Fine-scale observations of spatio-spectro-temporal dynamics of bird vocalizations using robot audition techniques

Shinji Sumitani, Reiji Suzuki, Shiho Matsubayashi, Takaya Arita, Kazuhiro Nakadai & Hiroshi G. Okuno

Graduate School of Informatics, Nagoya University, Chikusa, Nagoya, Aichi, Japan
Graduate School of Engineering, Osaka University, Suita, Osaka, Japan
Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan
Honda Research Institute Japan Co., Wako, Saitama, Japan
Graduate School of Fundamental Science and Engineering, Waseda University, Shinjuku, Tokyo, Japan

Keywords
Bird songs, ecoacoustics, robot audition, sound source localization, soundscape, t-SNE

Abstract
Ecoacoustics needs sophisticated acoustic monitoring tools to extract a wide level of features from an observed mixture of sounds. We have developed a portable acoustic monitoring system called 'HARKBird' which consists of a laptop PC and an inexpensive commercial microphone array with the robot audition software HARK. HARKBird can extract acoustic events in a recording, and we can obtain the begin and end timings, the spatial information (e.g., position or direction from the microphone array), and the spectrogram of the sound separated from the original recording. In this study, we report how robot audition techniques contribute to monitoring spatio-spectro-temporal dynamics of bird behaviors, using an extended and minimal system based on multiple microphone arrays. The dimension reduction of separated sounds is important to integrate the information from multiple microphone arrays. As a dimension reduction algorithm, we use t-SNE to help manual annotation of each sound and to generate the vocalization distribution automatically. We conduct playback experiments to Spotted Towhee (Pipilo maculatus) to simulate different cases of territorial intrusions (song/call/no playback). Our hypothesis in playback experiments is that playback of conspecific vocalizations would invoke aggressive responses of males against song playbacks and the effects would be more prominent than those of call playbacks. Our primary aim is to test whether our system can extract the necessary information on the aggressiveness of target individuals to examine our hypothesis. We show the system with manual annotation of vocalizations can extract their different spatio-spectro-temporal dynamics in different conditions, which supported our hypothesis. We also consider the spectral affinity-based automatic matching of localized sounds from different microphone arrays. The relative number of localized songs depending on the playback conditions reflected a similar trend to those in the manual approach, implying that we can grasp the long-term dynamics of vocalizations without costly annotations.

Introduction
Ecoacoustics is an interdisciplinary science that investigates natural and anthropogenic sounds and their relationship with the environment over multiple scales of time and space (Farina and Gage 2017). Songbirds are appropriate target taxa for the purpose of ecoacoustic researches (Gasc et al. 2016; Stowell 2018) because their vocalizations (1)
can provide a suite of useful information about the environment for monitoring, (2) enable individuals to interact in complex ways, behaving as complex systems of communication (Suzuki and Cody 2019; Stowell et al. 2016), and (3) have a rich and complex variety of structures (Catchpole and Slater 2008), which can be used as benchmarks for solving classification and clustering challenges (e.g., BirdCLEF (Goëau et al. 2016), DCASE (Stowell et al. 2018)).

Recent progress in the development of an autonomous recording unit (ARU) enabled us to obtain long-term acoustic data that can benefit such ecoacoustic research (Shonfield and Bayne 2017). There have been several approaches such as single-channel sound separation based on deep neural networks (Kong et al. 2019) and a distance estimation of sound sources using sound levels (Yip et al. 2019). However, it is still not easy to obtain the spatio-temporal information of multiple sound events (e.g., distance or direction) or the structure of soundscape using single-channel (monaural) microphone.

Sound source localization techniques using microphone arrays have been recently investigated as a promising approach (Blumstein et al. 2011; Goëau et al. 2016). There have been various studies on using microphone arrays to localize bird vocalizations because the spatial position of vocalizations has been recognized as an important information in both bioacoustics and ecoacoustics (Mennill et al. 2006, 2012; Collier et al. 2010; Harlow et al. 2013; Stepanian et al. 2016; Araya-Salas et al. 2017; Gayk and Mennill 2019). Hedley et al. (2017) discussed that the direction of arrivals (DOA) estimation using a microphone array may improve the ability to assess abundance in biodiversity surveys by facilitating more accurate localization of sounds. However, microphone arrays are still not widely used by field researchers because of the limited availability of both software and hardware.

Based on such scientific and engineering motivations described above, we believe the robot audition techniques. Robot audition technique concept aims to realize the recognition of noisy speech such as simultaneous speech using microphones embedded on the robot (Nakadai et al. 2000). Thus, it is obvious that the techniques can be applicable to ecoacoustics field that investigates complex acoustic events in natural sounds. However, there have been few attempts so far. We are developing a portable system, HARKBird, to localize birdsongs (Suzuki et al. 2017; Sumitani et al. 2019). HARKBird can extract acoustic events in a recording, and we can obtain the begin and end timings, the spatial information (e.g., position or direction from the microphone array), and the spectrogram of the sound separated from the original recording. HARKBird consists of a standard laptop PC with an open-source software for robot audition HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) (Nakadai et al. 2017), combined with a low-cost and commercially available microphone array. In previous research, we have shown effects of conspecific song playback on the directional changes and song-types of an individual of Japanese Bush Warbler Horornis diphone (Suzuki et al. 2018b) using a single 8-channel microphone array named TAMAGO (System in Frontier Inc.). We also conducted spatial localization of song posts of Great Reed Warblers Acrocephalus arundinaceus using three 16-channel microphone arrays named DACHO (System in Frontier Inc.) developed by the research group of HARK (Suzuki et al. 2018a).

More recently, we proposed a real-time 2D bird song localization system with two microphone arrays and showed that online localization was useful because appropriate locations to place microphone arrays based on the real-time localization results can be found (Sumitani et al. 2018). An off-line analysis with manual annotations of vocalization types based on t-SNE (t-distributed Stochastic Neighbor Embedding (Van der Maaten and Hinton 2008)) (Sumitani et al. 2019) could extract the spatial distribution of bird songs using a short recording in a preliminary playback experiment on a single individual of songbird (Sumitani et al. 2018). However, it is still unclear whether such a portable and easily deployable system can be used to obtain sufficient information to observe the spatial dynamics of bird behaviors depending on ecological and biological conditions such as habitat and territorial structures.

The purpose of this paper is to further discuss how robot audition techniques contribute to monitoring spatio-spectro-temporal dynamics of bird behaviors, focusing on the dimension reduction of separated sounds, which is a critical problem to integrate the information from multiple microphone arrays. This paper is separated into two parts. In the first part (approach), we show detailed analyses on the dynamics of individual behaviors under different conditions of song or call playback on multiple individuals. This was designed to show that the system with two microphone arrays, with manual annotation of sounds using t-SNE, can quantitatively extract changes in the spatial, temporal and spectral patterns between playback conditions.

Our hypothesis in playback experiments is that a playback of conspecific vocalizations would invoke aggressive responses of males against song playbacks and the effects would be more prominent than those of call playbacks. This is because song playbacks are often used to simulate an intrusion by a rival (Catchpole and Slater 2008; Harlow et al. 2013). Our primary aim in the experiments is to test whether our system can extract the necessary
information on the aggressiveness of target individuals to examine our hypothesis. The distance from a loudspeaker is one of the popular measures to estimate this degree (McGregor 1992; Catchpole and Slater 2008), and our system is expected to be able to extract more detailed information in spatial and temporal dimensions.

In the second part (approach), an automatic extraction of the properties of the soundscape from long-term recordings is introduced. This is essential for understanding the global dynamics of the soundscape and biodiversity. In addition to common indices for measuring biodiversity (e.g., acoustic complexity index; Pieretti et al. 2011), new measures are proposed to capture the components of the whole soundscape using novel theoretical methods such as sparse coding (Eldridge et al. 2016). The sound source localization approach using robot audition techniques is expected to be beneficial in the sense that we can obtain separated sounds directly from the directional information of localization. Thus, we further consider fully automatic extraction of the soundscape around microphones in the proposed system. We adopt t-SNE in a different way, which is simple but one of the highly winning techniques for dimension reduction. t-SNE is used in various applications to the spatial visualization of sound clips including bird vocalizations (Tan and McDonald 2017). We use this to estimate the affinity between the spectrograms of separated sounds to avoid mismatched localization of simultaneously occurring sounds, to exclude noises, and to visually represent the spectral property of each sound.

Specifically, we conducted playback experiments on two individuals of Spotted Towhee Pipilo maculatus, SPTO in California with several types of replayed vocalizations, as simulated and varied ecological environments to simulate different cases of territorial intrusions, using two microphone arrays as a minimal and portable observation system. We then extracted the spatio-spectro-temporal patterns of the target individuals with a help of an annotation tool based on the feature map of localized sounds created by t-SNE, which varied under different playback conditions. We further compare the obtained distributions from the latter automatic extraction method with the ones obtained from the former method based on manual classification.

Materials and Methods

Playback experiments

We conducted playback experiments to observe responses of individual SPTOs in response to playback sounds replayed around their territories at our field site in a mixed conifer-oak forest near Volcano, California, USA (38°48'N, 120°63'W), May 2018. We focused on this species because they have several types of short songs and also have a cat-like mew call, which are simple and clear and thus expected to be localized easily compared to the vocalizations of other species inhabit at the same site.

We used a laptop PC (TOUGHBOOK CF-C2; Panasonic), two 8-channel microphone arrays (TAMAGO; System in Frontier Inc., Tokyo, Japan (http://www.sifi.co.jp/en/)), and a loudspeaker (A3143; ANKER) for the 2D localization and playback (Fig. 1). In this figure, a white and egg-shaped device on a tripod is a microphone array we used in this study. It has eight channels placed in a circular manner. We used two microphone array units that were connected with a single laptop PC. We implemented a system that can record with two microphone arrays simultaneously for 2D localization, which is run on a single laptop by slightly modifying the Python scripts of HARKBird. Each microphone array was connected to the laptop via USB cable and placed on the top of a tripod. We adopted Ubuntu Linux 16.04 in which HARK and hark-python were installed to execute sound source localization processes using HARKBird.

We focused on two individuals of Spotted Towhee in this site and they were neighboring individuals inhabited approximately 150 m away from each other. For each individual, we deployed a system as shown in Figure 2. These places were the territories of these individuals. The distance between each microphone array was approximately 23.2 m and the loudspeaker approximately 15.4

Figure 1. The HARKBird system composed of a laptop PC, microphone arrays, and a loudspeaker. A white and egg-shaped device on a tripod is a microphone array unit we used in this study. It has 8 channels placed in a circular manner. We used two microphone array units that were connected with a single laptop PC. There is one microphone array in this picture although there is another one outside the picture.
and 15.5 m away from each microphone array in the case of individual A. For individual B, the distance between each microphone array was approximately 13.9 m and the loudspeaker approximately 14.6 and 8.0 m away from each microphone array. We determined deployment positions of the system so that the bird vocalizations could be localized clearly by using the real-time 2D localization function of the system in advance, like active monitoring.

In terms of song playback, we adopted the recordings of a song (‘Song1’ in Fig. 3(a)) and a call (‘Call’ in Fig. 3(k)) of individual A (not B), which were recorded in advance. The noise reduction and normalization were applied to the wave files of both recordings using Audacity 2.1.2 (http://audacityteam.org/). The pressure level of each playback sound was approximately 83.5 dB measured from 1 m away, which was intended to make it sound well for the target individual. It was expected to be biologically plausible but slightly larger than the original pressure levels of SPTO’s vocalizations. We designed six types of experimental settings: song playback to individual A (SA, 60 min), call playback to individual A (CA, 60 min), song playback to individual B (SB, 60 minutes), call playback to individual B (CB, 60 min) and no-playback for each individual (NA, 25 min and NB, 45 min). In no playback conditions (NA and NB), we used the shorter recordings in NA and NB because we could only obtain the shorter continuous recordings than the ones in playback conditions due to some troubles in the system (e.g., cable connection error or unexpected stop of recording). We replayed the two consecutive songs or calls with a 120-sec interval during each experimental period. We observed 10 types of songs and 1 type of call as shown in Figure 3, which were classified by the experimenter by visual and acoustic inspection.

**DOA estimation and sound source separation using HARKBird**

We used HARKBird to estimate the DOA of sound sources. HARKBird is a collection of python scripts that enable us to record using microphone arrays connected to the laptop PC, and localize and annotate sound sources in recordings using HARK. This runs on Ubuntu Linux in which HARK, HARK-Python, PySide, etc. are installed (see Nakadai et al. (2017); Suzuki et al. (2017); Sumitani et al. (2019). The codes and documents are available from websites (http://www.alife.cs.i.nagoya-u.ac.jp/~reiji/HARKBird/) and GitHub (https://github.com/HARKBird-project/HARKBird/).

The sound source localization algorithm is based on the MUltiple SIgnal Classification (MUSIC) method (Schmidt 1986) using multiple spectrograms with the Short Time Fourier Transformation (STFT). We extracted all separated localized sounds as wave files (16 bit, 16 kHz) using GHSSS (Geometric High-order Decorrelation-based Source Separation) method (Nakajima et al. 2010).
We specified the parameters of HARKBird\(^1\) in order not to fail to localize the vocalizations of the target individuals, although the parameters also enabled to localize faint noises and the vocalizations of some other songbirds.

HARKBird also provides several ways to visualize and analyze the results of localization as showing an example result of sound source localization and separation of a recording with one microphone array (Fig. 4). From these visualization results, we can see whether the vocalizations of the target individuals are localized properly. We use the time, duration, DOA and sound clip of each localized sound for 2D localization of sources.

### 2D sound source localization using t-SNE

We adopted a modified version of a simple 2D sound source localization algorithm based on triangulation of DOAs of each sound from two arrays used in Sumitani et al. (2018), and the position is shown on an aerial photograph (Fig. 5; see Sumitani et al. (2018)).

Figure 6(A) shows the schematic image of the method of two-dimensional localization using a pair of DOAs. In this

\(^{1}\)THRESH \(= 25.0\), NUM SOURCE \(= 1\), UPPER BOUND FREQUENCY \(= 8,000\) and LOWER BOUND FREQUENCY \(= 2,200\).
method, we calculated the intersection between the half straight lines issued from the positions of arrays toward the DOAs and regarded the position of the intersection as the spatially localized position (Figure 6(A) (top)). We adopted the localized position at the middle of the successfully localized duration during which the intersection existed, as illustrated in Figure 6(A) (bottom).

There are two major problems to be solved in conducting 2D sound source localization using multiple microphone arrays in fields. One is to identify the pair of DOAs of a single source from different arrays and exclude cases of mislocalizations in which two microphone arrays localized different sound sources at one time and the system misrecognized them as a unique source. This is not straightforward when many sound sources were localized at one. The other is to exclude many noises localized by each array that are inevitable in field recordings when we estimate the spatial positions of vocalizations. We adopt the t-SNE (Van der Maaten and Hinton 2008) to solve these problems simultaneously by creating the feature map of separated sounds. The primary feature of t-SNE is to represent the proximity between two neighboring data points on the space by a probability distribution and it is especially useful for visualizing high-dimensional data.

In this paper, we examine two methods to determine a pair of sound sources (DOA) localized by different arrays, of which DOAs are further used for a triangulation process to estimate the spatial position of a sound source. We use the distribution of sounds on a feature map based on t-SNE for this purpose.

A manual classification and 2D localization based on an interactive feature map

In the first method, we used a manual classification of sounds (DOA) localized by each array. Figure 6(B) illustrates the schematic images of feature maps, and Algorithm 1 shows the pseudo-code of the method. First, for each recording from each array, we created a 100 × 64 pixels image of the STFT spectrogram of each separated sound and regarded all spectrograms as a dataset. We created a two-dimensional feature map of this dataset with t-SNE and plotted data points on the feature map to visualize their distribution. Note that we had two feature maps for each array as illustrated in Figure 6(B). We adopted the sci-kit-learn library for Python for conducting t-SNE, and used a grid search for appropriate parameter settings of the algorithm to divide sound sources into clusters.

An interactive annotation tool shown in Figure 7 enabled users to classify similar sounds by surrounding them with a frame (dotted polygon) and labeling the name of their class (Sumitani et al. 2018). To examine the properties of each sound, the user can replay the sound and visualize its spectrogram by clicking the corresponding dot on the map. We classified sounds into each vocalization type (‘Songs 1–10’ and ‘Call’), the playback sound from the loudspeaker (‘Playback’) and others (‘None’) on the feature map.

Then, we conducted 2D birdsong localization by combining the two classification results. We extracted all pairs of sounds such that they were in the same class but from different maps and had any overlap in their localized time.
frames. We regarded the sounds of each pair was from a unique sound source, and their DOAs were used for the 2D spatial localization of this source as introduced above.

**A spectral affinity-based 2D localization**

The other method is to use the feature map to estimate the affinity between the spectrograms of separated sounds extracted from different arrays to avoid mismatched localization of simultaneously occurring sounds. A schematic image of this map is illustrated in Figure 6(C), and Algorithm 2 shows the pseudo-code of the method. In this case, we regard all images of separated sounds from a pair of recordings by two microphone arrays as a large dataset for a feature map. Thus, we use a single map that has all the sounds from two arrays (Figure 6(C)).

To determine the pairs of sounds to be used for 2D localization, we calculated the distance on the feature map between the sounds from the different arrays that have any overlap in their localized time frames. We regarded the pair of which the distance was within \( L \), which is the parameter that determines the threshold of the affinity for being recognized as two sounds are from a unique source. Note that the two-dimensional localization of each pair was performed in the temporal order of their (DOA) localization events, and the pairs that have any sounds used for 2D localization were excluded from

---

**Figure 5.** A snapshot of real-time 2D localization. The color gradient represents the accumulated values of the MUSIC spectrum over the two microphones, assuming that each value radiated from each microphone array according to the value of the MUSIC spectrum and the corresponding direction toward each location (blue: low, red: high). Map data ©2019 Google.

**Figure 6.** An image of the 2D sound source localization. (A) 2D sound source localization based on triangulation on a geographic space. (B) A manual classification based on a spectral feature map. (C) A spectral affinity-based classification. In this example, a sound source is localized at A1-B1 because (B) both of sources A1 and B1 are classified into the same class and or (C) the distance between separated sources A1 and B1 is smaller than \( L \) on the feature map, and A1 and B1 temporally overlapped.
candidate pairs, to avoid duplicated use of sound source information.
We expect that this method can exclude unnecessary noises as well as identify the positions of vocalizations

**Algorithm 1 A manual classification-based 2D localization**

```plaintext
for each recording from microphone array do
  Conduct the sound source localization with a time frame of 0.2 sec. and obtain the DOA and duration of each localized sound source using HARKBird (MUSIC method)
  Extract the separated sound of each source using HARKBird (GHDSS method)
  for each microphone array do
    Make a dataset (100 x 64 grey-scaled images) for t-SNE using the separated sounds
    Conduct dimension reduction with t-SNE and obtain the distribution of sound sources on the two-dimensional feature map.
    Classify the separated sounds into vocalization types and playback sounds using the interactive feature map
  for each pair of the separated sounds from different maps (arrays) that have any overlap in their localized time frames do
    if the classified labels are the same then
      Estimate the spatial position by using the pair of DOAs of the separated sounds
      Adopt the localized position at the middle of the successfully localized duration
    Create a spatial distribution of sound sources on the geographical space using the 2D localized results.
```

**Algorithm 2 A spectral affinity-based 2D localization**

```plaintext
for each recording from microphone array do
  Conduct the sound source localization with a time frame of 0.2 sec. and obtain the DOA and duration of each sound source using HARKBird (MUSIC method)
  Extract the separated sound of each source using HARKBird (GHDSS method)
  Make a dataset (100 x 64 grey-scaled images) for t-SNE using all of the separated sounds obtained from both of the microphone arrays
  Conduct dimension reduction with t-SNE and obtain the distribution of sound sources on the two-dimensional feature map.
  for each pair of the separated sounds from different arrays that have any overlap in their localized time frames, in the temporal order of their localized events do
    Calculate the distance between the separated sources on the feature map
    if the distance < \( L \in \{1, 2, 3\} \) and both of the separated sounds have not been used for 2D localization yet then
      Estimate the spatial position by using the pair of the DOAs of the separated sounds
      Adopt the localized position at the middle of the successfully localized duration
    Create a spatial distribution of sound sources on the geographical space using the 2D localized results
```

because noises were generally widely distributed on the feature map and the distance between sources tends to be large. Furthermore, we mapped the x and y values of each localized source on the 2D color space and plotted each source with the corresponding color to reflect its approximated spectral feature.

**Analysis**

After an overview of the results of the manual classification using the feature map of sound sources, we examine the spatio-spectro-temporal dynamics of vocalizations of the target individuals extracted using the manual classification described above to grasp the whole tendency of their behaviors. We analyzed the localization error using the positions of playback sounds from the loudspeaker. We also compared their behavioral patterns between playback conditions focusing on the number of vocalization types and their frequency, the average distance of vocalizations from the loudspeaker and the amount of directional changes between sequentially occurring vocalizations to see whether our hypothesis is correct or not.

Then, we examined the spatial distribution of vocalizations extracted using the spectral affinity-based localization, and discuss the accuracy of successful localization of the vocalizations of the target individuals, regarding the localization results of the former approach as the approximation of the ground truth. The data on the localization results and scripts for analyses are in Data S1 and S2.
Results

A manual classification and the spatio-spectro-temporal dynamics of bird vocalizations

Figure 8 shows the results of the manual classification on the feature map of sound sources localized by a microphone array (Mic. 0) in each condition. In most of the experimental settings, the vocalizations of the target individual and playback sounds formed small clusters (‘Songs 1–10’, ‘Call’ and ‘Playback’). In this study, we classified the other sounds into ‘None’, which mostly composed of a large giant cluster of noises or surrounding clusters of songs of other species (e.g., songs of Blackheaded Grosbeak *Pheucticus melanocephalus*) which were
Figure 9. The spatio-spectral-temporal dynamics of vocalization of the individual A. (A) the spatial distribution of vocalization types, (B) the histogram and (C) the change of the distance of vocalizations from the loudspeaker and their types, and (D) the histogram and (E) the change of direction between sequentially occurring vocalizations from the loudspeaker and their types.
Figure 10. The spatio-spectral-temporal dynamics of vocalization of the individual B. (A) the spatial distribution of vocalization types, (B) the histogram and (C) the change of the distance of vocalizations from the loudspeaker and their types, and (D) the histogram and (E) the change of direction between sequentially occurring vocalizations from the loudspeaker and their types.
singing throughout experimental periods. It is likely this was caused by the settings of HARK in this experiment because it tried to localize many sources of noise to ensure the vocalizations of the target individuals were captured. Different types of songs were divided into different clusters on the feature map. It helped us to classify vocalization types easily. However, some vocalizations of the target species were contained within the large clusters of other sources (None), which is presumably attributed to the fact that the individual sang far in the distance and thus the sound was faint. We also found that the clusters were formed more clearly when the number of vocalizations was large. It indicates that such interactive annotation tools based on the feature map do not work well without a large number of target sound sources in case of that there are many noises in the recording.

Figure 11. The statistics of vocalizations in different playback conditions. (A) the localization error using the positions of playback sounds from the loudspeaker, (B) the histogram of the vocalization types, (C) the average distance of vocalizations from the loudspeaker, and (D) the average directional change from the loudspeaker between sequentially occurring vocalizations. SA/CA: song/call playback to the individual A, SB/CB: song/call playback to the individual B, NA/NB: no playback for the individual A/B. Each bar in (A), (C) and (D) indicates the standard deviation.
Using the results of manual classification explained above, we obtained spatio-spectro-temporal dynamics of vocalizations of the target individuals as shown in Figures 9–11. The spatial distributions of vocalizations (Figures 9(A) and 10 (A)) show that we could grasp the approximate positions of the vocalizations of individuals A and B, and their vocalization types. They tended to sing around the loudspeaker but also kept a distance from it. The average distance from the loudspeaker in each playback condition was as follows: SA: 13.5 m, CA: 27.8 m, NA: 17.4 m, SB: 15.7 m, CB: 26.1 m and NB: 18.8 m. In addition, the temporal patterns of the distance and direction of vocalizations from the loudspeaker (Figures 9 and 10(B–E)) show their typical singing tendency, that is, the target individuals sang the same type of vocalizations several times repeatedly while changing their position.

The behavioral responses of individuals A and B depended on the playback conditions. In Figure 11, both individuals A and B tended to (1) sing more different types of songs more frequently, (2) sing in close proximity to the loudspeaker, and (3) showed a greater change in position measured by the direction from the loudspeaker when songs were replayed (SA and SB), compared with the cases when calls were replayed (CA and CB). While those behavioral differences were not statistically significant (the bootstrap estimates \( P < 0.05 \) based on 100,000 bootstrap resamplings), they were consistent between the target individuals A and B. In no playback conditions, the number of songs was small for the individual B, while remained unaffected for individual A after taking the experimental period into account.

In addition, these results implied that individual B was more strongly affected by the playback than individual A. This behavioral difference could be caused by the vocalizations of individual A which was used for both experiments. For the individual B, it is speculated that this experimental condition was like a situation that a neighboring individual had invaded into his own territory because playback sounds we used were the vocalizations of individual A; hence, he might have moved around and sang many types of vocalizations frequently.

---

Figure 12. Spatial distribution of sound sources based on the spectral affinity-based localization. (A) The distribution of spatial positions calculate from the all possible pairs of DOAs at the same time frame. (B) The distribution of extracted pairs using the method \((L = 2)\). (C) The distribution of sound sources on the feature space. Each plot is colored according to the position of the middle point between paired sound sources on the feature space and its corresponding color on the color map (D).

© 2020 The Authors. Remote Sensing in Ecology and Conservation published by John Wiley & Sons Ltd on behalf of Zoological Society of London.
Table 1. The accuracy (affinity-based/manual annotation) of the spectral affinity-based 2D localization.

| Playback condition | Localized vocalizations (affinity-based) | Localized vocalizations (manual annotation) | Accuracy (%) |
|--------------------|------------------------------------------|----------------------------------------------|--------------|
| SA                 | 20                                       | 261                                          | 8.0          |
| 2.0                | 72                                       | 276                                          | 27.6         |
| 3.0                | 118                                      | 45.2                                         |              |
| CA                 | 0.5                                      | 40                                           | 35.1         |
| 0.75               | 54                                       | 47.4                                         |              |
| 1.0                | 59                                       | 51.8                                         |              |
| NA                 | 0.5                                      | 20                                           | 29.4         |
| 0.75               | 26                                       | 38.2                                         |              |
| 1.0                | 32                                       | 47.1                                         |              |
| SB                 | 1.0                                      | 100                                          | 36.4         |
| 2.0                | 135                                      | 49.1                                         |              |
| 3.0                | 165                                      | 60.0                                         |              |
| CB                 | 1.0                                      | 9                                            | 5.1          |
| 2.0                | 93                                       | 53.1                                         |              |
| 3.0                | 149                                      | 85.1                                         |              |
| NB                 | 0.5                                      | 0                                            | 0            |
| 1.0                | 0                                        | 0                                            |              |
| 1.5                | 0                                        | 0                                            |              |

We analyzed the localization error using the positions of playback sounds from the loudspeaker, which is shown in Figure 11(A). These were SA: 0.867 ± 0.616 m, CA: 1.613 ± 0.494 m, SB: 1.315 ± 0.813 m, and CB: 1.510 ± 0.26 m. Thus, the localization error was within 2.0 m.

We also see that the variation of the distance from the loudspeaker was very large during around 2300–2600 s in the case of CB (Figure 10). This is not natural because the variation occurred in a very short time. These are expected to be caused by the localized positions of vocalizations. They were observed at the vicinity of the straight line connecting the microphones and the localized position sometimes fluctuated between the true position of sources.

**Spectral affinity-based localization**

In Figure 12, we examined the extent to which the sound environment around microphone arrays could be extracted without classification. Here, we focus on the case of SB because the target individual sang and moved more actively compared with other experimental conditions as the vocalization count and the directional changes in Figure 11(B) and (D). Figure 12(A) shows the spatial distribution of the 2D localized positions for all possible pairs of sound sources from two microphones in the same time frame. There was a significant number of mismatched and unnecessary sound sources such as environmental noises or vocalizations of other species. On the other hand, the spectral affinity-based 2D localization with $L = 2$ in Figure 12(B) clearly reflects the distribution of vocalizations with the manual annotation (Figure 10(A), top). The distribution of colors of the localized sound sources, determined by the color map (Figure 12(D)) on the feature map, also show that there were loosely clustered regions of similar colors, reflecting the spectral structure of the acoustic environment (i.e., song-types).

To evaluate this localization method, we calculated the accuracy of the extraction of vocalizations of the target individuals, assuming that the extracted vocalizations with the manual annotation were ground truth. Table 1 shows that the accuracy (the number of localized vocalizations of the target species with this method/that with the manual annotation) strongly depends on the threshold for the affinity $L$ between sources. While the accuracy was not so high (53.1%), the distribution of vocalizations visually appeared to be the most similar to the one with the manual annotation when the parameter was small ($L = 2$), while the accuracy was higher (85.1%) when $L = 3$. This is because $L$ needed to be relatively small to exclude many other noises and vocalizations of other species (e.g., Black-headed Grosbeak) and make the distribution of the target vocalizations more visible. However, considering that the number of localized sound sources with each microphone array in each condition was about 3000–4000, this method can drastically suppress the localization of unnecessary sounds and indicate the soundscape without human annotations. It is also found that this method might not work sufficiently if there were few target vocalizations (NB).

**Discussion**

We proposed an integrated framework for bird song localization based on a robot audition technique (HARK) with microphone array units (an array of array units), and showed the framework will be useful for the fine-scale observation of spatio-spectro-temporal dynamics of bird vocalization. We evaluated our system by conducting playback experiments on Spotted Towhee, and discussed whether and how the effects of playback conditions could be measured quantitatively.

In particular, we considered two types of approaches and methods. In the first approach, we used an interactive interface to manually but less costly extract songs of the focal individuals, which enabled us to visually inspect a distribution of sound sources on their feature map derived from a dimension reduction algorithm (t-SNE), and manually label clusters of sources of interest (songs of the target individuals). We observed aggressive responses of target individuals in playback conditions,
which were consistently more prominent for songs than calls. These effects were quantitatively estimated by the vocalization frequency and types; and the distance and the direction from the loudspeaker to the individual. In the second approach, we used an affinity-based automatic matching of DOA localized sounds from different microphone arrays. The relative number of localized songs depending on the playback conditions reflected the similar trend to those in the first approach. This implies the possibility that we can grasp the long-term dynamics of bird vocalizations without costly annotations.

An essential difference from the other related works is that the proposed framework provides all the processes necessary for this approach: recording with multiple sounds, localizing and separating sound sources, and manually or automatically classifying sound sources, to extract spatio-temporal dynamics of bird songs from real and noisy data. Our experimental results and analyses clarified that all these components are essential for observing fine-scaled or long-term dynamics of bird vocalizations in fields, while there are possibilities for improving each component.

Recently, more attention has been paid to the use of autonomous recording units (ARU) to observe birds in the field, which is reflected on the increasing number of studies that compare observation results based on ARUs and direct field observation on foot (Digby et al. 2013; Vold et al. 2017; Abrahams 2019; Darras et al. 2019). It is reported that autonomous sound recording outperforms human observers to monitor birds and allow investigations at higher temporal and spatial resolution and coverage than human observers alone, and thus considered more cost-effective (Darras et al. 2019). However, using a microphone array unit, which is composed of a single unit with many channels, is still not adopted very often even though such array units are getting more easily available (e.g., ReSpeaker for Raspberry Pi (Seed Studio)) due to the evolution of IoT (Internet of Things)-related devices.

Among a few approaches that focused on use of the similar types of microphone array units in fields (Harlow et al. 2013; Tiete et al. 2014; Hedley et al. 2017), Collier et al. (2010) deployed a wireless sensor network of eight self-developed and four-channel microphone array units with self-surveying facility which can automatically determine node locations, called VoxNet, in a tropical rain forest and successfully localized the songs of Antbirds Formicarius moniliger in a convex area of nodes with the substantially smaller localization errors of vocalizations (0.199 m), compared with the ones of playback sounds in our experimental trials. We expect that extending our system to such a networked one with self-surveying that can deploy nodes in a large area can contribute to the better performance. We are developing a networked version based on nodes composed of a Raspberry Pi 3 and a USB microphone array (TAMAGO; System in Frontier Inc.). Preliminary observations showed that the position of nodes strongly affected the localization results. However, at the same time, the improvement of localization accuracy under a minimal configuration with two microphone arrays connected to a laptop would be of important, considering the portability and deployability of the system.

There have been some approaches for 3D bird song localization (Harlow et al. 2013; Stepanian et al. 2016; Gabriel et al. 2018; Gayk and Mennill 2019). Recently, Gayk and Mennill (2019) recently developed a system based on time-of-arrival differences and composed of multiple and wireless and ARU units with a height of about 7 m for three-dimensional triangulation of flying warblers on the wing. This implies the potential of extending our framework to 3D localization of flying birds with smaller systems based on 3D DOA estimation with HARK. Gabriel et al. conducted a preliminary case study of 3D localization of singing birds on a tree using HARK and multiple 16-channel version of self-developed microphone array units, each can estimate both azimuth and elevation of sound sources (Gabriel et al. 2018), showing successful localization of some bird vocalization on a tree.

Song playback experiments (Harlow et al. 2013) and speaker tests (Hedley et al. 2017) have been used to evaluate birdsong localization systems based on microphone arrays. Harlow et al. applied Collier et al.’s system to 3D localization of birds in a Mexican rainforest (Harlow et al. 2013). They successfully estimated territories and 3D movements of bird pairs of White-breasted Wood Wrens Henicorhina leucosticta, using the system composed of eight microphone array nodes. Our results show that our minimal system composed of two arrays can be used for such purpose to quantitatively observe bird behaviors against song playbacks.

The application of dimension reduction algorithms or latent space visualization of communication signals is also getting much attention in the wide field of bioacoustics. Sainburg proposed a package for this purpose, called animal vocalization generative network (AVGN) (Sainburg et al. 2019). They showed the various examples of the use of the latent space such as discrete latent projections of animal vocalizations and temporally continuous latent trajectories. We showed another type of application of the use of this type of methods. This is to solve practical problems that can occur when we use multiple microphone arrays in noisy or complex acoustic environments while these have not been discussed in previous related works. In our framework, the manual annotation tool of vocalization types worked effectively both for noise removal and vocalization extraction of the target
individuals. This implies that t-SNE is an efficient algorithm for understanding the global sound environment from recording. Applying related techniques such as variation autoencoder (VAE) (Kingma and Welling 2013) or uniform manifold approximation and projection for dimension reduction (UMAP) (McInnes and Healy 2018) to generate the feature map of localized sounds might contribute to the improvement of the annotation tool. In addition, we considered a fully automatic extraction of behavioral patterns in the soundscape around microphone arrays. Our proposed method succeeded in eliminating the mislocalization of non-existent sound sources that can occur in two-dimensional localization when multiple sound sources existed at the same time. The sound sources with clear features gathered densely by t-SNE, while the non-featured sound sources such as noise are distributed sparsely on the feature map, and thus the noise can be removed automatically by this method.

While the improvement of each component of the framework is necessary, our approaches can be extended to further experimental researches in different directions. The first approach based on manual classification can contribute to quantitatively obtain the spatio-spectral-temporal dynamics of bird vocalizations in bioacoustic researches such as playback experiments in complex soundscape environments or observations of social interactions among multiple individuals or species in fields. The second approach based on affinity-based localization can contribute to grasping the long-term dynamics of vocalizations without manual annotation costs. A visualization and detection of species in large audio data, by coloring spectrograms using a combination of various acoustic indices, has been discussed (Towsey et al. 2018). By color-coding in the feature map, we could grasp the sound environment around the system. The fact our method can acquire such complex information of the soundscape automatically is a great advantage and it could be used as a soundscape observation method that can be adapted to ecoacoustics. For example, we can create new acoustic indices that can reflect its spatio-spectral-temporal structure using the approximated and long-term information from the distribution of sound sources (e.g., the variation or entropy in DOA of localized sound sources).

In sum, we believe both approaches can contribute to the understanding of biodiversity in various time scales in view of acoustic behaviors of songbirds in fields.

Acknowledgments
We thank Charles E. Taylor, Martin L. Cody, and Yi-Ju Wang (UCLA) for supporting field experiments and comments. This work was supported in part by JSPS/MEXT KAKENHI Grant Numbers: JP16K00294, JP17H06841, JP18K11467, JP19KK0260, and JP17H06383 in #4903 (Evolinguistics).

References
Abrahams, C. 2019. Comparison between lek counts and bioacoustic recording for monitoring western capercaillie (tetrao urogallus L.). J. Ornithol. 160, 685–697.
Araya-Salas, M., K. Wojcizulas-Jakubas, E. M. Phillips, D. J. Mennill, and T. F. Wright. 2017. To overlap or not to overlap: context-dependent coordinated singing in lekking long-billed hermits. Anim. Behav. 124, 57–64.
Blumstein, D., D. J. Mennill, P. Clemins, L. Giros, K. Yao, G. Patricelli, et al. 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J. Appl. Ecol. 48, 758–767.
Catchpole, C. K. and P. J. B. Slater. 2008. Bird song: biological themes and variations. Cambridge University Press, Cambridge.
Collier, T. C., A. N. G. Kirschel, and C. E. Taylor. 2010. Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. J. Acoust. Soc. Am. 128, 182–189.
Darras, K., P. Batry, B. J. Furnas, I. Grass, Y. A. Mulyani, and T. Tschamntke. 2019. Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide. Ecol. Appl. 29, e01954.
Digby, A., M. Towsey, B. D. Bell, and P. D. Teal. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. Methods Ecol. Evol. 4, 675–683.
Eldridge, A., M. Casey, P. Moscoso, and M. Peck. 2016. A new method for ecoacoustics? Toward the extraction and evaluation of ecologically-meaningful soundscape components using sparse coding methods. Peerf 4, e2108.
Farina, A. and S. Gage. 2017. eds. Ecoacoustics: the ecological role of sounds. John Wiley and Sons, Hoboken, NJ.
Gabriel, D., R. Kojima, K. Hoshiba, K. Itoyama, K. Nishida, and K. Nakadai. 2018. Case study of bird localization via sound in 3d space. In Proceedings of the 36th Annual Conference of the Robotics Society of Japan, RSJ2018AC1I2–06.
Gasc, A., D. Francomano, J. B. Dunning, and B. C. Pijanowski. 2016. Future directions for soundscape ecology: the importance of ornithological contributions. Auk 134, 215–228.
Gayk, Z. G., and D. J. Mennill. 2019. Pinpointing the position of flying songbirds with a wireless microphone array: three-dimensional triangulation of warblers on the wing. Bioacoustics 134, 1–12.
Goëau, H., H. Glotinm, W. P. Vellinga, R. Planqué, and A. Joly. 2016. LifeCLEF Bird Identi-fi-cation Task 2016: the arrival of Deep learning. LifeCLEF 2016 working notes.

Harlow, Z., T. C. Collier, V. Burkholder, and C. E. Taylor. 2013. Acoustic 3d localization of a tropical songbird. Pp. 220–224 in China Summit and International Conference on Signal and Information Processing (ChinaSIP).

Hedley, R. W., Y. Huang, and K. Yao. 2017. Direction-of-arrival estimation of animal vocalizations for monitoring animal behavior and improving estimates of abundance. Avian Conserv. Ecol. 12, 6.

Kingma, D. P., and M. Welling. 2013. Auto-encoding variational bayes. arXiv preprint, page arXiv:1312.6114.

Kong, Q., Y. Xu, W. Wang, P. J. B. Jackson, and M. D. Plumbley. 2019. Single-channel signal separation and deconvolution with generative adversarial networks. Pp. 2747–2753 in Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI2019).

McGregor, P. K. 1992. Quantifying responses to playback: one, many, or composite multivariate measures? Pp. 79–96 in P. K. McGregor, ed. Playback and studies of animal communication Vol. 6. Springer US, Boston, MA.

McInnes, L., and J. Healy. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. ArXiv e-prints 1802.03426.

Mennill, D. J., M. Burt, K. M. Fristrup, and S. L. Vehrencamp. 2006. Accuracy of an acoustic location system for monitoring the position of duetting songbirds in tropical forest. J. Acoust. Soc. Am. 119, 2832–2839.

Mennill, D. J., M. Battiston, and D. R. Wilson. 2012. Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. Methods Ecol. Evol. 3, 704–712.

Nakadai, K., T. Lourens, H. Okuno, and H. Kitano. 2000. Active audition for humanoid. Pp. 832–839 in Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.

Nakadai, K., H. Okuno, and T. Mizumoto. 2017. Development, deployment and applications of robot audition open source software HARK. J. Robot. Mechatron. 27, 16–25.

Nakajima, H., K. Nakadai, Y. Hasegawa, and H. Tsujiino. 2010. Blind source separation with parameter-free adaptive step-size method for robot audition. IEEE Trans. Audio Speech Lang. Process. 18, 1476–1485.

Pieretti, N., A. Farina, and D. Morri. 2011. A new methodology to infer the singing activity of an avian community: the Acoustic Complexity Index (ACI). Ecol. Ind. 11, 868–873.

Sainburg, T., M. Thiellk, and T. Q. Gentner. 2019. Latent space visualization, characterization, and generation of diverse vocal communication signals. bioRxiv, 870311.

Schmidt, R. 1986. Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. 34, 276–280.

Shonfield, J., and E. M. Bayne. 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conserv. Ecol. 12, 14.

Stepanian, P. M., K. G. Horton, D. C. Hille, C. E. Wainwright, P. B. Chilson, and J. F. Kelly. 2016. Extending bioacoustic monitoring of birds aloft through flight call localization with a three-dimensional microphone array. Ecol. Evol. 6, 7039–7046.

Stowell, D. 2018. Computational bioacoustic scene analysis. Pp. 303–333 in T. Virtanen, M. D. Plumbley, and D. Ellis, eds. Computational analysis of sound scenes and events, chapter 11. Springer International Publishing, Cham, Switzerland.

Stowell, D., L. Gill, and D. Clayton. 2016. Detailed temporal structure of communication networks in groups of songbirds. J. R. Soc. Interface 13, 20160296.

Stowell, D., H. Glotin, Y. Styliamou, and M. Wood. 2018. Bird audio detection in dcase2018 challenge. http://dcase.challenges2018/task-bird-audio-detection

Sumitani, S., R. Suzuki, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno. 2018. Extracting the relationship between the spatial distribution and types of bird vocalizations using robot audition system hark. Pp. 2485–2490 in Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems.

Sumitani, S., R. Suzuki, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno. 2019. An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques – HARKBird 2.0. Pp. 8246–8250 in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing.

Suzuki, R., and M. L. Cody. 2019. Complex systems approaches to temporal soundspace partitioning in bird communities as a self-organizing phenomenon based on behavioral plasticity. Artif. Life. Robot. 24, 439–444.

Suzuki, R., S. Matsubayashi, R. W. Hedley, K. Nakadai, and H. Okuno. 2017. HARKBird: exploring acoustic interactions in bird communities using a microphone array. J. Robot. Mechatron. 27, 213–223.

Suzuki, R., S. Matsubayashi, F. Saito, T. Murate, T. Masuda, K. Yamamoto, et al. 2018a. A spatiotemporal analysis of acoustic interactions between great reed warblers (acrocephalus arundinaceus) using microphone arrays and robot audition software hark. Ecol. Evol. 8, 812–825.

Suzuki, R., S. Sumitani, S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno. 2018b. Field observations of ecoacoustic dynamics of a Japanese bush warbler using an open-source software for robot audition hark. J. Ecoacoust. 2, EYA146.

Tan, M., and K. McDonald. 2017. Bird sounds: 2D visualization of thousands of bird vocalizations. https://expe-riments.withgoogle.com/bird-sounds
Tiete, J., F. Domnguez, B. da Silva, L. Segers, K. Steenhaut, and A. Touha. 2014. Sound-compass: a distributed MEMS microphone array-based sensor for sound source localization. *Sensors* **14**, 1918–1949.

Towsey, M., E. Znidersic, J. Broken-Brow, K. Indraswari, D. M. Watson, Y. Phillips, et al. 2018. Long-duration, false-colour spectrograms for detecting species in large audio data-sets. *J. Ecoacoust.* **2**, IUSWUI.

Van der Maaten, L. and G. Hinton. 2008. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605.

Vold, S., C. M. Handel and L. B. McNew. 2017. Comparison of acoustic recorders and field observers for monitoring tundra bird communities: acoustic monitoring of tundra-breeding birds. *Wildl. Soc. Bull.* **41**, 566–576.

Yip, D. A., E. C. Knight, E. H. Audet, S. J. Wilson, C. Charchuk, C. D. Scott, et al. 2019. Sound level measurements from audio recordings provide objective distance estimates for distance sampling wildlife populations. *Remote Sens. Ecol. Conserv.* https://doi.org/10.1002/rse2.118.

**Supporting Information**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Data S1.** The localization results and scripts for analysis based on manual annotation.

**Data S2.** The localization results and scripts for analysis based on an affinity-based localization.