一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-055-2
(11/22)

生成モデルに基づく鳴き声を用いた 鳥類に対するプレイバック実験の試行

A playback experiment on songbirds using simulated vocalizations based on a generative model

炭谷晋司[1*]　鈴木麗璽[1]　松林志保[2]　有田隆也[1]　中臺一博[3,4]　奥乃博[5]
Shinji Sumitani[1]　Reiji Suzuki[1]　Shiho Matsubayashi[2]　Takaya Arita[3]
Kazuhiro Nakadai[3,4]　Hiroshi G. Okuno[5]

[1] 名古屋大学，Nagoya University
[2] 大阪大学, Osaka University
[3] 東京工業大学，Tokyo Institute of Technology
[4] （株）ホンダ・リサーチ・インスティチュート・ジャパン， Honda Research Institute Japan
[5] 早稲田大学，Waseda University

Abstract: 本稿は，鳥類の音声コミュニケーションにおける，音響特性が与える影響をより理解するための実験の枠組みの構築を目的とする．予備実験として，変分オートエンコーダ（Variational Autoencoder, VAE）で学習したウグイスの 2 種類（H・L 型）の歌の特徴空間を利用して，両者および中間の歌を生成し，それらの歌を用いたプレイバック実験を行い，複数のマイクアレイによる 2 次元音源定位で行動パターンを分析した．その結果，生成音に対しても個体は反応を示し，歌として認識することが示唆された．

1　はじめに

鳥類にとって歌（さえずり）は，縄張りの主張や求愛などに用いられる重要な意思伝達手段である [1]．歌行動に基づく種間・個体間相互作用は，歌うタイミングなどの時間的相互作用，なわばり関係等の空間的相互作用，音の種類や周波数の特徴などの音響的相互作用といった様々な次元を持つ．そのため，この理解は，生物由来の音と周囲の環境との関係を様々な次元で理解を試みる生態音響学 [2] では重要な主題とされており，我々は特に歌う鳥の集団を歌を介して相互作用する複雑系とみなして理解を試みている [14].

我々は，鳥類の歌行動理解に活用することを目的として，ロボット聴覚オープンソースソフトウェア HARK[8] と市販のマイクアレイで構築される，録音分析システム HARKBird[15] を構築し，鳥類の歌行動のタイミング，方向（位置），および，音源の自動抽出の試行や．その応用や発展の可能性について検討してきた [16, 17, 12]. 現在では，機械学習を利用した音源の分類ツールが搭載され，より容易に音源の分類が可能となっている [13].

その中でも，鳴き声の種類に基づく音響的相互作用により焦点を合わせた研究として，スピーカから鳥類の歌を再生し，再生音が鳥類個体のに与える影響を調査するプレイバック実験を行ってきた．特に，ウグイスに対して，同種の持つ 2 種の歌（H型：求愛やなわばりの宣言，L型：警戒）をタイミングを変更して様々な条件でプレイバック実験を行い，その時の注目個体が歌を発した方向を単一のマイクアレイを用いて分析を行った [17]．その結果，再生条件に応じて注目個体の歌の頻度やH型・L型の割合，マイクから見た個体の方向の変化の傾向に違いがあることが定量的に示された．これは，従来の人手による観測では容易でなかった，多様な次元の詳細な歌行動傾向を把握可能なことを示唆している．

一方，近年では機械学習を用いて鳥類の歌行動を理解する試みがあり，次元削減アルゴリズムを用いた鳥類の歌のシーケンスに関する調査手法や，生成モデルに基づいた歌コミュニケーションにおける知覚や行動に関する調査手法が提案されるなど，機械学習を用いた鳥類の音響的相互作用の理解が注目されている [10]. 特に，生成モデルに関しては，潜在空間から新しい歌構造，音素構造を生成できるため，プレイバック実験に盛り込むことで歌構造の役割や，その適応性に関してもさらに深く理解できることが期待できる．

*連絡先： 名古屋大学大学院情報学研究科
〒 464-8601 愛知県名古屋市千種区不老町
E-mail:sumitani@alife.cs.is.nagoya-u.ac.jp

本研究は，鳥類の音声コミュニケーションにおける時間的・空間的・音響的関係を詳細に抽出可能な観測実験分析フレームワークをロボット聴覚技術と機械学習を融合して実現することを目的とする．本稿では，その中でも，上記の生成モデルを利用した歌の音響特性が与える影響をより理解するための実験の枠組みの構築を検討する．具体的には，生成モデルの 1 つである，変分オートエンコーダ（Variational Autoencoder, VAE）[6] を用いて，ウグイス *Horornis diphone* の 2 種類（H・L 型）の歌を学習し，その潜在空間から人工的な歌の生成を行った．その歌を利用してプレイバック実験を行い，まずウグイス個体は生成音を歌として認識しうるか，認識した場合，その中でも役割に違いのある H 型と L 型の中間の合成（混合音）にどのような反応を示すかについて調査した．それぞれのプレイバック実験では，複数のマイクアレイによる録音を行い，2 次元での音源定位によりプレイバックの影響を調査した．

2 手法

2.1 対象種

ウグイス *Horornis diphone* は,「ホーホケキョ」と聞こえる高いピッチで歌う H 型と「ホーホホホケキョ」などとホーの部分が断続する低いピッチで歌う L 型の 2 種の歌を持つ [3]．スピーカから同種の歌のプレイバックを行い，その間に行われたウグイス個体のさえずり回数と種類の計測を行った百瀬の実験では，プレイバックのある間は再生を行っていない期間と比較して H 型の頻度が減少し，L 型の割合が増加することを人手による観測結果から示した．このことから，H 型には縄張りの主張や求愛，L 型には近隣個体への威嚇の意味があるとされている [7]．前述の我々のウグイスに対するプレイバック実験 [17] は，百瀬の実験結果と一致した上に，移動に関しては，L 型の割合が高いほど移動の頻度が高く，スピーカ上を通過するような飛び方を頻繁に行うこと，距離の大きな移動後には L 型を歌う傾向があることを示した．さらに，人工物と生物の相互作用の調査のために行ったボーカロイド初音ミクによって作成した L 型の真似音をプレイバックした結果，ウグイス個体は再生音に興味を示し，スピーカ上を飛び回るような挙動を示した [20]．

上記のように，ウグイスは 2 種の歌をうまく使い分けることで，意思伝達を行っている．従来，歌の構造の地域差等に関する議論はなされてきた [4, 5] が，音響特徴から見た 2 種の歌の間の関係やその適応的意義に関しては議論の余地があると考えられる，生成モデルによって H 型，L 型の中間音を作成することで，両者の関係や役割の理解に貢献しうると期待し実験を行った．

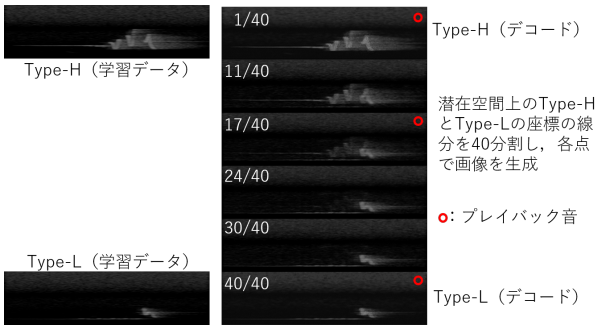

図 1: VAE 潜在空間から生成した歌

2.2 VAE による歌の生成

VAE[6] は，潜在変数が正規分布に従うように学習させるオートエンコーダである．その特徴から，デコーダ部分で生成される出力は潜在空間上で連続的な変化を伴うため，イメージの合成手法として応用される．今回は，ウグイスの H 型，L 型の歌それぞれの録音から 492×128 のグレースケールのスペクトログラム画像を作成し，それをデータセットとして学習を行った．得られた特徴空間から H 型，L 型および複合音の画像を生成し，その画像を再度音声信号に変換することで人工的な歌を生成した．

学習に用いた歌は，今回の実験対象となるウグイスとは異なるウグイス個体の H 型，L 型である．今回は，先行研究 [17] との知見の比較を念頭に，そこで実際にプレイバックに使用した H 型，L 型の歌それぞれ 1 個ずつからなる最小のデータセットを用いた．これは，生成モデルとしてはやや例外な手法であるが，予備的知見として，デコード・生成処理を経た上記のプレイバック音をウグイス個体が歌として認識するかを確かめたいということと，使用した歌の特徴空間において 2 種の中間の歌を容易に選択し生成可能ということによる．

ネットワーク構造には，エンコード側は 7 層のフィルタサイズ 3×3 の畳み込み層と 2 層の全結合層，デコード側はエンコード側と計算が逆となる同じ数の層を用いた．図 1 に学習に用いた 2 種のデータと，生成した画像を示す．生成した画像は H 型から L 型にかけて徐々に変化する様子が確認できる．

2.3 プレイバック実験

実験は，2019 年 5 月 4 日の午前中，名古屋大学大学院生命農学研究科付属フィールド科学教育研究センター稲武フィールド（愛知県豊田市稲武町）の森林で行った．予めウグイス 1 個体が鳴く場所を調査し，その場所にノート PC（TOUGHBOOK CF-C2; Panasonic）に

USB 接続した 2 つのマイクロホンアレイ（TAMAGO-03; System in Frontier Inc.）を三脚に固定し，約 7m ほど離して配置した．また，スピーカ（Mega; Tronsmart）は各マイクアレイからそれぞれ 7m，5m ほど離れた位置にある地上 2m 付近の枝の上に配置した（図 2）．

図 2: 実験風景

プレイバック音には，前節の VAE より生成した H 型，L 型および H 型と混合音 1 種を用いた（図 1 参照）．混合音に関しては，学習に用いた H 型，L 型のスペクトログラム画像をエンコードしたときに得られる H 型，L 型それぞれの潜在空間の座標間の線分を 40 分割し，H 型から 17 番目の座標を用いて生成した画像を音声に変換したものを用いた．これは実際に聞くことにより体感的に中間らしい音を選んだ．プレイバックの方法としては，各再生音を 2 度再生し，2 分間のインターバルを繰り返す方法を採用し，混合音（M），H 型，L 型，混合音，プレイバックなしの録音の順でそれぞれ 30 分間の実験を連続して行った．前のプレイバック実験や人の接近によるウグイス個体への影響を考慮するために，実験の間で数分程度間隔を開け，また，プレイバックを 30 分よりやや長めに行った．分析データは録音開始後 5 分からの 30 分間のデータを採用した

2.4 分析

まず，HARKBird を用いてウグイス個体の歌をうまく定位するように定位のパラメータを適宜調整し，音源の定位・分離を行った．短時間フーリエ変換によって得られた各チャネルのスペクトログラムから MUSIC 法 [11] を用いて音源定位を行い，その定位結果に基づいて GHDSS 法 [9] を用いて対応する音源を抽出した．その後，抽出した分離音源を 100×64 のグレースケール画像に変換し，t-SNE（t-distributed Stochastic Neighbor Embedding）[18] により 2 次元まで次元の圧縮を行い，得られた特徴空間を用いてプレイバック音およびウグイス個体の歌を分類した．分類した結果をもとに，同じラベルで分類された各マイクアレイの定位音源のペアで Sumitani らの MUSIC スペクトルに基づく三角測

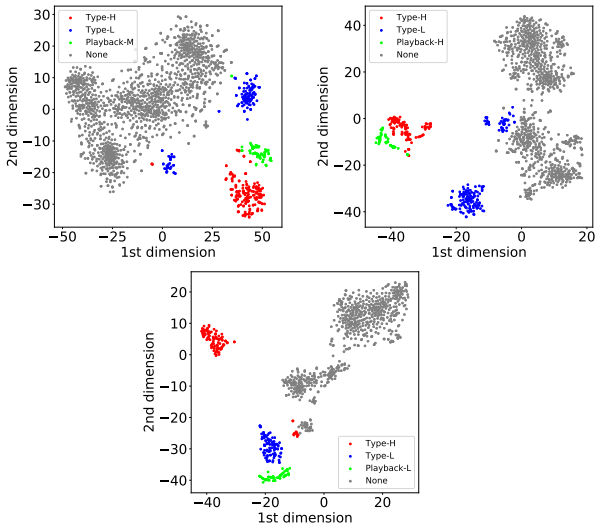

図 3: t-SNE 特徴空間上での音源分離（左上：混合音再生時，右上：H 型再生時，下：L 型再生時

量の手法 [12] で二次元定位を行い，移動距離や歌の頻度等の定量的観測を試みた．2 回目の混合音を用いたプレイバック実験では，ウグイス個体の歌う位置が，2 つのマイクアレイの直線状付近で二次元定位が困難であり，またプレイバックなしの録音ではほとんどウグイス個体が鳴いていなかったため，今回はそれらの結果の報告を割愛する．

3 結果

まず，各実験において定位（到来方向推定）された音源の音響的特徴とその分布について述べる．図 3 は各実験条件における，音源定位・分離結果に基づいて t-SNE[18] を実行し，その特徴空間上でウグイス個体の歌（H 型，L 型）とプレイバック音を抽出した結果である．多くの注目個体の歌やプレイバック音は，他の定位音源とは独立してクラスタを形成し同じ音同士で集まり分布した．また，プレイバック音に着目してみると，H 型のプレイバック音は個体の H 型近傍に，L 型のプレイバック音は個体の L 型近傍に，複合音は個体の H 型，L 型の中央付近に分布しており，構造間の関係をゆるやかに反映した分布であることが確認できる．“None” としてラベル付けされている音源の多くは，システム配置場所付近にあった 2 つの小川を流れる水音が占めていた．

次に，分類結果を用いて 2 次元定位を行った．図 4 は，各再生音をプレイバックしたときのウグイス個体の歌とプレイバック音の 2 次元定位結果を示す．概して，定位結果はシステム周りに集中しており，ウグイス個体はシステム周りでスピーカ音を警戒して動いて

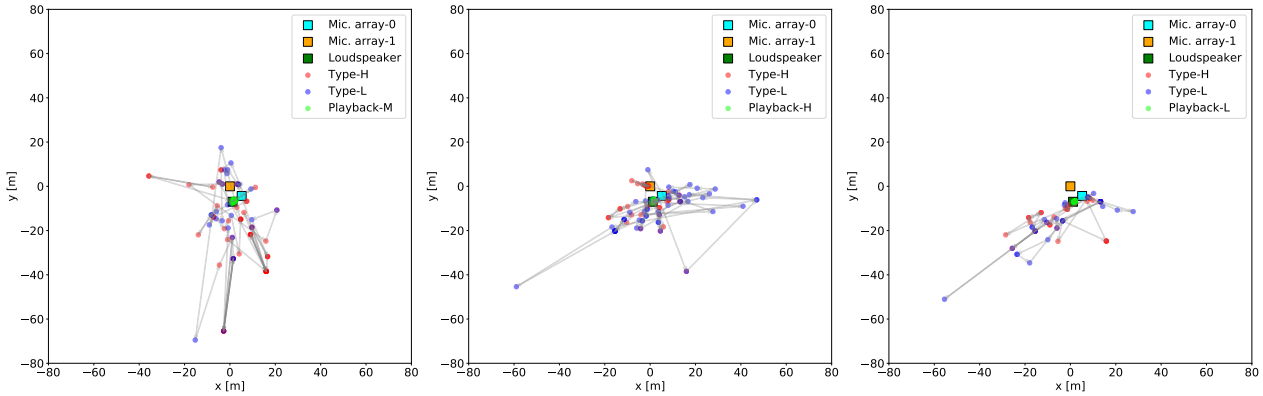

図 4: 2 次元定位結果（左：混合音再生時，中央：H 型再生時，右：L 型再生時）ウグイス個体の歌（赤点：H 型，青点：L 型）を繋ぐ矢印は，定位できたウグイス個体の歌を時系列順に繋いだもの

表 1: 各実験条件でのウグイス個体の歌行動．下線はプレイバックに対して最も警戒した反応（最高 L 型頻度・最低 H 型割合・最近平均距離）を示したと考えられる挙動．

再生音	Mic. array	H 型	L 型	合計	H 型の割合 H/(H+L)	2 次元定位数 All (H, L)	スピーカから歌定位位置までの平均距離 All（H，L）[m]
混合歌	0	83	60	143	0.580	88 (60, 28)	19.737（21.233，16.531）
	1	80	67	147	0.544		
H 型	0	56	81	137	0.409	98 (37, 61)	12.907（10.313，14.481）
	1	49	93	142	0.345		
L 型	0	60	49	109	0.550	65 (34, 31)	16.098（17.928，14.430）
	1	51	49	100	0.510		

いる様子が確認できる．先行研究 [17] において，プレイバックのない録音のみの分析では，ウグイス個体は樹上等のいくつかの場所で留まりくりかえし歌行動を行う傾向があり，プレイバックのある場合は頻繁にスピーカ周りを移動する傾向があった．このことを考慮すれば，ウグイス個体は VAE により生成された歌を種の歌として認識していると考えられる．

なお，全ての実験条件において，スピーカから大きく離れて定位されている定位音がいくつかあるが，これらの音源は全て注目したウグイス個体のものであった．それぞれのマイクアレイの定位結果を確認したところ，分離音源ははっきりとした波形であり，システムからそう遠くない場所で歌われた歌であることが考えられる．そのため，これらの結果は定位誤差により生じた例外である可能性がある．また，それらは図中下方に多く存在するため，その方面の環境の音響特性の影響を受けた可能性もある．

表 1 に，各実験条件でのウグイス個体の歌を定位した回数をまとめた．各マイクアレイで定位した歌の総数は，混合音の再生実験ではマイクアレイ 0，1 でそれぞれ 143 回，147 回の定位数，H 型再生時には，137 回，142 回，L 型再生時には，109 回，100 回と，実験が後になっていくにつれて減少傾向にあった．これは，実験を連続して行ったために，いわゆる学習効果によって，ウグイス個体がプレイバックに慣れていったことが考えられる．あるいは鳥類は朝方に頻繁に鳴き，次第に静かになる傾向があるため，実験対象としたウグイス個体に関しても歌う頻度が減少したことも考えられる．

各条件で歌った H 型と L 型の割合に着目みてみると，混合歌を再生した場合には，各マイクアレイの結果としてそれぞれ 0.580，0.544，H 型を再生した場合は 0.409，0.345，L 型を再生した場合は 0.550，0.510 であった．これらの結果は，我々の先行研究におけるプレイバックのない録音でのウグイス個体が歌った H 型の割合と比較しても小さい値であり，また，プレイバック実験時の割合に比較的似ていた．以上のことから，歌の割合からもウグイス個体がプレイバック音に対して反応を示しているといえる．

一方で，時間的学習効果を考慮すれば，冒頭に行った混合歌再生時の方がより警戒を示し，H 型の割合はより小さくなることが期待されるが，H 型，もしくは，L 型再生時における H 型の割合の方が小さかった．これは，混合歌は，H・L 型と比べて反応が低いことを示していると考えられる．

2 次元定位数は，各マイクアレイで定位できた音源数に比較すると数を減らした．それぞれのマイクアレイの音源定位結果を詳細に調べた結果，以下の原因によることが確認された：1．どちらか片側のマイクアレイで歌が定位できておらず 2 次元定位に至らなかった．

2．ペアとなる音源はあるが，各マイクアレイの定位方向が平行か、あるいは定位方向の延長線が広がる向きで定位しているため，交点ができなかった．これらは，実験場所が木々の多く生い茂る地点であり，かつ近隣に 2 つの河川が流れる地点であったため，音源定位がそれらの反響や雑音の影響を大きく受けたため起きてしまったと考えられる．また，二次元定位が困難な場所である，マイクアレイを繋ぐ直線上付近にウグイス個体のよく鳴く位置があったことも原因として挙げられる．

さらに，ウグイス個体が空間的にどのように歌っていたかを調査するために，各条件においてウグイス個体が歌った H 型，L 型あるいは双方における，スピーカから歌の定位位置までの平均距離を求めた．3 つの条件を比較すると，H 型，L 型合計での平均距離は，混合音再生時（19.737 m），L 型再生時（16.098 m），H 型再生時（12.907 m）の順であった．これは，H 型の割合が大きい順と一致しており，混合歌再生時のプレイバック音に対する警戒の度合いの小ささを反映することが示唆される．また，これらの距離の違いは特に H 型を歌う場合に顕著みられ，L 型には差が大きくない．H 型は縄張りの宣言や求愛の意味を持つため，不特定多数に対する歌であり，一方で L 型は近隣個体への威嚇の意味を持ち，その相手個体に対して発する歌であるとされている．そのため，固定されたスピーカに対しては L 型は同等の距離を保ちがちである一方，H 型の歌はより広い範囲で歌われた結果，歌う位置に差がみられたことが考えられる．

以上から，今回再生した生成音 3 種に対してウグイスはいずれも明らかな反応を示したが，H 型と L 型の混合音に対しては最も反応が弱く，遠くで H 型の歌を歌いがちであることが推測された．混合音は対象個体にとっては同種に近いが聞きなれない歌であるため，警戒しつつも戸惑っている状況であった可能性もあると考えている．

4　おわりに

音響特性が与える影響をより理解するための実験の枠組みの構築検討のために，ウグイスの 2 種類（H・L 型）の歌を VAE で学習し，その潜在空間から人工的な歌の生成を試み，その歌を利用してプレイバック実験を行った．その結果，ウグイス個体は，H 型，L 型の混合音も歌として認識することがわかった．その影響の理解にはより詳細な分析が必要であるが，生成モデルから生成した歌がプレイバック実験等の鳥類の歌コミュニケーションにおける音響的相互作用理解に利用可能であることを示した．今回は，2 種の歌の生成と，その中間音の生成のみの検討であったが，VAE の潜在空間ではある特徴を強調して生成することも可能であるため [19]，例えばウグイスに関していえば，H 型，L 型の特徴を強調した歌の生成も可能であると考える．現在，大きなデータセットを用いた学習も進めており，より VAE の特性を活かした，歌構造の役割やその適応性に関する知見を得たいと考えている．

謝辞

高部直紀氏（名古屋大学）のフィールド調査への協力と，Zhao Hao 氏（名古屋大学）のデータ分析への協力に謝意を表す．本研究の一部は JSPS 科研費 JP18K11467，JP19KK0260，JP17H06383（#4903）の助成を受けた．

参考文献

[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] S. Hamao, “Japanese bush warbler,” *Bird Research News*, vol. 4, no. 2, 2007.

[4] S. Hamao, “Acoustic structure of songs in island populations of the japanese bush warbler, cettia diphone, in relation to sexual selection,” *Journal of Ethology*, vol. 31, no. 1, pp. 9–15, Jan 2013. [Online]. Available: https://doi.org/10.1007/s10164-012-0341-1

[5] S. Hamao, “Rapid change in song structure in introduced japanese bush-warblers (cettia diphone) in hawai‘i,” *Pacific Science*, vol. 69, no. 1, pp. 59 – 66, 2015. [Online]. Available: https://doi.org/10.2984/69.1.4

[6] D. P. Kingma and M. Welling, “An introduction to variational autoencoders,” *CoRR*, vol. abs/1906.02691, 2019. [Online]. Available: http://arxiv.org/abs/1906.02691

[7] H. Momose, *Mechanism of maintaining territories by acoustic communication. Reproductive strategies of birds*. Toukaidaigaku-shuppankai, Tokyo, Japan, 2016.

[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. 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.

[10] T. Sainburg, M. Thielk, and T. Gentner, “Animal vocalization generative network (AVGN): A method for visualizing, understanding, and sampling from animal communicative repertoires,” in *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*, 2019, p. 3563.

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

[12] 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*, 2018, pp. 2485–2490.

[13] 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 ICASSP 2019*, 2019, pp. 8246–8250.

[14] R. Suzuki and M. Cody, “Complex systems approaches to temporal soundspace partitioning in bird communities as a self-organizing phenomenon based on behavioral plasticity,” *Artificial Life and Robotics*, pp. 1–6, 09 2019.

[15] 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.

[16] 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.

[17] R. Suzuki, S. Sumitani, Naren, 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.

[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.

[19] T. White, “Sampling generative networks: Notes on a few effective techniques,” *CoRR*, vol. abs/1609.04468, 2016. [Online]. Available: http://arxiv.org/abs/1609.04468

[20] 炭谷晋司，松林志保，鈴木麗璽，“マイクロホンアレイを用いたプレイバック実験に基づくウグイスのさえずりの方向分布分析”，**日本鳥学会** *2016* **年度大会講演要旨集**，p. 124，2016.