Applying machine learning to primate bioacoustics: review and perspectives

Jules Cauzinille, Benoît Favre, Ricard Marxer, Arnaud Rey

Applying machine learning to primate bioacoustics: review and perspectives

Jules Cauzinille[1, 2, 3], Benoit Favre[1, 3], Ricard Marxer[4, 3], and Arnaud Rey[2, 3]

[1]Aix-Marseille University, CNRS, LIS, Marseille, France

[2]Aix-Marseille University, CNRS, CRPN, Marseille, France

[3]Aix-Marseille University, ILCB, Marseille, France

[4]Université de Toulon, Aix-Marseille University, CNRS, LIS, Toulon France

Author Note

Jules Cauzinille https://orcid.org/0000-0002-8604-1801

Correspondence concerning this article should be addressed to Jules Cauzinille, ILCB, Aix Marseille University, 5 Av. Pasteur, 13100 Aix-en-Provence, France.

E-mail: jules.cauzinille@lis-lab.fr, Phone : +33679585636

Abstract

This paper provides a comprehensive review of the use of computational bioacoustics as well as signal and speech processing techniques in the analysis of primate vocal communication. We explore the potential implications of machine learning and deep learning methods, from the use of simple supervised algorithms to more recent self-supervised models, for processing and analyzing large datasets obtained within the emergence of passive acoustic monitoring approaches. In addition, we discuss the importance of automated primate vocalization analysis in tackling essential questions on animal communication and highlighting the role of comparative linguistics in bioacoustic research. We also examine the challenges associated with data collection and annotation and provide insights into potential solutions. Overall, this review paper runs through a set of common or innovative perspectives and applications of machine learning for primate vocal communication analysis and outlines opportunities for future research in this rapidly developing field.

Keywords: computational bioacoustics, primate vocal communication, passive acoustic monitoring, deep learning.

Applying machine learning to primate bioacoustics: review and perspectives

1 Introduction

Acoustic communication can be observed in many animal species and offers a diverse set of cues in the study of their behavior as well as a prolific insight for the monitoring of their activity. Primates are certainly no exception in this context, and the study of their vocalizations has been of great interest for the scientific community in recent years. Directly following their success in processing speech and audio, deep learning (DL) models were introduced to the field of computational bioacoustics through the ever growing availability of datasets allowed by technical advances in data recording, sharing and storage. This led to the now widely spread Passive Acoustic Monitoring (PAM) approach and, in turn, to an increasing need for efficient automated workflows in addition to hand-made annotations and analysis. This can be seen as a slight change of paradigm in the way primate vocal communication is treated and understood by researchers, which is also fairly recent and prone to evolve. In fact, deep learning methods developed for speech, audio or image processing as we currently understand them were seldom mentioned in computational bioacoustics reviews (Ganchev, 2017) until Stowell (2019, 2022). Fifteen years after the exploratory perspective paper from Zimmermann et al. (1995), one of the first studies mentioning the direct use of artificial neural networks applied to primate vocalizations was carried out by Pozzi et al. (2010). Besides that, simpler machine learning approaches developed for the processing of large unsegmented PAM datasets were not explored, to our knowledge, before the work by Kalan et al. (2015). The computational analysis of primate vocal communication systems is thus a young and rapidly growing field of study.

We hereby present a concise survey of the latest trends and approaches in machine learning applied to primate vocal communication research. In this perspective, we carefully

selected experiments, mostly published in the last three years, with additional earlier papers that we deemed interesting for a contextualized discussion. After introducing the key concepts and the general approach found in the literature, we describe three different approaches and the type of results they can provide to better understand primate vocal communication or to be used as monitoring tools. We then put forward different perspectives following the development of high-performance weakly supervised acoustic models and their potential use in primate communication research. Finally, we discuss data availability and ongoing efforts in collecting and sharing new exploitable datasets.

2 General considerations and methods

2.1 Passive Acoustic Monitoring

Directly observing primates in the scope of studying their communicative behavior in the wild can prove to be challenging depending on their species and the natural conditions encountered in their habitat. An essential limiting factor is their common tendency to flee on contact and the fact that human presence may affect their natural behavior upon direct observation (Crofoot et al., 2010). One solution to this problem is to focus on habituated or captive animals, but recent advances in technology also allowed researchers to resort to more passive behavioral data collection methods such as camera traps, drone technology or Passive Acoustic Monitoring (PAM). The emergence of high-storage and energy-efficient recording hardware and the rapid development of machine learning software is now turning PAM into a promising scientific tool to indirectly monitor either wild or captive animals. This approach can be summarized as follows: one or several, usually synchronized, microphones are placed at specific locations to record soundscapes over large spatiotemporal scales. This can be applied to any animal species displaying acoustic signals (Sugai et al., 2018) from insects such as mosquitoes (Kiskin et al., 2021) to birds (Pérez-Granados & Traba, 2021) and mammals including cetaceans

(Zimmer, 2011), deers (Enari et al., 2017) or primates (Do Nascimento et al., 2021).

A drawback of PAM experimental setups is that the collected data is inherently unimodal. However, it can be coupled with visual data from cameras, or an array of additional information (time of the day, meteorological conditions, location or expert annotations). PAM may be restricted in terms of modality, but it also presents some advantages. Regarding primate monitoring activities, for instance, it has been shown that PAM data is more valuable for the detection of primates than visual recordings from camera traps, as shown by Enari et al. (2019) for Japanese macaques (*Macaca fuscata*) and by Crunchant et al. (2020) for chimpanzees (*Pan troglodytes*). The approach may also allow researchers to rely on the extensive work, methods and software technologies developed for signal and speech processing. This makes PAM data a reliable source of answers for an array of ecological questions (Ross et al., 2023). In the perspective of primatology, computational bioacoustics can thus lead to significant discoveries regarding primate communication and vocal behavior. It allows researchers to process and filter large collections of sounds by relying on so-called machine learning methods which help them automatically analyze and interpret acoustic signals. Applied to primate vocalizations, these methods show impressive results in an array of essential tasks such as the denoising of recordings, the selection and segmentation of said recordings to extract meaningful or primate-only vocalizations from lengthy recordings, the detection and classification of species, individuals or specific types of calls, etc. (see Figure 1). We hereby present the different trends, approaches and benefits related to the application of machine learning to PAM datasets. To get a deeper understanding at the functioning of machine and deep learning algorithms, one may refer to specialized reviews such as Pichler and Hartig (2023).

2.2 Machine learning for bioacoustics

Machine learning is the implementation of artificial intelligence through algorithms and computer models trained to autonomously make predictions from data. This process

may be summarized as a search for parameters in mathematical functions that minimize the difference between predicted outcomes and actual (human) observations, enabling systems to generalize and make accurate decisions on new, unseen data. Deep learning is a subset of machine learning involving the use of artificial neural networks characterized by their large number of parameters interacting with data in hierarchical, complex and unforeseeable ways. For a more comprehensive exploration of machine learning, readers are encouraged to refer to James et al. (2023), which provides both introductory and in-depth coverage of the topic. Deep learning (DL) is thus particularly well-suited to handle large datasets containing complex patterns and unstructured information, such as those encountered in PAM datasets. For bioacoustics, machine and deep learning can be seen as a way to automatize, reproduce or enhance human annotations with computer models. These are built through the selection and development of algorithms adapted to the task at hand, and always rely on annotated data for their training and evaluation. Training a machine learning model consists in optimizing a system to reproduce a labeling process through trial and error by maximizing correct predictions and minimizing mistakes. Its performance is thus heavily dependent on the quality of the training data and on its ability to generalize the labeling procedure to previously unseen data. Experiments involving the use of machine and DL models trained on acoustic recordings of primates often follow standardized workflows. These, in turn, resemble methods developed for the study of other species and can show strong parallels with computational linguistics and speech processing research. As stated by Stowell (2022), "classification is indeed the main use of deep learning seen in computational bioacoustics." Although the task of classifying sound can be divided in various sub-tasks ranging from segmentation to labeling (see Section 3), it usually implies training a computational model on a given set of data for a specific task and evaluating its ability to perform the task when presented a different subset of the data. This common approach, which can be found in most of the experiments we hereby review, unfolds as follows (see Figure 2 for a visual description):

1. Acoustic data is collected from either captive or wild settings. This implies technical subtleties in terms of location of the recordings, number of acoustic sensors deployed, specific recording configurations such as sampling rate or frequency range, and the amount of data which can be obtained.

2. The data is then turned into processable inputs. It may be segmented into short clips, specifically tailored for the use of some DL models. It may also be transformed through feature extraction to facilitate its processing by the computer models (see Section 2.2).

3. Depending on the task at hand, annotation is needed to provide target labels which will be learned and predicted by the model. These can include several classes such as species, individual identities or call types, but also binary labels such as presence and absence of vocalizations, as well as dimensional labels.

4. The model architecture mostly depends on the task it needs to perform but also on the type of data it will be trained on. For instance, acoustic data in the form of spectrograms is efficiently processed by Convolutional Neural Networks (CNN). A CNN is an artificial neural network constructed as a stack of layers (the so-called *convolutional* filters) which efficiently extract meaningful information from audio or image inputs by recognizing patterns in the data. CNNs are by far the most popular choice in computational bioacoustics. With the advent of new DL models developed in the scope of processing longer sequences of speech or audio, some experiments now rely on state-of-the-art architectures such as Recurrent Neural Networks (RNN), or the popular transformer networks (Lin et al., 2022) which leverage attention mechanisms and further exploit longer and variable-rate relations in the data. Much simpler machine learning models such as Multi-Layer Perceptrons (MLP), Hidden Markov Models (HMM), Support Vector Machines (SVM) or clustering algorithms (see Pichler and Hartig (2023) for a typology of existing algorithms) can also show

interesting results. These simpler models may also be used as baselines, a voluntarily simple reference model used for comparison purposes.

5. The dataset is divided into *train*, *development* and *test* sets. The train set will be used during training to present examples of target labels associated with input data to the model, from which it will learn to extract cues and informative features. The development set is also used during training to make design choices that cannot be optimized using the machine learning method. It provides information in order to select from different models or tune functional aspects of the learning algorithm itself (often referred to as hyper-parameters). Finally, the test set, unseen by the model during training, will be used at the inference step to evaluate said performances. It is usually taken from a separate pool of data (different microphone, location or vocalizing individual) to ensure true generalisability of the model and avoid biased evaluations.

6. The evaluation of the model requires the selection of appropriate metrics depending on the task at hand. These are chosen to be as informative as possible in the context of the experiment and must reflect the prediction performance but also potential flaws in terms of false positives or false negatives. It must be chosen according to the way labels are balanced in the dataset as well as its size. To evaluate the automatic segmentation of acoustic data over time, for instance, authors will often rely on accuracy (the number of correctly predicted segments divided by the total number of segments). They may also use the F1-score to account for the balance between false negatives and false positives, which is ignored by the accuracy metric.

A great majority of computational bioacoustics experiments rely on spectral representations of sound prior to their automatic processing by machine learning models or their manual annotation. These spectral representations, relying on the short-time Fourier transform, encode the temporal evolution of acoustic energy over a range of frequencies in

a sound signal. Although several types of spectral transformations may be found in the literature, some can be turned into images to ease the manual analysis of acoustic material (as in the spectrogram from Figure 4). This type of sound representation is often used to speed up the manual annotation of bioacoustic data by human annotators. It can also be directly employed as the input of vision-based DL models which can reach high performances on a variety of tasks by processing sound as images. No particular representation of sound has been proven to work best across all species and tasks, and the use of a given method must be carefully justified because of its important implications on the performances of a computational model. Within spectral representations, a first distinction can be made between linear and logarithmic spectrograms, the latter being designed to mimic the way human ears process sound by emphasizing discriminability of lower frequencies and de-emphasizing it in higher ones. In this perspective, bioacoustic researchers often rely on mel-spectrograms, which tend to show promising results when used as features for animal vocalizations processing. Nonetheless, and despite their popularity, spectral representations are not always a preferred solution and other acoustic representation methods exist. As can be seen in Kiskin et al. (2020), wavelets can also show great benefits compared to the short-time Fourier transform, due to their ability to capture both fine details and broad trends in acoustic data. Although a less conventional solution, the authors show the advantages of wavelets when facing weak and noisy signals (such as mosquito sounds) and their ability to perform better across datasets compared to spectral solutions (with bird species classification).

With the advent of more powerful computational technologies, the use of the raw waveform also presents an array of advantages compared to spectral-based representations. Although directly using a waveform as input usually requires larger datasets and more computing power to efficiently train DL models, the approach allows researchers to bypass yet another manual pre-processing step. This also means a model can learn to extract any informative cues without the risk of loosing information through spectral transformations of

its training dataset. Additionally, training classifiers directly on the waveform can greatly simplify a classification pipeline, going from the acoustic data to the bioacoustic predictions in a straightforward manner, usually referred to as an "end-to-end" approach. Although it was not used on primate vocalizations to our knowledge, end-to-end approaches are gaining popularity in bioacoustics, an sound processing in general, with successful applications such as in Bravo Sanchez et al. (2021) on birds and Xie, Hu, et al. (2021) on frogs.

Finally, as we will further discuss in Section 4, acoustic representations may be extracted by pre-trained upstream models (see Figure 3). This is referred to as *pre-trained representation learning*, an increasingly popular solution in recent speech and audio processing research. The approach relies on the pre-training of large generalistic foundation models to enhance the performance of smaller task-specific ones. Once the pre-trained representations are learned and extracted, they may be used as traditional features containing useful information for an array of tasks.

A second optional step, directly preceding feature extraction, is *signal enhancement*, or *denoising*, a process which consists in filtering out non-informational signal from raw data. This signal processing method is quite popular in bioacoustics where clear recording conditions are rarely encountered (Xie, Colonna, & Zhang, 2021). The amount of noise which can affect the performances of computational models in processing primate vocalizations greatly depends on the recording location or microphone sensitivity, and may stem from an array of acoustic sources, from anthropogenic noise (vehicles, speech...) and natural soundscapes (rain, wind...) to other species or conspecifics vocalizations. Although denoising can be an essential tool, it is not always beneficial, as it may deprive the signal from essential information which could potentially be extracted by a computational model. The general approach consists in using simple fixed signal processing tools to perform the so-called signal enhancement directly on a spectrogram. Such tools have been extensively studied and engineered and are widely available through public softwares such as **noisereduce**, a spectral gating algorithm developed by Sainburg et al. (2020). A more

refined option from which computational bioacoustics could greatly benefit is DL-based denoising, a popular area of research in speech processing (Germain et al., 2019).

3 Tasks and applications

We hereby describe three main categories of tasks which can be tackled through the use of machine learning for primate vocalization analysis. For an overview of these categories and their applications in primate bioacoustics, see Figure 1.

3.1 Detection and segmentation

The most practical application of machine learning in bioacoustics, when facing large unlabeled recordings of natural soundscapes, is the detection of animal vocalizations among ambient noise. As we previously mentioned when introducing the PAM approach, the ever-increasing storage and battery life capabilities of microphones may result in recordings lasting several hours or days. Primates, however, are not constantly vocalizing and their calls usually span specific segments of time which need to be identified and extracted for their subsequent analysis. The manual segmentation of recordings (i.e. annotating start and end times of primate calls among a continuous audio clip) is an essential step in processing PAM data. An efficient way to carry out this segmentation is to directly inspect spectrograms of the recording in a specialized software such as PRAAT[1] or Raven Pro[2]. Although it results in precise annotations, manual segmentation may prove to be quite time-consuming depending on the length of the audio files and the nature of the recorded vocalizations in terms of frequency ranges, unit-rates, distances to the microphone and amounts of background noise. Automatic detection and segmentation were proposed as

[1] Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.17, retrieved 10 September 2023 from http://www.praat.org/

[2] K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology. (2023). Raven Pro: Interactive Sound Analysis Software (Version 1.6.4) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from https://ravensoundsoftware.com/

an answer to this issue. It may take at least three different forms:

- binary detection: the machine learning algorithm is given a segment of audio as input and outputs the probability of this segment containing a call. This may be referred to as "occupancy" or "presence" prediction.

- time-wise segmentation: the task can still be developed as a binary one but it results in a more fine-grained annotation of the input file with start and end time-codes of each call. This is often solved by making an occupancy prediction in short windows (10 or 50 ms) and merging consecutive positive decisions into a single segment.

- time and frequency-wise detection: directly inspired by image object detection, this task usually implies the use of spectrograms. The model is constructed as an object detector and outputs time and frequency boundaries of the target call as in Figure 4. To our knowledge, this approach was never explored for primate vocalizations and is scarcely applied to other species as well, although it could be used to identify various simultaneously vocalizing species in single segments.

Automatic segmentation is undoubtedly one of the most studied aspects of automatic detection in bioacoustics. The following examples exclusively focus on detecting gibbons, but similar directions are being explored on other primate species (Anders et al., 2021; Bonafos et al., 2023). Recently, approaches involving the use of deep learning are being predominantly adopted in automatic audio recognition and tend to replace the use of hand-crafted features and simpler machine learning algorithms. This scientific trend is widely adopted across bioacoustics, specifically through the use of CNN-based solutions from spectral inputs, as can be seen in the evolution of the DCASE challenge over the years (Mesaros et al., 2017, 2019). Although rarely explicitly compared with more simple statistical approaches, CNNs present an array of advantages in these tasks. Their primary benefit lies in their ability to generalize predictions across varying recording conditions. They also allow efficiently tackling tasks with noisy and unbalanced annotated datasets of

limited size (Anders et al., 2021), as is often the case with PAM recordings of primates in the wild. As we will see, they may also be coupled with so-called RNNs to account for the sequentiality of primates' vocalisations. This makes them especially effective at detecting primates with temporally dynamic calls, as is the case with gibbons. Finally, CNNs being a very popular option in the deep learning community, extensive research, publicly available resources and off-the-shelf solutions can be accessed with little expertise to develop fast and efficient models. All these advantages made CNNs a go-to solution in bioacoustic detection (Stowell, 2022), progressively replacing the use of simpler statistical algorithms. As we will see throughout this review, machine learning for primate bioacoustics follows a similar trend. Nonetheless, the efficiency and advantages of deep learning solutions come with some drawbacks, specifically in the interpretability of a model's predictions as well as in their need for higher computational ressources and larger datasets (see Figure 5). In recent bioacoustics papers, the specific reasons for choosing large deep learning models over lightweight statistical solutions are rarely explicited. However, they are often implicitly shown through comparison with simpler baseline performances.

An illustrative example in this perspective is a model developed by Dufourq et al. (2021) applied to the highly endangered Hainan Gibbons (*Nomascus hainanus*) from the *Bawangling National Nature Reserve*. The proposed model is a common one in bioacoustic event detection as it relies on the popular CNN architecture. In this case, the model is designed to differentiate between two classes of sounds : *non-primate background noise* and *primate vocalizations*. It is trained on mel-spectrogram representations of short sound segments which were previously labeled as such. The authors additionally resort to data augmentation, which consists in increasing the size and diversity of a dataset by slightly modifying it to create additional synthetic data. This method permits rendering the models robust to certain transformations that we know should not affect the system's prediction. Here, each segment is shifted in time to double the size of the initial dataset. The authors evaluate two types of architectures, namely a one-dimensional CNN leveraging

temporal patterns, and a two-dimensional one which captures frequency as well as time from the input spectrograms. One last step consists in post-processing the model predictions by removing unrealistic detections (such as isolated or very short calls). The 2-D CNN with data-augmentation paired with the post-processing step achieves 99.37% accuracy (compared to 97.60% without post-processing and 92.32% without data-augmentation as well)/. This shows that the CNN approach can be highly efficient in facilitating the segmentation of large PAM recordings of Gibbons, with an eight hours long test recording taking six minutes on average to be processed by the model. In addition to its high performance, this automated procedure also provides exhaustive quantitative information about Hainan gibbon's vocal behavior, including their preferred vocalization times, the amount of calls they produce in a day and their geographical distribution over the study site. By coupling the detections with additional metadata such as meteorological information and environmental parameters, this approach could lead to many more interesting observations.

As is common in machine learning experiments, the work by Dufourq et al. (2021) constitutes a baseline performance which was promptly improved by Ruan et al. (2022) with a slightly different architecture. This "baseline" can be considered as a performance milestone aimed at being improved upon and was included in one of the only bioacoustics benchmarks available to date: *the Benchmark of Animal Sounds* by Hagiwara et al. (2023). Ruan et al. (2022) approach relies on deep learning solutions, namely Residual Networks (a former state-of-the-art model, known for its high performances in image classification), SpecAugment (the random masking and warping of portions of the input spectrogram to improve the generalization ability of the model) and label smoothing (a method with similar results based on the addition of noise to the label distribution during training). These methods were originally developed for speech and image classification and allow the authors to propose *BPDnet*, a model with high performance on Hainan Gibbon's presence detection carried out on the same dataset as used by Dufourq et al. (2021) without the

need for manual intervention and post-processing. When compared with the baseline experimental setup, this new model improves the F1-score by 0.16 without post-processing and by 0.09 with post-processing.

The approach selected by Dufourq et al. (2021) corresponds to the most recommended and wide-spread approach for the segmentation of bioacoustic data. It relies on years of research into the CNN architecture and tackles major limitations through post-processing and data-augmentation in addition to the use of traditional spectral representations. Comparatively, Y. Wang et al. (2022) work on the same dataset with a rather innovative perspective relying on more complex state-of-the-art models. They implemented two solutions: a CNN stacked with a Hidden Markov Model (HMM) and a Convolutional Recurrent Neural Network (CRNN). The HMM architecture works as a post-processing step and allows for a correction of the CNN decision from contextual information of neighboring segments. The CRNN, in contrast, outputs decisions from sequential information rather than from a fixed segment. It still relies on a CNN for feature extraction which is subsequently passed to Gated Recurrent Units accounting for the temporal information. These particular types of models have been shown to be more effective at modeling long-term dependencies in sequential data (like sound), while also being computationally more efficient than traditional RNNs. Here, the choice of this architecture is motivated by the sequential nature of gibbon vocalizations which are known to produce varying sequences of notes combined into phrases. The authors show that the CRNN architecture improves performance compared to the CNN-HMM and that it is resistant to low Sound to Noise Ratio (SNR), a metric used to account for the amount of background noise in a given audio segment. Although the results are not comparable with previously mentioned experiments because of distinct evaluation metrics and dataset processing, CRNNs are a viable option for automatically processing large PAM recordings of gibbons and their resistance to noise can be seen as a major asset in the bioacoustic context.

Following both previously described frameworks, Tzirakis et al. (2020) conducted somewhat similar experiments on large recordings of Müller gibbons (*Hylobates muelleri*) from Malaysian Borneo. The authors built a publicly available model consisting of a 2-D CNN followed by a RNN which also capture longer-term temporal dynamics of the input signal. To validate their approach, the proposed architecture is compared to two publicly available toolkits for audio representation and analysis: End2You (a simpler yet similar model comprised of CNNs followed by Gated Recurrent Units) and openXBOW (based on the bag of words approach from computational linguistics and several machine learning models, including the Random Forest algorithm). The dataset was processed as a collection of positive (gibbon's presence) and negative (background noise) audio clips. The results indicate that the author's model reaches an Unweighted Average Recall of 93.3% on the test set compared to 84.8% for the best openXBOW model.

Although we only discussed three experiments on a single taxon, we can see that state-of-the-art segmentation of primate vocalization datasets mostly resort to closely related approaches and tend to show high performance if sufficient annotated data is provided. CNN solutions can be found successful for an array of other primates and animals: Stowell (2022) surveyed 83 such experiments for a variety of tasks including segmentation; and RNNs seem to be a viable option in further improving their performance. This approach also shows the benefit of potentially relying on off-the-shelf models made available by researchers in other fields. The limitation here is the availability of annotated data itself and the potential complexity of usage of such publicly available models for non-specialist practitioners.

3.2 Identification and density estimation

We have seen that segmentation and detection tasks relying on state-of-the-art deep learning architectures can be very effective for primate species shown to communicate vocally like gibbons, even in their noisy natural environment. Although the obtained

results and performances can be used in quantitative analysis of their vocal behavior and are essential for qualitative studies of the segmented calls, they lack in conservation value. In this perspective, identifying vocalizing individuals to estimate primate population density from their vocalizations and studying individual vocal signatures are both complex but potentially significant tasks. Few studies can be found exploring automatic density estimation relying on DL models, although human based detection seems to be a good option for this task. Using humans as acoustic detectors has indeed been proven successful in estimating the density of yellow-cheeked gibbons (*Nomascus gabriellae*) in Cambodia by Kidney et al. (2016). This means that acoustic data contains enough information to develop similar experiments with computational models.

In a similar perspective, several studies were conducted in developing automatic classifiers for caller identification. These often rely on recordings in captivity, during mark-recapture events or by focal recordings of individuals to allow for an easier extraction of the caller identity and the constitution of an annotated dataset. In fact, obtaining identity annotations from PAM data is a difficult process due to low control possibilities over the recordings. In captivity however, some solutions were explored by Bayestehtashk et al. (2014) who recorded groups of captive rhesus macaques (*Macaca mulatta*) using individual collars. The authors provide an interesting semi-automatic pipeline for the constitution of an ID-labeled corpus. To process the obtained data from multiple collars, they construct a segmentation model based on manually designed acoustic features often used in music and speech from the OpenSmile toolkit (Eyben et al., 2010). Their best performing model is a Support Vector Machine (SVM). SVMs are machine learning algorithms designed to separate samples of different class labels in the feature space. They are light machine learning models, easy to train, and show good performances on simple tasks. In this case, the SVM architecture reaches 88.9% accuracy on a manually segmented subset of the data. Once segmented, the obtained set of vocalizations is to be matched to the recordings from each collar in order to detect the one which most likely emitted a

specific vocalization. The authors rely on Dynamic Time Warping (Müller, 2007) to compute the acoustic similarity of each segment and show good results compared to manually aligned data. The obtained dataset can then be used to train classifiers for the automatic identification of individual monkeys.

In an opposite approach, highly territorial primates like Northern grey gibbons (*Hylobitae funereus*) can provide interesting datasets to perform this type of task on wild PAM recordings. By placing microphones inside individual group territories, Clink et al. (2017) successfully identified the acoustic parameters contributing to individuality in female's great calls. The authors manually extracted acoustic features from their PAM dataset and computed a Mahalanobis acoustic distance measure between pairs of vocalizations. They were able to discriminate between pairs of 33 females with a 95.7% accuracy using linear Discriminant Function Analysis, a method consisting in searching for linear combinations of the extracted features to separate the different individuals.

Machine learning algorithms such as Discriminant Function Analysis do not rely on deep learning, contrary to the CNNs and RNNs discussed in the previous section. For vocal signature classification, they seem to be a preferred approach with the advantage of demanding less computational power and data all the while yielding competitive results. We want to stress that resorting to deep learning solutions in bioacoustics is not always a preferred approach, especially when facing scarce annotated data, as is the case for identification of primate voice prints.

With a similar dataset of Northern grey gibbons and an approach involving Support Vector Machines and Mel-Frequency Cepstral Coefficients (MFCC), Lakdari et al. (2024) reached high performance on classifying female great calls by recognizing their emitter from a pool of 12 individuals. They further examined the performance of their approach by recording the calls in playback at varying distances to account for its resilience on low sound to noise ratios. They find that MFCCs are outperforming other feature extraction methods, namely acoustic indices or pre-trained DL models, when calls are recorded at

larger distances.

Another example of such solutions can be found in a study by Fedurek et al. (2016). The authors examine chimpanzees (*Pan troglodytes schweinfurthii*) pant-hoots and attempt at identifying the type of information acoustically embedded in them. Their experiment also relies on Support Vector Machines trained on MFCCs. They find that all four phases of the pant-hoot (introduction, build-up, climax and let-down) are associated with a variety of information, including individual identity, which is more specifically encoded in the introduction and climax. Despite these promising results, there seems to be a lack in the implementation of state-of-the-art DL models for vocal signature classification in primates. We have seen that this shortfall can be explained by difficulties in annotating data accordingly. It may also be explained by the more consistent amount of work put into identifying primates from visual data with face recognition (Guo et al., 2020; Schofield et al., 2019). Yet, few experiments relying on complex and innovative DL architectures from sound show promising results. We have mentioned Lakdari et al. (2024) who compared MFCCs feature extraction with deep learning models pre-trained on birds, speech or general sound. Leroux et al. (2021) also introduce transfer learning from DL models pre-trained on speech for chimpanzees voice print recognition. Both approaches will be further discussed in Section 4.

Nonetheless, a parallel task involving multi-label classification and voice prints with interesting machine learning solutions is primate species identification. As we have seen, most bioacoustic studies carried out on primates focus on single datasets from one species of interest. However, wild environments may host various cohabiting species which often end up overlapping in single PAM recordings. A first interesting study in this regard was conducted by Mielke and Zuberbühler (2013) with a combination of classification tasks for species, call type and caller identification. It relies on a MLP trained on a dataset of Stuhlman's blue monkey (*Cercopithecus mitis stuhlmanni*) vocalizations. This particular species allows for identity labeling because each group hosts a single male which also

happens to be the only producer of "pyow" calls. Additionally, other species' calls found in the same environment were added for the species discrimination task (olive baboons, *Papio anubis*; redtail monkeys, *Cercopithecus ascanius schmidti*, and guereza colobus monkeys, *Colobus guereza occidentalis*). After extracting MFCCs and training various MLPs with distinct hyperparameters, male identity classification resulted in 73% accuracy in average and species recognition resulted in 96% accuracy for the four classes. Despite the promising results, we must point out that substantial manual work had to be allocated for the pre-processing, segmentation and identification of the calls prior to the automated classification. Other early experiments by Kalan et al. (2015) and Heinicke et al. (2015) also showed interesting approaches with simpler algorithms including SVMs and Gaussian Mixture Models. Both papers focus on the identification of chimpanzees (*Pan troglodytes verus*), diana monkeys (*Cercopithecus diana*), red colobus (*Procolobus badius*) and king colobus (*Colobus polykomos*). Both SVMs and Gaussian Mixture Models were trained on MFCCs and other spectral information extracted PAM recordings. The algorithms show relatively low results with less than 5% of detected segments being true-positives for the best model.

A more recent approach involving Kernel Extreme Learning Machine was adopted by Zwerts et al. (2021). This particular type of model is a supervised learning algorithm using a kernel function to map input data into a high-dimensional space and allows for the learning of complex and non-linear relationships between input features and output targets. It was trained on MFCC representations of vocalizations from captive chimpanzees (*Pan troglodytes*), mandrills (*Mandrillus sphinx*), red-capped mangabeys (*Cercocebus torquatus*) and a mixed group of guenons (*Cercopithecus sp.*), with an additional class of background noise. The performances of the model are above chance (25% for the four species) with 76.7% accuracy in a four class setup and 69.7% accuracy with the addition of the noise class, but stay relatively low compared to more recent approaches.

As we have mentioned with other experiments, the publication of this new dataset

and baseline model may be seen as a benchmark which was promptly integrated as part of the INTERSPEECH 2021 Computational Paralinguistics Challenge (Schuller et al., 2021). In an attempt to tackle the species identification problem with state-of-the-art architectures (similar to what we have seen in Section 3.1), Pellegrini (2021) compared several DL models including CNNs, MobileNet and ResNets. They also revolved to data augmentation methods like SpecAugment and MixUp (another technique for data augmentation relying on the blending of pairs of training examples). The main difference between each model lies in the definition of their convolutional blocks. The first two models are standard CNN architectures with 6 and 10 layers respectively. The MobileNetV1 model relies on *depthwise separable convolutions* to reduce computational costs and gain efficiency, usually in the scope of being used in mobile and embedded devices. Finally, two CNN ResNet models make use of *residual connections*, allowing for a deeper network architecture to be trained without suffering from the vanishing gradient problem which affects models with many stacked layers, such as standard CNNs. The results show good improvement compared to Zwerts et al. (2021) baseline with an unweighted average recall of 92.5% achieved by the 10 layer CNN, closely followed by the large ResNet model. In this case, the 10 layer CNN is preferable to ResNet as it achieves better performance with a much smaller model size. The authors also note that the most common confusion made by their models regards the background noise class versus the primate vocalizations one. This confirms the importance of ongoing efforts in resolving "low level" tasks, such as the identification of primate vocalizations among natural noise.

3.3 Vocal repertoires and clustering

Despite this, "high level" tasks can be found in computational bioacoustics literature, with many relying on machine and deep learning-based solutions. One such task with great scientific value for primatology is the discovery or the classification of call types (i.e., the categories of calls produced by a species). This task may be carried out through

different approaches, including supervised classification (each class corresponding to a predefined call type) and unsupervised clustering (the grouping of similar acoustic objects into undefined call type categories). Each of these has been explored for various primate species using an array of machine learning algorithms. The relative success of one approach, especially in unsupervised contexts, is often seen as a form of validation of predefined expert descriptions of a species vocal repertoire. Call type classification thus serves the purpose of automatically processing large amounts of data while potentially questioning human bias in the definition of vocal repertoires and fostering replicable results across studies.

We hereby refer to "unsupervised" approaches to account for all experiments involving the training of a model with little to no reliance on expert labels and annotations. In the context of call type discovery, for instance, this means that a clustering model is trained on unlabeled acoustic samples and should discover its own typography of calls in order to categorize them. The related "semi-supervised" approach is one where a limited amount of information is given to the model prior to clustering, such as the number of categories to be discovered. Evaluating the results of such clustering approaches is a highly debated topic in computer science, as no single solution can objectively quantify the validity of a set of clusters compared to another. Von Luxburg et al. (2012) review the different issues related to clustering and its evaluation. Although not centered around bioacoustics, the paper draws inherent limits specific to the idea of automatic clustering:

- Evaluating clustering results is not problem-independent and must be related to the end-user intent and their scientific scope.
- As is always the case with high-dimensional data, selecting the features on which clustering will be carried out can greatly modify the output typology.
- A given clustering output can be found to be qualitatively reasonable for a specific research question but may be meaningless to others.

- Computing internal clustering quality scores (centrality of the clusters, likelihood scores, silouhette values, etc.) can be informative at the algorithm level but does not provide an objective and domain-specific evaluation of the results.
- Comparing the results of unsupervised clustering with predefined categories of calls should not be seen as undisputable proof of the validity of said "expert" categories, as both may be biased in different ways.

Clustering primate call-types should thus be seen as an exploratory approach, and experiments using it as a confirmatory solution to predefined human vocal repertoires should be taken with care.

To our knowledge, the first paper mentioning the use of Artificial Neural Networks for primate vocalization analysis, over and above the preliminary work of Zimmermann et al. (1995), is, in fact, aimed at call type classification on black lemurs (*Eulemur macaco*). Pozzi et al. (2010) compare the performances of supervised neural networks, statistical models and clustering algorithms in recognizing a set of predefined call types. They show that basic artificial neural networks trained to classify seven call types from which spectral (F0 and formants) and temporal (duration) acoustic features were extracted, can show high performances with a general accuracy of approximately 94%. Statistical analysis with Discriminant Function Analysis and K-means clustering showed slightly lower performances, with large disparities in classification accuracies for some call-types, potentially due to the unbalanced classes context. The authors thus give a first example of some advantages presented by deep learning methods compared to statistical approaches. They mention their ability to handle noisy recordings, to generalize human annotations to unseen data, and the reusability of a model's weights once it has been successfully trained. The authors also mention a set of limitations that can still be found in such experiments. These include the over-fitting problem where neural networks learn dataset-specific information related to individuals or to their sex rather than universal cues generalizable to the entire species. They also mention the problem of biases in the manual annotation of

datasets, which may greatly affect the evaluation of clustering.

As previously mentioned, statistical and machine learning algorithms that do not involve neural networks can show promising results in the analysis of call types. Turesson et al. (2016) investigated the use of seven different such models in addition to DL ones for the categorization of common marmoset (*Callithrix jacchus*) calls. The automatic identification of call types appears as an essential tool for marmosets, as they produce large amounts of characteristically complex and overlapped vocalizations on which annotation is rather tedious and time-consuming. The authors collected a dataset from captive monkeys' recordings with approximately 30 examples for each of the 11 call types investigated. They chose linear predictive coding as a feature extractor. This technique, traditionally used to encode the timbre of human voice signals in speech compression for telephony, consists of modeling the spectral envelope of sound samples as the weighted sum of previous samples from a given acoustic sample. The extracted features are used as an input to train several classifiers, namely an Optimum-Path Forest, an MLP, an SVM, a k-Nearest Neighbors clustering algorithm, etc. In addition, various proportions of the training set were tested to understand performance trade-offs relative to training size. The SVM, k-Nearest Neighbors and Optimum-Path Forest were found to be the best performing algorithms in both the smallest and largest training set sizes, with Optimum-Path Forest being parameter-free and requiring less computational resources. This suggests that simple statistical algorithms can show high performances when facing limited amounts of clean data, although we could argue that larger training datasets would increase performances in general and might be in favor of other more complex deep learning-based models.

The task of automatically discriminating between different types of calls may prove useful in quickly processing large amounts of data but it can also be used to infer new properties of primate communicative systems, especially when tackled with clustering. Erb et al. (2023) adapted different models to the classification of Bornean orangutans (*Pongo pygmaeus wurmbii*) pulse-types to investigate problematic elements in the specie's

predefined vocal repertoire. They collected a dataset of focal recordings from 23 individual males. Comparing human annotations and the unsupervised predictions of an SVM as well as soft and hard clustering algorithms, they showed that a set of six pulse-types gives rather poor results in terms of inter-annotator agreement as well as automatic predictions in this specific experimental setup. This negative result allows them to propose a new repertoire comprised of only three pulse-types, which shows higher classification accuracy and reproducibility. Finally, they highlight the importance of graded categories of signals, in opposition to strictly separated call types, in the typology of orangutan call types. This type of experiment shows how automatic clustering and classification, although not sufficient to objectively refute a predefined vocal repertoire, can still be used as an exploratory tool to identify its potential biases. Similarly, Wadewitz et al. (2015) question the discreetness of chacma baboons' (*Papio ursinus*) call type categories by investigating the results of a "fuzz" clustering algorithm. They argue that labeling primate vocal repertoires as being either fully discrete or fully graded may be considered an oversimplification. Hard clustering (found in K-means algorithms, for example) assigns each call to a single cluster or call-type. Fuzzy clustering (such as the C-means algorithm), allows separating different classes of calls in a gradual manner rather than a sharp one. Each call is given a membership value, ranging from 0 to 1, assigning it to each cluster. Intermediate membership values characterize calls ambiguously pertaining to multiple clusters. The authors find that, although hard K-mean clustering shows good alignment with predefined human labeled call types on chacma baboons, fuzzy clustering gives additional information regarding the atypicality of some of the species' calls. Again, when used as an exploratory tool, fuzzy clustering may question unforseen biases in the constitution of a species vocal repertoire.

We have seen that call-type classification and clustering can be carried out for different reasons with different algorithms, but the outcome of call-type clustering also greatly depends on the choice of acoustic features it is built on. A preferred approach is the

dimensionality reduction of spectral features, as in an experiment proposed by Sainburg et al. (2020) relying on the Uniform Manifold Approximation and Projection algorithm (UMAP), among others. This popular method has the benefits of being relatively efficient in discovering significant sound features and can result in informative visual representations of clusters. UMAP is one of the many dimensionality reduction algorithms used to classify vocalizations by taking high-dimensional features (like deep, spectral, or acoustic features) and mapping them to a lower-dimensional space while maintaining the underlying distances between different sounds. By reducing the dimensionality of spectrograms with said algorithm, the authors show the implications of automatic unsupervised clustering in a variety of topics related to primate vocalization analysis, including the discreteness of macaques vocal signatures (see Figure 6) or the apparent continuity of gibbon (*Hylobates sp.*) syllables (see Figure 7).

This approach should both be extended to other primates and explored through the use of different algorithms and input features. In fact, UMAP may struggle with capturing the global structure of acoustic data, particularly when dealing with complex and highly varied vocalizations, as is the case for primates. In addition, UMAP's performance may be strongly affected by the presence of outliers and noise in the data (as is often the case with PAM recordings), potentially leading to distorted representations and the absence of interpretable results.

As an alternative, Best et al. (2023) were inspired by deep representation learning and extended this methodological framework to an array of animal species, showing once again the great flexibility of the approach in validating vocal repertoires and alleviating their manual annotation. Contrary to the latter experiment, the features extracted prior to clustering are derived from a self-supervised CNN-based auto-encoder trained to *encode* informative components of a spectrogram through a bottleneck approach in order to subsequently *decode* input signals with minimal loss of information. This method, inspired by speech and image processing techniques, yields significant results in the unsupervised

clustering of call-types for a variety of taxa ranging from birds to marine mammals. The authors showed the benefits of working with UMAP and clustering algorithms based on deep representations of sound rather than spectral or handcrafted features. As was previously mentioned, the evaluation of clustering results through their comparison with pre-defined expert categories is rather exploratory and does not prove the objective validity of a given algorithm or feature extraction method (nore of said expert categories). Nevertheless, clustering solutions can be compared in terms of their alignment with human typologies to provide interesting insights on their ability to extract information deemed important by expert labelers. The authors thus demonstrate the ability of models pre-trained on non-bioacoustic datasets to extract features that encode sufficient information for an efficient unsupervised clustering of call-types. Although their Autoencoder architecture yields better results in most datasets, models like wav2vec (Schneider et al., 2019) and OpenL3 (Cramer et al., 2019), a model trained on audio/video correspondence from YouTube data, also show comparable agreement scores between found clusters and expert labels. This innovative paradigm, i.e., using large DL models pre-trained on non-bioacoustic data for bioacoustic tasks, seems to be an increasingly popular one in a variety of experiments, although it stays quite seldom explored for primate vocalizations. In the next section, we will discuss recent papers making use of this approach and see the potential implications and perspectives it may offer for the study of primate vocal communication.

4 Transfer learning and promising approaches

Fairly recently, the advent of so-called “Pre-Trained Models” (PTM) has undoubtedly revolutionized the use of deep learning for text, image and speech processing. For Natural Language Processing, PTMs such as BERT (Devlin et al., 2019) or GPT (Radford et al., 2018) have become a milestone in the field of artificial intelligence with large language models like ChatGPT showing impressive applications way beyond

computer science research by leveraging an ever-increasing access to high computational power and large amounts of data.

These models often rely on a self-supervised learning pre-training step, consisting in storing and extracting information from massive datasets which can then be reemployed in a variety of downstream tasks with great performance benefits compared to more traditional supervised approaches (X. Liu et al., 2023; Mohamed et al., 2022). In the acoustic domain, self-supervised models have also shown impressive capabilities in generalizing knowledge with performance gains across a wide range of domains. The typical approach in this regard is to pre-train a model on large unannotated datasets (which should be relatively close in nature to the target domain data) and to use the learned representation for downstream tasks on smaller manually labeled datasets. This process involves transfer learning, i.e., relying on the knowledge learned during the pre-training task for a new, potentially different, downstream task (see Figure 3). This approach was successfully carried out in a variety of domains including music or biomedical signal processing (Banville et al., 2021; Wu et al., 2021).

Bioacoustic tasks and use-cases are no exception here. Researchers in animal vocal communication progressively turned to this new paradigm in recent years by adapting methods initially developed for speech and sound processing to the analysis of acoustic data produced by animals. When it comes to primates, however, the success of self-supervised and transfer-learning approaches is yet to be confirmed and widely adopted. However, several such experiments can be found relying on a variety of parallel approaches, each showing its own benefits. We hereby discuss three main solutions that arise from using transfer learning for automatic primate vocalizations processing.

4.1 Retraining

In the field of machine learning, transfer learning refers to the *pre-training* of a model on a given dataset or task and the development of a downstream model aimed at

performing a specific downstream task on a different (annotated) dataset. In bioacoustics, however, a slightly different popular approach, partly relying on knowledge transfer, is the *retraining* of a model initially developed for similar tasks but on a different species. This retraining approach is not to be confused with the pre-training of a single foundation model in which previously learned weights may be reused for several applications. For primate vocalizations, a good example of knowledge transfer through retraining is the work by Romero-Mujalli et al. (2021). Here, the authors show the benefits of retraining the ultrasonic vocalization detector model *DeepSqueak* (originally developed for rodents) on a gray mouse lemur (*Microcebus murinus*) vocalizations dataset. Both taxa, rodents and gray mouse lemurs, show relative similarity in the frequency range and general spectral dynamics of their vocal communication. This similarity, the simplicity of the retraining approach, the efficiency of *DeepSqueak*'s Faster-RCNN and the user-friendly environment of the software allows yielding competitive results by training DeepSqueak on lemur's vocalization for their segmentation, classification and the unsupervised clustering of call types. The approach reaches high accuracy in the detection of calls (with 91% of correctly identified calls from a training set containing $\approx$ 2,000) with interesting insights on the effects of recording quality and inter-individual variation.

In a second part of the experiment, the authors also turn to transfer learning through pre-training. After having trained DeepSqueek on a gray mouse lemur dataset, they test its robustness in the detection of calls from Goodmans mouse lemurs (*M. lehilahytsara*), a closely related species which was never seen by the model during its training, achieving very similar results. This is thus a first example of how a model trained on one species or one dataset can be leveraged in processing a second species or dataset with the assumption that information extracted from the first task can be efficiently reemployed in the latter.

4.2 Pre-training

Surprisingly, relying on pre-trained models knowledge from taxonomically related primates is not a preferred approach in computational bioacoustics. Rather, most transfer learning experiments are built upon speech-based models with the underlying assumption that human and non-human primates share, at least, some vocal characteristics and that models are sufficiently resistant to such a domain shift. In addition to this, pre-trained self-supervised models for speech have been extensively explored in recent years and state-of-the-art solutions are now publicly available and easy to access through dedicated APIs like HuggingFace, S3PRL or SpeechBrain. We hereby give some of the few examples of how large speech-based PTM such as HuBERT (Hsu et al., 2021), wav2vec2 (Schneider et al., 2019) or DeepTone can be used to efficiently process primate vocalizations, either as frozen feature extractors replacing mel-spectrograms and engineered features or as foundation models aimed at offering a unified solution to multiple tasks and species.

An essential part in the development of transfer learning models is the comparison of their performance with more traditional approaches, as was done by Jiang et al. (2023). Here, the authors train a Long Short-Term Memory (LSTM) model and a transformer model on sound event detection: the segmentation and the automatic identification of call sequences from continuous vocalizations of bonobos (*Pan paniscus*), chimpanzees (*Pan troglodites*) and orangutans (*Pongo pygmaeus*). They focus their experiment on comparing performances across three feature extraction processes as input to the models: the raw waveform, spectrograms and wav2vec embeddings. Wav2vec (Schneider et al., 2019) is a speech based PTM, relying on self-supervised representation learning from raw audio and initially developed for speech recognition. The model consists in a multi-layer CNN trained on a noise contrastive binary classification task: it learns to extract informative representations of short sound frames of 30 ms from their context by differentiating them from other, randomly sampled, sound frames. This learned “latent representation”, also known as pre-trained embedding, can be seen as a fixed-length vector of 768 elements (for

wav2vec LARGE), supposedly encoding essential information from the input audio data. The encoded information was proven to be useful for its initial intended purpose of automatic speech recognition, but it may also incorporate other acoustic properties from speech such as the identity of a speaker or voice print, language information or even emotional expressivity from prosodical content (Y. Wang et al., 2021).

In the paper by Jiang et al. (2023), the assumption is that a wav2vec representation, although initially trained on speech, can also encode enough acoustic information to distinguish between great apes call types or to differentiate them from background forest noise. The authors find that training an LSTM on these embeddings yields better results compared to spectrograms or the raw waveform. They show the benefits of balancing classes when facing small annotated datasets and give an example of how pre-trained representation can also be used in *zero-shot* classification contexts by training a model on the orangutan dataset and using it to classify bonobo's call types without further training. These results thus show how the use of speech-based models is a promising solution for zero or few-shot learning from small primate datasets, even for cross-species classification.

Leroux et al. (2021) give an example of transfer learning from speech for primate vocal signature classification. They formulate a hypothesis for the existence of an acoustically encoded unique individual signature across call types in chimpanzees (*Pan troglodytes*) and test it through automatic classification of individuals. In doing so, the authors train several shallow classifiers on top of DeepTone Identity embeddings, a model pre-trained on 10,000 unique utterances from human IDs, and compare performances with MFCC inputs (a spectral representation often used for speaker identification). Despite the classifier relying on a simple SVM architecture and being trained on a rather unbalanced dataset of calls from three individual chimpanzees including three different call types, they reach 80% accuracy with consistently higher performances from DeepTone embeddings compared to MFCCs. Additionally, the transfer learning approach tends to show higher results compared to the spectral approach in low training data contexts, reaching a

maximally higher accuracy when using only 40 training examples. Again, this shows the value of pre-trained embeddings for few-shot learning when annotated data is scarce. The approach also gives a hint into the potential acoustic similarity between great apes vocalizations and human speech as well as the great generalizability and domain transfer abilities of speech based PTMs. In contrast, Lakdari et al. (2024) show that MFCCs can outperform pre-trained embeddings from wav2vec when used as input to an SVM for gibbon vocal identity classification. Yet, their experiment relies on modified verions of said embeddings (with embeddings averaged on several dimensions) which may result in an important loss of information prior to classification. We think that these modifications and the use of a single model may impair fair comparisons and do not properly reflect the abilities of pre-trained speech models (Jiang et al., 2023; Leroux et al., 2021; Sarkar & Doss, 2023).

4.3 Pretext tasks and pre-training data

We have seen that relying on learned representations from self-supervised PTMs can boost primate vocalization classification performances compared to using the raw waveform or spectral representations, even when said PTMs were initially trained on speech. With this in mind, an open question remains on the influence of different PTM architectures and pre-training datasets on the performances of downstream classifiers. Currently, speech processing state-of-the-art models are mostly self-supervised PTMs, and recent years have seen the emergence of innovative architectures frequently improving benchmark performance with new pre-training datasets, larger numbers of parameters or different "pretext tasks".

These pretext tasks are proxy tasks used to pre-train models on raw acoustic datasets without the need for human supervision (hence the self-supervised nature of these models). They consist in generating supervision from the data itself, requiring informative data representation learning from the model which will automatically learn to capture

structural information and acoustic patterns to reach low losses and higher predictive performances during training. A first example of such a pretext task is masked modeling, an approach introduced for the textual language model BERT (Devlin et al., 2019), which consists in masking portions of the data (either text, images or sound) and reconstructing said portions from their surrounding context. This approach was successfully implemented for speech in HuBERT by Hsu et al. (2021). As we have already seen for wav2vec, contrastive predictive coding is another pretext task consisting in predicting future sound frames from previous ones, and yields similar results compared to HuBERT.

Many such examples exist in the literature and could result in different performances gains when adapted to bioacoustic classification. This was tested by Sarkar and Doss (2023) who compared 11 speech-based PTMs, all trained on similar speech datasets (i.e., Librispeech for 10 of them and Libri-Light for Modified-CPC) on a common marmoset (*Callithrix jacchus*) caller detection task. These models include wav2vec and HuBERT as well as other state-of-the-art models including APC (Chung & Glass, 2020), Mockingjay (A. T. Liu et al., 2020) or WavLM (Chen et al., 2022), each presenting some specificity in their architecture, sizes and pretext tasks. As in both previously mentioned experiments, all PTM weights are kept frozen (the models are not further trained on unlabeled data) and used as feature extractors for downstream classifiers: SVMs and an LSTM for binary caller classification. The downstream models thus predict if two calls are uttered from the same individual or not. As can be seen in Figure 8, the authors test the performance of the downstream model in terms of Area Under the Curve and compare it with PTM size and pretext task (also referred to as the pre-training objective). The autoregressive reconstruction implemented in APC (Chung & Glass, 2020) and its vector quantized variant VQ-APC (Chung et al., 2020) seem to perform slightly better despite smaller model sizes. Surprisingly, Data2vec (Baevski et al., 2022) which was the most successful masked model for several speech tasks at the time of the experiment, performs lower than the rest, thus showing weaker representation learning capabilities in a domain adaptation context.

Following the question of model architectures and pretext tasks, the nature of the pre-training dataset could be considered as an essential part of the process, for the reason that PTMs are inherently conditioned to capture knowledge dependent on their training data. We should point out that the idea of using pre-training models from speech to perform bioacoustic tasks is not solely related to a theoretical similarity between speech and animal vocalizations. The approach can also be explained by the extensive availability of speech data in recent years, when the size of a pre-training dataset is an essential prerequisite to the success of self-supervised models. Yet, the effect of the nature of a pre-training dataset on bioacoustic tasks performances remains an open question. Although an intuitive answer to this second question would be that the closest in domain a pre-training dataset is to the downstream one, the preliminary results recently showcased by Hagiwara (2023) seem to indicate a more complicated situation. With their self-supervised model AVES, the authors go a step further from using speech-based PTMs as feature extractors and test performance gains in terms of pre-training data for an array of downstream tasks (classification and detection on marine and terrestrial mammals, amphibians, birds and primates). Heavily inspired by the HuBERT architecture, they entirely retrain the masked modeling transformer on several curated datasets including animal vocalizations, speech and general sound. They propose different pre-training sets by filtering audioset and VGGsound: two collections of several millions of 10 second audio clips drawn from YouTube videos with corresponding categories. By filtering said categories, they build 4 distinct data subsets :

- *core*: a configuration containing 153 hours of general sounds
- *bio*: the core configuration with added sounds corresponding to the animal label in VGG sound (360 hours)
- *non-bio*: a similarly sized control dataset containing random sounds from all categories except the animal one (360 hours)

- *all*: a dataset containing all types of sounds on top of the core configuration, making up to 5,054 hours of audio.

To further test the performance gains of their models, they compare them with VGGish and ResNet (both PTMs developed for general purpose audio-tagging) which were further trained in a supervised manner on the tasks at hand. The authors find that the AVES version trained on bioacoustic data (*bio*) outperforms other PTMs, including the supervised topline from VGGish and ResNet on most tasks. Although these results seem promising and show the validity of the approach, they must be taken carefully as the bioacoustic pre-training set only increases performance by a small margin. Furthermore, the primate detection task, carried out on Müller gibbons (*Hylobitae muelleri*), shows slightly lower mean average precision compared to the supervised and unsupervised versions of ResNet. This result may be explained by the scarcity of vocalizations contained in the gibbon dataset and might not entirely reflect the advantages of AVES which can be seen in most of the other tasks.

In a broader perspective, the authors compute t-scores to compare the average results obtained from the four pre-training datasets. Surprisingly, *bio* and *non-bio* reach very close performance and improve upon *all* despite their much smaller sizes. This indicates that selecting reduced curated datasets may give better results in a pre-training configuration rather than opting for very large and miscellaneous collections of sounds. The authors thus show their model's ability to generalize well across domains. This might be seen as a counterargument towards the need for a specific bioacoustic pre-training dataset for transfer learning from self-supervised models. It could also mean speech-based models are not successful in bioacoustics because of some acoustic resemblance between speech and animal vocalizations but rather because of their ability to transfer knowledge across acoustic domains. In any case, such assumptions will need to be further tested to account for the many technical limitations which might also explain these counter-intuitive results.

Lastly, the results obtained by Ghani et al. (2023), although not specifically tailored

for primate vocalization analysis, give a good example of transfer learning across species. In their experiments, large models pre-trained for bird sound classification (namely BirdNet 2.3 and Perch) are compared to general audio tagging models pre-trained on AudioSet (YAMNet, VGGish and AudioMAE). This comparison is carried out through probing: a method consisting in training simple linear layers on the pre-trained embeddings to understand how much of the information needed for the downstream task they are able to linearly encode. This gives a better account for the ability of a PTM to capture information for a given task, as the downstream model (a simple linear probe) adds very little knowledge to what was effectively captured by the PTM. In this case, results show that both bird-based models outperform the general event-detection ones by a good margin in detecting and classifying bird sounds as well as other animals such as frogs, cetaceans and bats. This is also the case in few-shot learning, as both models are still on the topline when downstream datasets are reduced in size, thus showing that pre-training models on bird sounds may be a viable option for few-shot learning on other scarcely annotated species. The authors state that this performance gain may be explained by the rich and diverse sounds produced by birds which occupy a broad range both temporally and in the spectral domain with great frequency, harmonic and rhythmical complexity. Added to this, the large amount of publicly available bird song datasets, in par with what can be found for speech compared to the scarcity of primate recordings, makes it another viable option for bioacoustic transfer learning across species. We think that, in addition to model testing and the development of primate-only PTMs, a good amount of work still needs to be put into understanding the influence of pre-training datasets for automatic primate vocalization analysis.

5 Discussion

5.1 Available datasets

As an encouragement for researchers to partake in further testing of the many options we have surveyed so far, we draw a non-exhaustive list of some publicly available primate vocalization datasets. These can be used either as pre-training data for transfer learning models or as manually annotated datasets for supervised downstream classification. Some also include open-source models and code to be used as inspiration or as baselines for performance evaluation. See Table 1 for a list of the previously-mentioned papers with code and dataset availability.

5.2 General lack of publicly available data

As can be seen in Table 1, few annotated datasets of primate vocalization recordings are made publicly available (compared to other taxa, or to the amount of public speech datasets). Despite this, our list in not exhaustive and more unpublished data could be found in addition to ongoing efforts still being carried out to this day in the recording of new datasets and their annotation. This lack of published data could be said to hinder research efforts into primate vocalization analysis and similar issues can be found in the whole field of bioacoustics (Baker & Vincent, 2019). Furthermore, within primate related research, some species are clearly underrepresented for various reasons, including the lack of interest put into the study of apparently poorly complex vocal systems, the remoteness of their habitat or the scarcity of endangered species. This is why the development of efficient machine learning solutions for the processing of primate vocalizations should always be made in parallel with annotation and recording projects as well as a substantial amount of work put into their deposition as supplementary material into public websites. The annotation itself should be thoroughly documented with an emphasis on reduced bias and reproducible methods. Finally, codes and models need to be published in open-source

(as is often the case). This encourages their reuse or adaptation, especially when developing large pre-trained foundation models which weights can be reemployed without the need for time and energy-consuming re-training.

5.3 Ethical and environmental concerns

In the field of computational bioacoustics applied to primates, we can observe a general tendency towards the use of PAM paired with deep learning approaches. This may be seen as a promising direction in terms of ethical and environmental concerns. In fact, PAM is considered as a non-invasive solution to the study of animal communication, and its automatic processing with machine learning methods leads to great opportunities for conservation and monitoring projects. Yet, several drawbacks should also be mentioned.

First of all, the lack of control over the elements recorded during PAM may lead to privacy concerns when human speech is picked up by the acoustic sensors. This problem, in turn, can be easily circumvented with similar machine learning methods as the ones used for the animal vocalization analysis. As we have mentioned before, speech processing methods for the automatic detection of speech have shown impressive results in the recent years and their implementation is strongly facilitated by the availability of user friendly open-source models. Employing such models to filter out speech, especially when facing recording of animals in captivity should be included as a preprocessing step in such experiments (Janetzky et al., 2021).

A second limit is the well-known environmental impact of AI, although computational bioacoustics stays a relatively niche domain of study compared to the research effort put into computer science for image or natural language processing. This important issue has been thoroughly addressed by specific reviews and studies (Van Wynsberghe, 2021). We should mention that promising solutions include the reuse of weights from pre-trained models which limits training time, and the development of foundation models for transfer learning as we discussed in Section 4.

Finally, the malicious use of automatic primate monitoring tools should always remain an important concern. As stated in Piel et al. (2022): "The ability of remote sensing tools to incidentally (or deliberately, in the case of poachers) reveal the location, movement and behavior of individuals raises concerns about informed consent, privacy, civil liberties, and fear of arrest". This problem is unfortunately embedded in the advocacy for open-source models, and no single solution exists. More work needs to be carried out in evaluating the impact of open-source animal detection models and datasets, as was done by Lennox et al. (2020) for biotelemetric data sharing. In contrast however, the detection of poaching activity may be tackled through machine learning solutions and seems to be an actively addressed problem in recent computational bioacoustics studies with tasks like gunshot, chainsaw or illegal cattle farming detection from acoustic data (Pérez-Granados & Schuchmann, 2023; Sethi et al., 2020).

6 Conclusion

Primate vocal communication research has seen a significant shift with the advent of machine learning and artificial neural networks, inspired by bioacoustics studies on other taxa, or sound and speech processing. The use of passive acoustic monitoring and the availability of large annotated datasets have paved the way for innovative automated workflows, reducing our reliance on manual annotations and analysis and questioning some aspects of human bias. This paradigm shift has seen the emergence of new automated tasks with important scientific implications and is starting to turn into valuable monitoring and conservation tools. We have provided a concise survey of the recent directions taken in computational bioacoustics, highlighting emerging approaches and the valuable insights they offer for the study of primate communication. We have discussed recent interests for state-of-the-art deep learning models and their prolific application to primate bioacoustic research, all the while reaffirming the validity and greater transparency of earlier statistical approaches. Looking ahead, the development of high-performance weakly supervised

transfer learning models holds promise for further advancements in the field, but challenges remain in understanding the behavior of these *black box* models for their subsequent use as scientific tools. Challenges also remain in terms of data availability, and in turning existing solutions into user-friendly light-weight models accessible to primatologists. As the field of computational bioacoustics evolves, we can expect deep learning research to bring further exciting developments that will continue to enrich our understanding of primate vocal communication and its role in their social interactions and ecological contexts.

Author contributions

Conceptualization: Jules Cauzinille, Ricard Marxer, Arnaud Rey, Benoit Favre; writing and preparation of original version: Jules Cauzinille; revision and editing: Ricard Marxer, Arnaud Rey, Benoit Favre. All authors have read and agreed to the published version of the manuscript

Acknowledgments

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). This work was also supported by the HEBBIAN (ANR-23-CE28-0008) and COMPO (ANR-23- CE23-0031) ANR projects.

Financial disclosure

This research received no external funding. The authors state that the views expressed in the submitted article are their own and do not constitute an official position of the institution or funder.

Conflict of interest

The authors declare no conflict of interests.

Glossary

clustering is a machine learning approach which consists in grouping similar objects into different subsets (or clusters). Although many algorithms exist, they generally work by partitioning a dataset into said clusters according to some similarity and/or dissimilarity metric. In the case of primate vocalizations, this approach is usually employed to examine repertoires of call-types. 8, 22–28, 30

CNN (Convolutional Neural Networks) are a popular DL architecture in bioacoustics, specifically designed to extract meaningful information from image inputs (such as spectrograms) by recognizing patterns in the data. 7, 13–17, 19, 21, 22, 27, 31

DL (Deep Learning) models refer to a category of machine learning algorithms based on neural networks and capable of learning complex patterns from data such as primate vocalizations. 3, 6, 7, 9–11, 17, 19–21, 24, 28

embedding (or latent representation) refers to a numerical vector outputted by a deep learning model and supposed to encode relevant informative features from input soundframes. Embeddings are usually extracted from a pre-trained model and used as input for downstream models in the transfer learning approach (see Figure 3). 31–33, 37

F1-score is a metric used to evaluate the performances of a machine learning model in binary classification tasks. Ranging from 0 (worst) to 1 (best), it is computed as a balance between precision and recall, making it particularly useful in scenarios with inbalanced classes or to account for false positives and false negatives. 8

mel-spectrogram provides a detailed visual representation of the frequency content of an audio signal. They can be considered as the conversion of a traditional spectrogram into the Mel scale, which is better aligned with human auditory perception. 9, 14, 31

MFCC (Mel-Frequency Cepstral Coefficients) are a spectral feature extraction method commonly used in sound processing. They allow capturing essential characteristics of an audio signal by converting the frequency domain into the so-called "Mel" scale. This makes MFCCs quite difficult to interpret visually but a very successful feature extraction method for machine learning models. 19–21, 32, 33

self-supervised learning is a pre-training approach relying on pseudo-labels found within the data itself, without human interventions (as in designing models that will predict future sound frames given a context). 27–29, 31, 33, 35, 36

supervised learning is a machine learning approach involving the use of annotated or labeled data as training material for a given algorithm. Weak supervision is more closely related to the unsupervised approach and involves using little expert knowledge during this process (a pre-defined number of call-types during unsupervised call-type clustering for example). 4, 21, 22, 24, 29, 36, 38, 40

SVM (Support Vector Machines) are a class of "simple" machine learning algorithms designed to linearly separate datapoints into different categories. They are especially popular in bioacoustics where the use of more complex deep learning models might not be necessary to reach acceptable performances on a given task. 8, 18, 21, 25, 32–34

transformer networks are a type of neural network specifically designed to model long-range dependencies in sequential data such as sound or text. Their popularity grew in recent years within natural language processing research but their application to bioacoustic data remains underexplored. 7, 31, 35

unsupervised learning is a machine learning approach which does not rely on expert annotations. The data used for unsupervised learning is unlabeled and the machine learning algorithms must categorize it according to the structure of the data itself or by recognizing specific patterns in said data. 22, 23, 25, 27, 28, 30, 36

References

Anders, F., Kalan, A. K., Kühl, H. S., & Fuchs, M. (2021). Compensating class imbalance for acoustic chimpanzee detection with convolutional recurrent neural networks. Ecological Informatics, 65, 101423. https://doi.org/https://doi.org/10.1016/j.ecoinf.2021.101423

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., & Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. International Conference on Machine Learning, 1298–1312. https://doi.org/10.48550/arXiv.2202.03555

Baker, E., & Vincent, S. (2019). A deafening silence: A lack of data and reproducibility in published bioacoustics research? Biodiversity Data Journal. https://doi.org/10.3897/BDJ.7.e36783

Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A., & Gramfort, A. (2021). Uncovering the structure of clinical eeg signals with self-supervised learning. Journal of Neural Engineering, 18(4). https://doi.org/10.1088/1741-2552/abca18

Bayestehtashk, A., Shafran, I., Coleman, K., & Robertson, N. (2014). Detecting vocalizations of individual monkeys in social groups. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Soci 4775–4779. https://doi.org/10.1109/EMBC.2014.6944692

Best, P., Paris, S., Glotin, H., & Marxer, R. (2023). Deep audio embeddings for vocalisation clustering. Plos one, 18(7), e0283396. https://doi.org/10.1371/journal.pone.0283396

Bonafos, G., Pudlo, P., Freyermuth, J.-M., Legou, T., Fagot, J., Tronçon, S., & Rey, A. (2023, October). Detection and classification of vocal productions in large scale audio recordings [working paper or preprint]. https://doi.org/10.48550/arXiv.2302.07640

Bravo Sanchez, F. J., Hossain, M. R., English, N. B., & Moore, S. T. (2021). Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep

learning architecture [Publisher: Nature Publishing Group]. Scientific Reports, 11(1), 15733. https://doi.org/10.1038/s41598-021-95076-6

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505–1518. https://doi.org/10.1109/JSTSP.2022.3188113

Chung, Y.-A., & Glass, J. (2020). Generative pre-training for speech with autoregressive predictive coding.
ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (IC
3497–3501. https://doi.org/10.48550/arXiv.1910.12607

Chung, Y.-A., Tang, H., & Glass, J. (2020). Vector-Quantized Autoregressive Predictive Coding. Proc. Interspeech 2020, 3760–3764.
https://doi.org/10.21437/Interspeech.2020-1228

Clink, D. J., Bernard, H., Crofoot, M. C., & Marshall, A. J. (2017). Investigating Individual Vocal Signatures and Small-Scale Patterns of Geographic Variation in Female Bornean Gibbon (Hylobates muelleri) Great Calls.
International Journal of Primatology, 38(4), 656–671.
https://doi.org/10.1007/s10764-017-9972-y

Cramer, A. L., Wu, H.-H., Salamon, J., & Bello, J. P. (2019). Look, listen, and learn more: Design choices for deep audio embeddings.
ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (I
3852–3856. https://doi.org/10.1109/ICASSP.2019.8682475

Crofoot, M., Lambert, T., Kays, R., & Wikelski, M. (2010). Does watching a monkey change its behaviour? quantifying observer effects in habituated wild primates using automated telemetry. Animal Behaviour, 80, 475–480.
https://doi.org/10.1016/j.anbehav.2010.06.006

Crunchant, A.-S., Borchers, D., Kühl, H., & Piel, A. (2020). Listening and watching: Do camera traps or acoustic sensors more efficiently detect wild chimpanzees in an open habitat? Methods in Ecology and Evolution, 11(4), 542–552. https://doi.org/10.1111/2041-210X.13362

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Compu 4171–4186. https://doi.org/10.18653/v1/N19-1423

Do Nascimento, L. A., Pérez-Granados, C., & Beard, K. H. (2021). Passive acoustic monitoring and automatic detection of diel patterns and acoustic structure of howler monkey roars. Diversity, 13(11). https://doi.org/10.3390/d13110566

Dufourq, E., Durbach, I., Hansford, J. P., Hoepfner, A., Ma, H., Bryant, J. V., Stender, C. S., Li, W., Liu, Z., Chen, Q., et al. (2021). Automated detection of hainan gibbon calls for passive acoustic monitoring. Remote Sensing in Ecology and Conservation, 7(3), 475–487. https://doi.org/10.1101/2020.09.07.285502

Enari, H., Enari, H., Okuda, K., Yoshita, M., Kuno, T., & Okuda, K. (2017). Feasibility assessment of active and passive acoustic monitoring of sika deer populations. Ecological Indicators, 79, 155–162. https://doi.org/https://doi.org/10.1016/j.ecolind.2017.04.004

Enari, H., Enari, H. S., Okuda, K., Maruyama, T., & Okuda, K. N. (2019). An evaluation of the efficiency of passive acoustic monitoring in detecting deer and primates in comparison with camera traps. Ecological Indicators, 98, 753–762. https://doi.org/10.1016/j.ecolind.2018.11.062

Erb, W., Ross, W., Kazanecki, H., Mitra Setia, T., Madhusudhana, S., & Clink, D. (2023, April). Vocal complexity in the long calls of Bornean orangutans (preprint). Animal Behavior and Cognition. bioarXiv. https://doi.org/10.1101/2023.04.05.535487

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459–1462. https://doi.org/10.1145/1873951.1874246

Fedurek, P., Zuberbühler, K., & Dahl, C. D. (2016). Sequential information in a great ape utterance. Scientific Reports, 6(1), 38226. https://doi.org/10.1038/srep38226

Ganchev, T. (2017). Computational bioacoustics: Biodiversity monitoring and assessment (Vol. 4). Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9781614516316

Germain, F. G., Chen, Q., & Koltun, V. (2019). Speech denoising with deep feature losses. Proc. Interspeech 2019, 2723–2727. https://doi.org/arXiv.1806.10522

Ghani, B., Denton, T., Kahl, S., & Klinck, H. (2023, July). Feature Embeddings from Large-Scale Acoustic Bird Classifiers Enable Few-Shot Transfer Learning [arXiv:2307.06292 [cs, eess]]. Retrieved July 20, 2023, from http://arxiv.org/abs/2307.06292

Guo, S., Xu, P., Miao, Q., Shao, G., Chapman, C. A., Chen, X., He, G., Fang, D., Zhang, H., Sun, Y., Shi, Z., & Li, B. (2020). Automatic identification of individual primates with deep learning techniques. iScience, 23(8), 101412. https://doi.org/https://doi.org/10.1016/j.isci.2020.101412

Hagiwara, M. (2023). Aves: Animal vocalization encoder based on self-supervision. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (I 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095642

Hagiwara, M., Hoffman, B., Liu, J.-Y., Cusimano, M., Effenberger, F., & Zacarian, K. (2023). Beans: The benchmark of animal sounds. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (I 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096686

Heinicke, S., Kalan, A. K., Wagner, O. J., Mundry, R., Lukashevich, H., & Kühl, H. S. (2015). Assessing the performance of a semiautomated acoustic monitoring system for primates (K. Jones, Ed.). Methods in Ecology and Evolution, 6(7), 753–763. https://doi.org/10.1111/2041-210X.12384

Hsu, W.-N., Bolte, B., Tsai, Y.-H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units.
IEEE/ACM Transactions on Audio, Speech, and Language Processing, PP, 1–1. https://doi.org/10.1109/TASLP.2021.3122291

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in python. Springer Nature. https://doi.org/10.1007/978-3-031-38747-0

Janetzky, P., Davidson, P., Steininger, M., Krause, A., & Hotho, A. (2021). Detecting presence of speech in acoustic data obtained from beehives.
Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop 26–30.

Jiang, Z., Soldati, A., Schamberg, I., Lameira, A. R., & Moran, S. (2023). Automatic Sound Event Detection and Classification of Great Ape Calls Using Neural Networks. https://doi.org/https://doi.org/10.48550/arXiv.2301.02214

Kalan, A. K., Mundry, R., Wagner, O. J., Heinicke, S., Boesch, C., & Kühl, H. S. (2015). Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. Ecological Indicators, 54, 217–226. https://doi.org/10.1016/j.ecolind.2015.02.023

Kidney, D., Rawson, B. M., Borchers, D. L., Stevenson, B. C., Marques, T. A., & Thomas, L. (2016). An efficient acoustic density estimation method with human detectors applied to gibbons in cambodia. PLOS ONE, 11(5), 1–16. https://doi.org/10.1371/journal.pone.0155066

Kiskin, I., Sinka, M., Cobb, A. D., Rafique, W., Wang, L., Zilli, D., Gutteridge, B., Dam, R., Marinos, T., Li, Y., et al. (2021). Humbugdb: A large-scale acoustic mosquito dataset. arXiv e-prints. https://doi.org/10.5281/zenodo.4904800

Kiskin, I., Zilli, D., Li, Y., Sinka, M., Willis, K., & Roberts, S. (2020). Bioacoustic detection with wavelet-conditioned convolutional neural networks. Neural Computing and Applications, 32(4), 915–927. https://doi.org/10.1007/s00521-018-3626-7

Lakdari, M. W., Ahmad, A. H., Sethi, S., Bohn, G. A., & Clink, D. J. (2024). Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. Ecological Informatics, 80, 102457. https://doi.org/10.1016/j.ecoinf.2023.102457

Lennox, R. J., Harcourt, R., Bennett, J. R., Davies, A., Ford, A. T., Frey, R. M., Hayward, M. W., Hussey, N. E., Iverson, S. J., Kays, R., Kessel, S. T., Mcmahon, C., Muelbert, M., Murray, T. S., Nguyen, V. M., Pye, J. D., Roche, D. G., Whoriskey, F. G., Young, N., & Cooke, S. J. (2020). A Novel Framework to Protect Animal Data in a World of Ecosurveillance. BioScience, 70(6), 468–476. https://doi.org/10.1093/biosci/biaa035

Leroux, M., Al-Khudhairy, O. G., Perony, N., & Townsend, S. W. (2021, December). Chimpanzee voice prints? Insights from transfer learning experiments from human voices. https://doi.org/https://doi.org/10.48550/arXiv.2112.08165

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. AI Open, 3, 111–132. https://doi.org/https://doi.org/10.1016/j.aiopen.2022.10.001

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., & Lee, H.-y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders.

ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (IC
6419–6423. https://doi.org/10.1109/ICASSP40776.2020.9054458

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2023). Self-supervised learning: Generative or contrastive. IEEE Transactions on Knowledge; Data Engineering, 35(01), 857–876. https://doi.org/10.1109/TKDE.2021.3090866

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B., & Virtanen, T. (2017). Dcase 2017 challenge setup: Tasks, datasets and baseline system. DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events. https://doi.org/doi:10.1109/TASLP.2019.2907016

Mesaros, A., Heittola, T., & Virtanen, T. (2019). Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups. Workshop on Detection and Classification of Acoustic Scenes and Events. https://doi.org/10.33682/m5kp-fa97

Mielke, A., & Zuberbühler, K. (2013). A method for automated individual, species and call type recognition in free-ranging animals. Animal Behaviour, 86(2), 475–482. https://doi.org/10.1016/j.anbehav.2013.04.017

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T., & Watanabe, S. (2022). Self-supervised speech representation learning: A review. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1179–1210. https://doi.org/10.1109/JSTSP.2022.3207050

Müller, M. (2007). Dynamic time warping. Information retrieval for music and motion, 69–84. https://doi.org/10.1007/978-3-540-74048-3_4

Pellegrini, T. (2021). Deep-Learning-Based Central African Primate Species Classification with MixUp and SpecAugment. Interspeech 2021, 456–460. https://doi.org/10.21437/Interspeech.2021-1911

Pérez-Granados, C., & Schuchmann, K.-L. (2023). The sound of the illegal: Applying bioacoustics for long-term monitoring of illegal cattle in protected areas. Ecological Informatics, 74, 101981. https://doi.org/https://doi.org/10.1016/j.ecoinf.2023.101981

Pérez-Granados, C., & Traba, J. (2021). Estimating bird density using passive acoustic monitoring: A review of methods and suggestions for further research. Ibis, 163(3), 765–783. https://doi.org/10.1111/ibi.12944

Pichler, M., & Hartig, F. (2023). Machine learning and deep learninga review for ecologists. Methods in Ecology and Evolution, 14(4), 994–1016. https://doi.org/10.1111/2041-210X.14061

Piel, A. K., Crunchant, A., Knot, I. E., Chalmers, C., Fergus, P., Mulero-Pázmány, M., & Wich, S. A. (2022). Noninvasive Technologies for Primate Conservation in the 21st Century. International Journal of Primatology, 43(1), 133–167. https://doi.org/10.1007/s10764-021-00245-z

Pozzi, L., Gamba, M., & Giacoma, C. (2010). The use of artificial neural networks to classify primate vocalizations: A pilot study on black lemurs. American Journal of Primatology: Official Journal of the American Society of Primatologists, 72(4), 337–348. https://doi.org/10.1002/ajp.20786

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Robakis, E., Watsa, M., & Erkenswick, G. (2018). Classification of producer characteristics in primate long calls using neural networks. The Journal of the Acoustical Society of America, 144(1), 344–353. https://doi.org/10.1121/1.5046526

Romero-Mujalli, D., Bergmann, T., Zimmermann, A., & Scheumann, M. (2021). Utilizing DeepSqueak for automatic detection and classification of mammalian vocalizations:

A case study on primate vocalizations. Scientific Reports, 11(1), 24463. https://doi.org/10.1038/s41598-021-03941-1

Ross, S. R.-J., O'Connell, D. P., Deichmann, J. L., Desjonquères, C., Gasc, A., Phillips, J. N., Sethi, S. S., Wood, C. M., & Burivalova, Z. (2023). Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Functional Ecology, 37(4), 959–975. https://doi.org/10.1111/1365-2435.14275

Ruan, W., Wu, K., Chen, Q., & Zhang, C. (2022). ResNet-based bio-acoustics presence detection technology of Hainan gibbon calls. Applied Acoustics, 198, 108939. https://doi.org/10.1016/j.apacoust.2022.108939

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS computational biology, 16(10), e1008228. https://doi.org/10.1371/journal.pcbi.1008228

Sarkar, E., & Doss, M. M. (2023, May). Can Self-Supervised Neural Networks Pre-Trained on Human Speech distinguish Animal Callers? [arXiv:2305.14035 [cs, eess]]. https://doi.org/https://doi.org/10.48550/arXiv.2305.14035

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. Proc. Interspeech 2019, 3465–3469. https://doi.org/10.21437/Interspeech.2019-1873

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. Science Advances, 5(9), eaaw0736. https://doi.org/10.1126/sciadv.aaw0736

Schuller, B. W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., … Kaandorp, C. (2021). The INTERSPEECH 2021 Computational Paralinguistics

Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. https://doi.org/10.48550/arXiv.2102.13468

Sethi, S., Jones, N., Fulcher, B., Picinali, L., Clink, D., Klinck, H., Orme, D., Wrege, P., & Ewers, R. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. Proceedings of the National Academy of Sciences, 117, 202004702. https://doi.org/10.1073/pnas.2004702117

Stowell, D. (2019). State of the art in computational bioacoustics and machine learning: How far have we come? Biodiversity Information Science and Standards, 3, e37227. https://doi.org/10.3897/biss.3.37227

Stowell, D. (2022). Computational bioacoustics with deep learning: A review and roadmap. PeerJ, 10. https://doi.org/10.7717/peerj.13152

Sugai, L. S. M., Silva, T. S. F., Ribeiro, J., José Wagner, & Llusia, D. (2018). Terrestrial Passive Acoustic Monitoring: Review and Perspectives. BioScience, 69(1), 15–25. https://doi.org/10.1093/biosci/biy147

Turesson, H. K., Ribeiro, S., Pereira, D. R., Papa, J. P., & de Albuquerque, V. H. C. (2016). Machine learning algorithms for automatic classification of marmoset vocalizations (M. Smotherman, Ed.). PLOS ONE, 11(9), e0163041. https://doi.org/10.1371/journal.pone.0163041

Tzirakis, P., Shiarella, A., Ewers, R., & Schuller, B. W. (2020). Computer Audition for Continuous Rainforest Occupancy Monitoring: The Case of Bornean Gibbons Call Detection. Interspeech 2020, 1211–1215. https://doi.org/10.21437/Interspeech.2020-2655

Van Wynsberghe, A. (2021). Sustainable ai: Ai for sustainability and the sustainability of ai. AI and Ethics, 1(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

von Luxburg, U., Williamson, R. C., & Guyon, I. (2012, February). Clustering: Science or art? In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.),

Proceedings of icml workshop on unsupervised and transfer learning (pp. 65–79, Vol. 27). PMLR.

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing vocal repertoireshard vs. soft classification approaches. PLOS ONE, 10(4), 1–16. https://doi.org/10.1371/journal.pone.0125785

Wang, Y., Boumadane, A., & Heba, A. (2021). A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. CoRR, abs/2111.02735. https://doi.org/10.48550/arXiv.2111.02735

Wang, Y., Ye, J., & Borchers, D. L. (2022). Automated call detection for acoustic surveys with structured calls of varying length. Methods in Ecology and Evolution, 13(7), 1552–1567. https://doi.org/10.1111/2041-210X.13873

Wu, H.-H., Kao, C.-C., Tang, Q., Sun, M., McFee, B., Bello, J. P., & Wang, C. (2021). Multi-task self-supervised pre-training for music classification. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (I 556–560. https://doi.org/10.1109/ICASSP39728.2021.9414405

Xie, J., Colonna, J. G., & Zhang, J. (2021). Bioacoustic signal denoising: A review. Artif. Intell. Rev., 54(5), 3575–3597. https://doi.org/10.1007/s10462-020-09932-4

Xie, J., Hu, K., Guo, Y., Zhu, Q., & Yu, J. (2021). On loss functions and cnns for improved bioacoustic signal classification. Ecological Informatics, 64, 101331. https://doi.org/https://doi.org/10.1016/j.ecoinf.2021.101331

Zimmer, W. M. X. (2011). Passive acoustic monitoring of cetaceans. Cambridge University Press. https://doi.org/10.1017/CBO9780511977107

Zimmermann, A., Zimmermann, E., Newman, J. D., & Jürgens, U. (1995). Artificial neural networks for analysis and recognition of primate vocal communication. In Current topics in primate vocal communication (pp. 29–46). Springer US. https://doi.org/10.1007/978-1-4757-9930-9_2

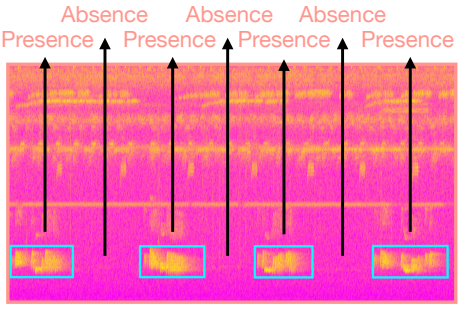
Zwerts, J. A., Treep, J., Kaandorp, C., Meewis, F., Koot, A. C., Kaya, H., et al. (2021). Introducing a central african primate vocalisation dataset for automated species classification. INTERSPEECH 2021, 466–470. https://doi.org/10.21437/Interspeech.2021-154

Table 1

Summary of cited experiments with data and code availability.

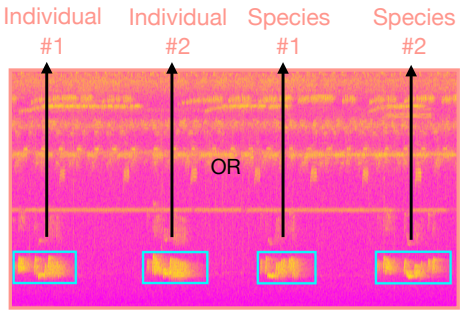
Citation	Species	Task	Architecture	Features	Availability
Dufourq et al. (2021)	Hainan Gibbons	Segmentation	supervised CNN		Dataset + code
Ruan et al. (2022)	Hainan Gibbons	Segmentation	ResNet		Subset + code
Y. Wang et al. (2022)	Hainan Gibbons	Segmentation	CRNN		Dataset + code
Tzirakis et al. (2020)	Müller gibbons	Segmentation	CRNN		Code
Bayestehtashk et al. (2014)	rhesus macaques	Segmentation	SVM	OpenSmile	
Robakis et al. (2018)	emperor tamarins + saddleback tamarins	Individual classification	Linear layers	Handcrafted	
Clink et al. (2017)	Müller gibbons	Individual discrimination	DFA	Handcrafted	Dataset
Fedurek et al. (2016)	chimpanzees	Individual classification	SVM	MFCC	
Mielke and Zuberbühler (2013)	Stuhlman's blue monkeys + others	Individual classification + species recognition	MLP		
Zwerts et al. (2021)	chimpanzees + mandrills + red-capped mangabeys + guenons	species recognition	KELM	MFCC	Dataset + code
Pellegrini (2021)	chimpanzees + mandrills + red-capped mangabeys + guenons	species recognition	CNN + Mobilenet + Resnet		Code
Pozzi et al. (2010)	black lemurs	call type classification	ANN + DFA + clustering		
Turesson et al. (2016)	common marmoset	call type classification	Seven models (SVM, kNN, OPF...)		Dataset + Code
Erb et al. (2023)	Bornean orangutans	call type clustering	SVM + clustering		
Sainburg et al. (2020)	gibbons + rhesus macaques + other taxa	call type clustering	UMAPS	spectrograms	Dataset
Best et al. (2023)	no primates	call type clustering	SSL CNN auto-encoder + UMAPS	spectrograms + PTMs	Code
Romero-Mujalli et al. (2021)	gray mouse lemurs + goodmans mouse lemurs	segmentation + classification + call type clustering	DeepSqueek (FastCNN)		Dataset + code
Jiang et al. (2023)	bonobos + chimpanzees + orangutans	segmentation + classification	CNN	waveform + spectrograms + wav2vec embeddings	Code
Leroux et al. (2021)	chimpanzees	Individual classification	ANNs	DeepTone	
Sarkar and Doss (2023)	common marmosets	individual discrimination	SVM, LSTM	11 speech PTMs	
Hagiwara (2023)	gibbons + other taxa	segmentation + other	AVES (bioacoustic HuBERT)		Dataset + code

3.1 Detection and segmentation



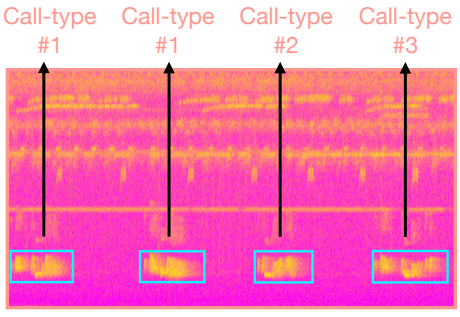
- Extracting vocalizations from long PAM recordings
- Analyzing vocalization frequency, localization, length, etc.
- Accelerating manual dataset collection for qualitative analysis

3.2 Identification and density estimation



- Counting or recognizing individuals / species in the wild
- Analyzing inter-individual / inter-sexual / inter-species variations
- Monitoring bioacoustic diversity and environmental pressures

3.3 Vocal repertoires and clustering



- Discovering new vocal repertoires / call-types
- Assessing the validity of pre-defined expert vocal repertoires
- Analyzing discreteness / continuity of vocalization typologies

Figure 1
The three main categories of tasks tackled with machine learning for primate bioacoustics.

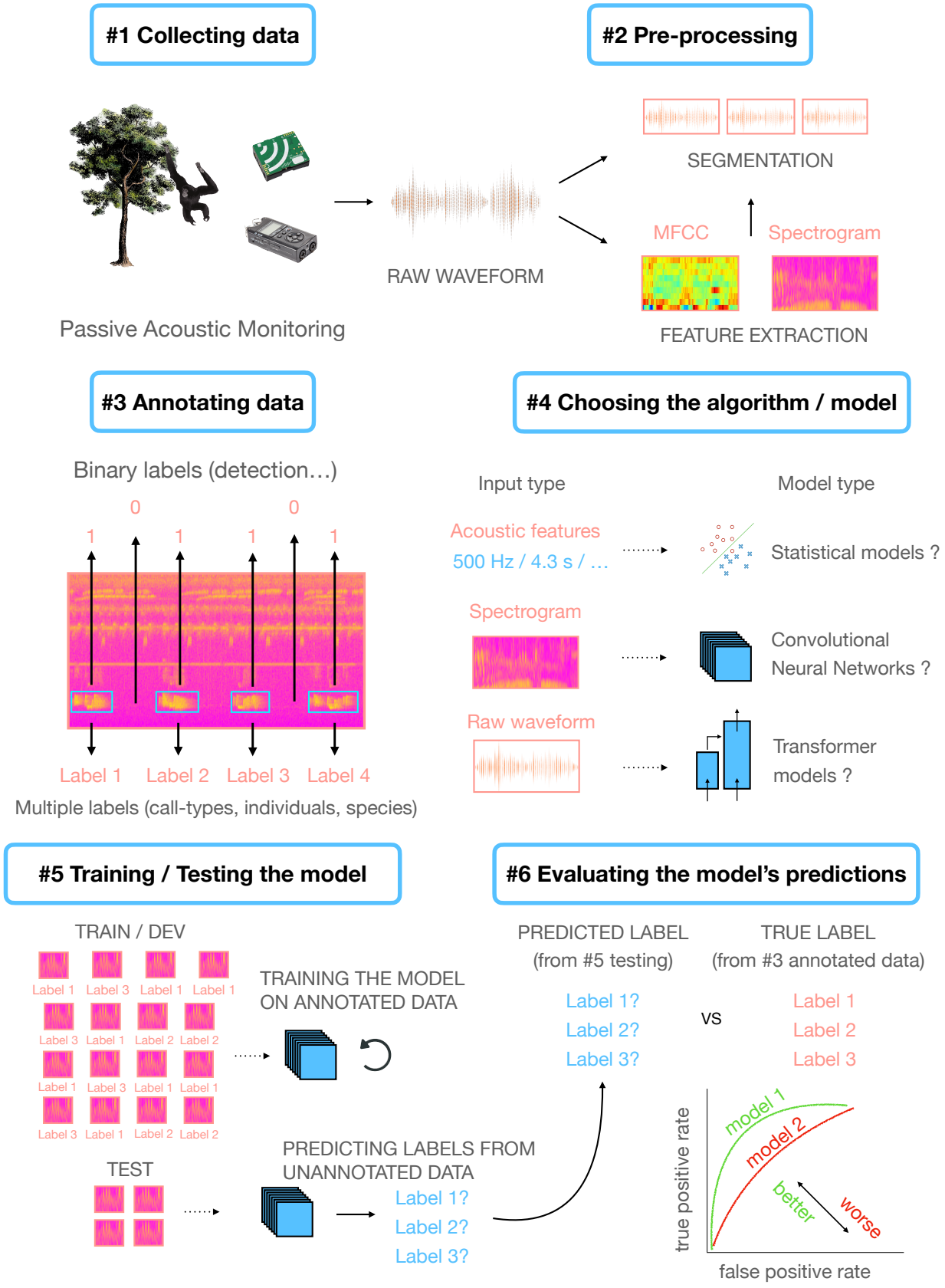


Figure 2
Machine learning workflow for primate bioacoustics.

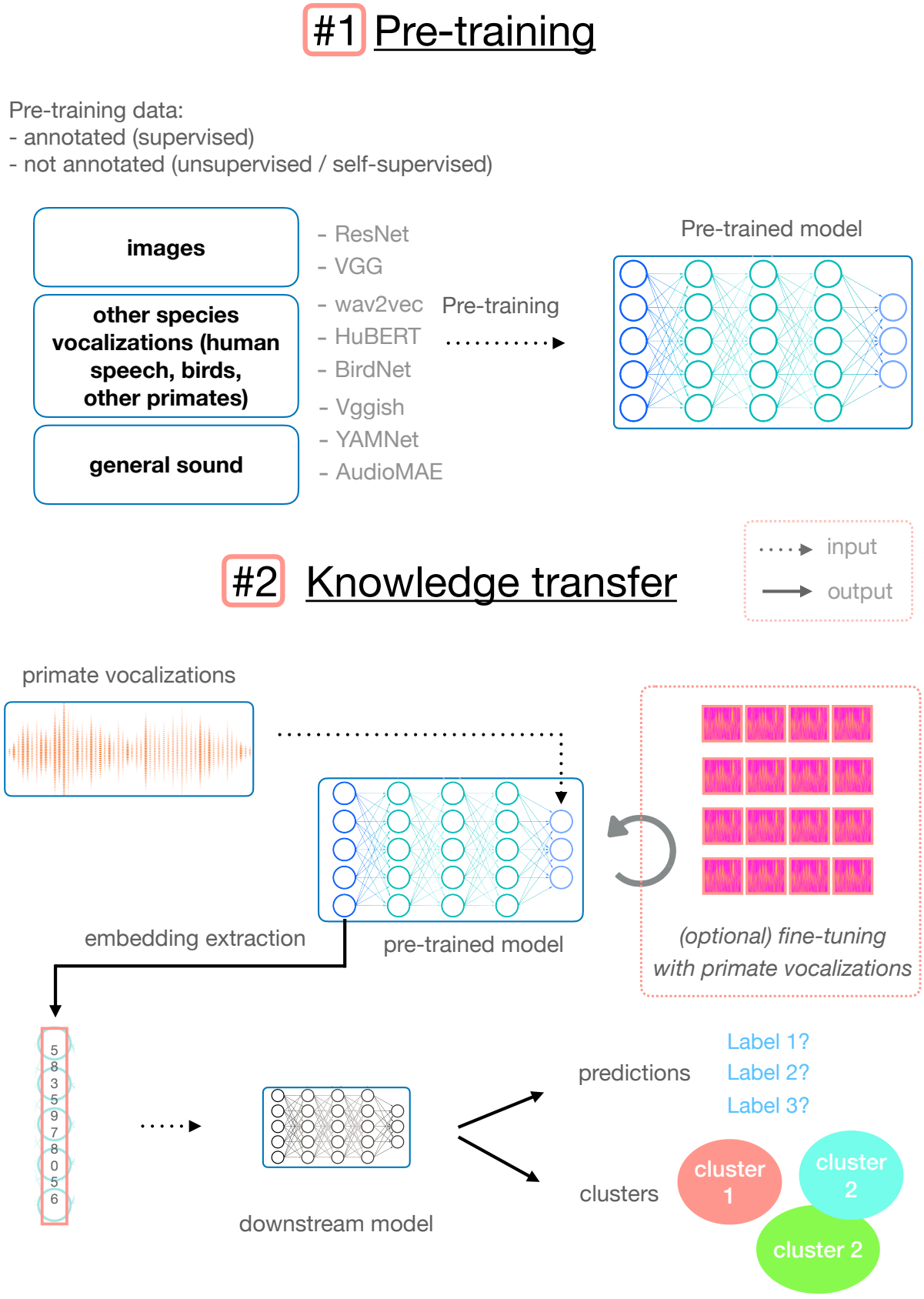


Figure 3

The transfer learning approach: Using a pre-trained model to classify primate vocalizations (ResNet, VGG, wav2vec, etc. are examples of publicly available pre-trained models).

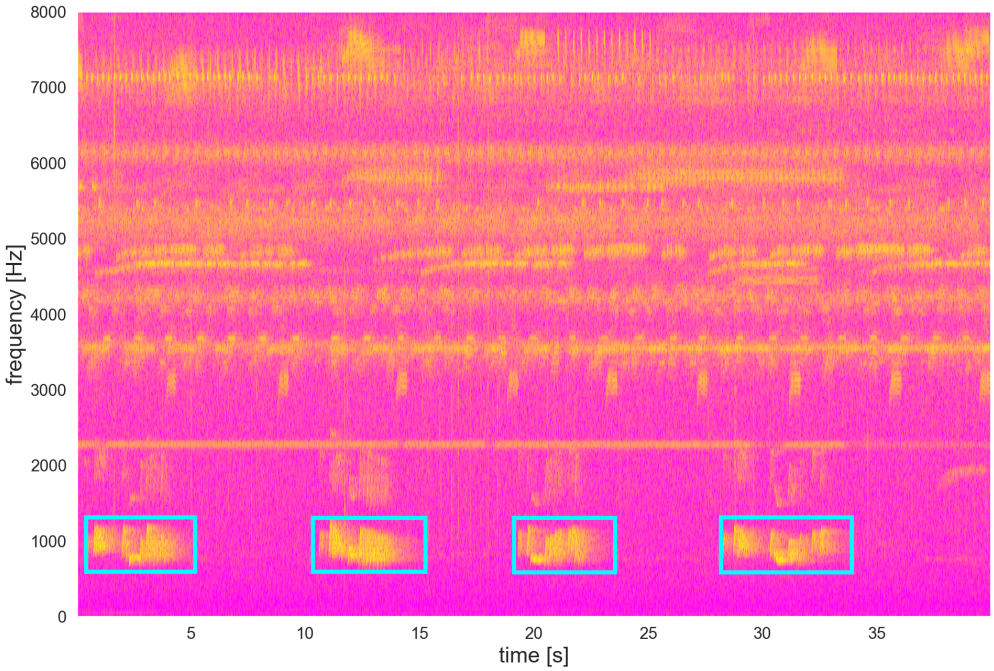


Figure 4

Spectrogram of a Müller Gibbon call. The blue boxes correspond to time and frequency boundaries of the calls. Data from Clink et al. (2017).

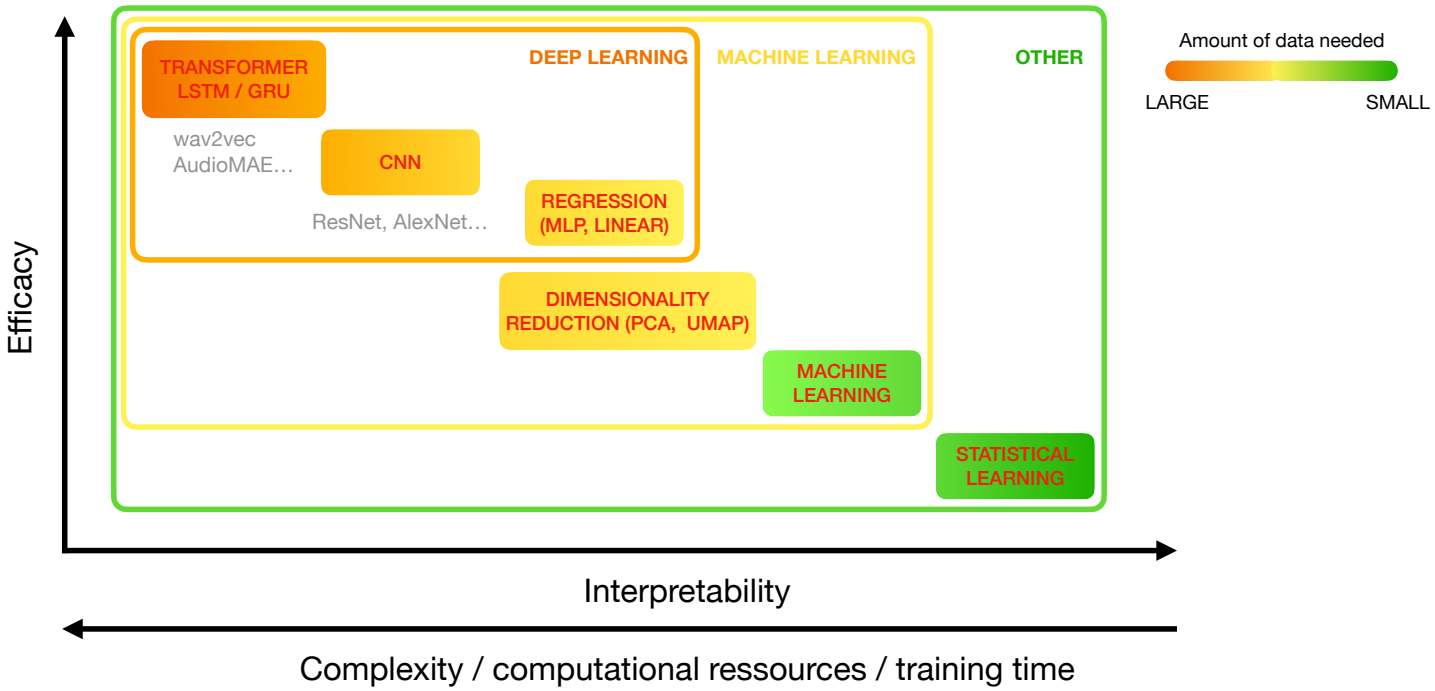


Figure 5

The efficacy / explainability tradeoff between different machine learning architectures (wav2vec, AudioMAE, ResNet and AlexNet are examples of popular deep learning architectures used in bioacoustics).

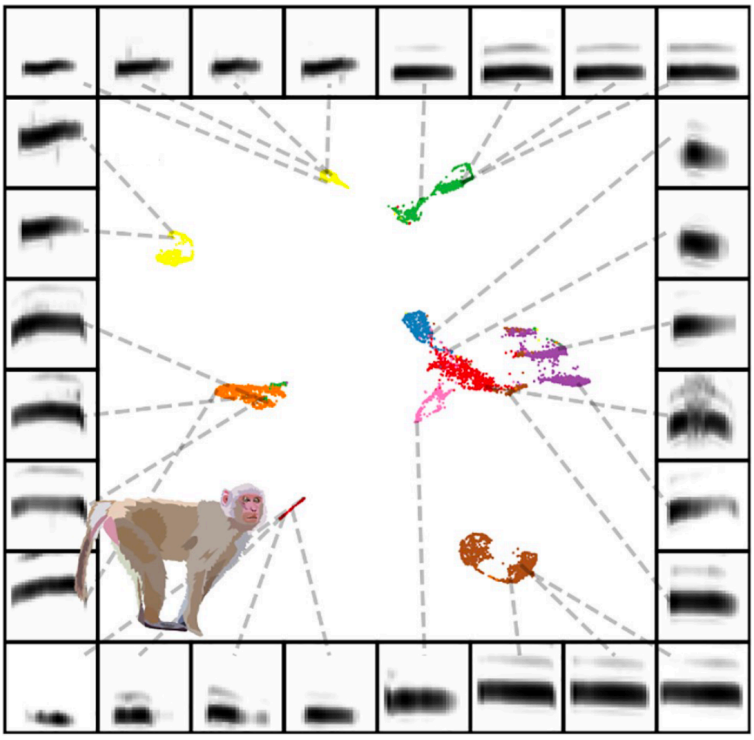


Figure 6

Spectrograms of Rhesus macaques (Macaca mulatta) vocal elements discretized and embedded into a 2D UMAP space. Scatter plot points are colored by individual identity. Image from Sainburg et al. (2020).

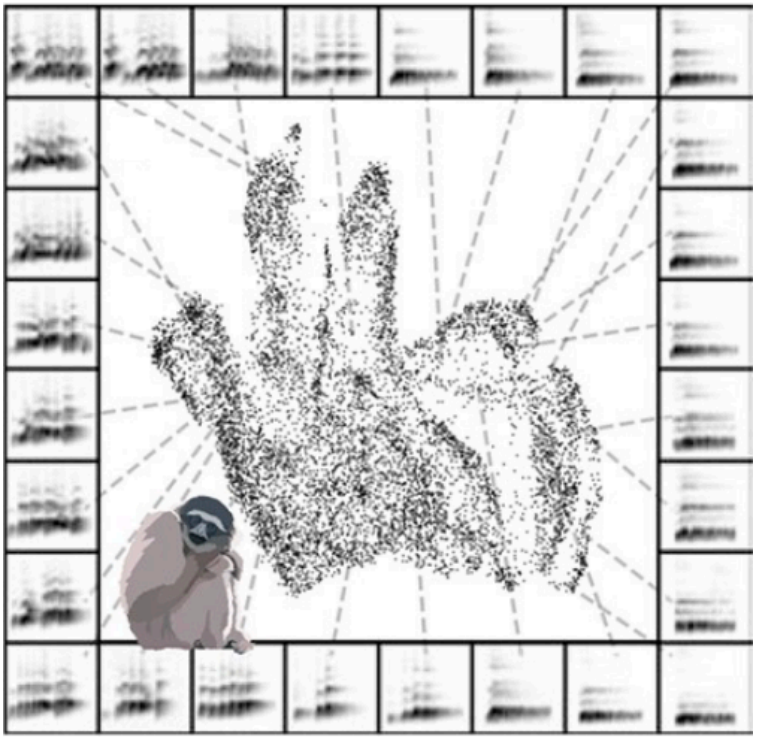


Figure 7

Gibbon syllable spectrograms embedded into a 2D UMAP space. Image from Sainburg et al. (2020).

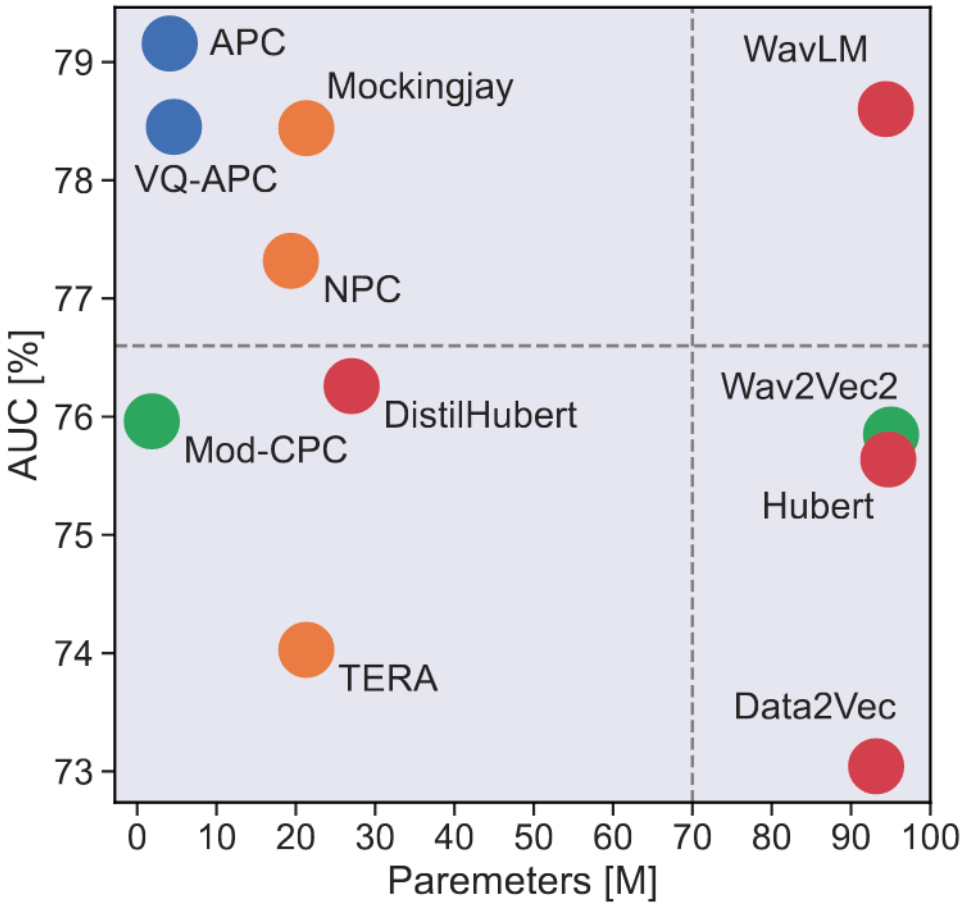


Figure 8

Model size against performance on a primate bioacoustics task. Model pre-training objective denoted as: Masked prediction (red). Autoregressive reconstruction (blue). Contrastive (green). Masked reconstruction (orange). AUC is the Area Under Curve evaluation metric. Figure from Sarkar and Doss (2023).