Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments[†]

Kotaro Hoshiba[*]

Abstract

In search and rescue activities, unmanned aerial vehicles (UAV) should exploit sound information to compensate for poor visual information. This paper describes the design and implementation of a UAV-embedded microphone array system for sound source localization in outdoor environments. Four critical development problems included water-resistance of the microphone array, efficiency in assembling, reliability of wireless communication, and sufficiency of visualization tools for operators. To solve these problems, we developed a spherical microphone array system (SMAS) consisting of a microphone array, a stable wireless network communication system, and intuitive visualization tools. The performance of SMAS was evaluated with simulated data and a demonstration in the field. Results confirmed that the SMAS provides highly accurate localization, water resistance, prompt assembly, stable wireless communication, and intuitive information for observers and operators.

1 Introduction

Research on remote sensing techniques involving unmanned aerial vehicles (UAV) is important to improve search and rescue in disaster-stricken areas because such technologies enable prompt action regardless of the terrain. Search and rescue tasks with UAV rely mainly on vision, which is vulnerable to poor lighting conditions or occlusions. A UAV-embedded microphone array system is expected to be effective for the detection of people needing assistance in disaster-stricken areas. Since a UAV-embedded microphone array system receives rotor and wind noise as well as environmental sounds, the target sound is contaminated by ego-noise and other noise. Sound source processing should be able to localize, and discriminate a target sound from noise. Robot audition software [1–3] has been developed to cope with a mixture of sounds contaminated by noise. In particular, the open source robot audition software HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [4, 5] provides noise-robust sound processing functions: sound source localization, source separation and recognition of separated sounds. Basic technologies of robot audition have been developed for use in indoor environments, and it is necessary to advance these technologies for use in outdoor environments, such as for search and rescue tasks using UAV. Five main challenges to developing such a system for UAV include:

1. sound source localization;
2. sound source separation and sound enhancement;
3. sound source classification;
4. real-time processing and intuitive visualization tools;
5. robustness of the device in outdoor environments.

The first challenge has been addressed in recent years in several studies including as a main research topic to find people in disaster situations, e.g., localization of an emergency signal from a safety whistle [6], and that of speech with a low signal-to-noise ratio (SNR) [7–9]. To locate the source of a sound, algorithms based on multiple signal classification (MUSIC) [10] are often used because they can effectively localize sound sources in highly noisy environments. In particular, MUSIC based on incremental generalized singular value decomposition with correlation matrix scaling (iGSVD-MUSIC-CMS) developed by Ohata et al. demonstrated good performance under dynamically changing noise [9]. iGSVD-MUSIC-CMS could localize sound sources in a low SNR environment, −15 dB. There is a severe trade-off between the speed and performance of the signal processing. Since only offline processing was reported in their evaluation, evaluation in real time is necessary for application to search and rescue tasks.

The second challenge, in order to identify a target sound source in extremely noisy environments, is also important, in two goals: to improve SNR of the target sound and to improve intelligibility of the separated signals. The first goal of the present study was to improve sound source classification (the third challenge). Recent studies on restoration of distorted signals have been reported including: sound source separation with a linear process [11] and integrated frameworks of sound source separation and classification using end-to-end training [12, 13]. While the first goal targeted machine listening, the second goal targets human listening, that is, an operator tries to identify a target sound source manually, e.g., by inspecting sound spectrograms or by listening to separated sounds. In this case, intelligibility is a primary requirement. This second goal has not been reported as a function for UAV, although it is important to the UAV operator.

The third challenge, to effectively discriminate a target sound source, such as a human-induced sound, from other sound sources has been inves-

[*] 助教 電気電子情報工学科

Assistant Professor, Dept. of Electrical, Electronics and Information Engineering

[†]This is the copy of Design of UAV-Embedded Microphone Array System for Sound Source Localization in Outdoor Environments, K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, H. G. Okuno, *Sensors*, Vol. 17, No. 11, pp. 1–16, 3 Nov. 2017.
http://www.mdpi.com/1424-8220/17/11/2535

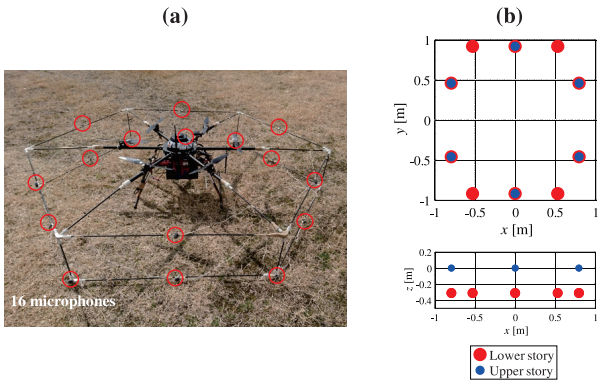

Figure 1 HMAS (hexagonal microphone array system). (**a**) the 16 microphones marked as red circles; (**b**) coordinates of the microphone positions in the HMAS.

tigated [11–13], albeit these studies only reported offline processing and did not mention real-time processing.

Regarding the fourth challenge, latency in visualizing flight and sound source information should be as short as possible for efficient operation of UAV. Since UAV operators may be situated far from the UAV, visualization tools should be capable of displaying accurate information regarding UAV location and sound source. Finally, regarding the fifth challenge, to ensure system efficiency, all-weather acoustic sensors and reliability of wireless communication using a Wi-Fi signal that carries acoustic signals, which are necessary in outdoor environments, should be proven.

In this paper, we report the development of a UAV-embedded microphone array system that resolved four of the above five challenges. The third challenge, sound classification, which we regard to be at a higher level than the other four, will be investigated in a separate study. The remaining of the paper is organized as follows: Section 2 describes the design method and details of the UAV-embedded microphone array system. Section 3 evaluates and discusses the performance of the system. Section 4 is the conclusion.

2 Methods

2.1 Design of Water-Resistant Microphone Array for Use Onboard UAV

To address the first and fifth challenges in Section 1, we designed and developed a microphone array.

Figure 1 shows our prototype hexagonal microphone array system (HMAS). Sixteen MEMS (Micro Electro Mechanical Systems) microphones are set on a two-story hexagonal frame whose diagonal length is 1.8 m. The microphones and cables, being exposed, were vulnerable to water, and risk of disconnection. Additionally, the complexity of the frame demanded a lot of time for assembly of the HMAS. To solve these problems, we designed a spherical microphone array system (SMAS), which is water resistant and simple to assemble in a UAV (Figure 2). As shown in Figure 2a or Figure 2c, twelve MEMS microphones are embedded in a spherical body with a diameter of 0.1 m. Since a single

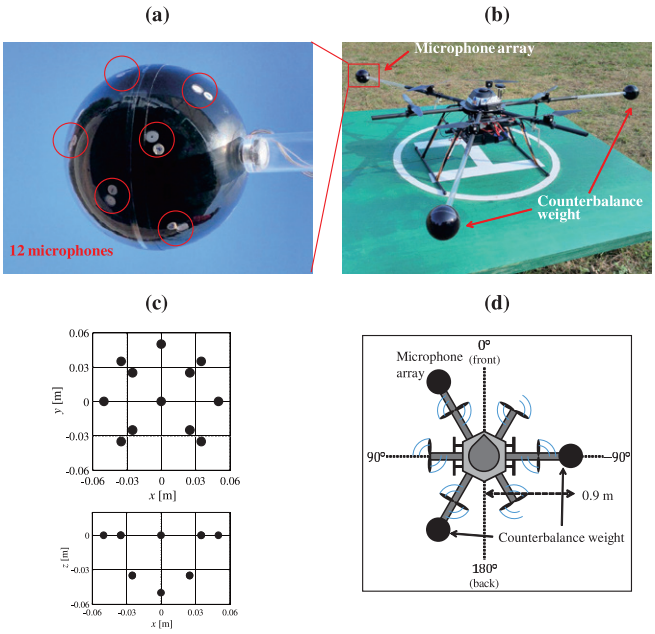

Figure 2 SMAS (spherical microphone array system). (**a**) the 12 microphones, and six of them marked as red circles; (**b**) UAV (unmanned aerial vehicles) with SMAS and two counterbalance weights; (**c**) coordinates of the microphone positions in the SMAS; (**d**) layout of the SMAS and two counterbalance weights in the UAV.

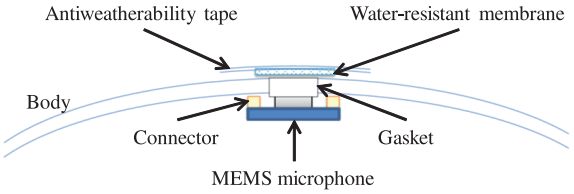

Figure 3 Internal structure of the SMAS.

strut and one cable connects the array to the UAV, its assembly is simple and the risk of disconnection is reduced. Unlike with the HMAS, where the microphones are equidistant around the UAV, the weight of the UAV is unbalanced with the SMAS. To solve this, we added two weights, each of the same size and mass, to counterbalance the SMAS. As shown in Figure 2d, the SMAS and weights are set at intervals of 120°, and the direction of the SMAS is 30° on the UAV coordinates. Figure 3 shows internal structure of the SMAS. To embed microphones into the body, gaskets were used. The microphone was attached to the gasket so that holes of the microphone and the gasket were coincident. When there is a gap between the microphone and the body, the microphone cannot receive acoustic signals precisely because of reverberations in the body. To fill gaps between microphones and the body, ring-shaped connectors were used. To ensure water resistance of the SMAS, holes of gaskets were covered with a water-resistant membrane and an antiweatherability tape. Since a water-resistant membrane and an antiweatherability tape are enough thin to pass acoustic signals through, they do not influence signals received by microphones.

Figure 4 (**a**) antennas on the UAV marked as red circles; (**b**) the Yagi antenna at a ground station.

2.2 Stabilization of Wireless Communication

To resolve the fifth challenge, to stabilize wireless communication, we incorporated a high-gain antenna that could receive a Wi-Fi signal from the UAV, which carries acoustic signals at a ground station. In addition, a communication protocol was also implemented to improve robustness.

For sound source localization, acoustic signals recorded by SMAS on the UAV are sent via wireless communication using a Wi-Fi signal to a ground station, and processed by a computer. A network system was constructed by assuming that the distance between the UAV and the ground station is short. However, in an outdoor environment, the network communication has the potential to be unstable as the distance increases. To ensure reliable wireless communication, two improvements were made.

First, we replaced an antenna at the ground station with the Yagi antenna (FX-ANT-A5, CONTEC (Osaka, Japan)) [14] to improve throughput of communication [15]. For acoustic signals recorded by 12 microphones, throughput of approximately 5 Mbps is necessary. Therefore, a high gain antenna was used for reliable wireless communication in outdoor environments. Figure 4 shows antennas (FX-ANT-A7, CONTEC) [16] on the UAV and the Yagi antenna at a ground station. On the UAV, two antennas was assembled to arms of UAV. At a ground station, the Yagi antenna was set on a tripod.

Second, we changed the communication protocol from TCP (Transmission Control Protocol) to UDP (User Datagram Protocol). In wireless communication tests at tens of meters of distance between the hovering UAV and the ground station, packet loss occurred approximately 400 times per minute. With TCP, each packet loss caused a retransmission request, greatly reducing acoustic signal throughput. Hence, the protocol was changed to UDP. Because UDP provides no guarantee of data integrity, no retransmission request is sent and throughput is maintained. When a packet loss occurs, the ground station receives defective acoustic signals. However, because the minimum frame size for sound source localization is much larger than one packet, the impact is negligible.

2.3 Development of Intuitive Visualization Tools for Operators

For the second and fourth challenges, we developed three visualization tools that display information regarding the sound source on the UAV coordinates and acoustic signals before and after their enhancement.

Essential to the system is a visualization tool to display sound source

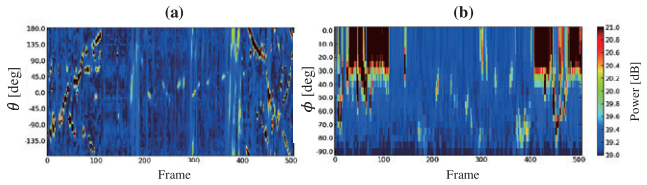

Figure 5 MUSIC (multiple signal classification) spectrum. (**a**) azimuth direction; (**b**) elevation direction.

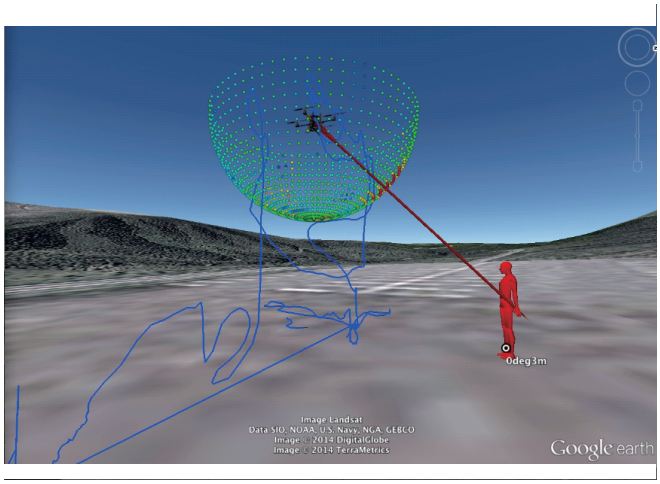

Figure 6 Previous visualization tool based on Google Earth™.

localization results. Several groups have developed such a tool for sound source localization [17–19]. Figure 5 shows MUSIC spectra produced by the MUSIC method [10]. It visualizes sound power arriving from each direction. The horizontal axis represents the frame number (time) and the vertical axis represents the azimuth angle θ (**a**) or the elevation angle ϕ (**b**). The sound power is represented as a color map. It is difficult to quickly determine the direction and time of a sound source with the MUSIC spectrum. In order to visualize sound source localization and the UAV location and orientation, we developed a tool to display such data on Google Earth™ (Google (CA, USA)) (Figure 6) [20]. Because users can change their viewpoint freely on Google Earth™, they can intuitively grasp the situation of the environment, the UAV, and the sound source. However, this tool is for observers only and not for the operator. Unlike an indoor environment, in an outdoor environment, the distance between the UAV and the operator may be large, necessitating a tool for its effective operation. Therefore, we developed visualization tools for operators.

Because an operator controls the UAV with reference to its coordinates, a user friendly method would be to display sound source directions on the UAV' s coordinates. Therefore, the coordinate system shown in Figure 7 was defined. Forward direction of the UAV is defined as the positive direction of the y-axis, and the azimuth and elevation are projected to the circumferential and radial directions, respectively. Using this coordinate system, two visualization tools to display sound source directions were developed as shown in Figure 8. Figure 8a shows the MUSIC spectrum. The sound power in each direction is depicted by a color map. Figure 8b illustrates only the sound directions after threshold processing for the

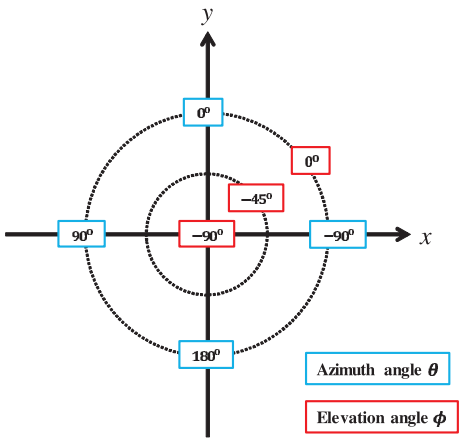

Figure 7 Visualization tool coordinate system.

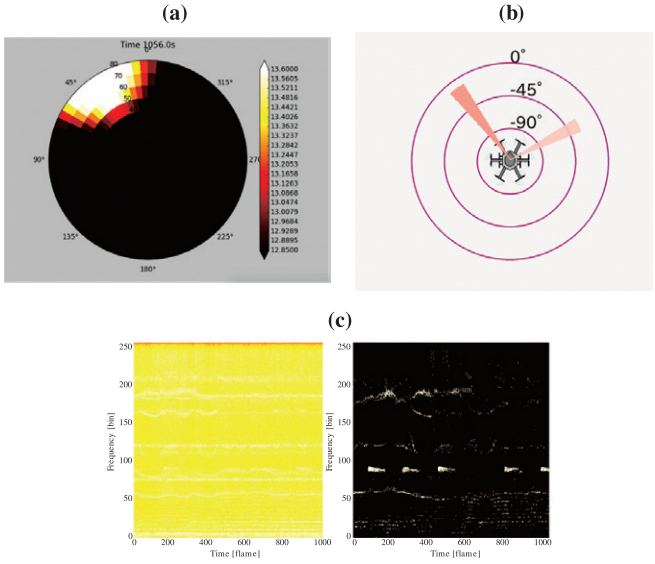

Figure 8 Visualization tools. (**a**) MUSIC spectrum; (**b**) sound direction; (**c**) spectrograms of captured sound (**left**) and after enhancement (**right**).

MUSIC spectrum. In addition, spectrograms of the recorded sound and after sound enhancement by online robust principal component analysis (ORPCA) [21] are displayed as in Figure 8c. The left panel shows a spectrogram of the recorded acoustic signal, and the right panel shows it after enhancement. The horizontal and vertical axes represent time and frequency, respectively. By viewing these three sets of data in real time, even when located far from the UAV, the operator knows the relationship between the UAV and the sound source.

2.4 Sound Source Localization Method

We used two methods of sound source localization, namely SEVD-MUSIC (MUSIC based on Standard Eigen Value Decomposition), which is an original broadband MUSIC method [10], and iGSVD-MUSIC [9]. SEVD-MUSIC has low noise robustness and low computational cost, while iGSVD-MUSIC has high noise robustness and high computational cost. Either of these can be selected according to the circumstances. Algorithms of SEVD-MUSIC and iGSVD-MUSIC are described below.

2.4.1 SEVD-MUSIC

M channel input sound signals of the f-th frame are Fourier transformed to $\boldsymbol{Z}(\omega, f)$, from which a correlation matrix $\boldsymbol{R}(\omega, f)$ is defined as follows:

$$\boldsymbol{R}(\omega, f) = \frac{1}{T_R} \sum_{\tau=f}^{f+T_R-1} \boldsymbol{Z}(\omega, \tau)\boldsymbol{Z}^*(\omega, \tau). \tag{1}$$

ω is the frequency bin index, T_R is the number of frames used for the correlation matrix calculation, and $\boldsymbol{Z}^*$ is a complex conjugate transpose of $\boldsymbol{Z}$. The SEVD-MUSIC method calculates eigenvectors through an SEVD of the obtained $\boldsymbol{R}(\omega, f)$:

$$\boldsymbol{R}(\omega, f) = \boldsymbol{E}(\omega, f)\boldsymbol{\Lambda}(\omega, f)\boldsymbol{E}^*(\omega, f). \tag{2}$$

$\boldsymbol{\Lambda}(\omega, f)$ is a matrix with diagonal components that are eigenvalues in a descending order. $\boldsymbol{E}(\omega, f)$ is a matrix containing eigenvectors corresponding to $\boldsymbol{\Lambda}(\omega, f)$. Using $\boldsymbol{E}$, and a transfer function, $\boldsymbol{G}(\omega, \psi)$, corresponding to the sound source direction, $\psi = (\theta, \phi)$ in the UAV coordinates, the MUSIC spatial spectrum, $P(\omega, \psi, f)$, is calculated:

$$P(\omega, \psi, f) = \frac{|G^*(\omega, \psi)G(\omega, \psi)|}{\sum_{m=L+1}^{M} |G^*(\omega, \psi)e_m(\omega, \psi)|}. \tag{3}$$

L is the number of target sound sources, and e_m is the m-th eigenvector contained in $\boldsymbol{E}$. $P(\omega, \psi, f)$ is average over ω direction to estimate the direction of the sound source:

$$\bar{P}(\psi, f) = \frac{1}{\omega_H - \omega_L + 1} \sum_{\omega=\omega_L}^{\omega_H} P(\omega, \psi, f). \tag{4}$$

ω_H and ω_L are indices corresponding to the upper and lower limits of the used frequency bin, respectively. Threshold processing and peak detection is performed for $\bar{P}(\psi, f)$ and ψ of the obtained peak is detected as the sound source direction.

2.4.2 iGSVD-MUSIC

In iGSVD-MUSIC, for the f-th frame, the section of the length of T_N frames from the $f - f_s$-th frame is assumed to be a noise section, and the noise correlation matrix $\boldsymbol{K}(\omega, f)$ is calculated:

$$\boldsymbol{K}(\omega, f) = \frac{1}{T_N} \sum_{\tau=f-f_s-T_N}^{f+f_s} \boldsymbol{Z}(\omega, \tau)\boldsymbol{Z}^*(\omega, \tau). \tag{5}$$

The iGSVD-MUSIC method estimates noise in each frame and responds to dynamic change in noise. The noise component can be whitened by multiplying $\boldsymbol{K}^{-1}$ to $\boldsymbol{R}$ from the left. The iGSVD-MUSIC method calculates singular vectors through the GSVD of $\boldsymbol{K}^{-1}(\omega, f)\boldsymbol{R}(\omega, f)$:

$$\boldsymbol{K}^{-1}(\omega, f)\boldsymbol{R}(\omega, f) = \boldsymbol{Y}_l(\omega, f)\boldsymbol{\Sigma}(\omega, f)\boldsymbol{Y}_r^*(\omega, f). \tag{6}$$

$\boldsymbol{\Sigma}(\omega, f)$ is a matrix with diagonal components of singular values in a descending order. $\boldsymbol{Y}_l(\omega, f)$ and $\boldsymbol{Y}_r(\omega, f)$ are matrices containing singular vectors corresponding to $\boldsymbol{\Sigma}(\omega, f)$. Then, the MUSIC space spectrum is

calculated:

$$P(\omega,\psi,f)=\frac{|G^{*}(\omega,\psi)G(\omega,\psi)|}{\sum_{m=L+1}^{M}|G^{*}(\omega,\psi)y_{m}(\omega,\psi)|}. \quad (7)$$

y_m is the m-th singular vector contained in Y_l. $P(\omega,\psi,f)$ is averaged over ω direction to estimate the direction of the sound source:

$$\bar{P}(\psi,f)=\frac{1}{\omega_H-\omega_L+1}\sum_{\omega=\omega_L}^{\omega_H}P(\omega,\psi,f). \quad (8)$$

Threshold processing and peak detection is performed for $\bar{P}(\psi,f)$ and ψ of the obtained peak is detected as the sound source direction.

Both sound source localization methods based on MUSIC basically assume an acoustic far-field. However, by using the transfer function G according to the distance to sound sources, it is possible to localize sound sources at any distance. In addition, at the altitude at which a UAV flies normally (at least a few meters), an acoustic field is a far-field. Therefore, the accuracy of sound source localization depends on a SNR of an acoustic signal rather than a distance between a microphone array to a sound source.

2.5 Structure of Microphone Array System

By integrating the above components, the microphone array system was constructed. Figure 9 shows the SMAS configuration. The microphone array on the UAV was connected to a multi-channel sound signal recorder, RASP-ZX (System In Frontier (Tokyo, Japan)) [22] for synchronous recording of 12 ch sound signals. The sound signals were recorded at a sampling frequency of 16 kHz, and a quantization bit rate of 24 bits. Recorded acoustic signals, images from the wireless camera and data from a GNSS/IMU (Global Navigation Satellite System/Inertial Measurement Unit) sensor were transmitted through a wireless network to the ground station. Different frequencies were used for the wireless communications to prevent cross talk. In the SMAS, data from a GNSS/IMU sensor and images from the wireless camera were not used; therefore, only recorded acoustic signals were received by the Yagi antenna. The received data was integrated using ROS (Robot Operating System) to provide general versatility. The acoustic signals were processed by a PC using a sound source localization method. HARK was used for the algorithm implementation. The data after processing was shared by three PCs via a router. To reduce the processing load of one computer for real-time visualization, visualization tools were displayed using three laptops. PC1, PC2 and PC3 displayed the MUSIC spectrum (Figure 8a), sound direction (Figure 8b) and enhanced sound (Figure 8c), respectively. Since the SMAS is a separate system from the UAV, including its power supply, it can be applied to various UAVs.

3 Results and Discussion

The performance of the SMAS was evaluated using numerical sound simulation and by demonstration in an outdoor environment.

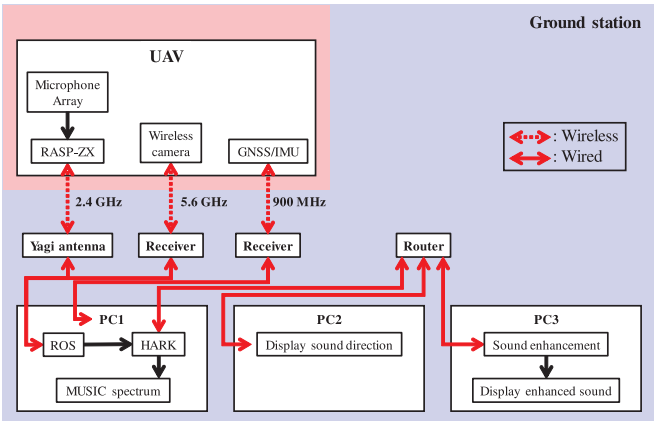

Figure 9 Configuration of SMAS.

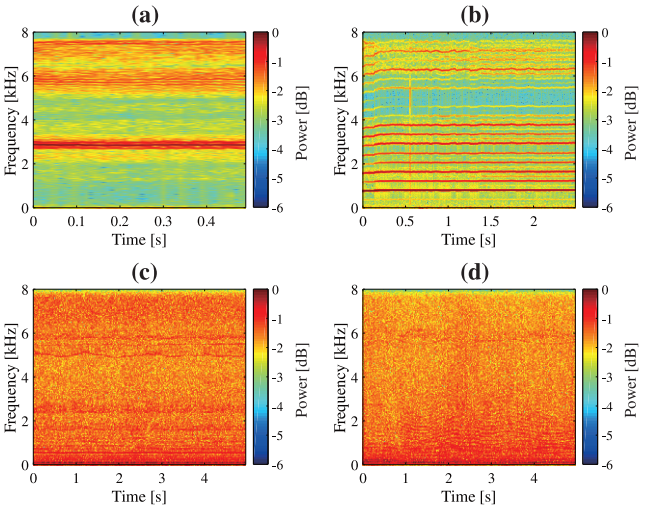

Figure 10 Spectrograms. (**a**) whistle; (**b**) voice; (**c**) noise of UAV recorded by hexagonal microphone array; (**d**) noise of UAV recorded by spherical microphone array.

3.1 Evaluation Procedure

Sound localization performance was evaluated using acoustic signals created in a numerical simulation. Using transfer functions corresponding to two types (hexagonal and spherical) of microphone array and sound samples, acoustic signals arriving from every direction were created. Recorded noise of an actual flying UAV was added to the created signals. The direction was set as every 5° in the azimuth range from −180° to 180° and the elevation range from −90° to 0°. As sound sources, a whistle and human voice were used. A Mini Surveyor MS-06LA (Autonomous Control Systems Laboratory (Chiba, Japan)) was used as the UAV. Spectrograms of the sound sources and the noise of the UAV recorded by each of the two microphone arrays are shown in Figure 10. Simulated signals were created in different SNR, −20, −10, 0, 10 and 20 dB. Simulated signals were processed by the SMAS and results were evaluated. Performance was also evaluated by demonstration in the field.

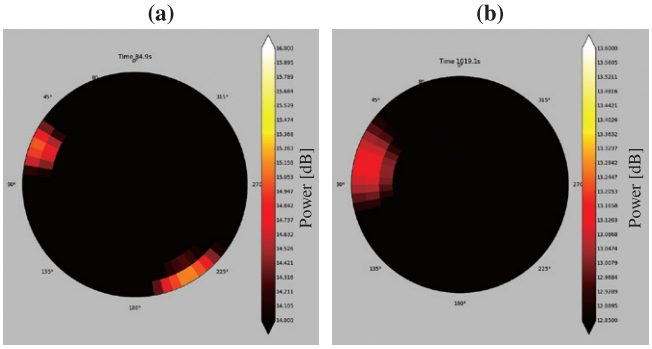

Figure 11 MUSIC spectra. (**a**) SEVD-MUSIC (MUSIC based on Standard Eigen Value Decomposition); (**b**) iGSVD-MUSIC (MUSIC based on incremental generalized singular value decomposition).

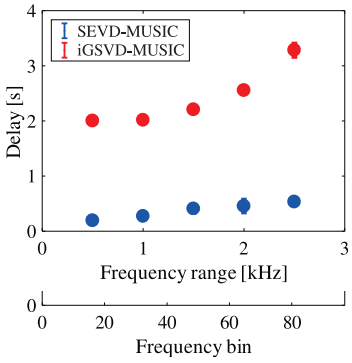

Figure 12 Time delay of the system.

3.2 Results of Simulation

The main differences between SEVD-MUSIC and iGSVD-MUSIC are noise robustness and computational cost. Since its computational cost is low, SEVD-MUSIC has a short delay; however, it has poor noise tolerance. Since iGSVD-MUSIC includes noise whitening, it has noise tolerance but a long delay. Thus, real-time property and noise tolerance are in a trade-off relationship. Figure 11 shows MUSIC spectra processed by SEVD-MUSIC (**a**) and by iGSVD-MUSIC (**b**) using spherical microphone array. The target sound is located around $\theta = 80°$. In both MUSIC spectra, the target sound source power can be seen. However, in Figure 11a, the noise power of the UAV can also be seen. Figure 12 shows the delay in the system when the frequency range, which is used in the MUSIC method, is changed. The horizontal axis represents the frequency range and the number of the frequency bin ($\omega_H - \omega_L + 1$), and the vertical axis represents delay. As shown in Figure 12, iGSVD-MUSIC has a time delay of 2 to 3 seconds longer than that of SEVD-MUSIC. In addition, as the frequency range increases, the time delay increases. Based on these results, localization performance was evaluated by its success rate. The success rate was calculated based on the UAV coordinates. When the angle of the maximum value of the MUSIC spectrum is matched with a set angle, it is defined that sound source localization succeeded. All simulated sounds were processed using sound source localization method, and the success

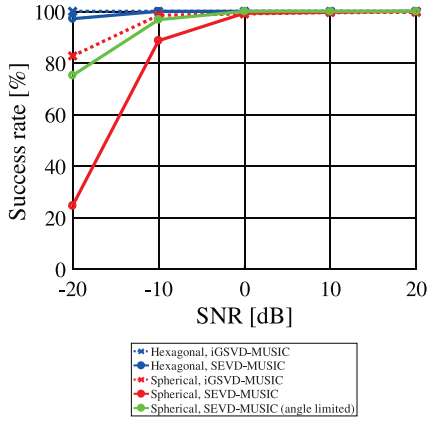

Figure 13 Success rate of localization.

rate was calculated. Figure 13 shows the success rate of localization for the hexagonal and a spherical microphone arrays, processed by SEVD-MUSIC and iGSVD-MUSIC. The frequency range used in the MUSIC method was from 500 to 3000 Hz. In the hexagonal array, the success rate of both MUSIC methods was almost 100% even with SNR less than 0 dB. In the spherical array, the success rate was lower than that of hexagonal. In particular, the success rate of SEVD-MUSIC was less than 30% when the SNR was −20 dB. This lower success rate was considered due to the smaller aperture diameter in the spherical array at 0.1 m compared to 1.8 m in the hexagonal. Therefore, the detection area of the MUSIC spectrum was limited to increase the accuracy of localization with SEVD-MUSIC. As shown in Figure 11a, noise of the UAV appear in one direction constantly as directional noise, in this case at the azimuth angle of around −150°. To avoid the effect of such noise, the detection azimuth angle was limited to $-60° \leq \theta \leq 120°$. The success rate in a case when the detection angle was limited is plotted as the green line in Figure 13. By limiting the detection angle, the success rate of short-delay SEVD-MUSIC using the spherical microphone array with SNR -10 dB could be increased to approximately 97%. Since the SNR, when blowing a whistle or speaking to an actual flying UAV from a distance of around 10 m was approximately −10 to 0 dB, it was considered sufficient localization performance. This technique can be used with microphones located at one site on the UAV, unlike the HMAS in which microphones are dispersed around the UAV. Due to the location of our microphone array, parameters for sound source localization could be easily tuned to attain accurate localization with a small latency.

3.3 Results of the Demonstration

Regarding efficiency in assembling the system, the HMAS took two hours to assemble, and especially time consuming was assembly of the frame and electric cables. In contrast, the SMAS took 40 min to assemble and 2 min to take off after switching on the UAV. Regarding water resistance, although the demonstration was performed in light rain, the SMAS worked without failure. To assess the reliability of wireless communication, throughputs were compared among four different antennas. Figure

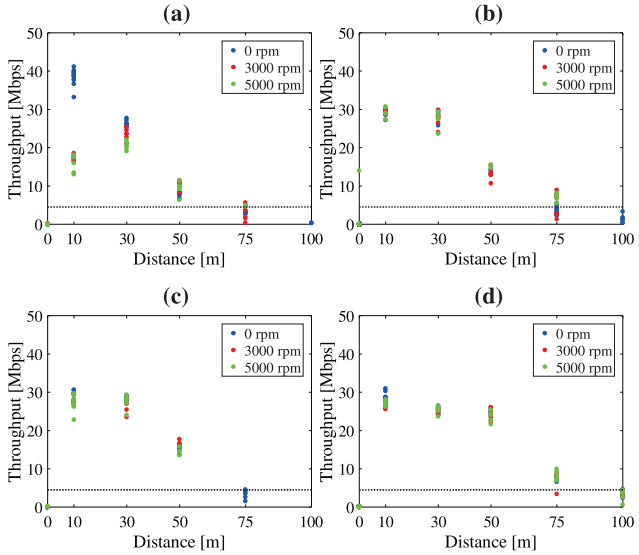

Figure 14 Throughputs by antennas type. (**a**) diversity; (**b**) Yagi (small); (**c**) collinear; (**d**) Yagi (large).

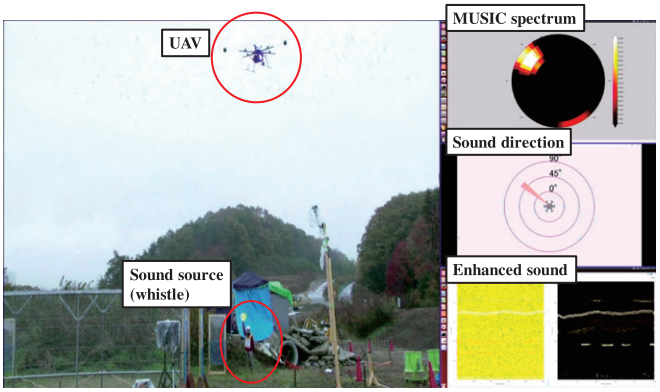

Figure 15 Data visualized in the demonstration.

14 shows the results of throughputs by antenna type: Diversity (FX-ANT-A1, CONTEC) [23], small Yagi (FX-ANT-A3, CONTEC) [24], Collinear (FX-ANT-A2, CONTEC) [25] and large Yagi (used in SMAS). Throughputs were measured by fixing the UAV on the ground at distances, 10, 30, 50, 75 and 100 m, with propeller rotation speeds, 0, 3000 and 5000 rpm. The required throughput (5 Mbps) is shown as a dotted line in Figure 14. It was found that throughput surpassed 5 Mbps even at 75 m by using the large Yagi-antenna. In the demonstration, the wireless network worked without disconnecting in the distance of tens of meters. To examine the intuitiveness of visualization tools, camera image, MUSIC spectrum, sound direction, and enhanced sound data were displayed as in Figure 15. These visualization tools provided directions of sound sources and other data in real time for the audience and operator, intuitively.

3.4 Discussion

Before the demonstration, we conducted over 10 test flights, and all sound source localization trials were successfully completed. Thus, usability of the SMAS was verified. Tables 1 and 2 show a summary of pros and cons of each microphone array system and sound source localization method. For the microphone array system, the HMAS provides high accurate localization; however, it does not have water resistance and efficiency in assembling. The SMAS provides lower accurate localization than the HMAS; however, we can increase the accuracy of localization depending on sound source localization method. For sound source localization method, SEVD-MUSIC has low noise tolerance and a small latency, while iGSVD-MUSIC has high noise tolerance and a large latency. Angle-limited SEVD-MUSIC can have high noise tolerance only when microphones located at one site on the UAV like the SMAS. Thus, because of their characteristics, we can select them according to the situation. In the demonstration, sound sources could be localized in real time with high accuracy using the SMAS and angle-limited SEVD-MUSIC because the SNR of the recorded acoustic signal was over -10 dB. However, in order to develop the system for the detection of people in a disaster-stricken area, a new sound source localization method with higher noise robustness and lower computational cost is needed. In addition, since there are several sound sources at an actual site, it is necessary to separate and identify human-related sound from recorded sounds. In future work, we will integrate the proposed sound source identification method using deep-learning [11–13] to the SMAS.

Table 1 Pros and cons of the HMAS and the SMAS.

	Accuracy of Localization	**Water Resistance**	**Efficiency in Assembling**
HMAS	○	×	×
SMAS	△	○	○

Table 2 Pros and cons of SEVD-MUSIC, iGSVD-MUSIC and angle-limited SEVD-MUSIC.

	Noise Tolerance	**Latency**
SEVD-MUSIC	×	○
iGSVD-MUSIC	○	×
Angle-limited SEVD-MUSIC	△	○

4 Conclusions

In this paper, we developed a UAV-embedded microphone array system for an outdoor environment. First, a novel microphone array was designed to ensure water resistance and efficiency of assembly. A 12 ch microphone array, including a spherical body of simple structure, was designed. By using coated microphones and a simple structure, water resistance and efficiency of assembly were ensured. Second, the antenna and communication protocol were changed to obtain reliable wireless communication. To improve throughput, the antenna at the ground station was changed to the Yagi antenna. To avoid reducing throughput, the communication protocol was changed from TCP to UDP. Third, intuitive visualization tools for a UAV operator were developed. By integrating the above im-

provements, the microphone array system was constructed. Tests showed that our microphone array system for an outdoor environment that is independent from the UAV provides highly accurate sound source localization performance in real time, and has effective intuitive operator visualization tools.

Acknowledgements The authors would like to thank Kai Washizaki, Mizuho Wakabayashi, Takahiro Ishiki, Makoto Kumon, Yoshiaki Bando, Daniel Gabriel, Kazuhiro Nakadai, Hiroshi G. Okuno (the authors of the original article) and the members of System In Frontier Inc. for their support. This work was supported by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant Nos.16H02884, 16K00294, and 17K00365, and also by the ImPACT (Impulsing Paradigm Change through Disruptive Technologies Program) of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).

References

[1] Nakadai, K.; Lourens, T.; Okuno, H.G.; Kitano, H. Active audition for humanoid. In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin, TX, USA, 30 July–3 August 2000; pp. 832–839.

[2] Okuno, H.G.; Nakadai, K. Robot audition: Its rise and perspectives. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), Brisbane, QLD, Australia, 19–24 April 2015; pp. 5610–5614.

[3] Okuno, H.G.; Nakadai, K. Special Issue on Robot Audition Technologies. *J. Robot. Mech.* **2017**, *29*, 15–267, doi:10.20965/jrm.2017.p0015.

[4] Nakadai, K.; Takahashi, T.; Okuno, H.G.; Nakajima, H.; Hasegawa, Y.; Tsujino, H. Design and Implementation of Robot Audition System 'HARK'—Open Source Software for Listening to Three Simultaneous Speakers. *Adv. Robot.* **2010**, *24*, 739–761, doi:10.1163/016918610X493561.

[5] http://www.hark.jp/

[6] Basiri, M.; Schill, F.; Lima, P. U.; Floreano, D. Robust acoustic source localization of emergency signals from Micro Air Vehicles. In Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), Vilamoura, Portugal, 7–12 October 2012; pp. 4737–4742.

[7] Okutani, K.; Yoshida, T.; Nakamura, K.; Nakadai, K. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), Vilamoura, Portugal, 7–12 October 2012; pp. 3288–3293.

[8] Furukawa, K.; Okutani, K.; Nagira, K.; Otsuka, T.; Itoyama, K.; Nakadai, K.; Okuno, H.G. Noise correlation matrix estimation for improving sound source localization by multirotor UAV. In Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), Tokyo, Japan, 3–8 November 2013; pp. 3943–3948.

[9] Ohata, T.; Nakamura, K.; Mizumoto, T.; Tezuka, T.; Nakadai, K. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), Chicago, IL, USA, 14–18 September 2014; pp. 1902–1907.

[10] Schmidt, R.O. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280, doi:10.1109/TAP.1986.1143830.

[11] Sugiyama, O.; Uemura, S.; Nagamine, A.; Kojima, R.; Nakamura, K.; Nakadai, K. Outdoor Acoustic Event Identification with DNN Using a Quadrotor-Embedded Microphone Array. *J. Robot. Mech.* **2017**, *29*, 188–197, doi:10.20965/jrm.2017.p0188.

[12] Morito, T.; Sugiyama, O.; Kojima, R.; Nakadai, K. Reduction of Computational Cost Using Two-Stage Deep Neural Network for Training for Denoising and Sound Source Identification. In Proceedings of the IEA/AIE 2016 Trends in Applied Knowledge-Based Systems and Data Science Volume 9799 of the Series Lecture Notes in Computer Science, Morioka, Japan, 2–4 August 2016; pp. 562–573.

[13] Morito, T.; Sugiyama, O.; Kojima, R.; Nakadai, K. Partially Shared Deep Neural Network in Sound Source Separation and Identification Using a UAV-Embedded Microphone Array. In Proceedings of the IEEE/RSJ International Conference on Robots and Intelligent Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 1299–1304.

[14] https://www.contec.com/products-services/computer-networking/flexlan-fx/fx-accessories/fx-ant-a5/

[15] Ishiki, T.; Kumon, M. Continuous transfer of sensor data from multi-rotor helicopter. In Proceedings of the 33-th Annual Conference of the RSJ, Tokyo, Japan, 3–5 September 2015; RSJ2015AC1L3-03. (In Japanese)

[16] https://www.contec.com/products-services/computer-networking/flexlan-fx/fx-accessories/fx-ant-a7/

[17] Sasaki, Y.; Masunaga, S.; Thompson, S.; Kagami, S.; Mizoguchi, H. Sound Localization and Separation for Mobile Robot Tele-Operation by Tri-Concentric Microphone Array. *J. Robot. Mech.* **2007**, *19*, 281–289, doi:10.20965/jrm.2007.p0281.

[18] Kubota, Y.; Yoshida, M.; Komatani, K.; Ogata, T.; Okuno, H.G. Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking. In Proceedings of the Tenth IEEE International Symposium on Multimedia (ISM), Berkeley, CA, USA, 15–17 December 2008; pp. 468–476.

[19] Mizumoto, T.; Nakadai, K.; Yoshida, T.; Takeda, R.; Otsuka, T.; Takahashi, T.; Okuno, H.G. Design and Implementation of Selectable Sound Separation on the Texai Telepresence System using HARK. In Proceedings of the IEEE International Conference on Robots and Automation (ICRA), Shanghai, China, 9–13 May 2011; pp. 2130–2137.

[20] Hoshiba, K.; Sugiyama, O.; Nagamine, A.; Kojima, R.; Kumon, M.; Nakadai, K. Design and assessment of sound source localization system with a UAV-embedded microphone array. *J. Robot. Mech.* **2017**, *29*, 154–167, doi:10.20965/jrm.2017.p0154.

[21] Feng, J.; Xu, H.; Yan, S. Online robust PCA via stochastic optimization. In Proceedings of the Neural Information Processing Systems Conference (NIPS), Stateline, NV, USA, 5–10 December 2013; pp. 404–412.

[22] http://www.sifi.co.jp/system/modules/pico/index.php?content_id=36&ml_lang=en

[23] https://www.contec.com/products-services/computer-networking/flexlan-fx/fx-accessories/fx-ant-a1/

[24] https://www.contec.com/products-services/computer-networking/flexlan-fx/fx-accessories/fx-ant-a3/

[25] https://www.contec.com/products-services/computer-networking/flexlan-fx/fx-accessories/fx-ant-a2/