An Expert-in-the-Loop Toolbox for Explainable AI in Animal Communication

Álvaro Vega-Hidalgo[1], **Artem Abzaliev**[1], **Nicole Guisneuf**[2], **Thore Bergman**[2,3], **Rada Mihalcea**[1]
[1]Computer Science and Engineering, University of Michigan
[2]Department of Psychology, University of Michigan
[3]Department of Ecology and Evolutionary Biology, University of Michigan
{alvarovh, abzaliev, nickigui, thore, mihalcea}@umich.edu

Abstract

Explainable AI (XAI) remains underdeveloped in bioacoustics, despite the growing reliance on high-performance black-box models. We evaluate the explainability of state-of-the-art models for capuchin monkey individual identification and introduce new methods to make bioacoustic classifiers more interpretable. Our approach combines participatory evaluation with domain experts through a web-based interface, with quantitative metrics that assess correspondence between saliency maps and expert annotations. Specifically, we report metrics on ranking quality, spatial overlap and distributional similarity. Each metric is computed under two complementary formulations of feature importance. To facilitate annotation, we introduce a web interface for pixel-level spectrogram labeling that provides interactive foreground and background audio playback, allowing experts to listen separately to masked regions, along with optional AI-assisted segmentation. These tools provide a reproducible framework for benchmarking explainability in bioacoustic models, advancing toward more transparent, collaborative, and biologically meaningful AI for animal communication.

1 Introduction

Motivation. There is no established toolbox for evaluating how bioacoustic models attend to meaningful spectro-temporal features at the pixel level, nor a participatory framework that allows domain experts to validate these models through time–frequency spectrogram annotations. This gap hinders scientific discovery and cross-disciplinary trust.

In order to dig deeper into the unknown semantic spaces of other species, AI models need to be carefully designed with appropriate architectures that enable effective domain expert input. Nevertheless, these two priorities conflict with each other as the trend is to develop increasingly complex models [18] for which internal representations are not interpretable by design but represent the state of the art in classification performance [22]. In bioacoustics, explainability has been addressed only sparingly; recent XAI studies have explored custom models [21], while most researchers continue to rely on black-box pretrained models [3] due to their powerful transfer learning capabilities. These models achieve high performance, but their learned representations are difficult to interpret, limiting biological insight, responsible use, and trust from cross-domain expertise.

Capuchin monkeys produce over 27 distinct call types and exhibit cultural evolution and complex social cognition, making them an excellent model for studying animal communication [7]. Recent advances in joint cross-species embedding models have unlocked superior classification performance of caller identity for this species, offering new opportunities for remote monitoring and analysis [23]. However, to uncover which features of their rich vocal repertoire convey individual identity, new

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

methodological approaches are needed. XAI provides one such avenue, motivating this study and offering a strong test case for advancing interpretable methods in bioacoustics. While recent work has emphasized system scalability [4] and the use of large language models for cross-modal representation learning [18], little attention has been given to model explainability. Our work addresses this gap by introducing an XAI toolbox and participatory framework to interpret black-box bioacoustic models in collaboration with domain experts.

Our work contributes (1) an XAI toolbox for bioacoustic models, combining spectrogram feature importance maps with simple, interpretable saliency evaluation metrics adapted from computer vision to compare model attention against expert annotations, and (2) a web-based annotation platform that enables pixel-level denoising and participatory validation of model attention.

2 Background and Related Work

Explainable AI. Explainability can play an invaluable role in scientific exploration by identifying and refining target phenomena, motivating hypotheses and guiding inquiry [24]. On the other hand, explainability has become a central concern in AI as models grow in size and complexity, and society increasingly questions the consequences of their inner workings, with XAI techniques as a deciding factor for user trust and adoption [20]. In animal communication studies, ground-truth labels are usually tied to observed behavioral states and contexts hypothesized to motivate specific signals, and statistical models have long been used to test such hypotheses about semantics and linguistics phenomena. As the field shifts from using simple statistical descriptions to complex AI models, explainability becomes crucial for it to situate algorithmic insights within the rich ecological and evolutionary context of biological signals.

Compared to language and vision, interpretability in audio models, and especially in bioacoustics, is far less developed. Most work in audio has focused on acoustic event detection [13, 17, 10]. However, for bioacoustic applications explainability remains scarce. Models often operate as black-box classifiers, offering little insight into what acoustic features drive decisions. This lack of interpretability not only limits scientific understanding of animal communication but also hampers trust in deployed systems for conservation and ecological monitoring. Addressing this gap requires adapting or developing interpretability frameworks that are sensitive to the unique structure and semantics of acoustic signals in biological contexts.

More recently, interpretability has started to gain traction. Heinrich et al. [9] proposed incorporating interpretability directly into the model architecture, demonstrating how a network can learn prototypical patterns for bird species. In contrast, Silva et al. [21] focus on post-hoc analysis of trained models, using SHAP to interpret learned features. Our work follows this post-hoc perspective, as we find it more practical to study interpretability after the model has already been trained.

Participatory Design and Human–AI Collaboration. Explainability is not only a technical concern but also a design principle for effective human-AI collaboration. According to established guidelines [2], systems should support transparency, provide rationales, and enable meaningful human control. In wildlife monitoring, where technologies often interact with communities in overseas territories, justice-oriented design principles are equally important to strengthen governance, community agency, and cultural appropriateness [15, 16].

3 Approach

Multi-Grid Spectrogram Occlusion. We generate *saliency maps* through a *multi-grid spectrogram occlusion* procedure that systematically perturbs localized time-frequency regions to reveal the spectro-temporal patterns most influential for model predictions.[1]

[1] Throughout this paper, we use the term *occlusion* in the sense established by explainable AI research [25], referring to the systematic perturbation of localized input regions to estimate their influence on model predictions. In the audio domain, this operation is implemented as *masking* (e.g., silencing or replacing time–frequency regions with noise), and both terms are used interchangeably. Saliency maps, introduced later in the paper, are obtained through this occlusion process and visualized as spectrogram heatmaps [19, 12, 6]. Regions of high saliency are interpreted as *feature importance*, i.e., spectro-temporal components most critical for model predictions.

The method systematically masks local spectro-temporal regions of the input and measures the change in model confidence. For a waveform x sampled at 48 kHz, we compute its spectrogram and partition it into grids with fixed cell sizes of 75 ms in time (t_w) and 3 kHz in frequency (f_w). To avoid biasing explanations to a single grid origin, we generate multiple translated grids by shifting the partition along time and frequency (by Δt, Δf). This increases effective resolution when aggregating results, similar to adaptive strategies in other domains [5].

Each perturbed input $\tilde{x}_i$ is produced by occluding a single cell. In our implementation, occlusion is applied directly in the time domain: the band-limited signal corresponding to the selected time-frequency window is extracted using zero-phase forward-reverse filtering to avoid phase distortion [8], tapered with short Tukey ramps to smooth boundaries and prevent spectral artifacts [1], and set to silence. This approach ensures that only the target region is modified while the remainder of the waveform remains undistorted.

The trained classifier for acoustic individual identification is then applied to perturbed inputs. For each occlusion we obtain class probabilities $p(y \mid \tilde{x}_i)$, to be compared with the unperturbed prediction $p(y \mid x)$. Feature importance is quantified in two complementary ways: (i) *distributional change* via Jensen–Shannon divergence (JS Div) [14] between the two predictive distributions, and (ii) *label-specific change* via the difference in cross-entropy with respect to the true label (ΔCE). Aggregating these values across all grids yields a prediction-drop heatmap mapped to the original spectrogram, providing a saliency map with higher resolution.

While our experiments employ silence-based occlusion, the procedure is fully parameterizable and readily extensible to alternative masking strategies (e.g., pink or band-limited noise), as well as to different window sizes and grid translation steps, depending on the requirements of other tasks or domains.

Web-Based Annotation Toolbox. Pixel-level annotation is well established in computer vision, but the analogous task of segmenting spectrograms into time and frequency bins remains largely absent in bioacoustics. Our lightweight web interface, built on Flutter and Firebase, sequentially serves spectrogram–audio pairs to annotators. Users can draw masks manually or provide foreground and background points that trigger AI-assisted segmentation with Meta's Segment Anything (SAM) [11], refining suggestions with tools such as eraser, brush, and opacity controls. To aid validation, the interface also supports playback of masked foreground and background audio at variable speeds (e.g., 0.3× for capuchin calls). Each completed mask is stored as a binary map with metadata for subsequent evaluation. Although originally designed for explainability, the expert-drawn masks can also serve as byproducts for audio denoising or source separation, providing an additional practical benefit. The design prioritizes ease of use and remote collaboration, consistent with principles of human–AI interaction and participatory bioacoustics [2, 16]. The source code will be released upon publication.

Evaluation Metrics. We evaluate the correspondence between saliency maps and expert annotations using three complementary metric families: ranking-based, overlap-based, and correlation-based measures. For ranking quality, we report the Area Under the Precision–Recall Curve (AUPRC), which assesses how well the saliency map ranks expert-annotated pixels above non-annotated ones. For spatial overlap, we threshold saliency maps at 0.2 and compute Intersection-over-Union (IoU) and Coverage, capturing how much of the annotated region is recovered and how precisely it is localized. Finally, for distributional similarity, we compute Pearson correlation between the continuous saliency map and the binary expert mask [12]. All metrics are evaluated under both importance formulations: JS Div and ΔCE.

4 Results

Qualitative: spectrogram saliency heatmaps. Vocal production in primates arises from mechanical and physiological processes that generate distinctive acoustic patterns, enabling individual recognition. AI explainability techniques should be capable of revealing this information by highlighting salient spectro-temporal regions. As shown in Fig. 1, the Whisper–Perch MRMR joint embedding model isolates specific portions of the spectrogram that may be informative, even when their biological relevance is not yet established. Expert annotations on frequency–time bins (spectrogram pixels) can highlight known salient features such as the call itself (manual source separation).

While essential, these annotations capture only what humans can interpret from the calls, whereas models, either individually or through combinations such as MRMR joint embeddings [23], can surface complementary perspectives. In turn, these saliency maps can themselves become analyzable objects, supporting statistical methods to test hypotheses about which acoustic features may carry semantic or individual identity cues.

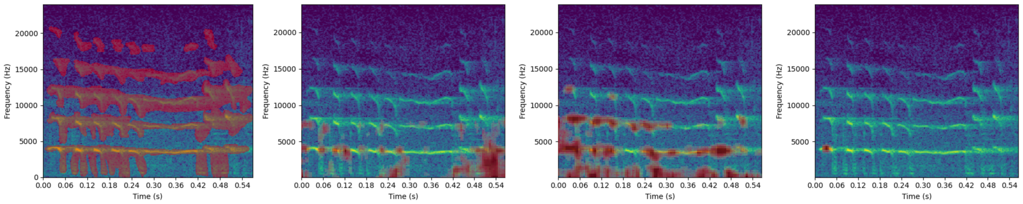

Figure 1: Qualitative evaluation of feature importance for acoustic individual identification in capuchin monkeys using saliency maps (ΔCE). Heatmaps (overlaid on spectrograms) highlight spectro-temporal regions that most influence model decisions about caller identity. From left to right: Annotated Mask, Whisper, Perch, and Whisper–Perch MRMR joint embedding. Note the variation in saliency across models, and how the Whisper–Perch MRMR joint embedding plot could motivate hypothesis testing if the pattern is consistent.

Model	**Importance**	**AUPRC**	**IoU$_{0.2}$**	**Coverage$_{0.2}$**	**Pearson Corr.**
Whisper–Perch MRMR joint embedding	JS Div	0.576 ± 0.128	0.013 ± 0.011	0.014 ± 0.011	0.214 ± 0.082
	ΔCE	0.469 ± 0.145	0.028 ± 0.018	0.029 ± 0.019	0.203 ± 0.084
Google Perch 2	JS Div	$\mathbf{0.605} \pm 0.140$	$\mathbf{0.103} \pm 0.038$	$\mathbf{0.110} \pm 0.045$	$\mathbf{0.397} \pm 0.097$
	ΔCE	0.571 ± 0.135	0.050 ± 0.025	0.052 ± 0.026	0.325 ± 0.101
Whisper Large V3	JS Div	0.372 ± 0.172	0.033 ± 0.020	0.037 ± 0.022	0.055 ± 0.122
	ΔCE	0.386 ± 0.161	0.030 ± 0.020	0.032 ± 0.021	0.055 ± 0.102
Baseline	JS Div	0.347 ± 0.148	0.026 ± 0.011	0.029 ± 0.011	0.001 ± 0.047
	ΔCE	0.352 ± 0.148	0.017 ± 0.010	0.018 ± 0.011	0.000 ± 0.047

Table 1: Performance comparison of model architectures across expert correspondence evaluation metrics using different feature importance methods.

Quantitative: expert annotation correspondence. Across all metrics, Perch corresponds most closely with expert annotations, achieving the highest scores in ranking, spatial overlap, and correlation (Table 1). Compared to Whisper, it improves overlap by about threefold and correlation by about sevenfold. Between importance formulations, JS Div generally outperforms ΔCE, with Perch showing the clearest advantage. Although the Whisper–Perch MRMR joint embedding achieves the best individual classification accuracy [23], it lags behind Perch in explainability. These results suggest that models can excel at classification while still diverging from human-recognizable cues, highlighting the importance of explainability as a complementary evaluation dimension.

5 Conclusion

This work introduces a comprehensive framework for evaluating and visualizing model interpretability in animal communication, combining expert–in–the–loop validation with quantitative XAI metrics. By linking saliency maps to expert annotations, our approach transforms explainability tools into instruments for scientific discovery of biologically meaningful cues. Beyond benchmarking model transparency, this framework enables new forms of collaboration between AI researchers and field experts, advancing toward interpretable and ethically grounded bioacoustics. Although originally designed for explainability, the expert-drawn masks produced through this process can also serve as byproducts for audio denoising or source separation, providing an additional practical benefit.

Limitations and Future Work. While the reported metrics currently measure agreement with human intuition rather than absolute ground truth, they provide a valuable bridge between model behavior and expert interpretation. Future work will integrate interpretability directly into model architectures, broaden validation across species and ecological contexts, and apply statistical analyses of saliency maps to test hypotheses about signal meaning. Expanding participatory annotation to more diverse user communities, exploring the role of explainable models in conservation decision-making, and conducting human–computer interaction studies to understand how users envision and engage with XAI tools remain exciting next steps.

References

[1] J.B. Allen and L.R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977. doi: 10.1109/PROC.1977.10770.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[3] Jules Cauzinille, Benoit Favre, Ricard Marxer, and Arnaud Rey. Applying machine learning to primate bioacoustics: Review and perspectives. *American Journal of Primatology*, 86(10): e23666, 2024.

[4] Vincent Dumoulin, Otilia Stretcu, Jenny Hamer, Lauren Harrell, Rob Laber, Hugo Larochelle, Bart van Merriënboer, Amanda Navine, Patrick Hart, Ben Williams, et al. The search for squawk: Agile modeling in bioacoustics. *arXiv preprint arXiv:2505.03071*, 2025.

[5] Daniel Fink, Theodoros Damoulas, and Jaimin Dave. Adaptive spatio-temporal exploratory models: Hemisphere-wide species distributions from massively crowdsourced ebird data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 1284–1290, 2013.

[6] Tristan Gomez, Thomas Fréour, and Harold Mouchère. Metrics for saliency map evaluation of deep learning explanation methods. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 84–95. Springer, 2022.

[7] Julie J Gros-Louis, Susan E Perry, Claudia Fichtel, Eva Wikberg, Hannah Gilkenson, Susan Wofsy, and Alex Fuentes. Vocal repertoire of cebus capucinus: acoustic structure, context, and usage. *International Journal of Primatology*, 29(3):641–670, 2008.

[8] Fredrik Gustafsson. Determining the initial states in forward-backward filtering. *IEEE Transactions on signal processing*, 44(4):988–992, 2002.

[9] René Heinrich, Lukas Rauch, Bernhard Sick, and Christoph Scholz. Audioprotopnet: An interpretable deep learning model for bird sound classification, 2024. URL `https://arxiv.org/abs/2404.10420`.

[10] Stefan Kahl, Tom Denton, Holger Klinck, Vijay Ramesh, Viral Joshi, Meghana Srivathsa, Akshay Anand, Chiti Arvind, Harikrishnan Cp, Suyash Sawant, et al. Overview of birdclef 2024: Acoustic identification of under-studied bird species in the western ghats. CEUR-WS, 2024.

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[12] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

[13] Jinhua Liang, Inês Nolasco, Burooj Ghani, Huy Phan, Emmanouil Benetos, and Dan Stowell. Mind the domain gap: A systematic analysis on bioacoustic sound event detection. *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1257–1261, 2024. URL `https://api.semanticscholar.org/CorpusID:268723684`.

[14] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002.

[15] Joycelyn Longdon. Environmental data justice. *The Lancet Planetary Health*, 4(11):e510–e511, 2020.

[16] Joycelyn Longdon, Michelle Westerlaken, Alan F Blackwell, Jennifer Gabrys, Benjamin Ossom, Adham Ashton-Butt, and Emmanuel Acheampong. Justice-oriented design listening: Participatory ecoacoustics with a ghanaian forest community. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2024.

[17] Inês Nolasco, Burooj Ghani, Shubhr Singh, Ester Vidaña-Vila, Helen Whitehead, Emily Grout, Michael G. Emmerson, Frants Havmand Jensen, Ivan Kiskin, Joe Morford, Ariana Strandburg-Peshkin, Lisa F. Gill, Hanna Pamula, Vincent Lostanlen, and Dan Stowell. Few-shot bioacoustic event detection at the dcase 2023 challenge. *ArXiv*, abs/2306.09223, 2023. URL `https://api.semanticscholar.org/CorpusID:260472804`.

[18] David Robinson, Marius Miron, Masato Hagiwara, Benno Weck, Sara Keen, Milad Alizadeh, Gagan Narula, Matthieu Geist, and Olivier Pietquin. Naturelm-audio: an audio-language foundation model for bioacoustics. *arXiv preprint arXiv:2411.07186*, 2024.

[19] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336 – 359, 2016. URL `https://api.semanticscholar.org/CorpusID:15019293`.

[20] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International journal of human-computer studies*, 146:102551, 2021.

[21] Larissa De Andrade Silva, Juan G Colonna, Bernardo B Gatto, and João Marcelo Protázio. Impacts of anthropogenic noise on the house wren's song: An xai approach to bioacoustic insights. In *2025 IEEE Symposium on Trustworthy, Explainable and Responsible Computational Intelligence (CITREx)*, pages 1–7. IEEE, 2025.

[22] Bart van Merriënboer, Vincent Dumoulin, Jenny Hamer, Lauren Harrell, Andrea Burns, and Tom Denton. Perch 2.0: The bittern lesson for bioacoustics. *arXiv preprint arXiv:2508.04665*, 2025.

[23] Álvaro Vega-Hidalgo, Artem Abzaliev, Thore Bergman, and Rada Mihalcea. Acoustic individual identification of white-faced capuchin monkeys using joint multi-species embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 645–659, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.51. URL `https://aclanthology.org/2025.acl-short.51/`.

[24] Carlos Zednik and Hannes Boelsen. Scientific exploration and explainable artificial intelligence. *Minds and Machines*, 32(1):219–239, 2022.

[25] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.