https://doi.org/10.20965/jrm.2017.p0188

Paper:

Outdoor Acoustic Event Identification with DNN Using a Quadrotor-Embedded Microphone Array

Osamu Sugiyama[*1], **Satoshi Uemura**[*2], **Akihide Nagamine**[*3], **Ryosuke Kojima**[*2], **Keisuke Nakamura**[*4], **and Kazuhiro Nakadai**[*2,*4]

[*1]Preemptive Medicine & Lifestyle-Related Disease Research Center, Kyoto University Hospital
54 Kawaharacho, Syogoin, Sakyo-ku, Kyoto City 606-8507, Japan
E-mail: sugiyama@kuhp.kyoto-u.ac.jp
[*2]Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan
[*3]Department of Electrical and Electronic Engineering, School of Engineering, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan
[*4]Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako, Saitama 351-0188, Japan
[Received July 25, 2016; accepted December 27, 2016]

This paper addresses *Acoustic Event Identification (AEI)* of acoustic signals observed with a microphone array embedded in a quadrotor that is flying in a noisy outdoor environment. In such an environment, noise generated by rotors, wind, and other sound sources is a big problem. To solve this, we propose the use of a combination of two approaches that have recently been introduced: *Sound Source Separation (SSS)* and *Sound Source Identification (SSI)*. SSS improves the *Signal-to-Noise Ratio (SNR)* of the input sound, and SSI is then performed on the SNR-improved sound. Two SSS methods are investigated. One is a single channel algorithm, *Robust Principal Component Analysis (RPCA)*, and the other is *Geometric High-order Decorrelation-based Source Separation (GHDSS-AS)*, known as a multichannel method. For SSI, we investigate two types of deep neural networks namely *Stacked denoising Autoencoder (SdA)* and *Convolutional Neural Network (CNN)*, which have been extensively studied as highly-performant approaches in the fields of automatic speech recognition and visual object recognition. Preliminary experiments have showed the effectiveness of the proposed approaches, a combination of GHDSS-AS and CNN in particular. This combination correctly identified over 80% of sounds in an 8-class sound classification recorded by a hovering quadrotor. In addition, the CNN identifier that was implemented could be handled even with a low-end CPU by measuring the prediction time.

Keywords: robot audition, sound source localization, sound source separation, sound source identification, unmanned aerial vehicle

1. Introduction

Outdoor scene analysis is an essential research topic in robotics. In vision, a large number of outdoor scene analysis studies have been done, because high-performance, robust sensors such as a camera *LIght Detection And Ranging (LIDAR)*, and the *Global Positioning System (GPS)* have become available. The technologies for visual scene analysis have had many applications, such as autonomous cars [1]. In auditory processing, robot audition has been studied for more than a decade. It focuses mainly on an indoor environment for human-robot interaction, and only a few studies have been conducted on *Outdoor Computational Auditory Scene Analysis (OCASA)*. In such a situation, OCASA with an *Unmanned Aerial Vehicle (UAV)* has begun to receive much attention in the past few years, since it assists the location of people in a disastrous situation. OCASA is made up of two technologies, *Acoustic Event Detection (AED)* and *Acoustic Event Identification (AEI)*.

AED can extract "where" and "when" information, performing sound source localization and sound activity detection. Okutani et al. reported on AED using a Parrot AR.Drone by installing an 8 ch microphone array that consisted of a small and lightweight microphone and a multichannel A/D converter [2]. They proposed *MUltiple SIgnal Classification based on incremental Generalized EigenValue Decomposition (iGEVD-MUSIC)* [3]. The iGEVD-MUSIC achieved noise-robust AED by incrementally whitening high-power noise generated by the rotation of propellers and wind. Furukawa et al. extended their method, in particular, the performance of noise correlation matrix estimation using motion information obtained from *Inertial Measurement Unit (IMU)* [4]. Ohata et al. focused on the computational cost of their method, achieving real-time processing by proposing *MUltiple SIgnal Classification based on incremental Generalized*

Singular Value Decomposition with Correlation Matrix Scaling (iGSVD-MUSIC-CMS) [5].

Although iGSVD-MUSIC-CMS is able to detect a sound source 10–20 m away from a UAV in real time, they did not deal with the issue of ascertaining the type of the sound source detected. To find people in a disastrous situation, the system has to distinguish speech sources from the other sound sources detected, which can be solved by AEI. Unlike AED, AEI has not been studied, including in an outdoor environment. Since a sound signal captured with a UAV-embedded microphone array is heavily contaminated with noise, we take two approaches for AEI: *Sound Source Separation (SSS)* and *Sound Source Identification (SSI)*. SSS extracts the target sound source from the mixture of sound sources, which has a low *Signal-to-Noise Ratio (SNR)*. Although there are many algorithms, for SSS, we investigate two methods. One is a single channel algorithm called *Robust Principal Component Analysis (RPCA)*, and the other is *Geometric High-order Decorrelation-based Source Separation with Adaptive Stepsize control (GHDSS-AS)* known as a multichannel method. Since SSS improves the SNR of the input sound, the performance of SSI is expected to improve. It is inevitable that the separated sound contains distortions to some extent because SSS is an ill-posed problem. This means that we have to select an SSI method with high performance. For SSI, we investigate two types of *Deep Neural Networks (DNNs)*, namely *Stacked denoising Autoencoder (SdA)* and *Convolutional Neural Network (CNN)*, which have been extensively studied as highly-performing approaches in the fields of automatic speech recognition and visual object recognition, and recently also in acoustic event detection [6–9]. We construct a prototype system for OCASA by integrating AEI by combining SSS and SSI, and AED based on iGSVD-MUSIC-CMS. The combinations of two SSS methods and two SSI methods are evaluated using real acoustic signals recorded with a UAV that has a 16 ch circular microphone array.

The rest of this paper is organized as follows: Section 2 introduces two techniques for AEI, namely SSS and SSI. Section 3 illustrates AEI system architecture using SSS and SSI. Section 4 evaluates the system, and the last section presents our conclusions.

2. Acoustic Event Identification

The proposed AEI, which consists of SSS and SSI, will now be explained.

For SSS, GHDSS-AS and RPCA are selected. GHDSS-AS is selected as high-performance multichannel noise robust sound source separation method [10]. On the other hand, RPCA is selected as high-performance blind noise robust sound source separation method [11]. For SSI, SdA and CNN which are well-known methods in DNN are introduced. SdA is selected for training the feed-forward neural network with a limited data set; CNN is selected as a high-performance image classification method by regarding a spectrogram of audio signal as an image feature.

2.1. Geometric High-Order Decorrelation-Based Source Separation with Adaptive Stepsize Control (GHDSS-AS)

GHDSS-AS [10] is an SSS algorithm based on microphone array processing. Thus, noise sources are not diffused; instead, directional noise sources are effectively suppressed. Two major approaches to sound source separation based on microphone array processing are beamforming and blind source separation. Beamforming uses a transfer function between a sound source and a microphone array, and a large number of microphones are necessary to obtain high performance. On the contrary, blind source separation does not need a transfer function for separation, and better performance can be achieved with a smaller number of microphones. However, the separated signals are difficult to track. That is, it is difficult to know which separated signal should be connected to which sound stream. This is called a permutation problem. In GHDSS-AS, to solve this problem, both blind separation and beamforming techniques are used in a hybrid way, meaning that a separation matrix is estimated using two cost functions of beamforming and blind source separation. Furthermore, GHDSS-AS supports online processing by introducing a step-size parameter. The step-size parameter has a fixed value in normal online SSS algorithms, but GHDSS-AS always controls this parameter, optimizing it so that SSS can work properly, even in a dynamic environment. Therefore, GHDSS-AS is high-performant and robust for both real and simulated data. Actually, open source robot audition software HARK (**H**RI-JP **A**udition for **R**obots with **K**yoto University) [12] provides eleven SSS algorithms, and, in most cases, GHDSS-AS has the best performance among them.

2.2. Robust Principle Component Analysis (RPCA)

RPCA was proposed to reduce the brittleness of *principal component analysis (PCA)* in terms of grossly corrupted observations [13]. RPCA can be applied to a single channel noise suppression even when the observations are highly noise-contaminated [14]. RPCA is defined as a solution for a convex optimization problem as follows:

$$\begin{aligned} &\text{minimize} \quad ||L||_* + \lambda ||S||_1 \\ &\text{subject to} \quad L + S = M \end{aligned} \quad \ldots\ldots\ldots\ldots (1)$$

where $M \in \Re^{n_1 \times n_2}$, $L \in \Re^{n_1 \times n_2}$ and $S \in \Re^{n_1 \times n_2}$ show observation, noise, and target signals, respectively. $||\cdot||_*$ and $||\cdot||_1$ denote the nuclear norm (sum of singular values) and the $L1$-norm (sum of absolute values of matrix entries), respectively. $\lambda > 0$ is a trade-off parameter between the rank of L and the sparsity of S.

RPCA assumes that S is sparse enough, and L should have a low-rank. These assumptions can be maintained in our application, AEI. For example, utterances such as asking for help, and other acoustic events are sparse and dynamic (high-rank) signals, while noise sources such as

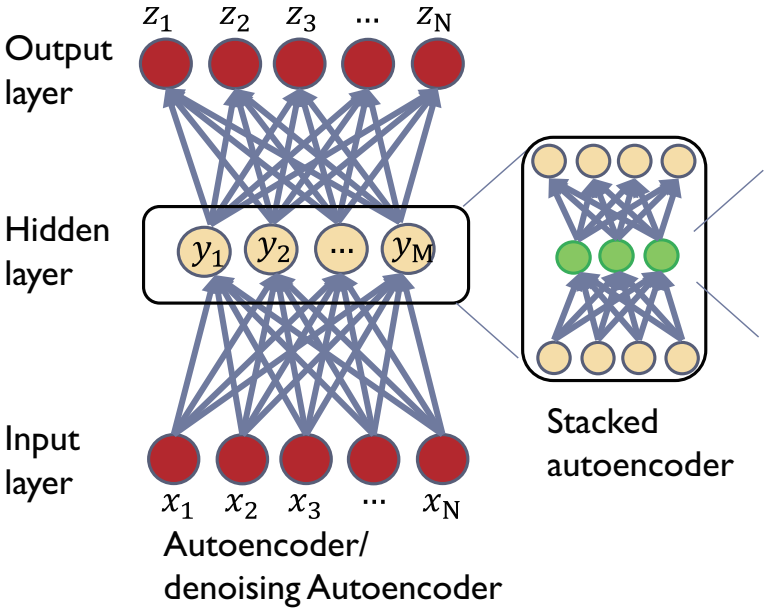


Fig. 1. Stacked denoising autoencoder for pre-training.

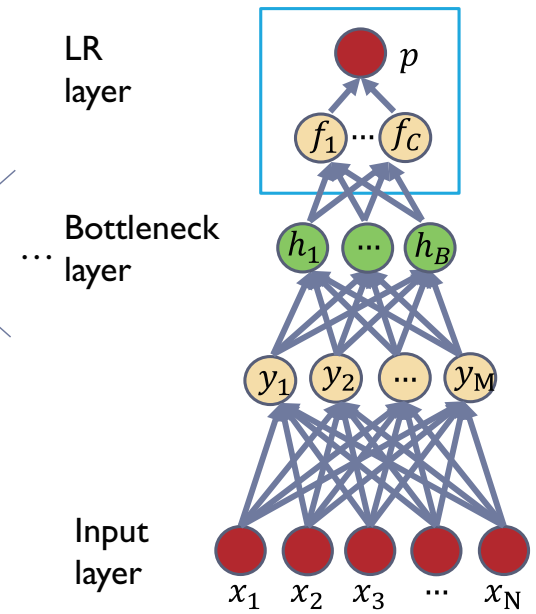


Fig. 2. Fine-tuning with logistic regression.

propellers and wind are continuous and stationary (low-rank) signals.

2.3. Stacked Denoising Autoencoder (SdA)

Since it is difficult to collect a large number of labeled training data with a UAV embedded microphone array, we trained the DNN classifier with SdA, with which we can train the classifier with unsupervised learning. The training of SdA has two stages, that is, unsupervised and supervised learning. The unsupervised learning stage is called "pre-training," and it adjusts the weights and biases of a neural network. After pre-training, the supervised learning stage is performed to adjust the parameters of the whole network precisely. This adjustment is called "fine-tuning." In particular, pre-training plays an essential role, training DNNs to avoid getting stuck into a local minimum. In this configuration, we used *Stacked denoising Autoencoder (SdA)* for pre-training, and *Logistic Regression (LR)* for fine-tuning.

SdA is built by stacking *denoising Autoencoders (dA)* in a nested way. It is said that SdA can represent highly a non-linear model by increasing the number of layers. Since dA is based on autoencoder, we start with autoencoder.

Autoencoder is a unsupervised learning method for a neural network. **Fig. 1** depicts a three-tiered network, consisting of input, hidden, and output layers. It looks like a normal neural network, the only difference being in the learning criteria of autoencoder, as the values of output nodes should be equal to those of the input nodes.

When an N-dimensional vector $\boldsymbol{x} = \{x_1, \ldots, x_N\}$ is given to the input layer, an M-dimensional vector $\boldsymbol{y} = \{y_1, \ldots, x_M\}$ and an N-dimensional vector $\boldsymbol{z} = \{z_1, \ldots, z_N\}$ can be calculated by Eqs. (2), and (3) in the hidden and output layers, respectively.

$$\boldsymbol{y} = \text{relu}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}), \quad \ldots\ldots\ldots\ldots (2)$$

$$\boldsymbol{z} = \text{relu}(\boldsymbol{W}'\boldsymbol{y} + \boldsymbol{b}'), \quad \ldots\ldots\ldots\ldots (3)$$

$$L(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - z_i)^2 \quad \ldots\ldots\ldots\ldots (4)$$

where relu applies a rectified linear unit function, $\text{relu}(x) = \max(0, x)$, for each element of an input vector. $\boldsymbol{W}$ and $\boldsymbol{W}'$ are $N \times M$ and $M \times N$ weight matrices, and $\boldsymbol{b}$ and $\boldsymbol{b}'$ are bias vectors for $\boldsymbol{y}$ and $\boldsymbol{z}$. $L(\boldsymbol{x}, \boldsymbol{z})$ is a mean squared error function to make the difference between $\boldsymbol{x}$ and $\boldsymbol{z}$ as small as possible. In autoencoder, the transposed matrix of $\boldsymbol{W}$ is used for $\boldsymbol{W}'$ to improve learning efficiency. A set of three parameters $[\boldsymbol{W}; \boldsymbol{b}; \boldsymbol{b}']$ is updated by Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments [15].

The dA is just an extension of autoencoder. In the case of dA, in the unsupervised learning stage, noise is added to each node in the hidden layer to make the neural network more noise-robust.

SdA is built by replacing the hidden layer of the learned dA with another autoencoder shown in **Fig. 2**. Every time this nesting process is performed, the number of layers is increased and the nested neural network should be learned. This step-by-step learning process achieves deep learning to avoid converging to local solutions [16].

After pre-training with SdA, LR is performed for fine-tuning in a supervised way. Usually, an LR layer is connected to the output layer or the middle of the hidden layers, which is also called the bottleneck layer. **Fig. 2** is an illustration of the LR layer with the bottleneck layer when they are connected. Since multi-class SSI should be considered, softmax-based LR is used and it is defined as

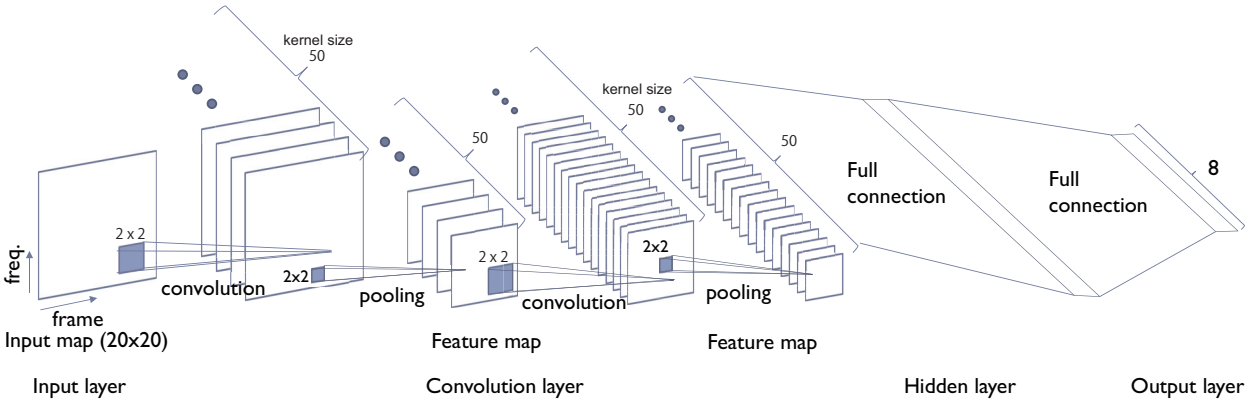


Fig. 3. Convolutional neural network (CNN).

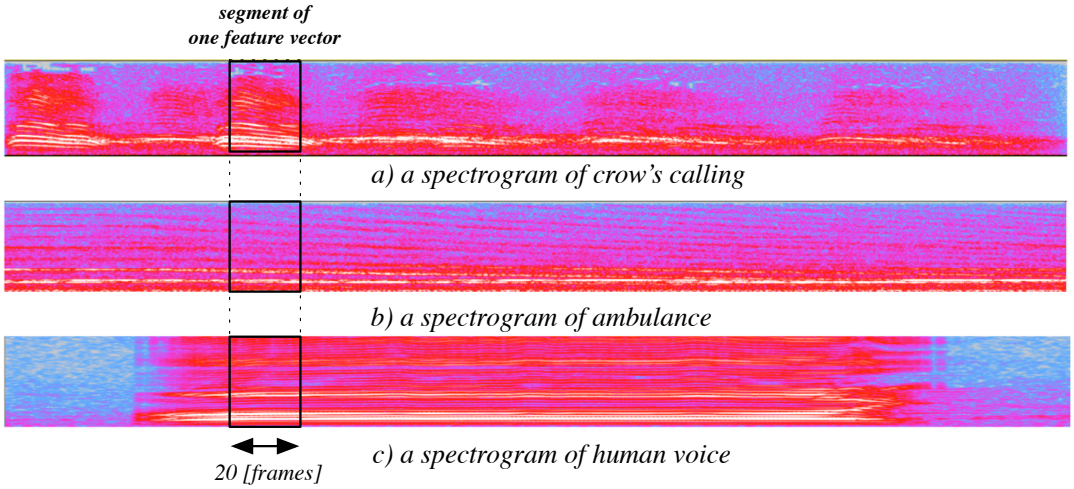


Fig. 4. Spectrograms of sound events.

follows,

$$\hat{p} = \underset{c}{\operatorname{argmax}} f_c \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

$$f_c = \frac{e^{\boldsymbol{W}_c'' \boldsymbol{h} + \boldsymbol{b}_c''}}{\sum_{i=1}^{B} e^{\boldsymbol{W}_i'' \boldsymbol{h} + \boldsymbol{b}_i''}}, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

where $\boldsymbol{h}$ is a B-dimensional vector of the bottleneck layer, $\boldsymbol{b}''$ is a bias vector in the LR layer, $\boldsymbol{W}''$ is a weight matrix between the bottleneck and the LR layers, and p is the class for an input signal given as a teacher label.

2.4. Convolutional Neural Network (CNN)

CNN is a multi-layer neural network consisting of input, convolution, hidden layers, and output layers shown in **Fig. 3** [17]. CNN is well-known for its high performance in image identification. In our study, we regarded a sound spectrogram as an image feature and tried to classify the spectrograms with CNN. The sample images of the sound spectrograms are shown in **Fig. 4**. As shown in **Fig. 4**, the spectrograms are spatially and temporarily different among the sound events; these differences become the cues for finding the targets. In this study, we set the segment length of spectrogram features to 20 frames, which we consider to enough for capturing the temporal transition of the sound spectrum.

A unique feature of CNN is the convolution layer, in which response maps are computed with two operations; convolution and pooling. The convolution operation calculates convolution from the input maps with multiple filters.

Let x be a spectral-temporal 2D map input, which is given to the convolution layer, and the k-th feature map h^k can be defined as follows:

$$h_{ij}^k = \text{relu}\left((W^k * x)_{ij} + b_k\right), \quad \ldots\ldots\ldots \quad (7)$$

where i and j are indices of the feature map, and "$*$" is the convolution operator. We selected relu (Rectified Linear Unit function) as an activation function to minimize the computational cost.

For the pooling operation, "max pooling," which downsamples the convoluted map with a max filter is introduced. When h^k is split into a set of $L \times L$ non-overlapping regions, $G(h^k)_{ij}$, where i and j are indices specifying the

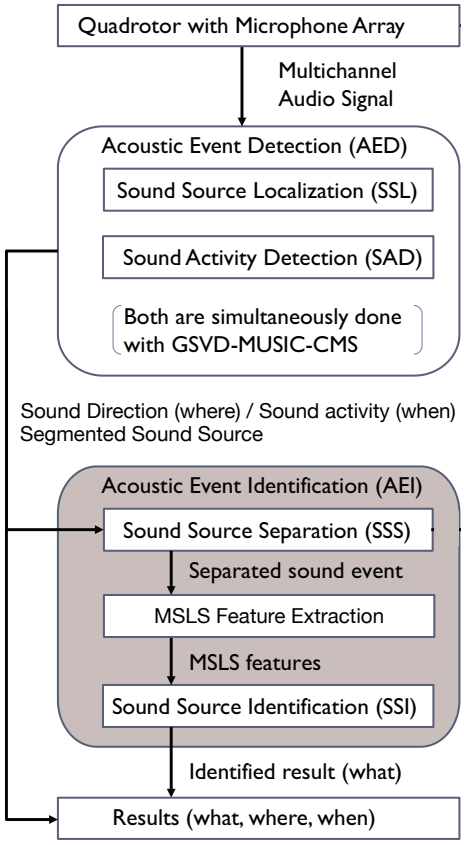


Fig. 5. Software architecture for OCASA with proposed AEI.

(i, j) region of the feature map, a responsive map $\tilde{h}^k$ can be computed by applying a max operation to each region.

$$\tilde{h}_{ij}^{k+1} = \max_{\acute{h}_{\acute{i}\acute{j}}^k \in G(h^k)_{ij}} \acute{h}_{\acute{i}\acute{j}}^k. \quad \cdots\cdots\cdots\cdots (8)$$

On top of the feature maps finally generated, an LR network with hidden and output layers is connected to form an SSI classifier. The learning algorithm of these layers is the same as that of the LR layer in SdA. That is, the parameters of weight matrices and biases are updated with mini-batch-iterated in Adam. To prevent overfitting problem [18], we also applied dropout, a mechanism to randomly dropping out units (along with their connections) from the neural network during training.

3. System Architecture

Figure 5 shows the software architecture for OCASA created by integrating the proposed AEI with the previously-reported AED [5]. The input to the system is multichannel audio signals captured with a quadrotor that is fitted with a 16 ch microphone array (**Fig. 6**).

The quadrotor is based on AscTech Pelican, the maximum payload of which is 650 g. The microphone array system was installed on the top of Pelican. It consists of a multi-channel A/D converter RASP developed by System-in-Frontier Inc., and 16 MEMS microphones attached to the surface at the positions of black hair seen in **Fig. 6**. The layout of the microphone array was designed to be large in diameter. This is because it is known that the main lobe of a larger microphone is sharper, meaning that it has better resolution in sound source localization and separation. Although SSS is introduced in this paper, the wind noise is still a big problem. To deal with this problem, at the position of each microphone, we added wind protection in the form of a material that resembles hair. It is known to be one of the most effective materials for wind protection because the wind power is absorbed by waving of the hair, yet this waving produces no acoustical noise.

Fig. 6. Quadrotor with microphone array.

AED estimates the direction and the activity of each sound event, i.e., *where* and *when* information with iGSVD-MUSIC-CMS [5]. AEI, which is focused on in this paper, includes two modules; SSS and SSI. As mentioned in Section 2, GHDSS-AS and RPCA can be used for SSS, and SdA and CNN can be used for SSI. SSS receives the multichannel audio signal with the direction and activity of the detected sound event, and SSS is then performed.

After that, a 20-dimensional *Mel-Scale Log Spectrum (MSLS)* acoustic feature is extracted from each 32 ms frame of the separated sound event every 10 ms, which is known to be one of the best acoustic features for microphone array processing [19]. An input $\boldsymbol{x}$ for SdA is a 400 dimensional vector consisting of 20-frame MSLS features, and CNN uses a 20×20 matrix, the axes of which are the frame and MSLS dimension, for $\boldsymbol{x}$. Both SdA and CNN classify the input acoustic feature into one of the pre-defined classes, i.e., *what* information. Finally, the OCASA system outputs a sound event with *where*, *when* and *what* information.

For implementation, we used the open source software robot audition software HARK [12]. HARK provides online and real-time algorithms for robot audition including multichannel recording, frequency analysis, iGSVD-MUSIC-CMS, GHDSS-AS, acoustic feature extraction among others. RPCA was newly implemented with MATLAB, and SdA and CNN were implemented with Python. This means that the whole system works off-line, although a part of the system implemented with HARK works in real time.

Table 1. Parameters to be trained with SdA and CNN.

	SdA	CNN
Parameters to be trained	$\{W_l, W'_l, b_l, b'_l\}_{l \in L_h}$	$\{W_l^k\}_{l \in L_{conv}}^{k \in K, N_K = 10,30,50}$ $\{W_l, b_l\}_{l=L_{mlp}}$
Layers	$N_{L_h} = 1,2,3$	$N_{L_{conv}} = 2$, $N_{L_{mlp}} = 1$

*N_K represents a number of feature maps in the convolution layer.
*N_{L_h}, $N_{L_{conv}}$, $N_{L_{mlp}}$ represent a number of hidden layers in SdA, convolution layers in CNN and multi-layer perceptron layers in CNN, respectively.
*A Total of 19,138 data are used. 15,310 features are used for training, and another 3,828 are used for testing.
*Both SdA and CNN were trained 200 epochs with batch size 50.

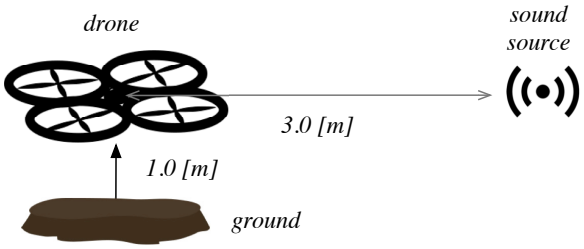


Fig. 7. Environmental settings.

4. Evaluation

We evaluated the OCASA system mainly in terms of AEI. Since our AEI consists of SSS and SSI, combinations of the SSS and SSI methods described in Section 2 were compared. For SSS methods, either GHDSS-AS or RPCA was selected as a multichannel or blind noise robust sound source separation method, respectively. GHDSS-AS was evaluated as a multichannel sound source separation method robust to noise by comparing its results in speech recognition task with those of other methods proposed in previous studies [20]. RPCA was also reported to be effective as a blind-noise-robust sound source separation method for a rescue robot [11]. GHDSS-AS is based on microphone array processing, multichannel audio singals captured with the microphone array embedded in quadrotor. Another input for GHDSS-AS is sound directions sent from AED. Using the multichannel signals and the sound directions, SSS can be performed using a transfer function between each sound source and the microphone array. The transfer function was obtained in advance by measuring impulse responses. In this paper, the measurements were performed on the level plane of the quadrotor at the 5° intervals. Since RPCA is a single channel noise suppression method, a single channel from multichannel, which is closest to the target sound source in place of the multichannel audio, was used. The single channel sound signal was segmented based on the sound source localization result with *iGSVD-MUSIC-CMS*.

For SSI methods, SdA and CNN were used. Every method used 20-dimensional MSLS features for 20 frames as an input acoustic feature, as mentioned in Section 3. The structure of CNN used in the evaluation is shown in **Fig. 3**. The parameters for the learning of SdA and CNN are shown in **Table 1**. In order to compare deep and shallow identification methodologies, we also used GMM (Gaussian Mixture Model), which is widely used to classify sound sources.

Additionally, in order to compare the performance differences with DNN parameters, we set up three conditions for both SdA and CNN. In the SdA conditions, we set up dA with three layers (400, 100 and 400 dimensions, respectively), SdA with five layers (400, 200, 100, 200 and 400 dimensions each) and SdA with seven layers (400, 300, 200, 100, 200, 300 and 400 dimensions each), respectively. While in CNN conditions, we changed the kernel size (see **Fig. 3**, convolution layer) of the convolution layers from 10 to 50, since it is not easy to change a layer size due to the image size of the feature vector.

Eight sound sources were selected from *RWCP Sound Scene Database in Real Acoustical Environments* and *JEIDA Noise Database*, including a request to ask for help, the ring of an alarm clock, hand claps, a car horn, cymbal crash, crow's call, a mobile phone's ringtone, and an ambulance's siren. These sound sources were recorded at a 16 kHz sampling rate while the quadrotor was hovering outdoors.

The multichannel audio signals in the experiment were captured in the environment shown in **Fig. 7**, where sound sources were output with the portable speaker 3.0 m away from the quadrotor. The quadrotor was fixed at a point 1.0 m point from the ground, its rotors rotating to keep it aloft. Though the sound sources in the experimental setup were not so far from the quadrotor as they were in the iGSVD-MUSIC-CMS evaluation [5], still the S/N rate of the recorded multichannel audio signals were approximately -20.0 dB. That is, we evaluated the proposed system in a low S/N rate acoustic environment. The length of each sound event was 3–4 seconds, and each event was recorded 15 times. In total, approximately 7-minutes of data were collected. Since the frame shift length for feature extraction was 10 ms, the total number of features was 19,138. The main noise sources were the propeller and wind in the recording.

In the evaluation, eight-class identification was conducted using five-fold cross validation. The eight sound sources directly correspond to the eight classes.

A frame-by-frame basis metric called *SSI Correct Rate (SSR)* was measured. SSR is denoted as follows:

$$SSR = \frac{1}{C} \sum_{i=1}^{C} \frac{\sum_{j=1}^{E_i} F_c(i,j)}{\sum_{j=1}^{E_i} F_a(i,j)}, \quad \cdots\cdots\cdots\cdots \quad (9)$$

where C is the number of classes, i is the class index.

Table 2. SSR for real data, 8-class SSI.

SSS/SSI	GHDSS-AS	RPCA
GMM (20 comp.)	0.662	0.504
dA (400-100)	0.742	0.482
SdA (400-200-100)	0.765	0.477
SdA (400-300-200-100)	0.695	0.464
CNN (10 kernel)	0.774	0.507
CNN (30 kernel)	0.803	0.588
CNN (50 kernel)	**0.842**	0.645

Each value in **Table 2** *was an average of SSR in 5-fold cross validation. SdA (400-200-100)* means the stacked denoising autoencoder, which consists of *400-200-100-200-400* dim. input, hidden and output layers respectively. *CNN (50 kernel)* means a convolutional neural network consisting of two convolutional and pooling layers with *50* kernels and one hidden layer.

Table 3. Confusion matrix of CNN with GHDSS-AS.

L0	L1	L2	L3	L4	L5	L6	L7
416	0	12	37	26	2	0	1
0	**951**	21	1	1	1	0	0
10	38	**705**	19	55	9	15	2
18	1	24	**248**	17	8	1	0
15	14	71	8	**728**	12	4	2
8	0	12	11	18	**85**	8	2
0	0	11	1	4	9	**37**	5
1	0	5	0	10	7	5	**19**

L0, L1, L2, L3, L4, L5, L6, L7 are *telephone ringtone, ambulance siren, crow call, human voice, car horn, hand clap, cymbal crash* and *clock alarm* respectively, in 3828 frames in the test datasets.

E_i is the number of sound events for the i-th class, and j is the index for the sound events included in the i-th class. $F_a(i,j)$ is the total number of frames and $F_c(i,j)$ is the number of successfully-identified frames for the j-th sound event of the i-th class.

4.1. Results

Table 2 shows the AEI results. GHDSS-AS had better performance than RPCA under SSS conditions. Since the propeller and wind noise also has the sparse noise components, the target sound signal was not properly separated by the RPCA. The target sound signal was relatively well separated by the GHDSS-AS.

On the other hand, CNN obviously performed better than did SdA. There would be two reasons for this. One is that SdA does not consider temporal information well, while CNN uses a convolution operation, which considers temporal information. The other is that we used a bottleneck layer for classification. In ASR based on DNN, it is known that the use of an output layer for classification has better performance, and this would also be true of SSI. The CNN had also better performance than did GMM (CNN $>$ SDA $>$ GMM), and these comparisons also indicate the importance of considering temporal information in the sound source classification.

As for the comparison under SdA conditions with GHDSS-AS, the performance of SdA with five layers had a higher performance than that of SdA with seven layers.

One of the reasons why the performance of SdA with seven layers was lower than that with five layers was the lack of a training data set since we evaluated the performance with a limited number of training and test data (**Table 1**). With a larger training data set, the performance of SdA with seven layers could exceed that with five layers.

Table 3 shows a confusion matrix of the audio event identification per each frame based on CNN with GHDSS-AS. Each event was properly identified, while the detection number of the events was biased. The longer and characteristic audio events such as phone, ambulance, and human voice, could be better detected and identified than the shorter audio events such as the clap, clock and cymbals. These characteristic audio events are also considered to be helpful in finding a target to be rescued in a disaster scenario.

4.2. Prediction Time of Acoustic Event Identification with GHDSS-AS-CNN for Real-Time Processing

We also examined the prediction time of the CNN identifier (50 kernels) with the combination of GHDSS-AS, the combination that had the best performance in the previous section. Since the components in the OCASA system other than the CNN identifier have the ability to work in real-time, the bottleneck for achieving the real time processing is the prediction time of CNN, which is widely known as a weak point of the deep neural network. The comparison was done with the GPU (Tesla K20c), a high-end CPU (3.2 GHz Xeon CPU), and low-end CPU (1.1 GHz Intel Core M). The first two are for high-performance computing, and the last one is to be embedded on the quadrotor.

Figure 8 shows the prediction time of each core for SSI with 68-frame datasets, which is the average frame number for the human voices in the test datasets. As shown in **Fig. 8**, the prediction time of the CNN identifier with the GPU was better than those of CPUs. From the perspective of comparing the CPUs, the performance of CPUs seemed to be linear with the CPU base frequency. As a consequence, the prediction time of SSI with GPU was around 67 times shorter than that of the low-end CPU. However, even the prediction time with the low-end CPU is around 94.8 msec, which is around 1/10 of the length of the human voice. With the assumption that the frequency of acoustic events is sparse and the SSI process is not always required, we can handle the incremental SSI processing with the additional implementation of the processing queue of SSI, which stores the acoustic event features and processes the SSI one by one.

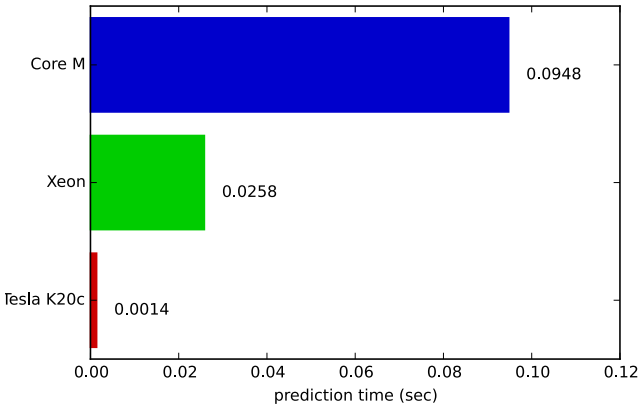


Fig. 8. Comparison of Tesla K20c, Xeon 3.2 GHz, and Core M 1.1 GHz prediction times.
Note: Each prediction time is an average of scores over five trials.

5. Conclusion

This paper presents *Acoustic Event Identification (AEI)* for acoustic signals observed with a quadrotor-embedded microphone array in a noisy outdoor environment in which a quadrotor is flying. Since the contamination of noise generated by the propellers and wind is a primary problem, we proposed the use of a combination of sound source separation and sound source identification. For sound source separation, we selected GHDSS-AS and RPCA as the most advanced multichannel and single-channel methods, respectively. For sound source identification, two deep learning techniques – SdA and CNN – were used, since SdA and CNN are known as powerful methods for audio and image processing, respectively. We assessed the proposed AEI methods by testing combinations of sound source separation and identification methods to validate the proposed method. Using audio signals recorded in an outdoor environment in which a quadrotor is operating, a sound source identification success of over 80% was achieved by using combinations of GHDSS-AS and CNN. We also measured the prediction time of a trained CNN identifier with GHDSS-AS, which showed that the CNN identifier could work even with a low-end CPU (Core M 1.1 GHz). Future work will include further exploration of the optimal parameter settings, and the implementation of an online and real-time system with the contributions of this paper.

Acknowledgements

This work was supported by JSPS KAKENHI Grant No.24220006, 16H02884, and 16K00294, and also by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).

References:

[1] P. Ross, "Robot, you can drive my car," IEEE Spectrum, Vol.51, No.6, pp. 60-90, 2014.
[2] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," IEEE/RSJ IROS, pp. 3288-3293, 2012.
[3] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2009), pp. 664-669, 2009.
[4] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor uav," Proc. of the IEEE/RSJ Int. Conf. on Robots and Intelligent Systems (IROS), pp. 3943-3948, 2013.
[5] T. Ohata, K. Nakamura, T. Mizumoto, T. Tezuka, and K. Nakadai, "Improvement in outdoor sound source detection using a quadrotor-embedded microphone array," 2014 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp. 1902-1907, 2014.
[6] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 171-175, 2015.
[7] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 3704-3708, 2014.
[8] H. Phan, M. Maas, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol.23, No.1, pp. 20-31, 2015.
[9] I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol.23, No.3, pp. 540-552, 2015.
[10] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Correlation matrix estimation by an optimally controlled recursive average method and its application to blind source separation," Acoustical Science and Technology, Vol.31, No.3, pp. 205-212, 2010.
[11] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H. G. Okuno, "Human-voice enhancement based on online rpca for a hose-shaped rescue robot with a microphone array," 2015 IEEE Int. Symposium on Safety, Security, and Rescue Robotics (SSRR), pp. 1-6, 2015.
[12] 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, pp. 739-761, 2010.
[13] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" J. of the ACM, Vol.58, No.3, Article No.11, 2011.
[14] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," J. of Computational Information Systems, Vol.10, No.10, pp. 4403-4410, 2014.
[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[16] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," L. Bottou (Ed.), The J. of Machine Learning Research, Vol.11, pp. 3371-3408, 2010.
[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Procs. of the IEEE, Vol.86, No.11, pp. 2278-2324, 1998.
[18] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 8609-8613, 2013.
[19] Y. Nishimura, K. Nakadai, M. Nakano, H. Tsujino, and M. Ishizuka, "Speech recognition for a humanoid with motor noise utilizing missing feature theory," Procs. of the 2006 IEEE-RAS Int. Conf. on Humanoid Robots (Humanoids 2006), pp. 26-33, 2006.
[20] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, "Blind source separation with parameter-free adaptive step-size method for robot audition," IEEE Trans. on Audio, Speech, and Language Processing, Vol.18, No.6, pp. 1476-1485, 2010.

Name:
Osamu Sugiyama

Affiliation:
Kyoto University Hospital

Address:
54 Kawaharacho, Shogoin, Sakyo-ku, Kyoto City 606-8507, Japan

Brief Biographical History:
2007-2009 SONY
2009-2013 Advanced Telecommunications Research Institute International
2013-2016 Kyoto University and Tokyo Institute of Technology
2016- Kyoto University Hospital

Main Works:
• O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model," Connection Science (Special issues on android science), Vol.18, No.4, pp. 379-402, 2006.

Membership in Academic Societies:
• The Robotic Society of Japan (RSJ)
• The Japanese Society for Artificial Intelligence (JSAI)

Name:
Satoshi Uemura

Affiliation:
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Address:
2-12-1-W8-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Brief Biographical History:
2016 Received Master of Engineering in Computer Science from Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)

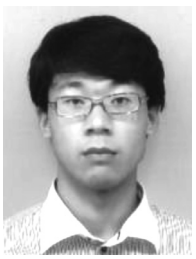
Name:
Akihide Nagamine

Affiliation:
Department of Electrical and Electronic Engineering, School of Engineering, Tokyo Institute of Technology

Address:
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Brief Biographical History:
2016- Department of Electrical and Electronic Engineering, School of Engineering, Tokyo Institute of Technology

Membership in Academic Societies:
• The Japan Society of Applied Physics (JSAP)



Name:
Ryosuke Kojima

Affiliation:
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Address:
2-12-1-W8-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Brief Biographical History:
2014 Received Master of Engineering in Computer Science from Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2014- Doctoral program, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

Main Works:
• R. Kojima, O. Sugiyama, and K. Nakadai, "Multimodal Scene Understanding Framework and Its Application to Cooking Recognition," Applied Artificial Intelligence, Vol.30, No.3, pp. 181-200, 2016.
• R. Kojima and T. Sato, "Goal and Plan Recognition via Parse Trees Using Prefix and Infix Probability Computation," Inductive Logic Programming, Springer, LNAI, Vol.9046, pp. 76-91, 2015.

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)
• The Japanese Society for Artificial Intelligence (JSAI)

Name:
Keisuke Nakamura

Affiliation:
Senior Scientist, Honda Research Institute Japan Co., Ltd.

Address:
8-1 Honcho, Wako-shi, Saitama 351-0114, Japan

Brief Biographical History:
2010- Joined Honda Research Institute Japan Co., Ltd.
2013 Received Ph.D. in Informatics from Kyoto University

Main Works:
• "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," Advanced Robotics, Vol.27, No.12, pp. 933-945, 2013.

Membership in Academic Societies:
• The Institute of Electrical and Electronic Engineers (IEEE)
• The Robotics Society of Japan (RSJ)

Name:
Kazuhiro Nakadai

Affiliation:
Honda Research Institute Japan Co., Ltd.
Tokyo Institute of Technology

Address:
8-1 Honcho, Wako-shi, Saitama 351-0188, Japan
2-12-1-W30 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Brief Biographical History:
1995 Received M.E. from The University of Tokyo
1995-1999 Engineer, Nippon Telegraph and Telephone and NTT Comware
1999-2003 Researcher, Kitano Symbiotic Systems Project, ERATO, JST
2003 Received Ph.D. from The University of Tokyo
2003-2009 Senior Researcher, Honda Research Institute Japan Co., Ltd.
2006-2010 Visiting Associate Professor, Tokyo Institute of Technology
2010- Principal Researcher, Honda Research Institute Japan Co., Ltd.
2011- Visiting Professor, Tokyo Institute of Technology
2011- Visiting Professor, Waseda University

Main Works:
• K. Nakamura, K. Nakadai, H. and G. Okuno, "A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition," Advanced Robotics, Vol.27, Issue 12, pp. 933-945, 2013 (Received Best Paper Award).
• H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based Online Calibration for Asynchronous Microphone Array," Advanced Robotics, Vol.26, No.17, pp. 1941-1965, 2012.
• R. Takeda, K. Nakadai, T. Takahashi, T. Ogata, and H. G. Okuno, "Efficient Blind Dereverberation and Echo Cancellation based on Independent Component Analysis for Actual Acoustic Signals," Neural Computation, Vol.24, No.1, pp. 234-272, 2012.
• K. Nakadai, T. Takahashi, H. G. Okuno et al., "Design and Implementation of Robot Audition System "HARK"," Advanced Robotics, Vol.24, No.5-6, pp. 739-761, 2010.
• K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino, "Improvement of recognition of simultaneous speech signals using AV integration and scattering theory for humanoid robots," Speech Communication, Vol.44, pp. 97-112, 2004.

Membership in Academic Societies:
• The Robotics Society of Japan (RSJ)
• The Japanese Society for Artificial Intelligence (JSAI)
• The Acoustic Society of Japan (ASJ)
• Information Processing Society of Japan (IPSJ)
• Human Interface Society (HIS)
• International Speech and Communication Association (ISCA)
• The Institute of Electrical and Electronics Engineers (IEEE)