Automated Annotation of Bioacoustic Soundscapes in the Wild

Anonymous Author(s)
Affiliation
Address
email

Abstract

Automated analysis of bioacoustic recordings is essential for monitoring biodiversity and ecosystem health, yet current methods struggle with the complexity of natural soundscapes and the scarcity of labeled data. We introduce a Bioacoustic Masked Autoencoder, a self-supervised framework designed to learn robust audio representations from large-scale, unlabeled recordings. Pretrained on over 15,000 hours of diverse terrestrial and marine audio, our model, a 1B-parameter Vision Transformer encoder paired with a 500M-parameter decoder, learns representations that generalize across species and habitats. When evaluated zero-shot on multiple bioacoustic benchmarks, our model outperforms state-of-the-art models in vocalization detection and species classification. We further demonstrate the benefits of combining supervised and unsupervised contrastive objectives for species-aware embeddings. Our contributions include (1) a large-scale unified dataset of bioacoustic recordings, (2) a pretrained foundation model for bioacoustic analysis, and (3) evidence that self-supervised learning enables scalable, label-efficient monitoring of global biodiversity. More results and visuals can be found at [LINK.

1 Introduction

Bioacoustic monitoring has emerged as a critical tool for ecological research, wildlife conservation, and biodiversity assessment Stowell et al. [2016], Marques et al. [2012]. By recording and analyzing animal vocalizations, researchers can track population dynamics, detect species presence, and monitor ecosystem health without physical intervention. However, the automated analysis of bioacoustic data presents significant challenges, particularly in natural environments where recordings contain diverse species, background noise, and complex acoustic events Stowell and Plumbley [2014], Mesaros et al. [2020].

In this paper, we introduce a specialized Masked Autoencoder for bioacoustic data that learns robust audio representations without reliance on labeled examples. Our approach builds upon recent advances in self-supervised audio representation learning, particularly Audio-MAE Huang et al. [2022], while incorporating several innovations tailored specifically to the challenges of bioacoustic analysis: We also leverage a diverse dataset of bioacoustic recordings spanning terrestrial and marine environments to ensure our model learns representations applicable across varied ecological contexts.

Our primary contributions are the following: **Model**. We release a 1 billion parameter Vision Transformer (ViT) encoder model trained heavily on bioacoustic data capable of handling long-sequences of audio. **Unified Dataset**. We release a collection of dozens of bioacoustic datasets with unified annotations. **New Benchmark Results**. We show that our model is able to achieve state-of-the-art results on bioacoustic benchmarks.

Submitted to the AI for Science workshop (NeurIPS 2025)

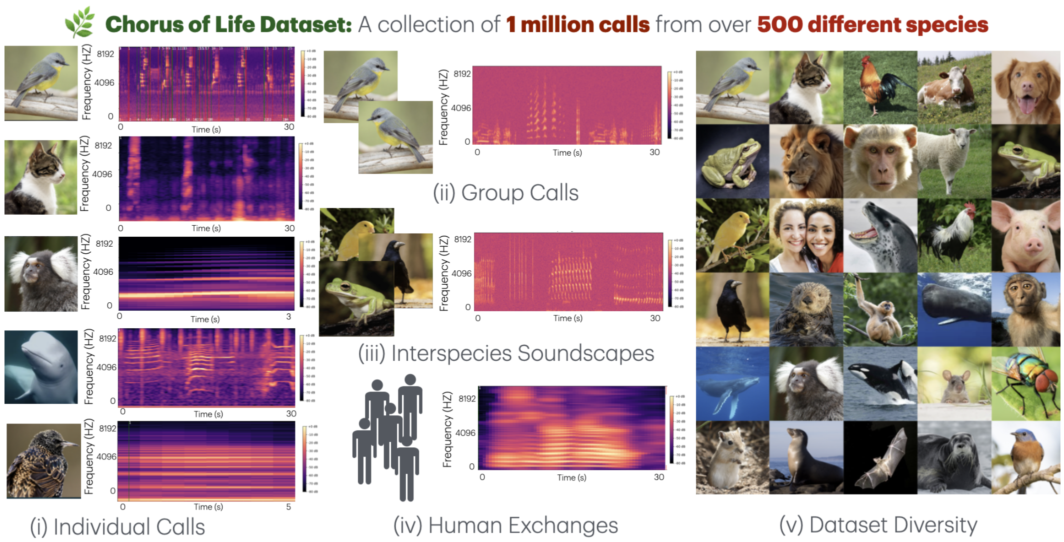

Figure 1: Overview of the Chorus of Life dataset. This dataset consists of over 1 million calls from 500+ species, showcasing a wide range of acoustic contexts. **(i)** Individual Calls: Examples of isolated vocalizations from different species, represented as spectrograms. **(ii)** Group Calls: Recordings of multiple individuals of the same species vocalizing together, highlighting overlapping patterns. **(iii)** Interspecies Soundscapes: Complex acoustic environments where calls from multiple species co-occur, mimicking real-world habitats. **(iv)** Human Exchanges: Speech interactions included in the dataset to support human-animal communication studies. **(v)** Dataset Diversity: Visual representation of species diversity, including birds, mammals, marine animals, and humans.

2 Related Work

The advent of foundation models has begun to transform the field. NatureLMaudio is the first foundational audio language model specifically designed for bioacoustics Robinson et al. [2024]. Unlike earlier approaches that were typically tailored to a single taxon or task, NatureLM-audio is trained on diverse text–audio pairs spanning bioacoustics, speech, and music. This broad training regime enables the model to generalize in a zero-shot manner to unseen species and novel tasks such as call-type prediction and individual counting—capabilities that are critical for conservation and ecological research.

Collectively, these studies illustrate an evolution from specialized, species-specific classifiers to general-purpose, large-scale models capable of cross-domain transfer. Our work builds on these advances by integrating the strengths of foundation model architectures. In doing so, we aim to provide a unified framework that can robustly detect, classify, and interpret animal vocalizations across a wide range of taxa and real-world conditions.

3 Dataset Creation

We curated a comprehensive pretraining dataset and also a large collection of labeled bioacoustic datasets by combining recordings from multiple bioacoustic sources covering diverse taxonomic groups and ecological environments. The primary datasets incorporated are included in Table 1. We curated more than 7000 hours of audio and over 1 million annotated calls across 30 genera and 500 species.

Data Augmentation In order to teach our model a more diverse set of bioacoustic data during both pre-training and fine-tuning, we leveraged a multitude of data augmentation strategies in order to increase both the size and diversity of our dataset: (i) `Mixing`: Audio is additive by nature and so we are able to easily add multiple calls together to simulate overlapping calls., (ii) `Stitching`: In order

Dataset	Num Calls	Duration	Dataset	Num Calls	Duration
AudioSet Gemmeke et al. [2017]	0	5,800 h	Giant Otters Mumm and Knörnschild [2014]	441	1 h
Animal Sounds Jayaya [2025]	809	1 h	Hainan Gibbons Dufourq et al. [2020]	1233	104 h
Anuraset Cañas et al. [2023]	16089	27 h	Hawaii Birds Navine et al. [2022]	59583	51 h
Bengal Finch Nicholson et al. [2017]	1215	5 h	HICEAS Yano et al. [2018]	796	13 h
BIRDeep Márquez-Rodríguez et al. [2024]	3749	9 h	Macaques Fukushima et al. [2015]	7285	1 h
BirdVox Lostanlen et al. [2018]	35402	18 h	Infant MarmosetsVox Sarkar and Magimai.-Doss [2023]	169318	59 h
Domestic Canaries Belzner et al. [2009]	14407	4 h	Multimodal Birds Kumar et al. [2023]	6524	4 h
Columbia/CR Álvaro Vega-Hidalgo et al. [2023]	7338	35 h	Northeast US Kahl et al. [2022a]	50760	285 h
DARPA dar	1718	5 h	Orca Sounds Internet Archive	398	1 h
Avian Dawn Weldy et al. [2024]	41183	132 h	Pig Sounds Briefer et al. [2022]	6887	1 h
DCASE	7206	18 h	Rainforest Yassin et al. [2020]	1216	21 h
Fruit Bats Prat et al. [2017]	90000	38 h	Rodent Sounds Tachibana [2019]	4576	1 h
ENA Birds Chronister et al. [2021]	16052	7 h	Sierra Nevada Clapp et al. [2023]	10976	17 h
ESC Piczak	400	1 h	Southwest Amazon Hopping et al. [2022]	16482	22 h
Rook Birds Martin et al. [2022]	21662	29 h	SSW Van Horn et al. [2022]	3861	11 h
			Watkins Marine Sounds Sayigh et al. [2016]	15152	30 h
			Western US Kahl et al. [2022b]	20147	33 h
			Sperm Whales	14764	250 h

Table 1: Curated datasets with the number of annotated calls (a single annotation of any length is an annotated call) and duration.

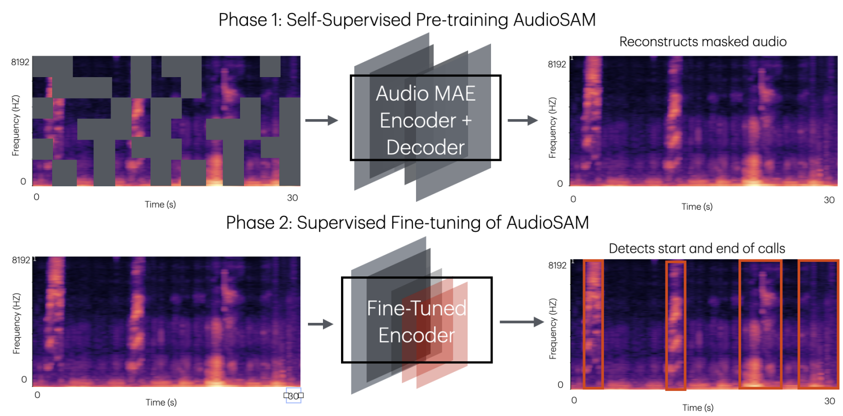

Figure 2: Two-phase training pipeline of our model. Our approach consists of **(i)** a self-supervised pre-training phase and **(ii)** a supervised fine-tuning phase. **(i)** Self-Supervised Pre-training: An Audio Masked Autoencoder (Audio MAE) is trained on spectrograms with randomly masked patches (left) to reconstruct the original audio (right). This step learns general acoustic representations without requiring labels. **(ii)** Supervised Fine-tuning: The pre-trained encoder is fine-tuned on labeled data to detect structural boundaries within calls (right), such as the start and end times of vocalizations.

to simulate long-term calls with short audio segments, we utilize audio stitching., (iii) `Amplitude Modulation`: We leverage changes in amplitude to simulate vocalizations being further or closer to a given audio source., (iv) `Noise addition and reduction`: This adds variety to training data and (v) `Varying FFT window`: Larger nFFT values provide a finer frequency resolution because more frequency bins are created. Since animals communicate with diverse frequency and temporal characteristics, it makes sense to vary the nFFT across training.

4 Pretraining

To learn robust audio representations without reliance on labeled data, we implement a Masked Autoencoder (MAE) pre training framework for bioacoustic data. Our approach is based on work on image pretraining and recent work on self-supervised learning for audio Huang et al. [2022], He et al. [2021].

Model Architecture The encoder processes only the visible (unmasked) portions of the input spectrogram, reducing computational requirements significantly during pretraining. We implement a transformer architecture with self-attention mechanisms that capture long-range dependencies in the

Model	dcase	enabirds	hiceas	rainforest	gibbons	esc	watkins
LLM w/o audio	0.000	0.001	0.210	0.000	0.013	0.020	0.041
SALMONN	0.005	0.004	0.097	0.002	0.005	0.320	0.041
BioLingual	0.036	0.109	**0.429**	0.004	0.018	0.307	0.041
NatureLM-audio	0.058	0.314	0.336	0.025	0.005	0.600	0.257
Our Model	**0.282**	**0.902**	0.304	**0.111**	**0.041**	**0.719**	**0.431**

Table 2: Zero-shot performance on multiple bioacoustic benchmarks. Columns *dcase*, *enabirds*, *hiceas*, *rainforest*, and *gibbons* report F1 scores for vocalization detection Robinson et al. [2024], while *esc* and *watkins* report accuracy on classification tasks. Best score per column is bolded.

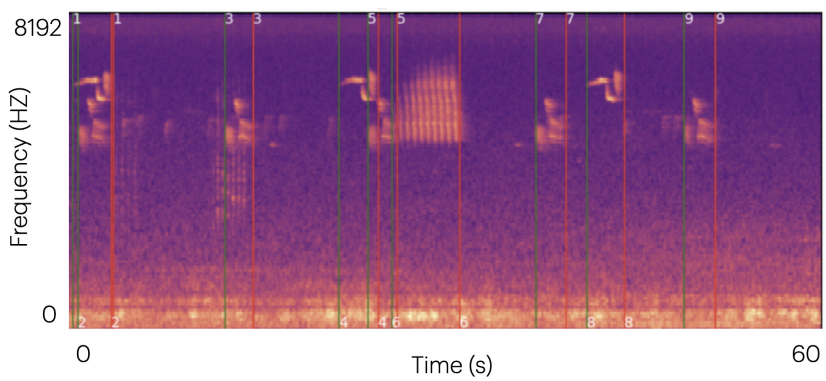

Figure 3: Model detection results on sample animal vocalization. Green bars indicate the start of a call and red bars indicate the end. Calls are also numbered.

audio signal. The decoder then reconstructs the full spectrogram, including the masked regions, from the encoded representations combined with positional embeddings.

Pre Training Dataset We pre-train our model on a diverse collection of audio recordings primarily from AudioSet Gemmeke et al. [2017], but also from synthetically generated datasets using augmentation techniques outlined in Section 3 comprising approximately 15,000 hours of unlabeled audio. The data set includes a wide range of acoustic environments, animal vocalizations, natural sounds, sounds of things, and more. This provides rich contextual variety to learn robust representations.

5 Model Evaluation

We trained our model on all the datasets mentioned in Table 1, excluding those included in the evaluation. The DCASE, Enabirds, HICEAS, Rainforest, and Hainan Gibbons datasets were withheld so that we can evaluate our model's zero-shot performance on those datasets.

Vocalization Detection We first load the pre-trained weights from the ViT and then attach a binary event detection head to the output embeddings from the ViT and then perform full fine-tuning. The results are shown in Table 2. We also provide an example of qualitative results in Figure 3.

6 Conclusion

This work introduces a scalable self-supervised framework for bioacoustic monitoring, enabling robust representation learning from complex and noisy soundscapes without reliance on manual annotation. By leveraging large-scale audio data and contrastive objectives, our approach significantly improves event classification and species identification performance across diverse ecosystems. Future directions include integrating multi-modal environmental signals and deploying lightweight models for real-time field applications, advancing automated biodiversity monitoring at global scale.

References

Dan Stowell, Mike Wood, Yannis Stylianou, and Herve Glotin. Bird detection in audio: A survey and a challenge, 2016.

Tiago Marques, Len Thomas, Stephen Martin, David Mellinger, Jessica Ward, David Moretti, Danielle Harris, and Peter Tyack. Estimating animal population density using passive acoustics. *Biological Reviews*, 88(2): 287–309, 2012.

Dan Stowell and Mark Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, page 488, 2014.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Bioacoustic challenges in dcase 2020, 2020. Dataset.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *NeurIPS*, 2022.

David Robinson, Marius Miron, Masato Hagiwara, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. *arXiv preprint arXiv:2411.07186*, 2024.

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.

Christina Mumm and Mirjam Knörnschild. The vocal repertoire of adult and neonate giant otters (pteronura brasiliensis). *PLoS ONE*, 9:e112562, 11 2014. doi: 10.1371/journal.pone.0112562.

Maulana Akbar Dwi Jayaya. Animal sounds dataset. Kaggle, 2025. Accessed on March 5, 2025.

Emmanuel Dufourq, James Hansford, Amanda Hoepfner, Heidi Ma, Jessica Bryant, Christina Stender, Wenyong Li, Zhiwei Liu, Qing Chen, Zhaoli Zhou, and Samuel Turvey. Automated detection of hainan gibbon calls for passive acoustic monitoring, 09 2020.

Juan Sebastián Cañas, Maria Paula Toro-Gómez, Larissa Sayuri Moreira Sugai, Hernán Darío Benítez Restrepo, Jorge Rudas, Breyner Posso Bautista, Luís Felipe Toledo, Simone Dena, Adão Henrique Rosa Domingos, Franco Leandro de Souza, Selvino Neckel-Oliveira, Anderson da Rosa, Vítor Carvalho-Rocha, José Vinícius Bernardy, José Luiz Massao Moreira Sugai, Carolina Emília dos Santos, Rogério Pereira Bastos, Diego Llusia, and Juan Sebastián Ulloa. Anuraset: A dataset for benchmarking neotropical anuran calls identification in passive acoustic monitoring, 2023. URL `https://arxiv.org/abs/2307.06860`.

Amanda Navine, Stefan Kahl, Ann Tanimoto-Johnson, Holger Klinck, and Patrick Hart. A collection of fully-annotated soundscape recordings from the island of hawai'i. *Dataset on Zenodo, September*, 2022.

David Nicholson, Jonah E. Queen, and Samuel J. Sober. Bengalese Finch song repository, 10 2017. URL `https://figshare.com/articles/dataset/Bengalese_Finch_song_repository/4805749`.

Kymberly M Yano, Erin Marie Oleson, Jennifer L Keating, Lisa Taylor Ballance, Marie Chapla Hill, Amanda L Bradford, Ann N Allen, Trevor W Joyce, Jeffrey E Moore, and Annette Elizabeth Henry. Cetacean and seabird data collected during the hawaiian islands cetacean and ecosystem assessment survey (hiceas), july–december 2017, 2018.

Alba Márquez-Rodríguez, Miguel Angel Mohedano-Muñoz, Manuel Jesus Marin-Jimenez, Eduardo Santamaria-Garcia, Giulia Bastianelli, Pedro Jordano, and Irene Mendoza. Birdeepaudioannotations (revision 4cf0456), 2024. URL `https://huggingface.co/datasets/GrunCrow/BIRDeep_AudioAnnotations`.

Makoto Fukushima, Alex M Doyle, Matthew P Mullarkey, Mortimer Mishkin, and Bruno B Averbeck. Distributed acoustic cues for caller identity in macaque vocalization. *Royal Society open science*, 2(12):150432, 2015.

Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Birdvox-full-night: a dataset and benchmark for avian flight call detection. In *Proc. IEEE ICASSP*, April 2018.

Eklavya Sarkar and Mathew Magimai.-Doss. Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers? In *Proc. INTERSPEECH 2023*, pages 1189–1193, 2023. doi: 10.21437/Interspeech.2023-1968.

Sandra Belzner, Cornelia Voigt, Clive K Catchpole, and Stefan Leitner. Song learning in domesticated canaries in a restricted acoustic environment. *Proceedings of the Royal Society B: Biological Sciences*, 2009.

Sumit Kumar, B. Anshuman, Linus Rüttimann, Richard H.R. Hahnloser, and Vipul Arora. Balanced deep cca for bird vocalization detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094650.

Álvaro Vega-Hidalgo, Stefan Kahl, Laurel B. Symes, Viviana Ruiz-Gutiérrez, Ingrid Molina-Mora, Fernando Cediel, Luis Sandoval, and Holger Klinck. A collection of fully-annotated soundscape recordings from neotropical coffee farms in colombia and costa rica, 2023.

Stefan Kahl, Russell Charif, and Holger Klinck. A collection of fully-annotated soundscape recordings from the northeastern united states. *Dataset on Zenodo, August 2022a. URL https://doi. org/10.5281/zenodo*, 7079380, 2022a.

The darpa timit acoustic-phonetic continuous speech corpus.

Internet Archive. Orcas classification.

Matthew Weldy, Tom Denton, Abram Fleishman, Jaclyn Tolchin, Matthew McKown, Robert Spaan, Zachary Ruff, Julianna Jenkins, Matthew Betts, and Damon Lesmeister. Audio tagging of avian dawn chorus recordings in california, oregon and washington., 2024.

EF Briefer, CCR Sypherd, LMC Leliveld, M Padilla de la Torre, and C Tallet. The soundwel database: a labeled pig vocalization repository [data set]. *Scientific Reports*, 12:3409, 2022.

Bourhan Yassin, inversion, Mahreen Qazi, and Zephyr Gold. Rainforest connection species audio detection. `https://kaggle.com/competitions/rfcx-species-audio-detection`, 2020. Kaggle.

Yosef Prat, Mor Taub, Ester Pratt, and Yossi Yovel. An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Scientific Data*, 4:sdata2017143, 10 2017. doi: 10.1038/sdata.2017.143.

Ryosuke O Tachibana. Dataset for usvseg performance test, 2019.

Lauren Chronister, Tessa Rhinehart, Aidan Place, and Justin Kitzes. An annotated set of audio recordings of eastern north american birds containing frequency, time, and species information. *Ecology*, 102, 05 2021. doi: 10.1002/ecy.3329.

Mary Clapp, Stefan Kahl, Erik Meyer, Megan McKenna, Holger Klinck, and Gail Patricelli. A collection of fully-annotated soundscape recordings from the southern sierra nevada mountain range. *Dataset on Zenodo, January*, 2023.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL `http://dl.acm.org/citation.cfm?doid=2733373.2806390`.

W Alexander Hopping, Stefan Kahl, and H Klink. A collection of fully-annotated soundscape recordings from the southwestern amazon basin. 1. *Zenodo. URL: https://doi. org/10.5281/zenodo*, 7079124, 2022.

Killian Martin, Olivier Adam, Nicolas Obin, and Valérie Dufour. Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks. *Ecological Informatics*, 72, 12 2022. doi: 10.1101/2022.02.19.481011.

Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. In *European Conference on Computer Vision (ECCV)*, 2022.

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. In *Proceedings of Meetings on Acoustics*, volume 27. AIP Publishing, 2016.

Stefan Kahl, Connor M Wood, Philip Chaon, M Zachariah Peery, and Holger Klinck. A collection of fully-annotated soundscape recordings from the western united states. *Dataset on Zenodo, September 2022c. URL https://doi. org/10.5281/zenodo*, 7050014, 2022b.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.