Multiple Sound Source Position Estimation by Drone Audition Based on Data Association Between Sound Source Localization and Identification

Mizuho Wakabayashi, Hiroshi G. Okuno, and Makoto Kumon

***Abstract*—Drone audition, or auditory processing for drones equipped with a microphone array, is expected to compensate for problems affecting drones' visual processing, in particular occlusion and poor-illumination conditions. The current state of drone audition still assumes a single sound source. When a drone hears sounds originating from multiple sound sources, its sound-source localization function determines their directions. If two sources are very close to each other, the localization function cannot determine whether they are crossing or approaching-then-departing. This ambiguity in tracking multiple sound sources is resolved by data association. Typical methods of data association use each label of the separated sounds, but are prone to errors due to identification failures. Instead of labeling by classification, this study uses a set of classification measures determined by support vector machines (SVM) to avoid labeling failures and deal with unknown signals. The effectiveness of the proposed approach is validated through simulations and experiments conducted in the field.**

***Index Terms*—Aerial systems, localization, mapping, perception and Autonomy, robot audition.**

I. Introduction

DISASTER robotics is intended to improve the promptness and effectiveness of search-and-rescue missions [1]. From the viewpoint of a well-known guidance "*First 72 Hour Response*," drones are considered to be useful for monitoring, logistics, and search-and-rescue missions, because they are unaffected by most ground conditions [2]. To improve the performance of search-and-rescue missions, drones require more sensory capabilities. The current drones use light detection and ranging (LiDAR) and cameras. Since vision sensors are not a panacea and have problems such as occlusion and inability to work without illumination, drones should be able to exploit other sensor modalities to compensate for such weaknesses.

Manuscript received September 10, 2019; accepted December 26, 2019. Date of publication January 9, 2020; date of current version January 23, 2020. This letter was recommended for publication by Associate Editor Dr. I. Gilitschenski and Editor C. C. Lerma upon evaluation of the reviewers' comments. This work was supported in part by JSPS under Grants KAKENHI 19H00750 and 17K00365, and in part by JST under ImPACT Tough Robotics Challenge. *(Corresponding author: Makoto Kumon.)*

M. Wakabayashi is with GSST, Kumamoto University, Kumamoto 860-8555, Japan (e-mail: m.wakabayashi@ick.mech.kumamoto-u.ac.jp).

H. G. Okuno is with Graduate Program for Embodiment Informatics, Waseda University, Shinjuku Tokyo 162-0079, Japan (e-mail: okuno@nue.org).

M. Kumon is with the Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto 860-8555, Japan (e-mail: kumon@gpo.kumamoto-u.ac.jp).

Digital Object Identifier 10.1109/LRA.2020.2965417

In this letter, we focus on sound processing to compensate for visual processing (hereinafter, *drone audition*). Nakadai and Okuno proposed a hearing capability for robots with their own microphones, i.e., *robot audition* [3]. Its three main functions are *sound source localization (SSL)*, *sound source separation (SSS)*, and *recognition of separated sound sources*. Assuming that the speech is captured by microphones close to the speaker, robust automatic speech recognition (*ASR*) is possible, thanks to the recent development of deep learning. However, if speech is contaminated by noise or other speech signals, the performance of ASR deteriorates drastically [4]. In fact, there remains a big gap in performance between clean data and contaminated data, the latter of which is quite common in robotics and real-world applications. For bridging a gap, the open-source robot audition software called "HARK" [5] has been developed, which is used in this study.

SSL is the most important function because most of the current applications are based on it. For example, a robot detects the direction of sound sources, or speakers, and pays attention to one of them. SSL is also used to estimate the positions of sound sources. Sasaki [6] proposed to estimate the positions of multiple sound sources by triangulation using a large microphone array (hereinafter, *a mic array*) on a mobile robot. Kumon [7] proposed a mobile cart control method based on the correlation matrix of an extended Kalman filter for SSL. Martinson [8] proposed auditory evidence grids to estimate a position of moving sound sources in a room. Sekiguchi [9] proposed to estimate the positions of multiple sound sources by using multiple robots each equipped with an independent mic array. Since visual, auditory, and other perception of a robot carries various uncertainties, this is tackled by simultaneous localization and mapping (SLAM) in Sasaki [6] and Evers [10]. For drone audition, Basiri [11] proposed to mount three microphones on a small fixed-wing aircraft; their system could localize an emergency whistle call on the ground using particle filters. Wang [12] proposed to mount a circular 8-ch mic array on a 3DR IRIS quadcopter; their system was shown by simulation that it could localize a speech signal by calculating DoA at each time-frequency bin and filtering time-frequency bins spatially. Hoshiba [13] proposed to use a 12-channel mic array to localize and estimate the position of a human caller for help on the ground. It should be noted that robot audition usually needs the horizontal direction (hereinafter, *azimuth*), while drone audition needs the vertical direction (hereinafter, *elevation*) in addition to azimuth.

II. Related Work

A significant problem of drone audition is noise caused mainly by the rotors and airflow around the drone (hereinafter, *ego-noise*). Normal noise suppression algorithms cannot cope with such dynamically changing ego-noise under a low signal-to-noise ratio (*SNR*).

A. Sound Source Localization (SSL)

SSL is extensively surveyed in [14], but most efforts have been devoted to normal robots, not to drones. The most common SSL with a mic array is based on multiple signal classification (MUSIC) [15], which estimates the direction of a sound source using the subspace method that is based on the orthogonality between the signal space and noise space and is robust against directional noise [4]. A practical usage of the original MUSIC requires the number of sound sources to be given in advance and that the power of the noise is weaker than that of the sound sources in order to discriminate sound sources from noise.

Many extensions have been proposed to relax the above requirements as the power of ego-noise of a drone exceeds that of the target signal in most cases. Generalized eigenvalue decomposition for MUSIC (*GEVD-MUSIC*) can deal with high-power noise by introducing a noise correlation matrix (*NCM*) and GEVD even when the SNR is less than 0 dB [16]. NCM can be incrementally estimated to adapt to dynamic changes in noise (*i*GEVD-MUSIC). However, this has two major drawbacks: high computational cost and overestimation of NCM. Generalized singular value decomposition for MUSIC (*GSVD-MUSIC*) reduces these two drawbacks [17]. Since (*i*)GEVD/GSVD-MUSIC of HARK provides SSL in real-time, we chose to use GSVD-MUSIC for SSL in our study.

B. Tracking by Data Association

Another critical problem is *multiple sound sources* from a wide range of acoustic fields. Since SSL provides temporal fragments of localization, these fragments must be tracked in order to estimate the position and trajectory of sound sources. For example, if two sound sources cross each other, the tracking function may estimate either that they are crossing or that they are approaching and then departing, which causes uncertainty in tracking. This is also the case when the drone is flying and the relative positions of the two sound sources change. This kind of ambiguity in tracking multiple sound sources can be dealt with by *data association*.

Nakadai [18] proposed to integrate localization and facial recognition as a means of data association in tracking moving speakers with binaural mics. The separated sounds are used to identify the speaker and then direction data of the same speaker are collected to estimate the trajectory. Data association based on labeling is prone to malfunction because incorrect labeling will generate the wrong trajectory.

Multi hypothesis tracking (MHT) [19] and joint probability data association (JPDA) [20] are sophisticated data associations for signals with uncertainty. The concept of JPDA takes the all possible combinations into account, which requires significant computation. Also, to cope with noisy observations, MHT must maintain a huge number of hypotheses, which significantly increases the amount of calculation and makes real-time processing difficult. Furthermore, auditory observations are more uncertain and irregularly obtained, as opposed to the case of laser-scanned images. Therefore, we propose to extend the *Global Nearest Neighbor (GNN)* method [21] that can track multiple objects with a small amount of computation.

C. Sound Source Separation (SSS)

SSS methods with a mic array are classified into three categories: *beam-former* based on spatial sparseness, *temporal-frequency masking* based on spatial and temporal sparseness, and *blind separation* including independent component analysis and non-negative matrix factorization (NMF). Blind separation usually assumes that the number of sound sources is less than that of microphones and is too computationally expensive to run in real-time or even with a small latency. Zegers [22] proposed a method to jointly perform source separation and speaker identification using NMF. Because of our extensive experience on real-time SSS with HARK [4], we chose geometric high-order decorrelation-based source separation (GHDSS) of HARK, a variation of blind separation with spatial constraints [23].

D. Sound Source Identification (SSI)

Data association needs not only sound source direction but also various information from acoustic signals. Sound source identification (SSI) is a promising candidate. If the identification information is used in data association, an improvement in tracking accuracy can be expected. Examples of SSI include Bayesian classifiers [24], fuzzy classifiers [25], artificial neural networks [26], Gaussian mixture models [27], hidden Markov models [28], and support vector machine (SVM) [30]. We chose to use SVM because we need a set of the likelihood for various sources, not a label of the source.

III. Position Estimation System

A. System Overview

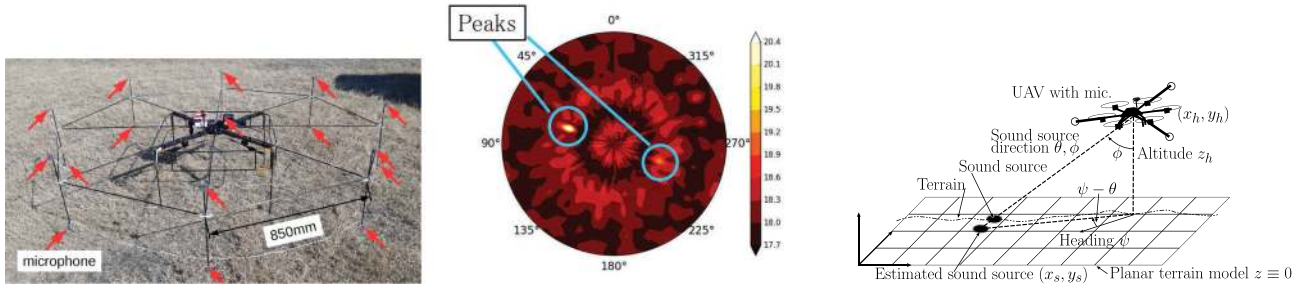
The system consists of the following modules (see Fig. 2).

1) Drone with a 16-channel mic array (Fig. 1(a)).
2) SSL (Fig. 1(b)).
3) Position estimation (Fig. 1(c)).
4) Single source tracking
5) SSS and extraction of MSLS features
6) SSI and extraction of feature vectors
7) Multiple source tracking by data association

B. Drone and Sound Source Localization (SSL)

A drone (ZION-PG560 of enRoute Inc., Japan) equipped with a 16-channel mic array (Acoustic Processing Unit, RASP-ZX of SystemInFrontier Inc., Japan) captures sounds and records them as a 16-ch 24 bit wave file with 16 kHz sampling (Fig. 1(a)).

GSVD-MUSIC calculates the MUSIC spectrum, and its peaks are filtered by threshold to determine the directions of the sound



(a) Drone with 16-ch Microphone Array (b) MUSIC Spectrum (c) Drone and a Sound Source in 3D Space

Fig. 1. Drone, MUSIC spectrum, and geometrical configuration of sound source position estimation.

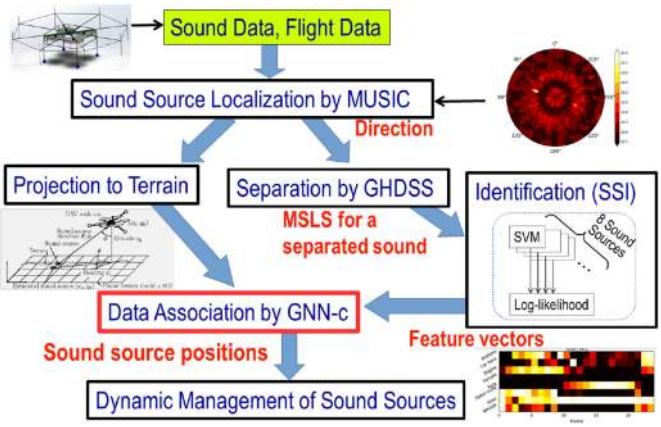


Fig. 2. Flowchart of the position estimation system.

sources (Fig. 1(b)). Each direction is projected to the ground to estimate the position on the ground (Fig. 1(c)). Since this estimation carries with it uncertainty due to ego-noise, the coarse resolution of SSL (5°), and the uncertainty of GPS and the IMU, a Kalman filter with data association is used to remove them.

As shown in Fig. 1(c), let z, ψ, θ, and ϕ be the observed altitude, yaw, azimuth, and elevation of the drone, respectively. Let $\boldsymbol{u} = (\sin(\psi - \theta), \cos(\psi - \theta))^T$ be the unit vector pointing to the sound source based on SSL, where T indicates the transpose operator. First, the sound source position, $\boldsymbol{x}_s = (x_s, y_s)^T$, is initially estimated using a planar terrain model as in our previous work [29] as follows:

$$\boldsymbol{x}_s \approx \boldsymbol{x}_h + \bar{z}_h \tan(\bar{\phi})\bar{\boldsymbol{u}} + \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{P})$, that is, $\boldsymbol{w}$ follows a normal distribution with mean of 0 and variance of $\boldsymbol{P}$ and measured values are denoted by $\bar{\cdot}$. Let $\boldsymbol{v} = (\cos(\psi - \theta), -\sin(\psi - \theta))^T$. $\boldsymbol{P}$ is defined as follows:

$$\boldsymbol{P} = \left((\sigma_z \tan(\bar{\phi})^2 + \left(\frac{\sigma_\phi \bar{z}_h}{\cos^2 \bar{\phi}} \right)^2 \right) \bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^T + (\sigma_\theta \bar{z}_h \tan(\bar{\theta})^T \bar{\boldsymbol{v}}\bar{\boldsymbol{v}}^T + \boldsymbol{\Sigma}_\varepsilon). \tag{2}$$

To cope with the uncertainty in the position estimation, an extended Kalman filter is used to get stable estimates [34].

C. Sound Source Identification (SSI)

The obtained signals are separated by GHDSS for SSI. GHDSS separates sound sources by using the directions obtained by SSL. A 50-dimensional mel-scale log spectrum (*MSLS*) [31] is extracted from the 0∼6,000 Hz band of the separated sound. The MSLS is obtained in the method of calculating the mel-frequency cepstral coefficient (*MFCC*) *without* applying a discrete cosine transform (DCT). Not applying a DCT is critical to improving the performance of ASR and SSI of contaminated sounds because DCT distributes local contamination in a frequency region to all frequency regions, which degrades the performance.

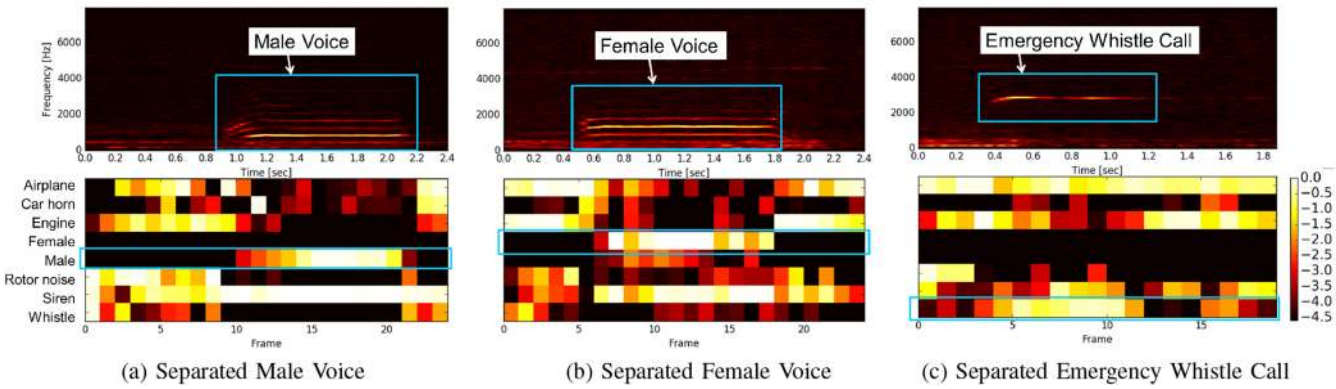
SSI is realized by using a set of SVMs to generate a feature vector instead of a label of a target sound. First, eight two-class SVMs are constructed to identify one of the following eight classes: male voice, female voice, emergency whistle, airplane, car horn, engine, siren, and rotor noise. Since we use AudioSet,[1] the above eight classes are chosen as rather independent bases. The design of more uncorrelated sound bases will be an important future work.

The training data are created by simulation with transfer functions generated mathematically. First, sound signals at a distance of 5 m from the drone are simulated, and ego-noise recorded from a real hovering flight test is added to the signal. Then, the captured sounds are separated by GHDSS with the specified direction, and a 50-D MSLS is calculated for each separated sound.

The multi-class classifier is realized with a set of SVMs. For a particular class, one SVM returns a positive label and the other SVMs return a negative label. Let $C_1, \ldots, C_8$ be the eight classes and l be the number of MSLS frames used for SSI. MSLS is represented by a vector $\boldsymbol{M}^T = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_l]$. The MSLS of the separated sound is applied to all SVM classifiers to obtain the class posterior probability $p_i = P(y = i|\boldsymbol{M})$. Here, the class posterior probability is treated as the likelihood $P(\boldsymbol{m}_i|C_j)$ of the observation of the class C_j. Thus, the *average log-likelihood (ALLD)* $s(C_j)$ of class C_j is obtained by the following equation.

$$s(C_i) = \frac{1}{l} \sum_{i=1}^{l} \log P(\boldsymbol{m}_i|C_j) \tag{3}$$

[1][Online]. Available: https://research.google.com/audioset/index.html



(a) Separated Male Voice (b) Separated Female Voice (c) Separated Emergency Whistle Call

Fig. 3. Spectrum and feature vectors of three sounds.

TABLE I
PERFORMANCE OF SSI ESTIMATED BY FIVE-FOLD CROSS-VALIDATION

Class	Accuracy
Male Voice	98.9 %
Female Voice	99.1 %
Whistle	93.3 %
Airplane	84.0 %
Car Horn	96.5 %
Engine	83.9 %
Siren	91.0 %
Rotor Noise	96.1 %

The resulting vector $\boldsymbol{s} = [s(C_1), \ldots, s(C_8)]^T$ represents the feature vector of the separated sound.

To evaluate the performance of each classifier and confirm that the set of eight base sound sources are mutually independent, k-fold cross-validation (k = 5) was performed. The results are shown in Table I. The cross-validation shows that every class is identified with an accuracy of over 80% and that the set can be used safely as a set of pseudo-orthogonal base sound sources.

An evaluation with open data [32] was conducted on male voices, female voices, and emergency whistle calls with an SNR of -15 dB. Fig. 3(a)–(c) depicts the spectrum in the upper part and a color map of the feature vector (ALLD calculated by Equation (3)) in the lower part for the separated sounds. The separated male voices had a high ALLD for male voices and low ALLD for emergency whistle calls, while the separated emergent whistle calls had a small ALLD for both male and female voices. Therefore, these three sounds were well discriminated.

IV. MULTIPLE SOUND SOURCE POSITION ESTIMATION BY DATA ASSOCIATION

A. Data Association for Sound Source Tracking

For tracking and estimating the position of multiple sound sources, data association is used to associate a tracked sound source with an observation. In this letter, we propose the *Global Nearest Neighbor with classification measurements (GNN-c)* based on GNN [21]. The idea of GNN-c is to incorporate information of sound source identification into geometrical criteria of data association.

1) Single Sound Source Correspondence: First, let us explain the case when only a single source is detected within a given range of the tracked target, which is called the *effective area*. When the observation j is within the given range of the tracked source i, i is updated using j.

The above range is computed using a geometric distance and feature-based distance as follows. The geometric distance at time step k, ${d^2_{k,ij}}^{(M)}$, is calculated in terms of the Mahalanobis distance between the tracked source i and the observation j as follows: ${d^2_{k,ij}}^{(M)} = \tilde{\boldsymbol{z}}^T_{k,ij} \boldsymbol{S}^{-1} \tilde{\boldsymbol{z}}_{k,ij}$ where $\tilde{\boldsymbol{z}}_{k,ij}$ is the observation error of the Kalman filter and $\boldsymbol{S}^{-1}$ is its co-variance matrix. The feature-based distance is measured by the following metric. Let $\boldsymbol{s}_{k,i}$ and $\boldsymbol{s}_{k,j}$ be feature vectors of the tracked source i and observation j, respectively. We define the Euclidean distance between these feature vectors as ${d^2_{k,ij}}^{(E)} = (\boldsymbol{s}_{k,i} - \boldsymbol{s}_{k,j})^T (\boldsymbol{s}_{k,i} - \boldsymbol{s}_{k,j})$. With these two distances, we in turn define the distance between the tracked sound source i and the observation j and the threshold, as follows:

$$d^2_{k,ij} = {d^2_{k,ij}}^{(M)} + w {d^2_{k,ij}}^{(E)} \tag{4}$$

$$G = G^{(M)} + w G^{(E)} \tag{5}$$

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 the two measures.

The value of $G^{(M)}$ is determined from a χ-squared distribution with two degrees of freedom, since the Mahalanobis distance follows a χ-squared distribution. Because $G^{(E)}$ depends on the classifier and the number of classes, its value is determined empirically. We set $G^{(E)} = 5.5$. Finally, if the condition,

$$d^2_{k,ij} < G \tag{6}$$

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

2) Multiple Sound Source Correspondence: Next, let us consider the case when multiple observations exist in the effective area of one tracked source, or when one observation is within the

effective areas of multiple tracked sources. In such situations, there remains ambiguity in the data association, which degrades the performance of the sound-source position estimation.

Suppose that n sources are being tracked at time k and m observations are obtained under the assumption of multiple sound sources. In addition to assigning the observation to known sources, it is also important to determine whether any new observation is either from the source currently being tracked or from a new source.

We verify for all combinations of tracked sources and observations whether Inequality (6) is satisfied and solve the assignment problem given by the cost matrix C below:

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1m} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nm} \end{bmatrix} \tag{7}$$

$$c_{ij} = \left\{ E \quad (d_{ij}^2 > G) \quad \text{or} \quad d_{ij}^2 \quad (\text{otherwise}) \right\} \tag{8}$$

where $E > G$ is a constant value to indicate a non-active state. With this cost matrix, we find a solution that minimizes the sum of distances in Equation (7) by combining the tracked sources and observations. We solved this problem by Munkres method [33]. Note that E in Equation (8) is set to a sufficiently large value.

B. Tracking Multiple Sound Sources

We introduce a management method for stable sound-source tracking.

1) 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 the Kalman filter starts tracking it. To avoid tracking a short fragment of a sound source, tracking continues only if consecutive observations exceed a continuous observation threshold N_1. If a tracked sound source is obtained a sufficient number of times, that is, more than the sound-source detection threshold N_2 ($N_2 > N_1$), the tracked sound source is labeled *valid*.

2) Maintenance of Multiple Sound Source Tracking: Sound source tracking continues while observations are being obtained, and the source is labeled *active*.

When an observation has not been obtained for a long time, for example, when no sound source has generated a signal, or the drone is too far from the sound source to capture any signal, we face the problem of a large accumulated uncertainty due to continuous expansion of the effective area. To avoid this problem, when an observation of a *valid* sound source that had been tracked for a sufficient time (N_3) has not been obtained for a period T_1, the Kalman filter stops tracking it; the source is labeled *dormant*. Then, the effective area is reinitialized to a constant radius r. If the source is observed again, the tracking resumes; the source is labeled *active* again (see Fig. 4).

3) Termination of Sound Source Tracking: If the number of observations is not less than N_1 and not more than N_3 and the sound source tracking is *dormant* during T_2, the tracking is terminated.

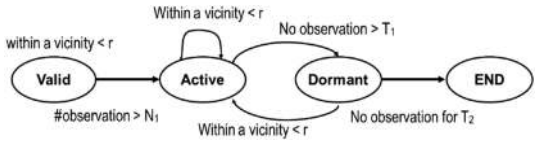


Fig. 4. State transition diagram for tracking.

V. Evaluation

A. Simulated Experiments

The performance of the proposed system was evaluated by comparing GNN-c with identification information and GNN without identification information while changing the power threshold of the MUSIC spectrum. The power threshold of the MUSIC spectrum was used to reduce the faulty detections of the sound source by discarding directions in which the MUSIC power is less than the threshold due to noise.

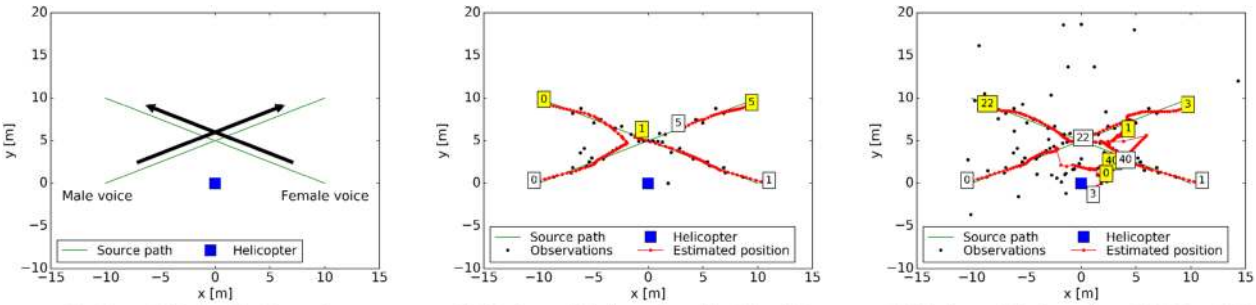
In the simulation, two voice sources (denoted as Male and Female) moved linearly at 0.5 m/s. Their two paths crossed in the middle (Fig. 5(a)), and the drone was located 5 m from the crossing point. Ego-noise recorded on actual flights was added to simulate the acoustic signal, whose SNR was about -15 dB during the utterances. The speech from the sources contained pauses, which lead to false detections because of the ego-noise. $G^{(M)}$ and $G^{(E)}$ were set to 5.991 and 5.5, respectively. The weight w in Equation (5) was set to 1.4 for the proposed method.

Preliminary tests confirmed that thresholds of 19.5 dB to 20.0 dB were suitable for detecting MUSIC spectrum peaks, as this range provided appropriate observations with fewer false positives.

The results of GNN without ID information are shown in Fig. 5(b) and (c). The figures show a top view of the field, including the detected source positions (black dots) and estimated sources with labels (red lines with numbers). For the case of a 20 dB threshold (Fig. 5(b)), there were three source estimates: Male was initially detected as ID 0, but it was estimated to be in the approaching-then-departing scenario; Female was detected as ID 1, but it was terminated in the middle; and a new estimate (ID 5) was generated in the second half. Relaxing the threshold to increase detections (Fig. 5(c), i.e., setting a 19.5 dB threshold) caused misestimations because of more clutter.

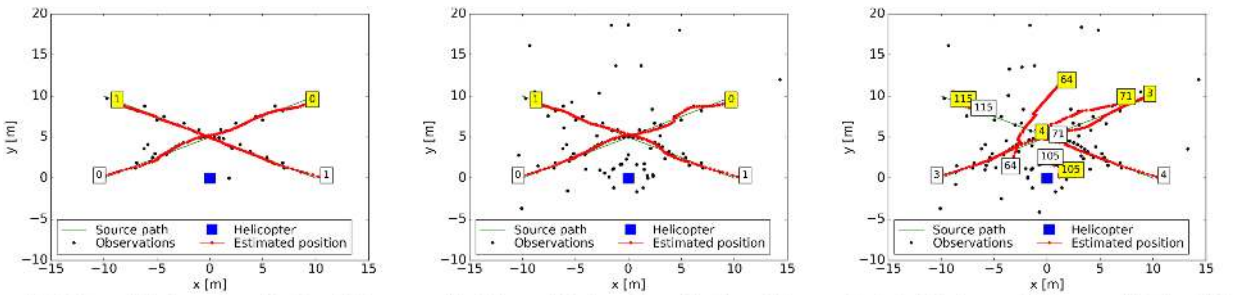
The proposed GNN-c with ID information worked appropriately for the 20 dB and 19.5 dB thresholds, as shown in Fig. 6(a) and (b), respectively. Even the proposed method failed to associate observations in the very noisy situation with the 19 dB threshold (Fig. 6(c)). In that case, Male was tracked properly, but Female was lost around the crossing point, and there were many faulty detections. The effectiveness of the proposed method can thus be seen by comparing Fig. 5(b), (c) and Fig. 6(a), (b).

The role of the feature vector can be explained by analyzing the case of the 19.5 dB threshold (Fig. 6(b)) where false positives existed. Feature vectors corresponding to the source candidates that were detected more than N_1 times are displayed in Fig. 7(a), where ID 0 and 1 correspond to the actual sources. As the figure shows, all the feature vectors are different from each other,



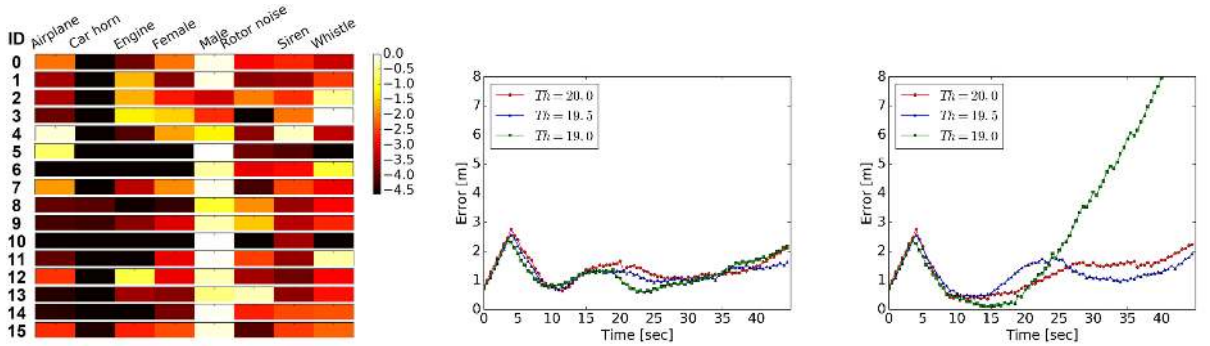
(a) Ground Truth of Trajectories (b) Estimated Trajectories with $Th = 20.0$ (c) Estimated Trajectories with $Th = 19.5$

Fig. 5. Ground truth and estimated trajectories by GNN without ID information.



(a) Estimated Trajectories with $Th = 20.0$ (b) Estimated Trajectories with $Th = 19.5$ (c) *Failed* Trajectory Estimates with $Th = 19.0$

Fig. 6. Trajectories estimated by GNN-c with ID information.



(a) Feature Vectors of ID 0–15 (b) Tracking Errors of Male Voice (ID0) (c) Tracking Errors of Female Voice (ID1)

Fig. 7. Feature vectors for ID0–15 and tracking errors.

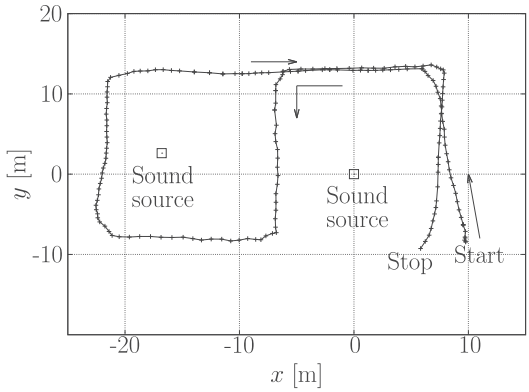
which prevented faulty observations from being integrated into the detected sources, and the system became robust.

The tracking error of the proposed method is summarized in Fig. 7(b) and (c). As the Kalman Filter had not converged in the initial stage of tracking, the error increased in the first several steps; the method managed to track the target within a 2 m range for the 20.0 dB and 19.5 dB thresholds once it had converged. Under noisy conditions (19.0 dB threshold), the system failed to track Female after 20 sec. when the error increased (Fig. 7(c)).

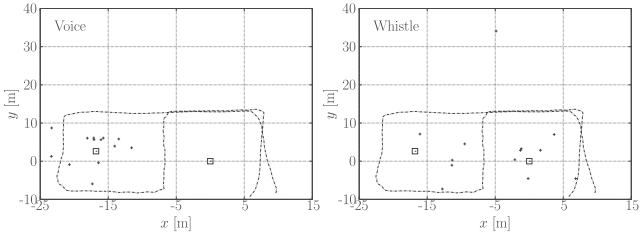
B. Field-Test Validation

Data obtained in the field test of our previous work [34] were utilized in a field experiment to validate the proposed approach. A drone with a 16-ch omnidirectional mic array (Fig. 1(a)) was commanded to fly a path over a field containing two static sound sources (left: voice, right: whistle), as shown in Fig. 8(a). The scenario was designed to demonstrate that the proposed approach was feasible and that the association with sound source classification would be effective for sound-source localization from the drone. Although the setup of the experiment was simpler than the numerical simulations as sound sources were static and located about 18 m away, it is valuable to test the system in the field because there were external noise sources including the reflection of the ego-noise from the ground.

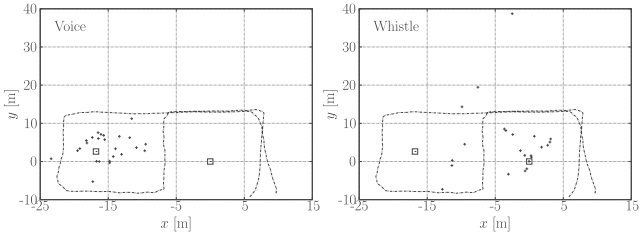
Fig. 8(b–1) and (b–2) show the observations utilized for the SSL based on the proposed algorithm (c–1) and that without acoustic signal classification (c–2). Fig. 8(c) shows the evolution of the estimation errors over time, and the final estimated positions are depicted in Fig. 8(d). As fewer observations



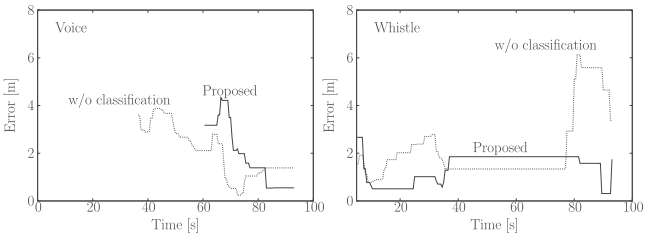
(a) Experiment setup



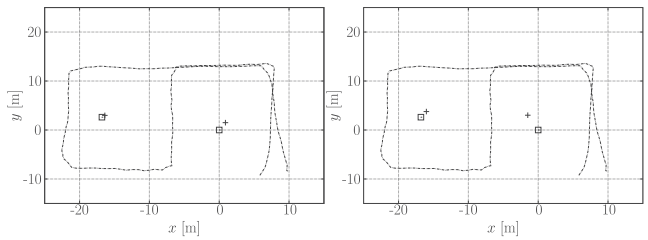
(b-1) Observations used by the proposed method



(b-2) Observations used by SSL without classification



(c) Estimation error



(d) Final estimates: proposed(left) and SSL w/o classification(right)

Fig. 8. Field test results. A loud speaker (voice) and a whistle tone were used as static sound sources emitting intermittent signals. The source at the origin generated beeps of a whistle while the other source emitted a human voice.

contributed to SSL with the proposed method comparing to the SSL without classification, the proposed method took longer time to find the left source (male voice), but the proposed method provided accurate and stable estimates of both two sources. The estimates by the SSL without classification deviated because of improper observations.

This result indicates that the proposed association method with classification made the localization robust even with a limited number of observations available under noisy conditions. This means it will be useful for practical applications, because it allows the auditory drone to localize targets without having to fly around them, which is how conventional approaches such as the one of Washizaki [35] cope with uncertain noisy observations.

Selection of a more restrictive threshold to filter the observations would suppress false observations and improve estimates of static sound sources even without classification, but it is worth emphasizing that the selection of such a threshold is sensitive and thus not always possible.

The sensor readings from the drone and acoustic signals were not perfectly synchronized because of the limitation of the computational resources. This imposes uncertainty in projecting the DoA on the ground, and it may affect the localization performance. Further investigation of such uncertainty is necessary although we experienced that the proposed method demonstrated better performance than the localization without association through multiple runs under different computational loads.

VI. Conclusion

We showed that acoustic information can be used to enhance sound source localization and tracking of multiple sound sources by drone audition and proposed data association between localization and identification of sound sources by GNN-c. The feature vectors are a set of average log-likelihoods obtained by multiple SVMs. Since multiple SVMs provide the probability of the source class with which to characterize *any* signal, this approach does not require the target signal *in advance*. Our claim is that identification of the target signals even in a significantly noisy environment helps localization. We expect that better signal identifiers such as DNN-based sound classifiers would improve performance, and, hence, it will be one of our future works to integrate recent machine learning techniques into our approach. An ongoing detailed evaluation involving more actual flights will be reported separately.

References

[1] 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., Berlin, Germany: Springer, 2019, pp. 3–22.

[2] K. Nonami *et al.*, "Recent R&D technologies & future prospective of flying robot in tough robotics challenges," *Disaster Robotics - Results from the ImPACT Tough Robotics Challenge*, S. Tadokoro Ed., Berlin, Germany: Springer, 2019, pp. 77–142.

[3] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proc. Conf. Amer. Assoc. Artif. Intell.*, 2000, pp. 832–839.

[4] H. G. Okuno and K. Nakadai, "Robot audition: Its rise & perspective," in *Proc. IEEE Int. Conf. Acoust., Speech Signal*, 2015, pp. 5610–5614.
[5] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design & implementation of robot audition system "HARK"," *Adv. Robot.*, vol. 24, no. 5-6, pp. 739–761, 2010.
[6] Y. Sasaki, S. Kagami, and H. Mizoguchi.,"Multiple Sound source mapping for a mobile robot by self-motion triangulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 380–385.
[7] M. Kumon and S. Uozumi., "Binaural localization for a mobile sound source," *J. Biomech. Sci. Eng.*, vol. 6, no. 1, pp. 26–39, 2011.
[8] E. Martinson and A. Schultz, "Auditory evidence grids," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 1139–1144.
[9] K. Sekiguchi, Y. Bando, K. Nakamura, K. Nakadai, K. Itoyama, and K. Yoshii, "Online simultaneous localization & mapping of multiple sound sources & asynchronous microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 1973–1979.
[10] C. Evers and P. A. Naylor, "Acoustic SLAM," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1484–1498, Sep. 2018.
[11] M. Basiri, F. S. Schill, P. U. Lima, and D. Floreano, "Robust acoustic source localization of emergency signals from micro air vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 4737–4742.
[12] L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 496–500.
[13] K. Hoshiba *et al.*, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, Nov. 2017, Art. no. E2535.
[14] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robot. Auton. Syst.*, vol. 96, pp. 184–210, Oct. 2017.
[15] R. Schmidt, "Multiple emitter location & signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
[16] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 3288–3293.
[17] T. Ohata, K. Nakamura, T. Mizumoto, T. Tezuka, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor embedded mmicrophone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5985–5990.
[18] K. Nakadai, K. Hidai, H. Mizoguchi, H.G. Okuno, and H. Kitano, "Real-time auditory & visual multiple-object tracking for robots," in *Proc. Int. Joint Conf. Artif. Intell.*, 2001, pp. 1425–1432.
[19] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4696–4704.
[20] S.H. Rezatofighi, A. Milan, Z. Zhang, Q. Shi, A. Dick, and I. Reid, "Joint probabilistic data association revisited," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 3047–3055.
[21] P. Konstantinova, A. Udvarev, and T. Semerdjiev, "A study of a target tracking algorithm using global nearest neighbor approach," in *Proc. CompSysTech*, 2003, pp. 290–295.
[22] J. Zegers and H. Van hamme, "Joint sound source separation & speaker recognition," in *Proc. Interspeech*, 2016, pp. 2228–2232.
[23] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1476–1485, Aug. 2010.
[24] M. F. Duarte and Y. H. Hu, "Vehicle classification in distributed sensor networks," *J. Parallel Distrib. Comput.*, vol. 64, no. 7, pp. 826–838, 2004.
[25] Q. Pan, J. Wei, H. Cao, N. Li, and H. Liu, "Improved DS acoustic-seismic modality fusion for ground-moving target classification in wireless sensor networks," *Pattern Recognit. Lett.*, vol. 28, no. 16, pp. 2419–2426, 2007.
[26] Q. Huang, T. Xing, and H. T. Liu, "Vehicle classification in wireless sensor networks based on rough neural network," in *Proc. Int. Symp. Neural Netw.*, 2006, pp. 58–65.
[27] Y. Kim, S. Jeong, D. Kim, and T. S. López, "An efficient scheme of target classification & information fusion in wireless sensor networks," *Pers. Ubiquitous Comput.*, vol. 13, no. 7, pp. 499–508, 2009.
[28] W. J. Roberts, H. W. Sabrin, and Y. Ephraim, "Ground vehicle classification using hidden Markov models," Defense Tech. Info. Center, Belcamp, MD, USA, DoD S&T Rep. ADA409368, 2001.
[29] K. Nakadai *et al.*, "Development of microphone-array-embedded uav for search and rescue task," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5985–5990.
[30] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Tech.*, vol. 2, no. 3, pp. 27.1–27.27, 2001.
[31] S. Yamamoto *et al.*, "Real-Time robot audition system that recognizes simultaneous speech in the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 5333–5338.
[32] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 776–780.
[33] J. Munkres, "Algorithms for the assignment & transportation problems," *J. Soc. Indust. Appl. Math.*, vol. 5, no. 1, pp. 32–38, 1957.
[34] M. Wakabayashi and M. Kumon, "Position estimation of multiple sound sources on the ground by multirotor helicopter with microphone array (in Japanese)," Japanese Soc. Artif. Intell., Tech. Rep. SIG-Chall-049-3, 2017, pp. 15–22.
[35] K. Washizaki, M. Wakabayashi, and M. Kumon, "Position estimation of sound source on ground by multirotor helicopter with microphone array," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 1980–1985.