AI Decodes Chimpanzee Vocabulary Items

Antoine Valet[1,2], Quentin Bacquelé[4], Cédric Girard-Buttoz[4], Roman M Wittig[1-3], Catherine Crockford[1-3]

[1]CNRS Institute for Cognitive Sciences, 67 Bld Pinel, 69500, Bron, Lyon,
[2]Taï Chimpanzee Project, Centre Swiss de Recherche, Ivory Coast.
[3]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[4]ENES, Bioacoustics Research Laboratory, Saint-Etienne, France

Abstract

Research continues to reveal complexity in animal vocal utterances (emission of vocalizations comprising one or more units). However, identifying what animal utterances mean remains an on-going challenge. Critical steps include mapping a clearly defined signal form on to clearly defined signal usage, where signal usage is assigned through a combination of the context of production and the receivers' response. Here, we address the problem of defining the signal form: the acoustic features which distinguish a call with one usage from a call with another usage, broadly analogous to identifying vocabulary items. Deciphering audio-recordings from species that live in noisy environments, like tropical forest, are particularly challenging, a task that may be aided by AI. To date, AI use in animal vocalizations recorded in natural environments has shown success in classifying vocalizations by species, individual, or call type. Much rarer are studies classifying vocalizations by the specific socio-ecological contexts in which they are emitted, such as greet, feed or play. We used the ANIMAL-SPOT algorithm with demonstrated success in call classification in noisy environments at the species level, on a sample of 1396 calls from 60 wild chimpanzees, Taï Forest, Ivory Coast. We successfully challenged ANIMAL-SPOT in a particularly hard task of classifying two acoustically similar noisy, graded 'grunt' calls to the correct context of production, either feed or greet, with >80% correct classification, when randomized simulations gave around 50% accuracy. We conclude, AI may emerge as a new tool for unravelling vocal communication in graded, noisy vocal systems, to automatically and rapidly distinguish subtle and hard to detect acoustic differences. Specifically, we demonstrate initial steps in applying AI to decode vocabulary in chimpanzees and other animals.

Introduction

Research continues to reveal complexity in animal utterances. Birds learn their songs from a parent, whales learn their group identity song; elephant rumbles, dolphin whistles and marmoset calls encode their own identity and that of their conversational partner, chimpanzee call combinations modify, add and generate new meanings compared to the composing calls[1–5].

Nonetheless, a crucial part of communication, identifying what animal utterances mean remains an on-going challenge. Key steps include mapping a clearly defined signal form onto clearly defined signal usage, where signal usage is assigned through a combination of the context of production and the receivers' response[6–8]. Here we address the problem of defining the signal form: the acoustic features which distinguish a call with one usage from a call with another usage, broadly analogous to identifying vocabulary items.

Particularly challenging is when calls elicit different responses but are acoustically similar to each other, as current tools can struggle to identify the acoustic features that define each call. Whilst traditional automated acoustic analyses procedures to assign call variants to contexts have been successful with loud calls[9–12], this proves highly challenging for low amplitude 'close' calls, especially for species living in noisy environments, like tropical forest[13], or for measuring certain variables, like formants, which can be difficult to measure, for example in calls with high fundamental frequencies[14,15], but which nonetheless likely convey information to receivers[15]. In such cases, either fully manual measurements of acoustic variables are more reliable[13], or manual checking of each automated acoustic variable[14], both highly time-consuming, and unfeasible with large datasets.

We are interested to see how well AI fairs in classifying call types (e.g., barks, hoos) or call variants (e.g., tonal versus noisy barks[10]) by context of production. We tested AI on a sample of chimpanzee acoustically graded, low-amplitude grunt calls from the Taï Forest, Ivory Coast, collected whilst conducting focal or ad libitum observational sampling from four habituated chimpanzee communities living in a natural tropical forest environment. Audio and behavioral sampling follows[16,17]. We included grunts emitted in one of two contexts, feed or greet.

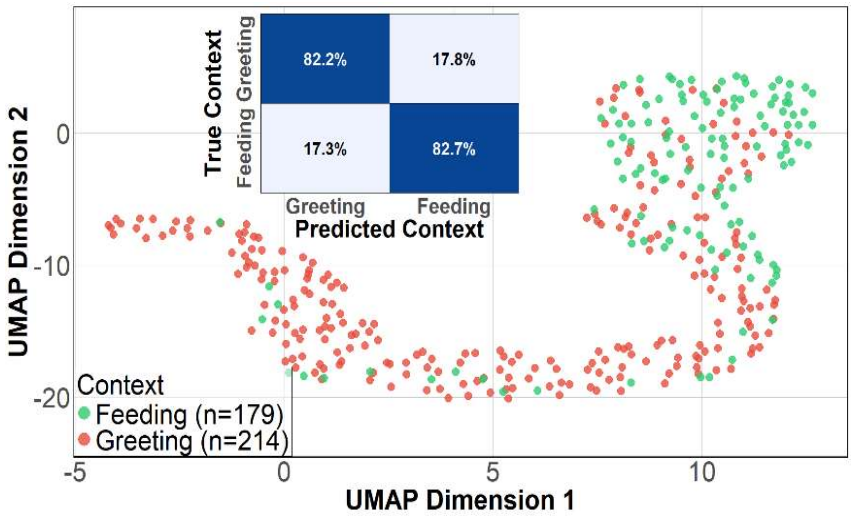

Figure 1. AI classification of two chimpanzee grunt variants by context of production (feed or greet), visualized using UMAP of the model embeddings, with associated confusion matrix. UMAP clustering (neighbors = 15, min_dist = 1.0), visualization of embeddings, colored by behavioral context, generated using a supervised convolutional neural network (ANIMAL-SPOT). The model includes grunts from 60 wild chimpanzees and was trained on datasets comprising 829 training samples, 174 validation samples, and 393 test samples, achieving an AUC of 0.872 for both classes.

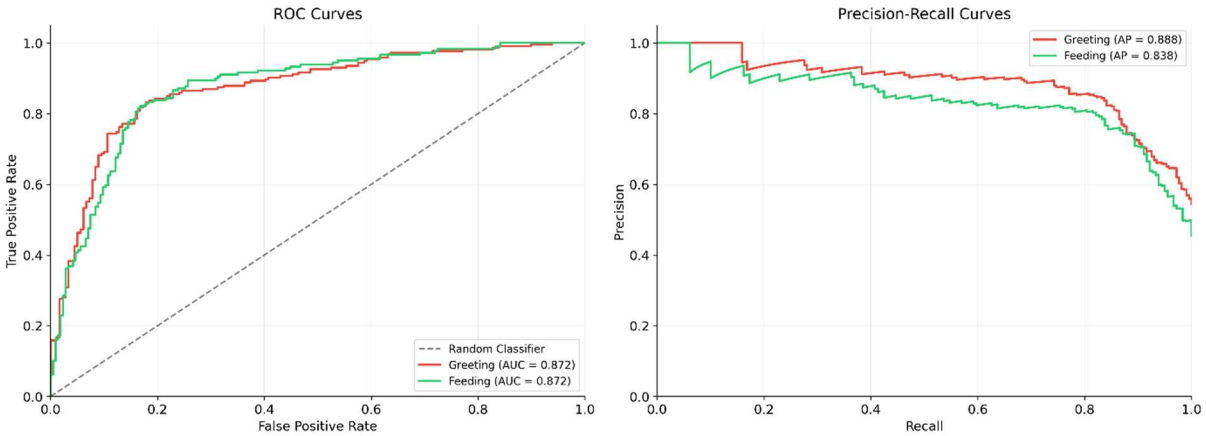

Figure 2. ROC and precision-recall curves for call classification. The left panel shows ROC curves for each call context, with grunts emitted in greeting contexts in red and grunts emitted in feeding contexts in green**.** The curves illustrate the trade-off between true positive rate and false positive rate for the model's predictions, with the diagonal dashed line representing random chance. Area Under the Curve (AUC) values are indicated for each context, reflecting the model's overall discriminability and showing performance above chance. The right panel shows Precision-Recall curves for the same contexts, highlighting the model's ability to maintain high precision across different recall thresholds. These curves are particularly informative for imbalanced classes.

Results & Discussion

AI starts to decode chimpanzee vocabulary

Working well for identification of species and individuals with song and long calls[6,18–21], few studies have tried to use AI to assign calls to contexts of production, a necessary step to link acoustic properties of calls to meaning[6,22]. As our goal was to address a biological question rather than optimize a deep learning model, we used ANIMAL-SPOT, a supervised convolutional neural network[18] (CNN), to classify two acoustically similar grunt variants, according to their context of production: food (N = 45 individuals, 654 calls) or greet (N = 49 individuals, 742 calls). This CNN has previously been validated with recordings from the Taï National Park, Ivory Coast [18], ensuring that it performs robustly with our dataset (also recorded in the Taï National Park).

Whilst observations suggest that chimpanzees hear a difference between the two target grunt variants[23,24], no study has yet successfully shown that these acoustically noisy and highly graded grunts emitted in different social contexts are acoustically distinct. Our AI results show classification of these two grunt variants, by context, with 82% accuracy (82.2% for greet grunt and 82.7 for feed grunts, **Figs. 1 & 2**), where simulated random assignments were < 55% accurate, demonstrating that ANIMAL-SPOT is not simply exploiting background noise or recording artifacts. To complement and validate our deep learning approach, we performed a Linear Discriminant Analysis (LDA) to identify which acoustic parameters contributed most to distinguishing feed vs greet grunt variants. This offers indirect insight into discriminating acoustic features potentially utilized by our deep learning process.

Due to the noisiness and gradation of chimpanzee grunts produced in each context, some acoustic features, such as the fundamental frequency or pitch, are difficult to measure reliably from spectrograms. We therefore focused on alternatives such as the Mel- frequency features (MFCC). The LDA achieved 83.6% classification accuracy for grunts in the two contexts, with bootstrap 95% confidence intervals [80.1-86.7%] (**Fig. S1**). A permutation test confirmed that these results are not biased by individual identify ($\mu = 0.696$ vs. 0.861 for the original score), showing a highly significant effect (364.64, $p < 0.001$). The ten most important acoustics features (selected via univariate analysis) were Mel-frequency coefficients, with the lowest frequency bands contributing the most (**Fig. S1**).

This first, successful challenge to classify two acoustically similar and noisy calls by contexts food or greet, suggests AI could also classify acoustically divergent calls, and hence much of the chimpanzee vocal repertoire by context. In sum, AI offers a critical step in decoding the chimpanzee vocal repertoire. Our next aim is to identify the features CNN relies on for classification by generating attention maps of grunt spectrogram critical areas. Our overarching goal is for deep learning or similar analysis to provide biologically meaningful insights into the acoustic features of calls emitted in different contexts that convey divergent information to chimpanzees.

Methods

Data preparation and annotation

To extract individual grunt elements, we used PRAAT software to segment each vocalization from the recordings. Each element was defined with precise start and end points corresponding to the onset and offset of the vocalization, minimizing the inclusion of background noise. As the ontogeny of these vocal variants in chimpanzees is not fully understood, only adult individuals (≥12 years old) were included in the analysis. Each element was annotated with metadata including caller identity, vocalization type, social group, and context of production. What signalers were doing whilst emitting vocalizations determined the assignment of the context of production. 'Greet' contexts were strictly assigned to grunt vocalizations emitted by subordinates to dominants whilst one individual approached the other. 'Feed' contexts were assigned when individuals grunted whilst eating or arriving at a food source[17,24]. To ensure accuracy, we excluded any grunt elements produced in ambiguous contexts, for example, those associated simultaneously both feed and greet, or greet and aggression contexts. Only calls with clear and unambiguous contexts, known signaler, and relatively clean spectrogram quality with little overlap, were retained for analysis.

The final dataset comprised 1,396 grunt elements (653 feed calls from 45 individuals and 743 greet calls from 49 individuals) produced by 60 unique individuals from four different neighbor communities, distributed across training (N = 20 chimpanzees), validation (N = 5 chimpanzees), and test (N = 35 chimpanzees) sets, with 25 individuals from the training and validation sets contributing calls in both contexts in a balanced manner. To minimize use of individuals from the same community between sets, and hence the possibility of ANIMAL SPOT identifying features related to non-context related acoustic features, such as group identity markers, individuals from the East and South groups were assigned to the training set, individuals from the North group to the validation

set, and individuals from the North-East group to the test set. The test set was the only data set to include any individuals from North-East group (N = 19 unique individuals), requiring generalization of context-related acoustic features for correct context assignment. Because the North-East group was underrepresented for feed calls, we complemented the test set with additional feed calls from individuals belonging to the other three groups but not assigned in training or validation set.

Dataset organization and splitting

We developed a Python script to reorganize the grunt dataset for the deep learning training. The script automatically structured the data into train, validation, and test sets while ensuring that no individual appeared in more than one set, preventing information leakage. Only callers with sufficient vocal material (≥3 utterances per context) were included in the training and validation sets, while other individuals were either excluded or allocated to the test set. To balance the dataset, a maximum of 30 elements per context per individual was sampled, prioritizing utterances with five or more elements. The resulting dataset maintained a binary classification framework**,** with *Feed* calls as the target class (46.8% of the dataset) and *Greet* calls as noise (53.2%). As a result, we obtained three subsets: 829 elements for training, 174 for validation, and 393 for testing.

Audio processing

Audio files were processed at a 44.1 kHz sampling rate, with spectrograms computed using a 1024-point FFT**, a** 172-sample hop length, and 256 frequency bins. Several trials were conducted to determine the optimal combination of hop length and frequency resolution. Optional preprocessing steps included filtering out broken audio files and applying min-max normalization[25].

Model training with ANIMAL-SPOT

We used ANIMAL-SPOT[26], an open-source supervised convolutional neural network framework for species-independent bioacoustics signal detection and classification, built on a ResNet-18 architecture[27,28] derived from ORCA-SPOT[29]. The network was implemented in PyTorch and trained on an NVIDIA A10 GPU (24 GB GDDR6, Ampere architecture)**.** Training used the Adam optimizer (initial learning rate = 1×10^{-5}, $\beta1 = 0.5$), a batch size of 16, and early stopping with a patience of 20 epochs. Datasets were loaded using PyTorch DataLoader objects with multi-worker support, and optional augmentation was applied to the training set when specified. Performance metrics, including accuracy**,** precision**,** recall**,** and F1-score**,** were evaluated on the validation and test sets. Training duration was approximately 15 minutes per run, depending on the dataset size.

Embedding Visualization and Classifier Evaluation

Embeddings were extracted[30] from the penultimate layer of the trained model, specifically after Layer 4 of the encoder, followed by global average pooling across the spatial dimensions. These embeddings were then projected into two dimensions using UMAP[31-33] to visualize the structure of the call space. UMAP was configured with 15 neighbors and a minimum distance of 1.0 to balance the preservation of local neighborhoods with the global separation of call contexts.

Model performance was assessed using a confusion matrix normalized by true class percentages, highlighting the proportion of correctly and incorrectly classified calls for each context. Additionally, ROC and precision-recall curves were computed for each vocal context (Greet in red, Feed in green) by comparing predicted probabilities with true labels. The area under the ROC curve (AUC) and average precision (AP) summarize classifier performance (**Fig. 2**), providing complementary measures of discriminative ability and precision across decision thresholds.

Acoustic features extraction and analysis

The following feature categories were extracted in Python with Librosa v0.10: Spectral features, temporal and energy features, cepstral features (MFCCs), Chroma and tonnetz features, and Mel-spectral features.

To determine which acoustic features distinguished the two context-specific grunt variants, we used a Linear Discriminant analysis (LDA). The dataset was split into training, validation, and test sets with a 70/15/15% ratio. In the training set, we performed a univariate score, using the one-standard-error rule, to obtain the optimal number of features. Model performance was quantified using classification accuracy and bootstrap confidence intervals (95%CI) on 1000 resamples. The statistical significance of discrimination between contexts was evaluated using a permutated Discriminant Function Analysis with 1000 random label shuffles to generate the null distribution of accuracy scores. Relative acoustic feature importance was derived from standardized LDA coefficients, identifying the most discriminating acoustics features.

Limitations

Future analyses will incorporate background noise segments (pre and post call) to test whether the CNN relies primarily on call structure rather than environmental noise, further validating the robustness of the acoustic signal classification. Given that individuals are repeatedly sampled across contexts, days, seasons and years and across their territory, we anticipate no overarching background noise features substantially contribute to the classification scores achieved. Assessing the model's generalizability will also be essential and will require testing it on data from different chimpanzee populations.

Conclusions and wider relevance

By identifying context-specific acoustic grunt variants, we provide further evidence[13] that chimpanzees' subtle acoustic differences encode quite different, functionally relevant information in their calls – a mechanism which expands their vocal repertoire and its messaging potential. We successfully challenged AI to classify acoustically similar noisy, graded calls to the correct context of production, making AI a future, potentially valuable, tool to decode the chimpanzee vocal repertoire. We demonstrate AI offers a powerful and promising approach for unravelling the communication of animals with graded vocal systems living in noisy environments, such as chimpanzees. Using an ethical approach whereby data collection does not negatively impact animals, we show the first steps in the possible use of AI in chimpanzee, and more broadly, animal, vocabulary decoding, a critical step in uncovering meanings embedded in animal conversation. However, this work represents only an initial step in a much longer and challenging journey towards fully applying AI to the decoding of animal communication.

Ethics: Data was sourced by us from the Taï National Park, Ivory Coast with our research permits and funding. We are strongly committed to non-invasive research. Data were collected whilst maintaining a 7 m distance from the chimpanzees, and following the IUCN health regulations for observing wild great apes (devised at this field site).

Checklist

1. Claims: are they justified? Yes
2. Limitations: are they discussed? Yes
3. Theory assumptions: NA
4. Experimental result reproducibility: are methods fully disclosed? No experiment was conducted. We used an algorithm already validated for species identification of calls, including for chimpanzee calls recorded in the same forest.
5. Open access to code and data: Not yet, but intended soon when we publish the preprint of this work.
6. Experimental settings/details: Included
7. Experimental statistical significance. Included.
8. Experiments computer resources: sufficient information on computer resources given to reproduce study? Yes.
9. Code of ethics: Yes
10. Broader aspects: positive and negative societal impacts discussed? Yes
11. Safeguards: for responsible release of data? Safeguards will be considered carefully before any data release.
12. Licenses for existing assets: others resources properly credited? Yes
13. New Assets: NA
14. Crowdsourcing and human involvement: NA
15. Institutional review board: NA
16. Declaration of LMM usage: Not used in any capacity.

References

1. Janik, V. M. & Slater, P. J. B. The different roles of social learning in vocal communication. *Animal Behaviour* **60**, 1–11 (2000).
2. King, S. L. & Janik, V. M. Bottlenose dolphins can use learned vocal labels to address each other. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13216–13221 (2013).
3. Girard-Buttoz, C., Neumann, C., Bortolato, T., Zaccarella, E., Friederici, A. D., Wittig, R. M., & Crockford, C. (2025). Versatile use of chimpanzee call combinations promotes meaning expansion. *Science Advances*, *11*(19), eadq2879.

4. Pardo, M. A. *et al.* African elephants address one another with individually specific name-like calls. *Nat Ecol Evol* **8**, 1353–1364 (2024).
5. Oren, G. *et al.* Vocal labeling of others by nonhuman primates. *Science* **385**, 996–1003 (2024).
6. Prat, Y., Taub, M. & Yovel, Y. Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. *Sci Rep* **6**, 39419 (2016).
7. Amphaeris, J., Blumstein, D. T., Shannon, G., Tenbrink, T., & Kershenbaum, A. A multifaceted framework to establish the presence of meaning in non-human communication. *Biological Reviews*, *98*(6), 1887-1909.*Biological Reviews* **91**, 13–52 (2016).
8. Seyfarth, R. M. & Cheney, D. L. How monkeys see the world. Chicago University Press (1990).
9. Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. Characterizing vocal repertoires—hard vs. soft classification approaches. *PloS one*, *10*(4), e0125785 (2015).
10. Crockford, C. & Boesch, C. Context-specific calls in wild chimpanzees, Pan troglodytes verus: analysis of barks. *Animal Behaviour* **66**, 115–125 (2003).
11. Crockford, C., Herbinger, I., Vigilant, L. & Boesch, C. Wild Chimpanzees Produce Group-Specific Calls: a Case for Vocal Learning? *Ethology* **110**, 221–243 (2004).
12. Fischer, J., Noser, R., & Hammerschmidt, K. (2013). Bioacoustic field research: a primer to acoustic analyses and playback experiments with primates. *American journal of primatology*, *75*(7), 643-663.
13 Crockford, C., Gruber, T., & Zuberbühler, K. Chimpanzee quiet hoo variants differ according to context. *Royal Society open science*, *5*(5), 172066 (2018).
14. Grawunder, S., Uomini, N., Samuni, L., Bortolato, T., Girard-Buttoz, C., Wittig, R. M., & Crockford, C. Chimpanzee vowel-like sounds and voice quality suggest formant space expansion through the hominoid lineage. *Philosophical Transactions of the Royal Society B*, *377*(1841), 20200455 (2022).
15. Fitch, W. T., Anikin, A., Pisanski, K., Valente, D. & Reby, D. Formant analysis of vertebrate vocalizations: achievements, pitfalls, and promises. *BMC Biol* **23**, 92 (2025).
16. Girard-Buttoz, C., Zaccarella, E., Bortolato, T., Friederici, A. D., Wittig, R. M., & Crockford, C. Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. *Communications Biology*, *5*(1), 410 (2022).
17. Bortolato, T., Friederici, A. D., Girard-Buttoz, C., Wittig, R. M. & Crockford, C. Chimpanzees show the capacity to communicate about concomitant daily life events. *iScience* **26**, 108090 (2023).
18. Bergler, C. *et al.* ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. *Sci Rep* **12**, 21966 (2022).
19. Cauzinille, J., Favre, B., Marxer, R. & Rey, A. Applying machine learning to primate bioacoustics: Review and perspectives. *American J Primatol* **86**, e23666 (2024).
20. Chitayat, A. *et al.* Acoustics, balance, and chimpanzees – The ABCs of developing a deep learning-based automated acoustic detector for wild chimpanzee (Pan troglodytes) loud calls. Preprint at https://doi.org/10.22541/au.173814850.06519206/v1 (2025).
21. Fedurek, P., Zuberbühler, K. & Dahl, C. D. Sequential information in a great ape utterance. *Sci Rep* **6**, 38226 (2016).
22. Amit, Y., & Yovel, Y. Bat vocal sequences enhance contextual information independently of syllable order. *Iscience*, *26*(4) (2023).
23. Slocombe, K. E. & Zuberbühler, K. Functionally Referential Communication in a Chimpanzee. *Current Biology* **15**, 1779–1784 (2005).
24. Crockford, C. Why does the chimpanzee vocal repertoire remain poorly understood? *The chimpanzees of the Taï forest: 40 years of research*. Cambridge University Press (2019).
25. Knight, E. C., Hernandez, S. P., Bayne, E. M., Bulitko, V., & Tucker, B. V. (2020). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics,* 29(3), 337–355. https://doi.org/10.1080/09524622.2019.1606734
26. Bergler, C., Smeele, S.Q., Tyndel, S.A. et al. ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. *Sci Rep* 12, 21966 (2022). https://doi.org/10.1038/s41598-022-26429-y
27. Hershey, S., Chaudhuri, S., Ellis, J., Gemmeke, D., Jansen *et al*. CNN Architectures for Large-Scale Audio Classification. arXiv:1609.09430
28. Salamon, J., Bello, J. P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification arXiv:1608.04363
29. Bergler, C., Schröter, H., Cheng, R.X. *et al*. ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning. *Sci Rep* 9, 10997 (2019). https://doi.org/10.1038/s41598-019-47335-w
30. Thomas, M., Jensen, F. H., Averly, B., Demartsev, V., Manser, M. B., Sainburg, T., ... & Strandburg-Peshkin, A. A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *Journal of Animal Ecology*, *91*(8), 1567-1581 (2022). DOI: 10.1111/1365-2656.13754
31. Cominelli S, Bellin N, Brown CD, Rossi V, Lawson J. Acoustic features as a tool to visualize and explore marine soundscapes: Applications illustrated using marine mammal passive acoustic monitoring datasets. *Ecol Evol.* 21;14(2):e10951. (2022) doi: 10.1002/ece3.10951. PMID: 38384822; PMCID: PMC10880131.
32. Arnaud V, Pellegrino F, Keenan S, St-Gelais X, Mathevon N, Levréro F, Coupé C. Improving the workflow to crack Small, Unbalanced, Noisy, but Genuine (SUNG) datasets in bioacoustics: The case of bonobo calls. *PLoS Comput Biol*. Apr 13;19(4):e1010325 (2023). doi: 10.1371/journal.pcbi.1010325. PMID: 37053268; PMCID: PMC10129004.
33. Parra-Hernández RM, Posada-Quintero JI, Acevedo-Charry O, Posada-Quintero HF. Uniform Manifold Approximation and Projection for Clustering Taxa through Vocalizations in a Neotropical Passerine (Rough-Legged Tyrannulet, Phyllomyias burmeisteri). *Animals* (Basel). 2020 Aug 12;10(8):1406. doi: 10.3390/ani10081406. PMID: 32806680; PMCID: PMC7460062.

Supplementary Materials

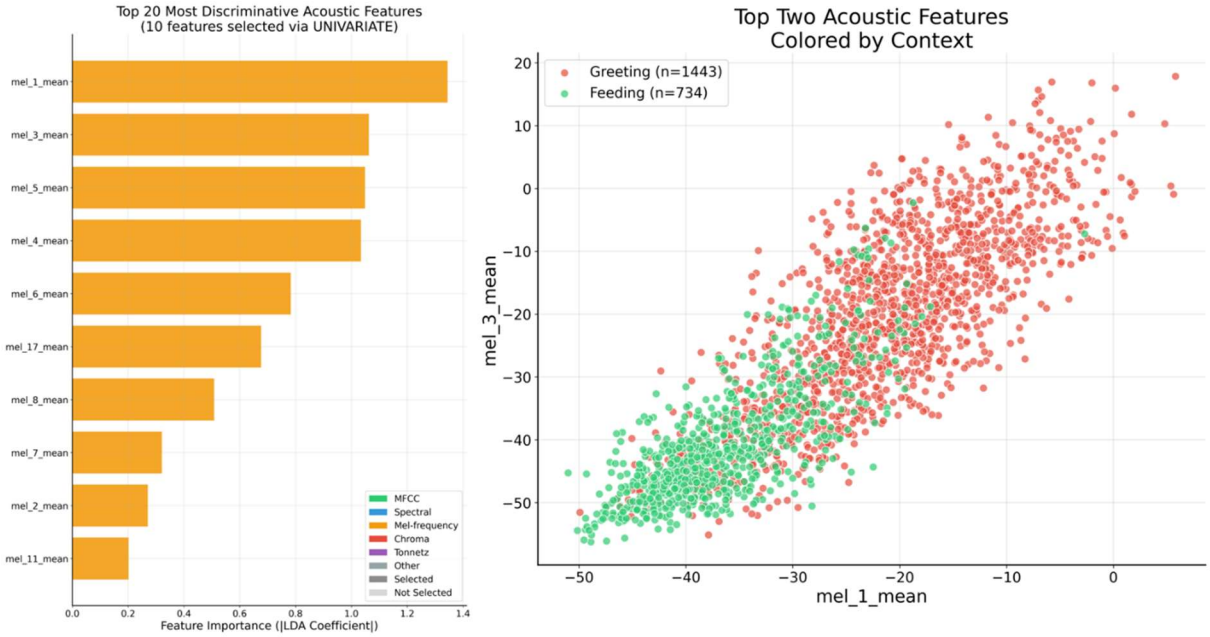

Figure S1. Acoustic features discriminating grunt variants in feed and greet contexts. Left: the ten most discriminating acoustic features identified through Linear Discriminant Analysis and selected using univariate-scores for grunts produced in feed and greet contexts. Right: distribution of grunts along the two most discriminating features, mel_1_mean and mel_3_mean, highlighting the acoustic separation between grunts emitted in greet (red dots) and feed (green dots) contexts.