Towards Differentiable Motor Control of Bird Vocalizations

Vincent Lostanlen

Towards Differentiable Motor Control of Bird Vocalizations

Vincent Lostanlen[1]

[1] Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

vincent.lostanlen@ls2n.fr

Abstract

Machine learning is ready to transform the experimental protocol of birdsong acquisition and playback in ethology and integrative neuroscience . An emerging methodology, known as differentiable digital signal processing (DDSP), allows to train neural networks for machine listening so as to fit the synthesis parameters which correspond to unlabeled audio data. In this short article, I present the value and of extending DDSP, initially developed for speech and music processing, to avian bioacoustics. The main two challenges reside in the definition of a suitable decoder and learning objective. I review some prior publications in biomechanical models of vocal production for passerines, similarity computing, and differentiable solvers of ordinary differential equations. Together, these publications hint at the feasibility of a fully automated and unsupervised algorithm for biologically plausible resynthesis of birdsong.

Index Terms: birdsong, model-based deep learning, physical modeling synthesis

1. Extended abstract

Over the past decade, the renewed interest for deep learning in signal processing has led to a new generation of systems for passive acoustic monitoring [1]. For example, BirdNET is a deep neural network which detects bird vocalizations from acoustic sensor data and recognizes the corresponding species according to a predefined taxonomy [2]. Comparable solutions exist for flight calls [3] and for open taxonomies [4]. Yet, in these examples, the machine listening system reduces birdsong to a sequence of time segments whose boundaries align with the onset of offset of each song bout [5]. In doing so, it erases spectrotemporal patterns which are attributable to intraclass variability.

Although per-species timings may suffice for ecologists who study wild avian populations, ethologists and neuroscientists often depend on a richer description of birdsong content as part of their research protocols. There is abundant literature on the evolutionary and developmental aspects of vocal learning in songbirds: e.g., zebra finches, canaries, and budgerigars. Moreover, a well-known study by Pepperberg *et al.* has shown the exceptional abilities of an African gray parrot in terms of functional vocalizations when interacting with humans in English [6]. In these studies, automating species classification would be useless, since the specimens are known and kept in an aviary. Rather, a valuable source of information on animal behavior is found in the fundamental frequency (f_0) contours of animal vocalizations. Unfortunately, f_0 tracking is more difficult for birdsong than for solo music or speech, due to higher rates of amplitude and frequency modulation. Hence, if f_0 tracking of birdsong is to be automated in the future, it requires a dedicated approach.

Despite the proven merits of machine learning in bioacoustic detection and classification, the task of f_0 tracking comes with a challenge of its own: that of collecting training data. Indeed, the expert annotation of f_0 contours is even more costly and time-consuming than that of species-specific vocal activity detection. For lack of available ground truth, the task must be approached via unsupervised learning techniques. Historically, some of these techniques have been successfully applied to marine bioacoustics (e.g., [7]) but rarely ever to birdsong, with the notable exception of spherical k-means [8]. Still, up to recently, unsupervised representation learning algorithms were unsuitable for highly time-varying and spectrally rich signals such as birdsong.

The situation has changed recently with the introduction of a new methodological framework for unsupervised learning in speech and music, known as differentiable digital signal processing (DDSP). The key idea behind DDSP is to train an autoencoder whose encoder contains learnable parameters but whose decoder does not, while both are compatible with automatic differentiation. Minimizing the reconstruction of error of the autoencoder over a training set of unlabeled natural sounds is tantamount to solving an inverse problem whose associated direct problem is specified by the decoder [9]. In its earliest version, the DDSP decoder was a simple additive sinusoidal model with random Gaussian noise and reverberation. More recently, a broader range of decoders has been developed, directly mimicking the state of the art in acoustical simulation and virtual analog audio effects: let us refer to [10] for a review. Therefore, DDSP is a kind of "model-based deep learning" in the sense that it hybridizes physics-driven and data-driven insights so as to learn an informative representation of natural sounds [11].

I propose to adapt the DDSP framework to the long-standing problem of unsupervised representation learning of birdsong. DDSP has already been successfully applied to f_0 estimation in music signals, under the name of DDSP-inv [12]. My scientific hypothesis is that DDSP-inv has the potential to improve the state of the art in analysis–synthesis of birdsong, currently held by hidden Markov models (HMM) [13] and, more recently, WaveNet [14]. However, I believe that the standard formulation of DDSP, based on sinusoidal models and multiscale spectrogram loss (MSS), is not suitable to birdsong. Indeed, even so the authors of DDSP have presented a demonstration of birdsong analysis–resynthesis as part of their "Paint With Music" outreach project, the result does not sound naturalistic [1]. To serve the needs of ethologists and neuroscientists working on captive birds, the components of DDSP must be redesigned.

On one hand, the groundbreaking publications of Mindlin, Laje, Amador, Sitt, Perl, and colleagues have laid the ground-

[1] Link to "Paint With Music" project:
https://magenta.tensorflow.org/paint-with-music

work for a comprehensive physical description of the vocal apparatus in some well-studied songbirds, e.g., zebra finch and canary [15, 16, 17, 18, 19]. The commonality between these publications is to model the syrinx as a nonlinear dynamical system whose parameters have a biomechanical interpretation. For example, [18] apply the theory of Takens–Bogdanov bifurcations to present a dynamical system governed by the following second-order ordinary differential equation (ODE):

$$\ddot{\boldsymbol{x}} = \gamma^2\alpha + \gamma^2\beta\boldsymbol{x} + \gamma\boldsymbol{x}^2 - \gamma\boldsymbol{x}\dot{\boldsymbol{x}} - \gamma\boldsymbol{x}^3 - \gamma, \qquad (1)$$

where $\boldsymbol{x}$ represents the departure of the midpoint position of the oscillating labia in the syrinx, α and β are functions of the air sac pressure and the activity of the ventral syringeal muscle, and γ is a time scaling factor. Although a Python implementation is available[2] to compute $\dot{x}$ from $\boldsymbol{\theta} = (\alpha, \beta, \gamma)$, it depends on NumPy; as such, it is not interoperable with neural network training. We propose to reimplement this synthesizer in PyTorch, a Python framework for differentiable computing. More precisely, the torchdiffeq library [20] allows to program solvers for ordinary differential equation in which the solution (x) may be differentiated with respect to the parameters ($\boldsymbol{\theta}$). Via reverse-mode automatic differentiation, it will be possible to evaluate the gradient of a function of x may with respect to neural network weights $\mathbf{W}$ where $\boldsymbol{\theta}$ is defined as $f_{\boldsymbol{W}}(\boldsymbol{x})$ and $f_{\mathbf{W}}$ is the encoder.

On the other hand, a new generation of differentiable time–frequency representations have the potential to improve the conditioning of the inverse problem in DDSP, which may accelerate gradient-based optimization when training the encoder. For example, a differentiable implementation of the joint time–frequency scattering transform (JTFS) has recently been released as part of the Kymatio package [21]. Prior work on synthetic chirps has confirmed that, with JTFS, parameter estimation is faster, more accurate, and less susceptible to random initialization than MSS [22]. Although there is a gap in acoustical complexity between synthetic chirps and real birdsong, this result is encouraging because it directly addresses the issue of unsupervised learning in the presence of fast spectrotemporal modulations. Another option would be to use a pretrained neural network as feature map for similarity computing between the natural signals and its autoencoded version.

In conclusion, I have described the promise and challenge of learning to control a physical model of birdsong without supervision and have outlined the necessary steps to get there. Beyond the fundamental interest of advancing differentiable digital signal processing (DDSP), its application to birdsong would unlock new research protocols in ethology and neuroscience.

2. Acknowledgements

This work is supported by ANR projet nIrVAna (ANR-23-CE37-0025). I thank Michael Newton and Yining Xie for helpful discussions and thank anonymous reviewers for their feedback.

3. References

[1] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022.

[2] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[3] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, p. e0214168, 2019.

[4] I. Nolasco, S. Singh, V. Morfi, V. Lostanlen, A. Strandburg-Peshkin, E. Vidaña-Vila, L. Gill, H. Pamuła, H. Whitehead, I. Kiskin *et al.*, "Learning to detect an animal sound from five examples," *Ecological informatics*, vol. 77, p. 102258, 2023.

[5] V. Lostanlen and B. Mcfee, "Efficient evaluation algorithms for sound event detection," in *Proceedings of the International Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2023.

[6] I. M. Pepperberg, "Functional vocalizations by an african grey parrot (psittacus erithacus)," *Zeitschrift für Tierpsychologie*, vol. 55, no. 2, pp. 139–160, 1981.

[7] P. Li, X. Liu, H. Klinck, P. Gruden, and M. A. Roch, "Using deep learning to track time× frequency whistle contours of toothed whales without human-annotated training data," *The Journal of the Acoustical Society of America*, vol. 154, no. 1, pp. 502–517, 2023.

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

[9] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable digital signal processing," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[10] B. Hayes, J. Shier, G. Fazekas, A. McPherson, and C. Saitis, "A review of differentiable digital signal processing for music and speech synthesis," *Frontiers in Signal Processing*, vol. 3, p. 1284100, 2024.

[11] G. Richard, V. Lostanlen, Y.-H. Yang, and M. Müller, "Model-based deep learning for music information research," *arXiv preprint arXiv:2406.11540*, 2024.

[12] J. Engel, R. Swavely, L. H. Hantrakul, A. Roberts, and C. Hawthorne, "Self-supervised pitch detection by inverse audio synthesis," in *Proceedings of the ICML Workshop on Self-supervision in Audio and Speech*, 2020.

[13] L. Gutscher, M. Pucher, C. Lozo, M. Hoeschele, and D. C. Mann, "Statistical parametric synthesis of budgerigar songs," in *Proceedings of INTERSPEECH*, 2019.

[14] R. R. Bhatia and T. H. Kinnunen, "An initial study on birdsong re-synthesis using neural vocoders," in *International Conference on Speech and Computer*. Springer, 2022, pp. 64–74.

[15] G. B. Mindlin and R. Laje, *The physics of birdsong*. Springer Science & Business Media, 2005.

[16] A. Amador and G. B. Mindlin, "Beyond harmonic sounds in a simple model for birdsong production," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 18, no. 4, 2008.

[17] J. Sitt, A. Amador, F. Goller, and G. Mindlin, "Dynamical origin of spectrally rich vocalizations in birdsong," *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, vol. 78, no. 1, p. 011905, 2008.

[18] Y. S. Perl, E. M. Arneodo, A. Amador, and G. B. Mindlin, "Nonlinear dynamics and the synthesis of zebra finch song," *International Journal of Bifurcation and Chaos*, vol. 22, no. 10, p. 1250235, 2012.

[19] A. Amador and G. B. Mindlin, "Low dimensional dynamics in birdsong production," *The European Physical Journal B*, vol. 87, pp. 1–8, 2014.

[20] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[21] J. Muradeli, C. Vahidi, C. Wang, H. Han, V. Lostanlen, M. Lagrange, and G. Fazekas, "Differentiable Time-Frequency Scattering On GPU," in *Digital Audio Effects Conference (DAFx)*, 2022.

[22] C. Vahidi, H. Han, C. Wang, M. Lagrange, G. Fazekas, and V. Lostanlen, "Mesostructures: Beyond spectrogram loss in differentiable time-frequency analysis," *Journal of the Audio Engineering Society*, vol. 71, no. 9, pp. 577–585, 2023.

[2] Python implementation: https://github.com/zekearneodo/syrinxsynth