AN INTEGRATED FRAMEWORK FOR FIELD RECORDING, LOCALIZATION, CLASSIFICATION AND ANNOTATION OF BIRDSONGS USING ROBOT AUDITION TECHNIQUES — HARKBIRD 2.0

S. Sumitani[(1)]*, R. Suzuki*[(1)]*, N. Chiba*[(1)]*, S. Matsubayashi*[(2)]*, T. Arita*[(1)]*, K. Nakadai*[(3),(4)] *and H. G. Okuno*[(5)]

(1) Nagoya University, Japan, (2) Osaka University, Japan
(3) Tokyo Institute of Technology, Japan, (4) Honda Research Institute Japan Co., Ltd., Japan
(5) Waseda University, Japan

ABSTRACT

Bird vocalizations are one of the important subjects in ecoacoustics because birds communicate diversely using various vocalizations such as songs and calls. We have developed a portable system, HARKBird to provide a basic function, i.e., birdsong localization, which automatically extracts sound sources and their direction of arrivals (DOA) using robot audition techniques based on HARK. In this paper, we introduce HARKBird 2.0 which is empowered for higher understanding of birdsongs. A new soundscape annotation tool for localization results is enhanced by an interactive interface for song classification based on an unsupervised feature mapping t-SNE. We show that HARKBird 2.0 provides bird researchers with an integrated framework to analyze spatio-spectro-temporal dynamics of birdsongs using the song analysis of Japanese bush warbler (*Horornis diphone*).

Index Terms— Ecoacoustics, birdsong, robot audition, sound classification, HARK

1. INTRODUCTION

Ecoacoustics is an interdisciplinary science that investigates natural and anthropogenic sounds and their relationships with the environment over multiple scales of time and space [1]. In particular, understanding of vocalizations of birds is one of the important subjects in this field because birds communicate diversely using various vocalizations such as a song, which is a long species-specific vocalization mainly by males of songbirds to advertise their territory or attract females in a breeding season, and a call, which is a shorter and simpler vocalization which tends to be produced by both sexes in particular contexts [2]. For a deeper understanding of ecological functions and semantics [3] of these vocalizations, it is important to clarify the fine-scaled and detailed relationships among their characteristics (e.g., spectral properties, song or call types) and behavioral contexts (e.g., direction, location, neighboring relationships). However, it takes a lot of time and effort to obtain such data using conventional recordings or by human observation. There have been some approaches for birdsong monitoring using sound source localization techniques (e.g., [4], [5]), but they are not easily available because of the limitation of both hardware and software while applications of new non-invasive recording devices are of focus in ecoacoustics [6].

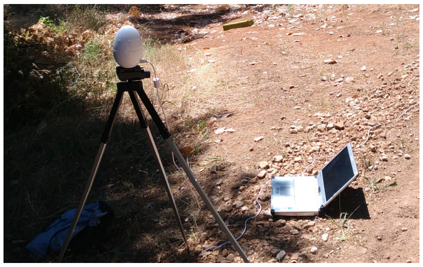
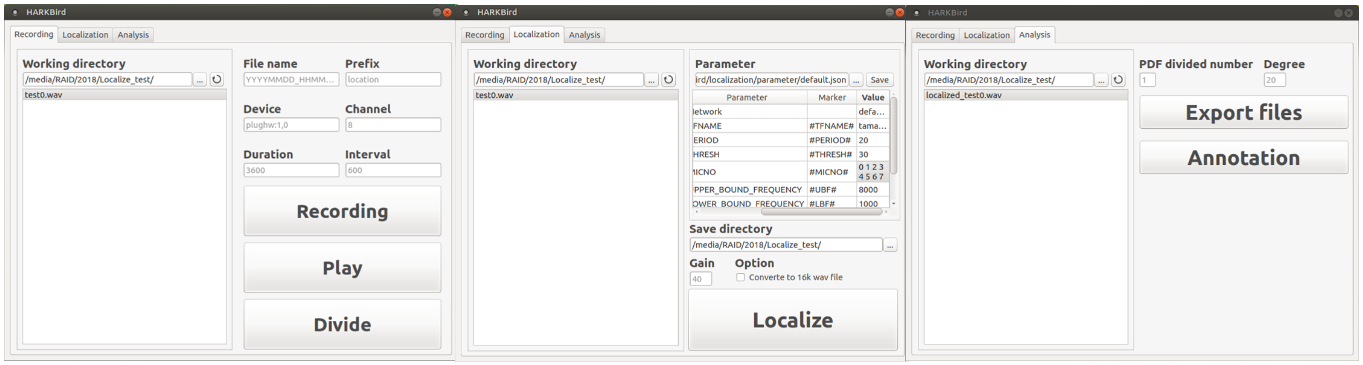


Fig. 1. The HARKBird system in the field.

We have been developing an easily available and portable system for bird song localization called HARKBird [7]. HARKBird consists of a laptop PC with an open source software for robot audition HARK (Honda Research Institute Japan Audition for Robots with Kyoto University)[8, 9] combined with a low-cost and commercially available microphone array (Fig. 1). This system can automatically extract sound sources and the Direction Of Arrival (DOA) of each localized sound using HARK. The initial version of HARKBird was developed to provide bird researchers with a minimal interface to record, localize and separate sound sources observed in their field. Using this system, we showed the existence of temporal overlap avoidance in singing behaviors of some forest species in Japan [7]. That system has been extended to spatially localize song posts of great reed warblers with multiple and self-developed microphone arrays [10], and analyses of effects of conspecific song playbacks on the directional changes and song-types of an individual of Japanese bush warbler [11].

From these experimental trials, it turned out that several important functions are necessary for practical ecological researches using HARKBird. In particular, what we call



Working directory: /media/RAID/2018/Localize_test/
test0.wav
File name: YYYYMMDD_HHMM...
Prefix: location
Device: plughw:1,0
Channel: 8
Duration: 3600
Interval: 600

(a) Recording tab

Working directory: /media/RAID/2018/Localize_test/
test0.wav
Parameter: ird/localization/parameter/default.json

Parameter	Marker	Value
letwork		defa...
FNAME	#TFNAME#	tama...
ERIOD	#PERIOD#	20
HRESH	#THRESH#	30
ICNO	#MICNO#	0 1 2 3 4 5 6 7
PPER_BOUND_FREQUENCY	#UBF#	8000
OWER_BOUND_FREQUENCY	#LBF#	1000

Save directory: /media/RAID/2018/Localize_test/
Gain: 40
Option: □ Converte to 16k wav file

(b) Localization tab

Working directory: /media/RAID/2018/Localize_test/
localized_test0.wav
PDF divided number: 1
Degree: 20

(C) Analysis tab

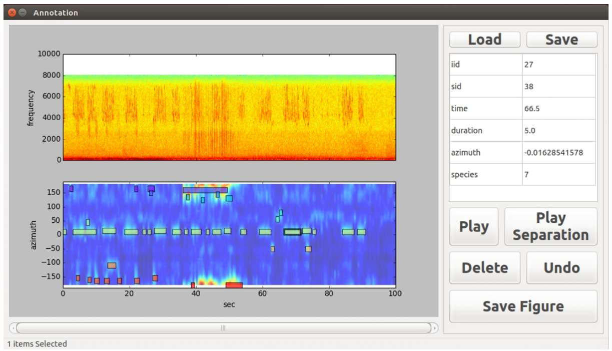
Fig. 2. The GUI of HARKBird.

"a sound source of interest/non-interest problem" is critical. Recordings in fields, especially in forests, include not only vocalizations of target species to be analyzed but also many other sound sources such as vocalizations of other species, wind and water sounds, etc. However, there is no way to discriminate between them automatically because what sounds should be picked up or eliminated depends on the specific purpose of each researcher.

To solve this problem, we provide a new sophisticated annotation tool for editing localization results, which is essential for practical use. However, it still costs a lot of manual effort to remove unnecessary sound sources one by one, by listening to separated sounds and looking at the spectrograms. Thus, we further integrate an interface for sound source classification on the feature space of spectrogram into the annotation tool. The basic idea is that a user can overview what kind of sounds are in the recording to understand the components of the soundscape by looking at the distribution of sound sources on their 2D feature space first, and then the user can classify them by simply grouping clusters of similar sounds on the space. We adopt such an interactive procedure because boundaries of clusters between necessary and unnecessary sources are not always clear in our cases, and thus the user should specify them.

We focus on unsupervised feature learning, which is recently recognized that it can improve automatic large-scale classification of bird sounds (e.g., [12]). We use a dimension reduction algorithm called t-SNE (t-distributed Stochastic Neighbor Embedding) [13] which has been used for arranging sound sources on a two-dimensional space by reducing the dimension of spectrograms (2D images) directly [14, 15, 16]. For example, Tan and McDonald created a two-dimensional visualization map of thousands of bird vocalizations using t-SNE [14]. In this map, similar sounds were placed closer. We also successfully classified several vocalization types of Spotted Towhee (*Piplio maculatus*) using t-SNE combined with a clustering algorithm DBSCAN with manual grouping of clusters [16].

In this paper, we introduce HARKBird 2.0 which is empowered for higher understanding of birdsongs, as described above. This provides bird researchers with an integrated framework to analyze dynamics of birdsongs by enabling them to overview, edit, classify and annotate sound sources on a "directional-spectro-temporal" space. We further introduce the song analysis of Japanese bush warbler (*Horornis diphone*) based on the proposed framework.



iid	27
sid	38
time	66.5
duration	5.0
azimuth	-0.01628541578
species	7

1 items Selected

Fig. 3. A snapshot of the annotation tool.

2. UPDATED FEATURES IN HARKBIRD 2.0

HARKBird is a collection of python scripts that enable us to record using microphone arrays connected to the laptop PC and analyze the recording using HARK. This runs in Ubuntu Linux on which HARK, HARK-Python, PySide, etc. are installed. The detailed description of HARKBird and the scripts are available from our website[1]. We briefly introduce newly added or refined functions of HARKBird 2.0.

2.1. Basic functions

The latest version of HARKBird can be operated with a new GUI composed of 3 tabs for basic function: "Recording", "Localization" and "Analysis" as shown in Fig. 2. In each

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

tab, a user can select a file or folder to be operated in the left side of the window, and can change the operation setting or execute each function in the right side.

In the recording tab, the user can start recording, with 16 kHz and 16 bit format, using microphone arrays and replay or divide the recordings. We assume that the user uses a 8-ch USB microphone array TAMAGO (System in Frontier Inc.). The current system supports simultaneous recording with multiple microphone arrays connected to the laptop by specifying multiple device names. This function has become necessary because we have started to conduct simultaneous recordings for 2D localization (e.g., [10], [16]). Furthermore, we can divide a recording file into several recording files by setting the number of divisions of the file or the recording time per file. In field observations, we often conduct long-term recordings and it takes much time to localize the whole long recorded files. Therefore, the file division is useful to analyze them in the short term. They can also be used for short term analysis to find the appropriate parameter settings for localization before analyzing long-term recordings.

The localization tab enable us to conduct sound source localization based on the MUltiple SIgnal Classification (MUSIC) method [17] using multiple spectrograms with the Short Time Fourier Transformation (STFT) and extract separated sounds as a wave file for each localized sound using GHDSS (Geometric High order Decorrelation based Source Separation) method [18]. We can set values of the essential parameters related to localization and separation of bird songs using HARK in a list in the right side. It is possible to set more parameters than the previous version and additional parameters can be added to the list by defining the parameter name and the corresponding tag in the network file of HARK. This helps us to localize the sound sources more appropriately depending on the acoustic properties of environments and target sounds. The overall results of localization and sound separation are outputted in a folder.

In the analysis tab, we can visualize and analyze the directional and temporal distribution of sound sources. With "Export files" button, the system outputs the spectrogram and the result of localization in PDF format with a specified number of total pages, which is useful to overview the results with an appropriate time scale. The data of all sound sources including their directions and their duration, etc. is output as a JSON file, which can be loaded in an annotation tool.

2.2. Soundscape annotation and song classification

We newly created a sophisticated tool for editing and annotating localization results, by refining the usability and functions of the one in the initial version of HARKBird, which can be used for practical data analyses of soundscape. The annotation window (Fig. 3) pops up when "Annotation" button is clicked. The top panel shows the STFT spectrogram of the recording. The time scale (x-axis) and the focal time period can be adjusted. The bottom panel shows the corresponding MUSIC spectrum and sound sources on the space of time and DOA. Each box represents the start (left edge) and end (right edge) time, and the DOA (y-coordinate) at the start time of the corresponding source. The color of each box represents its class. By clicking the box of each localized sound, the information of sound localization is displayed in the right side. The information can be edited manually, and the corresponding file of the separated sound or the duration in the original recording can be replayed. The tool supports undo process for these editing operations. The position of each source can also be modified by dragging the corresponding box. The user can save the modified data as a JSON file.

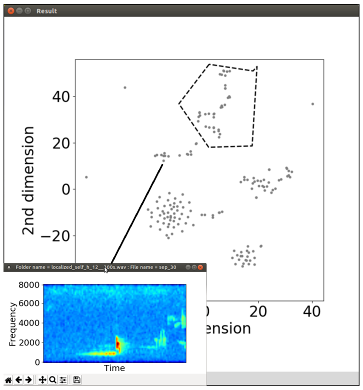


Fig. 4. A snapshot of the interactive classification tool.

We further developed a song classification tool by making use of the acoustic property of separated sounds to make the whole annotation process easier. This tool is composed of the python scripts for dimension reduction and interactive classification, which should be placed and executed in the directory of the localization results of the recording.

The user can conduct dimension reduction of localized sources by using the STFT spectrogram images (100 x 64 pixels) of all separated sounds as a data set. We adopted the scikit-learn library for conducting t-SNE to reduce the dimension of these data and plot them on a 2-dimensional plane to visualize their distribution. We also implemented a grid search for parameter settings because the distribution of sound sources on the two-dimensional feature space by t-SNE largely depends on the parameter setting and need to pick out an appropriate result in order to classify the localized sound easily. The perplexity, learning rate (epsilon), number of iterations, etc. are used as parameters for a grid search.

After picking out an appropriate result of the dimension reduction, we can use an interface for visualization and annotation of sound sources on the feature space. Fig. 4 shows an interface for the classification tool, and each node shows separated sound source. By clicking each node, the spectrogram is shown in another window and the separated sound is replayed. A group of nodes can be classified into a class by surrounding them with a frame (dotted polygon), and speci-

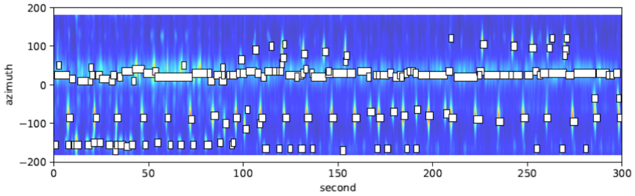


Fig. 5. The non-classified sound sources on time and DOA space.

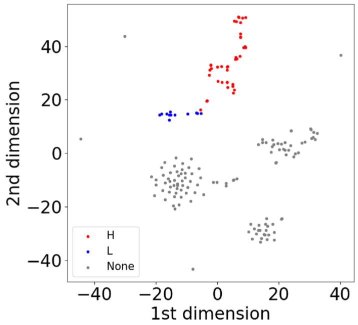


Fig. 6. The result of classification on the 2D feature space.

fying the name of their class. This grouping can be done with a simple keyboard operation and mouse operation, allowing the user to classify similar sounds at one time. By closing the window, the classified data is saved in eps format and also JSON file for the annotation tool of HARKBird is exported.

3. AN EXAMPLE ANALYSIS

We introduce a typical example of data creation using the tools described above. We use the recording tool of HARKBird during a playback experiment on Japanese bush warbler (*Horornis diphone*). This was recorded while we were replaying the type-H song of the same species from a loudspeaker in the territory of one individual, expecting aggressive responses of the focal individual against replayed songs because they would recognize replayed songs as vocalizations of an intruder (see (Suzuki et al., 2018) for detail explanations). As expected, the focal individual responded actively, singing two types of songs around the microphone array and the loudspeaker.

We extracted the initial 300 seconds in the recording and localized sound sources using HARKBird. Note that we adopted the slightly lower value of the threshold parameter for source tracking (i.e., the minimal power of the MUSIC spectrum with which a sound source is localized), in order to localize all songs of the focal individual and replayed songs while other sources were localized. Fig. 5 shows the result of the sound localization output. There are many sounds in various directions, at around 0 degree in particular. It is difficult to distinguish which sound are songs of the target individual.

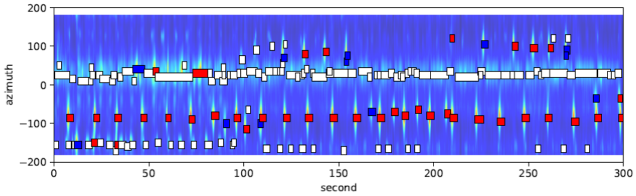


Fig. 7. The classified sound sources on the space of time and DOA.

Fig. 6 shows the distribution of these sources on the feature space of spectrogram. We see several clusters of sources, showing that each has its own spectral properties. We could classify some sounds into two classes of sounds: type-H (red) and type-L (blue) songs by replaying songs around the boundaries of clusters. While both song types have quite similar characteristics, our tool could separate them. We also see some clusters, each roughly reflect spectral properties of sounds (e.g., songs of Blue and white Flycatcher (middle right), noise appears to be induced by water flow (bottom right), faint or mixed sounds (middle center)), which enable us to grasp overall components of the soundscape. Fig. 7 shows the result of classification on the space of time and DOA. Clearly, the type-H songs at around the 100 degree are the sounds replayed from the loudspeaker and the other type-H songs and all type-L songs are those of the target individual. Thus, we can extract the songs and behaviors of the target individual and replayed songs and eliminate unnecessary sound sources, at one time.

4. CONCLUSION

In order to observe fine scaled spatio-spectro-temporal dynamics of bird vocalizations, we need to combine various techniques such as sound localization, separation, and classification. HARKBird 2.0 provides an integrated framework for such ecological data analysis based on HARK. We refined the basic functions of HARKBird and implemented the annotation tool by integrating the feature mapping techniques. It contributes to the extraction of necessary sound sources and reduction of time cost for classification and it helps us to know the soundscape around the microphone array. As far as we know, there have been no other tools with a series of functions from recording to classification, which are implemented in our framework, although some are partially developed [5]. We believe that this further helps us to better understand roles of songs and behaviors of birds in more detail.

5. ACKNOWLEDGEMENTS

This work was supported in part by JSPS/MEXT KAKENHI: JP16K00294, JP17H06841, JP18K11467 and JP17H06383 in #4903.

6. REFERENCES

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

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

[3] K. Daimon, R. W. Hedley, and C. E. Taylor, "Semantic inference of bird songs using dynamic bayesian network," in *the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017, pp. 4911–4912.

[4] D.J. Mennill, M. Battiston, and D.R. Wilson, "Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. methods in ecology and evolution," *Methods in Ecology and Evolution*, vol. 3, pp. 704–712, 2012.

[5] C.E. Taylor, Y. Huang, and K. Yao, "Distributed sensor swarms for monitoring bird behavior: an integrated system using wildlife acoustics recorders," *Artificial Life Robotics*, vol. 21, pp. 268–273, 2016.

[6] A Farina, "Perspectives in ecoacoustics: A contribution to defining a discipline. journal of ecoacoustics," *Journal of Ecoacoustics*, vol. 2, pp. TRZD5I, 2018.

[7] R. Suzuki, S. Matsubayashi, R. W. Hedley, 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.

[8] 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.

[9] H.G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *2015 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 2015, pp. 5610–5614.

[10] S. Matsubayashi, R. Suzuki, F. Saito, T. Murate, T. Masuda, K. Yamamoto, R. Kojima, K. Nakadai, and H.G. Okuno, "Acoustic monitoring of the great reed warbler using multiple microphone arrays and robot audition," *Journal of Robotics and Mechatronics*, vol. 27, pp. 224–235, 2017.

[11] R. Suzuki, S. Sumitani, 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, pp. EYAJ46, 2018.

[12] D. Stowell and M.D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[13] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Advanced Robotics*, vol. 24, pp. 739–761, 2010.

[14] M. Tan and K. McDonald, "Bird sounds: 2d visualization of thousands of bird vocalizations," 2017, `https://experiments.withgoogle.com/bird-sounds`.

[15] E. Benjamin and J. Altosaar, "MusicMapper: Interactive 2d representations of music samples for in-browser remixing and exploration," in *the International Conference on New Interfaces for Musical Expression*, 2015, pp. 325–326.

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

[17] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, 1986.

[18] 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.