Audio Perception in Robotic Assistance for Human Space Exploration: A Feasibility Study

Marco Sewtz, Werner Friedl, Adrian Bauer, Anne Köpken, Florian Lay, Nicolai Bechtel, Peter Schmaus, Rudolph Triebel, and Neal Y. Lii

German Aerospace Center (DLR)
Institut of Robotics and Mechatronics
Muenchener Str. 20, 82234 Weßling, Germany
{*Firstname.Lastname*}@dlr.de

***Abstract*—Future crewed missions beyond low earth orbit will greatly rely on the support of robotic assistance platforms to perform inspection and manipulation of critical assets. This includes crew habitats, landing sites or assets for life support and operation.**

Maintenance and manipulation of a crewed site in extra-terrestrial environments is a complex task and the system will have to face different challenges during operation. While most may be solved autonomously, in certain occasions human intervention will be required. The telerobotic demonstration mission, Surface Avatar, led by the German Aerospace Center (DLR), with partner European Space Agency (ESA), investigates different approaches offering astronauts on board the International Space Station (ISS) control of ground robots in representative scenarios, e.g. a Martian landing and exploration site.

In this work we present a feasibility study on how to integrate auditory information into the mentioned application. We will discuss methods for obtaining audio information and localizing audio sources in the environment, as well as fusing auditory and visual information to perform state estimation based on the gathered data. We demonstrate our work in different experiments to show the effectiveness of utilizing audio information, the results of spectral analysis of our mission assets, and how this information could help future astronauts to argue about the current mission situation.

TABLE OF CONTENTS

1. INTRODUCTION

Accomplishing the goals of bringing humankind to the Moon and Mars is some of the greatest challenges ahead for the space community. To help meet these challenges, robotic assistance will be key, particularly for the construction and support of habitat infrastructure, as well as for carrying out scientific tasks. However, due to the long distances, communication round trip will cause delays of 20min to several hours between Earth and Mars.

Surface Avatar, a telerobotic technology validation mission led by German Aerospace Center (DLR) with partner European Space Agency (ESA), gives astronauts on board the International Space Station (ISS) control over robotic assets [1]. It investigates a combined approach offering scalable autonomy through multi-modal teleoperation to perform tasks in different scenarios. These can range from simple surveillance to complex maintenance tasks which often include a search for failure in which the astronaut has to detect an anomaly in the environment. The astronaut has to investigate multiple objects to observe their state, often accompanied by detailed inspection and manipulation of inner components.

Figure 1: Integrated audio perception into the telerobotic system of Surface Avatar. The robot in the experimental area detected a sound event with an unknown spectral profile and requests manual action from an astronaut on board the International Space Station (ISS).

Audio perception provides an additional modality that may decrease crew time to find the anomaly in extra-terrestrial environments with an atmosphere like Mars. The direction of arrival of a sound event received by the system can be estimated and displayed to the astronaut. Furthermore, the robot's knowledge of the world can be used to infer the current state of a known object remotely and detect failures. All of these can be displayed to the crew as illustrated in the simulated view in Figure 1.

In this feasibility study, we aim to show our preliminary results on using audio perception, in the context of a telerobotic mission, to help understand the world around the robot and propose an approach to:

- detect sound events
- localize sound sources
- fuse sound input with vision sensors and prior knowledge
- obtain spectral knowledge and infer objects' state based on the received data

978-1-6654-9032-0/23/$31.00 ©2023 IEEE

 DOI: 10.1109/AERO55745.2023.10116018

2. Related Work

Early research in the field of sound source localization has focused on the imitation of binaural audio perception of humans and animals [2][3][4][5]. They are based on the interaural phase difference (IPD) and interaural intensity difference (IID) of received signals. The inclusion of the head-related transfer function [6] and the modeling of the reverberation of the environment [7][8] increases the robustness further. However, these approaches require an accurate calibration process, where deviations and unexpected components in the environmental modeling greatly influence the outcome.

Successive work has been carried out on the estimation of Direction of Arrival (DoA) of a signal [9][10]. Incorporating a delay and sum beamformer (DSBF) these approaches estimate the direction using the time delay between the input signal of individual sensors. But low signal to noise ratio (SNR) environments or varying spectral profiles of the sound sources prevent usable results. Approaches based on deep learning [11][12][13][14][15] promise to overcome the mentioned problems, but require dedicated data sets for specific sources for training or immense data for generalization.

More recently, research attention has shifted toward subspace-based approaches like multiple signal classification (MUSIC) [16] or Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) [17]. To overcome the limitations and constraints of the chosen sampling frequency, they offer increased robustness and angular resolution [18][19][20]. The initial high computational demand could be decreased with recent advances in offering real-time estimations for outdoor [21] and indoor [22] environments.

The field of acoustic monitoring is well established in the area of ecological research, especially for ornithology [23][24]. Semi-automated analysis [25][26][27] are utilized for temporal and spatial estimation of bird behavior, which has been developed to detect and monitor audio events. However, expert knowledge is necessary to label received audio fragments. Full-automation methods [28][29][30][31] offering an unsupervised approach, which requires intense training. These methods have been applied toward factory and technical applications for process monitoring for additive manufacturing [32][33]. Furthermore, convolutional neural networks have been added for detecting the degradation state of robotic system [34]. However, the unknown spectral profiles or signals with high variances are still problematic.

In this work, we aim to show that acoustic perception can be effectively used as an additional modality in telepresence applications by implementing it in a ISS-to-Earth demonstration missions, Surface Avatar [1]. It depends on knowledge gained in previous space-to-ground missions, Analog-1 [35][36][37] and Meteron Supvis Justin [38]. We focus on a system that extends the immersion of the robot operator to obtain more knowledge about the environment and which keeps the astronaut in the loop.

3. System Overview

This work is intended to be integrated into DLR's Rollin' Justin [39]. It is a dexterous humanoid robot with a mobile wheel-base, which has served in a wide array of research toward space exploration and terrestrial applications [38] [40]. Equipped with an Intel Realsense D435i RGBD camera, it is able to visually perceive its environment. The sensor is mounted on the head to mimic human-like anatomy and follows the head movement to stay aligned with the visual processing pipeline.

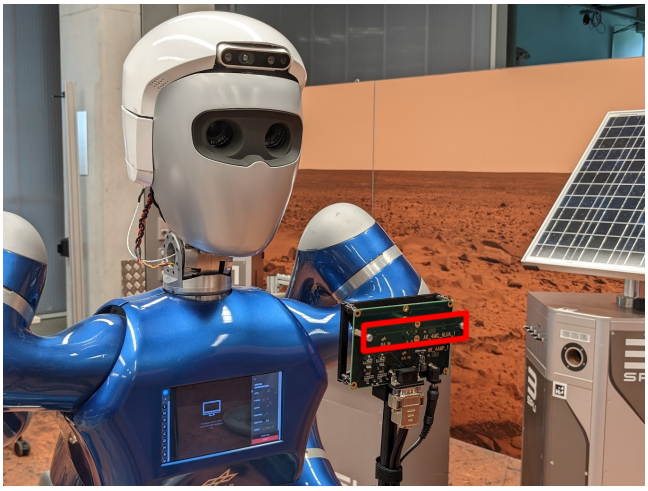

Figure 2: DLR's dexterous humanoid robot, Rollin' Justin, and the microphone array (red) used in this work.

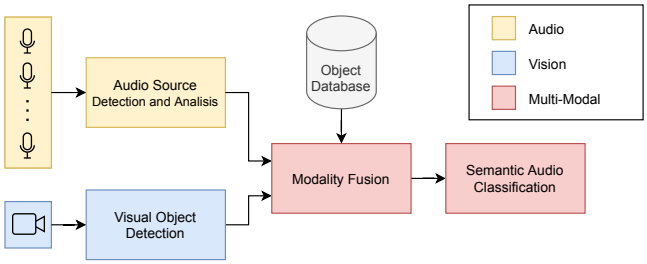

Figure 3: Overview of the system architecture. The approach is divided into auditoral, visual and prior knowledge.

We utilize a four-sensor microphone array as depicted in Figure 2 to receive audio information on the environment. The sensors are arranged linearly with located $d = [0.00, 0.015, 0.06, 0.09]$ cm along the x-axis and enables broadband estimation of signals in the audible range. We investigated a future integration of the array into a novel head design [41] consisting of eight microphones heterogeneously placed on the forehead of the robot. The estimated directivity patterns (-3dB at $\pm 40°$) are integrated into this feasibility study to assure the applicability.

Furthermore, the robot operates in an environment at the DLR simulating a Martian exploration and science site [42]. The environment includes a mechanical mock-up of a lander, several Smart Payload Units (SPUs) for scientific experiments and monitoring and a visual representation of Martian setting. All objects a marked with Apriltags [43] for easy identification and localization.

All data are recorded and pre-processed before fusing them together. Afterwards, using prior knowledge on the environment, the semantic information on the current perception of the world will be jointly inferred. An overview is given in Figure 3.

Audition

Audio is captured using the microphone array. For processing it is essential to have synchronous data acquisition.

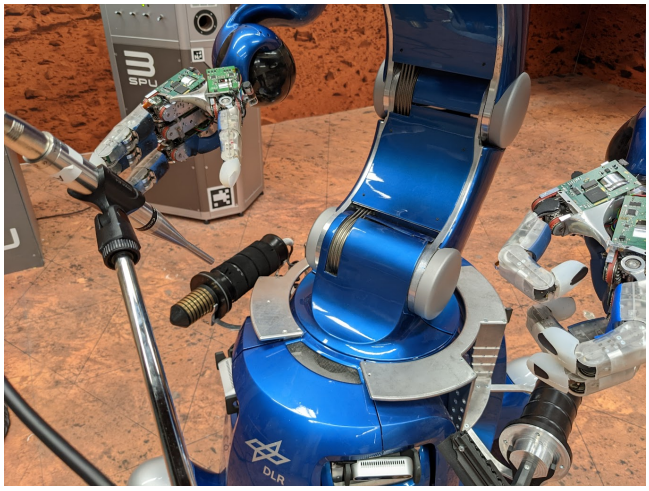

Figure 4: Estimation of the background noise profile for the robot environment. An audio probe is used to capture a highly accurate frequency spectrum that can be used for spectral subtraction in noise filtering.

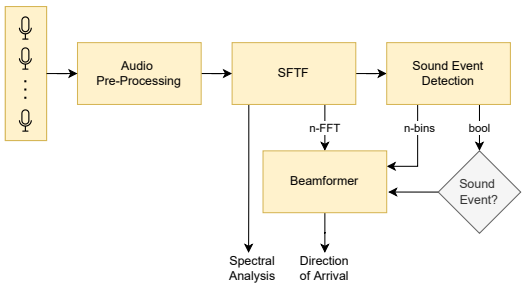

Figure 5: Audio processing branch. After pre-processing the received signals are transformed into the frequency domain and the presence of a sound source is estimated. Afterwards, a possible source is localized.

Therefore, the analog-to-digital conversion is triggered on hardware side. The sampling rate is set to 44100Hz to capture the full spectrum of most signals available in our environment.

Background noise such as wind or system noise created by mechanical components, e.g. cooling fans, induce an omnipresent spectral component that is always accumulated to the received signal. A prior statistical profile is estimated using a sound probe as shown in Figure 4 to obtain an accurate recording of the actual noise. Then, a Fourier analysis is performed to obtain the gains of the spectral components. These can be applied later for noise reduction by spectral subtraction. To prohibit unnecessary detection and estimation efforts that may lead to false positive results in subsequent modules, the presence of a suitable input signal is detected. An evaluation of the power equivalent of the sound signal is performed, comparing the active input to the previously acquired noise spectrum. The received response is used to classify the audio as *noise* or *sound event*. Afterwards, the DoA of the signal is estimated to obtain the spatial information of the sound source used later in the fusion process. The chain of modules is shown in Figure 5.

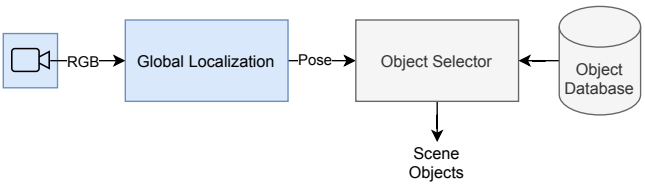

Figure 6: Components of the vision processing. The camera data is used for a global position strategy based on a Simultaneous Localization and Mapping (SLAM) approach and refinement using AprilTags. This information is used to receive current scene objects from a knowledge base.

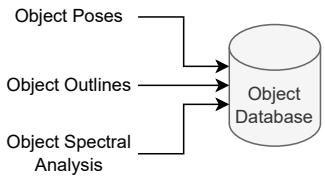

Figure 7: The object database is storage of the robots perception and knowledge about the world. It included the poses of known objects, their geometric outlines and previously obtained spectral profiles of observed states.

Visual Perception

The vision system is primarily used to obtain the localization information of the system. The coarse ego-pose estimation is retrieved by a SLAM system based on a multi-camera approach [44] in the base. Further refinement of the pose is obtained by using visible AprilTags in the environment. Finally, based on the current localization, all scene objects are loaded from a central object database. It is noteworthy that the query returns more objects than visible to the camera as the auditory system is capable of perceiving more of the world than the field-of-view of the camera. As seen in Figure 6 the system returns the list of scene objects needed for the fusion process.

Multi-Modal Fusion and Processing

In this step the information of the audio and visual branch are fused together to obtain a multi-modal description of the world. The DoA estimation retrieved from the audio beamforming module is used to cast a ray from the current position of the robot and infer the 3D position of the sound source using the known geometric outline of scene objects obtained from the vision branch. If a sound source can be located within an object, the relevant spectral information of the given entity is loaded from the database. Finally, this is compared to the received spectrum and the state is inferred.

Object Database

The aforementioned object database is a storage of prior knowledge obtained before the operation of the system (Figure 7). It contains for each object in the environment its exact position, orientation and geometric outline. Furthermore, it also contains a list of spectral information of different states. Each consists of the median and an acceptance band of normalized frequency spectra, e.g. Figure 8 displays the characteristic spectrum of a running drill. This is used to estimate the state of an object or infer if the observed situation is unknown.

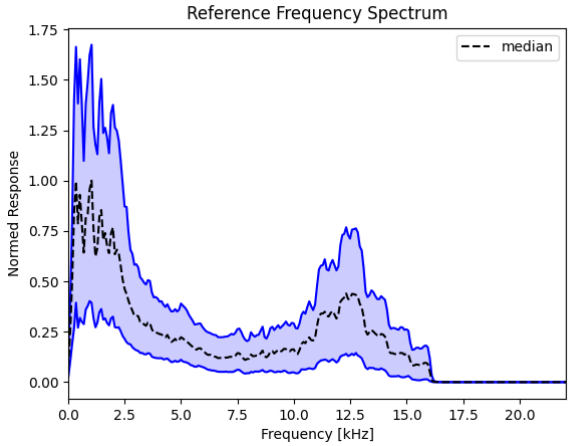

Figure 8: Normalized frequency spectrum of a running drill. The median is depicted as a dashed line. The acceptance band of the sound spectrum is shown by the 20th and 80th percentile.

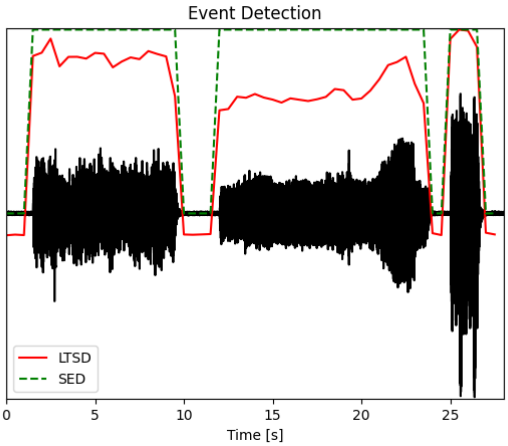

Figure 9: LTSD response for three different sound sources. The input audio is separated into *sound event* (SED=high) and *noise* (SED=low).

4. METHODOLOGY

The following sections describes the key aspects of the selected approaches and applied customizations in depth.

LTSD Power Evalutation

The module for detecting sound events is based on the voice activity detection (VAD) approach by Ramırez et al. [45]. The received input signal is analyzed on smaller chunks. Each is further divided into overlapping subframes, which are transformed into the frequency domain using a short-term Fourier transform. We estimate the spectral envelope for the chunk for the frequency bin l on N subframes as

$$\mathrm{LTSE_N(k)} = \max\left(\mathrm{X(k,0)}, \mathrm{X(k,1)}, ..., \mathrm{X(k,N)}\right) \tag{1}$$

with $X(k,n)$ representing the k-th bin of the n-th subframe.

Each long-term spectral envelope (LTSE) value represents the current maximal gain for each frequency bin in the envelope. To receive information on the overall spectrum differs from the noise reference ξ, we calculate the long-term spectral divergence (LTSD) as given by

$$\mathrm{LTSD_N} = 10\log_{10}\left(\frac{1}{\mathrm{n_{FFT}}}\sum\frac{\mathrm{LTSE^2(k)}}{\xi^2(\mathrm{k})}\right) \tag{2}$$

with $\mathrm{n_{FFT}}$ as the amount of frequency bins in each subframe analysis. Subsequently inserting the audio chunks, we receive a temporal trend of the LTSD responses. A typical result can be seen in Figure 9. Furthermore, we exploit Equation (2) and retrieve the m-most deviating frequency bins compared to the reference ξ and propagate this information to the beamformer module.

MUSIC DoA Estimation

We integrate a modified implementation [22] of the MUSIC algorithm [16] [21] to locate sources using the directed subspaces of the frequency domain. Considering the complex short-term input signal $s_k(t)$ for the k-th frequency band, we get

$$\begin{aligned} s_k(t) &= \lambda_k(t)e^{i2\pi f_k t} \\ &= \lambda_k(t)e^{i\omega_k t} \end{aligned} \tag{3}$$

For a linear microphone array of N sensors where each signal is delayed by

$$\Delta_n = \frac{d_n \sin(\theta)}{c_0} \tag{4}$$

with the DoA θ and the speed of sound c_0, we can construct the system equation as

$$\begin{bmatrix} 1 \\ e^{i\omega_k\Delta_1} \\ e^{i\omega_k\Delta_2} \\ \vdots \\ e^{i\omega_k\Delta_N} \end{bmatrix} s_k(t) =: \boldsymbol{a_k} s(t) \tag{5}$$

We denote $\boldsymbol{a_k}$ as the *steering vector* of the sound source, describing the angular dependency of the received signal to the direction of arrival. As described in the referenced work, the source subspace $\boldsymbol{U}_S$ of the received signal is extracted. The aforementioned *steering vector* is an element of the signal subspace, therefore

$$\boldsymbol{a_k} \in \boldsymbol{U}_S, \tag{6}$$

$$\Rightarrow \boldsymbol{a_k} \perp \boldsymbol{U}_\Sigma \tag{7}$$

of the noise subspace $\boldsymbol{U}_\Sigma$. We can formulate the response equation as

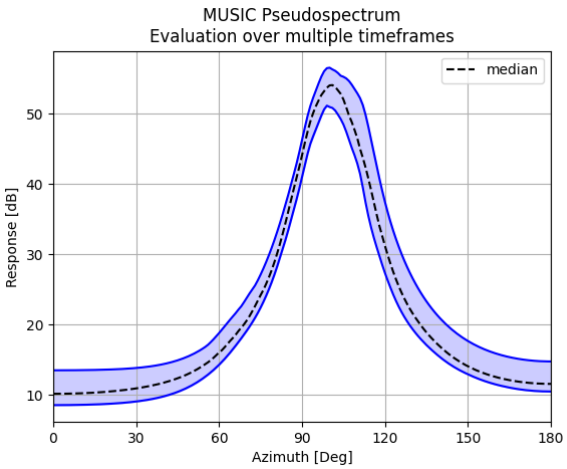

Figure 10: Pseudospectrum as returned from the custom MUSIC implementation. The frequency evaluation is adapted to the current received spectrum and the DoA can be reconstructed from the signal maximum.

$$P(\theta) = 10 \log_{10} \sum_{k=1}^{K_N} \frac{1}{\langle \boldsymbol{a_k}, \boldsymbol{U}_\Sigma \rangle^2} \tag{8}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. We further only examine the N-most deviating frequency bins as calculated in the LTSD power evaluation to integrate into the final response. This reduces the amount of resources needed to process the data while increasing the robustness in low SNR scenarios. An exemplary pseudospectrum is displayed in Figure 10 showing a detected sound source at $\approx 100°$.

Modality Fusion

Processing the separate modalities independently, the modality fusion combines both branches and estimates the joint state. Based on the global position of the system, a set of scene objects is loaded from the object database. The received geometry is projected on the 2D ground plane as the microphone array is only capable to distinguish between azimuth but not elevation angles. As ray is casted starting at the microphones reference position and with the estimated orientation. The ray is tested with each outline of the scene objects for an intersection. The point is reprojected to the microphone array and checked against the sensor accuracy to take measurement tolerances into account. Finally, after testing all lines, the intersection with the shortest ray length is taken as the source position.

Spectral Classification

As a last step, the spectral information of the object is examined. The received audio is compared in the frequency domain with already obtained spectral profiles. For each profile, a audio sample is recorded with a duration of at least 5s. These audio samples are transformed with a short-term Fourier transform (SFTF) using small overlapping subframes with a hop-parameter of 32 samples. The median spectrum P_{50} is calculated over all received spectra. The highest value of the median is used to normalize the spectrum and constraint it to $[0, 1]$. Afterwards, the 20th percentile P_{20} and the 80th percentile P_{80} for each frequency bin are taken as the lower and upper bound of the acceptance band. When receiving a new and unclassified spectrum, first the spectral components of the background noise is subtracted from the input signal. Afterwards, the median spectrum is estimated and normalized. We calculate the sum of the squared differences of frequencies that are within the acceptance band range of each bin.

$$s = \sum_{k \in K} \frac{1}{s_k} \tag{9}$$

$$s_k = \begin{cases} 0 & X_k < P_{20,k} \\ (X_k - P_{50,k})^2 & P_{20,k} \leq X_k \leq P_{80,k} \\ 0 & X_k > P_{80,k} \end{cases} \tag{10}$$

The received score describes the similarity of two frequency spectra within the acceptance band. Further we can set a threshold τ for recognizing known profiles. A analyzed spectrum is only considered if the $s \geq \tau$, ultimately leading to the assumption, if no score passes the threshold, the spectrum originates from an unknown source.

5. Evaluation

For the evaluation, we consider the scenario of a dexterous mobile robot operating in a Martian environment. During the final ISS-Earth experiment session of METERON SUPVIS Justin [38] [46], ESA astronaut Alexander Gerst was tasked with finding, and replacing a failed component in a SPUs in the simulated Martian environment on ground. To recover to nominal operation, the operator first had to search for the problem with visual inspection of all components in the environment. This failure investigation and maintenance (shown in Figure 11), was, as expected, time-consuming. This inspired us to consider other modes of surveying the environment to achieve faster failure detection and localization. This desire turned us to audio perception, to remotely infer the state of an object.

We start with an evaluation in a simulated environment showing the applicability of our method for audio perception and finally show experiments conducted in our laboratory to show the transferability to actual applications.

Simulation

We use a simulated environment of a room with a rectangular floor shape of $W = 8\text{m}$, $L = 8\text{m}$ and a constant height of $H = 4\text{m}$. Further, we define the absorption properties of the walls, the floor and the ceiling based on the data in [47] to mimic the acoustic behavior of our lab. The floor is constructed of rigid plywood with a linoleum surface. The northern and eastern wall are of hard surfaces. The ceiling and southern, as well as the western wall are with high absorption to reflect open space. All parameters are shown in Table 1. We design a reverberation time of $t_{60} = 0.5\text{s}$ for our evaluation.

We further placed three sound sources (an *engine*, a *press*, and an unknown *air valve*) in the room, each emitting a different pre-recorded sound. An eight-sensor microphone array with the same directivity pattern as the future integrated sensor array of the system is placed at the south wall of the room. The resulting room is shown in Figure 12 and the source-specific room impulse response (RIR) in Figure 13.

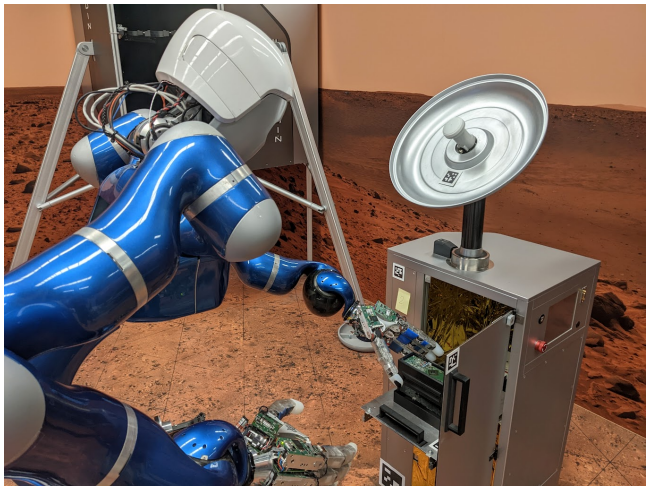

Figure 11: Detecting and replacing a failed component in a simulated Martian habitat. Prior missions required visual inspection of the enclosed modules for failure detection. Robot audition can enable remote detect the components' state, which can speed up anomaly detection.

Table 1: Material absorption properties at different frequencies were used for the simulation.

Element	250Hz	500Hz	1kHz	2kHz	4kHz	8kHz
Floor	0.21	0.10	0.08	0.06	0.06	0.06
Ceiling	0.45	0.55	0.60	0.90	0.86	0.75
Wall N	0.02	0.03	0.03	0.04	0.05	0.05
Wall E	0.02	0.03	0.03	0.04	0.05	0.05
Wall S	0.93	1.00	1.00	1.00	1.00	1.00
Wall W	0.93	1.00	1.00	1.00	1.00	1.00

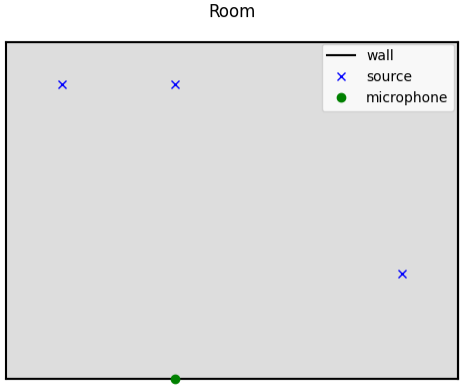

Figure 12: Simulated room environment. Displayed are the three sound sources, the position of the first microphone of the sensor array and the dimensions of the room. Absorption properties of all elements can be extracted from Table 1.

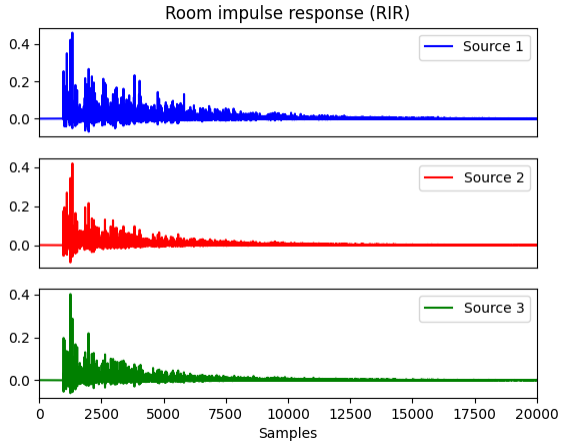

Figure 13: Estimated RIR of the simulated environment in Figure 12. The graphs show the propagation delay each signal needs to reach the first microphone. Further, echo can be identified as the following peaks in the graph. The slow drop after the impulse is due to the reverberation of $t_{60} = 0.5$s.

The simulated audio data is loaded into the proposed processing pipeline. Positions of the system and scene objects are altered by an uncertainty of 10cm. An exemplary result of the data fusion is shown in Figure 14 and shows the localization of a simulated source. For classification, we evaluate the naïve approach of comparing the sum of squared differences (SSD) and our proposed method of calculating the difference in the acceptance band.

The resulting score distribution is shown in Figure 15. The SSD approach for classification yields to individual class scores that are mostly in the range of $[10, 20]$. In general, narrow spectral profiles like *drill* or *saw* result in similar scoring results. Since the complete spectrum is compared, and in the case of a narrow-band signal, most of the spectral components are the background noise which scores a high similarity in this approach. Contrary, our approach takes the variance of the pre-recorded profile into account. While still a fairly simple approach, it results in high deviating class scores and is more robust to narrow-band profiles.

Further, we expect unforeseen sound events to occur and the spectral information of those is unknown. Since our classification approach is explicitly designed to handle this case, it estimates the score only on the acceptance band, thus yielding a significantly lower score compared to known sound profiles. An example can be seen in Figure 16. SSD scores in a comparable range as in the case of a known source. In the given example, it results in the selecting the *saw* class as it is a highly narrow-band profile and therefore more frequency bins with only the background noise. Our approach scores higher values for wide-band profiles like *engine* or *press* due to the higher probability of components of the unknown source laying coincidentally within the acceptance band. However, the overall scoring range is below 1 and by deploying a threshold of $\tau = 5$ including a safety margin, we can safely classify the input signal as *unknown*.

We further investigate the impact of the SNR and the number of simultaneously emitting sources on the successful inference in the modality fusion outcome. We place one, two and three sources in the room and artificially change the SNR

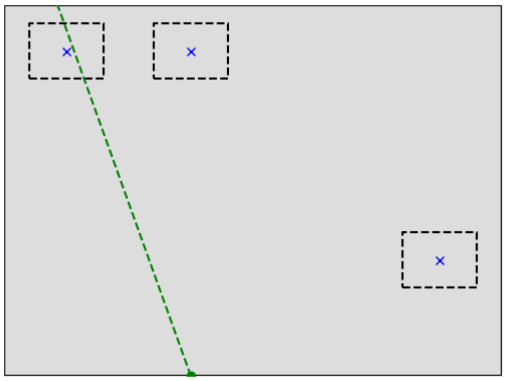

Figure 14: Illustration of the simulated room including three objects and the microphone array. The estimated DoA is shown as a dashed line. By using ray tracing, the source can be located within the object on the upper left-hand side.

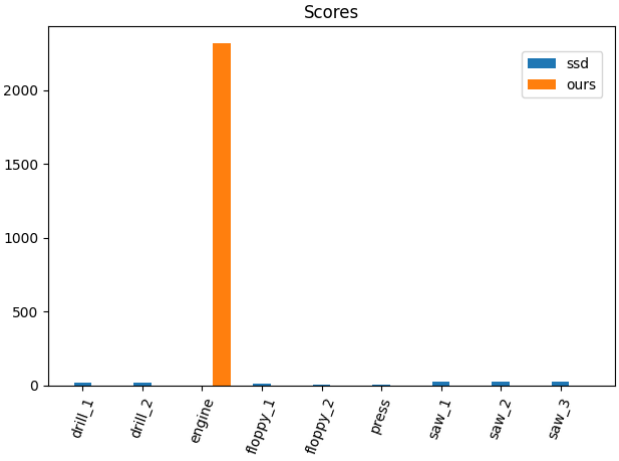

Figure 15: Results of the classification process for a known spectral profile of an engine. While the naïve approach SSD performs poorly and only small deviations between different sound profiles are recognizable, our approach correctly classifies the profile.

of the target in the simulation. The noise sources are set to be at $\mathrm{SNR} = 10\mathrm{dB}$ compared to the background noise. We sample 50 different scenarios where the sources a placed at random positions in the $3\mathrm{m}$ cone as defined in [41] at distances in the range of $[0.5\mathrm{m}, 5.0\mathrm{m}]$. We define a threshold of 0.90 for the desired hit rate as this is a good trade-off on correctly detected sound events and misses in our scenario. The results of the evaluation are shown in Figure 17. While in the single source case the threshold is already reached at $\mathrm{SNR}_{\mathrm{target},1} \approx 5\mathrm{dB}$, additional sources decrease the performance. For two active sources the minimum ration is increased to $\mathrm{SNR}_{\mathrm{target},2} \approx 20\mathrm{dB}$, for three sources the threshold is reached at $\mathrm{SNR}_{\mathrm{target},3} \approx 30\mathrm{dB}$.

Concluding with respect to a future use-case within the Surface Avatar mission, the results show that the perception of audio events and the fusion of the different modalities is feasible. The simulated signals could be identified for their origin

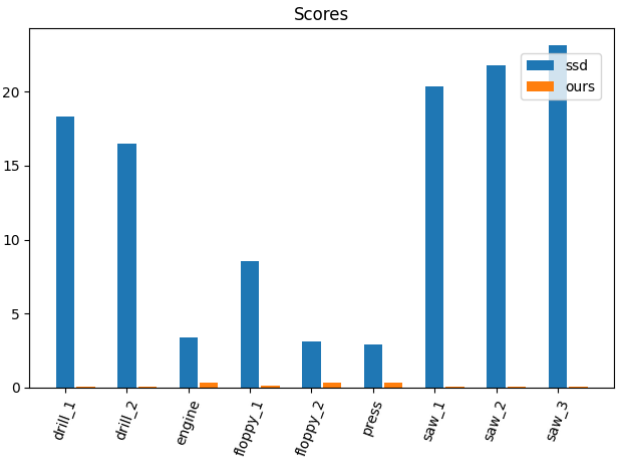

Figure 16: Results of the classification process for an unknown sound event. Compared to the results in Figure 15 the score of our approach is significantly lower and it can be easily stated the system received an unknown spectral profile.

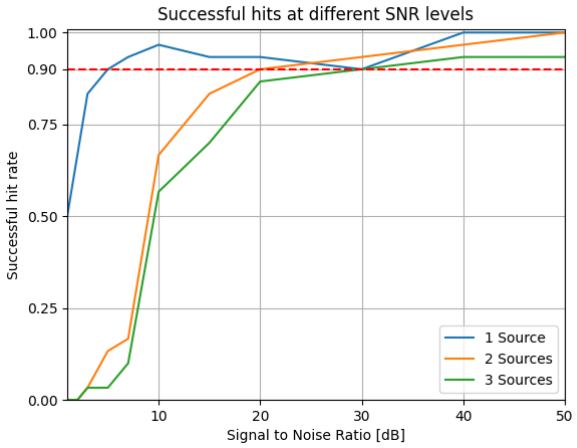

Figure 17: Evaluation of the relation between SNR, the number of active emitting sources and the success rate of detecting the target object. The three curves (blue, orange, green) show the rates for one, two and three sources respectively. We added a threshold of 0.90 as the minimum success rate for use in our scenario. It can be seen that additional sources increase the minimum SNR for successful operation.

and a simple yet effective approach based on an acceptance band in the spectral profile led to the successful estimation of different states. However, the presence of additional sources in the environment affects the performance of the processing pipeline and the rate of successful identification of emitting objects. Assuming that the robotic system itself will be acting as an emitting source in the world, a SNR of at least $20\mathrm{dB}$ must be assured for operation.

Lab Evaluation

Further evaluation is conducted on recordings taken in a laboratory environment. In this experiment, we aim to show the transferability of our approach into a realistic scenario.

A speaker is placed inside of one of the scene objects and

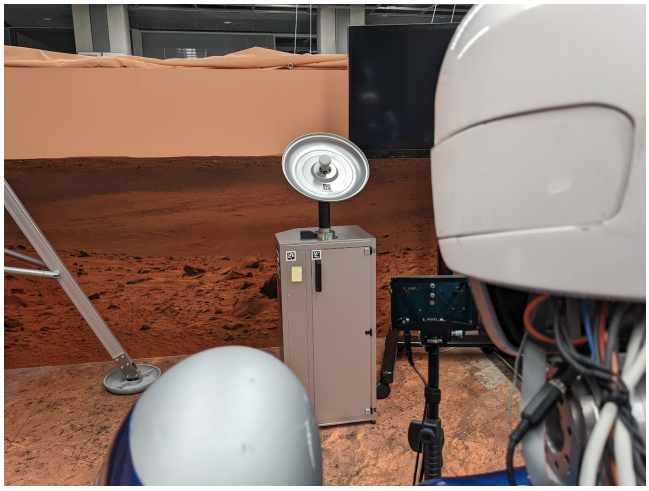

Figure 18: Audio perception evaluation in the METERON environment. A speaker is placed inside the SPU, shown with an antenna mounted on top, and is emitting a recording of running drill at low and high speed. The system shall differentiate between both states.

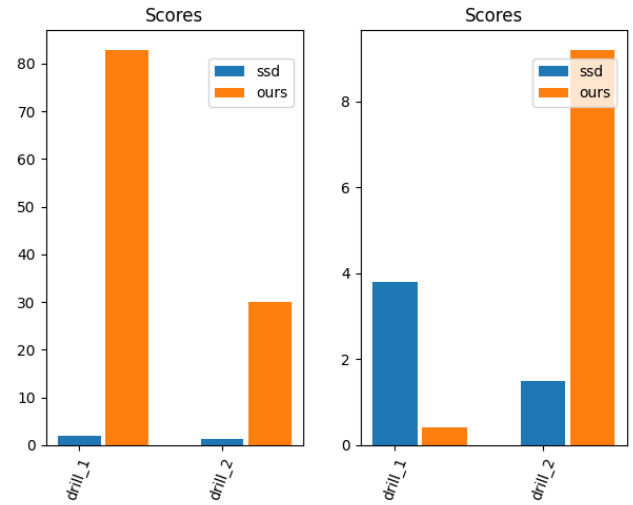

Figure 19: Scoring results for the state estimation of the drill object. The approach correctly estimated the correct states of the object on *low* and *high* speed. As the system identified the object according to its prior knowledge, only the associated sound profiles are loaded for estimation.

is set to alternately emit the sound of the pre-recorded drill at low and high speed. As discussed before, the robot is emitting noise and is a sound source itself in the environment resulting in a minimum of at least two sources at the same time. The speaker is set to transmit at an average of $50\mathrm{dB}$ taking into account the transmission from inside the object and over the distance to the sensor array to meet the requirement of $\mathrm{SNR}_{\mathrm{target},2} \approx 20\mathrm{dB}$. The sensor array is positioned in front of the robot facing the same direction as the camera interface. The setup is shown in Figure 18 including the robot, the sensor array and the target object. All data is fed into the proposed processing pipeline, including the prior knowledge of the positions of objects, possible spectral profiles and the mapping of object's emitting frequencies. The source object is detected in the localization module, the robot's knowledge is updated according to the database content and the state is determined based on the received audio signal.

An exemplary result is shown in Figure 19. Based on the prior knowledge, the system reduced the total amount of possible spectral profiles to two, *drill_1* and *drill_2*. The classification resulted in correct associations with the emitting profiles. However, the yielded scores a significantly lower than compared to the simulated ones. This is due to further sources in the environment emitting sounds that are overlaying the audio signal and induce further spectral noise. The transmission through the scene objects and the frequency depending sampling accuracy of the microphones were not simulated. Nevertheless, the preliminary results already show that the desired estimation can be achieved under lab conditions.

6. Conclusion and Outlook

In this work we presented a first study on the integration of audio perception into the context of the Surface Avatar mission led by DLR with partner ESA. We aimed to show the usability of audio input as an additional perception modality to improve situational awareness of the surface environment where the robot is operating in. This can offer further information on the world and the state of objects in the robot's surrounding.

Our approach is divided into an audio and vision branch, which eventually are fused into a single state estimation of located sources withing known objects. We further introduces a method to compare the received spectral information with prior learned profiles. Moreover, the system is able to detect unknown profiles which are not part of the set of known data.

We showed the performance of our proposed method in simulation as well as the transferability in a real scenario. The sound source localization yields high accuracy in combination with the visual perception and results in robust fusion of the two modalities for spectral profile and state estimation. Deployed to our laboratory, the system was able to detect and estimate the current state of the target object.

The presented system shall be integrated into the perceptional system of the robotic assistance platforms and provide the audio modality in the upcoming Surface Avatar ISS-Earth telerobotic experiment sessions in 2023-2024.

References

[1] Lii, N. Y. et al., "Introduction to Surface Avatar: the First Heterogeneous Robotic Team to be Commanded with Scalable Autonomy from the ISS," in *Proceedings of the 73rd International Astronautical Congress (IAC)*. International Astronautical Federation, September 2022.

[2] L. A. Jeffress, "A place theory of sound localization." *Journal of comparative and physiological psychology*, vol. 41, no. 1, p. 35, 1948.

[3] J. Huang, N. Ohnishi, and N. Sugie, "Building ears for robots: sound localization and separation," *Artificial Life and Robotics*, vol. 1, no. 4, pp. 157–163, 1997.

[4] K. Nakadai, K.-i. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-

object tracking for humanoids," in *International Joint Conference on Artificial Intelligence*, vol. 17, no. 1. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001, pp. 1425–1436.

[5] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: Robust sound source localization and extraction," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 2. IEEE, 2003, pp. 1147–1152.

[6] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4290–4296, 2008.

[7] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3-d localization based on hrtfs," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5. IEEE, 2006, pp. V–V.

[8] I. Kossyk, M. Neumann, and Z.-C. Marton, "Binaural bearing only tracking of stationary sound sources in reverberant environment," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 53–60.

[9] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, vol. 2. IEEE, 2003, pp. 1228–1233.

[10] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 1. IEEE, 2004, pp. 1033–1038.

[11] E. Mumolo, M. Nolich, and G. Vercelli, "Algorithms for acoustic localization based on microphone array in service robotics," *Robotics and Autonomous systems*, vol. 42, no. 2, pp. 69–88, 2003.

[12] R. Roden, N. Moritz, S. Gerlach, S. Weinzierl, and S. Goetze, *On sound source localization of speech signals using deep neural networks*, 2015.

[13] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[14] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[15] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 603–609.

[16] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[17] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[18] S. Argentieri and P. Danes, "Broadband variations of the music high-resolution method for sound source localization in robotics," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2007, pp. 2009–2014.

[19] F. Asono, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot" jijo-2"," in *Proceedings. 1999 IEEE/SICE/RSJ. International Conference on Multisensor Fusion and Integration for Intelligent Systems. MFI'99 (Cat. No. 99TH8480)*. IEEE, 1999, pp. 243–248.

[20] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 2027–2032.

[21] K. Nakamura, K. Nakadai, and G. Ince, "Real-time super-resolution sound source localization for robots," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 694–699.

[22] M. Sewtz, T. Bodenmüller, and R. Triebel, "Robust music-based sound source localization in reverberant and echoic environments," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2474–2480.

[23] L. S. M. Sugai, C. Desjonqueres, T. S. F. Silva, and D. Llusia, "A roadmap for survey designs in terrestrial acoustic monitoring," *Remote Sensing in Ecology and Conservation*, vol. 6, no. 3, pp. 220–235, 2020.

[24] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, and D. Llusia, "Terrestrial passive acoustic monitoring: review and perspectives," *BioScience*, vol. 69, no. 1, pp. 15–25, 2019.

[25] D. Llusia, R. Márquez, and R. Bowker, "Terrestrial sound monitoring systems, a methodology for quantitative calibration," *Bioacoustics*, vol. 20, no. 3, pp. 277–286, 2011.

[26] E. P. Kasten, S. H. Gage, J. Fox, and W. Joo, "The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology," *Ecological informatics*, vol. 12, pp. 50–67, 2012.

[27] R. Kojima, O. Sugiyama, R. Suzuki, K. Nakadai, and C. E. Taylor, "Semi-automatic bird song analysis by spatial-cue-based integration of sound source detection, localization, separation, and identification," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1287–1292.

[28] A. Digby, M. Towsey, B. D. Bell, and P. D. Teal, "A practical comparison of manual and autonomous methods for acoustic monitoring," *Methods in Ecology and Evolution*, vol. 4, no. 7, pp. 675–683, 2013.

[29] R. Suzuki, S. Matsubayashi, R. W. Hedley, K. Nakadai, and H. G. Okuno, "Harkbird: Exploring acoustic interactions in bird communities using a microphone array," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 213–223, 2017.

[30] C. Astaras, J. M. Linder, P. Wrege, R. D. Orume, and D. W. Macdonald, "Passive acoustic monitoring as a law

enforcement tool for afrotropical rainforests," *Frontiers in Ecology and the Environment*, vol. 15, no. 5, 2017.

[31] J. S. Ulloa, T. Aubin, D. Llusia, C. Bouveyron, and J. Sueur, "Estimating animal acoustic diversity in tropical environments using unsupervised multiresolution analysis," *Ecological Indicators*, vol. 90, pp. 346–355, 2018.

[32] L. W. Koester, H. Taheri, L. J. Bond, and E. J. Faierson, "Acoustic monitoring of additive manufacturing for damage and process condition determination," in *AIP Conference Proceedings*, vol. 2102, no. 1. AIP Publishing LLC, 2019, p. 020005.

[33] M. S. Hossain and H. Taheri, "In situ process monitoring for additive manufacturing through acoustic techniques," *Journal of Materials Engineering and Performance*, vol. 29, no. 10, pp. 6249–6262, 2020.

[34] J. Bynum and D. Lattanzi, "Combining convolutional neural networks with unsupervised learning for acoustic monitoring of robotic manufacturing facilities," *Advances in Mechanical Engineering*, vol. 13, no. 4, p. 16878140211009015, 2021.

[35] Carey, W. et al., "Analog-1: A touch remote," in *Proceedings of the 73rd International Astronautical Congress (IAC)*. International Astronautical Federation, September 2022.

[36] Wedler, A. et al., "Finally! Insights into the ARCHES lunar planetary exploration analogue campaign on etna in summer 2022," in *Proceedings of the 73rd International Astronautical Congress (IAC)*. International Astronautical Federation, September 2022.

[37] T. Krueger, E. Ferreira, A. Gherghescu, L. Hann, E. den Exter, F. P. van der Hulst, L. Gerdes, A. Pereira, H. Singh, M. Panzirsch *et al.*, "Designing and testing a robotic avatar for space-to-ground teleoperation: the developers' insights," in *71st International Astronautical Congress, IAC 2020*. International Astronautical Federation, 2020.

[38] N. Y. Lii, D. Leidner, P. Birkenkampf, B. Pleintinger, R. Bayer, and T. Krueger, "Toward scalable intuitive telecommand of robots for space deployment with meteron supvis justin," 2017.

[39] C. Borst, T. Wimbock, F. Schmidt, M. Fuchs, B. Brunner, F. Zacharias, P. R. Giordano, R. Konietschke, W. Sepp, S. Fuchs *et al.*, "Rollin'justin-mobile platform with variable base," in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 1597–1598.

[40] V. et al., "An ecosystem for heterogeneous robotic assistants in caregiving: Core functionalities and use cases," *IEEE Robotics & Automation Magazine*, vol. 28, no. 3, pp. 12–28, 2021.

[41] M. Sewtz, T. Bodenmüller, and R. Triebel, "Design of a microphone array for rollin justin," in *ICRA Workshop*, 2019.

[42] R. Bayer, P. Schmaus, M. Pfau, B. Pleintinger, D. Leidner, F. Wappler, A. Maier, T. Krueger, and N. Y. Lii, "Deployment of the solex environment for analog space telerobotics validation," in *Proceedings of the International Astronautical Congress, IAC*, 2019.

[43] J. Wang and E. Olson, "Apriltag 2: Efficient and robust fiducial detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4193–4198.

[44] M. Sewtz, X. Luo, J. Landgraf, T. Bodenmüller, and R. Triebel, "Robust approaches for localization on multi-camera systems in dynamic environments," in *2021 7th International Conference on Automation, Robotics and Applications (ICARA)*. IEEE, 2021, pp. 211–215.

[45] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[46] P. Schmaus, D. Leidner, T. Krüger, R. Bayer, B. Pleintinger, A. Schiele, and N. Y. Lii, "Knowledge driven orbit-to-ground teleoperation of a robot coworker," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 143–150, 2019.

[47] M. Vorländer, *Auralization*. Springer, 2020.

BIOGRAPHY

Marco Sewtz *received his B.Eng. degree in electrical engineering at the University of Applied Sciences of Munich and his M.Sc. degree at the Technical University of Munich. He works at the Institute for Robotics and Mechatronics at the German Aerospace Center (DLR) as a researcher since 2018. His interests focuses on SLAM and multi-model perception of the environment. Before his current role, he worked as an electrical designer for high-performance processing modules for space hardware at Airbus Defence and Space.*

Werner Friedl *received his Dipl.-Ing.(FH) in Mechatronic at the University of applied science In Munich and starts at DLR in 2004. 2006 he develop the torso of DLR Humanoid Justin. In the DLR Hand-Arm- project he developed the forearm of the AWIWI hand and AWIWI II. Since 2015 he is responsible for the mechanical hand development at DLR. His main research focus includes variable stiffness actuation, tendon driven hands and grasping.*

Adrian Bauer *received a Bachelor in Mechanical Engineering in 2012, a Bachelor in Cognitive Sciences in 2015, and a master in Robotics, Cognition, Intelligence from TU Munich in 2018. Currently he is pursuing a PhD in robotics at the German Aerospace Center. His interest is in enabling robotics to generate meaningful symbolic plans in presence of epistemic uncertainty.*

Anne Köpken *received a Bachelor in Electrical Engineering from TU Munich in 2019, and a Master in Robotics, Cognition, Intelligence from TU Munich in 2021. She spent one semester at the JSK Laboratory at the University of Tokyo in 2019/20. Currently she is pursuing a PhD in robotics at the German Aerospace Center in Oberpfaffenhofen near Munich. She is interested in enabling robots to cope with unexpected situations and finding ways to prevent and recover from failures.*

Florian S. Lay *received his B.Sc. degree in Engineering Science in 2018, and his M.Sc. in "Robotics, Cognition, Intelligence" in 2020 both from the Technical University of Munich. Since 2020 he is pursuing a PhD in robotics at the German Aerospace Center (DLR). His interests range from symbol grounding and emergence for task and motion planning to multi-robot world representations.*

Nicolai Bechtel *received his Master of Science in Computational Engineering from the University of Applied Sciences Munich in 2018. Since then, he has been conducting research in the field of haptics and virtual reality as a research assistant at the Center for Robotics and Mechatronics of the German Aerospace Center (DLR) in Oberpfaffenhofen. His research focuses on haptics, multi-body-dynamic simulations, and development of virtual reality environments. He is currently working on topics such as Model-Based Teleoperation and GUI development involving Augmented Reality.*

Peter Schmaus *received his M.Sc. Degree in "Robotics, Cognition, Intelligence" from Technical University of Munich, Germany, in 2013. He joined the German Aerospace Center (DLR) Institute of Robotics and Mechatronics in 2011 where he was involved in the ISS-to-ground telerobotics projects Kontur-2, METERON SUPVIS Justin, and became Co-Investigator of the Surface Avatar experiment suite. His main interests lie in Shared Autonomy and effective Human-Robot Interaction.*

Rudolph Triebel *received his PhD in 2007 from the University of Freiburg in Germany. From 2007 to 2011, he was a postdoctoral researcher at ETH Zurich, where he worked on machine learning algorithms for robot perception. From 2011 to 2013 he worked in the Mobile Robotics Group at the University of Oxford. In 2015, he was appointed as leader of the Department of Perception and Cognition at the Robotics Institute of DLR.*

Neal Y. Lii *is the domain head of Space Robotic Assistance, and the co-founding head of the Modular Dexterous (Modex) Robotics Laboratory at the German Aerospace Center (DLR). Neal received his BS, MS, and PhD degrees from Purdue University, Stanford University, and University of Cambridge, respectively. He has served as the principal investigator of the ISS-to-Earth telerobotic experiments, Surface Avatar, and METERON SUPVIS Justin. Neal is primarily interested in the use of telerobotics in both space and terrestrial applications.*