Retrieval-Augmented Bioacoustics: Evidence-Guided Generation for Animal Communication

Bodhisatta Maiti
Independent Researcher
Orlando, Florida, United States
bodhisatta.iitbhu@gmail.com

Abstract

Animal vocalizations carry important information about communication, context, and behavior, but most current AI approaches in bioacoustics focus on narrow tasks such as species classification or call detection. A gap remains in methods that can help researchers interpret and summarize acoustic data in a grounded and transparent way. This proposal introduces Retrieval-Augmented Bioacoustics (RAB), a framework that combines acoustic embeddings with retrieval from call libraries and generative modeling. Retrieval provides concrete evidence, while generation produces outputs such as annotation suggestions, monitoring summaries, cross-species communication hypotheses, and prototype call synthesis. Two design choices strengthen the framework: adapting the number of retrieved neighbors depending on signal quality, and citing retrieved calls directly in generated outputs to increase transparency. RAB offers a model-agnostic approach that can be applied on top of existing or future embedding models, with potential impact on both ethological research and conservation applications.

1 Introduction

Animal communication has long been a central theme in understanding biodiversity, behavior, and ecology. Vocalizations in particular carry rich information about social interactions, environmental context, and survival strategies. In recent years, large amounts of bioacoustic data have become available through passive acoustic monitoring systems, public repositories, and conservation studies [1,2]. At the same time, advances in artificial intelligence have created new opportunities for analyzing such data. Self-supervised audio encoders, multimodal models, and large language models have shown strong performance in speech, music, and general audio understanding [3]. These developments suggest the possibility of applying similar techniques to non-human animal communication at scale. Despite this progress, most AI methods in bioacoustics remain focused on narrow prediction tasks. Common applications include species classification, detection of vocalizations in long recordings, or call-type identification [4]. While these approaches provide useful signals, they do not directly support interpretation, summarization, or hypothesis generation for researchers. In addition, current systems often act as black boxes, producing outputs without clear evidence or transparency for domain scientists. To address this gap, this proposal introduces Retrieval-Augmented Bioacoustics (RAB), a framework designed to combine retrieval from call libraries with generative modeling. In this setup, retrieval provides acoustic evidence in the form of nearest neighbors or cluster exemplars, while generation produces outputs that are more directly useful to ethologists. Example applications include annotation assistance, monitoring summaries, cross-species communication hypotheses, and prototype call synthesis. The framework is model-agnostic and can operate with embeddings derived from existing encoders such as wav2vec2, BEATs, or AudioMAE [5], as well as future foundation models. Two design choices strengthen the framework. First, retrieval depth can adapt to the quality of the signal or the confidence of the embedding model, ensuring robustness in noisy or uncertain

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

settings. Second, generated outputs can cite the retrieved calls that informed them, providing transparency and traceability for experts. These features make RAB distinct from simple classification or detection pipelines. While the proposal emphasizes new opportunities that arise when retrieval and generation are combined, it is important to recognize that many routine tasks in bioacoustics are already well served by traditional detection pipelines and metadata-based aggregation. The RAB framework is therefore intended as a complementary layer rather than a replacement, particularly in settings where researchers benefit from seeing concrete examples, resolving ambiguous segments, or exploring patterns that extend across datasets. In addition, the generative component should be interpreted within the constraints of current models: most audio–language systems handle non-human vocalizations inconsistently, and multi-example conditioning remains limited. These considerations help scope the framework toward applications where evidence-guided interpretation is more valuable than direct automation. Overall, RAB represents a forward-looking direction for AI in bioacoustics. By combining retrieval and generation, it highlights how AI systems can move from black-box predictions toward evidence-guided outputs that support interpretation and collaboration with domain scientists.

2 Proposed Framework: Retrieval-Augmented Bioacoustics (RAB)

The Retrieval-Augmented Bioacoustics (RAB) framework is designed to connect existing advances in acoustic representation learning with generative modeling in order to support interpretation of animal vocalizations. The framework is organized in two main stages: retrieval and generation. Retrieval stage: Bioacoustic signals are first represented using embeddings learned from models such as self-supervised audio encoders or multimodal audio–language systems [6]. These embeddings are indexed in a library of calls, which may include annotated datasets, species-specific repositories, or recordings collected from passive monitoring devices. When a new query vocalization is presented, the system retrieves the most similar calls from the library, providing a set of candidate examples and associated metadata. Generation stage: The retrieved calls act as evidence for the generative component. Depending on the task, the generative module may produce textual explanations, ecological summaries, or even synthetic audio exemplars. For example, generation can provide a rationale for a suggested label, produce a short report summarizing activity within a time window, or create a prototype call representing a cluster of vocalizations. At the same time, the design of this stage must reflect the practical constraints of current audio–language models. Most domain-specific systems accept only a single audio segment at a time, and general-purpose models vary in how reliably they capture non-human acoustic structure. As a result, generation in RAB may rely on summaries grounded in spectrogram snippets, metadata associated with retrieved calls, or lightweight fine-tuning on domain text rather than direct multi-audio conditioning. This keeps the generative component aligned with information that can be reliably extracted from the retrieved neighbors. Design choices: Two simple design elements strengthen this workflow. First, retrieval depth can adapt to the quality of the input signal. In clean, high-confidence conditions, a small set of neighbors may be sufficient, while noisy or uncertain inputs may require a larger retrieval pool. This adaptive-k strategy ensures that retrieval remains robust under varied recording environments. In this context, adaptive-k is used purely as a practical engineering choice to make the workflow more flexible, rather than as a methodological contribution of its own. Second, generated outputs can explicitly cite the retrieved calls that influenced them. Including identifiers, timestamps, or spectrogram snippets provides transparency for domain experts and allows verification of the evidence behind each output. Overall, the RAB framework is model-agnostic and can operate on top of existing embedding methods while remaining flexible enough to incorporate future foundation models in bioacoustics. By coupling retrieval with generation, it shifts the focus from narrow prediction tasks toward evidence-guided outputs that are more interpretable for ethologists and conservation researchers.

3 Applications of RAB

3.1 Annotation Assistance

Annotating large bioacoustic datasets is a major bottleneck in animal communication research. A single recording campaign can produce thousands of vocalizations that require labeling, often at the level of species, call type, or behavioral context [7]. Manual annotation by experts is slow and resource-intensive, and automated classifiers, while helpful, typically act as black boxes that provide

little justification for their outputs [8]. This lack of transparency reduces trust and makes it harder for experts to adopt automated tools in their workflow. The RAB framework provides a way to accelerate annotation while maintaining interpretability. A new vocalization is embedded and compared to a library of previously labeled calls. Similar examples are retrieved, and the generative module produces a suggestion that links the query to the retrieved evidence. For instance, a vocalization might be described as resembling known alarm calls of a given species, with the relevant retrieved calls displayed alongside the suggestion. Two design features are especially important here. Adaptive-k retrieval allows the number of neighbors to expand when the input signal is noisy, while focusing on fewer examples when confidence is high. Retrieval attribution ensures that each generated suggestion cites the retrieved calls that informed it, using identifiers, timestamps, or spectrogram snippets. These features transform the system from a black-box classifier into a semi-automated assistant, reducing annotation workload while giving experts the ability to verify and refine each decision. In many annotation scenarios, the primary benefit comes from quickly reviewing similar examples rather than from fully generated explanations. For this reason, the generative component should be viewed as optional and used mainly to highlight distinguishing features or summarize retrieved context, allowing experts to retain control over final decisions.

3.2 Summarization of Passive Acoustic Monitoring

Passive acoustic monitoring has become a common approach in ecology and conservation, producing long, continuous recordings of natural environments [9]. These recordings can span weeks or months, capturing thousands of vocalizations across many species. While this data is valuable, it is difficult for researchers to process manually, and even detection-based pipelines often return overwhelming volumes of output that still require interpretation [10]. The RAB framework can be applied to generate higher-level summaries of passive monitoring data. Individual calls are first detected and embedded, then indexed for retrieval. Retrieval organizes the recordings by time, location, or species, and the generative module produces concise summaries. For example, the system might produce a daily report describing the frequency of whale calls in a marine dataset or highlight changes in bird vocal activity compared to previous weeks. Two aspects of the design are especially helpful in this context. Adaptive-k retrieval allows faint or uncertain calls to be included in the retrieval pool, ensuring that low-amplitude but ecologically important signals are not overlooked. Retrieval attribution ensures that each generated summary links back to the retrieved calls, so researchers can verify the evidence underlying reported patterns. Together, these features shift monitoring analysis from raw detections toward structured, evidence-guided ecological reporting. For large, continuous recordings, conventional detection and simple aggregations often remain the most efficient tools for producing daily or weekly summaries. RAB becomes more useful when recordings contain overlapping species, ambiguous low-amplitude signals, or segments that are difficult to separate through automated chunking alone. In such settings, retrieval helps surface comparable events, and generation can condense their shared patterns into short, interpretable notes.

3.3 Cross-Species Hypothesis Generation

Comparative research in animal communication often seeks to identify functional similarities in vocalizations across species. For example, alarm calls in birds and mammals can share temporal or acoustic structures even when the species are not closely related [11]. Such comparisons can provide insights into convergent evolution, communication strategies, and the ecological role of signals. However, systematically identifying these parallels is difficult, as it requires searching across large, diverse datasets and drawing connections that are not obvious from raw recordings. The RAB framework can support hypothesis generation in this area by linking retrieval with generative explanation. A query call is embedded and compared against a cross-species library of vocalizations. Retrieved calls may come from entirely different taxa but exhibit acoustic similarity. The generative component then produces a hypothesis, such as noting that a dolphin whistle shares structural properties with known primate alarm calls. Retrieval attribution is especially important here. Each hypothesis can cite the specific calls that were retrieved, along with metadata such as species, location, or context. This allows researchers to evaluate the plausibility of the suggestion rather than treating the output as a definitive claim. In this way, RAB acts as a hypothesis-support tool, surfacing candidate parallels for expert review and further study, rather than attempting to replace scientific interpretation. Here, generation is not intended to furnish definitive interpretations but to translate retrieved acoustic parallels into a concise hypothesis that can be examined by experts. The value of

this step depends entirely on the strength of retrieval and the visibility of the evidence that supports each suggestion.

3.4 Prototype Call Synthesis

Bioacoustic datasets often contain clusters of calls that are acoustically similar but vary in quality due to background noise, distance from the recorder, or overlapping signals [12]. For researchers, listening to dozens of near-identical recordings can be inefficient, especially when the goal is to characterize a call type or compare across contexts. Having a single representative "prototype" for a cluster can make data exploration and labeling more manageable. Within the RAB framework, prototype generation can be achieved by combining retrieval with generative modeling. Similar calls are grouped into clusters during the retrieval stage. Instead of presenting each call individually, the generative module can produce a synthetic exemplar that captures the shared features of the cluster. This prototype might be rendered as a spectrogram or as a short audio token that approximates the cluster average. Although generative audio models have shown progress in domains such as speech and music [13], their application to non-human communication is still limited. RAB extends this direction by tying prototype synthesis to retrieval, ensuring that the generated call is anchored to real examples rather than produced freely. For researchers, this approach could speed up the process of cataloging call types and reduce the burden of navigating large, noisy datasets. In many cases, identifying a representative call directly from the retrieved neighbors may be sufficient, especially when clusters are small or relatively clean. Prototype synthesis becomes useful only when researchers want a noise-reduced exemplar that captures structure shared across multiple recordings. Any synthesized prototype should be treated as an interpretive aid and validated against the underlying calls.

3.5 Ethologist-in-the-Loop Interfaces

While automated methods can process large volumes of audio efficiently, expert input remains critical in bioacoustics research. Human interpretation is often required to validate call types, contextual meanings, or ecological significance [14]. Systems that provide results without transparency risk being underutilized, as researchers may hesitate to rely on outputs they cannot verify. The RAB framework can be integrated into interactive interfaces that keep ethologists in control. In this setup, retrieval presents a set of relevant calls and associated metadata, while the generative component offers a preliminary explanation or summary. The expert can then accept, reject, or adjust the output, with corrections feeding back into the system for future refinement. For example, a researcher might be shown retrieved calls labeled as possible mating signals, along with a generated summary, and then confirm whether this interpretation holds in context. Both adaptive-k retrieval and retrieval attribution are valuable in such interfaces. Adjusting the retrieval depth ensures that experts see a sufficient variety of candidates even when inputs are noisy, while explicit attribution shows exactly which calls influenced each output. This interactive workflow combines the scalability of AI with the interpretive expertise of domain scientists, supporting collaboration rather than replacement. These interfaces are most effective when they foreground the retrieved evidence and treat generative outputs as tentative suggestions. Expert review remains central, and the system's role is to organize information in a way that accelerates, rather than replaces, human interpretation.

4 Limitations and Ethical Considerations

The Retrieval-Augmented Bioacoustics framework is at this stage a conceptual design rather than a fully implemented system. Several limitations must be acknowledged. First, the quality of results depends heavily on the availability of representative call libraries. Many species remain under-sampled, and datasets are often biased toward birds or other well-studied taxa [15]. This imbalance may reduce the generality of retrieval-based approaches. Second, generative models can introduce artifacts or produce misleading outputs if not carefully constrained, especially in the case of synthetic prototype calls. Such limitations highlight the importance of using RAB as a support tool rather than as a definitive source of scientific claims. Ethical considerations are equally important in the application of AI to animal communication. Passive monitoring data that reveals the presence of rare or endangered species can be misused, for example by facilitating poaching or habitat disturbance [16]. The ability to generate summaries or hypotheses must therefore be paired with safeguards,

such as limiting access to sensitive geographic metadata. Another concern is anthropomorphism. Even when assisted by retrieval or generative tools, interpretations of animal vocalizations should be treated cautiously, since inferred meanings can easily be overstated without clear supporting evidence [17]. Ensuring that retrieval attribution is always included in outputs can help mitigate this risk by keeping the evidence visible to researchers. Several elements of the framework should be regarded as exploratory. Tasks such as prototype synthesis or long-form summarization are intended to illustrate how retrieval and generation might interact, not to replace established detection-and-aggregation pipelines. The generative stage in particular requires careful validation, as current models vary in their handling of non-human audio and may need constrained prompting or fine-tuning to operate reliably. These constraints reinforce the importance of expert oversight and highlight that RAB is best positioned as an interpretive tool rather than a fully automated system. Finally, RAB assumes collaboration between AI systems and domain experts. Without expert oversight, there is a risk that generated outputs could be misinterpreted or used outside their intended context. Maintaining an ethologist-in-the-loop approach and clear communication of limitations is essential for responsible deployment.

5 Conclusion

This proposal introduced Retrieval-Augmented Bioacoustics (RAB), a framework that combines acoustic retrieval with generative modeling to support the study of animal communication. The framework emphasizes evidence-guided outputs, making use of adaptive-k retrieval to handle noisy conditions and retrieval attribution to ensure transparency. Five applications were outlined: annotation assistance, summarization of passive acoustic monitoring, cross-species hypothesis generation, prototype call synthesis, and ethologist-in-the-loop interfaces. Together, these examples highlight how retrieval and generation can extend beyond classification tasks to provide interpretive and transparent support for researchers. Although RAB is currently conceptual, it builds on recent progress in self-supervised audio representation learning [3], which provides a foundation for evidence-guided retrieval workflows. Future work will require implementation, collaboration with domain experts, and evaluation across multiple taxa. If developed responsibly, the framework has the potential to accelerate ecological research, support conservation, and broaden understanding of non-human animal communication.

References

[1] Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. (2021) BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics* **61**:101236.

[2] Kahl, S., Stöter, F.-R., Goëau, H., Glotin, H., Planqué, R., Vellinga, W.-P. & Joly, A. (2019) Overview of BirdCLEF 2019: Large-scale bird recognition in soundscapes. In *Proceedings of the CLEF 2019 Conference and Labs of the Evaluation Forum*.

[3] Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33*, pp. 12449–12460.

[4] Kahl, S., Denton, T., Klinck, H., Glotin, H., Goëau, H., Vellinga, W.-P., Planqué, R. & Joly, A. (2021) Overview of BirdCLEF 2021: Bird call identification in soundscape recordings. In *Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Vol. 2936, pp. 1437–1450.

[5] Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z. & Wei, F. (2022) BEATs: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

[6] Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F. & Feichtenhofer, C. (2022) Masked Autoencoders that Listen. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

[7] Vellinga, W.-P. & Planqué, R. (2015) The Xeno-canto collection and its relation to sound recognition and classification. In *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum*, CEUR-WS.org, Vol. 1391.

[8] Stowell, D., Wood, M., Stylianou, Y. & Glotin, H. (2016) Bird detection in audio: A survey and a challenge. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2016)*, pp. 1–6. IEEE.

[9] Marques, T.A., Thomas, L., Martin, S.W., Mellinger, D.K., Ward, J.A., Moretti, D.J., Harris, D. & Tyack, P.L. (2013) Estimating animal population density using passive acoustics. *Biological Reviews* **88**(2):287–309.

[10] Stowell, D., Petrusková, T., Šálek, M. & Linhart, P. (2019) Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *Journal of the Royal Society Interface* **16**(153):20180940.

[11] Blumstein, D.T. (2007) The evolution, function, and meaning of marmot alarm communication. *Advances in the Study of Behavior* **37**:371–401.

[12] Priyadarshani, N., Marsland, S. & Castro, I. (2018) Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology* **49**(5): jav-01447.

[13] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. (2016) WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

[14] Sueur, J. & Farina, A. (2015) Ecoacoustics: the ecological investigation and interpretation of environmental sound. *Biosemiotics* **8**:493–502.

[15] Stowell, D. (2022) Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**:e13152.

[16] Krause, B. & Farina, A. (2016) Using ecoacoustic methods to survey the impacts of climate change on biodiversity. *Biological Conservation* **195**:245–254.

[17] Janik, V.M. & Slater, P.J.B. (1997) Vocal learning in mammals. *Advances in the Study of Behavior* **26**:59–99.