INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Localizing Bird Songs Using an Open Source Robot Audition System with a Microphone Array

Reiji Suzuki[1], *Shiho Matsubayashi*[1], *Kazuhiro Nakadai*[2] *and Hiroshi G. Okuno*[3]

[1]Graduate School of Information Science, Nagoya University, Japan
[2]Honda Research Institute Japan Co., Ltd., Japan
[3]Graduate Program for Embodiment Informatics, Waseda University, Japan
reiji@nagoya-u.jp, mt.shiho@gmail.com, nakadai@jp.honda-ri.com, okuno@aoni.waseda.jp

Abstract

Auditory scene analysis is critical in observing bio-diversity and understanding social behavior of animals in natural habitats because many animals and birds sing or call and environmental sounds are made. To understand acoustic interactions among songbirds, we need to collect spatiotemporal data for a long period of time during which multiple individuals and species are singing simultaneously. We are developing HARKBird, which is an easily-available and portable system to record, localize, and analyze bird songs. It is composed of a laptop PC with an open source robot audition system HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) and a commercially available low-cost microphone array. HARKBird helps us annotate bird songs and grasp the soundscape around the microphone array by providing the direction of arrival (DOA) of each localized source and its separated sound automatically. In this paper, we briefly introduce our system and show an example analysis of a track recorded at the experimental forest of Nagoya University, in central Japan. We demonstrate that HARKBird can extract birdsongs successfully by combining multiple localization results with appropriate parameter settings that took account of ecological properties of environment around a microphone array and species-specific properties of bird songs.

1. Introduction

Auditory scene analysis is critical in observing bio-diversity and understanding social behavior of animals in natural habitats because many animals and birds sing or call and environmental sounds are made. Sound information, however, has not been utilized so much compared to visual information in environmental monitoring and wildlife management. In ornithology or observing birds, bird songs give a critical clue of monitoring.

In forests, many male birds produce long vocalizations called songs to advertise their territory or attract females in a breeding season [1]. There have been empirical studies on the temporal partitioning or overlap avoidance of singing behaviors of songbirds with various time scales [2, 3, 4, 5, 6]. We are interested in clarifying its underlying dynamics as an example of complex systems based on adaptive behavioral plasticity from both theoretical [7] and empirical standpoints [8, 9].

We need to collect spatiotemporal data for a long period of time during which multiple individuals and species are singing simultaneously to understand such complex interaction processes. It has been recognized that acoustic monitoring of animals using a microphone array is a promising approach [10]. There are various ways to understand behaviors of birds using microphone arrays. For example, they have been used to track the movement of individuals both in 2D [11] and 3D [12] spaces. However, monitoring birds using microphone arrays are still not well-adapted to field researchers because of the limited availability of both software and hardware.

To solve this problem, we are developing an easily-available and portable system called HARKBird. HARKBird consists of a standard laptop PC with an open source robot audition system HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) [13] and a commercially available low-cost microphone array. It helps us with annotating recordings and grasping the soundscape around the microphone array by extracting the direction of arrival (DOA) of sound sources and its separated sound automatically. HARKBird helps us annotate bird songs and grasp the soundscape around the microphone array by providing the DOA of each localized source and its separated sound automatically. A significant benefit of using HARK is that it has constantly been updated since its original release in 2010 where we can find the latest algorithms for sound source localization, separation, and even recognition.

Mennill et al. constructed a system composed of an array of multiple commercial stereo recorders (Songmeter SM2 with GPS; Wildlife Acoustics Inc.) [15]. Recorded sounds are synchronized to generate 8-channel data and bird or animal call is extracted manually. Then its 2D location is estimated based on a cross-correlation method [14] in MatLab. They showed a high-level accuracy to localize variety of sounds, including bird songs replayed by a loud speaker, under ideal conditions in which a single target sound source was played in a relatively quiet environment. On the other hand, our system aims at grasping more realistic representation of the soundscape in which multiple individuals or species sing at the same time in noisy environments. A notable feature that distinguishes HARKBird from other works with the similar motivation is the simplicity of the system. It allows us to conduct recordings and necessary analyses in such complex conditions with a single system even in real-time while others based on standard recorders cannot.

In this paper, we briefly introduce our system, and show an example analysis of a track recorded at the experimental forest of Nagoya University in central Japan. Despite challenging ecological and acoustic properties of the environment, we successfully extracted songs of different bird species singing simultaneously by adjusting parameters of HARKBird for properties of both target species' songs (e.g. frequency) and the surrounding environment (e.g., vegetation, water flow, winds and obstacles).

http://dx.doi.org/10.21437/Interspeech.2016-782

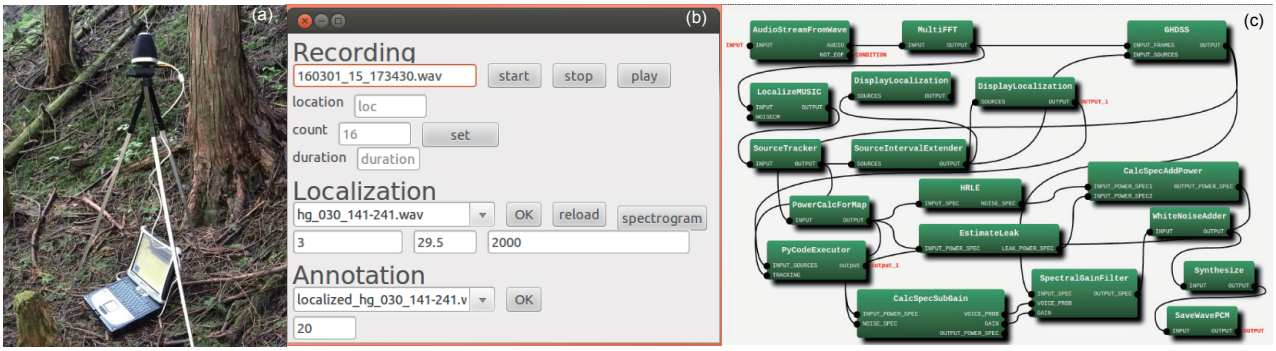

Figure 1: *An overview of HARKBird. (a) A snapshot of the system. (b) The GUI interface. (c) The network of HARK, which conducts sound source localization of a recording with a MUSIC (Multiple Signal Classification) method using multiple spectrograms with FFT, and then separates localized sounds with a GHDSS (Geometric High order Decorrelation based Source Separation) method in real-time.*

2. HARKBird

Fig. 1 (a) shows a snapshot of the system. We used TOUGHBOOK CF-C2 (Panasonic) and the Microcone (Dev-Audio)[1], a 7-channel microphone array, placed on a tripod. The whole software system is composed of HARK and a set of scripts of python with major modules (e.g., wxPython, PySide) and some standard softwares for sound processing (e.g., sox, arecord, aplay). The information for installation is available from our website [2]. The GUI interface (Fig. 1 (b)) allows us to start / stop recording; localize[3] and separate sound sources; and export the results for annotation, using the network of HARK (Fig. 1 (c), see the caption and the documentation of HARK[4] for detail.).

HARK's main sound source localization algorithm is based on MUltiple SIgnal Classification (MUSIC) [16]. Since MUSIC has sharper peaks for sound source directions than conventional beamformers such as a delay-and-sum beamformer, it is a noise-robust algorithm. HARK's MUSIC accepts several parameters to control the behavior of MUSIC and source tracking. Three main parameters are important for localizing bird songs successfully. Firstly, *the expected number of sound sources for MUSIC* (NS) determines the basic number of localized sounds throughout the track.

Secondly, *the lower bound frequency for MUSIC* (LB) is important because it can significantly reduce localization of noises. In recording in forests, noises are usually caused by leaves, waters and winds and their main frequencies are lower than bird songs targeted in our study. Thus, we can reduce unnecessary localization of such noises by setting a higher value to LB. Needless to say, there is a trade-off between localization of lower-frequency songs and surrounding noises.

Thirdly, *the threshold for source tracking* (TS) is important for ignoring a localized sound source of which power is less than TS. There is also a song/noise trade-off to set this parameter. See the documentation of HARK for detail. We particularly focus on LB and TS to obtain better localization results by taking account of the surrounding environment around the microphone array as discussed later.

These parameter tuning is not easy, because ground truth is not available for songscape; for example, which bird of which species sing when, where and how long? In our preliminary experiments with recording at a park in Japan in 2013, we compared two results obtained by Bayesian non-parametrics for microphone array processing (BNP-MAP) [17] and HARK [18] (unpublished results). Since BNP-MAP assumes an infinite number of sound sources, it usually separated more sound sources than HARK. By scrutiny of those results by ornithologists, we concluded that the result obtained by BNP-MAP can be treated as ground truth. By tuning HARK parameters, HARKBird attained almost similar performance as BNP-MAP. In this paper, the default setting of parameters of HARKBird has been determined based on various experiences including the above case.

HARKBird lastly generates a PDF file that shows the spectrogram of a channel of the original recording; the MUSIC spectrum; and the directional and temporal pattern of sound localization in which each localized sound is represented as a line in the space of time and direction of arrival (DOA). Figs. 2 (a) and (b) are examples. This PDF file is useful to overview the long-period pattern of the acoustic environment.

Additionally, while we do not discuss details in this paper, HARKBird has other following features: a sound separation of localized sources, an interactive interface that shows both the spectrogram and the localization result in which each separated sound can be replayed, an exportation of some files for annotation in a JSON format, and a simple and minimal annotation tool for editing and classifying the localization results.

3. An example analysis of bird songs considering the parameters of localization

In this section, we introduce an example analysis of bird songs using HARKBird. More specifically, we discuss how to set the parameters to extract songs of specific species or individuals in a track successfully, taking account of the acoustic environment around the microphone array.

We recorded bird songs at the Inabu field, the experimental forest of Field Science Center, Graduate School of Bioagricultural Sciences, Nagoya University, in central Japan (May 2015). The forest is mainly composed of conifer plantation (Japanese cedar, Japanese cypress, and red pine), with small patches of

[1]Microcone is discontinued, but TAMAGO, a low-price 8-channel USB microphone array, is available from System In Frontier (http://www.sifi.co.jp/en/).

[2]http://www.alife.cs.is.nagoya-u.ac.jp/~reiji/HARKBird/

[3]In this paper, we use the term "localize" as an estimate of the direction of arrival in 2D without a distance information.

[4]http://www.hark.jp/document/hark-document-en/

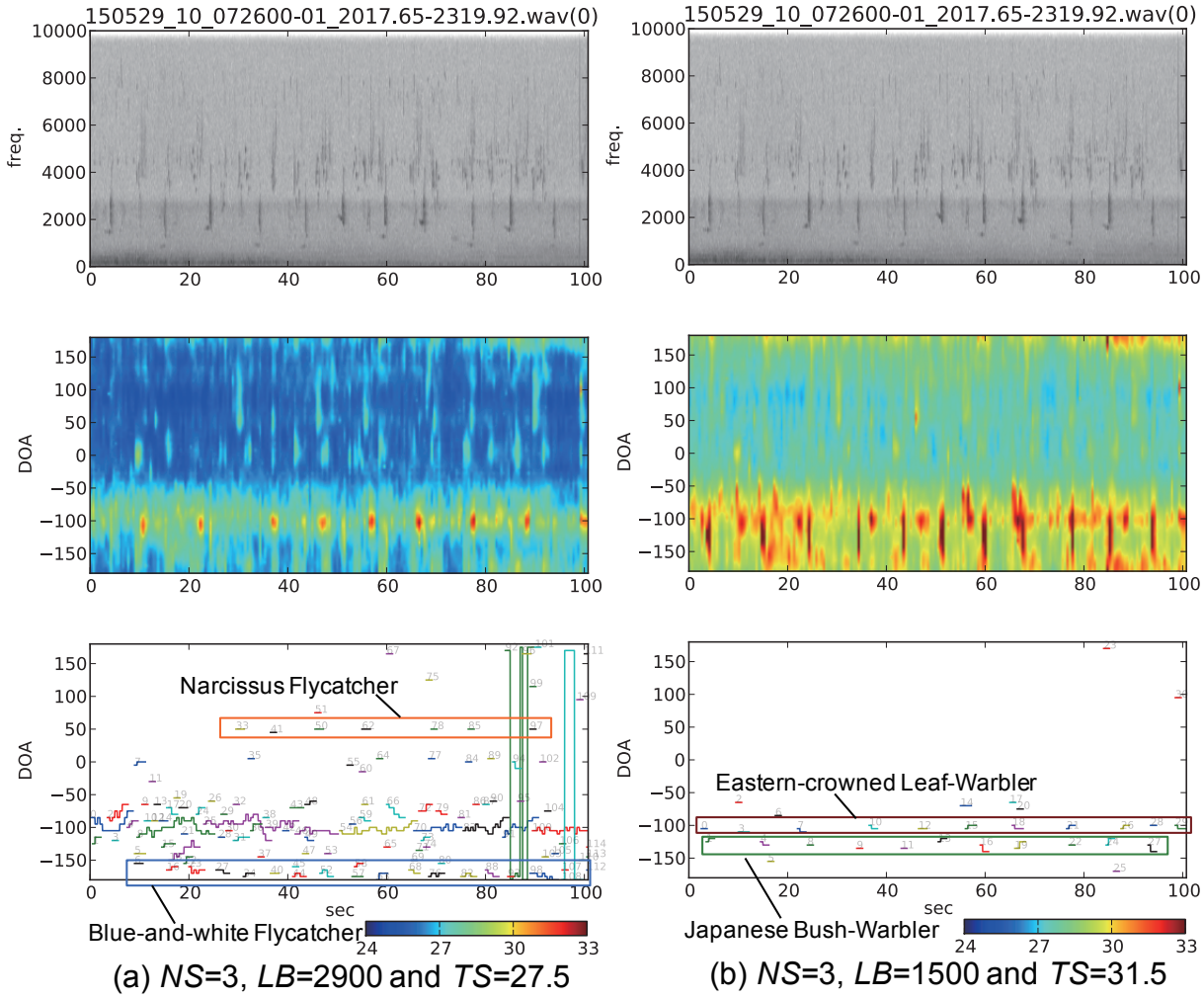

Figure 2: *Localization results of the first 100 seconds in an approximately 5 minutes recording at the Inabu field, the experimental forest of Field Science Center, Graduate School of Bioagricultural Sciences, Nagoya University, in central Japan (May 2015). We used the following parameter settings in Section 1: (a) NS=3, LB=2900 and TS=27.5; (b) NS=3, LB=1500 and TS=31.5. Top: the spectrogram of a channel of the original recording. Middle: the MUSIC spectrum. Bottom: the directional (DOA) and temporal pattern of sound localization. In (a), the songs of Narcissus Flycatcher (Ficedula narcissina) and Blue-and-white Flycatcher (Cyanoptila cyanomelana) were localized successfully. In (b), the songs of East-crowned Leaf Warbler (Phylloscopus coronatus) and Japanese Bush-Warbler (Horornis diphone) were localized successfully. Note that the classification of species was conducted manually.*

broadleaf trees (Quercus, Acer, Carpinus, etc.). In this forest, common bird species are known to vocalize actively during this season. We selected approximately five minute recording duration in which four species were singing actively. Those four species included Narcissus Flycatcher (*Ficedula narcissina*, NAFL), Blue-and-white Flycatcher (*Cyanoptila cyanomelana*, BAWF), East-crowned Leaf Warbler (*Phylloscopus coronatus*, ECLW) and Japanese Bush-Warbler (*Horornis diphone*, JBWA). We believe that a unique individual vocalized its species-specific songs repeatedly for each species.

We conducted preliminary localization analysis and found that the two different settings of the parameters worked well to extract songs of the different species in this track: (a) The higher LB (=2900) and the lower TS (=27.5) to localize songs of NAFL and BAWF; and (b) The lower LB (=1500) and the higher TS (= 31.5) to localize songs of ECLW and JBWA. Figs. 2 (a) and (b) depict the localization results of the first 100 seconds in this recording with the parameter settings of (a) and (b), respectively.

The top panels in Fig. 2 show the spectrograms of one out of seven channels of the original recording. They helps us distinguish different types of songs. The middle panels depict the MUSIC spectrum. In both parameter settings, higher power was observed between -150 and -50 degrees of the DOA. The bottom panels also illustrate the DOA of each localized sound. As shown in the bottom panels of Fig. 2, the distribution pattern of localized sounds varied significantly under two different settings. Each rectangle indicates the spaciotemporal pattern of the vocalizations of the corresponding species. The classification of species was conducted manually by examining separated sound sources.

We used the setting (a) to particularly localize songs of NAFL. The songs of this species were a little faint compared to songs of other species recorded in this track. The vague-

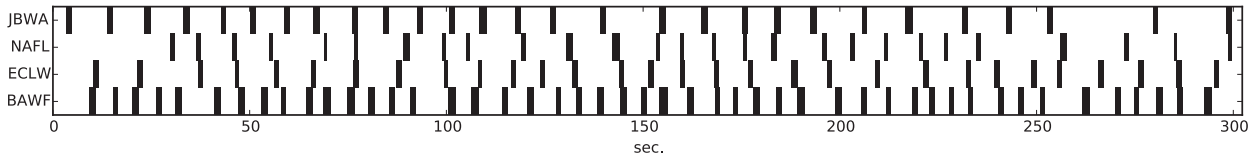

Figure 3: *The annotated track by human with the help of the localization results. Each box represents the timing and duration of a song of the corresponding species. NAFL: Narcissus Flycatcher (Ficedula narcissina), BAWF: Blue-and-white Flycatcher (Cyanoptila cyanomelana), ECLW: East-crowned Leaf Warbler (Phylloscopus coronatus) and JBWA: Japanese Bush-Warbler (Horornis diphone) .*

ness could be attributed to the location of the individuals. For example, this individual could be singing in a bush or in a far distance. To localize those faint songs with the lower power, we used the lower value for TS (27.5). Instead, we limited to localize high-frequency sources by adjusting LB to the higher value (2900) because the frequency of songs of NAFL was relatively higher than other environmental noises in this case. As a result, the songs of NAFL were constantly localized at around 50 degrees as shown in Fig. 2 (a). Also, the songs of BAWF, which have the similar frequency property to those of NAFL, were localized successfully at around -170 degrees. Note that there are sound sources repeatedly localized in the direction of 0 degrees. It turned out that these were not sound sources from real birds but songs of BAWF reflected by a neighboring red pine or the wall of an old prefabricated hut located at around 0 degrees.

The system however failed to localize bird songs and instead localized numerous noisy sources including fractions of songs between -150 and -50 degrees. These noisy sources, both in short and long duration, could be caused by neighboring vegetation such as thick bamboo bush or water flow in that direction. They might have made continuous noises, which was reflected to the higher power of MUSIC spectrum in that direction compared to other directions.

To minimize the influence of such noises, we decided to use the other setting (b) with a much higher TS value (31.5). At the same time, we adopted the lower LB value (1500) to localize the whole songs of JBWA whose song contains an introductory component that consists of a low frequency sound. As a result, the songs of both ECLW and JBWA were constantly localized at around -100 and -130 degrees, respectively, as shown in Fig. 2 (b). The songs of NAFL and BAWF were ignored in this case.

These results clearly show that HARKBird can successfully localize songs of various species in different ecological environments by correctly specifying appropriate settings of parameters for localization.

4. Accuracy of localization

Finally, to evaluate the overall localization accuracy generated by the two different settings discussed above, we further conducted fine-grained annotation of the whole five minute recording by human referring to the localization results. Fig. 3 shows the annotated timing of singing behaviors of each species. The separated songs and their directional information were particularly beneficial to minimize the probability of misclassification or overlooking of bird songs, that often occurs when multiple individuals or species sing simultaneously. We defined the success rate of localization for each species as "the ratio of the number of localized songs by HARKBird to that of actual songs recognized by human or HARKBird. We used the localization results with the setting (a) for NAFL and BAWF; and (b) for ECLW and JBWA for five minutes. As shown in Table 1, more than 88 % of the songs and calls were localized successfully.

Table 1: The accuracy of localization of songs for five minutes. NAFL: Narcissus Flycatcher (*Ficedula narcissina*), BAWF: Blue-and-white Flycatcher (*Cyanoptila cyanomelana*), ECLW: East-crowned Leaf Warbler (*Phylloscopus coronatus*) and JBWA: Japanese Bush-Warbler (*Horornis diphone*).

species	NAFL	BAWF	ECLW	JBWA
parameter setting	(a)	(a)	(b)	(b)
actual song	27	47	30	28
localized song	24	45	27	28
success rate	88.9	95.7	90.0	100.0

5. Conclusions

We introduced HARKBird and discussed how it can localize bird songs by adjusting its parameter settings for both target species' songs and their surrounding environment. Also, by combining multiple localization results with appropriate parameter settings, over 88% of songs were localized as sound sources. This result can be useful to examine potential song overlap avoidance among species. In fact, our preliminary analysis of the annotated data in Fig. 3 shows a statistical significance in overlap avoidance among these species in Fig. 3 (df=3, χ^2=7.10, P=0.03) when compared to a random case based on the duty cycle method [6, 19].

We are currently conducting a 2D location estimation by extending our system to be able to record with multiple microphone arrays. Furthermore, because HARK can localize sounds in real time, we are also extending HARKBird to an interactive system that can respond to acoustic events. We believe that further development of HARKBird contributes to better understand complex acoustic interactions in bird communities.

6. Acknowledgements

The authors thank Mami Toyoshima (Nagoya University) for developing a pilot version of the system; Takashi Kondo and Naoki Takabe (Nagoya University) for supporting field works in Japan; Charles Taylor and Martin Cody (University of California, Los Angeles) for supporting bird song projects. This work was supported in part by JSPS KAKENHI 15K00335, 16K00294 and 24220006.

7. References

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

[2] M. L. Cody and J. H. Brown, "Song asynchrony in neighbouring bird species," *Nature*, vol. 222, pp. 778–780, 1969.

[3] R. Planqué and H. Slabbekoorn, "Spectral overlap in songs and temporal avoidance in a peruvian bird assemblage," *Ethology*, vol. 114, pp. 262–271, 2008.

[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] X. Yang, X. Ma, and H. Slabbekoorn, "Timing vocal behaviour: Experimental evidence for song overlap avoidance in Eurasian Wrens," *Behavioural Processes*, vol. 103, pp. 84–90, 2014.

[6] C. Masco, S. Allesina, D. J. Mennill, and S. Pruett-Jones, "The song overlap null model generator (song): a new tool for distinguishing between random and non-random song overlap," *Bioacoustics*, vol. 25, pp. 29–40, 2016.

[7] R. Suzuki and T. Arita, "Emergence of a dynamic resource partitioning based on the coevolution of phenotypic plasticity in sympatric species," *Journal of Theoretical Biology*, vol. 352, pp. 51–59, 2014.

[8] 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," in *Proceedings of the 20th International Symposium on Artificial Life and Robotics*. ALife Robotics Corporation Ltd., 2015, pp. 11–15.

[9] R. Suzuki, R. Hedley, and M. L. Cody, "Exploring temporal soundspace partitioning in bird communities emerging from inter- and intra-specific variations in behavioral plasticity using a microphone array," in *Abstract Book of the 2015 Joint Meeting of the American Ornithologists' Union and the Cooper Ornithological Society*, 2015, p. 86.

[10] D. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirshel, "Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus," *Journal of Applied Ecology*, vol. 48, pp. 758–767, 2011.

[11] T. C. Collier, A. N. G. Kirschel, and C. E. Taylor, "Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network," *The Journal of the Acoustical Society of America*, vol. 128, pp. 182–189, 2010.

[12] Z. Harlow, T. Collier, V. Burkholder, and C. E. Taylor, "Acoustic 3d localization of a tropical songbird," in *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, 2013.

[13] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "Design and implementation of robot audition system 'HARK' —open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, pp. 739–761, 2010.

[14] D. J. Mennill, J. M. Burt, K. M. Fristrup, and S. L. Vehrencamp, "Accuracy of an acoustic location system for monitoring the position of duetting songbirds in tropical forest," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2832–2839, 2006.

[15] 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*, pp. 704–712, 2012.

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

[17] T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, "Multiple emitter location and signal parameter estimation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 2, pp. 493–504, 2014.

[18] Y. Bando, T. Otsuka, K. Itoyama, K. Yoshii, Y. Sasaki, S. Kagami, and H. G. Okuno, "Challenges in deploying a microphone array to localize and separate sound sources in real auditory scenes," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 2015, pp. 723–727.

[19] R. W. Ficken, M. S. Ficken, and J. P. Hailman, "Temporal pattern shifts to avoid acoustic interference in singing birds," *Science*, vol. 183, no. 4126, pp. 762–763, 1974.