A robot audition approach toward understanding social interactions among songbirds in a semi-free flight environment

Shinji Sumitani[1], Reiji Suzuki[1], Kazuhiro Wada[2], Takaya Arita[1], Kazuhiro Nakadai[3,4] and Hiroshi G. Okuno[5]

[1]Graduate School of Informatics, Nagoya University, Japan
[2]Department of Biological Sciences, Faculty of Science, Hokkaido University, Japan
[3]Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Japan
[4]Honda Research Institute Japan Co., Ltd., Japan
[5]Graduate School of Creative Science and Engineering, Faculty of Science and Engineering, Waseda University, Japan

[1]sumitani@alife.cs.is.nagoya-u.ac.jp

Abstract: We discuss the applicability of robot audition approach (i.e., sound source localization and separation with microphone arrays) to understand social and acoustic interactions in a population of songbirds as a complex system. In order to better extract natural spatio-temporal-spectral patterns of vocalizations using our bird song localization system called HARK-Bird, we developed a semi-free flight experimental environment which consists of an out-door mesh tent with nests, perches, and microphone arrays. We conducted preliminary recordings of vocalizations of a few Zebra Finch (*Taeniopygia guttata*) in the environment, and triangulation of the direction of arrivals of sources that have similar spectral features measured with t-SNE, using a pair of arrays spatially suitable for localization of sources in the corresponding area. We found that vocalizations were localized at around several positions close to perches and nests, and also observed multiple individuals sang each other in a short time period. This means that we can extract such fine-scaled data to understand the group-level dynamics of their vocalizations.

Keywords: songbird, social interaction, robot audition, t-SNE, Zebra Finch

1 INTRODUCTION

Songbirds communicate diversely using various types of vocalizations. Songs, which are relatively long and complex vocalizations, are used for territorial defense and courtship toward females [1], and calls, which are short and simple ones, are used to exchange more specific information such as warning of predators and also used to form social bonds. Because of their diversity in acoustic communications in temporal, spatial and spectral dimensions, songbirds are regarded as suitable species of researches in ecoacoustics, which is an interdisciplinary science that investigates natural and anthropogenic sounds and their relationship with the environment over multiple scales of time and space [2].

We are interested in a population of songbirds which communicate with each other via vocalizations as a complex system and have tried to understand the interactions and the structures of their societies. For example, we focused on the interspecific temporal overlap avoidance of singing behaviors based on the acoustic niche hypothesis (ANH) [3]. We discussed the effects of the length of species-specific songs on the evolution of the temporal overlap avoidance by combining a computational experiment of the coevolution of behavioral plasticity and behavioral observation of wild songbirds in California [4, 5], implying that the species who have longer songs can become a driver species that tends to dominate the temporal dynamics of acoustic interactions. However, it is not easy to obtain the information for analyzing such detailed and complex acoustic interactions by conventional recordings using a single microphone or by human observation.

Based on these backgrounds, we are developing a portable system, HARKBird, to localize birdsongs in fields, which automatically extracts sound sources and the information such as their direction of arrivals (DOA) or their timings [6, 7]. HARKBird consists of a standard laptop PC with an open-sourced software for robot audition HARK, combined with a low-cost and commercially available microphone array. We showed effects of conspecific song playback on the directional changes and song-types of the vocalizations of a Japanese Bush Warbler (*Horornis diphone*) using a single 8-channel microphone array [8]. We also conducted spatial localization of song posts of two Great Reed Warblers (*Acrocephalus arundinaceus*) using three 16-channel microphone arrays [9], which clarified the asymmetric effects from one individual to another in temporal overlap avoidance of their songs as well as successful estimation of their song posts.

Recently, there are also researches for understanding intraspecific social interactions among a larger population of songbirds. Some researchers created their communication networks focusing on their calls by extracting call events when multiple individuals vocalized at one time, using autonomous recording devices deployed in fields [10] or on-bird recording devices [11, 12]. Gill et al. investigated the call communication among Zebra Finch (*Taeniopygia guttata*, ZF) by using recording with on-bird microphone transmitters and they revealed their vocal interactions change based on breeding stages. Although their method enabled us to obtain cooccurrences of vocalizations comprehensively, the on-bird recording is invasive and difficult to obtain their spatial information while the use of vocalizations can depend on their spatial situations [13].

The purpose of this research is to discuss the applicabil-

Fig. 1: An experimental environment with HARKBird.

ity of robot audition approach to the understanding of the social and acoustic interactions in a population of songbirds in fields, focusing on such intra-specific communication dynamics of vocalizations in a group of ZF. This species is well known as a model animal for vocal learning and uses various calls to form social relationships as explained above. Thus, we can discuss various roles of the spatio-temporal contexts of their vocalizations in non-invasive conditions.

In order to obtain sufficient information about bird populations while keeping the condition as natural as possible, we adopted a mesh tent as a semi-free flight experimental environment. We arranged nest boxes, perches and multiple microphone arrays in the tent and released several individuals of ZF and conducted recordings. Then, by using the localized results from each microphone array, we estimated the spatial distribution of vocalizations of ZF individuals, which showed the potential of our approach as we could obtain the fine-scaled data to understand the group-level dynamics of their vocalizations.

2 METHOD

2.1 Experimental setting and recording

We developed a recording environment in an approximately 7m × 7m rectangular mesh tent (Fig. 1) located in a field on the campus of Hokkaido University (43°04’18.3”N, 141°20’28.4”E). In this environment, we arranged 5 nest boxes and 4 perches at approximately 1.5m above the ground, and food and water space were provided near the center of the tent as shown in Fig. 2, which is a sufficient environment for a few songbirds can fly around freely. Then, we placed five microphone arrays (TAMAGO-03, System in Frontier Inc.) connected to a laptop PC (TOUGHBOOK CF-C2; Panasonic) and arranged them at the four corners and the center of the tent.

We released several individuals of ZF and conducted recording trials from June to September 2019. The preliminary data analysis in this paper focused on the 10 minutes recording recorded on June 3, 2019, in which 5 male individuals were released and singing in the experimental environment. In this recording, their social relationships might not be stable because the recording started less than one hour after the release of the individuals in the tent. Thus, we expected to observe active interactions by using vocalizations and movements.

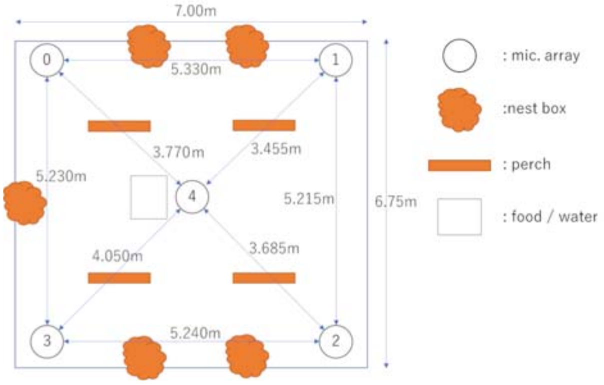

Fig. 2: The arrangement of the experimental environment.

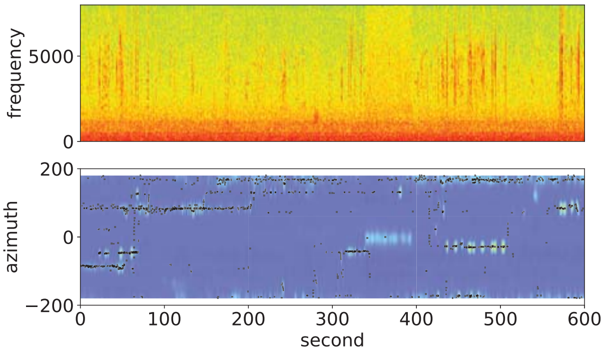

Fig. 3: The visualization of localization based on microphone array 4. The top panel represents the spectrograms with STFT. The bottom panel represents the MUSIC spectrum (heat map) and the distribution of the DOA and duration of localized sounds (black bars).

2.2 Spatial division-based 2D localization

We conducted localization and separation of sound sources in the recording with each microphone array by using HARKBird and obtained the DOAs and timings of localized sound sources. The sound source localization algorithm was based on the MUltiple SIgnal Classification (MUSIC) method [14] using multiple spectrograms with the Short-Time Fourier Transformation (STFT). We extracted all localized and then separated sounds as wave files (16 bit, 16 kHz) using GHDSS (Geometric High-order Decorrelation-based Source Separation) [15]. See [6, 7] for more details of HARKBird[1] and [16] for HARK. We adjusted the parameters for HARK to localize vocalizations of ZF individuals as much as possible[2]. As a result, we obtained the localized results for a single recording from each microphone array such

[1] http://www.alife.cs.i.nagoya-u.ac.jp/˜reiji/HARKBird/

[2] PERIOD = 5, THRESH = 28, UPPER_BOUND_FREQUENCY = 8000, LOWER_BOUND_FREQUENCY = 3000, NUM_SOURCE = 2

Algorithm 1 Spatial division-based 2D localization

```
Initialize an empty list $Lp$ for keeping localized positions.
Initialize an empty list $Ls$ for keeping separated sounds.
for each pair of microphone array (the centered and one of the peripheral arrays) do
    Make a dataset for t-SNE using the separated sounds by two arrays.
    Conduct t-SNE and obtain the source distribution on the feature space.
    for each sound source $S_i$ from the peripheral array do
        Find a separated sound $S_c$ from the centered array which is the closest to $S_i$ on the feature space and has any overlap of localization time to $S_i$.
        Calculate the distance between $S_i$ and $S_c$ as $d_i$.
        if $d_i < 10$ then
            Localize the position of the source $p_i$ using their DOAs.
            if $p_i$ is in the corresponding area of the pair then
                Append $p_i$ to $Lp$
                Append the separated sound of $S_c$ to $Ls$
Create a 2D localization map using $Lp$.
Make a dataset for t-SNE using $Ls$.
Conduct t-SNE using $L_s$, and obtain the source distribution on the feature space.
```

as Fig. 3. At the period where the signal of the spectrogram around 2000–6000 Hz appears strongly in the top panel, the many black bars, each indicating a localized source, are displayed in the bottom panel. It was confirmed by human inspection that these localized sounds included the almost all vocalizations of ZF individuals sang during the recording.

In order to estimate the spatial positions of vocalizations, we used a simple 2D sound localization method based on the triangulation of DOAs of sound sources, localized by two microphone arrays [17]. Algorithm 1 shows the procedures of 2D localization and analyses we adopted in this preliminary analysis. In our previous research, we observed that the localization of sources near the straight line connecting the two microphone arrays was difficult and not stable. To solve this problem, we divided the whole space in the experimental tent into several areas and chose a pair of microphone arrays for each area (Fig. 4). In each area, we adopted the pair of the microphone array 4 and another one which corresponds to the color of the area. This spatial division-based 2D localization enabled us to estimate more appropriate positions of sound sources because each area was relatively close to the corresponding pair of microphone arrays as well as to avoid localization of sources at difficult positions explained above.

Furthermore, when multiple sources were localized at once, we need to specify a pair of DOA information of a unique source from two microphone arrays. Here, we used a dimension reduction algorithm t-SNE [18] in order to select the pair of separated sounds to be localized which have similar spectral features. We created grey-scaled 100 × 64 pixel images of STFT spectrogram of separated sounds from two microphone arrays as a data set for dimension reduction with t-SNE and then plotted them into a 2D plane of resultant feature space. Then, for each spectrogram of the separated sound source obtained by the peripheral microphone array (0-3), we searched for the separated sound localized by the centered microphone array 4 which is the closest to the focal source on the feature space. We conducted the 2D localization using the pair of these sources if the distance between them was within 10 on the feature space. From these localization results, we extracted the sound sources which were localized in the corresponding area of the pair. This method enabled us to estimate spatial positions using the information of a unique source without explicit classification.

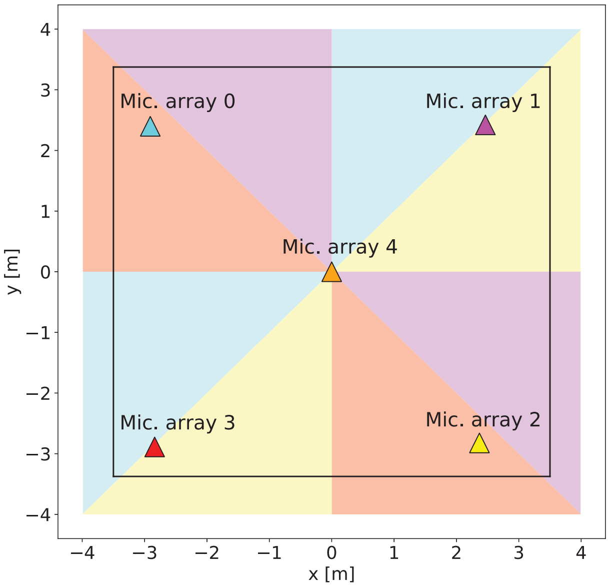

Fig. 4: The spatial division of the tent area for 2D localization. For each area, we conduct 2D localization by using the pair of microphone array 4 and another one which corresponds to the area color.

After estimating the spatial positions of sound sources with each pair, we integrated them and obtained the final result of 2D localization. In order to eliminate the sound sources from the outside of the tent, we limited the localization range to a square, 4m each side, centered on the center of the tent. We regarded them as the vocalizations of ZF individuals and extract the acoustic features by conducting dimension reduction with t-SNE using the separated sound sources obtained by the recording from microphone array 4 which were localized in 2D localization phase. By combining the localization results and the acoustic features, we finally obtained the integrated spatial, temporal and spectral dynamics of vocalizations among ZF individuals.

3 RESULTS

3.1 Spatial and temporal dynamics

Fig. 5 shows the spatial distribution of localized vocalizations of ZF individuals. The graph shows that sound sources were localized at around several positions close to perches and nests while some sound sources were also localized at the places where there were no objects because some individuals sang on the ground. We also see that the temporal change of distribution of localized sound sources. For example, the sound sources during the initial 100 sec., depicted with blue dots, were localized at 5 or more positions, while the sound sources during 400-500 sec., depicted with orange

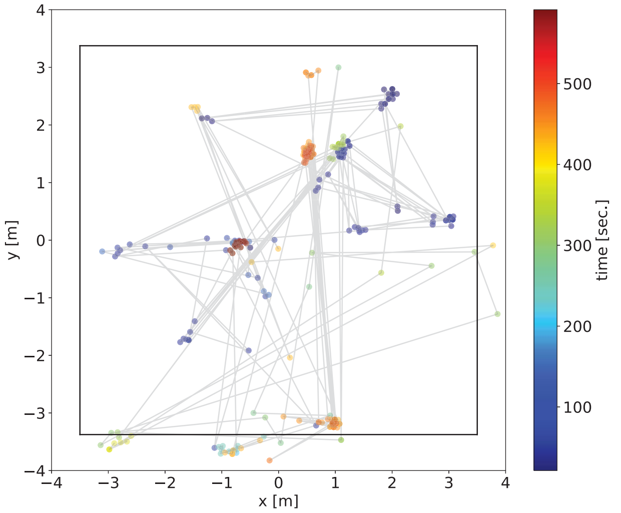

Fig. 5: The spatial distribution of localized sound sources. The color corresponds to the localized time represented by the color bar. Each localized sound source is connected to the subsequent one with a grey line.

dots, were mainly localized at other 2 positions.

In addition, we represented the temporal changes in localized positions by connecting a position of the localized source and its subsequent one with a link. We observe that there were many links between two clusters (orange dots) during 400-500 sec. We confirmed that some individuals sang in this time period by detailed analyses which will be explained later. Thus, it implies some ZF individuals interacted with each other repeatedly by using their vocalizations.

On the other hand, some sound sources were localized at the out of the tent, which were observed in the lower left in Fig. 5. It can be caused by the calibration error of the arrangement of the microphone array 3 because all of the sound sources localized at the area by using the microphone array 3 were slightly far from a nest or the edge of the tent.

3.2 Spectral relationships

Fig. 6 represents the feature distribution of sound sources obtained by the final integration of localization results. The localized sound sources were distributed widely on the feature space. We also observe that some of them formed clusters, each of which was composed of many sources that had similar acoustic property in spectrograms as illustrated in Fig. 6. This implies that the distribution reflected the acoustic property of sound sources. Furthermore, we confirmed that almost all localized sound sources in the feature space were vocalization types of ZF individuals although some noises such as food-pecking sound were included. It can help us to understand the acoustic properties and contexts of vocalizations of ZF individuals and their relationships.

3.3 Spatial, temporal and spectral dynamics

We combined the spatio-temporal dynamics (Fig. 5) and the spectral properties (Fig. 6) of vocalizations of ZF individuals to see their overall spatial, temporal and spectral dynamics. Fig. 7 shows when and where ZF individuals sang and what types of vocalizations they used. The source distribution in Fig. 7 (bottom) was similar to the one of the centered microphone array in Fig. 3 (bottom), and thus this result reflects the whole dynamics of vocalizations in the tent. We can see changes in their group-level behaviors from this figure. For example, the sound sources were localized at multiple directions and distances during 0-100 and 300-500 sec., implying that some ZF individuals vocalized at various positions or moved actively. On the other hand, the sound sources were localized at around similar directions and distances during 100-200 and 500-600 sec. This situation can be presumed that a single or a smaller number of ZF individuals vocalized at a unique position. It also should be noted that the acoustic properties of localized sources tended to correlate with such spatio-temporal dynamics in that spatially or temporally closer sources tended to share their spectral properties (i.e., their colors) at least in part.

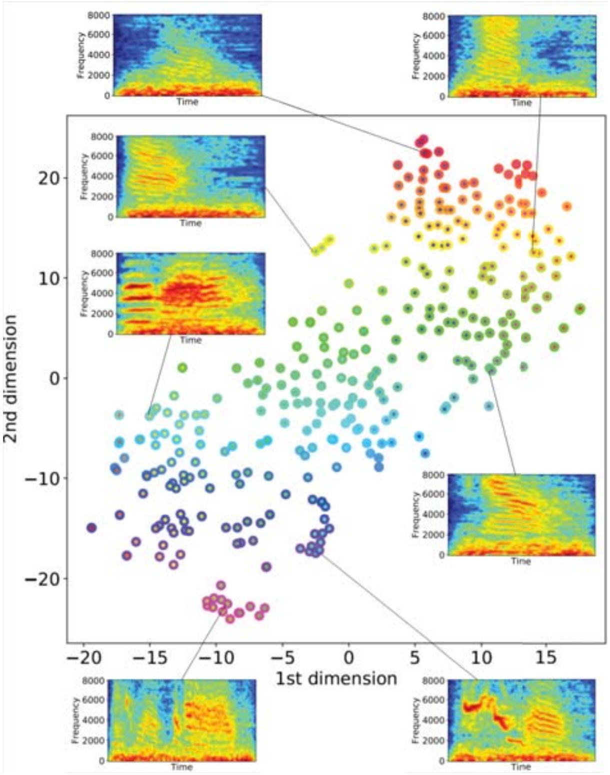

Fig. 6: The distribution of localized sound sources from the recording with microphone array 4 in the feature space of t-SNE. All of them were localized in 2D localization phase. The face color of each dot represented by hsv color map corresponds to the position at 1st dimension. The edge color represented by gist-rainbow color map corresponds to the position at 2nd dimension.

We further investigated two different patterns of the former acoustic interactions in more detail. Fig. 8 and 9 show durations in which multiple ZF individuals vocalized at multiple directions from the centered array: around 135 deg., -30

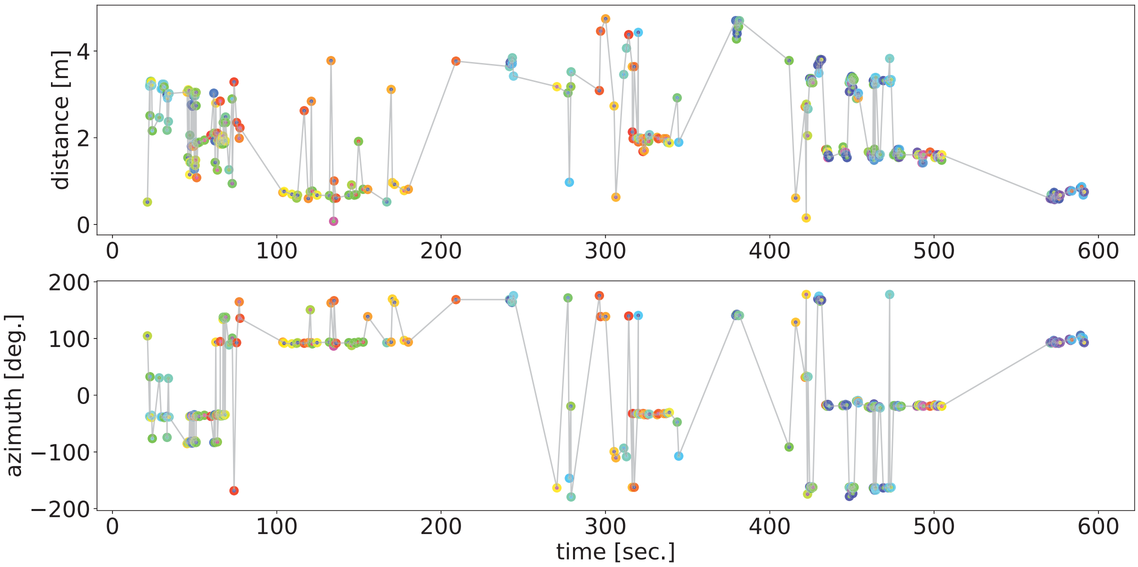

Fig. 7: The spatial, temporal and spectral dynamics of vocalization among ZF individuals. The top figure represents the temporal change of distance of localized sources from the microphone array 4. The bottom figure represents the temporal change of direction of localization results from the microphone array 4. The color of dots corresponds to the color in Fig. 6.

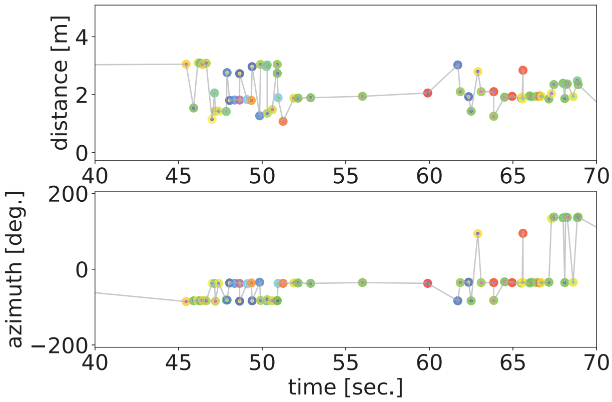

Fig. 8: The detailed results of Fig. 7 during 40–70 sec. Multiple individuals singing alternately at some directions. The color of dot corresponds to the color in Fig. 6.

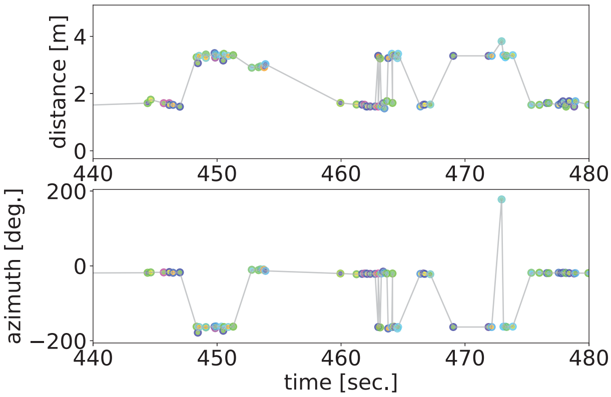

Fig. 9: The detailed results of Fig. 7 during 440–480 sec. Multiple individuals singing alternately at approximately -10 degree and -180 degree. The color of dot corresponds to the color in Fig. 6.

deg. and -80 deg. during 40-70 sec; and around -10 deg. and -180 deg. during 440-480 sec., respectively. It is expected that different individuals vocalized at different positions (not a single individual moving actively) in both cases because there were some localized sound sources at multiple directions in a very short period of time. It should be noted that they tended to interact differently between these cases: they vocalized simultaneously in the former case while alternately in the latter case. We also see that the acoustic properties of vocalizations they used were relatively similar in each time period. Specifically, their vocalization types were mainly indicated by light green-colored dots in Fig. 8 while they were indicated by blue-colored dots in Fig. 9. This suggests that ZF individuals interacted with each other by using specific vocalization types, which is expected to depend on their social situations.

4 CONCLUSION

In order to further consider social and acoustic interactions among a population of songbirds in fields by extending our robot audition approach, we conducted preliminary recordings of vocalizations of ZF individuals in a semi-free flight experimental environment, consists of an out-door mesh tent with some nests, perches and microphone arrays.

We proposed a spatial division-based 2D localization, performed by dividing the whole space in the experimental tent into several areas and choosing a pair of microphone arrays for each area, to reduce the localization error caused by the positions of sound sources. It enabled us to extract vocal positions of ZF accurately as reflecting the positions of arrangements such as nests or perches. We also analyzed the acoustic properties of their vocalizations and the spatio-temporal interactions of them. As a result, we could extract two specific patterns of the interactions that multiple individuals vocalized simultaneously/alternately. It can help us to understand their social interactions themselves and the semantics or functions of the vocalizations.

These results indicate that our approach can contribute to further understanding of roles of vocal learning in communicative interactions of songbirds by considering temporal changes in the properties of their vocalizations although there are still challenges to sophisticate the classification of vocalizations. This might give us insights into the general dynamics of animal communication systems including human languages.

ACKNOWLEDGEMENTS

This work was supported in part by JSPS/MEXT KAKENHI: JP19KK0260, JP18K11467 and JP17H06383 in #4903 (Evolinguistics).

REFERENCES

[1] C. K. Catchpole and P. J. B. Slater, *Bird Song: Biological Themes and Variations*. Cambridge University Press, 2008.

[2] A. Farina and S. H. Gage, *Ecoacoustics: The Ecological Role of Sounds*. John Wiley and Sons, 2017.

[3] B. Krause, “Bioacoustics: Habitat ambience and ecological balance,” *Whole Earth Review*, vol. No 57, 12 1987.

[4] R. Suzuki, C. E. Taylor, and M. L. Cody, “Soundscape partitioning to increase communication efficiency in bird communities,” *Artificial Life and Robotics*, vol. 17, no. 1, pp. 30–34, 2012.

[5] R. Suzuki and M. L. Cody, “Complex systems approaches to temporal soundspace partitioning in bird communities as a self-organizing phenomenon based on behavioral plasticity,” *Artificial Life and Robotics*, vol. 24, no. 4, pp. 439–444, 2019.

[6] R. Suzuki, S. Matsubayashi, K. Nakadai, and H. G. Okuno, “HARKBird: Exploring acoustic interactions in bird communities using a microphone array,” *Journal of Robotics and Mechatronics*, vol. 27, pp. 213–223, 2017.

[7] S. Sumitani, R. Suzuki, S. Matsubayashi, K. Arita, T. Nakadai, and H. G. Okuno, “An integrated framework for field recording, localization, classification and annotation of birdsongs using robot audition techniques - harkbird 2.0,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, in press.

[8] R. Suzuki, S. Sumitani, N. , S. Matsubayashi, T. Arita, K. Nakadai, and H. G. Okuno, “Field observations of ecoacoustic dynamics of a japanese bush warbler using an open-source software for robot audition hark,” *Journal of Ecoacoustics*, vol. 2, p. EYAJ46, 2018.

[9] R. Suzuki, S. Matsubayashi, F. Saito, T. Murate, T. Masuda, Y. Yamamoto, R. Kojima, K. Nakadai, and H. G. Okuno, “A spatiotemporal analysis of acoustic interactions between great reed warblers (acrocephalus arundinaceus) using microphone arrays and robot audition software hark,” *Ecology and Evolution*, vol. 8, pp. 812–825, 2018.

[10] D. R. Farine, L. M. Aplin, B. C. Sheldon, and W. Hoppitt, “Interspecific social networks promote information transmission in wild songbirds,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 282, no. 1803, pp. 2014–2804, 2015.

[11] L. F. Gill, W. Goymann, A. Ter Maat, and M. Gahr, “Patterns of call communication between group-housed zebra finches change during the breeding cycle,” *eLife*, vol. 4, p. e07770, 2015.

[12] D. Stowell, L. Gill, and D. Clayton, “Detailed temporal structure of communication networks in groups of songbirds,” *Journal of The Royal Society Interface*, vol. 13, no. 119, p. 20160296, 2016.

[13] D. Todt and M. Naguib, “Vocal interactions in birds: The use of song as a model in communication,” *Advances in the Study of Behavior*, vol. 29, pp. 247–296, 2000.

[14] R. Schmidt, “Bayesian nonparametrics for microphone array processing,” *IEEE Transactions on Antennas and Propagation (TAP)*, vol. 34, no. 3, pp. 276–280, 1986.

[15] H. Nakajima, K. Nakadai, Y. Hasegawa, and H. Tsujino, “Blind source separation with parameter-free adaptive step-size method for robot audition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, pp. 1476–1485, 2010.

[16] K. Nakadai, H. G. Okuno, and T. Mizumoto, “Development, Deployment and Applications of Robot Audition Open Source Software HARK,” *Journal of Robotics and Mechatronics*, vol. 27, pp. 16–25, 2017.

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

[18] L. van der Maaten and G. Hinton, “Viualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.