Recognition of In-Field Frog Chorusing Using Bayesian Nonparametric Microphone Array Processing

Yoshiaki Bando†, Takuma Otsuka‡, Ikkyu Aihara*, Hiromitsu Awano†, Katsutoshi Itoyama†, Kazuyoshi Yoshii†, and Hiroshi G. Okuno**
†Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan.
‡ NTT Communication Science Laboratories, Kyoto, 351-0114, Japan.
∗Graduate School of Life and Medical Sciences, Doshisha University, Kyoto, 610-0321, Japan.
∗∗Graduate School of Embodiment Informatics, Waseda University, Tokyo, 169-0072, Japan.
{*yoshiaki, ohtsuka, itoyama, yoshii, okuno*}*@kuis.kyoto-u.ac.jp*,
ikkyu.aihara@gmail.com, awano@easter.kuee.kyoto-u.ac.jp

Abstract

In this paper, we exploit Bayesian nonparametric microphone array processing (BNP-MAP) for analyzing the spatio-temporal patterns of the frog chorus. Such analysis in real environments is made more difficult due to unpredictable sound sources including calls of various species of animals. An application of conventional signal processing algorithms has been difficult because these algorithms usually require the number of sound sources in advance. BNP-MAP is developed to cope with auditory uncertainties such as reverberation or unknown number of sounds by using a unified model based on Bayesian nonparametrics. We exploit BNP-MAP for analyzing the sound data of 20 minutes captured by a 7-channel microphone array in a paddy rice field in Oki Island, Japan, and revealed that two individuals of Schlegel's green tree frog (*Rhacophorus schlegelii*) called alternately with anti-phase. This result is compared with the video data captured by a video camera with 18 units of sound-imaging devices called *Firefly* deployed along the bank of the rice field. The auditory result provides more detailed patterns of the frog chorus in higher temporal resolutions. This higher resolution enables to analyze fine temporal structures of the frog calls. For example, BNP-MAP reveals the trill-like calling pattern of *R. schlegelii*.

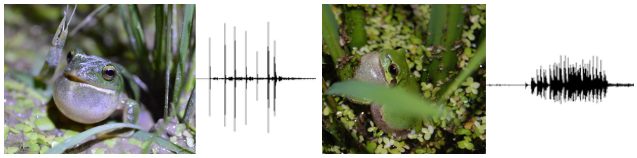
Introduction

In singing the following German folksong "Der Froschgesang" (Froggy Song):

> *Ganze Sommer næchtelang, hören wir den Frosch gesang; quak quak quak quak, kæ kæ kækæ kæ kæ kæ quak quak quak.*

How do singers sing "*quak quak quak quak, kæ kæ kæ kæ kæ kæ kæ quak quak quak.*"? Do they sing the whole phrase together or by two groups? For the latter, one group sings "*quak quak quak quak,* ⟨SILENCE⟩ *quak quak quak.*", while the other one sings "⟨SILENCE⟩ *kæ kæ kæ kæ kæ kæ kæ* ⟨SILENCE⟩ ."

This has been an open problem of the frog chorus for a long time. Aihara et al. modeled frog chorusing as a coupled non-linear oscillator system and obtained the following results by simulation: two-frog chorusing has anti-phase synchronization, and three-frog chorusing has two stable states; 2:1 anti-phase, and triphase synchronization. These simulated results are confirmed by recording choruses of Japanese tree frog (*Hyla japonica*) (see Figure 1(b)) *in a laboratory* and analyzing the recordings by independent component analysis (Aihara et al. 2011). This observation treats only dyadic or triadic interactions among neighbors in a laboratory.

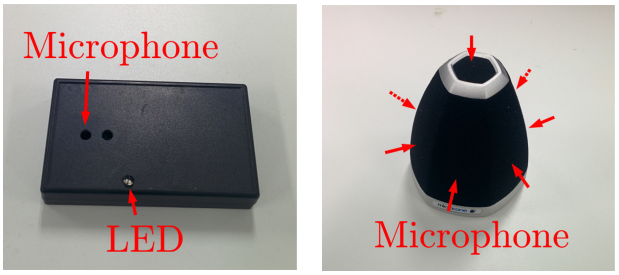


(a) Schlegel's green tree frog (*Rhacophorus schlegelii*) (b) Japanese tree frog (*Hyla japonica*)

Figure 1: Two species of tree frogs and their calls

Recently, Aihara et al. discovered that *H. japonica* call alternately with anti-phase *in paddy rice fields* (Aihara et al. 2014). This observation was obtained by using 40 units of sound-imaging device called *Firefly* shown in Figure 2(a) developed by Mizumoto et al. (Mizumoto et al. 2011), instead of using a microphone array. Since frog chorusing are truly dynamic environments for social communication, in-field recordings contain a lot of sounds. Microphone array processing for such recordings is very difficult due to various uncertainties caused by dynamic environments. Conventional signal processing algorithms assume the number of sound sources in advance, but this assumption does not hold in a real-world environment.

The 60th anniversary essay of *Animal Behaviour* advocates the importance of recent signal processing technologies (Bee, Schwartz, and Summers 2013); Bee et al. pointed out that Firefly and microphone array processing are promising directions to explore the complexity of chorus organiza-



(a) Firefly (b) Microcone

Figure 2: Sound-imaging device Firefly and microphone array called Microcone (Dev-Audio)

tion. Recording and analyzing interactions over large spatial and temporal scales are really challenging. Microphone array processing not only localizes calling males but also separates each calling signals for subsequent acoustic analysis. Microphone array processing with conventional delayed-sum beamformer was used for frog chorusing (Jones, Jones, and Ratnam 2014; Megela Simmons, Simmons, and Bates 2008).

Recent advancement of statistical signal processing has gained a further applicability. A Bayesian nonparametric microphone array processing (BNP-MAP) simultaneously localizes and separates sound sources even when the number of sound sources is unknown in advance (Otsuka et al. 2014; 2012).

The extraction of calling sounds of targeted species from the observed mixture is essential to the analysis in the actual field, while the uncertainties such as an unknown number of sound sources in the observation is problematic for the robust extraction process. We employ BNP-MAP, which is a microphone array processing technique based on a probabilistic model, to deal with the extraction of target sound sources under the uncertainties.

Computational sustainability and Biodiversity

A goal of computational sustainability is to develop computational methods for a sustainable environment, economy and society (Gomes 2009). Some researchers focus on biodiversity and species conservation. eBird project is establishing a crowd sourcing for collecting bird observation. It uses machine learning to detect faults of observations and grade the capability of each human observer and then improve the skills of human as well as the system's capabilities (Kelling et al. 2013; Sullivan et al. 2009).

Cody et al. discovered that birds of the same species sing songs synchronously and stop singing to avoid soundspace overlap when other species are singing (Cody and Brown 1969). Suzuki et al. are developing "HARKBird" to localize and separate each bird song in natural environments and visualize auditory scene (Suzuki, Taylor, and Cody 2014). HARKBird uses a robot audition software called HARK (Nakadai et al. 2010) with a microphone array called Microcone (Dev-Audio) shown in Figure 2(b). Large-scale interactions by bird communities provide not only clues for understanding bird lives but also ideas for designing ICT

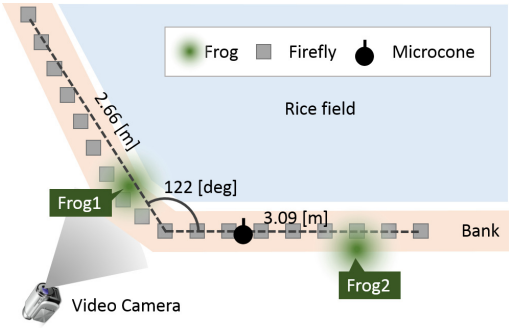


Figure 3: Setting of in-field experiment. 18 units of Fireflies are deployed on the two banks of a paddy rice field. One Microcone microphone array is used. Two frogs, Frog1 and Frog2, of *R. schlegelii* remain at almost same position during the recording.

protocols. This avoidance of soundspace overlap exploits temporal resource partitioning, which inspired a protocol for wireless sensor networks "DESYNC" (Dogesys et al. 2007).

Why do we focus on *Rhacophorus schlegelii*?

In this paper, we study the chorus structures of *R. schlegelii* (see Figure 1(a)). Choruses of *R. schlegelii* can be observed widely in Japan, from Kagoshima to Aomori prefecture (Maeda and Matsui 1999). They usually breed in paddy fields in April and May. Some male frogs of *R. schlegelii* start calling immediately after sunset, and then the chorus size becomes larger. Since the male frogs call in holes under the ground of the banks of paddy fields, it is difficult to precisely localize the positions of callers. In addition, calls of *R. schlegelii* show complicated temporal structures; namely, a single call consists of several pulses vocalized at the intervals of 40 msec (Maeda and Matsui 1999) as shown in Figure 1(a) . To reveal the acoustic interactions of *R. schlegelii* calling under the ground by using such complex calls, sophisticated experimental techniques are required. To our knowledge, this is the first study examining the chorus structures of *R. schlegelii* by combining microphone-array and sound-imaging techniques.

Experiments: Methods and Results

This section presents localization and separation results of the in-field experiment. The result of BNP-MAP is compared with the state-of-the-art sound separation methods: HARK and Firefly.

In-field Recording

We recorded frog choruses of two individuals of *R. schlegelii* in a paddy rice field in Oki Island in Japan this June between 22:20 and 22:40. As illustrated in Figure 3, we placed one Microcone microphone array, 18 units of Fireflies, and one video camera around the field. Microcone captured multichannel sound signals at 16 kHz sampling. The spatio-temporal light pattern of the Fireflies was captured using a

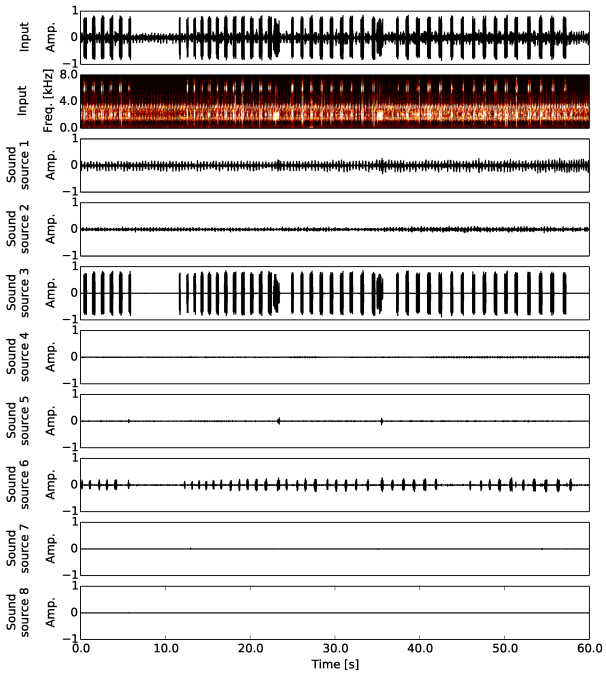


Figure 4: Input and separated sound sources by BNP-MAP. The top 2 figures are the waveform and its spectrogram of the sound captured by one microphone of Microcone. The next eight waveforms correspond to separated sound sources, *SS1*–*SS8*. Frog 1 and 2 correspond to *SS6* and *SS3*, respectively. *SS1*, *SS2*, and *SS5* are calls of three individuals of *H. japonica*. *SS4*, *SS7*, and *SS8* are other noise sources out of the video camera's visual field.

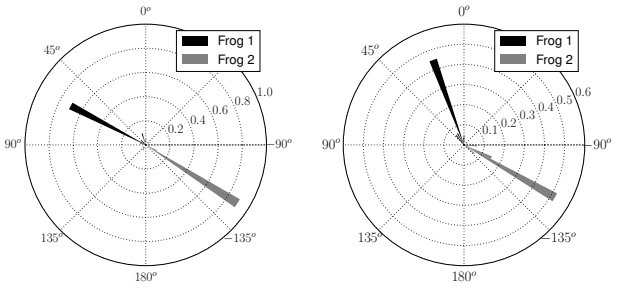
handy video camera, Sony HDR-XR550V, with HD quality at 29.97 fps sampling.

Since male frogs of *R. schlegelii* are calling in a hole under the ground, it is quite difficult to find them at night. Hence, we located one by one by carefully listening to each call for obtaining the ground truth of frog locations illustrated in Figure 3. During the recording, two individuals were confirmed to stay at almost the same position. We confirmed that all the frogs in the target area were covered with Fireflies and that the video camera captured the LED lights of all the Fireflies prior to the field recordings.

Single and AV-integrated Analysis

Sound sources separated by BNP-MAP and HARK are indexed by the direction ranging from -180 degree to 180 degree. The separated signals corresponding to the calls of the two frogs are labeled manually as Frog1 and Frog2 by consulting to the results of Firefly. Calls of *R. schlegelii* and *H. japonica* were distinguished based on their spectral and temporal properties mentioned in (Maeda and Matsui 1999).

To reduce the computational cost of BNP-MAP, the 20-minute recording is divided into one-minute segments. Each one-minute segment is analyzed by BNP-MAP. On the contrary, the whole recording is analyzed by HARK. The source number, a parameter for MUSIC in HARK, is set to 3 experimentally. Figure 4 depicts the sound sources separated by BNP-MAP. Calls of Frog1 and Frog2 are clearly separated as *SS6* and *SS3*, respectively. (For details, see its caption.)



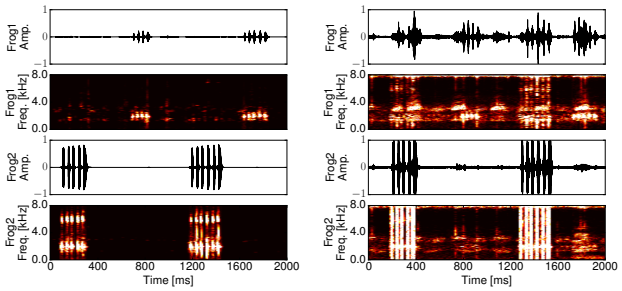
(a) BNP-MAP (b) HARK (GHDSS-AS)

Figure 5: Frequency of Frog1 and Frog2's calls obtained by BNP-MAP and HARK. In circular charts, the direction of segments indicate the calling direction and the length indicate how frequently each frog calls were detected throughout the recording.

Comparison of localization by BNP-MAP and by HARK The ground truth almost corresponds to the result of BNP-MAP. Figure 5 (a) and (b) depict the frequency of the directions of the frog calls separated by BNP-MAP and by HARK during the whole time. Localization of Frog2 by HARK is almost the same as by BNP-MAP, while localization of Frog1 by HARK is erroneous by about 20 [deg]. This is because GHDSS-AS uses the "standard" transfer functions contained in the HARK distribution. The standard transfer functions were measured each 5 degree in a quiet room, and thus differs from actual transfer functions in a paddy rice field. That's why the localization of Frog2 was consistent. Considering the fact that the amplitude of *SS6* of Frog1 calls is much smaller than that of *SS3* of Frog2 shown in Figure 4, BNP-MAP is more robust against the power of sound source than GHDSS-AS of HARK.

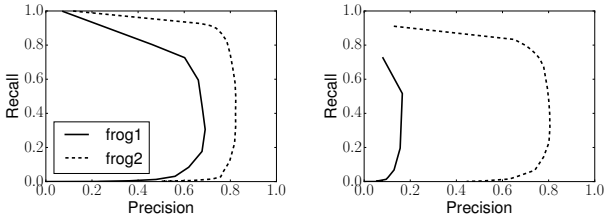
Comparison of separation Figure 6 shows the waveform and spectrogram of sound source separated by either BNP-MAP or GHDSS-AS for Frog 1 and Frog2. Waveforms by BNP-MAP are excerpts from Figure 4. While separated signals for Frog1 separated by BNP-MAP contain small noise, that of both Frog2 and Frog2 separated by GHDSS-AS contain larger noise. These noise are due to crosstalk of other sound sources such as three individuals of *H. japonica* and other noise sources. This demonstrates that BNP-MAP separation outperforms GHDSS-AS.

Comparison of separation performance in VAD For evaluating the separation performance, we use *voice activity detection (VAD)*, or the duration of calling, obtained by Firefly as the reference. For Frog1 and Frog2, one Firefly closest to each frog position shown in Figure 3 is selected. The



(a) BNP-MAP (b) HARK (GHDSS-AS)

Figure 6: Separated sound source for the two frogs by BNP-MAP and GHDSS-AS in waveform and spectrogram. Upper pairs are for Frog1, while lower for Frog2.



(a) BNP-MAP HARK (GHDSS-AS)

Figure 7: Precision-Recall curves of quality of separation in terms of VAD. VAD of Firefly is used as reference.

VAD results of Firefly is a temporal sequence of binary mask calculated by thresholding each temporal brightness pattern. The VAD results of BNP-MAP and HARK are calculated by thresholding the amplitude of the separated signal. Let α be the threshold to the separated signal, the precision and recall were calculated as follows:

$$Precision^i(\alpha) = \frac{\#\{t|F_t^i = 1 \text{ and } |S_t^i| > \alpha\}}{\#\{t|\ |S_t^i| > \alpha\}}$$

$$Recall^i(\alpha) = \frac{\#\{t|F_t^i = 1 \text{ and } |S_t^i| > \alpha\}}{\#\{t|F_t^i = 1\}}$$

where t, i, F_t^i, and S_t^i represent time, frog index (1 or 2), the VAD result of Firefly, and the amplitude of the separated signal by either BNP-MAP or HARK. $\#\{t|C\}$ denotes the duration where condition C is satisfied. Since the temporal resolution of the Firefly (video) and audio signal differ, the audio stream in 16 kHz recording and video with 29.97 fps are synchronized such that the sound amplitude is obtained with the rate of 29.97 fps.

Figure 7 shows the precision-recall curves by varying the threshold to the separated signals α from 0 to the maximum amplitude. The precision of Frog1 by GHDSS-AS is much lower than by BNP-MAP. Since GHDSS-AS requires the sound source direction in separation, worse localization deteriorates the separation performance. This property is common in most sound source separation algorithm except BSS. In addition, sound level of calls of Frog1 is weak. Therefore, the performance of separating Frog1' calls becomes poor.

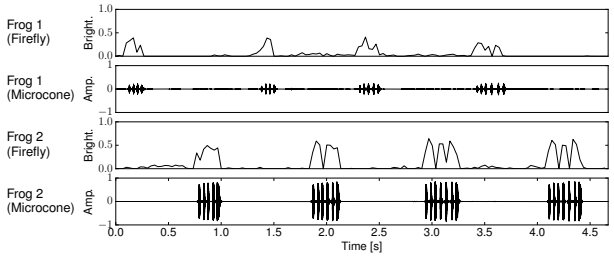


Figure 8: Temporal Synchrony shown by visual and auditory analysis for two individuals of *R. schlegelii*. Visual data by Firefly is shown in scaled brightness and auditory data by BNP-MAP is shown in scaled amplitude.

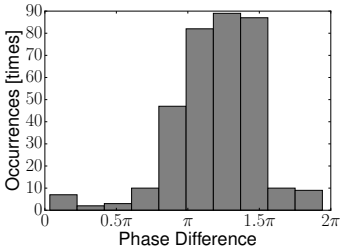


Figure 9: Histogram of phase difference $\phi_{1,2}$ between calls of Frog 1 and Frog2 for whole recording.

Temporal synchrony between BNP-MAP and Firefly Figure 8 shows the illumination patterns of two Fireflies deployed besides calling frogs and also the waveforms of separated sounds by BNP-MAP. It is demonstrated that the two frogs call alternately with anti-phase (Aihara et al. 2011). Figure 9 illustrates the phase difference $\phi_{1,2}$ between the calls of Frog1 and Frog2 which are accumulated by the same method as (Aihara et al. 2014). This figure also illustrates the anti-phase calling because the main peak of the histogram is on $\phi_{1,2} = \pi$ instead of $\phi_{1,2} = 0$.

Moreover, the separated sounds by BNP-MAP reveal that several pulses are included in each call of *R. schlegelii*. The interval of the pulses is about 40 ms, which corresponds to 25 Hz. Since it is more than the Nyquist frequency of the sampling rate of video data capturing the light pattern of Firefly (29.97Hz), Firefly fails to capture all pulses.

Conclusion

In this paper, we exploit BNP-MAP in analyzing the spatio-temporal patterns of the frog chorus. BNP-MAP succeeds in simultaneously estimating localization and separation of calls of two individuals of *R. schlegelii* in spite of frog calls of other species and noise sources. BNP-MAP outperforms GHDSS-AS in localization and separation. BNP-MAP increases the temporal resolution from video rate (29.97 fps) to audio rate (16 kHz). This higher resolution enables to analyze fine temporal structures of frog calls. For example, BNP-MAP reveals the trill-like calling pattern of *R. schlegelii*.

The next step is to design and implement an audio-visual

integrated system with BNP-MAP and Firefly. It is generally difficult for microphone array processing algorithms to separate sound sources whose directions are close to each other and estimate their distances. The audio-visual integration can be considered as combination of macroscopic analysis with Firefly and microscopic analysis with BNP-MAP. Such hybrid analysis should be developed for actual applications. We will further evaluate the performance of BNP-MAP and its audio-visual integration with Firefly in outdoor environments.

Acknowledgments

This study was partially supported by JSPS KAKENHI 24220006 and 26-258.

References

Aihara, I.; Takeda, R.; Mizumoto, T.; Otsuka, T.; Takahashi, T.; Okuno, H. G.; and Aihara, K. 2011. Complex and Transitive Synchronization in a Frustrated System of Calling Frogs. *Physical Review E* 83(031913).

Aihara, I.; Mizumoto, T.; Otsuka, T.; Awano, H.; Nagira, K.; Okuno, H. G.; and Aihara, K. 2014. Spatio-Temporal Dynamics in Collective Frog Choruses Examined by Mathematical Modeling and Field Observations. *Scientific Reports* 4(3891).

Bee, M. A.; Schwartz, J. J.; and Summers, K. 2013. All's well that begins wells: celebrating 60 years of *animal behaviour* and 36 years of research on anuran social behaviour. Animal Behaviour*, Anniversary Essay* 85(1):5–18.

Cody, M. L., and Brown, J. H. 1969. Song asynchrony in neighbouring bird species. *Nature* 222:778–780.

Dogesys, J.; Rose, A.; Patel, A.; and Nagpal, R. 2007. Desync: Self-organizing desynchronization and tdma on wireless sensor networks. In *In* Proceedings of Internaitonal Conference on Information Processing in Sensor Networks (IPSN), 11–20.

Gomes, C. P. 2009. Computational sustainability: Computational methods for a sustainable environment, economy, and society. *The Bridge* 39(4):5–13.

Jones, D. L.; Jones, R. L.; and Ratnam, R. 2014. Calling dynamics and call synchronization in a local group of unison bout callers. *Journal of Comparative Physiology A* 200(1):93–107.

Kelling, S.; Lagoze, C.; Wong, W.-K.; Yu, J.; Damoulas, T.; Gerbracht, J.; Fink, D.; and Gomes, C. 2013. Ebird: a human/computer learning network to improve biodiversity conservation and research. *AI Magazine* 34(1).

Maeda, N., and Matsui, M. 1999. *Frogs and toads of Japan. 3rd ed.* Bun-ichi Shobo Shuppan Co., Ltd., Tokyo.

Megela Simmons, A.; Simmons, J. A.; and Bates, M. E. 2008. Analyzing acoustic interactions in natural bullfrog choruses. *Journal of Comparative Psychology* 122(3):274–282.

Mizumoto, T.; Aihara, I.; Otsuka, T.; Takeda, R.; Aihara, K.; and Okuno, H. G. 2011. Sound imaging of nocturnal animal calls in their natural habitat. *Journal of Comparative Physiology A* 197(9):915–921.

Nakadai, K.; Takahashi, T.; Okuno, H. G.; Nakajima, H.; Hasegawa, Y.; and Tsujino, H. 2010. Design and implementation of robot audition system'hark'open source software for listening to three simultaneous speakers. *Advanced Robotics* 24(5-6):739–761.

Otsuka, T.; Ishiguro, K.; Sawada, H.; and Okuno, H. G. 2012. Bayesian unification of sound source localization and separation with permutation resolution. In *the 26th AAAI Conference on Artificial Intelligence (AAAI-12)*, 2038–2045.

Otsuka, T.; Ishiguro, K.; Sawada, H.; and Okuno, H. G. 2014. Bayesian nonparametrics for microphone array processing. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 22(2):493–504.

Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R. E.; Fink, D.; and Kelling, S. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142(10):2282–2292.

Suzuki, R.; Taylor, C. E.; and Cody, M. L. 2014. Temporal partitioning to avoid soundspace overlap by bird communities. In *the 26th International Ornithological Congress*, 41.