Acoustic Individual Identification of White-Faced Capuchin Monkeys Using Joint Multi-Species Embeddings

Anonymous ACL submission

Abstract

Acoustic individual identification of wild animals is an essential task for understanding animal vocalizations within their social contexts, and for facilitating conservation and wildlife monitoring efforts. However, most of the work in this space relies on human efforts, as the development of methods for automatic individual identification is hindered by the lack of data. In this paper, we explore cross-species pre-training to address the task of individual classification in White-Faced Capuchin monkeys. Using acoustic embeddings from birds and humans, we find that they can be effectively used to identify the calls from individual monkeys. Moreover, we find that joint multi-species representations can lead to further improvements over the use of one representation at a time. Our work demonstrates the potential of cross-species data transfer and multi-species representations, as strategies to address tasks on species with very limited data.

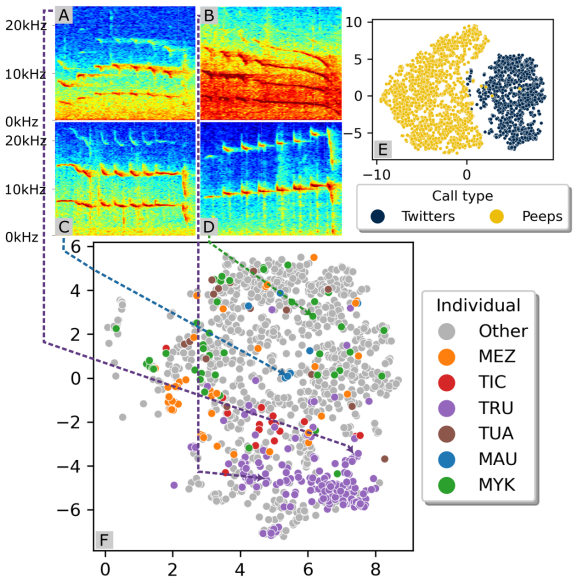

Figure 1: (A–D) Capuchin Twitter vocalizations show diverse structural variations. (E–F) t-SNE of Google Perch-Whisper embeddings. (E) Call type clusters. (F) Colored by individual, highlighting four diverse examples of Twitters.

1 Introduction

For a long time, researchers viewed the vocalizations of non-human species as mere reactions to internal emotional states (Lorenz, 1952). Consequently, early scientific methods in animal communication research largely overlooked individual differences and did not test for the presence of linguistic features (e.g., pragmatics, semantics, syntax) in animal communication systems. This simplified view of animal communication has been overturned by the growing evidence uncovering the presence of linguistics features in non-human animals (Bergman et al., 2019), leading to the emergence of Animal Linguistics as a formal interdisciplinary research field (Bowling and Fitch, 2015; Engesser et al., 2015; Suzuki, 2024; Berthet et al., 2023; Suzuki, 2021; Scott-Phillips and Heintz, 2023). This shift in perspective highlights the need for individual-level analysis, as it allows researchers to account for the social and environmental contexts in which vocalizations occur, ultimately improving our ability to test their linguistic capacities more rigorously.

Additionally, long-term, individual-level analyses are critical for understanding and protecting wildlife. Such analyses support key approaches like social network quantification, assessing animal cognition, and performing capture–recapture techniques for tracking population dynamics (Slater, 1981; Carlson et al., 2020). Over the past decade, acoustic monitoring has emerged as a widely adopted, cost-efficient strategy in conservation, leading to growing interest in acoustic individual identification. By enabling researchers to recognize individuals from their vocalizations, this approach paves the way for more nuanced insights into ecology, behavior, evolution, and conservation (Knight

et al., 2024).

In this paper, we address the task of acoustic individual identification (Acoustic ID) in White-Faced Capuchins. We collect a two-year dataset of individualized focal recordings, a labor-intensive yet optimized method that mitigates signal-to-noise and cocktail party problems in wild bioacoustic settings (Bermant, 2021; Miron et al., 2024). Using this dataset, we evaluate human speech- and bird bioacoustic-based pre-trained networks, comparing single-embedding models to ensembles that merge embeddings from distinct networks. We hypothesize that human speech embeddings, such as Whisper or HuBERT, complement bioacoustic embeddings like Google Perch or BirdNET–originally trained on bird sounds–and predict that heterogeneous embedding combinations will outperform single-embedding models.

Transfer learning has significantly advanced acoustic classification tasks in non-human animals (Miyaguchi et al., 2024; Kahl et al., 2023; Abzaliev et al., 2024). Recent studies on gibbons have explored the use of self-supervised speech models (e.g., HuBERT, Wav2vec 2.0), pre-trained bird classifiers (e.g., BirdNET, Perch), and non-transfer-learning deep models for primate Acoustic ID, finding that speech models most effectively capture individual vocal signatures, bird classifiers perform well in automated detection but are more susceptible to background noise, and non-transfer-learning models struggle when trained on small datasets (Cauzinille et al., 2024; Clink et al., 2024). Nevertheless, it remains unclear whether using multiple joint embeddings leads to better performance by exploiting complementary features from different training data domains.

This work makes three main contributions. First, we propose White-Faced Capuchin monkeys as a model organism for advancing computational research on animal communication. Second, we show that combining embeddings from human speech and bird bioacoustics models significantly improves acoustic identification performance in White-Faced Capuchins, outperforming single-embedding baselines. Finally, our findings show that acoustic diversity and soundscape similarity play a greater role than phylogenetic proximity. Smaller models trained on diverse bird vocalizations recorded in natural environments outperform much larger speech-trained models designed for humans, despite humans being more closely related to our study species. These results highlight the value of cross-species model development in achieving better generalization for the Acoustic ID task.

2 Study system: White-Faced Capuchin Monkeys in the Taboga Reserve, Costa Rica

White-Faced Capuchin Monkeys (*Cebus capucinus*) are ideal for studying animal communication, with 27 call types (Gros-Louis et al., 2008), complex social behavior and cognition including tool use (Goldsborough et al., 2024), complex social networks (Crofoot et al., 2011) and cultural transmission (Perry et al., 2017). Taboga hosts their highest known density (Tinsley Johnson et al., 2020).

Data collection. Our field team collected audio recordings of focal individuals by following them in the Taboga forest. We used directional microphones aimed at the subjects from January 2021 to December 2022 through the wet and dry seasons, with hours ranging from 5 am to 5 pm. Recordings were captured at $48\,\mathrm{kHz}$ and $16\,\mathrm{bit}$ resolution. These raw recordings were subsequently trimmed to isolate the precise moments when vocalizations were detected, and only the calls classified as either a "Peep" or "Twitter" were included in this dataset, according to established criteria in the literature (Gros-Louis et al., 2008).

Audio recordings. The full dataset consists of 1,257 Twitter recordings and 2,089 Peep recordings from 45 individuals, although 15% of the recordings were assigned to unknown individuals. We include data from individuals that had at least 30 recorded calls, while recordings from unidentified subjects encountered in the field are grouped into an "Unknown" class. For Peeps, this dataset includes 15 individuals, and for Twitters this dataset included 10 individuals (total sample=1609). Peep calls are typically short (mean $0.27\,\mathrm{s}$, SD $0.27\,\mathrm{s}$), whereas Twitter calls are more complex (Figure 1) and longer (mean $0.40\,\mathrm{s}$, SD $0.18\,\mathrm{s}$).

3 Cross-Species Embeddings for Individual Classification

Collecting focal audio recordings of wild animals in their natural habitat is a challenging and resource-intensive task. Even with dedicated field teams, building large enough datasets to fully exploit deep neural networks is difficult. As a result, transfer learning–which leverages the induc-

tive bias of models pre-trained on larger, related datasets–has emerged as the most effective strategy for achieving high performance in bioacoustic classification under low-data conditions (Ghani et al., 2023a).

Audio Representation Models. We extract pre-trained embeddings from Google Perch V8 (Ghani et al., 2023a), a model primarily trained on bird vocalizations, and Whisper (Radford et al., 2022), which was predominantly pre-trained on human speech. While additional embeddings were evaluated, we focus on these two in the main text for clarity, with results from five other models detailed in Appendix A. We apply mean-pooling to obtain lower-dimensional representations from large speech models like Whisper.

Minimum Redundancy Maximum Relevance. To combine representations from multiple species, we explore a feature-select model using Minimum Redundancy and Maximum Relevance (MRMR) (Ding and Peng, 2005), alongside simple concatenation and summation. Originally developed in cancer research for gene selection, MRMR improves feature selection in high-dimensional datasets by balancing two key criteria: maximizing relevance to the target variable (measured via mutual information) while minimizing redundancy (filtered using a correlation coefficient threshold). Our implementation starts with the feature that has the highest mutual information among both embeddings, removes any features with a correlation coefficient of 0.8 or higher, and then iteratively selects the next most informative feature. This process continues until 1024 embedding features are selected from both embeddings, ensuring an optimal balance of diversity and informativeness.

Experimental Setup. To ensure a fair comparison, we carefully control parameter counts and apply hyperparameter tuning. Single-embedding models and the MRMR model compress each input into 512 units, then reduced it to 64 for final classification. Concatenation and summation ensembles apply a 256-dimensional compression to each embedding separately, then sum or concatenate the outputs before another 64-unit layer. For a robust comparison, we generate 50 random train-test splits (10 recordings per individual in the test set) and train models with all seven single embeddings as well as all pairwise combinations (concatenation, summation, and MRMR).

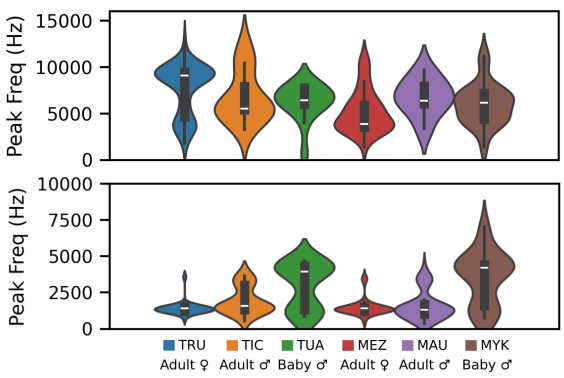

Figure 2: Peak frequency for the Peeps call type recordings of six capuchin monkey individuals

To identify the best hyperparameters for each model trained, we conduct a search over learning rates $\{1\text{e-}5, 5\text{e-}5, 1\text{e-}4, 5\text{e-}4, 1\text{e-}3, 5\text{e-}3\}$ and dropout rates $\{0.2, \ldots, 0.6\}$, evaluating 30 randomly sampled configurations for 100 epochs each with early stopping (patience=10, min ΔF1=0.001), and selected the highest F1-scoring setup. All models are trained using the Adam optimizer. After confirming normality and homoscedasticity, we compare each architecture's top-performing model via ANOVA and a post-hoc Tukey test.

Spectrogram annotations and measurements. To compare explainable acoustic features with non-interpretable deep embeddings, we manually measure Peak Frequency and other acoustic parameters from spectrograms, following standard bioacoustic methods. Using Raven Pro 1.6 (K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology, 2024), we select regions of interest and extract 30 interpretable features (see Appendix A), including Peak Frequency, Center Frequency, and Center Time. These measurements were taken from six individuals—one adult male, one adult female, and one infant from each of the two monkey troops—chosen for their distinct characteristics.

4 Results

Table 1 shows the results of the Acoustic ID task for selected models. We present F1 scores for the models trained on bird vocalizations and human speech data, together with their ensembles. While single-species vocalization models perform reasonably well, the models with the highest F1 scores are those that combine multiple embeddings (either using summation, concatenation or MRMR). Fur-

thermore, the best-performing ensemble combine models developed for bioacoustic vocalizations and models developed for human speech. This highlights the potential of cross-species pre-training in a limited data regime. Pre-training on human speech does not capture enough information for the bioacoustic domain, as shown by the performance of Whisper for both vocalization types. But combined, those two models achieve an F1 score of 0.70 for Peeps and 0.66 for Twitters. This improved performance suggests that combining speech-trained and bioacoustic-trained embeddings effectively leverages complementary information. We also present the results for other models in Appendix A.

Despite its smaller size and more limited training dataset, the bioacoustic model Perch outperforms the much larger Whisper model, which was developed for human speech. Domain relevance is more important than model size, training data set size, or phylogenetic proximity for the Acoustic ID task in Capuchins. Trained on data from noisy field conditions, Perch learns the acoustic variability of field conditions, contributing to its strong performance. Although our focal species is neither a bird nor a human, the top-performing models across architectures are trained using both bird- and human-derived embeddings, suggesting that joint multi-species embeddings provide better generalization for Capuchin acoustic classification tasks.

We visualize the embeddings of the best-performing model from table 1 using t-SNE (van der Maaten, 2009) in Figure 1. Different call types formed well-defined clusters (Figure 1E), whereas individual classifications appear more diffuse (Figure 1F), illustrating the difficulty of the Acoustic ID task (see A for more t-SNE visualizations). We also analyze the distribution of peak frequencies across individuals in Figure 2). Lower-pitched sounds characterize Peeps, while Twitters span a broader spectral range of peak frequencies. Notably, both call types exhibit bimodal distributions, with this pattern being more pronounced in certain individuals. This bimodal distribution could reflect two or more call subtypes with distinct pitches and should be investigated further to test for the existence of pragmatics or semantics in their communication system through pitch modulation. Variability within the Twitter call type extends beyond overall pitch modulation. Some Twitters exhibit an n-shaped pitch contour, a continuous descending note, a final lower-pitched note, or a rising pitch throughout the call (Figure 1-A,B,C,D, respectively). Empirical studies incorporating rich social and environmental contexts will be crucial for uncovering the functional significance of this variation in Capuchin calls.

Table 1: Top-performing models for Twitters and Peeps (Mean F1 Score ± SD), with significance assessed by comparison to the best simple model (Perch). Significance levels: * for $p < 0.05$ and ** for $p < 0.0001$ (Tukey's test).

Model	**F1 Score**
Twitters	
Perch (Simple)	0.61 ± 0.03
Whisper (Simple)	0.55 ± 0.03
Perch + Whisper (Concat)	0.63 ± 0.03
Perch + Whisper (Sum)	0.63 ± 0.03*
Perch + Whisper (MRMR)	**0.66 ± 0.03****
Peeps	
Perch (Simple)	0.66 ± 0.02
Whisper (Simple)	0.62 ± 0.03
Perch + Whisper (Concat)	0.67 ± 0.02*
Perch + Whisper (Sum)	0.68 ± 0.02**
Perch + Whisper (MRMR)	**0.70 ± 0.02****

5 Conclusion

This study examined Acoustic Individual Identification in two call types of White-Faced Capuchins. We established performance baselines for pre-trained embeddings and found that combining multiple embeddings (summation, concatenation, and minimum redundancy maximum relevance) improves classification performance. Our findings also indicate that domain relevance outweighs model size in noisy environments. Future work should extend these multi-species embeddings to other taxa, confirming broader applicability in bioacoustics and animal linguistics.

6 Limitations

While this study focused on Acoustic ID, a deeper investigation into the behavioral and social functions of these call types remains relevant for future work. While there are other ways of improving Acoustic ID, such as data augmentation (MacIsaac et al., 2024), we considered those techniques out of scope for the present study and focused on investigating the complementarity of joint multi-species

embeddings. Our primary goal with this dataset is to make it accessible to the broader scientific community. We anticipate making it publicly available in a forthcoming study with further analyses.

7 Ethical Considerations

No animals were harmed during this study. All research adhered to ethical guidelines for animal welfare, recognizing the importance of studying animal communication while prioritizing their well-being, particularly in the context of climate change and habitat loss affecting this species. Additionally, all individuals involved in data collection and processing were engaged in formal employment or academic research under ethical labor practices.

References

Artem Abzaliev, Humberto Pérez Espinosa, and Rada Mihalcea. 2024. Towards dog bark decoding: Leveraging human speech processing for automated bark classification. *arXiv preprint arXiv:2404.18739*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Thore J Bergman, Jacinta C Beehner, Melissa C Painter, and Morgan L Gustison. 2019. The speech-like properties of nonhuman primate vocalizations. *Animal Behaviour*, 151:229–237.

Peter C Bermant. 2021. Biocppnet: automatic bioacoustic source separation with deep neural networks. *Scientific Reports*, 11(1):23502.

Mélissa Berthet, Camille Coye, Guillaume Dezecache, and Jeremy Kuhn. 2023. Animal linguistics: a primer. *Biological reviews*, 98(1):81–98.

Daniel L Bowling and W Tecumseh Fitch. 2015. Do animal communication systems have phonemes? *Trends in Cognitive Sciences*, 19(10):555–557.

Nora V Carlson, E McKenna Kelly, and Iain Couzin. 2020. Individual vocal recognition across taxa: a review of the literature and a look into the future. *Philosophical Transactions of the Royal Society B*, 375(1802):20190479.

Jules Cauzinille, Benoît Favre, Ricard Marxer, Dena Clink, Abdul Hamid Ahmad, and Arnaud Rey. 2024. Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures. In *Interspeech 2024*, pages 132–136. ISCA; ISCA.

Dena J Clink, Hope Cross-Jaya, Jinsung Kim, Abdul Hamid Ahmad, Moeurk Hong, Roeun Sala, Hélène Birot, Cain Agger, Thinh Tien Vu, Hoa Nguyen Thi, et al. 2024. Benchmarking automated detection and classification approaches for monitoring of endangered species: a case study on gibbons from cambodia. *bioRxiv*, pages 2024–08.

Margaret C Crofoot, Daniel I Rubenstein, Arun S Maiya, and Tanya Y Berger-Wolf. 2011. Aggression, grooming and group-level cooperation in white-faced capuchins (cebus capucinus): Insights from social networks. *American Journal of Primatology*, 73(8):821–833.

Chris Ding and Hanchuan Peng. 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205.

Sabrina Engesser, Jodie MS Crane, James L Savage, Andrew F Russell, and Simon W Townsend. 2015. Experimental evidence for phonemic contrasts in a non-human vocal system. *PLoS biology*, 13(6):e1002171.

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

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

Zoë Goldsborough, Margaret C Crofoot, and Brendan J Barrett. 2024. Male-biased stone tool use by wild white-faced capuchins (cebus capucinus imitator). *American Journal of Primatology*, 86(4):e23594.

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

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2023. Masked autoencoders that listen. *Preprint*, arXiv:2207.06405.

K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology. 2024. Raven pro: Interactive sound analysis software (version 1.6.5). [Computer software].

Stefan Kahl, Tom Denton, Holger Klinck, Hendrik Reers, Francis Cherutich, Hervé Glotin, Hervé Goëau, Willem-Pier Vellinga, Robert Planqué, and Alexis Joly. 2023. Overview of birdclef 2023: Automated bird species identification in eastern africa. In *CLEF (Working Notes)*, pages 1934–1942.

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. 2021. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236.

Elly Knight, Tessa Rhinehart, Devin R de Zwaan, Matthew J Weldy, Mark Cartwright, Scott H Hawley, Jeffery L Larkin, Damon Lesmeister, Erin Bayne, and Justin Kitzes. 2024. Individual identification in acoustic recordings. *Trends in Ecology & Evolution*.

Konrad Lorenz. 1952. *King Solomon's Ring*. T. Y. Crowell, New York.

Jennifer MacIsaac, Stuart Newson, Adham Ashton-Butt, Huma Pearce, and Ben Milner. 2024. Improving acoustic species identification using data augmentation within a deep learning framework. *Ecological Informatics*, 83:102851.

Marius Miron, Sara Keen, Jen-Yu Liu, Benjamin Hoffman, Masato Hagiwara, Olivier Pietquin, Felix Effenberger, and Maddie Cusimano. 2024. Biodenoising: animal vocalization denoising without access to clean data. *arXiv preprint arXiv:2410.03427*.

Anthony Miyaguchi, Adrian Cheung, Murilo Gustineli, and Ashley Kim. 2024. Transfer learning with pseudo multi-label birdcall classification for ds@ gt birdclef 2024. *arXiv preprint arXiv:2407.06291*.

Susan E Perry, Brendan J Barrett, and Irene Godoy. 2017. Older, sociable capuchins (cebus capucinus) invent more social behaviors, but younger monkeys innovate more in other contexts. *Proceedings of the National Academy of Sciences*, 114(30):7806–7813.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Thom Scott-Phillips and Christophe Heintz. 2023. Animal communication in linguistic and cognitive perspective. *Annual Review of Linguistics*, 9(1):93–111.

PJB Slater. 1981. Individual differences in animal behavior. In *Perspectives in Ethology: Volume 4 Advantages of Diversity*, pages 35–49. Springer.

Toshitaka N Suzuki. 2021. Animal linguistics: exploring referentiality and compositionality in bird calls. *Ecological Research*, 36(2):221–231.

Toshitaka N Suzuki. 2024. Animal linguistics. *Annual Review of Ecology, Evolution, and Systematics*, 55.

Elizabeth Tinsley Johnson, Marcela E Benítez, Alexander Fuentes, Celia R McLean, Ariek B Norford, Juan Carlos Ordoñez, Jacinta C Beehner, and Thore J Bergman. 2020. High density of white-faced capuchins (cebus capucinus) and habitat quality in the taboga forest of costa rica. *American Journal of Primatology*, 82(2):e23096.

Laurens van der Maaten. 2009. Learning a parametric embedding by preserving local structure. In *International Conference on Artificial Intelligence and Statistics*.

A Appendix

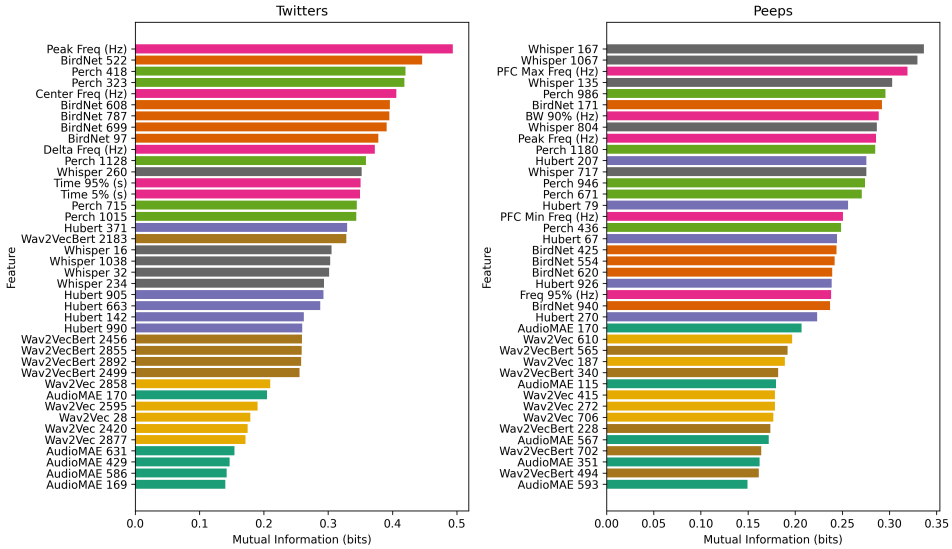

Figure 3: Mutual information of the top features for both call type datasets, spanning seven acoustic pre-trained embeddings. We display the five highest-performing features per pre-trained embedding, along with the top five interpretable features per model.

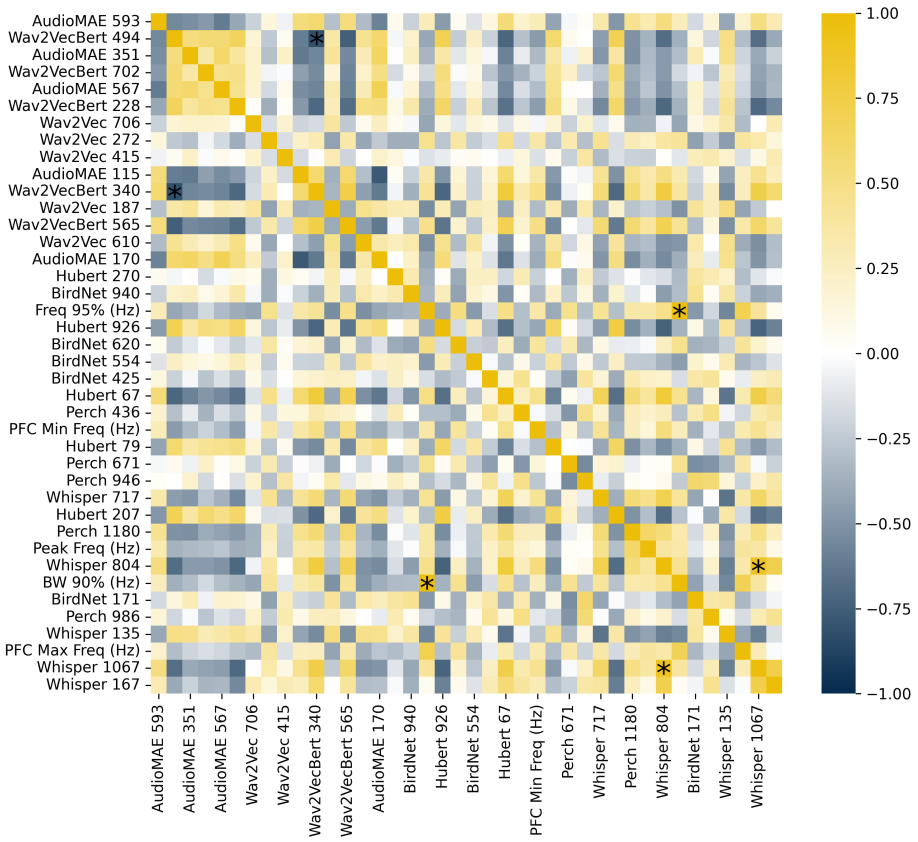

Figure 4: Mutual information of the top features in the Peeps call type dataset, spanning seven acoustic pre-trained embeddings. We display the five highest-performing features per pre-trained embedding, along with the top five interpretable features. Asterisks show correlation coefficients above 0.8.

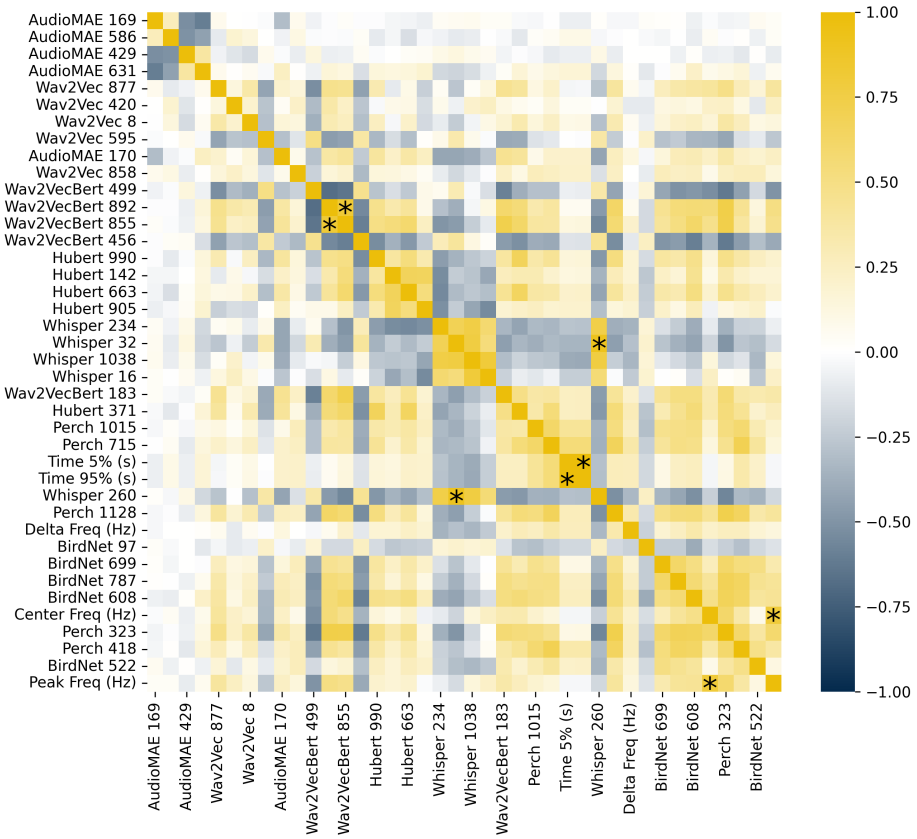

Figure 5: Mutual information of the top features in the Twitters call type dataset, spanning seven acoustic pre-trained embeddings. We display the five highest-performing features per pre-trained embedding, along with the top five interpretable features. Asterisks show correlation coefficients above 0.8.

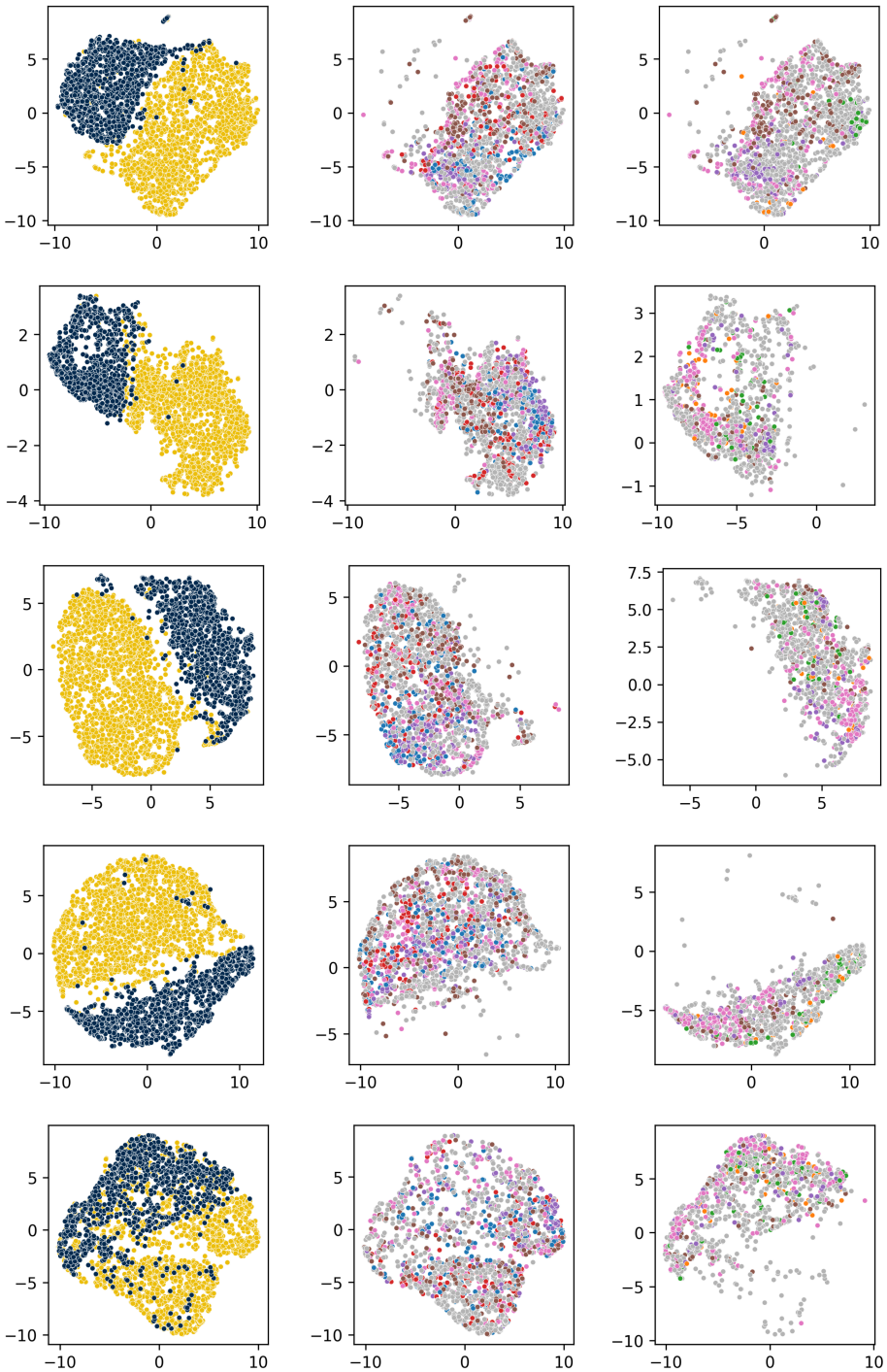

Figure 6: t-SNE visualizations of five pre-trained embeddings, primarily trained on human speech data (with AudioMAE also incorporating internet-sourced audio). The first column presents the t-SNE plot of call types (Peeps in yellow and Twitters in blue), while the second and third columns show the t-SNE projections of Twitters and Peeps, respectively, with points colored by individual identity. From top to bottom, the rows correspond to HuBERT, Wav2Vec, Wav2Vec BERT, Whisper, and AudioMAE.

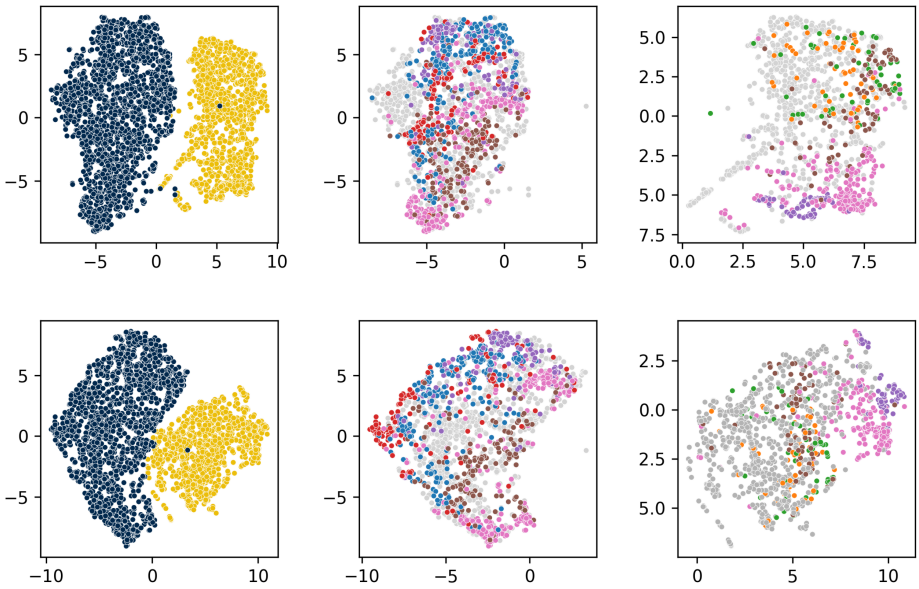

Figure 7: tSNE visualizations of five pre-trained embeddings, primarily trained on bioacoustics bird data (with BirdNET also incorporating other animals). The first column represents the t-SNE plot of call types (Peeps in yellow and Twitters in blue), while the second and third columns depict the t-SNE projections of Twitters and Peeps colored by individual, respectively. From top to bottom, the rows correspond to BirdNET and Perch, respectively.

Model	F1 Score (Mean ± Std)
Simple Network	
Perch	0.66 ± 0.02
Whisper	0.62 ± 0.03
BirdNet	0.60 ± 0.02
HuBERt	0.55 ± 0.02
Wav2Vec2	0.47 ± 0.02
Concatenation (2 Embeddings)	
BirdNET + Perch	0.67 ± 0.02
Perch + Whisper	0.67 ± 0.02
Perch + Hubert	0.66 ± 0.02
Perch + AudioMAE	0.64 ± 0.02
Perch + Wav2Vec2	0.64 ± 0.02
Summation (2 Embeddings)	
Perch + Whisper	0.68 ± 0.02
Perch + BirdNET	0.67 ± 0.02
Perch + HuBERT	0.66 ± 0.02
BirdNet + Whisper	0.64 ± 0.02
Perch + Wav2Vec2	0.64 ± 0.02
MRMR (2 Embeddings)	
Perch + Whisper	**0.70 ± 0.02**
Perch + BirdNET	0.69 ± 0.02
Perch + HuBERT	0.68 ± 0.02
Perch + Wav2Vec2	0.67 ± 0.02
Perch + Wav2Vec-bert	0.67 ± 0.02

Table 2: Performance of the top 5 models per method on the Acoustic ID task using the Peeps dataset (Mean F1 Score ± Standard Deviation).

Model	F1 Score (Mean ± Std)
Simple Network	
Perch	0.61 ± 0.03
BirdNet	0.60 ± 0.04
HuBERT	0.56 ± 0.04
Whisper	0.55 ± 0.03
Wav2Vec-bert	0.43 ± 0.03
Concatenation (2 Embeddings)	
BirdNET + Whisper	0.63 ± 0.03
BirdNET + Perch	0.62 ± 0.03
Perch + Whisper	0.62 ± 0.03
BirdNET + HuBERT	0.62 ± 0.03
Perch + HuBERT	0.61 ± 0.03
Summation (2 Embeddings)	
BirdNET + Whisper	0.63 ± 0.04
Perch + Whisper	0.63 ± 0.03
BirdNET + Perch	0.63 ± 0.03
BirdNET + HuBERT	0.62 ± 0.03
Perch + HuBERT	0.62 ± 0.03
MRMR (2 Embeddings)	
Perch + Whisper	**0.66 ± 0.03**
BirdNET + Whisper	0.65 ± 0.03
Perch + HuBERT	0.64 ± 0.03
BirdNET + Perch	0.64 ± 0.03
Perch + Wav2Vec2	0.64 ± 0.03

Table 3: Performance of the top 5 models per method on the Acoustic ID task using the Twitters dataset (Mean F1 Score ± Standard Deviation).

Measurement	Units	Definition
Center Freq	Hz	The frequency that divides the selection into two intervals of equal energy (i.e., the 50th percentile frequency) measured on each spectrogram slice.
Freq 25%	Hz	The 25th percentile frequency (first quartile) measured on each spectrogram slice.
Freq 75%	Hz	The 75th percentile frequency (third quartile) measured on each spectrogram slice.
Freq 5%	Hz	The 5th percentile frequency measured on each spectrogram slice, indicating the lower bound of the energy distribution.
Freq 95%	Hz	The 95th percentile frequency measured on each spectrogram slice, indicating the upper bound of the energy distribution.
BW 50%	Hz	The inter-quartile range bandwidth, computed as the difference between the 75th and 25th percentile frequencies (i.e., the bandwidth containing 50% of the energy).
BW 90%	Hz	The bandwidth encompassing 90% of the signal's energy, calculated as the difference between the 95th and 5th percentile frequencies.
Peak Freq	Hz	The frequency at which the maximum power (or peak power) occurs within the selection, as observed in each spectrogram slice.
Center Time	s	The time that divides the selection into two intervals of equal energy (i.e., the median or 50th percentile time) for the signal's energy distribution.
Time 25%	s	The time by which 25% of the total energy has been accumulated within the selection.
Time 75%	s	The time by which 75% of the total energy has been accumulated within the selection.
Dur 50%	s	The duration over which the central 50% of the signal's energy is distributed, computed as the difference between the 75th and 25th percentile times.
Time 5%	s	The time by which 5% of the total energy has been accumulated within the selection.
Time 95%	s	The time by which 95% of the total energy has been accumulated within the selection.
Dur 90%	s	The duration over which 90% of the signal's energy is distributed, computed as the difference between the 95th and 5th percentile times.
Delta Freq	Hz	The difference between the upper and lower frequency limits of the selection.
Delta Time	s	The difference between the beginning and ending times of the selection.
Time 5% Rel.	–	The relative time (as a proportion of total duration) at which 5% of the signal's energy is accumulated.
Time 25% Rel.	–	The relative time at which 25% of the signal's energy is accumulated.
Center Time Rel.	–	The relative time corresponding to the median (50%) of the signal's energy distribution.
Time 75% Rel.	–	The relative time at which 75% of the signal's energy is accumulated.
Time 95% Rel.	–	The relative time at which 95% of the signal's energy is accumulated.
Peak Time Relative	–	The time at which the peak amplitude occurs, expressed as a proportion of the total selection duration.
PFC Avg Slope	Hz/ms	The average slope of the peak frequency contour over time, computed as the mean of the differences between successive peak frequencies.
PFC Max Freq	Hz	The maximum frequency reached in the peak frequency contour.
PFC Max Slope	Hz/ms	The maximum rate of change (slope) observed in the peak frequency contour.
PFC Min Freq	Hz	The minimum frequency reached in the peak frequency contour.
PFC Min Slope	Hz/ms	The minimum rate of change (slope) observed in the peak frequency contour.
PFC Num Inf Pts	–	The number of inflection points in the peak frequency contour, indicating how frequently the slope changes sign.

Table 4: Summary of acoustic measurements derived from Raven Pro 1.6. Definitions are adapted from the Raven Pro manual.

Model name	Number of parameters	Training data (hours)	Reference
BirdNET	27M	8300	Kahl et al. (2021)
Hubert-Large	1B	60960	Hsu et al. (2021)
Perch	7.8M	<10k	Ghani et al. (2023b)
Wav2vec2	317M	54000	Baevski et al. (2020)
W2v-BERT 2.0	600M	60960	Hsu et al. (2021)
Whisper-Large-v2	1.55B	680000	Radford et al. (2022)
AudioMAE	304M	5500	Huang et al. (2023)

Table 5: List of considered models for acoustic embeddings, including their size, training data, and references.