https://doi.org/10.20965/jrm.2021.p0515

Paper:

Localization of Flying Bats from Multichannel Audio Signals by Estimating Location Map with Convolutional Neural Networks

Kazuki Fujimori*, Bisser Raytchev*, Kazufumi Kaneda*, Yasufumi Yamada*, Yu Teshima, Emyo Fujioka**, Shizuko Hiryu**, and Toru Tamaki*****

*Hiroshima University
1-4-1 Kagamiyama, Higashi-hiroshima, Hiroshima 739-8527, Japan
**Doshisha University
1-3 Tatara-miyakodani, Kyotanabe, Kyoto 610-0394, Japan
***Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555, Japan
E-mail: tamaki.toru@nitech.ac.jp
[Received December 18, 2020; accepted April 28, 2021]

We propose a method that uses ultrasound audio signals from a multichannel microphone array to estimate the positions of flying bats. The proposed model uses a deep convolutional neural network that takes multichannel signals as input and outputs the probability maps of the locations of bats. We present experimental results using two ultrasound audio clips of different bat species and show numerical simulations with synthetically generated sounds.

Keywords: bat, multichannel, ultrasound signal, CNN, location estimation

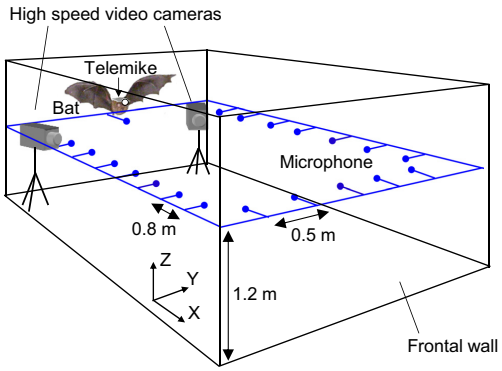

Fig. 1. Experimental setup in a laboratory room of the size 6 m × 4.5 m × 2.35 m [7].

1. Introduction

Bats use ultrasound to perceive their surrounding environment. Research on the ecology of bats is expected to elucidate the mechanism of biosonar, acoustic navigation, and complex swarm behaviors, as well as to develop new sensors and navigation systems. To this end, obtaining the trajectories of flying bats is important for studying the relationship between flight paths and target approaching or foraging [1, 2]. Currently, there are two types of methods for estimating bat trajectories. One is the arrival time difference using a microphone array [3–5], typically used in field settings. As bats regularly emit ultrasound pulses to locate their own positions, we can record the pulses with calibrated microphones and calculate the position of the bat from the difference in the arrival times of the signals at different microphones. However, it is time-consuming for human operators to process multichannel audio signals to calculate the time difference, particularly when two or more bats fly simultaneously. The other method is stereo camera measurements [6]. This approach can be applied for multiple flying bats in both indoor and outdoor settings; however, it is difficult to detect and track bats in stereo images automatically or even manually. This is because bats are difficult to identify in the dark in the field, or even during a single flight in a laboratory room, owing to the low lighting conditions and the limited sensor dynamic range of the stereo camera. A relatively brighter environment is necessary for the functioning of automatic tracking systems, but this is not the case in the field environment, and even in laboratory experiments.

Therefore, we focus on the difficulty of the approach using a microphone array because a larger flying space can be captured and pulse sounds (and also emitted directions) are important for research on bats. As a first step, in this study we tackle this problem with a microphone array installed in a laboratory room (see **Fig. 1**). We propose an automatic method for localizing bat positions and estimating flying trajectories using multichannel audio signals. We develop a deep convolutional neural network (CNN) model that takes multichannel signals as the input and outputs the coordinates of bats. As ground-truth coordinates to be estimated, we use the locations

obtained using a stereo camera with manual annotations. This approach considerably reduces the processing time from several hours with manual operation to a few minutes, once the model has been trained on a calibrated setting.

A naive approach is to minimize the loss (or cost function) between the ground-truth and estimated coordinates of the target bat (Section 3.1). It is straightforward, but not applicable to the case of multiple bats flying simultaneously. Therefore, we adopt another approach to estimate a probability map that represents the locations of multiple bats (Section 3.2). Multiple locations of bats can be obtained by detecting the peaks in the map. The above approaches take a clip of multichannel signals in a certain time window as input and estimate the bat location(s). To obtain the trajectory, we propose to estimate multiple maps at successive time steps simultaneously (Section 3.3), which is better than estimating locations at different times separately in terms of temporal coherency. In addition, we estimate the number of bats flying simultaneously by using an additional branch in the CNN model (Section 3.4).

Experimental results (Section 4.2) of two different species (*Rhinolophus ferrumequinum nippon* and *Miniopterus fuliginosus*) demonstrate the feasibility of the proposed method. Real data for these experiments are limited, and we show results with synthetically generated multichannel signals by using a simple audio simulation (Section 4.3). The simulation can generate a vast amount of synthetic signals useful for training the CNN model; therefore, this approach is a promising route to extend our method to a field setting with a predefined calibrated microphone array.

2. Related Work

2.1. Stereo Camera

Stereo cameras have been used to obtain the coordinates of objects and points in a 3D scene that are visible from both cameras. They are based on triangulation; typically, the cameras are tightly fixed on a stereo rig, and then, their parameters are calibrated prior to stereo measurements. This approach has been adopted to localize flying animals [6] by tracking each animal in each stereo image; however, it typically involves human annotations or corrections. This is because automatic (or semi-automatic) methods for detecting and tracking bats often fail due to the low visibility of black bats in a dark environment with low lighting conditions, or their high-speed motion compared with the shutter speed and frame rate of the camera, as well as the relatively low resolution of bats in the stereo images.

2.2. Difference of Arrival Times

Sound source localization (SSL) is a well-known problem in signal processing for estimating the location of a sound source. A basic approach is to use the difference in the arrival times of the same audio signal at three or more different microphones [8–10]. SSL has been studied for various applications, such as human interactive robots and speaker tracking in meetings. This approach has been used to localize bats flying in outdoor environments [3]. The ultrasound pulses emitted by bats were recorded at a very high sampling frequency (e.g., 500 kHz). These pulses are so highly directional that the microphone array is typically placed such that it surrounds the flying space of the bats, to record pulses in any direction. The problem is that processing multichannel audio signals involves many human operations to eliminate the mismatches of pulses in different channels, even with custom-made software for this task, and even in the case of a single flying bat (multiple bats worsen the situation because of the need for the separation of their mixed pulses).

2.3. Deep Learning Approaches

As it is difficult for naive SSL methods to handle real environments with noise and multiple sound sources, deep learning approaches have recently gained considerable attention. These approaches typically estimate the vertical and horizontal direction angles of sound sources using a stereo microphone [11–13].

Audio signals are often converted to spectrograms, which are 2D visualization of 1D signals, whose vertical and horizontal axes represent the frequency and time, respectively, and the intensity indicates the power of the signal at a specified frequency and time. Several studies [11, 12, 14–26] localized sound sources using spectrogram-based deep models; however, these studies estimated the direction of arrival of sound at the microphones, not the locations of the sound sources. Recently the work by Vera-Diaz et al. [27] has estimated the coordinates of a speaker by using two microphone arrays placed on a table. Their CNN model takes raw multichannel audio signals as input and outputs the coordinates; therefore, it is not applicable to multiple sound sources.

In this study, our CNN model localizes multiple bats from their raw multichannel audio signals directly, which is inspired by [27] applied to multichannel electroencephalogram (EEG) signals [28, 29]. Instead of estimating the coordinates of a single sound source, we estimate a probability map representing the locations of multiple sound sources, followed by peak detection in the map, which is a common approach for pose estimation [30, 31]. In addition, our main motivation is to obtain the trajectories of flying bats. This is a dynamic auditory situation and is very challenging. This is completely different from scenes including only static sound sources, which were considered in all the above studies with deep models.

3. Method

In this study, we focus on finding the 2D (x and y) coordinates of bats, although our framework can be extended to 3D localization with the 3D configurations of microphones.

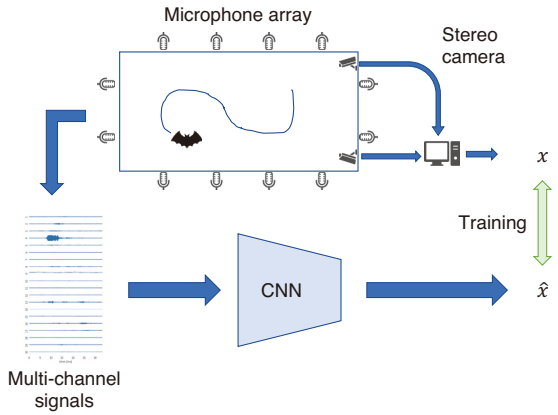

Fig. 2. Overview of coordinate estimation.

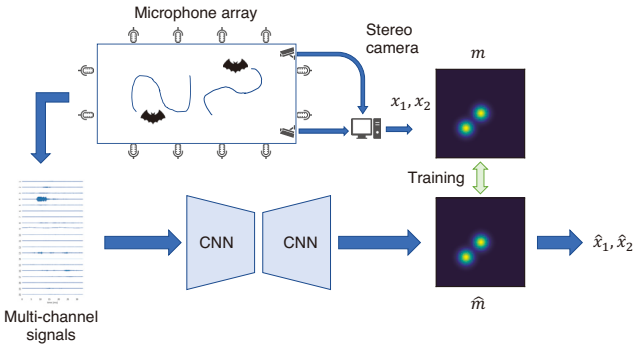

Fig. 3. Overview of map estimation.

3.1. Coordinate Estimation

First, we introduce a naive and intuitive approach as a baseline for the following approaches. In this case, we directly estimate the 2D coordinates of a single sound source from a fixed-size clip of the multichannel signals. An overview of the system is shown in **Fig. 2**.

Let $I^t \in R^{C\times T}$ be the input of multichannel signals with C channels (called a sound *clip*) and a duration T at time t, and let $x^t, \hat{x}^t \in R^2$ be the ground-truth and estimated 2D coordinates of a bat, respectively. We use a CNN model that takes I^t as the input and outputs the prediction $\hat{x}^t$, to minimize the following loss function:

$$L_{coord} = \frac{1}{2}\|x^t - \hat{x}^t\|_2^2, \quad \ldots\ldots\ldots\ldots \quad (1)$$

which is a commonly used mean-square-error (MSE) loss.

3.2. Map Estimation

Estimating a map to detect multiple points in an image has been used for image understanding [30, 31]. Here, we estimate the coordinates of multiple sound sources from a fixed-size clip of a multichannel signal. To this end, we estimate a 2D map of multiple bat locations instead of a single pair of 2D coordinates. This map has a higher value if a bat is likely to exist in each discretized location of the 2D grid. Ground-truth maps are generated using a Gaussian distribution via the following procedure. First, the actual coordinates of the bats are obtained using stereo camera measurements as before. Then, we create 2D Gaussian distributions with a fixed variance, whereas the mean locations of each Gaussian shape are specified by the ground-truth coordinates. Finally, all the Gaussian shapes are aggregated to create a single map. Examples are shown in **Fig. 3**. Note that the isometric Gaussian shape is used in this study for simplicity, and more complex shapes will be considered in the future.

Let $m^t, \hat{m}^t \in R^{H\times W}$ be the ground-truth and estimated maps at time t, respectively, and we minimize the MSE loss function, which is commonly used for pose estimation [30, 31], as follows:

$$L_{map} = \frac{1}{HW}\|m^t - \hat{m}^t\|_2^2, \quad \ldots\ldots\ldots\ldots \quad (2)$$

where H and W are the height and width of the map, respectively. From the predicted map $\hat{m}^t$, single or multiple peak(s) are detected to obtain the final estimation of the bat locations. Note that we only consider peaks in this study, whereas the map distribution may also have important information on the uncertainty of the bat locations.

Our CNN model for map estimation has an encoder-decoder architecture. The architecture of the encoder is almost the same as the coordinate estimation. The encoded feature is then reshaped into a 1D vector and fed to the fully connected (FC) layers. The resulting feature vector is then reshaped into a tensor of size $C \times H' \times W'$. The decoder subsequently increases the spatial resolution while reducing the channel dimensions. Finally, we use a 1×1 convolution with a single-channel output followed by a sigmoid, and obtain a 2D map estimation $\hat{m}^t$ of size $H \times W$.

3.3. Simultaneous N Map Estimation

The two aforementioned approaches have the same issue. If the sound sources are static (i.e., fixed in the auditory scene), then these approaches function as expected because an input multichannel sound clip corresponds to the ground-truth locations of the sound sources. However, our motivation is to obtain the trajectories of flying bats with a single audio clip, and in this case, a single ground-truth location for the audio clip is not suitable for training CNN models.

Two additional issues need to be addressed. 1) Silent clips: depending on the duration of the clip, no bat sounds are included in the clip. The clip is silent, and it should not be used because it has no information for the estimation. However, it is not easy to determine whether the clip is silent due to environmental noise. 2) Temporal incoherency: after the estimation of bat locations for a given clip at a certain time, the next locations are estimated from the next clip at the next time step. In this case, locations at different times are estimated independently, and may be inconsistent between successive time steps even if the ground-truth locations have a smooth trajectory over time.

Therefore, we propose the use of a relatively longer audio clip to avoid the issue of silent clips, and estimate N maps at multiple time steps, rather than a single time

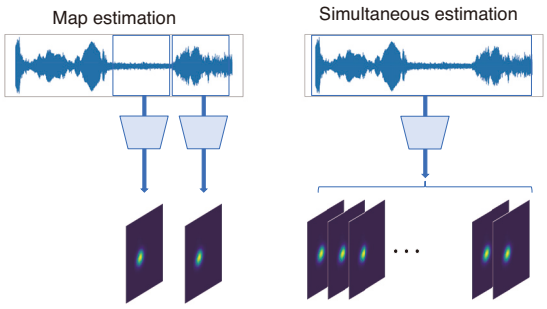

Fig. 4. Estimation of (right) a single map and (left) multiple N maps simultaneously.

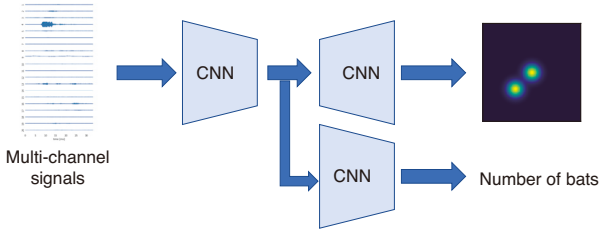

Fig. 5. Estimating the number of bats with an additional branch.

step (see **Fig. 4**). Let $I^t \in R^{C\times T'}$ be the input of multi-channel signals with C channels and a duration T' ($> T$) at time t, and $m^t, \hat{m}^t \in R^{H\times W\times N}$ be the ground-truth and estimated N maps, respectively. The MSE loss function is computed over N maps as follows:

$$L_{Nmap} = \frac{1}{NHW}\|m^t - \hat{m}^t\|_2^2. \quad \ldots\ldots\ldots (3)$$

3.4. Estimating the Number of Bats

The peak detection in the maps used herein is simple, and requires the number of peaks (bats). This number is known in indoor experiments; however, it is not known in advance in other situations, typically in outdoor scenes to which we intend to extend our method in the future. To this end, we add a branch to estimate the number of bats in parallel to the map estimation branch (see **Fig. 5**). The output of the number branch is not a single integer, but instead an n-dimensional categorical one-hot encoding vector $y = (p_0, p_1, \ldots, p_n)$, whose elements represent different integers from $0, 1, \ldots, n$. This is because we intend to represent probabilities over the number of bats as confidence.

The loss function for the number branch is a common cross entropy (CE) expressed as follows:

$$L_{num} = -\sum_{c=0}^{n} y_c \log p_c, \quad \ldots\ldots\ldots (4)$$

where p_c is the predicted probability that the number is c, normalized to be a unit $\sum_c p_c = 1$, and y_c is the c-th element of the ground truth y.

The final loss is the combination of the map loss and the number loss with a weight λ as follows:

$$L = \lambda L_{Nmap} + L_{num}. \quad \ldots\ldots\ldots (5)$$

Table 1. CNN architecture for coordinate estimation.

Stage	1	2	3	4	5
Blocks	2	2	3	3	3
Channels	32	32	64	64	64
1D filter size	11	11	11	17	17

4. Experimental Results

In this section, we report the experimental results using real data and numerical simulations. First, we describe the details of this setting.

4.1. Experimental Setting

4.1.1. Coordinate Estimation

To extract features from a clip I^t, the CNN model consists of five stages of blocks, as listed in **Table 1**. Each block has a 1D temporal convolution layer with the specified output channels and temporal filter size, followed by batch normalization (BN) and rectified linear unit (ReLU) activation. Max pooling layers were inserted between each stage to halve the temporal resolution. After reshaping the feature into a 1D vector, three FC layers with the output units of 5120, 1024, and 2 were used for predicting $\hat{x}^t$.

The time window T of the clip was set to 16,666 points based on the following considerations. The maximum flight speed of the bats in the laboratory was approximately 4–5 m/s, corresponding to 10–20 cm in 1/30 s. On average, bats emit ultrasound pulses every 30–40 ms. Furthermore, the maximum distance between the walls of the laboratory was approximately 6 m, which indicates that the sound waves reach all the microphones within approximately 22 ms, assuming the sound speed of 340 m/s. Hence, if the time window is set to 1/30 s = 16,666 points, the ultrasound pulse emitted by the bats at the beginning of the time window arrives at all the microphones in the same time window. Thus, localization is expected to be realized by using the difference of arrival times of the ultrasound pulse. Note that the window size should be modified accordingly when the field size is different from the current setting. For training, the clips were extracted by shifting 1024 points from a long original signal sequence. Therefore, 15,642 out of 16,666 points overlapped in successive clips.

4.1.2. Map Estimation

In this study, the map size was set to 256×256 pixels, which indicates that the size of a single grid of the map was approximately 20 mm compared to the room size. This grid size is small relative to the size of typical bats (20–50 mm, without wings). The standard deviation of the 2D Gaussian distributions for the ground-truth maps was $\sqrt{200} \simeq 14$ pixels.

The architecture of the encoder-decoder model for map estimation is shown in **Table 2**, which is almost the same

Table 2. Encoder (top) and decoder (bottom) in the CNN architecture for the map estimation.

Encoder stage	1	2	3	4	5	6
Blocks	2	2	3	3	3	3
Channels	32	32	64	64	64	64
1D filter size	21	21	21	35	35	35

Decoder stage	1	2	3	4
Blocks	2	2	2	2
Channels	32	16	8	4
2D filter size	3×3			

Table 3. Encoder (top) and decoder (bottom) in the CNN architecture for the simultaneous N map estimation. The 3D filters in the decoder have the dimensions $H \times W \times N$.

Encoder stage	1	2	3	4	5	6
Blocks	2	2	3	3	3	3
Channels	64	128	256	512	1024	1024
1D filter size	21	21	21	35	35	35

Decoder stage	1	2	3	4	5
Blocks	2	2	2	2	2
Channels	8	8	8	4	2
3D filter size	$3 \times 3 \times 35$				

as the coordinate estimation. The encoded feature was then reshaped into a 1D vector and fed to two FC layers with the output units of 8192. The resulting feature vector was then reshaped into a tensor of size $C \times H' \times W' = 32 \times 16 \times 16$.

We detected peaks from the estimated map as follows. First, a max filter of size 25 was applied to the estimated map. Then, the maximum location was detected as the first peak. Next, we removed map values larger than 0.5 times the value of the detected first peak. We then detected the next peak. This process was repeated until a specified number of peaks were obtained. This method is simple, but it fails when there are numerous bats; however, it is effective for the following experiments.

4.1.3. Simultaneous *N* Map Estimation

Table 3 shows the architecture of the CNN model. Here, we used the time window of $T' = 61{,}440$ points, corresponding to 120 ms, which is sufficiently long to avoid a clip being silent compared with the average emission intervals of 30–40 ms. The number of maps was set to $N = 60$; hence, maps were effectively estimated every 2 ms.

For training, the clips were extracted by shifting 1024 points from a long original signal sequence. Therefore, 60,416 out of 61,440 points overlapped in successive clips.

4.1.4. Number Estimation

In the following experiments, the maximum number of bats was $n = 3$, which was known in advance. The framework will be extended for a larger number of bats using probability in the future.

Table 4. Estimation errors (RMSE) in mm of coordinates, simultaneous N coordinates, map, and simultaneous N maps for *R. ferrumequinum nippon*.

	Train	Test
Coordinates	23.676	935.327
N coordinates	59.036	430.466
Map	16.237	531.698
N maps	20.573	221.708

4.2. Real Environment

We used two real ultrasound multichannel sounds, each for two different species; *Rhinolophus ferrumequinum nippon* and *Miniopterus fuliginosus*.

We compared the following variations: coordinates, map, simultaneous N maps, and simultaneous N maps with the number estimation branch.

4.2.1. Recording Setup

The multichannel ultrasound audio signals were recorded by 20 microphones (excluding broken ones) placed at regular intervals on the wall of the laboratory at the same height (1.2 m) from the floor level [7] as shown in **Fig. 1**. The sampling frequency was 500 kHz, and each channel of (for example) 60-s sound data was a time series of 500 kHz $\times$ 60 s = 30,000,000 points. Bats flew in this environment, and the task was to localize the coordinates of the bats from the 20-ch signals.

A calibrated stereo camera was installed at the edge of the room with a viewing angle of 30.4°H $\times$ 22.6°V. The bat locations in the stereo images were detected manually by operators or semi-automatically by a commercial tracking software. The 3D locations were recovered so that the origin of the coordinate system was almost the center of the room. The video and audio recordings were synchronized at the beginning with a mechanical trigger. The frame rate was 30 fps, and the bat locations were linearly interpolated to generate ground-truth coordinates.

4.2.2. Results for *R. ferrumequinum nippon*

In the data for *R. ferrumequinum nippon*, an ultrasound audio signal of a single flying bat was recorded for approximately 60 s. We evaluated the models in a hold-out setting, which is common for time series prediction, by using the fist 50 s for training and the remaining 10 s for the test.

Table 4 shows the root mean squared error (RMSE) of estimation for the training and test sets. For both coordinate and map estimations, the training errors increased with the simultaneous estimation, whereas the test errors decreased. This demonstrates that the single estimation is prone to overfitting the training set, probably due to the

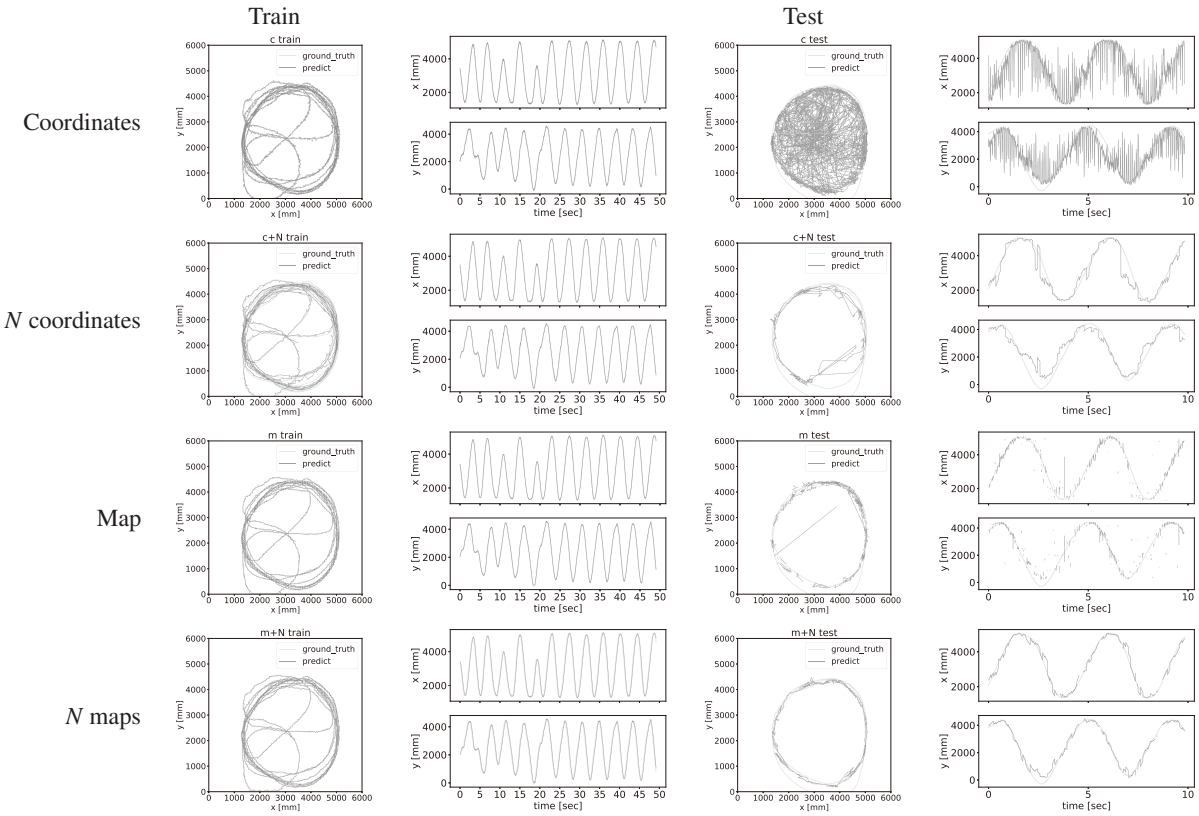

Fig. 6. Experimental results for *R. ferrumequinum nippon* (train and test). From top to bottom, the results of estimation of coordinates, simultaneous *N* coordinates, map, and simultaneous *N* maps. In the left, the estimated trajectories of training data are shown, whereas the trajectories of test data are shown in the right. In each setting, a 2D plot in *x*-*y* coordinates and two plots with *x*-*t* and *y*-*t* axes are shown. The predicted trajectories are shown in black, whereas the ground-truth trajectories are shown in gray.

silent clip issue. The *N* map estimation with the number branch performed the best for the test set. The error was approximately 221 mm, which is almost the minimum possible error because the bats move 10–20 cm in $1/30$ s as stated in Section 3.1.

The estimated trajectories are shown in **Fig. 6** for the training and test sets. We did not perform any post-processing (such as filtering); hence, the coordinate estimation suffers from temporal incoherency. In contrast, the simultaneous map estimation estimates the test trajectory with small visible errors. Note that peak detection often fails because a flat map without any maximum is predicted. Plots are missing in such cases, and map estimation is severely degraded by this issue, whereas *N* map estimation is less affected.

Note that the estimated trajectories of the test data miss the locations in the bottom part (lower part in the *y*-axis of the 2D plot). The ground-truth trajectory (gray curves) goes beyond the bottom of the plot, whereas the corresponding *y* coordinate of the predicted trajectories (black curves) does not follow. This discrepancy was due to the difference between the training and test data. A model learned from data is likely to overfit the lack of variation in location in the training data. However, this is inevitable in the case of our experiment because we can not control the flight of the bats. A possible approach is to use synthetically generated data with a large variety instead of real data with limited variety. We investigated the possibility of using this approach in the following subsection.

Table 5. Estimation errors (RMSE) in mm of coordinates, simultaneous *N* coordinates, map, simultaneous *N* maps, and with the number estimation branch for *M. fuliginosus*.

	Train
Coordinates	17.166
N coordinates	41.803
Map	29.553
N maps	83.632
N maps with number	81.508

4.2.3. Results for *M. fuliginosus*

In the data for *M. fuliginosus*, there were three bats in total. Initially, the first bat was released, and it flew for 30 s. Then, the second bat was released, and two bats flew for the next 20 s. Finally, the third bat was released, and three bats flew for 10 s. These data are too limited to be used for training in a hold-out setting; therefore, we report training errors only as a reference.

The training RMSEs are listed in **Table 5**. Similar to **Table 4**, using simultaneous *N* map estimation increased

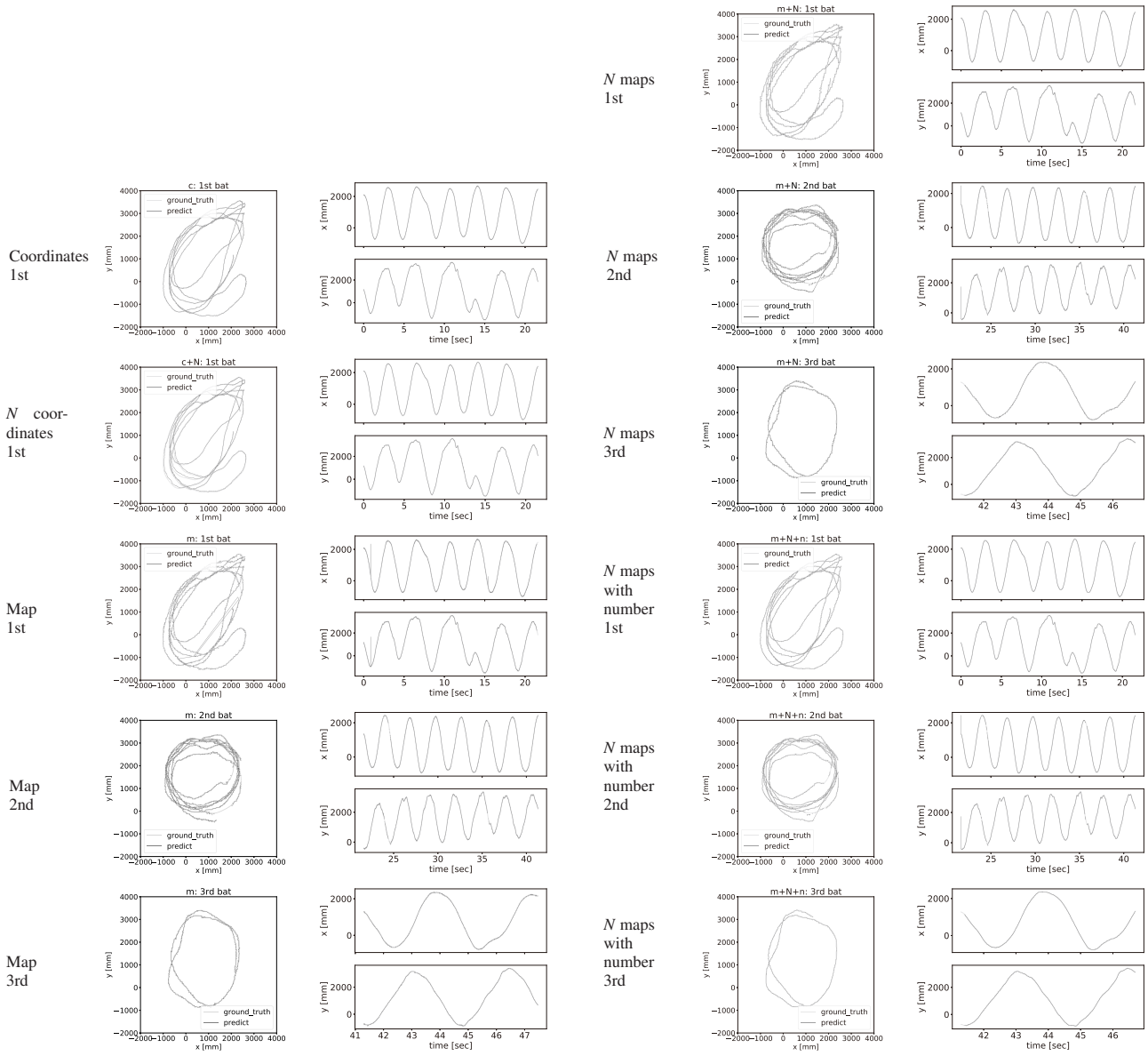

Fig. 7. Experimental results for *M. fuliginosus* (train only). From top to bottom and left to right, the results of estimation of coordinates, simultaneous *N* coordinates, map, simultaneous *N* maps, and simultaneous *N* maps with the number estimation branch for the 1st, 2nd, and 3rd bats, respectively. In each setting, a 2D plot in *x*-*y* coordinates and two plots with *x*-*t* and *y*-*t* axes are shown. The predicted trajectories are shown in black, whereas the ground-truth trajectories are shown in gray.

the training errors for both the coordinate and map estimations. However, adding the number estimation branch to the *N* map estimation performed better.

Figure 7 shows the estimated trajectories of the training set. For coordinate estimation, the trajectories are shown only for the first bat because a single pair of 2D coordinates can be predicted. Compared with **Fig. 6**, the trajectories are not in a regular circle, particularly for the first bat.

In reality, training errors do not reflect any aspect of test errors; however, annotation costs for these kinds of data are very high, and it is inevitable to use a small amount of data for analysis in ecology involving animals (as in our case) compared with big data in computer science. Therefore, we perform numerical simulations, as shown in the following subsection.

4.3. Numerical Simulation

Here, we describe the generation of synthetic pulses of bats. First, we prepared a single-channel clip of an ultrasound pulse for *R. ferrumequinum nippon* recorded using a nearby microphone. Then, we virtually made a bat fly in a 2D room of size 6 m $\times$ 6 m, in which microphones were set around the wall. The virtual bat was assumed to emit pulses inside the room at least 10 cm away from the wall at a random point inside the area of 5.9 m $\times$ 5.9 m, toward a random direction. To this end, we placed the

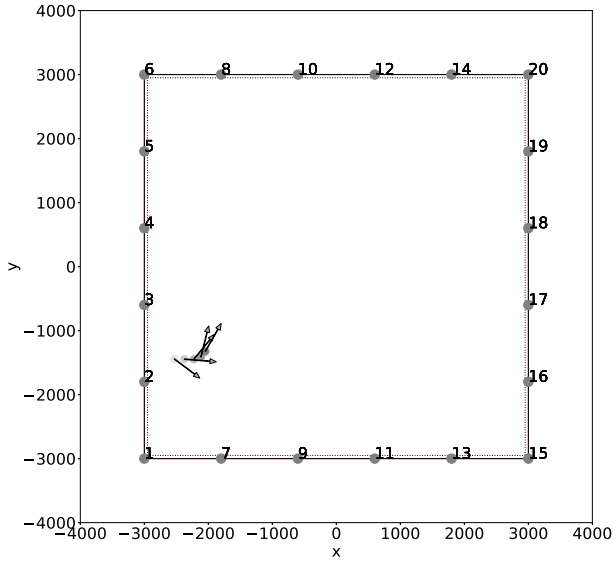

Fig. 8. An example of generated trajectories. The circles are bat locations, and the arrows show the pulse emission directions. The digits on the wall are microphone IDs.

Table 6. Estimation errors (RMSE) in mm in the numerical simulation.

	RMSE
N maps	70.748
N maps with number	85.494

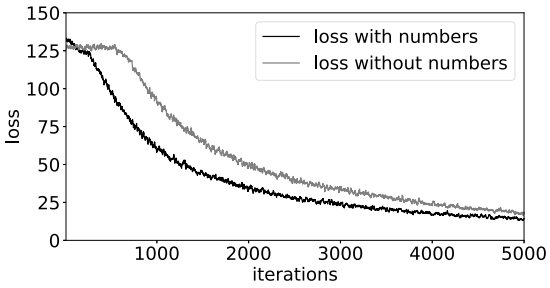

Fig. 9. Convergence of losses with and without the number branch.

sound source at the point and played the clip into the direction repeatedly for a certain duration.

The sound pulse was attenuated before being recorded by the microphones. We simulated the sound fall-off by using an attenuation with distance [32] as well as the directional fall-off of the pulse emission [33]. Let x be the angle from the emission direction, and the directional fall-off L_{dir} is represented by a Gaussian $L_{dir} = \exp\left(-\frac{x^2}{2\sigma^2}\right)$ with $\sigma = \frac{b}{2\sqrt{2\log 2}}$, where b is the fall-off width, which is set to 70° in this case. The attenuation is given by $L_{att} = L_{dif} * L_{abs}$, where L_{dif} is the diffusional decay $L_{dif} = r_0/r$ to the distance r with the reference $r_0 = 0.001$ m. L_{abs} is the absorption $L_{abs} = \exp\left(\frac{-r\alpha}{20\log_{10} e}\right)$ with a constant $\alpha = 1.5$ dB/m. The final fall-off of the signal is represented by $L_{foff} = L_{dir} * L_{att}$.

A virtually recorded signal for each microphone is the original signal at the emitting point, but attenuated by the final fall-off, and shifted in time for sound propagation at a speed of 340 m/s. In a sound clip of duration T, we moved the virtual bat by specifying five random points. The first point was randomly chosen in the valid area, and we then moved randomly toward the direction within an angle less than $\frac{1}{5}\pi$ from the previous moving direction (except for the first move). The distance to the next point was randomly sampled from a Gaussian with a mean of 114 mm and a standard deviation of 27 mm (these parameters were empirically chosen). The number of bats was randomly set to one or two. Note that, for simplicity, we ignored the Doppler effect (the flying speed was slow) and echoes from the walls (the pulse was short).

We generated 1000 synthetic clips for training, and estimated the trajectories of another 1000 clips for evaluation. An example is shown in **Fig. 8**. The RMSEs are listed in **Table 6**. This numerical simulation did not confirm the effectiveness of the number branch in terms of RMSE. Nevertheless, the addition of the branch considerably reduced the loss, as shown in **Fig. 9**, probably due to the help of the additional information from the number branch. This is promising for simulations with a large amount of synthetic data.

5. Conclusions

In this paper, we proposed CNN models to estimate bat locations by using multichannel ultrasound signals recorded using a microphone array. The experimental evaluation was limited; however, the experimental results showed that our map estimation approaches performed better than the coordinate estimation for real data of *R. ferrumequinum nippon*, as shown in the bottom row of **Fig. 6** and the rightmost column in **Table 4**. We plan to perform additional experiments to demonstrate the effectiveness of the proposed approach. Our approach is data-driven and automatically estimates the trajectories of flying bats from sound data. However, the limitation of this method is that we need a large and reasonable dataset for a better performance. In the current study, we also performed a numerical simulation to train the model with synthetically generated audio signals. We believe that extending this simulation would be helpful in training CNN models applicable to field settings, which will be our future work.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Numbers JP16H06540 and JP16H06542. We would like to thank Daisuke Ogawa for his help during the early stages of this study.

References:

[1] B. Tian and H.-U. Schnitzler, "Echolocation signals of the greater horseshoe bat (Rhinolophus ferrumequinum) in transfer flight and during landing," The J. of the Acoustical Society of America, Vol.101, pp. 2347-2364, 1997.

[2] K. Ghose, T. K. Horiuchi, P. S. Krishnaprasad, and C. F. Moss, "Echolocating bats use a nearly time-optimal strategy to intercept prey," PLOS Biology, Vol.4, No.5, 2006.

[3] E. Fujioka, I. Aihara, S. Watanabe, M. Sumiya, S. Hiryu, J. A. Simmons, H. Riquimaroux, and Y. Watanabe, "Rapid shifts of sonar attention by Pipistrellus abramus during natural hunting for multiple prey," The J. of the Acoustical Society of America, Vol.136, No.6, pp. 3389-3400, 2014.

[4] J. C. Koblitz, "Arrayvolution: using microphone arrays to study bats in the field," Canadian J. of Zoology, Vol.96, No.9, pp. 933-938, 2018.

[5] A.-M. Seibert, J. C. Koblitz, A. Denzinger, and H.-U. Schnitzler, "Scanning behavior in echolocating common pipistrelle bats (Pipistrellus pipistrellus)," PLOS ONE, Vol.8, No.4, pp. 1-11, 2013.

[6] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic, "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study," Proc. of the National Academy of Sciences, Vol.105, No.4, pp. 1232-1237, 2008.

[7] Y. Yamada, S. Hiryu, and Y. Watanabe, "Species-specific control of acoustic gaze by echolocating bats, Rhinolophus ferrumequinum nippon and pipistrellus abramus, during flight," J. of Comparative Physiology, A: Neuroethology, Sensory, Neural, and Behavioral Physiology, Vol.202, pp. 791-801, 2016.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol.24, No.4, pp. 320-327, 1976.

[9] R. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. on Antennas and Propagation, Vol.34, No.3, pp. 276-280, 1986.

[10] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," 1997 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Vol.1, pp. 375-378, 1997.

[11] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," 2018 IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 74-79, 2018.

[12] W. He, P. Motlicek, and J.-M. Odobez, "Joint localization and classification of multiple sound sources using a multi-task neural network," Proc. Interspeech 2018, pp. 312-316, 2018.

[13] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2386-2390, 2018.

[14] R. Takeda and K. Komatani, "Unsupervised adaptation of deep neural networks for sound source localization using entropy minimization," 2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2217-2221, 2017.

[15] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," IEEE Trans. on Industrial Electronics, Vol.65, No.8, pp. 6403-6413, 2018.

[16] S. Chakrabarty and E. Habets, "Multi-speaker localization using convolutional neural network trained with noise," Workshop on Machine Learning for Audio Processing (ML4Audio) at NIPS2017, 2017.

[17] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," J. of Robotics and Mechatronics, Vol.29, pp. 37-48, 2017.

[18] E. Ferguson, S. Williams, and C. Jin, "Sound source localization in a multipath environment using convolutional neural networks," 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2386-2390, 2018.

[19] T. Hirvonen, "Classification of spatial audio location and content using convolutional neural networks," 138th Audio Engineering Society Convention, Vol.2, 9294, 2015.

[20] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," 2018 26th European Signal Processing Conf. (EUSIPCO), pp. 1462-1466, 2018.

[21] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol.25, No.12, pp. 2444-2453, 2017.

[22] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting CNNS for improving acoustic source localization in noisy and reverberant conditions," IEEE Trans. on Emerging Topics in Computational Intelligence, Vol.2, No.2, pp. 103-116, 2018.

[23] W. Ma and X. Liu, "Phased microphone array for sound source localization with deep learning," Aerospace Systems, Vol.2, pp. 71-81, 2019.

[24] E. Thuillier, H. Gamper, and I. J. Tashev, "Spatial audio feature discovery with convolutional neural networks," 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 6797-6801, 2018.

[25] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," Computer Speech & Language, Vol.49, pp. 83-106, 2018.

[26] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," IEEE J. of Selected Topics in Signal Processing, Vol.13, No.1, pp. 34-48, 2019.

[27] J. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates," Sensors, Vol.18, No.10, 3418, 2018.

[28] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano, and M. Shah, "Generative adversarial networks conditioned by brain signals," Proc. of the IEEE Int. Conf. on Computer Vision, pp. 3410-3418, 2017.

[29] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah, "Deep learning human mind for automated visual classification," The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.

[30] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2017.

[31] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[32] B. D. Lawrence and J. A. Simmons, "Measurements of atmospheric attenuation at ultrasonic frequencies and the significance for echolocation by bats," The J. of the Acoustical Society of America, Vol.71, No.3, pp. 585-590, 1982.

[33] K. Motoi, M. Sumiya, E. Fujioka, and S. Hiryu, "Three-dimensional sonar beam-width expansion by Japanese house bats (Pipistrellus abramus) during natural foraging," The J. of the Acoustical Society of America, Vol.141, No.5, pp. EL439-EL444, 2017.

Name:
Kazuki Fujimori

Affiliation:
Signpost Corporation

Address:
4-12-20 Nihonbashi-honcho, Chuo-ku, Tokyo 103-0023 Japan

Brief Biographical History:
2018-2020 Hiroshima University
2020- Signpost Corporation

Main Works:
• K. Fujimori, B. Raytchev, K. Kaneda, E. Fujioka, S. Hiryu, and T. Tamaki, "Position estimation using multi-channel audio signals," ACML2019 workshop on Machine Learning for Trajectory, Activity, and Behavior (ACML-TAB), in conjunction with The 11th Asian Conf. on Machine Learning (ACML2019), WINC AICHI, Nagoya, Japan, November 17, 2019.

Name:
Bisser Raytchev

Affiliation:
Hiroshima University

Address:
1-4-1 Kagamiyama, Higashi-hiroshima, Hiroshima 739-8527, Japan

Brief Biographical History:
2000-2003 NTT Communication Science Laboratories
2003-2008 National Institute of Advanced Industrial Science and Technology (AIST)
2008- Hiroshima University

Main Works:
• B. Raytchev, A. Masuda, M. Minakawa, K. Tanaka, T. Kurita, T. Imamura, M. Suzuki, T. Tamaki, and K. Kaneda, "Detection of Differentiated vs. Undifferentiated Colonies of iPS Cells Using Random Forests Modeled with the Multivariate Polya Distribution," Springer Lecture Notes in Computer Science (LNCS), Vol.9901, Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2016, pp. 667-675, 2016.

Membership in Academic Societies:
• Association for Computing Machinery (ACM)
• The Association for Computational Linguistics (ACL)
• The Medical Image Computing and Computer Assisted Intervention Society (MICCAI)

Name:
Kazufumi Kaneda

Affiliation:
Hiroshima University

Address:
1-4-1 Kagamiyama, Higashi-hiroshima, Hiroshima 739-8527, Japan

Brief Biographical History:
1986- Hiroshima University
1991-1992 Visiting Researcher, Brigham Young University

Main Works:
• W. Yamamoto, B. Raychev, T. Tamaki, and K. Kaneda, "Spectral Rendering of Fluorescence using Importance Sampling," SIGGRAPH Asia 2018 Poster, Tokyo Int. Forum, Tokyo, Japan, December 4-7, 2018.

Membership in Academic Societies:
• Information Processing Society of Japan (IPSJ)
• The Institute of Electronics, Information and Communication Engineers (IEICE)
• Association for Computing Machinery (ACM)

Name:
Yasufumi Yamada

Affiliation:
Hiroshima University

Address:
1-3-2 Kagamiyama, Higashi-hiroshima, Hiroshima 739-0046, Japan

Brief Biographical History:
2017-2018 Research Associate, Doshisha University
2018-2021 Research Fellows, Hiroshima University
2021- Assistant Professor, Hiroshima University

Main Works:
• Y. Yamada, K. Ito, T. Tsuji, K. Otani, R. Kobayashi, Y. Watanabe, and S. Hiryu, "Ultrasound navigation based on minimally designed vehicle inspired by the bio-sonar strategy of bats," Advanced Robotics, Vol.33, No.3-4, pp. 169-182, 2019.

Membership in Academic Societies:
• Acoustical Society of Japan (ASJ)
• The Japan Society of Mechanical Engineers (JSME)
• Japanese Society for Mathematical Biology (JSMB)

Name:
Yu Teshima

Affiliation:
Doshisha University

Address:
1-3 Tatara-miyakodani, Kyotanabe, Kyoto 610-0394, Japan

Brief Biographical History:
2016- Ricoh Company, Ltd.

Main Works:
• Y. Teshima, T. Banda, Y. Mibe, and S. Hiryu, "Study of sensing strategy by visualization of bat recognition space," The 20th Conf. of the Society of Instrument and Control Engineers, System Integration Division (SI2019), Sunport Takamatsu, Kagawa, Japan, December 12-14, 2019.

Membership in Academic Societies:
• Acoustical Society of Japan (ASJ)
• The Society of Instrument and Control Engineers (SICE)

Name:
Emyo Fujioka

Affiliation:
Doshisha University

Address:
1-3 Tatara-miyakodani, Kyotanabe, Kyoto 610-0321, Japan

Brief Biographical History:
2013-2014 Researcher, JST FIRST Aihara Innovative Mathematical Modelling Project
2014-2016 Assistant Researcher, Doshisha University
2016- Post-doctoral Research Fellow, Doshisha University

Main Works:
• E. Fujioka, I. Aihara, M. Sumiya, K. Aihara, and S. Hiryu, "Echolocating bats use future-target information for optimal foraging," Proc. of the National Academy of Sciences, Vol.113, No.17, pp. 4848-4852, 2016.

Membership in Academic Societies:
• Acoustical Society of America (ASA)
• Acoustical Society of Japan (ASJ)
• The Society for Bioacoustics (SFBA)
• The Marine Acoustic Society of Japan (MASJ)
• Japanese Society of Bio-Logging Science (BLS)

Name:
Shizuko Hiryu

Affiliation:
Doshisha University

Address:
1-3 Tatara-miyakodani, Kyotanabe, Kyoto 610-0321, Japan

Brief Biographical History:
1999-2005 IBM Japan
2007-2008 JSPS Research Fellow (PD), The University of Tokyo
2008-2012 Assistant Professor, Doshisha University
2012-2017 Associate Professor, Doshisha University
2014-2018 JST PRESTO Researcher
2017- Professor, Doshisha University

Main Works:
• S. Hiryu, M. E. Bates, J. A. Simmons, and H. Riquimaroux, "FM echolocating bats shift frequencies to avoid broadcast-echo ambiguity in clutter," Proc. of the National Academy of Sciences, Vol.107, pp. 7048-7053, 2010.

Membership in Academic Societies:
• Acoustical Society of America (ASA)
• Acoustical Society of Japan (ASJ)
• Marine Acoustic Society of Japan (MASJ)
• Japan Ethological Society (JES)
• Japanese Society of Bio-Logging Science (BLS)

Name:
Toru Tamaki

Affiliation:
Nagoya Institute of Technology

Address:
Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555, Japan

Brief Biographical History:
2001-2005 Niigata University
2005-2020 Hiroshima University
2015-2016 Visiting Researcher, École Supérieure d'Ingénieurs en Électrotechnique et Électronique (ESIEE Paris)
2020- Nagoya Institute of Technology

Main Works:
• K. Terao, T. Tamaki, B. Raytchev, K. Kaneda, and S. Satoh, "An Entropy Clustering Approach for Assessing Visual Question Difficulty," IEEE Access, Vol.8, pp. 180633-180645, doi: 10.1109/ACCESS.2020.3022063, 2020.

Membership in Academic Societies:
• The Institute of Electronics, Information and Communication Engineers (IEICE)
• Information Processing Society of Japan (IPSJ)
• The Institute of Electrical and Electronics Engineers (IEEE)