in the noisy environment.

First, it was decided to use a built-in spectral subtraction algorithm in our recognition engine, which follows equation 7. The setup of the engine allows us to adjust only α and β parameters. The power exponent γ was set to *2*.

Based on these assumptions, more than 20 tests were done, where α was in interval *<0.5, 2>* and β was in interval *<0.1, 1>*. The results – whereby WER decreased significantly – can be seen in Table 4.

		WER [%]	
α	β	SNR = 10dB	SNR=0dB
without enh.		16.78	61.81
2	0.5	13.05	60.5
1.5	0.5	12.72	51.17
1	0.5	12.27	50.47
0.5	0.5	12.16	50.62
0.5	0.3	**12.12**	**48.79**

Table 4. Results of experiments with spectral subtraction enhancement with varying α and β parameters

As we can see in Table 4, the best recognition performance was reached when oversubtraction factor α was about *0.5* and flooring factor β was also about *0.3* (so-called "weak" spectral subtraction [29]).

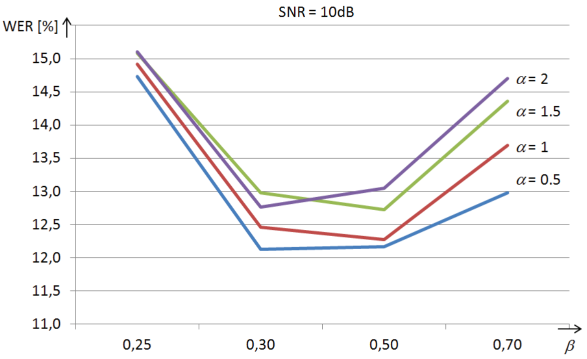

Figure 6. Dependency of WER from α and β for SNR = 10dB

The graph in Figure 6 shows results also for other values of the β parameter. We can see that the best values of WER were obtained when β was in the interval from 0.3 to 0.5. Outside of this interval, WER increases rapidly.

As we supposed, speech recognition accuracy is higher when spectral distortion of speech is rather low (α has lower value), there is a rather larger amount of remaining residual noise and a lower amount of musical noise, which is affected by factor β. The best result in our case was achieved when β was about *0.3*. The relatively high value of the flooring factor means that the larger amount of residual noise was left in the enhanced signal.

6.4 Experiments based on iterative spectral subtraction and modified LIMA framework

Although using a spectral subtraction algorithm brings some improvement, it is not sufficient for the robust speech interface (see Table 4) we intended to develop. Further improvement is also required when the speech has the same power as the noise (SNR is 0 dB) – the lowest WER is still too big (48.79%).

As presented in [12], [14], [27] and [30], iterative use of spectral subtraction can enhance the speech and decrease the amount of musical noise in the enhanced speech.

One promising solution was to join the iterative approach with the modified LIMA framework.

A new, stand-alone Speech Enhancement Toolkit (SET) was prepared for performing iterative spectral subtraction. We modified the algorithm proposed by Berouti et al. in [10], where the appropriate oversubtraction factor was computed automatically according to formulas based on SNR level. To use a modified LIMA framework, α and β parameters must be set manually.

A series of experiments were performed with two and three iterations and with different settings of α and β parameters in particular iterations. The number of possible combinations was reduced by using only the two best settings of first iteration (where $\alpha = 2$, $\beta = 0.5$ and $\alpha = 0.5$, $\beta = 0.3$). Table 5 contains the results of experiments with iterative SS, where some improvement was achieved.

1. iteration		2. iteration		WER [%]	
α	β	α	β	SNR = 10dB	SNR = 0dB
without enhancement				16.78	61.81
2	0.5	2	1	12.5	48.23
2	0.5	1	0.5	**9.81**	48.56
2	0.5	0.5	0.5	13.2	49.01
0.5	0.5	1	0.5	10.11	42.52
0.5	0.5	0.5	0.5	11.6	**42.37**
0.5	0.3	1	0.5	9.85	50.43
0.5	0.3	0.5	0.3	10	48.97

Table 5. The results of experiments with spectral subtraction enhancement with varying power and floor

Significant improvement was achieved in the second iteration for both levels of SNR. The defined criterion for robustness was fulfilled for the SNR level 10dB.

The obtained results signify that, in the case of worse SNR, more care was taken and the weak spectral subtraction was more successful. Conversely, better SNR enables the subtraction of more noise, producing a smaller amount of musical noise and distortion that is still acceptable for speech recognition purposes.

6.5 Verifying SS methods in real outdoor environment

The described experiments, which were done with recordings using artificially added street noise, helped us to tune up parameters of the spectral subtraction algorithm for improving the robustness of the SCORPIO speech interface.

For the verification of the proposed solution, we did an evaluation in a real outdoor environment. The evaluation took place in the car-park out the front of the laboratory, near the road and tram line. Commands for service robot were read and recorded by 44 test subjects (students) using the video eyewear. The overall length of recordings was 42 minutes.

Recordings were annotated as in previous experiments. During the manual correction of annotations, we detected some differences from the simulated outdoor environment recordings:

- The power of the background noise has higher fluctuations, as in the case of noise recordings used in previous experiments. The noise was less stationary.
- Test subjects increase volume when the noise increases. This is the main difference to recordings with artificially mixed noise.
- A higher level of environment noise was assumed. The SNR was around 15dB–20dB, because the location chosen for the experiment was less noisy than we expected.

At first, the base reference tests were done without using enhancement algorithms. Because of a relatively high SNR level, WER was only 6.43%. The obtained results were so good that it was difficult to obtain significant improvement. However, some improvement was achieved by using basic a spectral subtraction algorithm, although the iterative approach was not able to further improve robustness. The results of the evaluation can be seen in Table 6.

As we concluded earlier, the obtained results confirmed that in the case of better SNR, it is possible to subtract a large amount of noise. This means that for higher values of α better results are achieved (last row in Table 6).

α	β	WER [%]
without enh.		6.43
0.5	0.3	7.14
0.5	0.5	6.23
1	0.5	6.08
2	0.5	6.03

Table 6. The results of verifying experiments with spectral subtraction enhancement with varying α and β parameters

7. Conclusions

In this, paper the SCORPIO service robot and the research, development and testing of a robust speech interface was introduced.

The speech interface was integrated into the robot's portable briefcase remote control device with limited hardware power, making it possible to use voice to control the secondary functionalities of the robot. The main difficulty of using the speech interface is the working area of such service robots – noisy, outdoor environments with SNR in an interval from 20 to 10dB. To fulfil the environmental robustness of the speech interface, acceptable for the SCORPIO manufacturer, the WER has to be lower than 10%.

The unique combination of a modified LIMA framework and an iterative spectral subtraction algorithm was proposed, which decreases WER from 16.78% to 9.81% for SNR level 10dB. Significant improvement was also achieved for SNR level 0dB, when WER decreases from 61.81% to 42.37%, but the obtained level of WER is not sufficient for the speech interface to be usable. For such bad (and even worse) conditions, another approach has to be proposed. During real-time factor utilization tests, there was only a 0.15% increase in utilization observed during the preprocessing phase caused by the iterative SS algorithm.

We can conclude that the spectral subtraction algorithm, especially the combination of the modified LIMA framework and the iterative approach, is well suited for increasing robustness of speech recognition in noisy environments. Whilst a larger amount of noise can be subtracted all at once in the case of higher SNR (more than 15dB), when the SNR is lower, significant improvement can be achieved by iterative subtraction of smaller amounts of noise ("weak" subtraction). The proposed approach gives the best results in the case that the background noise is predominantly stationary, but it is not unusable when there is a limited amount of non-stationary noise.

Our future work will be focused on algorithms for automatic setting of spectral subtraction parameters according to confidence of the speech recognition process, which will be able to adapt to changing noise conditions. Other speech enhancement techniques and their modification will be also taken into consideration in future research. A multimodal interface using a gyro sensor and positioning algorithms are also planned to improving the robustness of the communication interface as described in [4] and [5].

8. Acknowledgments

The research presented in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under research projects No. 3928/2010-11 (30%) and VEGA 1/0386/12 (30%), and by the Research & Development Operational Program funded by the ERDF under the ITMS project 26220220141 (40%).

9. References

[1] Schraft R.-D., Schmierer G. (2000) Service robots: Products, Scenarios, Visions. Natick, MA: A K Peters/CRC Press.

[2] Prenzel O. (2009) Process Model for the Development of Semi-Autonomous Service Robots. Dissertation Thesis, University of Bremen.

[3] Stiefelhagen R., et al. (2004) Natural human-robot interaction using speech, head pose and gestures. IROS 2004, Sendai, Sep. 28-Oct. 2, Vol. 3, pp. 2422–2427.

[4] Johnson D.O. and Agah A. (2009) Human robot interaction through semantic integration of multiple modalities, dialog management, and contexts. Int. J. Soc. Rob., Vol. 1, no. 4, pp. 283–305.

[5] Breazeal C., et al. (2005) Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. IROS, Alberta, pp. 708–713.

[6] Perez-Vidal C., et al. (2012) Steps in the development of a robotic scrub nurse. Robotics and Autonomous Systems, Vol. 60, no. 6, June 2012, pp. 901–911.

[7] Ondáš S. and Juhár J. (2012) Improving robustness of the SCORPIO robot speech interface by iterative spectral subtraction. Journal of Electrical and Electronics Engineering, Vol. 5, no. 1, pp. 151–154.

[8] Loizou P.C. (2007) Speech Enhancement. Theory and Practice. Canada: CRC Press, Taylor & Francis Group.

[9] Lee S.-C., Chen B.-W., Wang J.-F. (2010) Noisy Environment-Aware Speech Enhancement for Speech Recognition in Human-Robot Interaction Application. IEEE SMC 2010, Istanbul, 10-13 Oct., pp. 3938–3941.

[10] Berouti M., Schwartz R., and Makhoul J. (1979) Enhancement of speech corrupted by acoustic noise. ICASSP '79, Washington, April 2-4, pp. 208–211.

[11] Kleinschmidt F.T., Sridharan S., Michael W. (2007) A Modified LIMA Framework for Spectral Subtraction Applied to In-Car Speech Recognition. ICSPCS 2007, Gold Coast, Dec. 17-19, pp. 335–338.

[12] Yamashita K., Ogata S., Shimamura T. (2005) Improved spectral subtraction utilizing iterative processing. Trans. of IEICE, J 88-A (11), pp. 1246–1257.

[13] Boll S. (1979) Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans. Acoust. Speech Signal Process, ASSP-27 (2), pp 113–120.

[14] Li X., Li G., Li X. (2008) Improved Voice Activity Detection Based on Iterative Spectral Subtraction and Double Thresholds for CVR. PEITS workshop, Guangzhou, pp. 153–156.

[15] Ondáš S., et al. (2011) Speech interface for controlling service robot SCORPIO. Journal of Electrical and Electronics Engineering, Vol. 4, no. 1, pp. 143–146.

[16] http://www.ztsvvu.eu/, Accessed 20 May 2012.

[17] Lee A., Kawahara T. and Shikano K. (2001) Julius – an open source real-time large vocabulary recognition engine. EUROSPEECH, Aalborg, pp. 1691–1694.

[18] http://www.ampltd.com/dload/t886.pdf, Accessed 2011.

[19] Pleva M., Lojka M., Juhar J. (2012) Modified Viterbi Decoder for Long-Term Audio Events Monitoring. Journal of Electrical and Electronics Engineering, Vol. 5, no. 1, ISSN: 1844–6035, pp. 195–198.

[20] Lihan S., Juhár J., Čižmár A. (2005) Crosslingual and Bilingual Speech Recognition with Slovak and Czech SpeechDat-E Databases. INTERSPEECH, Lisbon, Sept. 4-8, pp. 225–228.

[21] Pollak P., et al. (2000) SpeechDat(E) Eastern European Telephone Speech Databases. Proc. of LREC Satellite workshop XLDB, Athens, Greece, pp. 20–25.

[22] Johansen F.T., et al. (2000) The COST 249 SpeechDat Multilingual Reference Recogniser. LREC, Athens, Vol. 3, pp. 1351–1355.

[23] Young S., et al. (2006) The HTK Book. CUED, 271p.

[24] Lin J.-M., Cheng W.-G., Fang G.-M. (2008) Software integration for applications with audio stream. IIH-MSP Proceedings, Art. no. 4604242, pp. 1126–1129.

[25] Pleva M., Vozarikova E., Dobos L., Cizmar A. (2011) The Joint Database of Audio Events and Backgrounds for Monitoring of Urban Areas. Journal of Electrical and Electronics Engineering, Vol. 4, no. 1, pp.185–188.

[26] Hirsch H.G. (2011) FaNT – Filtering and Noise Adding Tool. Hochschule Niederrhein. http://dnt.kr.hs-niederrhein.de, Accessed Jan. 2011.

[27] Khan M.R., Hasan T. (2008) Iterative noise power subtraction technique for improved speech quality. ICECE, Dhaka, Dec. 20-22, pp. 391–394.

[28] Li S., et al. (2010) Iterative spectral subtraction method for millimeter-wave conducted speech enhancement. JBiSE, Vol. 3, pp. 187–192.

[29] Inoue T., et al. (2010) Theoretical analysis of iterative weak spectral subtraction via higher-order statistics. MLSP workshop, Kittila, Aug. 29–Sept. 1, pp. 220–225.

[30] Nishikawa K., et al. (2009) A study of the residual musical tone reduction on iteration-spectral subtraction. Autumn Meeting of ASJ, pp. 149–150.