scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A large annotated dataset of vocalizations by common marmosets

Charly Lamothe[1,2,3] ✉, Manon Obliger-Debouche[1,3] ✉, Paul Best[2], Régis Trapeau[1], Sabrina Ravel[1], Thierry Artières[2], Ricard Marxer[2] & Pascal Belin[1] ✉

Non-human primates, our closest relatives, use a wide range of complex vocal signals for communication within their species. Previous research on marmoset (*Callithrix jacchus*) vocalizations has been limited by sampling rates not covering the whole hearing range and insufficient labeling for advanced analyses using Deep Neural Networks (DNNs). Here, we provide a database of common marmoset vocalizations, which were continuously recorded with a sampling rate of 96 kHz from an animal holding facility housing simultaneously ~20 marmosets in three cages. The dataset comprises more than 800,000 files, amounting to 253 hours of data collected over 40 months. Each recording lasts a few seconds and captures the marmosets' social vocalizations, encompassing their entire known vocal repertoire during the experimental period. Around 215,000 calls are annotated with the vocalization type. We offer a trained classifier to assist future investigations. Finally, we validated our dataset by sampling 700 representative recordings and cross-examining them with four experts.

Background & Summary

Non-human primates, our close evolutionary cousins, exhibit various complex behaviors, including the extensive use of acoustically diverse vocal signals for communication within conspecifics. By conducting comparative research on non-human primates, valuable insights can be gained into the evolutionary development of speech and language. Although non-human vocalizations can be difficult for humans to decipher, large acoustic datasets may make it possible to identify important nuances that are critical to communication among animals but may be imperceptible to the human ear. There has been considerable interest recently in the common marmoset (*Callithrix jacchus*) as a neuroscientific model organism[1], and many attempts have been made to study and characterize its vocal repertoire[2–7].

Several studies have used machine learning to detect and/or segment and/or label marmoset vocalizations. For example, Turesson *et al.*[8] attempted to identify the most robust classifier for a small labeled vocalization dataset (~300 samples), while Phaniraj *et al.*[9] sought to optimize source identification from a small dataset size of labeled marmoset vocalizations (~7 K samples), and Zhang *et al.*[6] focused on finding the best supervised deep learning-based methods for detecting and classifying marmoset vocalizations in a medium dataset of labeled vocalizations (~20 K samples).

However, the existing studies present several important limitations. First, the audio recording setups did not allow recording above a sampling rate of 48 kHz, which would allow the full frequency range of marmoset vocalizations, corresponding to their hearing range from 125 Hz to 36 kHz[10], to be recorded. Second, the existing datasets did not provide a sufficient number of labeled vocalizations to leverage advanced analytical methods. Fine-grained statistical analyses, such as those based on deep learning for decoding animal communication[11], require substantial data (usually hundreds of thousands, e.g., Best *et al.*[12] for animal vocalizations and up to millions in state-of-the-art DNNs). Finally, the manually labeled vocalizations in most marmoset research studies are often not shared publicly. To address this gap, we employed signal processing and deep learning tools to segment automatically and cluster vocalizations, following methods detailed in recent computational

[1]La Timone Neuroscience Institute UMR 7289, CNRS, Aix-Marseille University, Marseille, France. [2]Laboratoire d'Informatique et Systèmes UMR 7020, CNRS, Aix-Marseille University, Marseille, France. [3]These authors contributed equally: Charly Lamothe, Manon Obliger-Debouche. ✉e-mail: charlylmth@gmail.com; obliger.manon@gmail.com; pascal.belin@univ-amu.fr

neuroethology literature[12–14]. We then implemented an iterative refinement process to label vocalizations across extensive recordings, minimizing the need for expert supervision.

Here, we present a large collection of vocalizations of marmosets. We have acquired and segmented over 800,000 vocalizations with a sampling rate of 96 kHz from a soundproofed animal facility room that contains three cages (~20 marmosets) over a period of three years. Marmosets are capable of producing a diverse array of vocalizations, including trills, phees, twitters, tsiks, seeps, and infant cries, even when kept in captivity[2,4,15,16].

Such a high-rendering method might produce noisier segmentation and labeling; therefore, we validated our dataset by selecting a representative sample of 700 recordings, which were reviewed and cross-examined by four independent experts to ensure accuracy and consistency. Yet, our method has the benefit of opening the way for four data-driven approaches. First, our comprehensive dataset allows for the characterization of the acoustical properties of the marmoset vocal repertoire at a group level, such as investigating whether certain call types often occur together[17]. This strategy has been proposed to instigate inter-species communication differences[13]. Second, future works could leverage such a large amount of data for studying the cortical processing of vocalizations using deep learning. In the last decade, DNN-based representations have emerged as the class of computational model that correlates best with human brain responses to a wide range of auditory stimulations, including speech[18,19], language[20,21], natural sounds[22], and even music[23]. They have also been proven proficient in untangling the neuro-computational mechanisms of face processing in macaques[24–26]. Leading computational neuroscience groups advocated that engineering new bio-inspired DNN-based representations would further help us understand the development, organization, and learning objective of sensory cortical processing[27–31]. A key parameter in training such models is the ecological relevance of the training data[32]. No study has investigated the possible similarity between monkey vocalization cortical processing and the representations learned by DNNs trained on monkey vocalizations. Indeed, training such networks would require substantial data, at least hundreds of thousands of input data[12], and up to millions for state-of-the-art approaches[33,34]–no dataset of this nature is, to date, publicly available for monkey vocalizations. We propose a dataset of 253 hours of segmented marmoset vocalizations acquired during social interactions within three family groups, allowing the above-mentioned training regimes to be used for the first time for supervised and unsupervised training, with the latter seeming to be the most bio-plausible[32]. Such DNN models trained on marmoset vocalizations could then be mapped with existing marmoset brain responses to vocalizations[35–38], using representational similarity analysis[39] (e.g., in[22,40]) or brain-scoring[41]. Third, another approach could be to focus on classifier training using labeled data for passive monitoring in natural settings[42]. Since our vocalizations were recorded in laboratory settings, a model trained with our dataset might need transfer learning to match the distribution of the auditory environment to monitor[43]. Finally, a recent key study in NeuroAI leveraged deep-learning-inspired learning approaches to map behavioral actions with neural activity[44], a line of research that could benefit from using a high number of vocalizations.

However, it is important to acknowledge certain limitations. The dataset lacks information about the sender's identity and the context in which the vocalization was produced. This absence of sender identification restricts the study of information encoded in vocalizations and the study of developmental processes (vocal ontogeny), such as the resemblance of infant to adult call types. Future research could benefit from monitoring systems that capture sender and contextual information.

Methods

Animal retrieval and cares. This study involved a total of thirty-five common marmosets (*Callithrix jacchus*) from a single colony initially structured into three families (same dialect). The marmosets were kept on a 12-hour light/dark cycle that began at 8 a.m. They were fed 3 times per day (pellets around 9 a.m., vegetables around 11 a.m., treats and gum arabic around 3 p.m.). They cannot hide in boxes or anything that can absorb vocalizations. Before entering the housing area, all staff (including researchers, veterinarians, and animal technicians) had to knock on the door to alert the animals and prevent any form of stress. Moreover, they were instructed not to talk in the presence of the marmosets. Most animals were present during the same period, except when a conflict or death occurred, at which point we re-established breeding pairs to maintain the colony. Consequently, while new family groups were formed over time, only three distinct families were in the room at any given moment, housed in three cages (each cage is 1.05 m long × 0.85 m wide × 2 m high). For more details on the periods of inclusion of each monkey, refer to Supplementary Table 1 and Supplementary Fig. 3. All animals included are the offspring of parents and grandparents that were born and raised in captivity for research purposes. All experimental procedures were in compliance with the European directive (2010/63/UE) and were approved by the Ethics Board of Institut de Neurosciences de la Timone (reference 2019010911313842).

Experimental setup. Acoustic recorders were set up in a lab with captive marmosets (Fig. 1). The recordings were made using one microphone (C-100, Sony Corporation, Japan, frequency response of 20 Hz to 50 kHz) placed directly in the room of three marmoset families (e.g., Supplementary Table 1). We used the tracks from the one with the best signal quality. The mixing desk (RME Fireface UFX II, RME, Germany) and the computer allowing the recording via Adobe Audition (Adobe, CA, USA) were located in an adjacent room. Husbandry and technical rooms are soundproofed from the rest of the laboratory animal facility. Audio data was recorded from December 2019 to April 2023, consisting of 997 hours of data recording (wav format, sampling rate: 96 kHz, depth: 32 bit).

Segmentation and labeling. To build a dataset of marmoset vocalizations annotated by type, we followed the pipeline shown in Fig. 2, which comprises three steps: Detection, Cluster-based labeling and Iterative label refinement. Each step is described in this section.

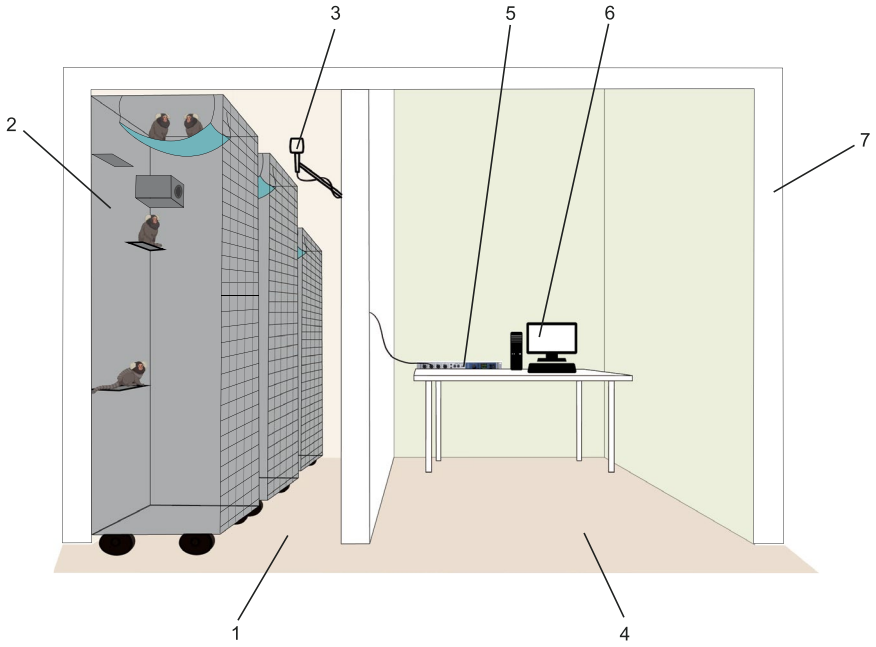

Fig. 1 Schematic of the recording system. The diagram shown here is a schematic drawing of the recording setup, and the relative sizes and positions of the components are not to scale. The husbandry room (1) contained three cages (only one visible here, (2)) and one microphone (3). The technical room (4) was separated by a wall and contained a mixing desk (5) and a computer (6) allowing the recording. Husbandry and technical rooms were soundproof thanks to specialized insulation (7).

The Detection phase first aims to isolate a first pool of vocalizations from background noise (examples of noise in Supplementary Fig. 5). We used a stationary noise reduction algorithm relying on spectral gating (*noisereduce* Python package[13]). We then partially identified (recordings from 2019-2020) the vocalization sound events using a dynamic-thresholding segmentation algorithm[13]. This algorithm dynamically sets a noise threshold based on the anticipated level of silence within a vocal behavior clip. It then identifies syllables as continuous vocal segments separated by noise (for more details, cf[13], *Segmentation*). This first phase allowed us to get approximately 100,000 segmented audio events (including some 'noise' audio events at this stage) (Fig. 2, blue panel; see hyperparameters in Supplementary Table 2). A measurement of the detection and segmentation accuracy is provided and exemplified in Supplementary Fig. 1.

Given the large number of utterances to label, we opted for a semi-automated procedure leveraging unsupervised and self-supervised machine learning strategies to explore the sound event space and label the vocalization types, as well as filter out the noisy sound events (Fig. 2, orange panel). A convolutional auto-encoder (network architecture and particularities of the training procedure are detailed in[12]) was trained on spectrograms of 0.5 seconds long acoustic extracts to encode them into a 16-dimensional latent space, allowing the measurement of vocalization similarity[10,13]. The representations were short-time Fourier Transforms (STFT) with a Hann window of 1,024, no FFT padding, and a hop size of 368), on which we applied a Mel-like bank of 128 triangular filters, logarithmically spread between 1 kHz and 48 kHz. The Mel scale is a popular choice of center frequencies aiming to mimic pitch perception characteristics of the human auditory system, which, in our case, we extend to higher frequencies to fully cover that of marmoset vocalizations. These representations were subsequently treated as points in a feature space after applying the dimensionality reduction algorithm UMAP[45]. We then clustered vocalizations close to one another in feature space, using a density-based algorithm[46], allowing the annotation of vocalizations by type (Fig. 2, orange panel, 'Clustered sound events'). Clusters, which encompass hundreds to thousands of sound events, were meticulously examined by experts. They associated these clusters with specific call types and filtered out any misclassifications. For each cluster, an expert reviewed a folder of spectrogram images, discarding any that did not align with the cluster's general trend. Subsequently, these cluster sounds were categorized either by vocalization type or labeled as 'noise.' This process yielded a partially labeled database, essential for the subsequent iterative label refinement procedure.

After compiling the initial database, we engaged in an iterative process: we trained a classifier and then improved its predictions by visually inspecting and manually correcting multiple spectrograms displayed simultaneously. These spectrograms were sampled from instances where the classifier misidentified labels with high confidence (Fig. 2., green panel).

We continued this process until no outliers were identified for each label. We empirically observed that each label could use the classifier's confidence score as a generalization threshold (Infant cry ≥ 0.5, Phee ≥ 0.7, Seep ≥ 0.86, Trill ≥ 0.86, Tsik ≥ 0.7, and Twitter ≥ 0.7). For example, if the classifier predicts that a given vocalization is a Phee call and the associated confidence score is greater than or equal to 0.7, we retained this prediction. Conversely, any vocalization with prediction confidence below these label-specific thresholds was

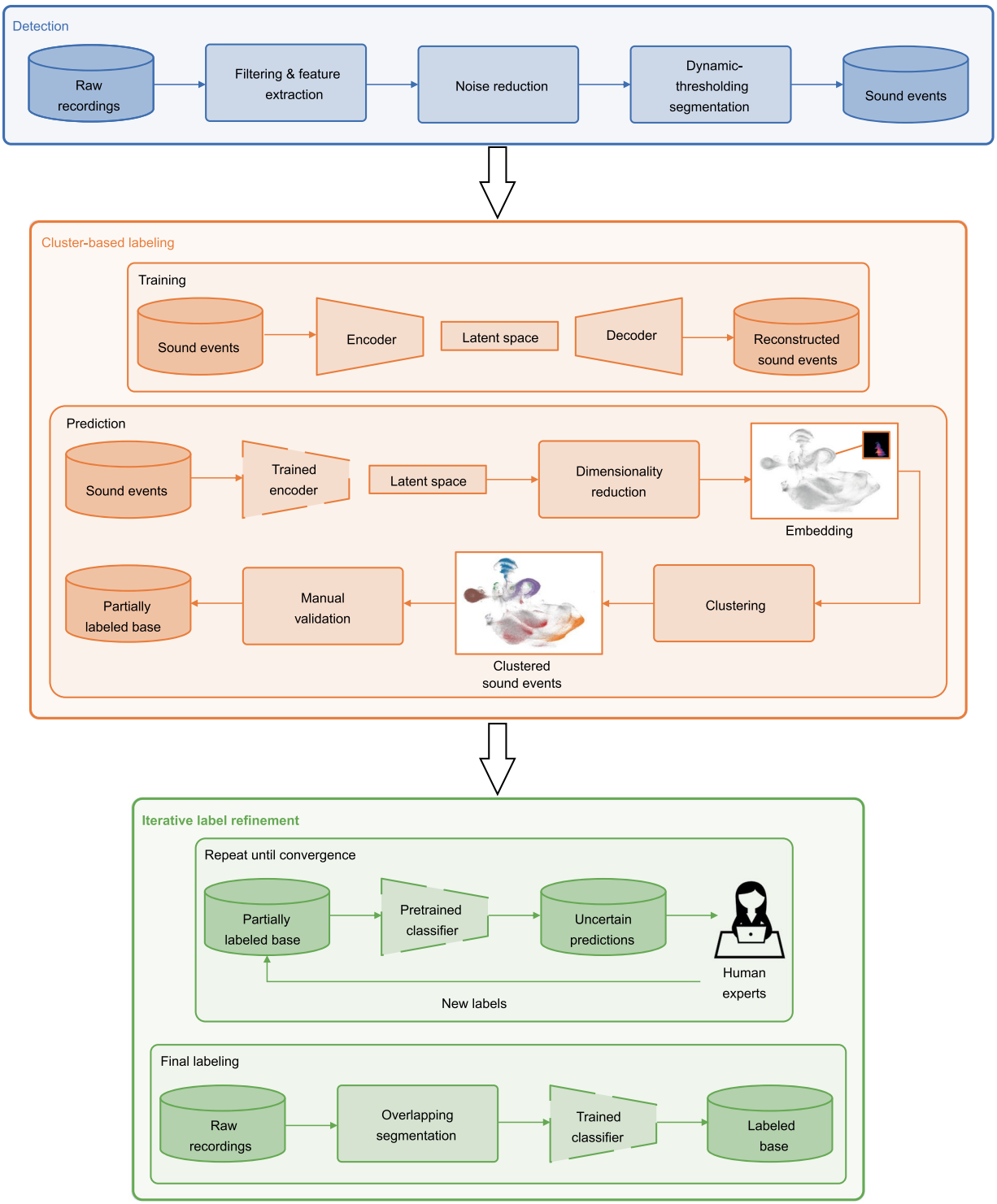

Fig. 2 Pipeline for creating the published database. The Detection phase aims to provide a first pool of segmented vocalizations to start the Cluster-based labeling phase. Later in the process, in the Iterative label refinement phase, all of the raw recordings are passed to the trained classifier.

re-labeled as 'Vocalization' (i.e., a vocalization of unknown type; examples can be found in Supplementary Fig. 6). At this stage, some call types were excluded because the classifier could not classify them robustly due to insufficient representation in the dataset (e.g., composite call types like Seep-Ek and Trill-Phee). Ultimately, we identified the six most certain vocalization types: Infant cry, Phee, Seep, Trill, Tsik, and Twitter (see examples in Supplementary Fig. 4).

Finally, in the last step (Iterative label refinement, box 2), we manually segmented all the raw recordings with overlapping and ran the trained classifier in each segment. We then calculate the maximum confidence points

Fig. 3 Latent projection of vocalizations. For each segmented vocalization, we computed a spectrotemporal representation. Using the trained encoder, we transformed these representations into a 16-dimensional space. From there, we employed the UMAP technique to map the data into latent feature spaces. The colored points denote the predictions where the classifier assigned a high confidence score.

to identify each vocalization onset and offset. At this stage, most of the onset and offset were slightly shifted. To correct for the classifier imprecisions, we empirically adjusted each vocalization's start and end times based on its predicted label post-classification (Supplementary Table 3). As a result of this process, we were able to segment 871,044 vocalizations (253 hours), of which 215,000 (72 hours) were identified as a specific type of vocalization (see Fig. 3 for the latent projection of all the vocalizations, colored by label). To allow users to investigate the dataset further and visualize the call types' relationships, we developed *Marmaudio Explorer*, an interactive visualization interface of the low-dimensional projection of the vocalizations (Supplementary Fig. 9). It allows users to select a group of points, which makes the spectrograms of the corresponding vocalizations appear on the Spectrograms panel, and saves the selected metadata and spectrogram images in a new folder, allowing for deep investigation.

With timestamps for each vocalization, the dataset could offer a starting point for exploring the sequential organization of the marmoset vocal repertoire at the group level, though not at the individual level. See Figs. 4, 5 for visual representations of the vocalizations' distribution; see Supplementary Table 4 for the distribution by label. One additional application could be to monitor the time the vocalizations are produced (see Supplementary Fig. 7 and the code to make it).

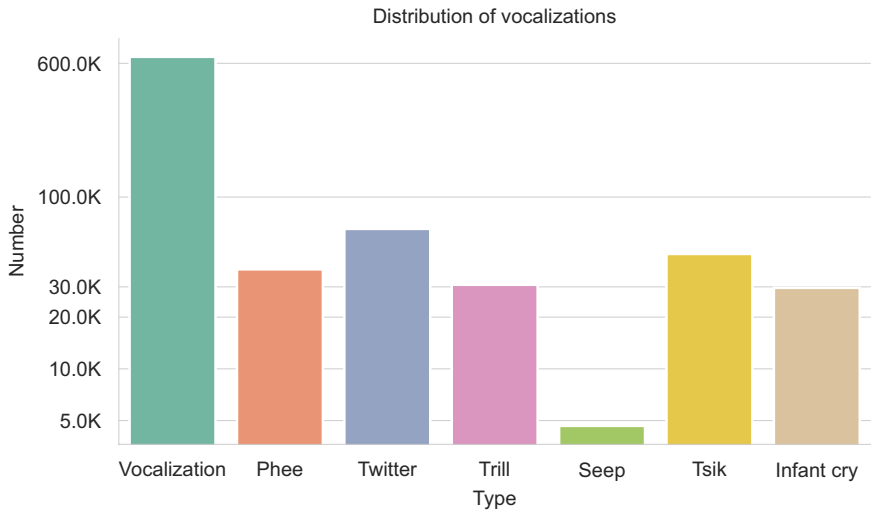

Fig. 4 Distribution of vocalizations. Distribution over call type of 871,044 vocalizations. Seeps are rarer than the other call types. Previous literature has described the Seep as a warning and an alarm call[2,4,51]. Therefore, we do not expect a higher Seep sample size.

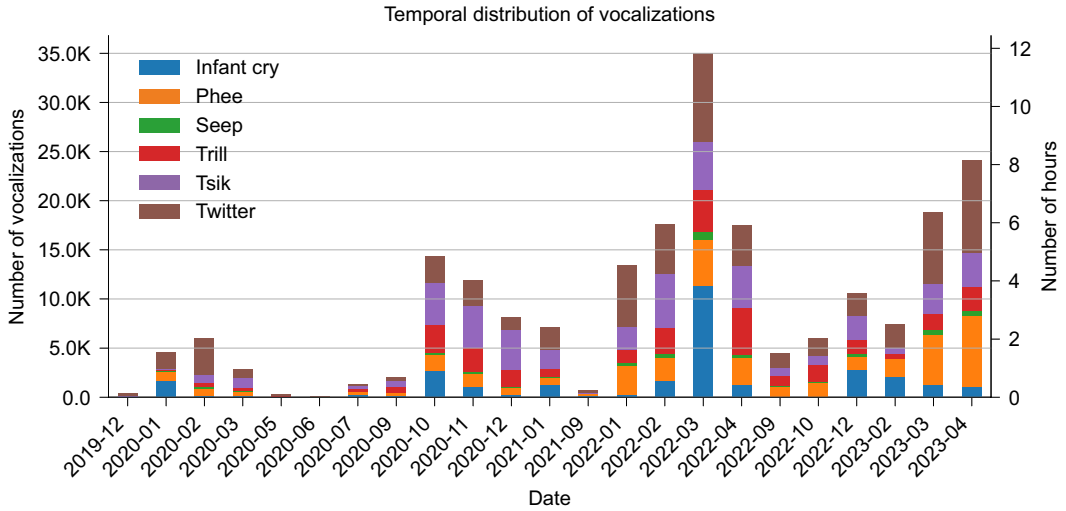

Fig. 5 Temporal distribution of vocalizations over time. Distribution over time of 215,000 labeled vocalizations (72 hours in total). For each month, the proportion of vocalization type is indicated in thousands of vocalizations and in hours. The proportion of labeled/unlabeled vocalization is 25/75% (unlabeled omitted here). Initial recordings began as a proof of concept in 2019–2021, confirming the feasibility of segmenting and labeling vocalizations from raw data. Recording frequency gradually increased in 2022 to capture more detailed patterns. Starting in 2022, precise time information (hour, minute, second, and millisecond) was saved. Temporal variation reflects occasional disturbances in group structure, including conflicts and deaths.

Since several cages are in the same room, a few overlaps can still occur; in such cases, we consider the sound extract 'noisy'. The recording is removed from the analysis to avoid misidentification. However, we know from previous studies there is a turn-taking system in marmoset conversations. For example, marmosets modulate their calls to reduce overlap[47] or favor call productions during periods of silence to avoid overlaps[48].

Data Records

The data are publicly available on the Zenodo data repository[49].The data consist of:

1. Vocalizations.zip: The 871,044 recorded audio files (FLAC format, sampling rate: 96 kHz, depth: 32 bit).
2. Annotations.tsv: The annotation file with 871,044 annotations. These annotations were obtained from the semi-automatic labeling (see above) and include details such as the predicted vocalization type. The content of each column in the annotation file is described in Table 1. Each annotation corresponds to a single vocalization in one file. Most files contain one vocalization, corresponding to one predicted call type—though a few files contain several vocalizations.

Column name	Description
parent_name	Name of the folder containing the vocalization file, using the format YYYY_MM_folder ID
file_name	Name of the.wav file containing the vocalization, using the format Type_ID.flac
label	Type of vocalization as classified by the model: Phee, Trill, Seep, Twitter, Tsik, Infant cry, or Vocalization by default
duration	Length of the vocalization file in seconds
year	Start year of vocalization
month	Start month of vocalization
day	Start day of vocalization
hour	Start hour of vocalization (since 2022)
minute	Start minute of vocalization (since 2022)
second	Start second of vocalization (since 2022)
millisecond	Start millisecond of vocalization (since 2022)

Table 1. Annotation details. Descriptions of each column of the annotation file.

Label	Error rate (%)	Accuracy (%)	CI lower (%)	CI upper (%)
Infant cry	0.00	100.00	100.00	100.00
Phee	0.00	100.00	100.00	100.00
Seep	21.00	79.00	71.00	87.00
Trill	17.00	83.00	76.00	90.00
Tsik	15.00	85.00	78.00	92.00
Twitter	1.00	99.00	97.00	100.00
Vocalization	7.00	93.00	88.00	98.00
Average	8.71	91.29	87.00	95.00

Table 2. Vocalization type error rates. 700 recordings (100 per label type) were re-reviewed by four experts. Errors were noted for inconsistencies or lingering uncertainties.

3. Metadata.pdf: A metadata file that details the annotation definitions, the vocalization type error rates, the description of the recorded subjects, the temporal distribution of vocalizations by label over time, and the individual presence over time with family group (respectively Table 1, 2, Supplementary Table 1, Supplementary Table 4, Supplementary Fig. 3).
4. Raw_audio_example_2020_03_05_0.wav: An example raw audio file of 5 minutes long.
5. Audio_Examples.zip: A set of audio example files (a small subset of representative examples sampled from the recorded audio files).
6. Noise_Examples.zip: A few representative audio example files of noise.
7. Code_Usage.zip: A code folder containing a sample Python code exemplifying data loading and plotting of vocalization spectrograms; a sample Python code exemplifying classifier loading and vocalization type prediction; the associated code to run the examples; a classifier trained on the six vocalization types: Phee, Trill, Seep, Twitter, Tsik, Infant cry (which can also be found in Supplementary Code 1,2).
8. Technical_Validation_Data.zip: 700 representative recordings and cross-examining them with four experts.

The recorded audio files are divided into folders by month of recording, with no more than 10,000 files per folder. The annotation and metadata files are in tabular separated-value format (TSV) to ease their use with automatic tools and allow direct upload into spreadsheet software. The metadata file includes descriptions of all identifiers in the annotation file. The example files contain several audio recordings that illustrate different recorded sounds. They are provided to help users become more familiar with the recorded data. These examples include Phee calls, Twitter calls, Infant cry, and examples of background noises.

Technical Validation

The annotation types were defined by (Manon Obliger-Debouche) M.O.D. and Sabrina Ravel (S.R.). The recordings were annotated semi-automatically by Charly Lamothe (C.L.) and Paul Best (P.B.1). These observers were certified after annotating several days of recordings, which were then validated by an expert. (M.O.D. or SR). 700 annotated recordings (100 per label type) were sampled randomly and were then carefully re-annotated by M.O.D., C.L., P.B.1. and S.R. The annotators were presented with each of the 700 randomly sampled vocalizations along with the prediction of the trained classifier, then had to reply if they agreed with it by yes/no (see example of the annotation interface in Supplementary Fig. 8).

Errors were counted when there was a discrepancy between the post-hoc and the original annotations or when the post-hoc examination concluded that some doubt still existed (e.g., if only 3 out of the 4 confirm it,

it's considered an error, procedure from[50]). The error rate was computed as the proportion of errors divided by the total annotations, resulting in an average of 9.43% (90.57% Confidence-Interval [CI]: 86.00–95.00%) for the vocalization type identification. Accuracy was calculated as 1-error rate (see Table 2 below for scores per label type). We quantified inter-rater reliability using the accuracy across various vocalization categories, showing high accuracy between specific raters (Supplementary Fig. 2).

Code availability

The code is available on https://github.com/swasun/MarmAudioDataset.

Received: 3 July 2024; Accepted: 3 April 2025;
Published online: 13 May 2025

References

1. Miller, C. T. *et al*. Marmosets: A Neuroscientific Model of Human Social Behavior. *Neuron* **90**, 219–233 (2016).
2. Epple, G. Comparative Studies on Vocalization in Marmoset Monkeys (Hapalidae). *Folia Primatologica* **8**, 1–40 (1968).
3. Pistorio, A. L., Vintch, B. & Wang, X. Acoustic analysis of vocal development in a New World primate, the common marmoset (Callithrix jacchus)a. *J. Acoust. Soc. Am.* **120**, 16 (2006).
4. Bezerra, B. M. & Souto, A. Structure and Usage of the Vocal Repertoire of Callithrix jacchus. *Int J Primatol* **29**, 671–701 (2008).
5. Agamaite, J. A., Chang, C.-J., Osmanski, M. S. & Wang, X. A quantitative acoustic analysis of the vocal repertoire of the common marmoset (*Callithrix jacchus*). *The Journal of the Acoustical Society of America* **138**, 2906–2928 (2015).
6. Zhang, Y.-J., Huang, J.-F., Gong, N., Ling, Z.-H. & Hu, Y. Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks. *The Journal of the Acoustical Society of America* **144**, 478–487 (2018).
7. Zhao, L., Roy, S. & Wang, X. Rapid modulations of the vocal structure in marmoset monkeys. *Hearing Research* **384**, 107811 (2019).
8. Turesson, H. K., Ribeiro, S., Pereira, D. R., Papa, J. P. & Albuquerque, V. H. Cde Machine Learning Algorithms for Automatic Classification of Marmoset Vocalizations. *PLOS ONE* **11**, e0163041 (2016).
9. Phaniraj, N., Wierucka, K., Zürcher, Y. & Burkart, J. M. Who is calling? Optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers. *Journal of The Royal Society Interface* **20**, 20230399 (2023).
10. Osmanski, M. S. & Wang, X. Measurement of absolute auditory thresholds in the common marmoset (Callithrix jacchus). *Hearing Research* **277**, 127–133 (2011).
11. Rutz, C. *et al*. Using machine learning to decode animal communication. *Science* **381**, 152–155 (2023).
12. Best, P., Paris, S., Glotin, H. & Marxer, R. Deep audio embeddings for vocalisation clustering. *PLOS ONE* **18**, e0283396 (2023).
13. Sainburg, T., Thielk, M. & Gentner, T. Q. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput Biol* **16**, e1008228 (2020).
14. Sainburg, T. & Gentner, T. Q. Toward a Computational Neuroethology of Vocal Communication: From Bioacoustics to Neurophysiology, Emerging Tools and Future Directions. *Frontiers in Behavioral Neuroscience* **15**, 330 (2021).
15. Remington, E. D., Osmanski, M. S. & Wang, X. An Operant Conditioning Method for Studying Auditory Behaviors in Marmoset Monkeys. *PLOS ONE* **7**, e47895 (2012).
16. *Marmosets and Tamarins: Systematics, Behaviour, and Ecology*. (Oxford University Press, Oxford, New York, 1993).
17. Bosshard, A. B. *et al*. Beyond bigrams: call sequencing in the common marmoset (Callithrix jacchus) vocal system. *R Soc Open Sci* **11**, 240218 (2024).
18. Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V. & McDermott, J. H. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98**, 630–644. e16 (2018).
19. Tuckute, G., Feather, J., Boebinger, D. & McDermott, J. H. Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *PLOS Biology* **21**, e3002366 (2023).
20. Caucheteux, C., Gramfort, A. & King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat Hum Behav* 1–12, https://doi.org/10.1038/s41562-022-01516-2 (2023).
21. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun Biol* **5**, 134 (2022).
22. Giordano, B. L., Esposito, M., Valente, G. & Formisano, E. Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nat Neurosci* 1–9, https://doi.org/10.1038/s41593-023-01285-9 (2023).
23. Güçlü, U., Thielen, J., Hanke, M., van Gerven, M. A. J. & van Gerven, M. A. J. Brains on Beats. *In Proceedings of the International Conference on Neural Information Processing Systems* 2101–2109 (2016).
24. Higgins, I. *et al*. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat Commun* **12**, 6456 (2021).
25. Zhuang, C. *et al*. Unsupervised neural network models of the ventral visual stream. *Proc Natl Acad Sci USA* **118**, e2014196118 (2021).
26. Dado, T. *et al*. Brain2GAN: Feature-disentangled neural encoding and decoding of visual perception in the primate brain. *PLOS Computational Biology* **20**, e1012058 (2024).
27. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356–365 (2016).
28. Richards, B. A. *et al*. A deep learning framework for neuroscience. *Nat Neurosci* **22**, 1761–1770 (2019).
29. Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J. & Hinton, G. Backpropagation and the brain. *Nat Rev Neurosci* **21**, 335–346 (2020).
30. Amunts, K. *et al*. *The Coming Decade of Digital Brain Research - A Vision for Neuroscience at the Intersection of Technology and Computing*. https://zenodo.org/record/7764003 (2023).
31. Zador, A. *et al*. Catalyzing next-generation Artificial Intelligence through NeuroAI. *Nat Commun* **14**, 1597 (2023).
32. Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N. & Kietzmann, T. C. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc Natl Acad Sci USA* **118**, e2011417118 (2021).
33. Baevski, A., Zhou, H., Mohamed, A. & Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *NIPS'20: 34th International Conference on Neural Information Processing Systems* (2020).
34. Défossez, A., Copet, J., Synnaeve, G. & Adi, Y. High Fidelity Neural Audio Compression. *Transactions on Machine Learning Research* (2023).
35. Sadagopan, S., Temiz-Karayol, N. Z. & Voss, H. U. High-field functional magnetic resonance imaging of vocalization processing in marmosets. *Sci Rep* **5**, 10950 (2015).
36. Jafari, A. *et al*. A vocalization-processing network in marmosets. *Cell Reports* **42**, 112526 (2023).
37. Belin, P., Trapeau, R. & Obliger-Debouche, M. A small, but vocal, brain. *Cell Reports* **42**, 112651 (2023).
38. Dureux, A., Zanini, A. & Everling, S. Mapping of facial and vocal processing in common marmosets with ultra-high field fMRI. *Commun Biol* **7**, 1–15 (2024).

39. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Sys. Neurosci*. https://doi.org/10.3389/neuro.06.004.2008 (2008).
40. Giordano, B. L. *et al*. The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nat Hum Behav* https://doi.org/10.1038/s41562-021-01073-0 (2021).
41. Schrimpf, M. *et al. Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like?* http://biorxiv.org/lookup/doi/10.1101/407007 (2018).
42. Morgan, M. M. & Braasch, J. Long-term deep learning-facilitated environmental acoustic monitoring in the Capital Region of New York State. *Ecological Informatics* **61**, 101242 (2021).
43. Dufourq, E., Batist, C., Foquet, R. & Durbach, I. Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics* **70**, 101688 (2022).
44. Schneider, S., Lee, J. H. & Mathis, M. W. Learnable latent embeddings for joint behavioural and neural analysis. *Nature* **617**, 360–368 (2023).
45. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3**, 861 (2018).
46. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *JOSS* **2**, 205 (2017).
47. Zhao, L., Rad, B. B. & Wang, X. Long-lasting vocal plasticity in adult marmoset monkeys. *Proc Biol Sci* **286**, 20190817 (2019).
48. Roy, S., Miller, C. T., Gottsch, D. & Wang, X. Vocal control by the common marmoset in the presence of interfering noise. *Journal of Experimental Biology* **214**, 3619–3629 (2011).
49. Lamothe, C. *et al*. MarmAudio: A large annotated dataset of the common marmoset vocal repertoire, *Zenodo*, https://doi.org/10.5281/zenodo.15017207 (2025).
50. Prat, Y., Taub, M., Pratt, E. & Yovel, Y. An annotated dataset of Egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Sci Data* **4**, 170143 (2017).
51. Lazaro-Perea, C. Intergroup interactions in wild common marmosets, *Callithrix jacchus*: territorial defence and assessment of neighbours. *Animal Behaviour* **62**, 11–21 (2001).

Acknowledgements

We thank Joël Baurberg for his assistance in reducing electromagnetic noise in the recordings. We thank BaNCo and Qarma team members for their useful discussions. We thank Jules Collenne, William Jonas and Rami Soussi for their contributions during their internships. This work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 788240). 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). PB1's contribution was funded by the ADSIL AI Chair 471 on bioacoustics ANR-20-CHIA-0014.

Author contributions

Charly Lamothe (C.L.) and Manon Obliger-Debouche contributed equally to this work. C.L., M.O.D., R.T., and P.B.2 (Pascal Belin) conceived and designed the experiments. M.O.D. conducted the experiments. C.L. and P.B.1 (Paul Best) created the processing and analysis tools and assembled the database. Thierry Artières (T.A.), Ricard Marxer (R.M.), P.B.2 and S.R. supervised the study. C.L., M.O.D. and P.B.1 wrote the manuscript and P.B.2, R.M. reviewed it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04951-8.

Correspondence and requests for materials should be addressed to C.L., M.O.-D. or P.B.