2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Noise Correlation Matrix Estimation for Improving Sound Source Localization by Multirotor UAV

Koutarou Furukawa[1], Keita Okutani[2], Kohei Nagira[1], Takuma Otsuka[1], Katsutoshi Itoyama[1], Kazuhiro Nakadai[2,3] and Hiroshi G. Okuno[1]

***Abstract*— A method has been developed for improving sound source localization (SSL) using a microphone array from an unmanned aerial vehicle with multiple rotors, a "multirotor UAV". One of the main problems in SSL from a multirotor UAV is that the ego noise of the rotors on the UAV interferes with the audio observation and degrades the SSL performance. We employ a generalized eigenvalue decomposition-based multiple signal classification (GEVD-MUSIC) algorithm to reduce the effect of ego noise. While GEVD-MUSIC algorithm requires a noise correlation matrix corresponding to the auto-correlation of the multichannel observation of the rotor noise, the noise correlation is nonstationary due to the aerodynamic control of the UAV. Therefore, we need an adaptive estimation method of the noise correlation matrix for a robust SSL using GEVD-MUSIC algorithm. Our method uses a Gaussian process regression to estimate the noise correlation matrix in each time period from the measurements of self-monitoring sensors attached to the UAV such as the pitch-roll-yaw tilt angles, xyz speeds, and motor control values. Experiments compare our method with existing SSL methods in terms of precision and recall rates of SSL. The results demonstrate that our method outperforms existing methods, especially under high signal-to-noise-ratio conditions.**

I. Introduction

Multirotor unmanned aerial vehicles (UAV) are a useful and universal sensing platform because they have agility and mobility regardless of the terrain conditions, and of indoor or outdoor spaces. While recent research on multirotor UAV-based airborne sensing has focused on visual information [1]–[3], the visual modality is unsuitable for detecting hidden and/or overlapping objects. In the research reported here, we focused on the use of auditory information for sound source localization (SSL), i.e., the detection of sound sources with a microphone array and the estimation of the direction of arrival (DOA) of the target sound with an algorithm.

Auditory detection using a UAV may be possible even if obstacles hinder visual detection. Additionally, auditory sensing is useful for detecting people and animals because they emit vocal sounds to communicate with others. This means that auditory sensing is beneficial in search and rescue tasks and environmental monitoring and surveillance [4]–[7]. An example application is shown in Fig. 1.

[1] Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, 606-8501, Japan {kfurukaw,knagira,ohtsuka,itoyama,okuno}@kuis.kyoto-u.ac.jp

[2] Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Meguro-ku, Tokyo, 152-8552, Japan okutani@cyb.mei.titech.ac.jp

[3] Honda Research Institute Japan Co., Ltd., Wako, Saitama, 351-0114, Japan nakadai@jp.honda-ri.com

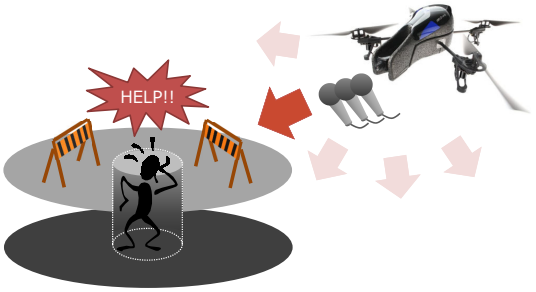


Fig. 1: Example application of SSL using a multirotor UAV. The best way to find a person who fell into a hole may be to detect the cries for help.

The main problem in using a multirotor UAV for SSL is that the ego noise degrades the SSL performance. This is because the ego noise has two characteristics in particular: (1) loudness and (2) nonstationarity. The ego noise of a multirotor UAV is generated mainly around the motors and rotors. Since the microphones have to be attached near these noise sources to prevent a loss of thrust, the noise power is high. The ego noise is nonstationary because the rotational speed of each motor dynamically changes in response to midair position control.

A generalized eigenvalue decomposition-based multiple signal classification (GEVD-MUSIC) algorithm [8] reduces the effect of ego noise. The GEVD-MUSIC algorithm uses a spatial correlation matrix of the noise component to cancel the noise signal during yielding the DOA spectrum. Since the noise correlation matrix is assumed to be stationary, adaptive estimation of the matrix is a critical issue for SSL under nonstationary ego noise conditions.

There remain several drawbacks in existing SSL methods using the GEVD-MUSIC algorithm, especially with respect to estimating noise correlation matrix. The iGEVD-MUSIC algorithm developed by Okutani et al. [9] regards the spatial correlation matrix of preceding observation as a noise correlation matrix. Since this algorithm is based on the assumption that the target sound changes more dynamically than noise, a stationary target sound might be incorrectly regarded as noise. Ince et al. [10] proposed an SSL method using a template database in order to suppress ego noise of a humanoid robot that stems from its joint motion. A template is composed of monitoring data from joint sensors and the spectrum of the ego noise. The noise corresponding to the joint sensor data that do not exist in the database may be incorrectly estimated because each template is found using a nearest neighbor search.

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

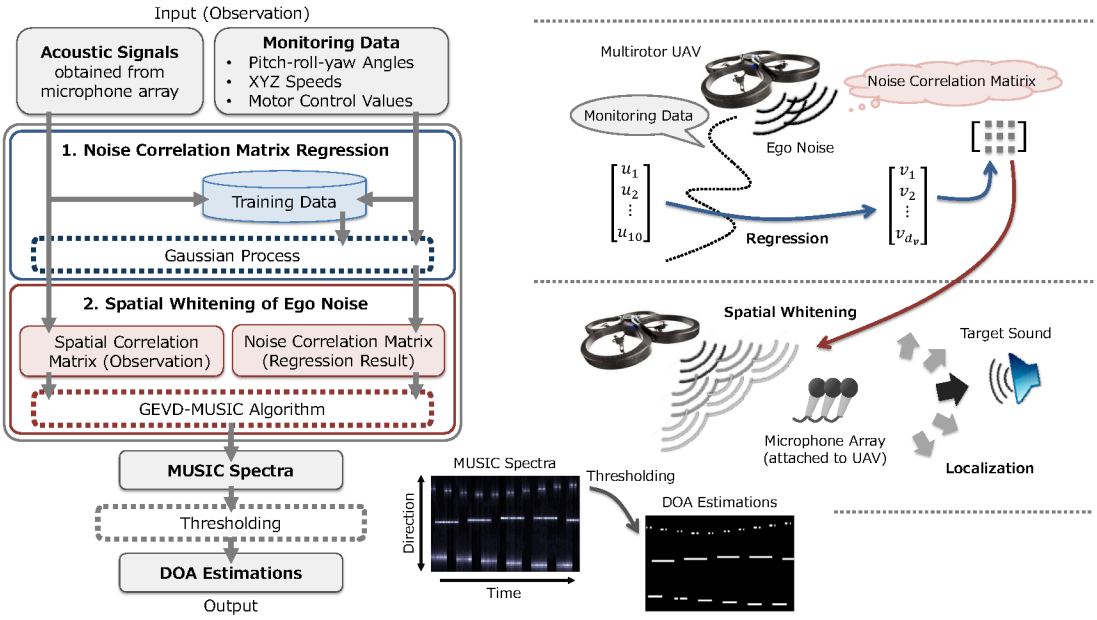


Fig. 2: Our SSL method has two parts: 1) regression of noise correlation matrix using a Gaussian process and 2) spatial whitening of ego noise using GEVD-MUSIC algorithm.

We propose a Gaussian process regression [11] of the noise correlation matrix with the data collected by self-monitoring sensors attached to a UAV body. Unlike iGEVD-MUSIC method, our method can prevent the mis-suppression of a stationary target sound because the matrix is estimated from the training data of noise unmixed with the target sound. Additionally, our model allows an interpolation of matrices for a robust noise correlation matrix estimation. This is important because a sensory measurement of the UAV cannot be exactly the same as those in the database.

Our SSL method consists of two parts, as shown in Fig. 2. The first part is regression of the noise correlation matrix using a Gaussian process. The other is an ego noise reduction with spatial whitening using a GEVD-MUSIC algorithm and the noise correlation matrix. We first describe how we use the Gaussian process regression to estimate the noise correlation matrices from the self-monitoring data. Then, we introduce an ordinary MUSIC algorithm, and extend it to the GEVD-MUSIC algorithm that uses the noise correlation matrices so as to suppress the effect of the noise signal.

II. Regression of Noise Correlation Matrix

We use a Gaussian process regression to estimate a noise correlation matrix because a complicated dependence between self-monitoring data and the matrix makes it difficult to obtain a parametric model. A Gaussian process [11], which is a stochastic process over continuous space, is used for regression because it describes the distribution over functions:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{1}$$

where m is the mean function, and k is the covariance function or kernel. The mean value of function $f(\mathbf{x})$ at location $\mathbf{x}$ is $m(\mathbf{x})$, and the covariance is $k(\mathbf{x}, \mathbf{x}')$, i.e., the distance between $\mathbf{x}$ and $\mathbf{x}'$ through the kernel. The parameters for the Gaussian process are learned from training data collected beforehand.

We use the self-monitoring data of a multirotor UAV, $\mathbf{u}_t$, as the features and the vectorized noise correlation matrix, $\mathbf{v}_{t,f}$, as the dependent variables:

$$\mathbf{u}_t = [u_{t,1}, \ldots, u_{t,10}]^{\mathrm{T}}, \tag{2}$$

$$\mathbf{v}_{t,f} = [v_{t,f,1}, \ldots, v_{t,f,M(M+1)/2}]^{\mathrm{T}}, \tag{3}$$

where subscript t is the time frame index, f is the frequency bin, and M is the number of microphones.

Each feature vector comprises the following elements:

- $u_{t,1}, \ldots, u_{t,3}$: pitch-roll-yaw angles,
- $u_{t,4}, \ldots, u_{t,6}$: xyz speeds,
- $u_{t,7}, \ldots, u_{t,10}$: pulse width modulation (PWM) values.

Pitch, roll, and yaw represent 3D rotation in the Euler angle coordinate system. Pitch is the up and down motion of the nose of the UAV. Roll is the left and right tilt, and yaw is the orientation of the head, as illustrated in Fig. 3, which also shows the XYZ axes. The PWM, a motor control value, is the ratio of the pulse width and pulse period of the motor input. As we used a quadrotor UAV, the PWM values had four elements.

The noise correlation matrix needs to be vectorized because a Gaussian process regression is unable to directly handle a matrix. Though a noise correlation matrix has Hermitian symmetry, this redundant structure is incompatible with the Gaussian process assumption that dependent variables have a Gaussian distribution. First, we perform Cholesky decomposition on noise correlation matrix $\mathbf{Q}_{t,f}$. We obtain an upper triangular matrix $\mathbf{L}_{t,f}$ to reduce redundancy:

$$\mathbf{Q}_{t,f} = \mathbf{L}_{t,f}^{\mathrm{H}} \mathbf{L}_{t,f}. \tag{4}$$

Fig. 3: Coordinate system of 3D rotation describing UAV's attitude.

Next we take the square roots of diagonal elements of $\mathbf{L}_{t,f}$ and concatenate the upper triangular elements of the column vectors into an $M(M+1)/2$-dimensional vector, $\mathbf{v}_{t,f}$. Taking the square roots ensures positive semi-definiteness of $\mathbf{Q}_{t,f}$. We assume that the mean of the prior distribution of $\mathbf{Q}_{t,f}$ is the identity matrix because using the identity matrix means the absence of directional noise in the GEVD-MUSIC algorithm. The relationship between $\mathbf{L}_{t,f}$ and $\mathbf{v}_{t,f}$ is given by

$$\mathbf{L}_{t,f} = \begin{bmatrix} v_{t,f,1}^2 & v_{t,f,2} & \cdots & v_{t,f,d_\mathbf{v}-(M-1)} \\ & v_{t,f,3}^2 & \cdots & v_{t,f,d_\mathbf{v}-(M-2)} \\ & & \ddots & \vdots \\ \mathbf{0} & & & v_{t,f,d_\mathbf{v}}^2 \end{bmatrix} - \mathbf{I}_M, \quad (5)$$

where $d_\mathbf{v} = M(M+1)/2$.

We compute as the regression result the mean of the conditional distribution of the dependent variable $\mathbf{v}_{T+1,f}$ given the training data $\mathcal{D}$ and a newly observed feature $\mathbf{u}_{T+1}$:

$$p(\mathbf{v}_{T+1}|\mathcal{D}, \mathbf{u}_{T+1}). \quad (6)$$

The training data set $\mathcal{D}$ is $\{(\mathbf{u}_t, \mathbf{v}_{t,f})\}_{t=1,\ldots,T}$, which is collected beforehand in the absence of target sound sources. We obtain the mean, $m(\mathbf{v}_{T+1,f})$, as

$$m(\mathbf{v}_{T+1,f}) = \mathbf{k}_T^\mathrm{T} \left(\mathbf{K}_T + \rho^2 \mathbf{I}_T\right)^{-1} \mathbf{V}_{T,f}^\mathrm{T}, \quad (7)$$

$$\mathbf{V}_{T,f} = [\mathbf{v}_{1,f}, \ldots, \mathbf{v}_{T,f}], \quad (8)$$

where ρ^2 is the variance of the additive noise in the feature vectors. In (7), $\mathbf{k}_T$ and $\mathbf{K}_T$ are given by

$$\mathbf{k}_T = \begin{bmatrix} k_{1,T+1} \\ \vdots \\ k_{T,T+1} \end{bmatrix}, \quad \mathbf{K}_T = \begin{bmatrix} k_{1,1} & \cdots & k_{1,T} \\ \vdots & \ddots & \vdots \\ k_{T,1} & \cdots & k_{T,T} \end{bmatrix}. \quad (9)$$

We abbreviate $k(\mathbf{u}_i, \mathbf{u}_j)$ as $k_{i,j}$. $\mathbf{K}_T$ is a Gramian matrix. We use with a little change the Mahalanobis kernel [12], which is based on a radial basis function kernel:

$$k(\mathbf{u}_i, \mathbf{u}_j) = \exp\left(-\frac{\gamma(\mathbf{u}_i - \mathbf{u}_j)^\mathrm{T} \boldsymbol{\Sigma}^{-1} (\mathbf{u}_i - \mathbf{u}_j)}{\dim(\mathbf{u}_i)}\right), \quad (10)$$

where $\boldsymbol{\Sigma}^{-1}$ is a matrix that has variances on the diagonal. We normalize the scale of each element of the feature vectors by using a Mahalanobis kernel.

III. MUSIC and GEVD-MUSIC Algorithms

The multiple signal classification (MUSIC) algorithm [13] is a subspace-based DOA estimation algorithm. It decomposes an observed noisy signal into the signal subspace and noise subspace to obtain the spatial spectrum of DOA, the MUSIC spectrum. A MUSIC spectrum has peak values corresponding to the directions from which sounds are coming. The noise used here is diffuse noise; directional noise might create peaks in the MUSIC spectrum. To remove any peaks created by directional noise, we use a GEVD-MUSIC algorithm [8] with additional information for noise: a noise correlation matrix. Let's take a look at these algorithms in detail.

We assume the situation is one in which we observe N target sounds $\mathbf{z}_{t,f}$ with M microphones. We consider the observed signal to be a time-frequency representation, $\mathbf{x}_{t,f}$, which is obtained by short-time Fourier transform (STFT):

$$\mathbf{x}_{t,f} = [x_{t,f,1}, \ldots, x_{t,f,M}]^\mathrm{T}, \quad (11)$$

$$\mathbf{z}_{t,f} = [z_{t,f,1}, \ldots, z_{t,f,N}]^\mathrm{T}. \quad (12)$$

We can write observed noisy signal $\mathbf{x}_{t,f}$ with additive diffuse noise $\mathbf{n}_{t,f}$ as

$$\mathbf{x}_{t,f} = \mathbf{A}_f \mathbf{z}_{t,f} + \mathbf{n}_{t,f}. \quad (13)$$

Let $\mathbf{A}$ be an $M \times N$ steering matrix containing prior knowledge of the microphone array geometry:

$$\mathbf{A}_f = [\mathbf{a}_{f,\theta_1}, \ldots, \mathbf{a}_{f,\theta_N}]. \quad (14)$$

Each column vector of the steering matrix is called a steering vector and describes the process of signal arrival from a particular direction. The spatial correlation matrix of the observed signal, $\mathbf{R}_{t,f}$, is given by

$$\mathbf{R}_{t,f} = \mathbb{E}\left[\mathbf{x}_{t,f}\mathbf{x}_{t,f}^\mathrm{H}\right] = \mathbf{S}_{t,f} + \sigma_w^2 \mathbf{I}_M, \quad (15)$$

$$\mathbf{S}_{t,f} = \mathbf{A}_f \mathbf{z}_{t,f} \mathbf{z}_{t,f}^\mathrm{H} \mathbf{A}_f^\mathrm{H}, \quad (16)$$

where σ_w^2 is the variance of the noise, and $\mathbf{I}_M$ is the $M \times M$ identity matrix. The superscript H is the adjoint operator. The $\mathbb{E}[\mathbf{x}_{t,f}\mathbf{x}_{t,f}^\mathrm{H}]$ is calculated as the time average of $\mathbf{x}_{t,f}\mathbf{x}_{t,f}^\mathrm{H}$.

Let $\lambda_{\mathbf{R},i}$ be the i-th largest eigenvalue of matrix $\mathbf{R}_{t,f}$, and $\mathbf{e}_{\mathbf{R},i}$ be the corresponding eigenvector. The eigenspace of $\mathbf{R}_{t,f}$ is decomposed into two orthogonal subspaces: the signal subspace $\mathrm{span}(\mathbf{e}_{\mathbf{R},1}, \ldots, \mathbf{e}_{\mathbf{R},N})$ and the noise subspace $\mathrm{span}(\mathbf{e}_{\mathbf{R},N+1}, \ldots, \mathbf{e}_{\mathbf{R},M})$. The latter is written as

$$\mathrm{span}(\mathbf{e}_{\mathbf{S},N+1}, \ldots, \mathbf{e}_{\mathbf{S},M}) = \mathrm{span}(\mathbf{a}_{f,\theta_1}, \ldots, \mathbf{a}_{f,\theta_N})^\perp. \quad (17)$$

We define the spatial spectrum, $\mathbf{p}_{t,f}$, based on the orthogonality of these subspaces as

$$\mathbf{p}_{t,f} = [p_{t,f,\theta_1}, \ldots, p_{t,f,\theta_A}]^\mathrm{T} \quad (18)$$

$$p_{t,f,\theta_i} = \frac{\|\mathbf{a}_{f,\theta_i}^\mathrm{H} \mathbf{a}_{f,\theta_i}\|}{\sum_{i=N+1}^{M} |\mathbf{a}_{f,\theta_i}^\mathrm{H} \mathbf{e}_{\mathbf{R},i}|^2}, \quad (19)$$

where A is the number of steering vectors. In this definition, as the denominator reaches zero when the direction of the steering vector and the DOA of the target sound are the same, we obtain the peak value in the spectrum. Since target sounds ordinarily cover a wide frequency range, we use the weighted sum of $\mathbf{p}_{t,f}$ in a certain frequency band:

$$\mathbf{p}_t = \sum_f w_{t,f} \mathbf{p}_{t,f}, \quad (20)$$

where $w_{t,f}$ is a weight, which is typically the principal eigenvalue $\lambda_{\mathbf{R},1}$.

In the MUSIC algorithm, directional noise can be incorrectly identified as the target sound because noise is assumed to be relatively small and diffuse. To enable target sounds to be distinguished from directional noise, we use a GEVD-MUSIC algorithm. Instead of standard eigenvalue decomposition (SEVD) as used in the MUSIC algorithm, generalized eigenvalue decomposition (GEVD) of $\mathbf{R}_{t,f}$ is used with noise correlation matrix $\mathbf{Q}_{t,f}$:

$$\mathbf{R}_{t,f}\mathbf{e}'_{\mathbf{R},i} = \lambda'_{\mathbf{R},i}\mathbf{Q}_{t,f}\mathbf{e}'_{\mathbf{R},i}, \tag{21}$$

where $\lambda'_{\mathbf{R},i}$ is the i-th largest generalized eigenvalue of matrix $\mathbf{R}$, and $\mathbf{e}'_{\mathbf{R},i}$ is the corresponding generalized eigenvector. In most cases, since $\mathbf{Q}_{t,f}$ is Hermitian and positive definite, $\mathbf{Q}_{t,f}$ is decomposed as

$$\mathbf{Q}_{t,f} = \mathbf{\Phi}^{\mathrm{H}}_{t,f}\mathbf{\Phi}_{t,f}. \tag{22}$$

Equation (21) is identical to the eigenvalue decomposition of $\mathbf{R}'_{t,f}$ after spatial whitening of directional noise using $\mathbf{\Phi}_{t,f}$.

$$\mathbf{R}'_{t,f} = \mathbf{\Phi}^{-\mathrm{H}}_{t,f}\mathbf{R}_{t,f}\mathbf{\Phi}^{-1}_{t,f} = \sum_{i=1}^{M}\lambda_{\mathbf{R}',i}\mathbf{e}_{\mathbf{R}',i}\mathbf{e}^{\mathrm{H}}_{\mathbf{R}',i}. \tag{23}$$

The way this whitening works is as follows.

When noise $\mathbf{n}_{t,f}$ in (13) is directional, spatial correlation matrix $\mathbf{R}_{t,f}$ is obtained as

$$\mathbf{R}_{t,f} = \mathbf{S}_{t,f} + \mathbf{Q}_{t,f} \tag{24}$$

instead of as shown in (15). Substitution of (24) into (23) yields

$$\mathbf{R}'_{t,f} = \mathbf{\Phi}^{-\mathrm{H}}_{t,f}\mathbf{S}_{t,f}\mathbf{\Phi}^{-1}_{t,f} + \mathbf{I}_M. \tag{25}$$

Since the rank of the first term, $\mathbf{\Phi}^{-\mathrm{H}}_{t,f}\mathbf{S}_{t,f}\mathbf{\Phi}^{-1}_{t,f}$, is N, we obtain the spatial spectrum using generalized eigenvectors $\mathbf{e}'_{\mathbf{R},i}$ in a manner similar to that discussed above. Thus, we verify the spatial whitening using GEVD.

IV. EVALUATION

We conducted several experiments to evaluate the performance of our SSL method. We used acoustic signal and monitoring data as input data, and we obtained MUSIC spectra and DOA estimations as output.

A. SSL System Construction

We constructed the SSL system as shown in Fig. 4. We used an AR.Drone[1] as the multirotor UAV. It has several kinds of built-in sensors: a 6-degrees of freedom inertial measurement unit, ultrasound telemeters, and cameras for ground speed measurement. The data collected by these sensors are called navigation data, which is abbreviated as *navdata*. We equipped the AR.Drone with a microphone array and a RASP-24[2] signal processing unit, as shown in Fig. 5. The microphone array had eight MEMS microphones

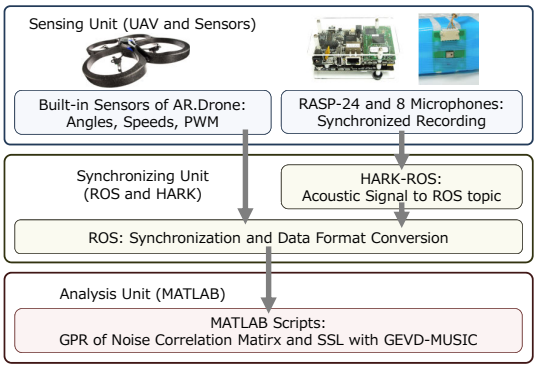


Fig. 4: SSL system used for evaluation.

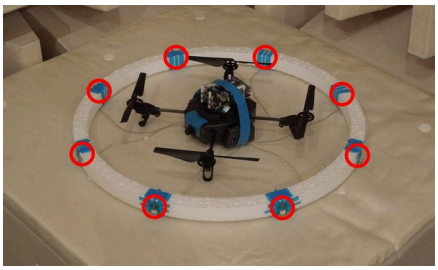


Fig. 5: Sensing unit consisting of an AR.Drone, a RASP-24 signal processing unit and a microphone array, which had eight microphones at the locations marked by red circles.

facing outward that were equally spaced in a circular framework. We reduced the weight of the AR.Drone by removing unneeded components because it originally lacked the ability to carry the signal processing unit and microphone array.

We used as the recording system the HRI-JP Audition for Robots with Kyoto University (HARK)[3] and the Robot Operating System (ROS)[4]. HARK is a collection of modules for robot audition that enabled us to publish a ROS topic containing a multichannel acoustic signal recorded using the signal processing unit. ROS is used as a platform for operating many kinds of robots. Using this software, we collected the acoustic signals corresponding to the navdata.

B. Experimental Conditions

We recorded the noise of the AR.Drone for approximately 200 s during hovering and 400 s during moving in an anechoic chamber. We used one-fifth of each flight for test data and the rest for training data. The test data, which contained the target sound and noise, was made by using a simulation mixture. As prior knowledge of the microphone array geometry, we computed 72 steering vectors using a time-stretched pulse response.

The sampling rate of the acoustic signals was 16,000 Hz, and the navdata was obtained per 60 ms on average. As we obtained a time frame of the acoustic signals per 16 ms after STFT, the navdata were linearly interpolated to fill the gaps of these sampling rates. The STFT frame length was 512 samples, and the shift length was 256 samples. We used a Hann window.

[1] ardrone.parrot.com
[2] www.sifi.co.jp
[3] winnie.kuis.kyoto-u.ac.jp/HARK/
[4] www.ros.org

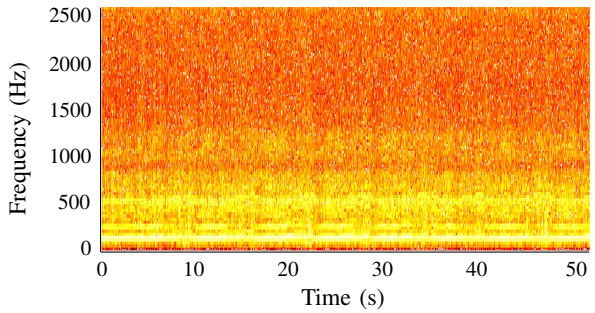


Fig. 6: Spectrogram of noise of AR.Drone during flight. The noise had higher energy in the low-frequency zone, and it peaked in several frequency bins.

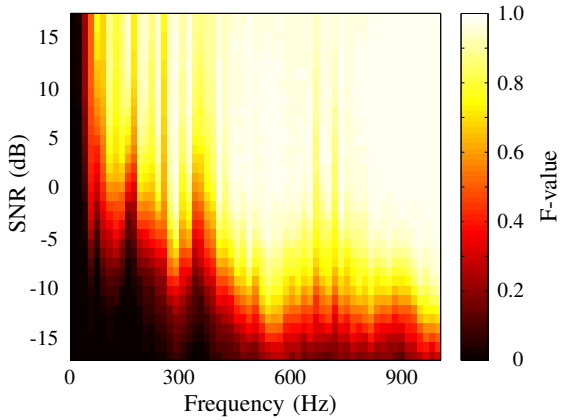


Fig. 7: F-values at equilibrium points under each condition of SNR and target sound frequency.

The spectrogram of the AR.Drone noise shown in Fig. 6 revealed that the noise energy was unevenly distributed across the frequency zones and tended to be concentrated in the low-frequency zone.

C. Frequency Response

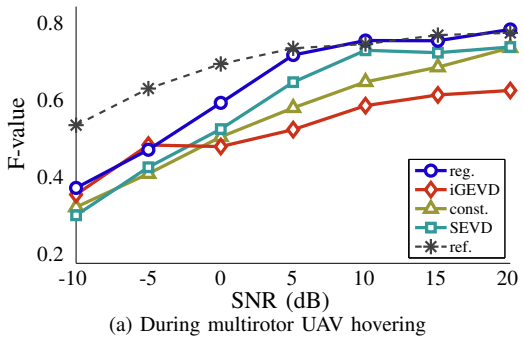
First we evaluated the SSL performance by changing the signal to noise ratio (SNR) and the frequency of the target sound, which was assumed to be a pure tone. The performance was evaluated by using the F-values at the equilibrium point, which is defined as the point where precision equaled recall. The precision and recall of the MUSIC spectra $\mathbf{P}$, which has $\mathbf{p}_t$ as column vectors, were calculated using

$$\mathrm{Pre}(\mathbf{P}) = \frac{\#\{(t,\theta) \mid p_{t,\theta} \geq \xi \text{ and } p'_{t,\theta} = 1\}}{\#\{(t,\theta) \mid p_{t,\theta} \geq \xi\}}, \quad (26)$$

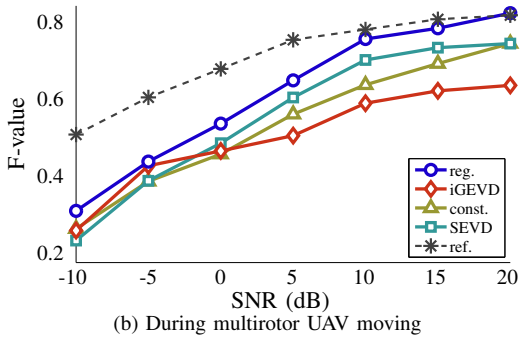
$$\mathrm{Rec}(\mathbf{P}) = \frac{\#\{(t,\theta) \mid p_{t,\theta} \geq \xi \text{ and } p'_{t,\theta} = 1\}}{\#\{(t,\theta) \mid p'_{t,\theta} = 1\}}, \quad (27)$$

where ξ is the threshold for $\mathbf{P}$, and $p'_{t,\theta}$ is an element of a reference spatial spectrum that has 1 in the correct DOA of each target sound. The # denotes the number of elements in a set.

Fig. 7 shows that the lower the SNR, the more difficult it was to detect a target sound with a low frequency. This result agrees with the spectrogram of the noise shown in Fig. 6 and thus suggests that noise likely masked target sounds that had a low frequency.

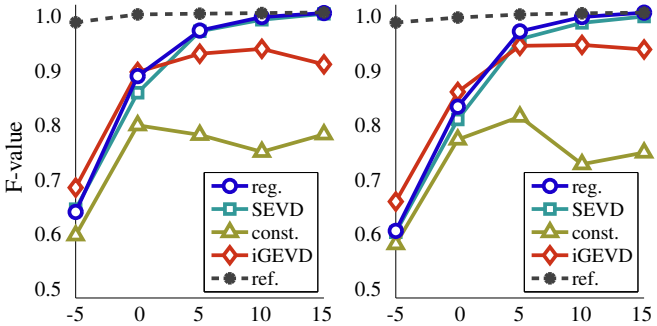


(a) During multirotor UAV hovering



(b) During multirotor UAV moving

Fig. 8: F-values of equilibrium points under various SNR conditions.



(a) Hovering. (b) Moving.

Fig. 9: F-values of DOA estimations by VBHMM-based thresholding under various SNR conditions.

D. Performance with Simulated Data

We experimentally compared the performance of our method with those of the existing methods. We generated two sets of test data by simulation using both the hovering noise and the moving noise. These sets contained three kinds of target sound data: human speech, pure tone, and white signal. Here, *white* means having a constant power spectral density in the frequency domain, not spatially. These target sounds arrived from different directions repeatedly. We compared our method to three existing methods. One uses an ordinary MUSIC algorithm, without spatial whitening of the noise. One uses the GEVD-MUSIC algorithm with a constant noise correlation matrix, which is the time average of the test data. The other uses the iGEVD-MUSIC algorithm, which regards the preceding observation as noise.

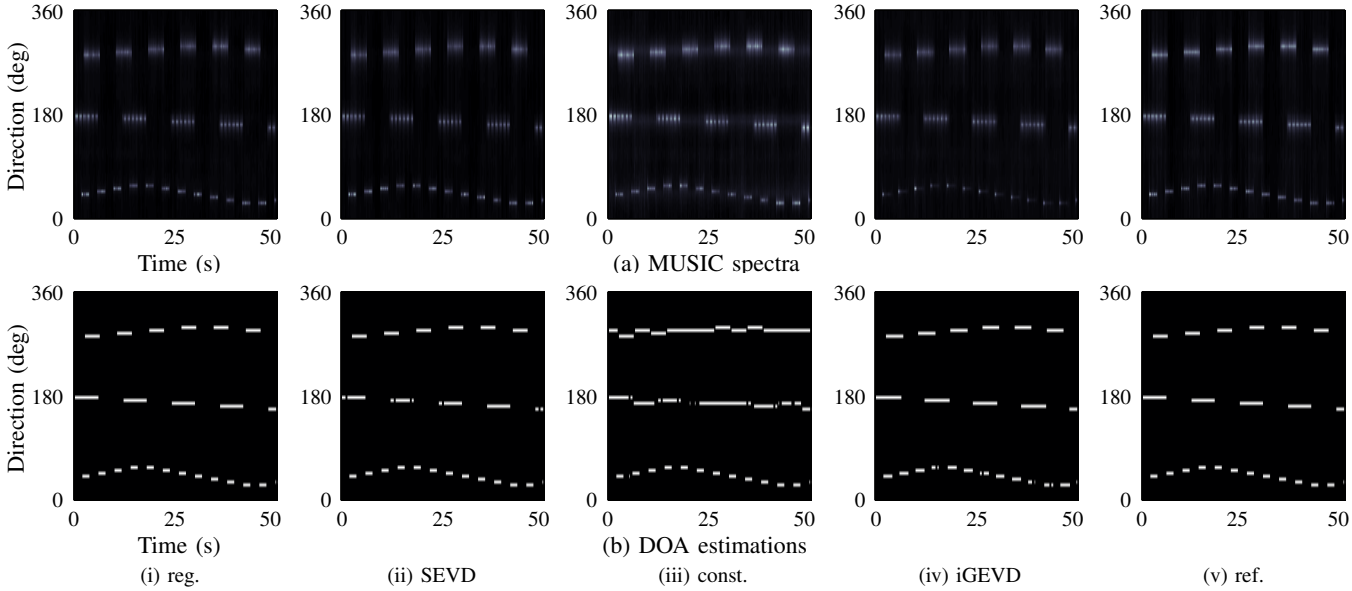


Fig. 10: MUSIC spectra and DOA estimations for each method of constructing noise correlation matrix. ("reg." denotes proposed method, "SEVD" denotes method using ordinary MUSIC, "const." denotes method using GEVD-MUSIC with constant noise correlation matrix, "iGEVD" denotes method using iGEVD-MUSIC, and "ref." denotes method using correct noise correlation matrix.)

We evaluated the performance on the basis of two criteria: the F-value of the equilibrium points described above and the F-value using variational Bayesian hidden Markov model (VBHMM)-based thresholding [14]. These results are shown in Figs. 8 and 9. The DOA estimations are obtained from the MUSIC spectra (Fig. 10).

Fig. 8 shows that our method created clearer peaks in the MUSIC spectra than the other methods. It is reasonable to suppose that under high SNR conditions, the other methods falsely suppress target sound components due to using incorrect noise correlation matrices. The F-values obtained by VBHMM-based thresholding show that our method slightly increases the number of correct DOA estimations under high SNR conditions (Fig. 9).

V. CONCLUSION

We have developed a method that improves SSL using a multirotor UAV equipped with a microphone array. The problem with a multirotor UAV is nonstationary ego noise emitted during its flight. Our method uses Gaussian process regression of the noise correlation matrix along with data collected by self-monitoring sensors. The regression result of the noise correlation matrix used in a GEVD-MUSIC algorithm as additional information on directional high-power noise.

Experimental results demonstrated that our method improves SSL performance, especially under high SNR conditions. Future work includes improving the accuracy of regression by optimizing feature selection, increasing the training data set, and evaluating SSL performance using real-world data.

VI. ACKNOWLEDGMENTS

This research was partially supported by JSPS Grant-in-Aid for Scientific Research (S) No. 24220006.

REFERENCES

[1] L. Meier, P. Tanskanen, F. Fraundorfer, and M. Pollefeys, "PIXHAWK: A system for autonomous flight using onboard computer vision," in *Proc. of IEEE ICRA*, 2011, pp. 2992–2997.

[2] M. W. Achtelik, S. Lynen, S. Weiss, L. Kneip, M. Chli, and R. Siegwart, "Visual-Inertial SLAM for a Small Helicopter in Large Outdoor Environments," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 2651–2652.

[3] A. Natraj, P. Sturm, C. Demonceaux, and P. Vasseur, "A Geometrical Approach For Vision Based Attitude And Altitude Estimation For UAVs In Dark Environments," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 4565–4570.

[4] M. Basiri, F. Schill, P. U. Lima, and F. Dario, "Robust Acoustic Source Localization of Emergency Signals from Micro Air Vehicle," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 4737–4742.

[5] H. Yoshinaga, K. Mizutani, Wakatsuki, and Naoto, "A sound source localization technique to support search and rescue in loud noise environments," vol. 67, pp. 11–16, 2012.

[6] H. Sun, P. Yang, L. Zu, and Q. Xu, "A Far Field Sound Source Localization System for Rescue Robot," in *Proc. of IEEE CASE*, 2011, pp. 1–4.

[7] S. Kimura, T. Akamatsu, K. Wang, D. Wang, S. Li, S. Dong, and N. Arai, "Comparison of stationary acoustic monitoring and visual observation of finless porpoises," vol. 125, pp. 547–553, 2009.

[8] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *Proc. of IEEE/RSJ IROS*, 2009, pp. 664–669.

[9] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor Auditory Scene Analysis Using a Moving Microphone Array Embedded in a Quadrocopter," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 3288–3293.

[10] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of general applicability of ego noise estimation," in *Proc. of IEEE ICRA*, 2011, pp. 3517–3522.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006.

[12] S. Abe, "Training of support vector machines with Mahalanobis kernels," in *Artificial Neural Networks: Formal Models and Their Applications (ICANN 2005)*, vol. 3697, 2005, pp. 571–576.

[13] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[14] T. Otsuka, K. Nakadai, T. Ogata, and H. G. Okuno, "Bayesian Extension of MUSIC for Sound Source Localization and Tracking," in *Proc. of Interspeech*, 2011, pp. 3109–3112.