Robot Audition and Computational Auditory Scene Analysis

Kazuhiro Nakadai* and Hiroshi G. Okuno

Robot audition aims at developing robot's ears that work in the real world, that is, machine listening of multiple sound sources. Its critical problem is noise. Speech interfaces have become more familiar and more indispensable as smartphones and artificial intelligence (AI) speakers spread. Their critical problems are noise and multiple simultaneous speakers. Recently two technological advances have contributed to significantly improve the performance of speech interfaces and robot audition. Emerging deep learning technology has improved noise robustness of automatic speech recognition, whereas microphone array processing has improved the performance of preprocessing such as noise reduction. Herein, an overview and history of robot audition are provided together with introduction of an open-source software for robot audition and its wide applications in the real world. Also, it is discussed how robot audition contributes to the development of computational auditory scene analysis, that is, understanding of real-world auditory environments.

1. Introduction

"Robot Audition" is a research area originating in Japan that aims at the construction of the robot's auditory functions.[1] At that time, many robots were announced one after another from Japanese companies, institutes, and universities for the 2005 Aichi Expo. Most of them were developed for mobility and execution of specific tasks such as cleaning. Some robots capable of verbal human–robot communication have appeared; users have to use a headset microphone worn close to their mouth because the robot cannot listen to sounds with its own ears due to noise problems. Robot audition was proposed to solve this extremely unnatural problem by claiming the following key requirements:[1] 1) Understanding General Sound: to assume that an input sound is always a mixture of sound sources, which derived from Computational Auditory Scene Analysis (CASA); 2) Active Audition: to use active motions to improve auditory functions by suppressing noises generated due to motions; 3) Multimodal Integration: to integrate multimodal sensory information for mutual disambiguation of missing or ambiguous sensory information; 4) Online and Real-Time Processing: to consider online and real-time processing for human–robot interactive scenarios.

To achieve the first requirement, it is necessary to deal with noise contamination of a target sound source and simultaneous sound sources. The standard approach is to use the amplitude and phase differences between multiple microphones to suppress noise and/or to separate sound sources because these differences change according to the direction/location of a sound source. Using this principle, many studies on sound-source localization, sound source separation, and automatic speech recognition have been reported in robot audition. There are mainly two approaches. One is called binaural processing, which is defined as using two microphones together with biologically inspired methods. The other is microphone array processing, which uses multiple microphones based on acoustic signal processing. Although these approaches have been well studied in acoustic signal processing, most have focused on exploration of human/animal hearing mechanisms or on mathematical formulations with numerical simulation. The other three requirements have not been considered and thus the existing techniques in acoustic signal processing could not be applied directly to robots. In robotics, among the aforementioned requirements, online and real-time processing has been pursued. Although a few studies related to multimodal integration has been reported, audio has been considered as a complement of vision. Before robot audition was proposed, the concepts of understanding general sound and active audition did not exist in robotics, and thus, a robot had to follow a strategy called “stop–perceive–act” to avoid motion noise generation on listening.[1] Therefore, considering all four of the aforementioned key requirements makes a clear distinction between robot audition and other related fields such as acoustic
signal processing and speech processing, whereas presenting robot audition a new research area in robotics.

For the first 10–15 years, elemental technology has been developed with binaural and microphone array processing. After that, deployment of robot audition has actively been performed in various fields, thanks to the release of open-source software for robot audition. More recently, it has started to expand to scene analysis and understanding considering higher-level cognition by integration with deep learning technology.

The rest of this article is organized as follows: Section 2 describes an overview and history of robot audition. Section 3 introduces open-source software for robot audition HARK (Honda Research Institute Japan Audition for Robots with Kyoto University), which plays an important role in collaboration and interdisciplinary research. Section 4 shows the deployment of robot audition with HARK. Section 5 presents recent research topics for more practical robot audition. The last section concludes the article.

2. History of Robot Audition

This section explains the history of robot audition after it had been proposed in 2000, by dividing it into three periods; CASA, Binaural Robot Audition, and Microphone Array-Based Robot Audition.

2.1. Computational Auditory Scene Analysis

The genealogy of robot audition goes back to “Auditory Scene Analysis (ASA)” written by A.S. Bregman in 1990. ASA aims to elucidate human auditory functions psychophysically based on the idea that humans perceive each sound as a sound stream. In a general environment where there are multiple sound sources, ASA claims that a mixture of the sound sources is perceived as multiple streams by “stream segregation” based on various cues. This concept of human perception when multiple sound sources exist simultaneously has drawn attention because mainly a single sound source has been considered before that. Inspired by ASA, CAS was proposed as a constructivist approach to elucidate human auditory functions in the 1990s. Most studies in CASA have been conducted with numerical simulations in the beginning, but it gradually addressed issues closer to the real world. In particular, music information processing has been actively promoted, with the International Conference of Musical Information Retrieval (ISMIR) being established in 2000.

2.2. Binaural Robot Audition

On the other hand, aiming at understanding a more general sound environment, which is not limited to specific domains such as music, Nakadai and Okuno proposed robot audition in 2000 as a new research area bridging artificial intelligence, robotics, and signal processing. In the early 2000s, a binaural approach that imitated human and animal auditory processing was predominant. As humans and animals have two ears, the idea behind this approach was that the auditory functions could be achieved with two microphones. “Active audition” was also proposed as a challenge unique to robot audition to use active motions to improve auditory functions by suppressing motion noises. Many studies on sound-source localization have been reported such as applications of the Jeffress’s model [2,3,6–10] and neural networks [11].

2.3. Microphone Array-Based Robot Audition

In the mid-2000s, acoustic signal processing using a microphone array consisting of multiple microphones has been applied to sound-source localization and separation for robots. This is caused by the fact that the performance of the binaural approach was poor to deal with the real-world problems of sound-source localization and separation. Researchers in robot audition believed that the usage of a large number of microphones could provide better performance than the binaural approach with two microphones. During this period, Japan had a robot boom stimulated by the upcoming 2005 Aichi Expo, which drove research on making robots with capabilities for human–robot communication in the real world where multiple sound sources exist. Microphone array processing with beamforming and independent component analysis has been extensively studied [1,12–14]. We developed sound-source localization and separation which can be integrated with automatic speech recognition and reported a robot that can listen to simultaneous utterances of 11 people who order food.

As a collection of developed robot audition functions, we also released the open-source software for robot audition, HARK. HARK plays a crucial role in collaboration and...
deployment of robot audition technology, and the next section will introduce HARK.

3. Open-Source Software for Robot Audition

In 2008, the methods developed for robot audition were collected as open-source software called HARK (Honda Research Institute Japan Audition for Robots with Kyoto University)[18] aiming at the audio-equivalent to OpenCV. HARK includes primary functions for robot audition, that is, sound-source localization, sound-source separation, and automatic speech recognition together with other necessary functions to construct a robot audition system such as sound-source tracking, feature extraction, frequency analysis, and so on. HARK is updated almost every year. After every update, we plan to have free tutorials and/or hackathons in Japan and abroad including international conferences such as IEEE Humanoids-2009 and IROS 2018. The total number of downloads has exceeded 16,000, and 17 HARK tutorials have been held with four hackathons as of December 2019.

We set up two design guidelines to develop HARK; user-friendliness and real-time processing. For the first one, we introduced a graphical user interface programming environment shown in the left of Figure 1. It can run on a web browser and the difference between operating system platforms is small. For programming, users just select functional modules (shown as green boxes) from a module list, put them on the panel, and connect between the modules.[16] We also prepared a manual and cookbook with over 300 pages in Japanese and English.

For real-time processing, all functional modules are designed to work online and in real-time. We also prepared a standard 8-ch circular microphone array called TAMAGO shown in the right of Figure 2, which can connect to a personal computer via a universal serial bus interface. It is helpful to make a real-time robot audition system with minimum cost because a set of transfer functions between the standard microphone array and target sound sources can be downloaded from the HARK web site, whereas users need to measure or calculate transfer functions when they want to use their own designed microphone array.

In addition, useful peripheral packages are available. HARK-ROS offers seamless integration with ROS[19] which is de facto standard middleware in robotics. It makes integration with a user’s existing system easier. HARK-OpenCV provides a wrapper for OpenCV,[20] which is a well-known library for computer vision. It enables us to make an audio-visual integrated system using HARK.

4. Deployment of Robot Audition

In the 2010s, research on the deployment of robot audition had started. This is because tools such as HARK to share knowledge

---

Figure 1. An example of HARK networks: programming on a browser, hierarchical programming supported.
and methodologies for deployment became available. Activities for deployment leads to expand the research area of robot audition as shown in Figure 3. Human–robot interaction was the original target application of robot audition, and it covers Information and Communication Technology (ICT) applications,[21] automotive applications,[22] search and rescue applications,[23,24] and ecology and ethology applications.[25]

For ICT applications, using a tablet device with an 8-ch microphone array, we validated the effectiveness of robot audition technology by showing augmented reality-based sound-source localization, support for communication with a hearing-aid, and support for multilingual communication.[21]

For automotive applications, we showed an always listening in-vehicle information (IVI) system based on HARK. For conventional IVI systems, a driver has to push a talk button and wait until the system becomes ready before speaking. Even recent IVI systems are triggered by a wake-up speech command every time a driver wants to speak to the system. Our developed system achieved always-listening function which accepts commands from multiple users sitting at the driver and passenger seats.[22]

For search and rescue applications, we installed a microphone array to a drone. Thanks to highly noise-robust functions of HARK, we successfully showed a live demonstration that the developed drone was able to detect and localize human utterances outdoors in a 3D space from the sky.[23] Another activity for search and rescue is for a hose-type robot that searches for survivors in debris (See the photo of the bottom center of Figure 3). We attached multiple microphones and speakers to the robot at certain intervals. The microphones and speakers are advantageous to support two functions; sound-based posture estimation and noise reduction of the robot’s vibration noise. Sound-based posture estimation significantly reduces drift errors which are inevitable with integral-type sensors such as inertial measurement units, and noise reduction enables the communication between an operator for the robot and a survivor.[24]

For ethology and ecology applications, we are working on automated extraction of when, where, and what information on bird songs. Such information was manually extracted by experts, which has difficulties in repeatability, quality, and area coverage due to the limitation of human listening capability. Using multiple microphone arrays, we developed 3D sound-source localization that can detect bird songs over a wide range of 100 m.[25] The robot audition technology is gradually spreading in the fields of ethology and ecology.[26,27]
5. Recent Progress toward More Practical Robot Audition

This section introduces recent progress in robot audition by picking up two topics. The first one is a search and rescue application using a drone mentioned in the previous section. It is now called “drone audition,” and its research community is growing all over the world. The world-first international symposium on noise from unmanned aircraft systems/unmanned aerial vehicles called Quiet Drones will be held in October 2020. The second one is related to multimodal integration which is mentioned as a requirement in robot audition in Section 1. Several studies have been conducted for multimodal integration such as audio–visual human tracking and audio–visual speech recognition. As a recent topic toward next-generation CASA, audio–visual reconstruction of scenes including dynamic and transparent objects is introduced in this article.

5.1. Drone Audition

When a disaster occurs, it is necessary to find survivors within three days. Otherwise, their chances of survival would be greatly reduced. It is necessary to search for such survivors day and night. However, in the disaster site, all roads are blocked, and emergency vehicles are of little use. Under an idea that a combination of drone and robot audition technology, that is, drone audition, can provide faster and more extensive exploration in such situations day and night, the first activity of drone audition was conducted in 2011 just after the Great East Japan Earthquake. It is also a challenging research topic to deploy robot audition outdoors.

In 2014, the Tough Robotics Challenge (TRC) which is a 5-year project of Impulsing Paradigm Challenge through Disruptive Technologies Program (ImPACT) was launched by the Japanese Cabinet and Japan Science and Technology Agency (JST). Drone audition was strongly supported by TRC, and sound-source localization and extraction with a microphone array mounted on a drone were achieved in real-time under highly noisy outdoor situations where drone noise and other environmental sounds exist. The developed techniques solved two problems. One is a noise problem because the drone equipped with a microphone array itself is a loud noise source. For this problem, we proposed a sound-source localization method that is robust for dynamically changing drone and wind noises. Another problem is that sound-source localization generally estimates a direction of arrival for the target sound source and it does not provide distance to the source. By assuming that the target sound source is at ground level in a 3D map, 3D sound-source localization was achieved. In Figure 4, blue circles are the obtained 3D sound source positions in a live demo. A 3D point cloud map was generated in advance, but real-time point cloud map generation will be realized in the near future.

The developed technology can detect a sound source from the distance of 12–15 m even when a signal-to-noise ratio (SNR) is around −15 dB. In a frame-based sound-source localization rate, obtained by calculating the number of successfully localized frames divided by the total number of signal frames, 80% and

![Figure 4](image-url)

**Figure 4.** Drone audition for search and rescue task: A drone equipped with a 16-ch spherical microphone array searched for people outdoors by detecting their voices from the sky. (a) Two operators control a drone flying. A person wearing a red jacket (top-left) and another in a pipe (bottom-left) are target sound sources. (b) The operator’s computer shows sound localization results as blue circles corresponding two targets on the previously measured point cloud map. (c) A subject in a pipe is successfully detected even when the lid is closed. (d) View for developers. Left shows radar view in the drone’s coordinates. The angle and radius of the circle represent azimuth and elevation, respectively. The white fan shape shows the area to be ignored because extremely high-power noise generated by drone’s propellers exist in this area. When sound sources are detected, bright area will appear in the black part. Right is a topview including frame-based sound-source candidates as red dots, integrated estimation of sound-source positions as blue dots (corresponding to blue circles in [b]), and drone position and trajectory as black dot and line. Adapted with permission. Copyright 2020, Nakadai Lab., Tokyo Tech.
above accuracy is achieved having a low SNR of \(-20\, \text{dB}\).\textsuperscript{[23]} By frame integration, the success rate significantly improves, and the localization errors are at most 3 m\textsuperscript{[35]} even when two speakers utter simultaneously. We, thus, decided that the error allowance for a sound source in a live demo, that is, a radius of blue circle in Figure 4, was set to 1.5 m.

For this search and rescue task, other sensors such as visual and thermal sensors can be considered. A visual sensor is effective for accurate tracking, and its use is limited to daytime when a target is uncovered, whereas drone audition can be applied at night even when the target is hidden in rubble. A thermal sensor can be used under these harsh conditions; however, it detects artifacts produced by other thermal objects such as heavy machinery operating in the field. On the other hand, the resolution and accuracy of tracking using these types of sensors outperform those of drone audition. Therefore, in practice, the development of more robust sensing that integrates drone audition with other sensors will be promising in the future.

### 5.2. Audio–Visual Scene Reconstruction

3D scene reconstruction has been mainly studied in the field of computer vision as Structure from Motion (SfM), which uses multiple photos taken from different viewpoints. This assumes a stationary scene for all taken photos, and thus there is a problem that it cannot deal with dynamic scenes where some objects are in motion. Another problem is that because it cannot extract the features of a transparent object because it relies only on photos taken by a camera. In particular, it is difficult to know whether a transparent object is hollow or solid. We are tackling these problems through audio–visual integration.

As an example of the first problem, let us consider a swiveling fan shown in Figure 5. When the fan is not swiveling as shown in Figure 5a, its 3D scene is properly reconstructed with SfM. However, once the fan starts swiveling as shown in Figure 5b, the part of the swiveling fan cannot be reconstructed. To solve this problem, we propose an audio–visual reconstruction method by assuming that motions generally generate sound.\textsuperscript{[16]} We used a device having a microphone array and a camera together, and captured audio signals and visual images at the same time. Sound-source localization using the captured audio signals estimates a sound source region in the corresponding image, and texture mapping is performed for the sound source region. Figure 5c is the finally obtained reconstruction with the proposed audio–visual integration. In this example, the swiveling motions are not reconstructed. We, thus, extended the method to reconstruct motions as well, which provides 4D scene reconstruction.\textsuperscript{[16]}

In the extended method, each image was divided into stationary and dynamic parts. These two parts are separately reconstructed, and two types of 3D reconstructed scenes are integrated into a 4D scene at the final stage. Figure 6 shows 4D scene reconstruction with the extended audio–visual reconstruction method. A microphone array is located at the center of each image, and a train is running on a circular toy rail. A camera is used to take photos from different angles to cover the whole scene shown in the upper panels in Figure 6. Using the microphone array, a sound region in each image is extracted. The lines from the microphone array in the lower panels of Figure 6 show sound-source localization results to extract the sound regions. Considering that the extracted regions are for a dynamic object, that is, train, 3D reconstruction with SfM is performed for all the extracted regions. The residual images are considered for a stationary background scene. Another 3D reconstruction with SfM is performed using all of these images. Finally, the two reconstructed scenes are integrated into a 4D scene by considering the positions of the dynamic object shown in the lower panels in Figure 6. The extended method still has a limitation that target sound sources always exist on the rail, and a further extension to relax this limitation is ongoing.

Next, the second problem, that is, reconstruction of transparent objects is considered. It is difficult to extract visual features of a transparent object because transparent, especially uniformly transparent, parts are less likely to be captured in the camera image. An acrylic plate exists inside the dotted rectangle in Figure 7a, and it is hard to be captured by a camera. In this situation, 16 images were taken by the camera by changing viewpoints shown in Figure 7b, and SfM was performed using these 16 images. Only the edges of the transparent object were reconstructed as shown in Figure 7c. To solve this problem, we use distance estimation with an audible sound. Indeed, ultrasonic sensors are commonly used for distance estimation; however, the onsets of ultrasonic sound often make annoying noise. In addition, the risk of ultrasonic exposure must be minimized because high-power ultrasonic sound is imperceptible.

![Figure 5. 3D reconstruction of a swiveling fan. a) Visual reconstruction of stationary fan. b) Visual reconstruction of swiveling fan. c) AV reconstruction of swiveling fan.](image-url)
to humans and it adversely affects human hearing ability whether it is mis-emitted or not. The distance estimation with audible sound is more beneficial when it is used in a daily environment where many people may exist. Figure 7d shows an audio–visual reconstruction result. The transparent plane was detected by a combination of an audible-sound-based distance estimation using cross-power spectrum and kurtosis-based plate detection, and the detected transparent plane was integrated with Figure 7c. The red part is the reconstructed transparent plane, and the coverage is around 50%. It is still a preliminary result, but it shows the effectiveness of audio–visual integration.

Figure 6. 3D reconstruction of a running train. The upper panels (1–4) were taken by a camera located at the fixed position. The egg-shaped microphone array was located at the center. Because a train is running, its position is different between the panels. The lower panels (1–4) were reconstructed images by audio-visual integration. Audio-visual reconstruction achieved 3D reconstruction of a dynamic scene, which cannot be done only with conventional SfM using visual information. Reproduced with permission. Copyright 2020, IEEE.

Figure 7. 3D reconstruction of a transparent plane. a) An image with a transparent plane. b) 16 images for SfM. c) Reconstructed only with SfM. d) Reconstructed with audio information.
6. Conclusion

This article presents an overview, history, deployment, and recent progress of robot audition together with the development of open-source software for robot audition called HARK. The next step of robot audition will aim at realizing more practical technology. As deep learning is emerging technology and its techniques have been actively introduced to signal and speech processing, integration of deep learning along with robot audition technology will lead to development of practical robot audition technology that fulfills the four requirements of robot audition. In addition, the following issues to deal with real environments should be considered; adaptive and prompt processing for dynamically changing acoustic environments, learning and recognizing acoustic environments efficiently and effectively in under-resourced situations, structuring acoustic environments in terms of spatial, temporal, and ontological ways, and constructing a framework to deal with higher level information on when, where, what, how, and why.

Acknowledgements

The authors would like to thank many people who are involved in robot audition, in particular, Dr. M. Kumon, Dr. K. Hoshiba, Dr. K. Itoyama, Dr. K. Nishida, T. Yamada, and M. Shakeel for their recent supports for robot audition. The authors also would like to thank T. Konno, Y. Okamoto, and H. Kishinami for their help regarding audio-visual integration. This work was partially supported by JSPS KAKENHI, grant nos. 19K12017, 19KK0260, and 20H00475.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

automatic speech recognition, multimodal integration, open-source softwares, robot audition, sound-source localization, sound source separation

Received: March 23, 2020
Revised: May 29, 2020
Published online: July 8, 2020

[1] K. Nakadai, T. Lourens, H. G. Okuno, H. Kitano, in Proc. of 17th National Conf. on Artificial Intelligence (AAAI-2000), AAAI, Austin, TX 2000, pp. 832–839.
[2] K. Nakadai, D. Matsuura, H. G. Okuno, H. Tsujino, Speech Commun. 2004, 44, 97.
[3] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, H. G. Okuno, IEEE Trans. Robot. 2007, 23, 742.
[4] A. S. Bregman, Auditory Scene Analysis, The MIT Press, Cambridge MA 1990.
[5] L. Jeffress, J. Comp. Physiol. Psychol. 1948, 41, 35.
[6] M. Kumon, Y. Noda, in Proc. of the 2011 IEEE/RSJ Int. Conf. on Robots and Intelligent Systems (IROS), IEEE, San Francisco, CA 2011, pp. 112–117.
[7] A. Portello, P. Danès, S. Argentieri, in Proc. of the 2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE, Tokyo, Japan 2013, pp. 2908–2913.
[8] K. Youssef, S. Argentieri, J.-L. Zarader, in Proc. of the 2013 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE, Tokyo, Japan 2013, pp. 2927–2932.
[9] H.-D. Kim, K. Komatanai, T. Ogata, H. G. Okuno, Adv. Robot. 2007, 23, 629.
[10] U.-H. Kim, H. G. Okuno, Adv. Robot. 2013, 27, 1161.
[11] H. Nakashima, T. Mukai, in Proc. of the 2005 IEEE Int. Conf. on Systems, Man and Cybernetics, IEEE, Waikoloha, HI, USA 2005, pp. 3534–3539.
[12] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, K. Shikino, IEEE Trans. Audio, Speech, Lang. Process. 2006, 14, 666.
[13] I. Haru, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, K. Yamamoto, in Proc. of the 2004 IEEE/RSJ Int. Conf. on Robots and Intelligent Systs (IROS), IEEE, Sendai, Miyagi, Japan 2004, pp. 2404–2410.
[14] M. Sato, A. Sugiyama, S. Ohnaka, IEICE Trans. 2005, 88-A, 2055.
[15] Demo for 11 simultaneous speech recognition, https://www.youtube.com/watch?v=NoiwXa6D3Uc (accessed: May 2020).
[16] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, H. Tsujino, Adv. Robot. 2010, 24, 739.
[17] K. Nakadai, H. G. Okuno, T. Mizumoto, J. Robot. Mechatronics 2017, 29, 16.
[18] HARK, https://www.hark.jp/ (accessed: May 2020).
[19] ROS, https://ros.org (accessed: May 2020).
[20] OpenCV, https://opencv.org (accessed: May 2020).
[21] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, H. G. Okuno, Proc. of the 2011 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE, Hamburg, Germany 2015, pp. 6137–6142.
[22] K. Nakadai, T. Mizumoto, K. Nakamura, in Proc. of the 2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE, Hamburg, Germany 2015, pp. 6129–6136.
[23] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. P. Gabriel, K. Nakadai, H. G. Okuno, Sensors 2017, 17, 1.
[24] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, T. Kawahara, H. G. Okuno, IEEE/ACM Trans. Audio, Speech, Lang. Process. 2018, 26, 215.
[25] D. Gabriel, R. Kojima, K. Hoshiba, K. Itoyama, K. Nishida, K. Nakadai, Adv. Robot. 2019, 33, 403.
[26] S. Matsubayashi, R. Suzuki, F. Saito, T. Murate, T. Masuda, K. Yamamoto, R. Kojima, K. Nakadai, H. G. Okuno, J. Robot. Mechatronics 2017, 29, 224.
[27] R. Suzuki, S. Matsubayashi, F. Saito, T. Murate, T. Masuda, K. Yamamoto, R. Kojima, K. Nakadai, H. G. Okuno, Ecol. Evol. 2017, 8, 812.
[28] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, H. Kitano, in Proc. of the 17th Int. Joint Conf. on Artificial Intelligence (IJCAI 2001), MIT Press, Seattle, Washington, DC 2001, pp. 1424–1432.
[29] T. Yoshida, K. Nakadai, Adv. Robot. 2012, 26, 1183.
[30] K. Okutani, T. Yoshida, K. Nakamura, K. Nakadai, in Proc. of the 2015 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE, Vilamoura, Algarve, Portugal 2012, pp. 3288–3293.
[31] T. Ohata, K. Nakamura, A. Nagamine, T. Mizumoto, T. Ishizaki, R. Kojima, O. Sugiyama, K. Nakadai, J. Robot. Mechatronics 2017, 29, 177.
[33] Disaster Robotics – Results from the ImPACT Tough Robotics Challenge (Eds: K. Nonami, K. Hoshi, K. Nakadai, M. Kumon, H. G. Okuno, Y. Tanabe, K. Yonezawa, H. Tokutake, S. Suzuki, K. Yamaguchi, S. Sunada, T. Takaki, T. Nakata, R. Noda, H. Liu, S. Tadokoro, In S. Tadokoro), Springer Tracts in Advanced Robotics, Vol. 128, Springer, Berlin/New York 2019, pp. 77–142.

[34] T. Suzuki, D. Inoue, Y. Amano, in Proc. of the 2019 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), IEEE, Macau, China 2019, pp. 4402–4408.

[35] M. Wakabayashi, H. G. Okuno, M. Kumon, IEEE Robot. Autom. Lett. 2020, 5, 782.

[36] T. Konno, K. Nishida, K. Itoyama, K. Nakadai, in Proc. of the 2020 IEEE/SICE Int. Symp. on System Integration (SII 2020), IEEE, Hawaii Convention Center, Honolulu, Hawaii 2020, pp. 802–807.

[37] C. H. Knapp, G. C. Carter, IEEE Trans. Acoust., Speech, Signal Process. 1976, 24, 320.

[38] Nakadai Lab, https://www.ra.sc.e.titech.ac.jp/ (accessed: May 2020).

[39] Drone audition demo movie, https://youtu.be/71jEYKAEnGc (accessed: May 2020).