Large-scale audio-language datasets for bioacoustics

Gagan Narula* Marius Miron David Robinson Milad Alizadeh
Masato Hagiwara Ellen Gilsenan-McMahon Sara Keen Benjamin Hoffman
Maddie Cusimano Emmanuel Chemla Matthieu Geist Olivier Pietquin
Earth Species Project

Abstract

We introduce bioacoustic datasets for training and evaluation of audio-language foundation models. The training dataset aggregates 22,000 hours of audio across 44 different tasks, and includes animal vocalizations, human speech, music, and environmental sounds from public sources. The benchmark dataset tests zero-shot transfer on species classification and detection, call-type classification, and other bioacoustic tasks. Most models for conservation, biodiversity monitoring, and ethology are predictive models trained on small datasets with limited species coverage. Our large-scale, cross-taxa multimodal datasets enable the transition to foundational generative models that demonstrate exceptional ability to handle novel data and tasks, exhibit in-context learning, and produce unconstrained output, capabilities that greatly benefit bioacoustics research. These datasets were used to train and evaluate the NatureLM-audio model, the first audio-text language model for bioacoustics that demonstrates effective zero-shot transfer across species and tasks. We explore several possibilities for extending these datasets and furthering the use of generative models in bioacoustics.

1 Introduction

Many vital research questions in behavioral ecology, conservation, and biodiversity monitoring center on analyzing the sounds produced by animals in an ecosystem, also known as *bioacoustics* [Bradbury and Vehrencamp, 1998, Rutz et al., 2023, Stevens et al., 2024]. Animal vocalizations provide biologists and conservationists with essential information about the health of an ecosystem, and are crucial to understand animal communication, behavioral patterns in response to threats, effects of environmental noise, etc. To do so, researchers must identify and extract the vocalizations from long, noisy audio recordings, which can be an arduous manual task. A classical machine learning approach is to develop specialized models for sound event detection and classification of species, call-types, and many more (Stowell [2022], Dufourq et al. [2021]). These task-specific models may speed up analysis; however, their scope is limited, and they often require fine-tuning, demanding machine learning expertise and compute resources.

Observing the recent developments in multimodal foundation models, we believe that jointly modeling diverse taxa and tasks is better than training specialized models in silos. The ability to monitor species across taxa, with finer granularity (call types, lifestage, etc.), and with fewer barriers to entry, would bring the progress LLMs have brought to other fields into conservation. Our solution to these issues is to: **(1)** massively scale up the size of audio datasets relevant to bioacoustic tasks, to improve both the in-domain and out-of-domain generalization of any model, and **(2)** employ audio-language generative models that can be queried via text and audio, without machine learning knowledge. In this work, we build on the datasets used in training NatureLM-audio (Robinson et al. [2025]). In particular, we focus on enhancing the usability and scalability of such a large dataset ($\approx 15TB$) by

*Corresponding author: gagan@earthspecies.org

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

extensively validating and reformatting the data and also integrating it with an existing Python library, `https://huggingface.co/datasets/EarthSpeciesProject/`.

2 Datasets

NatureLM-audio-training is a large and diverse audio-language dataset designed for training bioacoustic models that can generate a natural language answer to a natural language query on a reference audio recording. The dataset can be used for both generative and predictive modeling. Public sources such as Xeno-canto (Vellinga and Planqué [2015]), iNaturalist (Chasmai et al. [2024]), Watkins (Sayigh et al. [2016]), Animal Sound Archive (Museum für Naturkunde Berlin) and others were curated to obtain more than 26 million (audio, text) pairs including animal vocalizations, insects, human speech, music, and environmental sounds. The dataset contains more than 2 million audio files totaling 22,100 hours and 44 bioacoustic tasks. NatureLM-audio model trained on this dataset (Robinson et al. [2025]) shows strong transfer to unseen species and taxonomies in classification as well as call type classification and lifestage estimation. These tasks are part of **BEANS-Zero benchmark** which contains natural language queries and new bioacoustic tasks and datasets. Both are publicly available on the `https://huggingface.co/datasets/EarthSpeciesProject/` library.

3 Conclusion: the generative model data flywheel

Building and releasing large-scale datasets affords the following iterative refinement loop: **(1)** train a large generative model such as NatureLM-audio-training, evaluate on BEANS-Zero benchmark. **(2)** Apply it to a new, unreleased research dataset to obtain a weakly labeled dataset. **(3)** Use the weak labels in an active learning loop: fine-tune the large model (e.g., using adapters like LoRA Hu et al. [2021]) on the weakly labeled dataset. Use active learning methods (Tamkin et al. [2022]) to retrain the model, then label the dataset. Repeat this process until satisfactory performance is achieved on a small, held-out annotated set. **(4)** Finally, publish the research and contribute the dataset to the NatureLM-audio project, increasing its size and diversity, and building a robust data flywheel.

Extensions Several extensions can be made to NatureLM-audio-training and BEANS-Zero benchmark (see Appendix A for data sources). First, NatureLM-audio-training is biased towards the *Aves* group (birds) due to sampling biases in field recordings in citizen science databases like Xeno-canto and iNaturalist. This needs to be resolved by increasing the proportion of other taxonomic groups. A large number of unannotated Passive Acoustic Monitoring data (Gibb et al. [2019]) are available and need to be parsed and labeled with the procedure described earlier. Second, both NatureLM-audio-training and BEANS-Zero benchmark do not contain in-context learning formatted data (Brown et al. [2020]), to aid a model in performing a new task in a few-shot setting. Third, the released NatureLM-audio-training is a single sampling rate dataset, and ideally the next version will provide original sampling rates despite the increase in size. Fourth, generative models can also be trained with reinforcement learning (RL) to "think" longer (Snell et al. [2024]) and generate better, more interpretive answers, which opens another path towards greater test-time performance. To do so, we need to create training data that incorporate verifiable rewards for RL.

Impact Progress in the bioacoustic foundation data front should also lead to improvements in general large audio-language models like Qwen2-Audio (Chu et al. [2024]) due to cross-domain transfer (Ghani et al. [2023]). Foundation models for biodiversity and ecosystem monitoring will significantly impact research in climate science and ecology due to their interconnected nature (Pörtner et al. [2023], with positive spillover effects on urban planning, environmental policy, and the legal rights of nature (Epstein et al. [2023]) effectively reducing biodiversity loss due to human activity. Foundation datasets can (and should) be further extended to include other modalities such as videos, images, and biologger motion data such that a more comprehensive picture of animal behavior and ecosystem health can be obtained beyond bioacoustics.

Acknowledgments and Funding Disclosure

We would like to acknowledge the work of the citizen scientists who uploaded their recordings to xeno-canto.org and iNaturalist which forms a major portion of our datasets. This work was supported by Earth Species Project, a 501(c)(3) nonprofit organization.

References

Jack W Bradbury and Sandra Lee Vehrencamp. *Principles of animal communication*, volume 132. Sinauer Associates Sunderland, MA, 1998.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

Mustafa Chasmai, Alexander Shepard, Subhransu Maji, and Grant Van Horn. The inaturalist sounds dataset. *Advances in Neural Information Processing Systems*, 37:132524–132544, 2024.

Mustafa Chasmai, Alexander Shepard, Subhransu Maji, and Grant Van Horn. The inaturalist sounds dataset, May 2025. URL `https://arxiv.org/abs/2506.00343`.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL `https://arxiv.org/abs/2407.10759`.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020.

Emmanuel Dufourq, Ian Durbach, James P Hansford, Amanda Hoepfner, Heidi Ma, Jessica V Bryant, Christina S Stender, Wenyong Li, Zhiwei Liu, Qing Chen, et al. Automated detection of hainan gibbon calls for passive acoustic monitoring. *Remote Sensing in Ecology and Conservation*, 7(3):475–487, 2021.

Yaffa Epstein, Aaron M. Ellison, Hugo Echeverría, and Jessica K. Abbott. Science and the legal rights of nature. *Science*, 380(6646):eadf4155, 2023. doi: 10.1126/science.adf4155. URL `https://www.science.org/doi/abs/10.1126/science.adf4155`.

E Fonseca, X Favory, J Pons, F Font, and X Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876, 2023.

Rory Gibb, Ella Browning, Paul Glover-Kapfer, and Kate E Jones. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. *Methods in Ecology and Evolution*, 10(2): 169–185, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

Stefan Kahl, Russell Charif, and Holger Klinck. A collection of fully-annotated soundscape recordings from the Northeastern United States, September 2022. URL `https://doi.org/10.5281/zenodo.7079380`.

Jasper Kanes. Marine mammal phonations of Barkley Canyon: A publicly available annotated data set. *J. Acoust. Soc. Am.*, 150(4 Supplement):A48, October 2021. doi: 10.1121/10.0007587. URL `https://doi.org/10.1121/10.0007587`.

Charly Lamothe, Manon Obliger, Paul Best, Régis Trapeau, Sabrina Ravel, Thierry Artières, Ricard Marxer, and Pascal Belin. Marmaudio: A large annotated dataset of vocalizations by common marmosets, March 2025. URL `https://doi.org/10.5281/zenodo.15017207`.

Shun Lei, Yaoxun Xu, Zhiwei Lin, Huaicheng Zhang, Wei Tan, Hangting Chen, Jianwei Yu, Yixuan Zhang, Chenyu Yang, Haina Zhu, Shuai Wang, Zhiyong Wu, and Dong Yu. Levo: High-quality song generation with multi-preference alignment, 2025. URL `https://arxiv.org/abs/2506.07520`.

Museum für Naturkunde Berlin. Animal sound archive. `https://doi.org/10.15468/0bpalr`. Accessed via gbif.org 2023-05-09.

H.-O. Pörtner, R. J. Scholes, A. Arneth, D. K. A. Barnes, M. T. Burrows, S. E. Diamond, C. M. Duarte, W. Kiessling, P. Leadley, S. Managi, P. McElwee, G. Midgley, H. T. Ngo, D. Obura, U. Pascual, M. Sankaran, Y. J. Shin, and A. L. Val. Overcoming the coupled climate and biodiversity crises and their societal impacts. *Science*, 380(6642):eabl4881, 2023. doi: 10.1126/science.abl4881. URL `https://www.science.org/doi/abs/10.1126/science.abl4881`.

Lukas Rauch, Raphael Schwinger, Moritz Wirth, René Heinrich, Denis Huseljic, Marek Herde, Jonas Lange, Stefan Kahl, Bernhard Sick, Sven Tomforde, et al. Birdset: A large-scale dataset for audio classification in avian bioacoustics. In *The Thirteenth International Conference on Learning Representations*, 2025.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. In *The Thirteenth International Conference on Learning Representations*, 2025.

Christian Rutz, Michael Bronstein, Aza Raskin, Sonja C Vernes, Katherine Zacarian, and Damián E Blasi. Using machine learning to decode animal communication. *Science*, 381(6654):152–155, 2023.

Laela Sayigh, Mary Ann Daher, Julie Allen, Helen Gordon, Katherine Joyce, Claire Stuhlmann, and Peter Tyack. The Watkins marine mammal sound database: An online, freely accessible resource. *Proceedings of Meetings on Acoustics*, 27(1):040013, 2016. doi: 10.1121/2.0000358. URL `https://asa.scitation.org/doi/abs/10.1121/2.0000358`.

Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Abdelkader DEBBAH. How bad is training on synthetic data? a statistical analysis of language model collapse. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=t3z6UlV09o`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL `https://arxiv.org/abs/2408.03314`.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024.

Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

Alex Tamkin, Dat Nguyen, Salil Deshpande, Jesse Mu, and Noah Goodman. Active learning helps pretrained models learn the intended task, 2022. URL `https://arxiv.org/abs/2204.08491`.

Willem-Pier Vellinga and Robert Planqué. The xeno-canto collection and its relation to sound recognition and classification. *CLEF (Working Notes)*, 1391, 2015.

A Data creation pathway

Bioacoustic data can be aggregated from various sources with creative commons or similarly permissive licenses:

- Primary sources with excellent metadata are citizen science databases such as Xeno-canto and iNaturalist. A large portion of our datasets, and of other bioacoustic datasets such as BirdSet (Rauch et al. [2025]) and iNatSounds (Chasmai et al. [2025]) also derive their data from these primary sources. We standardized these data sources into a single format and added generated natural language prompts derived from the metadata.
- Soundscape recordings via passive acoustic monitoring are excellent resources for very long, high sampling-rate recordings of bioacoustics and environmental sounds. Datasets such as Barkley Canyon for marine mammals (Kanes [2021]) and Sapsucker Woods (Kahl et al. [2022]) for birds were included in NatureLM-audio-training. More sources such as Sanctsound (`https://sanctsound.ioos.us/`) and Orcasound (`https://registry.`

opendata.aws/orcasound/) are available for aggregation. These petabyte-scale datasets have a mix of annotated and unannotated recordings requiring significant preprocessing and standardization.

- Large research datasets from lab environments focusing on single species (such as MarmAudio for the common marmoset Lamothe et al. [2025]) with very fine-grained annotations of vocalization types, timing, etc., can be a very important addition to NatureLM-audio-training.
- General purpose, non-bioacoustic audio datasets such as FSD50K (Fonseca et al. [2021]), AudioCaps (Gemmeke et al. [2017]), Clotho (Drossos et al. [2020]) were added to NatureLM-audio-training. Larger datasets such as AudioSet (Gemmeke et al. [2017]) would be excellent additions to perform noise augmentations during training and to promote cross-domain transfer.
- Generative audio models (Lei et al. [2025]) have the ability to synthesize audio when conditioned on text or other audio. We could utilize these models to augment a dataset for rare taxa. Caution is needed when using synthetic data because too much synthetic data in the training mix can lead to degeneracy (Seddik et al. [2024])