To appear in *Advanced Robotics*
Vol. 00, No. 00, Month 201X, 1–14

FULL PAPER for Review

Drone Audition Listening from the Sky Estimates Multiple Sound Source Positions by Integrating Sound Source Localization and Data Association

Mizuho Wakabayashi[a], Hiroshi G. Okunoi[b†], and Kumon Makoto[a‡*]

[a] *Graduate School of Science and Engineering, Kumamoto University, Kumamoto 860-8555, Japan*;
[b] *Graduate Program for Embodiment Informatics, Waseda University, Tokyo 169-0072, Japan*

(*Submitted on July 2019*)

A multi-rotor helicopter (hereinafter, *drone*) with sensors for scene analysis is expected to improve real-world tasks including search and rescue tasks. In addition to visual information, acoustic information obtained by sound source localization, position estimation, and separation is critical for conducting such urgent tasks in order to compensate for the weakness of visual sensors due to darkness and occlusion. This paper focuses on the estimation of sound source positions from acoustic signals captured by a drone equipped with a microphone array. Due to noise generated by drone rotors, the estimation of the sound source position is obscured and prone to error. In addition, a drone is deployed to capture multiple sound sources, erroneous sound sources localization deteriorates the performance of such estimation. We cope with this problem by introducing data association for disambiguation to estimate multiple sound source positions. Since drone audition can be used for search and rescue tasks, real-time processing is critical. This paper presents the details of drone audition and demonstrates that the developed system can estimate multiple sound source positions with the accuracy of about 3 m.

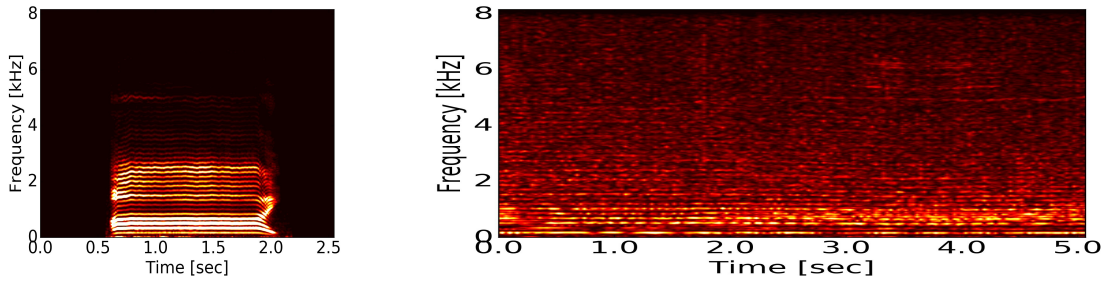
Keywords: drone audition; robot audition; sound source localization; sound source position estimation; computational auditory scene analysis; data association

1. Introduction

Recently, drones, or unmanned aerial vehicles or multi-rotor helicopters, have been put into practical use in various tasks such as measurement, logistics, monitoring, and search and rescue tasks. In order to realize a more advanced drone for autonomous monitoring and search and rescue tasks, sensor capabilities are critical. For example, drones are deployed after sunset to search and rescue potential victims of a disaster. Since any sensor has weak and strong points, a complimentary selection of sensors is mandatory. Image sensor, the most popular sensor for monitoring, provides a wide range of visual scene analysis information including object detection and recognition, but it also has drawbacks in darkness and occlusion. Other sensors, in particular, auditory sensors may compensate for its weakness. In this paper, we focus on the auditory sensor to search for sound sources on the ground.

This is a preprint of an article whose final and definitive form has been published in ADVANCED ROBOTICS [year of publication], copyright Taylor & Francis and Robotics Society of Japan, is available online at: http://www.tandfonline.com/Article DOI;10.1080/01691864.2020.1757506.

†Supported by ImPACT TRC and Kakenhi 19H00750.
‡Supported by ImPACT TRC and Kakenhi 17K00365 and 19H00750.
*Corresponding author. Email: kumon@gpo.kumamoto-u.ac.jp



(a) Female Voice. (b) Rotor Noise.

Figure 1. Spectrogram of a Female Voice and Rotor Noise

1.1 *Robot Audition and Drone Audition*

Robot audition proposed by Nakadai and Okuno [1] aims to the understanding of the sound environment by robots equipped with microphones. They have developed the open-sourced robot audition software called HARK [2] as an audio equivalent to OpenCV [3]. HARK has been applied to a wide range of applications from human-robot interactions, multi-person interactions, an in-car navigation system, bird song scene analysis, to search and rescue tasks. As another example of robot audition application, Sasaki proposed to localize multiple sound sources by triangulation using a large microphone array on a mobile robot [4].

Kumon [5] proposed a mobile cart control method based on the correlation matrix of an extended Kalman filter for sound source localization. For drones, Basiri [6] mounted three microphones on a small fixed-wing aircraft and proposed using a particle filter to localize sound source, i.e., an emergent whistle, on the ground. In the case of *drone audition*, that is, the application of robot audition to drones, *localization*, or extracting the directional information of a sound source, is a minimum function and should estimate the sound source in three-dimensions including not only the horizontal direction (hereinafter, *azimuth*) but also the vertical direction (hereinafter, *elevation*). The distance is difficult to estimate with a single microphone arrays.

Among drones, a multi-rotor helicopter is suitable for observing ground sound sources because it is capable of vertical take-off and landing, hovering, and low-speed flight with a simple structure combining fixed-pitch rotors. For example, if the sound source direction can be obtained by hovering stably and capturing sounds, the sound source position can be estimated by calculating the position of the drone under a simple assumption of height (altitude) and direction. For example, Okutani [7] and Ohata [8] proposed to estimate the position of the ground sound source based on the MUSIC (MUltiple SIgnal Classification) method [9] which estimates noise information sequentially to cancel it from the captured sounds and estimates the sound source direction. Several variation of MUSIC methods are incorporated into HARK. Washizaki [10] integrated the sound source direction obtained during the flight to estimate the sound source position in three-dimensions. Wakabayashi [11] and Yamada [12] proposed a method that makes the drone approach a target sound source to make the estimation of the sound source position more accurate.

1.2 *Problems with Drone Audition*

c

2. Drone Audition: Method, Platform, and System

This section discusses sound source localization with a microphone array, the drone we use, and the system architecture.

2.1 *Sound Source Localization with a Microphone Array*

According to the experience with HARK [22], MUSIC [9] is the best method for sound source localization (direction only) with a microphone array in real-world environments including indoor, outdoor, natural environments and the sky. MUSIC estimates the direction of a sound source based on the subspace method using orthogonality between signal space and noise space and is robust against directional noise. The original MUSIC assumes that the number of sound sources and that of noise are given in advance and that the power of noise is weaker than that of sound sources.

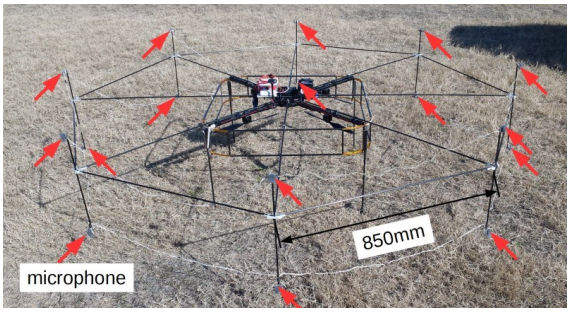
Many extensions are proposed to relax these assumptions as ego-noise of a drone exceeds the target signal in most cases. Okutani [7] proposed incremental Generalized EigenValue Decomposition (iGEVD-MUSIC), which can deal with high power noise by introducing a noise correlation matrix and GEVD even when the signal-to-noise ratio (hereinafter, SNR) is less than 0 B. In addition, the noise correlation matrix is incrementally estimated to adapt to dynamic changes in noise. However, it has two main drawbacks; high computational cost and over-estimation of noise correlation matrix. Ohata [8] extended MUSIC called incremental Generalized Singular Value Decomposition (iGSVD-MUSIC), which attains low computational cost. They also developed Correlation Matrix Scaling (CMS), i.e., soft whitening of noise, to solve the over-estimation problem. Experimental results show that iGSVD-MUSIC with CMS improves sound source localization drastically and achieves real-time processing. The both systems, (i)GEVD/GSVD-MUSIC, are incorporated in HARK. We, therefore, adopt GSVD-MUSIC to determine the sound source direction. Again, if ego noise is a non-stationary large signal and the target sound is far away from the drone and thus has low SNR, the performance of sound source localization degrades severely.

It is also important to note that searching for ground sound sources from the air by a drone is quite different from searching for sound sources on the ground by a mobile robot. Since there is usually no shielding between the microphone array in the air and ground sound sources, the reverberation may be ignored outdoors. Therefore, relatively simple sound transfer can be assumed without paying much attention to reflection and reverberation. At the same time, since sound signals may arrive to the drone from a wide range on the ground, *multiple sound sources* should be taken into consideration. Since the SNR of signals captured by a microphone array mounted on a drone is low due to ego noise, detected sound source should be verified whether it is a target sound source, say speech, a non-target sound source, or false detection. Since such sound source identification introduces another errors, we adpot a feature vector of sound separated by GHDSS (Geometric Constrained High-order Source Separation) [2, 22] to disambiguate the crossing problem.

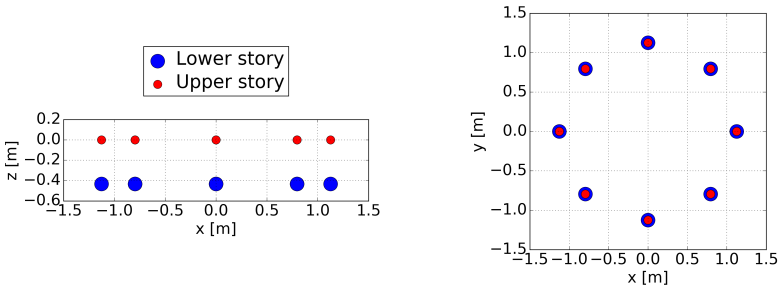
2.2 *Drone and Microphone Array*

In this paper, we use a quadrotor helicopter as a drone. The critical problem with drone audition at the platform level is the design of microphone configuration. The purpose of drone audition is to extract the direction of a sound source, that is, localization, in azimuth and elevation. Based on the previous exploitation of microphone configuration [17, 18], we adopt a two-story configuration of microphone array shown in Figure 2 with a drone of enRoute PG-560 [24]. Different approaches such as a spherically-packed microphone as in [20, 25] are also applicable as far as omni-directional observation is possible.

The microphone array captures sounds as multi-channel wave data by the on-board sound processing unit RASP-ZX [26], and transmitted to the ground station via a 2.4 GHz wireless LAN. The system scheme is depicted in Figure 3. The drone is capable of automatic flight by GPS with flight controller called Pixhawk [27], which provides the position, altitude, and posture of the drone in real-time. In this implementation, a ROS node of MAVROS (MAVLink extendable communication node for ROS with proxy for Ground Control Station) [28] is treated as an interface of aircraft information at the ground station, which is connected with a ROS



(a) Drone with a 2-Story Microphone Array.



(b) Side-view of Mic Configuration. (c) Overview of Mic Configuration.

Figure 2. Drone Platform: enRoute PG-560 Quadrotor helicopter with a Microphone Array

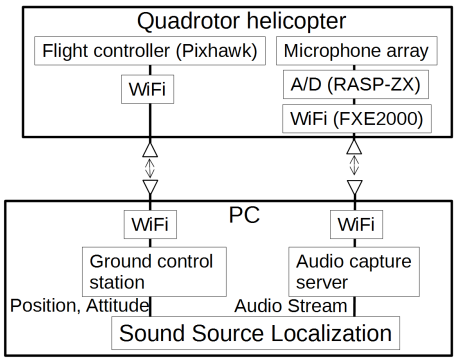


Figure 3. System Schme for Drone Audition

node to estimate the position in real-time. Figure 4 depicts a screenshot of GUI at the ground station, which shows a birds-eye view of the field, MUSIC spectrum, and sound source position estimation.

3. Estimation of Multiple Sound Source Positions

3.1 *Sound Source Position Estimation by a Kalman Filter*

To cope with uncertainty of observed signals in sound source localization, Kalman filter [29] is used to estimate positions of ground sound sources. As shown in Figure 5, let h, ψ, θ, and ϕ be the altitude, yaw, azimuth, and elevation of a sound source, respectively. To model uncertainty of estimation of each parameter, h, ψ, ϕ, it is divided into an observed value $\tilde{h}$ and uncertainty δh; i.e., h is represented as $h = \tilde{h} + \delta h$. Here, uncertainty $\delta \boldsymbol{g}$ represents the roughness to indicate how uneven the ground surface is.

Assume that uncertainty of the signal is sufficiently small and that m observations are obtained

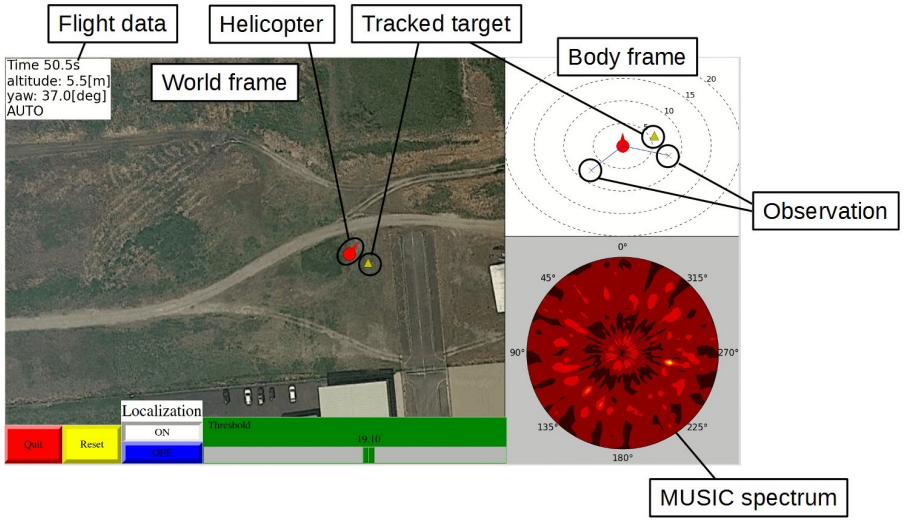


Figure 4. Screenshot of the developed GUI for Drone Audition

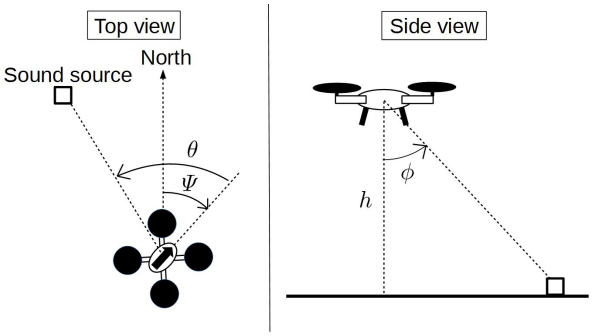


Figure 5. Geometric Relationship between Sound Source Position and its Estimated Direction

at time k, the sound source position $\boldsymbol{z}$ in the drone coordinate is calculated from the jth ($j = 1, 2 \cdots m$) observation as follows:

$$\boldsymbol{z}_{k,j} = (\hat{h}_k + \delta h) \tan(\hat{\phi}_k + \delta\phi)\boldsymbol{b}_k + \delta\boldsymbol{g}_k, \tag{1}$$

where $\boldsymbol{b}_k = \left[\sin(\psi_k - (\hat{\theta}_k + \delta\theta)) \cos(\psi_k - (\hat{\theta}_k + \delta\theta))\right]^T$.

Assume that each uncertainty follows a normal distribution, that is, $\delta h \sim \mathcal{N}(0, \sigma_h)$, $\delta\phi \sim \mathcal{N}(0, \sigma_\phi)$, $\delta\theta \sim \mathcal{N}(0, \sigma_\theta)$, and $\delta\boldsymbol{g} \sim \mathcal{N}(0, \sigma_{\boldsymbol{g}})$, Equation (1) can be rewritten as $\boldsymbol{z}_{k,j} = \boldsymbol{U}\hat{\boldsymbol{x}}_{k,j} + \boldsymbol{a}_k$, where $\hat{\boldsymbol{x}}_{k,j}$ is a sound source position estimated by observation, that is, $\hat{\boldsymbol{x}}_{k,j} = \boldsymbol{b}_k \hat{h}_k \tan\hat{\phi}_k$. $\boldsymbol{U}$ is an observation matrix, which is a unit matrix in this paper. $\boldsymbol{a}_k$ follows a normal distribution $\mathcal{N}(0, \boldsymbol{P})$ where co-variance matrix $\boldsymbol{P}$ is calculated based on the geometrical relation shown in Figure 5.

Since the position of a ground sound source is estimated in the drone coordinate, it should be converted into the world coordinate. The problem with such coordinate conversion is that dynamic fluctuation of the location and posture of the drone may force the estimated position unstable. To cope with this unstability of the position estimation, we use a Kalman filter to get a stable estimation. Let $\boldsymbol{u}_k^t$ and u_k^y a change in parallel movement of the drone and one in yaw in the world coordinate, respectively. The position of a ground sound source in the drone coordinate changes from $\boldsymbol{x}_{k-1}$ to $\boldsymbol{x}_k$ from time $k-1$ to k, and its movement vector is $\boldsymbol{v}_k$ as shown in Figure 6. Let $\boldsymbol{c}_k$ be the movement vector of a target sound source. Assuming $\boldsymbol{v}_k$ $\boldsymbol{c}_k$

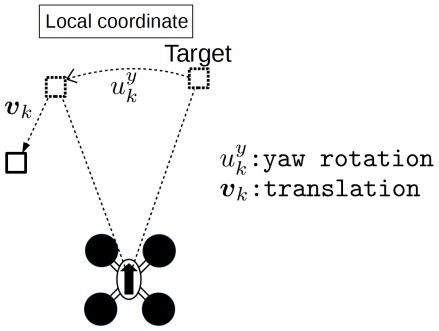


Figure 6. Update of the Estimated Target Position with drone flights

have uncertainty, the i-th ($i = 1, 2 \cdots n$) sound source $\boldsymbol{x}_{k,i}$ at time k is represented as follows:

$$\boldsymbol{x}_{k,i} = \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{x}_{k-1,i} + \hat{\boldsymbol{c}}_k + \delta\boldsymbol{c} + \hat{\boldsymbol{v}}_k + \delta\boldsymbol{v}, \tag{2}$$

where $\boldsymbol{R}(\hat{u}_k^y)$ is a rotation matrix that rotates around the yaw by a rotation angle $\hat{u}_k^y$.

Assuming that $\delta\boldsymbol{v} \sim \mathcal{N}(0, \boldsymbol{Q})$, Equation (2) can be written as $\boldsymbol{x}_{k,i} = \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{x}_{k-1,i} + \hat{\boldsymbol{c}}_k + \hat{\boldsymbol{v}}_k + \boldsymbol{t}_k$, where $\boldsymbol{t}_k \sim \mathcal{N}(0, \boldsymbol{Q})$. The co-variances used in the above Kalman filter are empirically determined. $\hat{\boldsymbol{c}}_k$ is estimated as $\hat{\boldsymbol{c}}_k = \hat{\boldsymbol{e}}_k \delta k$ by using an average speed $\hat{\boldsymbol{e}}_k$ calculated by the previous n steps.

Here, we introduce a threshold V for speed to discriminate between moving and stationary sound sources. If $||\hat{\boldsymbol{e}}_k|| < V$, $\hat{\boldsymbol{e}}_k = [0, 0]^T$. With the above model, the prediction and update steps of the Kalman filter for sound source i and observation j can be written as follows.

1. Prediction step

$$\bar{\boldsymbol{x}}_{k,i} = \boldsymbol{R}(\hat{u}_k^y)\boldsymbol{x}_{k-1,i} + \hat{\boldsymbol{c}}_k + \hat{\boldsymbol{v}}_k \tag{3}$$

$$\bar{\Sigma}_{k,i} = \boldsymbol{Q} + \boldsymbol{R}(\hat{u}_k^y)\Sigma_{k-1,i}\boldsymbol{R}(\hat{u}_k^y)^T \tag{4}$$

2. Update step

$$\tilde{\boldsymbol{z}}_{k,ij} = \boldsymbol{z}_{k,j} - \boldsymbol{U}\boldsymbol{x}_{k,i} \tag{5}$$

$$\boldsymbol{S}_k = \boldsymbol{U}\hat{\Sigma}_{k,i}\boldsymbol{U}^T + \boldsymbol{P} \tag{6}$$

$$\boldsymbol{K}_k = \bar{\Sigma}_{k,i}\boldsymbol{U}^T\bar{\boldsymbol{S}}_k^{-1} \tag{7}$$

$$\boldsymbol{x}_{k,i} = \bar{\boldsymbol{x}}_{k,i} + \boldsymbol{K}_k\tilde{\boldsymbol{z}}_{k,ij} \tag{8}$$

$$\Sigma_{k,i} = \bar{\Sigma}_{k,i} - \boldsymbol{K}_k\boldsymbol{U}\hat{\Sigma}_{k,i}, \tag{9}$$

3.2 *Data Association for Sound Source Tracking*

3.2.1 Effective Area and Distances

The critical problem with sound source localization in the presence of multiple sound sources (or simply sources) is disambiguation in *data association* that matches tracking sound sources with observations. In this paper, we propose the GNN-c (Global Nearest Neighbor with classification measurements based on GNN (Global Nearest Neighbor) [15, 16].

First, in order to reduce wrong correspondence between racking source and observation, we introduce the *effective area* as described below for each tracking source. If the observation j is in the effective area of the tracking source i, the tracking source i is updated using this observation j. The Mahalanobis distance between the tracking source i and the observation j is calculated by observation error specified by Equation (8) and co-variance matrix specified by Equation (9) as follows:

$$d^2_{k,ij}{}^{(M)} = \tilde{\boldsymbol{z}}^T_{k,ij} \boldsymbol{S}^{-1} \tilde{\boldsymbol{z}}_{k,ij}. \tag{10}$$

Let $\boldsymbol{s}_{k,i}$ and $\boldsymbol{s}_{k,j}$ be feature vectors of tracking source i and observation j, respectively. We define the Euclidean distance between these feature vectors as follows:

$$d^2_{k,ij}{}^{(E)} = (\boldsymbol{s}_{k,i} - \boldsymbol{s}_{k,j})^T (\boldsymbol{s}_{k,i} - \boldsymbol{s}_{k,j}). \tag{11}$$

With these distances, we define the distance between the tracking sound source i and the observation j and the threshold as follows:

$$d^2_{k,ij} = d^2_{k,ij}{}^{(M)} + w d^2_{k,ij}{}^{(E)} \tag{12}$$

$$G = G^{(M)} + w G^{(E)} \tag{13}$$

where $G^{(M)}$ and $G^{(E)}$ are thresholds corresponding to $d^2_{k,ij}(M)$ and $d^{(E)}_{k,ij}$ respectively and w is a weight that determines the relative importance of two distances.

The value of $G^{(M)}$ is determined from the χ-squared distribution with 2 degrees of freedom, since the Mahalanobis distance follows the χ-squared distribution. In this paper, in order to set the 95% confidence ellipse of the tracking source as the effective region, we set $G = 7.951$. Since the threshold $G^{(E)}$ depends on the classifier and the number of classes, its value is determined empirically. We set $G^{(E)} = 5.5$. Finally, if the condition,

$$d^2_{k,ij} < G \tag{14}$$

satisfies, observation j is assigned to the tracking source i. Then, the feature vector of the tracking sound $\boldsymbol{s}_{k,i}$ is updated by taking an average of feature vectors assigned to i.

In this study, Support Vector Machine (SVM) based feature classification [30] is utilized to compute the feature vector s_k from acoustic feature in Mel-Scale-Log Spectrum (MSLS) [2] of a separated sound. As mentioned above, GHDSS is used to separate sounds from a wave file transferred from the drone to the ground station via 2.4 GHz WiFi.

3.2.2 Tracking Multiple Sound Sources

If a single observation exists in the effective area of one tracking source, it can be assigned immediately. However, if multiple observations exist in the effective area of one tracking source, or if one observation is within the effective areas of multiple tracking sources, there remains ambiguity in data association, which deteriorates the performance of sound source position estimation.

Assum that n sources are being tracked at time k and m observations are obtained under the assumption of multiple sound sources. What is the criteria to determine whether a new observation is either from the source currently being tracked or from a new source? We verify for all combinations of tracking sources and observations whether or not the Equation (14) is

satisfied, and solve the assignment problem given by the cost matrix C below.

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ c_{21} & c_{22} & \cdots & c_{2m} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix} \tag{15}$$

$$c_{ij} = \begin{cases} E & d_{ij}^2 > G \\ d_{ij}^2 & d_{ij}^2 \leq G \end{cases} \tag{16}$$

where $E > G$ is a constant value to indicate non-active state. With this cost matrix, we find a solution that minimizes the sum of distances in Equation (16) by combining tracking sources and observations. In the research, this problem is solved using Munkres method [31]. However, the constant value E in Equation (16) is set to a sufficiently large value in order to solve this problem as finding the combination that minimizes the sum of distances.

3.3 *Tracking Multiple Sound Sources*

We introduce a management method for stable sound source tracking.

3.3.1 Generation of New Sound Source Tracking

If there is no sound source being tracked, or if the observation is not assigned to the sound source currently being tracked, a new tracking sound source is generated and tracking by the Kalman filter starts. To avoid tracking a short fragment of sound source, tracking continues only if observations are continuously obtained more than a continuous observation threshold N_1. If a tracking sound source obtains a sufficient number of times, that is, more than a sound source detection threshold N_2 ($N_2 > N_1$), the tracking sound source is labeled as *valid*.

3.3.2 Maintenance of Multiple Sound Source Tracking

Sound source tracking continues while observation is obtained, that is, *active*. When observation is not obtained for a long time, for example, no sound source generates a signal, or the drone is too far from the sound source to capture any signal, we have a problem of large accumulated uncertainty due to continuous expansion of effective area.

In order to avoid this, when observation is not during T_1 for a tracking sound source that is labeled as valid with a sufficient number of time (N_3), tracking by the Kalman filter is stopped, that is, *dormant*. Then, the effective area is reinitialized to a constant radius r. If observation is obtained again, sound source tracking resumes and becomes active.

3.3.3 Termination of Sound Source Tracking

If the number of observations is not less than N_1 and not more than N_3 and sound source tracking is dormant during T_2, the tracking is terminated.

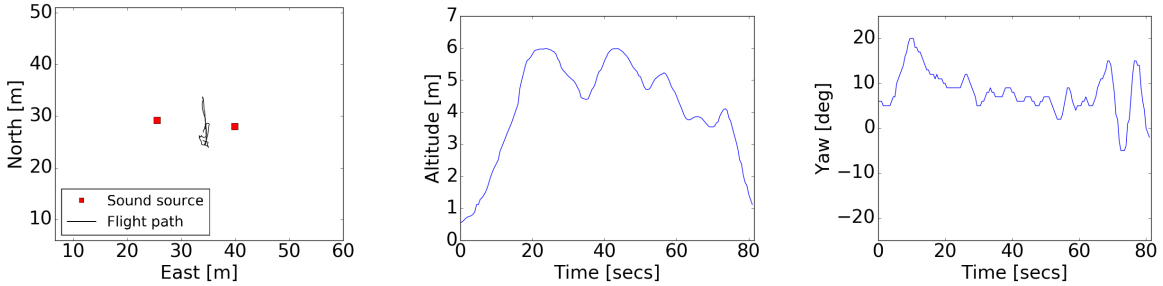
4. Physical Experiments and Evaluation

Flight experiments were conducted to verify the performance of the proposed system.

4.1 *Setting of Experiments*

We use the drone, that is a quadrotor helicopter equipped with a microphone array, to fly in two ways: hovering by manual steering assisted by posture stabilization by the onboard controller,

Figure 7. Snapshot of Experiment: two sound sources are a male speaker and an emergent whistle blown by a man

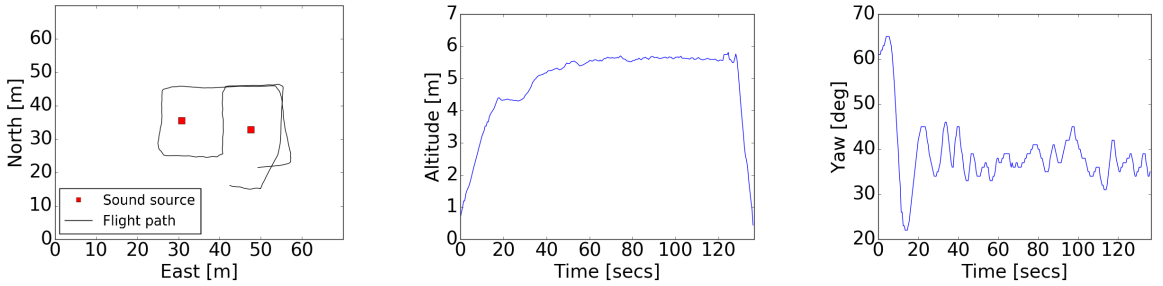


(a) Positions of Drone, Loudspeaker, and Whistler.

(b) Attitude.

(c) Yaw.

Figure 8. Manual hovering of Drone to estimate sound source position.



(a) Flight path and positions of sound sources.

(b) Attitude.

(c) Yaw.

Figure 9. Autopilot flight of drone to estimate sound source position.

and waypoint tracking by autopilot. For each flight, the drone was heading north and kept the altitude of approximately 5 m from the ground (see Figure 7).

We use two sound sources, a loudspeaker on the ground that utters man's scream and an emergent whistle sound blown by a standing man of 1.7 m tall. The positions of the two sound sources are shown in Figures 7 and 8. In this figure, while the drone is hovering, the loudspeaker and the man are located at the mirror positions from the drone at the distance of about 10 m (see Figure 8(a)). The route of autopilot is also shown in Figure 9(a). In both figures, red square points indicate the position of sound sources.

4.2 *Results of Experiments*

Flights for hovering and autopilot are shown in Figures 8 and 9, respectively. The flight time for hovering and autopilot are about 80 seconds and 130 seconds, respectively. The altitude of each

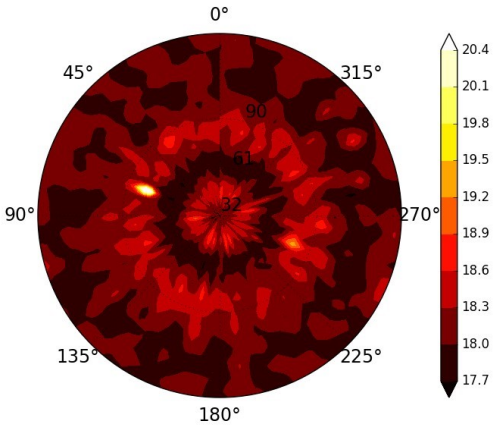
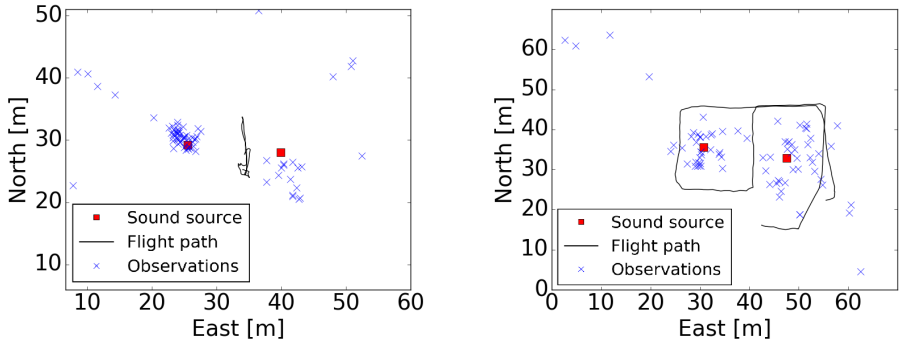


Figure 10. Example of MUSIC spectrum: two yellow peaks in 80° and 260°.



(a) Manual Hovering. (b) Autopilot Flight.

Figure 11. Observations projected on the ground.

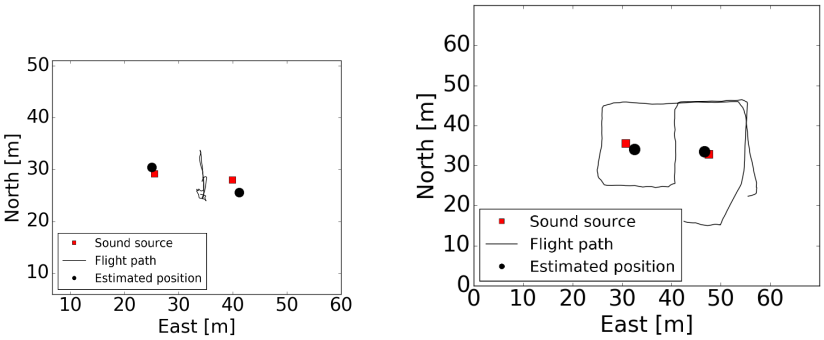
flight differs as is shown in Figures 8(b) and 9(b), because hovering is controlled manually. In autopilot, the drone is set to head to the north, or 0°, while actual flight was 40° to the west. This deviation in yaw does not matter because such fluctuation is taken into consideration in Equation 2.

Sound source localization is calculated every 0.5 sec by GSVD-MUSIC method to determine the direction in real time and sound source position is estimated by integrating the direction and drone data of attitude and posture. Since rotor noise increases during takeoff and landing and causes frequent false detection, localization by GNN-c is set to be performed only at the altitude of 3.5 m or more. The parameters described in Section 3.3 are set as follows: $N_1 = 2$, $N_2 = 4$, $N_3 = 6$, $T_1 = 10$ sec, $T_2 = 10$ sec, and $r = 1.5$ m.

Figure 10 shows MUSIC spectrum in azimuth in the circumferential direction and elevation angle in the radial direction when the drone is centered and viewed vertically from the above. In this figure, two peaks of the MUSIC spectrum appear in the directions of 80° and 260°. Since the purpose of this paper is to localize multiple sound sources, we use multiple peaks in the MUSIC spectrum as observations. The number of peaks used for observation is set equal to that of sound sources used in GSVD-MUSIC of HARK. It is set to 3 in this experiment. Furthermore, another threshold is introduced to reduce false detection, where peaks with small power are not used as observation. The threshold was set empirically to 19.1 in this experiment.

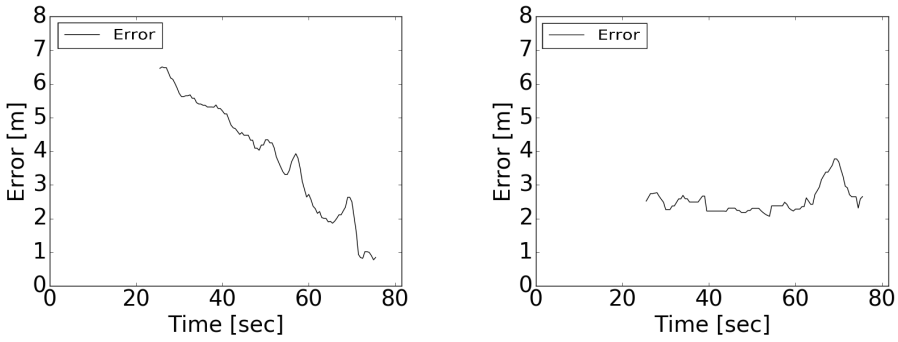
4.2.1 Results of sound source position

Figure 11 shows the observations of each flight. In Figure 11(a), dense observation is obtained near the loudspeaker, but the observation near the whistle has variation and is slightly offset



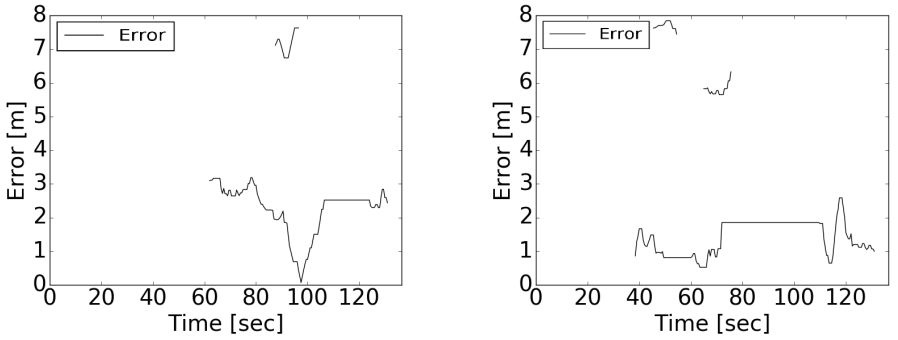
(a) Manual Hovering. (b) Autopilot Flight.

Figure 12. Final Estimation of Sound Source Position.



(a) Loudspeaker (Man's Scream). (b) Whistle.

Figure 13. Position Estimation Error in Manual Hovering)



(a) Loudspeaker (Man's Scream). (b) Whistle.

Figure 14. Position Estimation Error in Autopilot Flight

from the actual position. Furthermore, we can notice false positives in the lower left and right. In Figure 11(b), observations are scattered since the drone is moving. Although false positives appear in the upper left and lower right, most observations are obtained around the actual sound source positions.

The estimated sound source position by the Kalman filter at the end of each flight is shown by a black circle in Figure 12. The figure demonstrates that correct or nearly correct sound source positions are estimated in both flights.

Figures 13 and 14 show errors in the estimated position of the tracking source. In Figure 13(a), error is large at first, but error decreases as the drone approaches a sound source. In Figure 14(a), error appears in the upper part as a fragment. This is because tracking sound source management

Table 1. Estimation Error of Each Target.

Sound Source	Manual Hovering Error [m]	Autopilot Flight Error [m]
Loudspeaker	0.84 [m]	2.44 [m]
Whistle	2.65 [m]	1.00 [m]

proposed in Section 3.3.3 notices this false detection and then aborts its tracking. In addition, tracking sound source management proposed in Section 3.3.2 keeps error within a constant value. In other words, the prediction by the Kalman filter is stopped and the matching by data association is performed using the position saved in the world coordinate.

Table 1 shows the error of the estimated sound source positions after each flight. As a result, the error was within 3 m. This result demonstrates that localization of multiple sound sources is practical by the proposed real-time sound source localization and acoustic data transmission system.

5. Conclusion

In this paper, we estimate the position of multiple sound sources on the ground using acoustic information obtained by a quadrotor helicopter equipped with a microphone array. The effective area of each tracking sound source works well to prevent false correspondence between a tracking sound source and an observation.

Performance evaluation was conducted through experiments using manual hovering and autopilot flight. The error is large when the drone is far from a sound source, but the error decreases to less than 3 m when approaching a target. The flight path of the drone can be specified to fly around the sound sources in this experiment. However, this may not be the case in real-world problems because sound source positions are unknown in advance. As future work, real-time flight planning is mandatory with exploratory sound source searching.

References

[1] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In: *Proceedings of AAAI-2000*, pp.832-839.

[2] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and Implementation of Robot Audition System "HARK". *Advanced Robotics*, Vol.24, No.5-6 (2010), pp. 739-761.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, vol. 15, pp. 120-125, 2000.

[4] Y. Sasaki, S. Kagami, and H. Mizoguchi. Multiple sound source mapping for a mobile robot by self-motion triangulation. In: Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2006, pp. 380-385.

[5] M. Kumon and S. Uozumi. Binaural localization for a mobile sound source. *Journal of Biomechanical Science and Engineering*, vol. 6, no. 1 (2011), pp. 26-39.

[6] M. Basiri, F. S. Schill, P. Lima U., and D. Floreano. Robust acoustic source localization of emergency signals from micro air vehicles. In: Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012, pp. 4737-4742.

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

[8] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai. Improvement in outdoor sound source detection using a quadrotor embedded microphone array. In: Proceedings of 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 104, pp. 1902-1907.

[9] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, vol.34, no.3 (1986), pp276-280.

[10] K. Washizaki, M. Wakabayashi and M. Kumon. Position estimation of sound source on ground by multirotor helicopter with microphone array. In: Proceedings of 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 1980-1985.
[11] M. Wakabayashi, K. Washizaki, and M. Kumon. Sound Source Search by a Quadrotor Helicopter (in Japanese). In: Proceedings of the 34th Annual Convention of Robotic Society of Japan, 2016, RSJ20161C3-04.
[12] K. Yamada and M. Kumon. Map Management of Sound Source Search based on Grid based Recursive Bayes Filter by a Multi-rotor Helicopterr (in Japanese). In: Proceedings of the 35th Annual Convention of Robotic Society of Japan, 2017, RSJ2017AC3AC2-06.
[13] C. Kim, F. Li, A. Ciptadi, and J.M. Rehg. Multiple Hypothesis Tracking Revisited. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4696-4704.
[14] S.H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid. Joint Probabilistic Data Association Revisited. In: Proceedings of Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3047-3055
[15] P. Konstantinova, A. Udvarev, and T. Semerdjiev. A Study of a Target Tracking Algorithm Using Global Nearest Neighbor Approach. In: CompSysTech, 2003, pp. 290-295.
[16] M. Ozaki, K. Kakinuma, M. Hashimoto, and K. Takahashi. Laser-based pedestrian tracking in outdoor environments by multiple mobile robots. *Sensors*, Vol. 12 (2012), pp. 14489-14507.
[17] M. Kumon and T. Ishiki. A microphone array configuration for an auditory quadrotor helicopter system. In: Proceedings of the 12th IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR), 2014, p. 34.
[18] T. Ishiki and M. Kumon. Design model of microphone arrays for multirotor helicopters. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 6143-6148.
[19] S. Tadokoro. Overview of the ImPACT Tough Robotics Challenge and Strategy for Disruptive Innovation in Safety and Security. *Disaster Robotics - Results from the ImPACT Tough Robotics Challenge*, S. Tadokoro Ed., Springer Tracts in Advanced Robotics 128, pp. 3-22, 2019.
[20] K. Nakadai, M. Kumon, H.G. Okuno, K. Hoshiba, M. Wakabayashi, K. Washizaki, T. Ishiki, D. Gabriel, Y. Bando, T. Morito, R. Kojima, O. Sugiyama. Development of microphone-array-embedded UAV for search and rescue task. In: Proceedings of 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 5985-5990.
[21] K. Nonami, K. Hoshiba, 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, and H. Liu. Recent R&D Technologies and Future Prospective of Flying Robot in Tough Robotics Challenges. *Disaster Robotics*, STAR 128, Chapter 3, pp. 77-142, 2019.
[22] H.G. Okuno and K. nakadai. Robot Audition: Its Rise and Perspective. In: Proceedeings of 2015 International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5610-5614.
[23] T. Ishiki, K. Washizakki, and M. Kumon. Evaluation of microphone array for multirotor helicopters. *Journal of Robotics nad Mechatronics*, Vol. 29, No. 2 (Feb. 2017), pp. 154-176.
[24] enRoute Inc. ZION-PG560-SPECS. Available from: https://enroute1.com/portfolio-posts/zion-pg-700/zion-pg560-specs/
[25] K. Hoshiba, K. Washisaka, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H.G. Okuno. Design of UAV-embedded Microphone Array System for Sound Source Localization in Outdoor Environments. *Sensors*, Vol. 17, No. 11 (2017), 2535.
[26] System Infrontier Inc. Acoustic Processing Unit (RASP-ZX). Available from: http://www.sifi.co.jp/system/modules/pico/index.php?_id=36&ml_lang=en
[27] 3D Robotics Inc. Pixhawk. Available from: https://store.3dr.com/t/pixhawk
[28] ROS.org. MAVROS - – MAVLink extendable communication node for ROS with proxy for Ground Control Station. Available from: http://wiki.ros.org/mavros
[29] G. Welch and G. Bishop. An Introduction to the Kalman Filter. In: ACM SIGGRAPH Course Notes, 2001, Course 8.
[30] K. Makoto, Y. Ito, T. Nakashima, T. Shimoda, and M. Ishitobi. Sound source classification using support vector machine. *IFAC Proceedings Volumes*, vol. 40, issue 13, 2007, pp. 465-470.
[31] J. Munkres. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathmatics*, vol. 5, no. 1 (1957), pp. 32-38.
[32] K. Hoshiba, K. Nakadai, M. Kumon, and H.G. Okuno. Assessment of MUSIC-Based Noise-Robust Sound Source Localization with Active Frequency Range Filtering. *Journal of Robotics and Mecha-*

tronics, Vol. 30, No. 3 (June 2018), pp.426-435.