Who is calling? Optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers

Phaniraj, Nikhil ; Wierucka, Kaja ; Zürcher, Yvonne ; Burkart, Judith M

DOI: https://doi.org/10.1098/rsif.2023.0399

Who is calling? Optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers

Author(s):
Phaniraj, Nikhil; Wierucka, Kaja; Zürcher, Yvonne; Burkart, Judith M.

Publication date:
2023-10

Permanent link:
https://doi.org/10.3929/ethz-b-000639047

Originally published in:
Journal of the Royal Society. Interface 20(207), https://doi.org/10.1098/rsif.2023.0399

INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Phaniraj N, Wierucka K, Zürcher Y, Burkart JM. 2023 Who is calling? Optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers. *J. R. Soc. Interface* **20**: 20230399. https://doi.org/10.1098/rsif.2023.0399

Received: 12 July 2023
Accepted: 25 September 2023

Subject Category:
Life Sciences–Mathematics interface

Subject Areas:
computational biology, evolution

Keywords:
machine learning, hierarchical classifier, marmoset calls, bioacoustics, time series analysis, source identification

Author for correspondence:
Nikhil Phaniraj
e-mail: nikhil.phaniraj@uzh.ch

Electronic supplementary material, is available online at https://doi.org/10.6084/m9.figshare.c.6872089.

THE ROYAL SOCIETY PUBLISHING

Who is calling? Optimizing source identification from marmoset vocalizations with hierarchical machine learning classifiers

Nikhil Phaniraj[1,2,3], Kaja Wierucka[1,4], Yvonne Zürcher[1] and Judith M. Burkart[1,2,5]

[1]Institute of Evolutionary Anthropology (IEA), University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[2]Neuroscience Center Zurich (ZNZ), University of Zurich and ETH Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[3]Department of Biology, Indian Institute of Science Education and Research (IISER) Pune, Dr. Homi Bhabha Road, Pune 411008, India
[4]Behavioral Ecology & Sociobiology Unit, German Primate Center, Leibniz Institute for Primate Research, Kellnerweg 4, 37077 Göttingen, Germany
[5]Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich, Affolternstrasse 56, 8050 Zürich, Switzerland

NP, 0000-0002-9633-8705; JMB, 0000-0002-6229-525X

With their highly social nature and complex vocal communication system, marmosets are important models for comparative studies of vocal communication and, eventually, language evolution. However, our knowledge about marmoset vocalizations predominantly originates from playback studies or vocal interactions between dyads, and there is a need to move towards studying group-level communication dynamics. Efficient source identification from marmoset vocalizations is essential for this challenge, and machine learning algorithms (MLAs) can aid it. Here we built a pipeline capable of plentiful feature extraction, meaningful feature selection, and supervised classification of vocalizations of up to 18 marmosets. We optimized the classifier by building a hierarchical MLA that first learned to determine the sex of the source, narrowed down the possible source individuals based on their sex and then determined the source identity. We were able to correctly identify the source individual with high precisions (87.21%–94.42%, depending on call type, and up to 97.79% after the removal of twins from the dataset). We also examine the robustness of identification across varying sample sizes. Our pipeline is a promising tool not only for source identification from marmoset vocalizations but also for analysing vocalizations of other species.

1. Introduction

Comparative studies of primate communication are crucial for understanding the evolutionary origins of human speech [1–4]. Much progress has been achieved over the last decades, mainly by recording and simultaneously annotating the vocalizer or physically separating social partners to track individual contributions to conversations. The vocal communication of callitrichid monkeys appears particularly rich among non-human primates, with several features thought to be precursors of language. For instance, common marmosets (*Callithrix jacchus*) have been shown to possess superior control and flexibility in their calls, one of the requirements for speech. Among others, they can actively interrupt ongoing calls [5,6], make long-term changes to call frequency in response to noise [5], converge in vocal space to a social partner

[7] and engage in antiphonal call conversations similar to antiphonal speech in humans [8]. As immatures, they go through a babbling phase [9], and contingent vocal feedback from caregivers contributes to fully developing their repertoire [10,11], not described in any other primate and reminiscent of vocal learning—an essential building block for language. Marmosets may be particularly relevant because, like humans, they are cooperative breeders, which may have played a major role in language evolution [12]. However, our knowledge of how marmosets and other primates communicate under more naturalistic situations and beyond dyadic contexts is rather limited (but see [13]). Given the highly social nature of marmosets, the next frontier in studying marmoset vocal communication is to do so under more naturalistic conditions—in social groups, where they fully display their communication skills.

A major bottleneck for studying group-level communication in naturalistic settings without separating animals is to accurately determine the vocalizer (source identity). There are at least two approaches to doing this. On the one hand, microphone arrays can localize sounds in three-dimensional space. This information can then be integrated with visual monitoring or signal-based individual localization to match the vocalization to its source individual. This is mostly feasible in captive conditions. On the other hand, a more broadly applicable alternative is to use individual signatures in the vocalizations to classify calls.

Previous attempts to determine source identity from animal vocalizations have used a broad range of unsupervised (e.g. k-means [14], Gaussian mixture models [14], Bayesian functional mixed models [15], hidden Markov models [16]) and supervised machine learning (ML) classifiers (e.g. discriminant function analyses [17], support vector machines [18], random forests [19], artificial neural networks [20]) mostly with mixed results (16.26–91.5% accuracy, but see [21,22] for instances of greater than 95% accuracy). To improve classifier performance, the focus has largely been on hyperparameter optimization [23], feature selection [24] and data quality enhancement techniques [25]. However, for instances of supervised classification, training datasets available to researchers often contain not only individual identities but additional information such as the sex, age class or social status of the individual. Such additional information could aid in improving classifier performance when implemented in hierarchical machine learning algorithms (MLAs). Furthermore, increasing the number of features extracted before feature selection would give the classifier a more detailed representation of the vocalizations and a larger feature space. The current study aims to demonstrate this for source identification from marmoset vocalizations by addressing the two issues with (i) plentiful feature extraction to achieve detailed representations of the calls and (ii) hierarchical MLAs that take sex into account as the first hierarchical layer.

First, the traditional approach to representing animal vocalizations involves spectral feature extraction. Specific spectral features are chosen because we understand how sound modifications affect these feature values, making them easy to interpret. Software like Raven (The Cornell Lab of Ornithology, Ithaca, NY, USA), Avisoft SASLab Pro (Avisoft Bioacoustics, Germany) and Kaleidoscope (Wildlife Acoustics Inc., USA) offer easy solutions to extracting these features and have been widely used along with custom scripts for feature extraction from animal vocalizations. A drawback of this approach is that not all animal vocalizations show maximum variability along these 'pre-selected' feature values. The small feature space and reduced class separability of calls due to limited features prevent ML approaches from achieving their maximum potential for obtaining high classification accuracies. One workaround is to initially extract a vast number of acoustic features and let the MLAs decide which ones would be most helpful for their classification task. For this, recent advances in time series analyses allow for multiple operations to be performed that provide meaningful information about the nature of the time series [26]. This can be exploited by viewing the acoustic waveform as a time series of pressure points and performing time-series analyses on acoustic data. In contrast to spectrogram-based feature extraction software, time series analyses can extract up to 7700 features from a single vocalization [27]. In this paper, we implement time-series analysis for plentiful feature extraction and use tree-based classifiers to extract the most meaningful features. This enables us to provide detailed representations of the marmoset calls as input for the classifier.

Second, we hypothesize that using a hierarchical ML classifier will improve classification accuracies by breaking up the classification problem into a hierarchy of smaller ones. Callitrichid vocalizations contain information about the sex of the caller [28–30], which can potentially assist MLAs with efficient source identification. We, therefore, develop a hierarchical ML classifier that first determines the sex of the source, thus narrowing down the possible source individuals, and then determines the source identity. We compare its performance with the non-hierarchical approach by applying it to three major marmoset call types relevant to group coordination and affiliative behaviours: trills, phees and food calls. Trills are used as within-group close-contact signals. They are between 0.3 and 0.8 s in duration and are characterized by quick sinusoidal variations in pitch between 5 and 8 kHz throughout the call [31]. Phees are louder long-distance contact calls with variable durations between 0.5 and 2 s and a narrow frequency band between 6 and 8 kHz [31]. Food calls are extremely short (approx. 0.05 s), steep downsweeps (10–6 kHz) elicited by individuals when they see food, during feeding and to initiate food sharing [31,32]. Example spectrograms for each call type are provided in figure 1. Finally, we analyse the acoustic features used by the hierarchical classifier for the various steps in the hierarchy and compare the precisions and recalls of the non-hierarchical and hierarchical classifiers at different sample sizes (i.e. number of calls per individual).

2. Methods

2.1. Experimental subjects

This study used marmoset vocalizations collected by Zürcher *et al*. [7]. The data contained vocalizations from 20 adult common marmosets, 10 males and 10 females. They were housed in 2.4 × 1.8 × 3.6 m enclosures with at least one other individual, with each group having access to a personal outdoor enclosure of the same dimension and a common experimental room. Lighting was regulated to maintain a 12/12 h day/night cycle. Animals were fed a predetermined amount of vitamin-enriched mush in the morning, vegetables and fruits during noon, and one of either gum, boiled egg, cottage cheese or insects after noon. Ad libitum access to water was always provided. All

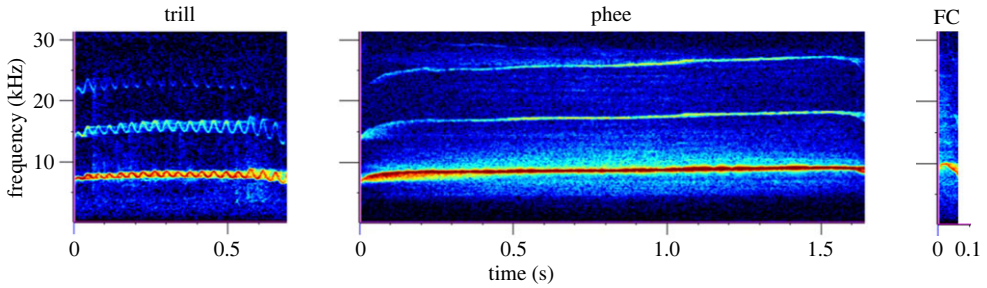

Figure 1. Marmoset call types. Example spectrograms of a single trill, phee and food call (FC, 0.08s in length) of common marmosets. Spectrograms were obtained using the Hann filter of size 512 samples, hop length of 256 samples and discrete Fourier transform (DFT) bin size of 512 samples.

experiments were approved by Zürich's cantonal veterinary office (licence ZH223/16).

2.2. Vocalization recordings and segmentation

Individuals were recorded in their home enclosures or in a separate experimental room in sessions lasting for approximately 30 min. Each individual went through approximately 17 recording sessions (17.3 ± 15.5 mean ± s.d.) spread over approximately 8 days (7.6 ± 7.6 mean ± s.d.). As phee calls are long-distance contact calls, focal individuals were separated from their group to elicit them. In this case, the focal individual was only visually isolated but acoustically in contact with other group members. For eliciting food calls, the focal individual's highly preferred food was provided until enough vocalizations were obtained during that session. Trills could be recorded without any such interference. A condenser microphone (CM16/CMPA, Avisoft Bioacoustics, Germany) connected to Avisoft UltraSoundGate 116H (Avisoft Bioacoustics, Germany) was used for recordings, and calls were labelled in real time using Avisoft Recorder (Avisoft Bioacoustics, Germany). Calls were recorded at a sampling rate of 62 500 Hz and segmented manually using Avisoft Pro (Avisoft Bioacoustics, Germany). The start and end of calls were determined by visual inspection of the spectrogram. Only those calls that were visible clearly on the spectrogram, had no interference with other calls and could be accurately assigned to the categories of trill, phee or food call were included for further analyses. See [7,33] for detailed information about the recording procedure and processing.

2.3. Datasets and feature extraction

An imbalanced dataset with a skewed data distribution biases the classifier towards the majority class, often preventing it from learning the underlying patterns that make the classes different and limiting its generalizability. Such a problem can be solved by balancing the dataset. As the original dataset was imbalanced (the number of calls per call type per individual was highly variable), a combination of majority class random undersampling and synthetic minority oversampling technique (SMOTE) [34,35] was used to create 18 datasets (one original and five generated, for each of the three call types). SMOTE is a data augmentation technique that synthesizes new data for minority classes without repeating the original datapoints to make them equal to the majority class. For this, first, a call belonging to a minority class was randomly selected (observation). Next, the nearest minority class 'neighbours' of this call in the high-dimensional feature space (and not necessarily belonging to the same individual) were determined. Then, the feature vector of the observation was subtracted from that of the nearest neighbours, multiplied by a random number between 0 and 1, and added to the feature vector of the observation. This effectively synthesized new data points within the hypervolume bounded by the neighbours. The datasets are listed as follows (X = T for trills, P for phees, F for food calls in the name of the dataset):

1. Original-X: The original dataset after feature extraction. Consisted of 1247 trills, 1443 phees and 4434 food calls from 20 individuals each.
2. Imbalanced-X: Marmosets with less than 25 calls per call type were removed from the Original-X datasets to make them suitable for ML processing. No undersampling was done. Consisted of 1207 trills from 16 individuals, 1374 phees from 10 individuals and 4419 food calls from 18 individuals.
3. Balanced-X: SMOTE was applied on Imbalanced-X to obtain balanced datasets. Consisted of 4528 trills, 3350 phees and 15 372 food calls.
4. Balanced197-X: From the Imbalanced-X dataset, classes with greater than 197 calls were undersampled to 197 calls. SMOTE was applied to this. Consisted of 3152 trills, 1970 phees and 3546 food calls.
5. Balanced99-X: From the Imbalanced-X dataset, classes with greater than 99 calls were undersampled to 99 calls. SMOTE was applied to this. Consisted of 1584 trills, 990 phees and 1782 food calls.
6. Balanced50-X: From the Imbalanced-X dataset, classes with greater than 50 calls were undersampled to 50 calls. SMOTE was applied to this. Consisted of 800 trills, 500 phees and 900 food calls.

The smaller Balanced197-X, Balanced99-X and Balanced50-X datasets were used to test the capability of the ML approach to classify calls in limited sample size scenarios.

For feature extraction, the MATLAB-based highly comparative time series analysis (HCTSA) [27] toolbox provides an architecture to extract over 7700 features from every call, and we implemented this for marmoset calls. When inputted, HCTSA views the acoustic waveform as a time series of pressure points, performs several time-series analyses on acoustic data, and provides a matrix of feature measurements. Features common across calls of all individuals for a given call type were used for further analyses. As every call is a point in a very high-dimensional feature space, we required a dimensional reduction technique to visualize the trill, phee and food call datasets. We did so using t-distributed stochastic neighbour embedding (t-SNE) [36], an unsupervised MLA for nonlinear dimensional reduction.

We trained individual multi-class adaptive boosting algorithms with decision trees as weak learners and 10-fold cross-validation (henceforth AdaBoost) for determining the important features to use for classification and for performing the classification itself. This method uses multiple weak learners to create a strong learner [37]. Unlike random forests, in which multiple trees are trained in parallel, and their collective decision (using a 'voting' system) is obtained, AdaBoost trains trees in a sequential manner wherein every new tree aims to be specialized

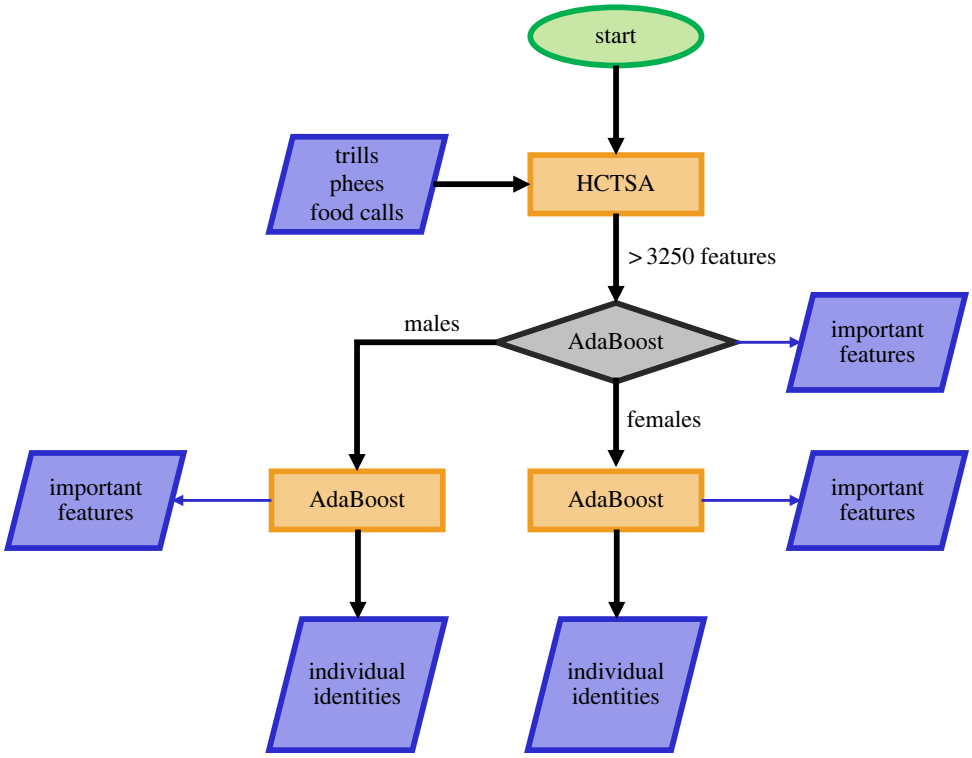

Figure 2. The hierarchical classification approach. Features from trills, phees, and food calls were extracted using HCTSA. These features were used to train AdaBoost to first classify calls based on sex and then individual identities. The oval denotes start, parallelograms inputs/outputs, rectangles processes, and the diamond a decision.

in correcting the errors of a previous tree. AdaBoost is considered better than random forest classifiers due to its higher accuracy and lower susceptibility to overfitting [38,39]. We first trained AdaBoost (using MATLAB's 'fitcensemble', 'AdaboostM1', and 'AdaboostM2' functions) on Imbalanced-X, Balanced-X, Balanced197-X, and Balanced99-X datasets to classify calls based on source identity—as a direct or 'non-hierarchical approach' (in contrast to the hierarchical approach that was used later)—to determine source identity from calls. We chose the number of trees and learning rate based on the observations of the classification loss function.

Although classification accuracy is the most widely used metric to assess MLAs, it does not represent the model's performance on class-imbalanced datasets. This is because the classifier can get away with a high accuracy score by simply predicting most data points as belonging to the majority class. In such cases, the receiver operating characteristic (ROC) curve can be used to evaluate the classifier's performance. The ROC curve visualizes how the true positive rate changes as a function of the false positive rate at various threshold values. The area under this ROC curve, simply 'area under curve' (AUC), can be a useful tool for examining classifier performance along with accuracy. Therefore, for assessing the performance of AdaBoost on Imbalanced-X datasets, ROC-AUC was calculated in a one versus rest setting for each class in the dataset, along with the accuracies.

Even while assessing the performance of MLAs on balanced datasets, accuracy does not represent how variable the predictions for each class (marmoset individuals) are. Class-specific precision and recall values represent individual-specific performance of the classifier and how they vary across individuals. Therefore, for the rest of the classifiers, precisions and recalls for each class were calculated using the formulae below, and the summary (means and standard deviations) of these values was used to assess their performance.

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

and

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}.$$

For inspecting if individual variability of calls within marmoset groups can be explained by variation in sex and whether MLAs could exploit this to perform better, individual AdaBoosts were trained on Imbalanced-X and Balanced-X datasets to classify calls based on sex, and then the source individual. Sex was chosen as a cue because previous studies in callitrichids have shown that they can discriminate calls based on the sex of the source [28–30], and in our case, the total number of classes could be split into half based on the sex of the individual (10 males and 10 females out of 20 individuals). Later, each dataset was divided into two sub-datasets based on the 'true' sex of the source individual, and separate AdaBoosts were trained on each of them. This was the hierarchical approach to determine first the sex and then the source identity from calls (figure 2). To assess the performance of the classifiers for the hierarchical classification approach, precisions and recalls for each individual were calculated for all classifiers as

$$Y^{\text{final}}(\text{sex},\text{ind}) = Y(\text{sex}) \times Y(\text{ind} \mid \text{sex}).$$

Y is the precision or recall, and ind is the individual identity of the marmoset. For example, the final precision for a female individual would be the precision for determining the sex as a female, multiplied by the precision for determining the source individual among females.

We performed Wilcoxon signed-rank test to compare precision and recall scores for every class between the two approaches because the classes, i.e. the individuals for both approaches, were the same.

2.4. Feature importance scoring

Along with classifying calls, each fold of AdaBoost also provides predictor importance scores for each feature, which represents

Table 1. Sample sizes of the datasets.

	trills			phees			food calls		
dataset	female	male	total	female	male	total	female	male	total
Original-X	895	352	1247	565	878	1443	1906	2528	4434
Imbalanced-X	895	312	1207	533	841	1374	1904	2515	4419
Balanced-X	2547	1981	4528	1420	1930	3350	7686	7686	15 372
Balanced197-X	1773	1379	3152	788	1182	1970	1773	1773	3546
Balanced99-X	891	693	1584	396	594	990	891	891	1782
Balanced50-X	450	350	800	200	300	500	450	450	900

Table 2. Non-hierarchical AdaBoost performance for imbalanced and balanced datasets. Imbalanced-X and Balanced-X datasets were used. Mean ± s.d. are provided for ROC-AUC scores.

		imbalanced			balanced		
call	classes (individuals)	sample size	ROC-AUC (%)	accuracy (%)	sample size	ROC-AUC (%)	accuracy (%)
trills	16	1206	92.63 ± 4.30	64.84	4528	98.95 ± 0.91	82.91
phees	10	1498	94.36 ± 4.54	70.96	3550	98.55 ± 1.42	83.92
food calls	18	4419	92.72 ± 5.08	60.8	15 372	97.04 ± 2.31	71.43

how important that feature was for AdaBoost in the classification task. First, we checked how (non-hierarchical) AdaBoost performed the feature selection task. For this, we ranked features by predictor importance scores given by the most accurate of the 10 models (run on 10 different folds) in the AdaBoost for each dataset. Then, for the three Balanced-X datasets, we used the top-20 features to visualize t-SNE clusters. t-SNE plots were generated using MATLAB's 'tsne' function with the Barnes–Hut algorithm keeping the Barnes–Hut trade-off parameter at 0.5 to increase processing speed for large datasets. Exaggeration was set to 4, perplexity to $n/100$, and learning rate to $n/12$ (where n is the total number of calls for that call type) as these values are shown to provide robust results when datasets are large [40]. We compared these with t-SNE plots of the corresponding datasets generated using 20 random features. For a more quantitative comparison, we selected 20 random features from each of the three datasets 100 times, plotted the histograms of the mean silhouette scores, fit Gaussians to these distributions, and calculated the probability of getting a mean silhouette score greater than that of the top-20 features by chance. For the hierarchical classifiers, we analysed the features used by the various levels in the hierarchy when implemented on the three Balanced-X datasets.

3. Results

3.1. Datasets and feature extraction

The sample sizes of each of the datasets obtained are listed in table 1. The Original-X and Imbalanced-X datasets have variable number of calls per call type and individual (electronic supplementary material, table S1).

We could extract between 3776 and 4553 features from each marmoset call in the Original-X datasets, with 3255, 3395 and 3477 features common across the calls of all individuals for trills, phees and food calls, respectively.

3.2. Machine learning classifiers and feature importance scoring

We monitored the classification loss function of AdaBoost with the addition of every new weak learner. We observed the loss function to plateau at approximately 500 trees while training AdaBoost to determine sex and approximately 2500 for other tasks, and these number of trees were therefore used.

3.2.1. Balanced versus unbalanced datasets

Classification accuracies ranged from 60.8% to 70.96% for imbalanced datasets versus 71.43% to 83.92% for balanced datasets (chance probabilities = 6.25% for trills, 10% for phees and 5.56% for food calls). Mean ROC-AUCs were higher for all the Balanced-X datasets compared with Imbalanced-X datasets (table 2).

3.2.2. Top-20 versus random-20 features

t-SNE plots obtained using top-20 features on the Balanced-X datasets showed visibly better clusters compared with t-SNE plots obtained using random-20 features on the corresponding datasets (figure 3, electronic supplementary material, tables S2–S4). The top-20 features provided by AdaBoost gave significantly greater silhouette scores than what would be obtained by chance for trills ($Z = 7.0154$, $p < 0.001$), phees ($Z = 4.8361$, $p < 0.001$) and food calls ($Z = 6.8922$, $p < 0.001$).

3.2.3. Hierarchical versus non-hierarchical classifiers at large sample sizes

The accuracies of AdaBoost to classify the following datasets based on sex were: Balanced-T = 99.4%, Balanced-p = 96.8%, Balanced-F = 96.7%. The hierarchical classification approach

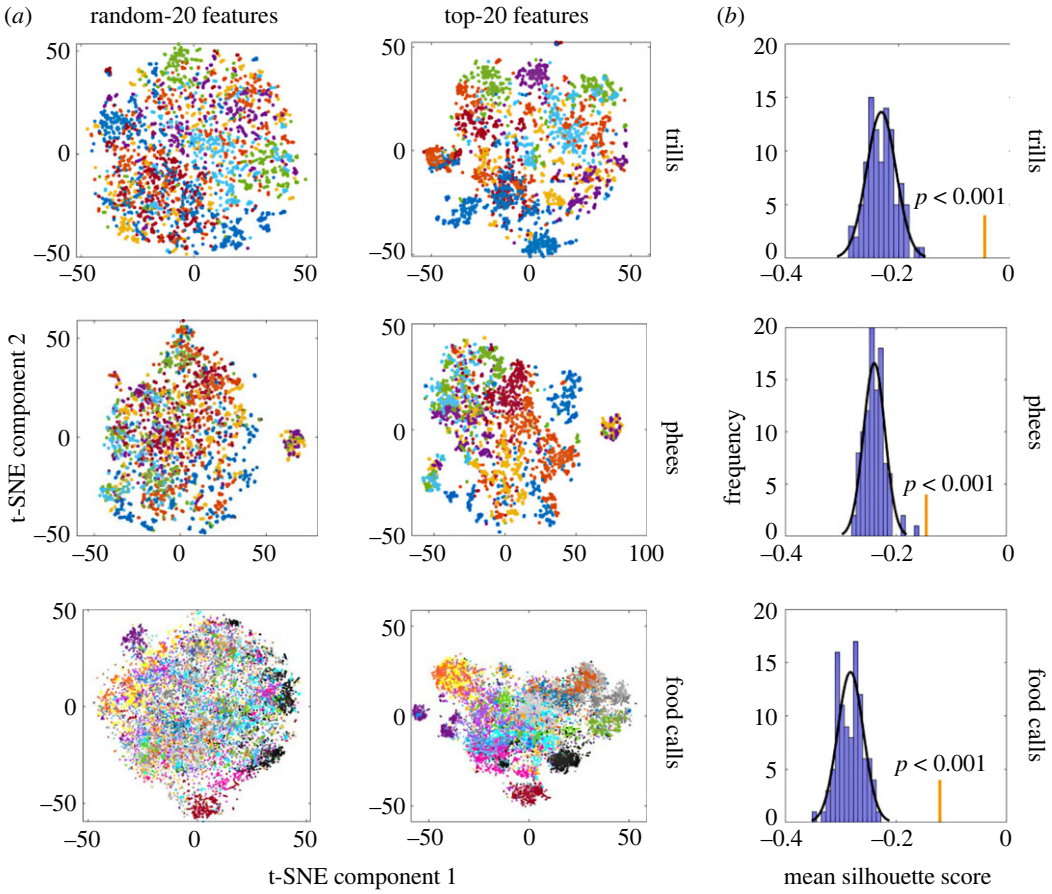

Figure 3. Qualitative and quantitative comparisons of random-20 and top-20 features for clustering data. (*a*) Qualitative comparison. Figures are t-SNE plots (squared Euclidean distance metric) of Balanced-X datasets using 20 randomly chosen features or the top-20 features for classification by AdaBoost for that call type. Each point is a call, coloured according to the source individual. (*b*) Quantitative comparison. Blue bars depict frequencies (histogram) of mean silhouette scores obtained after performing t-SNE using 20 random features selected 100 times for that call type (trills/phees/food calls) on Balanced-X datasets. The grey line depicts the Gaussian function fit to the histogram. Orange vertical bars denote the mean silhouette scores obtained after performing t-SNE using top-20 features for classification by AdaBoost for that call type. The *p*-value shown is the normalized area under the Gaussian function to the right of the orange bar.

Table 3. Comparing non-hierarchical and hierarchical approaches for classifying calls based on source identity. Mean ± s.d. precisions and recalls with corresponding *p*-values for testing the hypotheses: mean precisions/recalls of non-hierarchical = hierarchical. Both approaches were tested using the Balanced-X datasets.

		non-hierarchical		hierarchical		*p*-value	
call	classes (individuals)	precision (%)	recall (%)	precision (%)	recall (%)	for precision	for recall
trills	16	83.42 ± 11.48	82.87 ± 8.63	94.42 ± 9.61	94.19 ± 8.24	0.017	0.01
phees	10	84.15 ± 9.84	83.93 ± 8.63	92.57 ± 3.30	92.55 ± 3.79	0.027	0.037
food calls	18	72.23 ± 15.09	71.43 ± 13.13	87.21 ± 9.71	86.65 ± 5.62	0.01	0.002

gave significantly better precision and recall scores than the non-hierarchical approach for the corresponding Balanced-X datasets ($p < 0.05$ across all call types, Wilcoxon signed-rank test; table 3). A thorough evaluation of the precisions and recalls for each individual by the hierarchical classifier revealed that the trills of two individuals—Washington and Wisconsine (female twins)—had significantly lower scores (precisions: 74.23% and 67.46%, respectively, recalls: 75.07% and 72.99%, figure 4). The mean precision and recalls for trills for all except these two individuals were as high as 97.79% and 97.07%, respectively. The phee precision and recall scores for Washington and Wisconsine were also slightly lower than that of other females (electronic supplementary material, figures S1*a*,*b*). For food calls, the precision and recall scores of these two individuals were similar to or higher than the mean scores of all females (electronic supplementary material, figure S1*c*,*d*).

3.2.4. Hierarchical versus non-hierarchical classifiers at reduced sample sizes

With the decrease in sample size per class, the difference between the mean precisions and recalls of the hierarchical and non-hierarchical approaches was reduced for all three call types (figure 5 for precision, electronic supplementary material, figure S3 for recall).

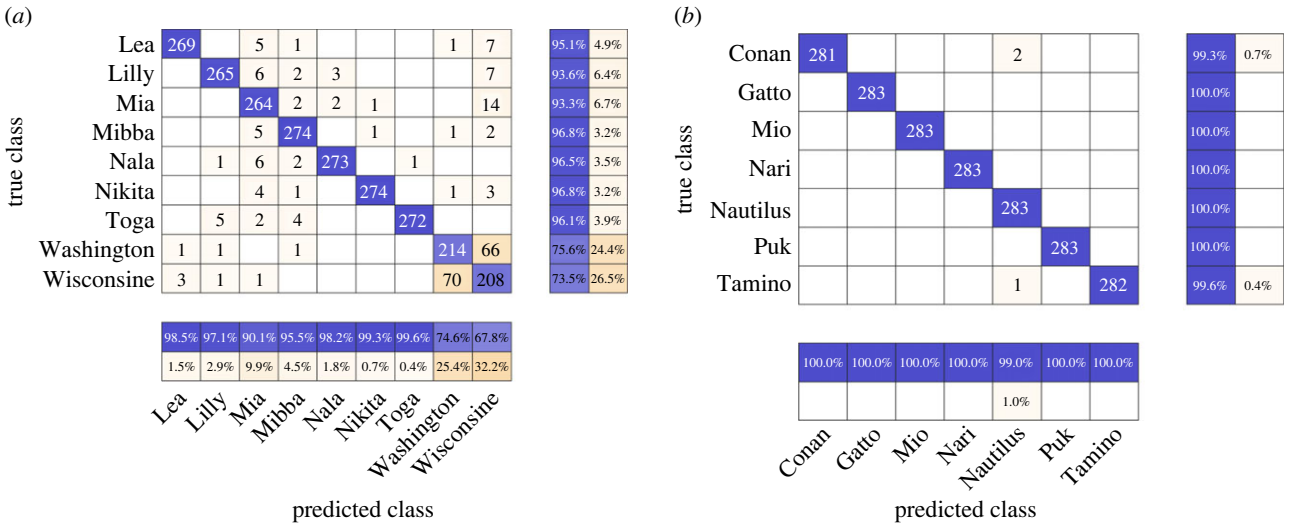

Figure 4. Individual precisions and recalls for determining the source identity from trills by the hierarchical classifier for females (*a*) and males (*b*). Confusion matrices are shown, with rows depicting the true source identity and columns representing the prediction made by the hierarchical classifier. The absolute number of calls is shown within the matrices, with those correctly classified highlighted in blue and those wrongly classified highlighted in orange (intensity proportional to the number for both). The rows and columns are summarized with the row summary depicting individual precisions in blue and the columns summary showing individual recalls in blue. The classifiers were tested on the Balanced-T datasets.

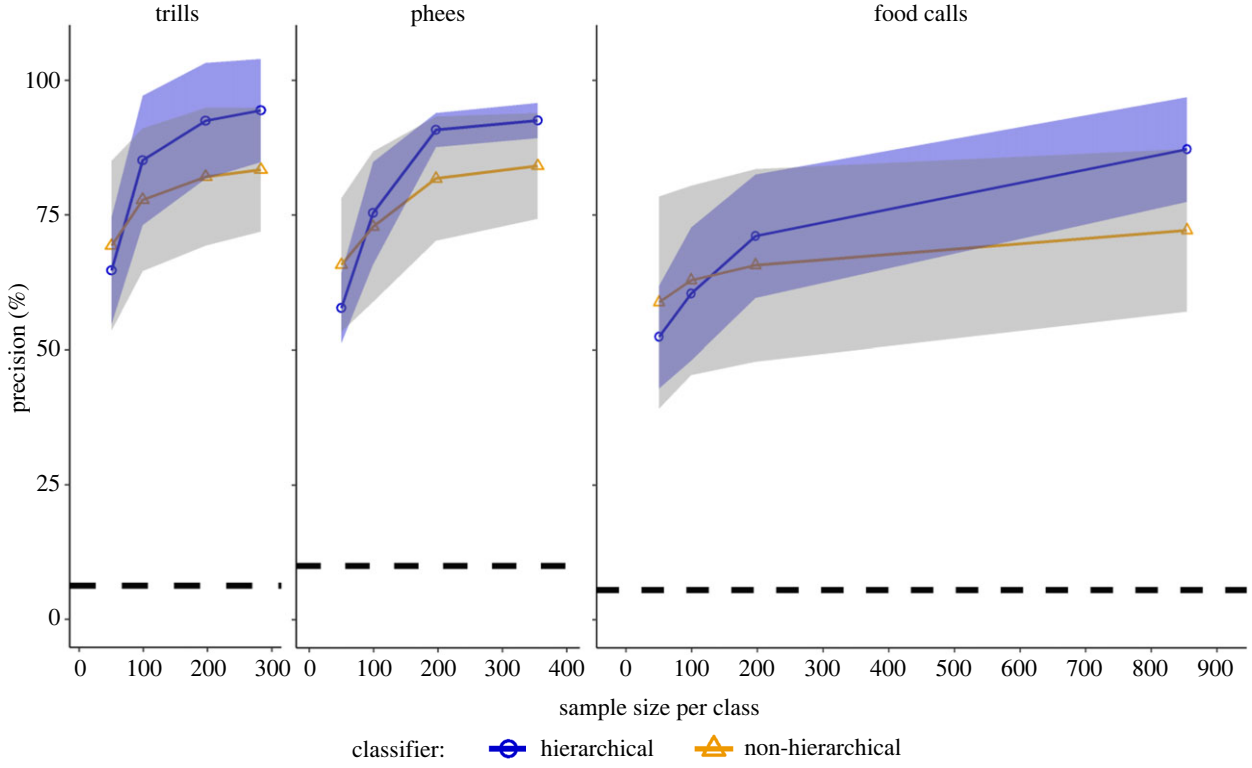

Figure 5. Classifier performance at different sample sizes. Precisions of AdaBoost as a function of sample size per class for trills, phees and food calls with s.d. represented as shaded regions around lines connecting means. Dashed black lines indicate the chance precision of classification for a given call type. Note that the highest sample size per class was obtained by oversampling the minority classes to be equal to the majority class (SMOTE, see Methods). See electronic supplementary material, table S1 for sample sizes in our original dataset.

3.2.5. Feature selection at various levels of the hierarchical classifier

The features selected by each level of the hierarchy varied with the task. Approximately 30%–60% of the features were used solely for determining sex across call types. Only a few features were common for determining individual identities among males and females (figure 6). The family of features from which most of the top-10 important features for each classification task came are presented in electronic supplementary material, table S5.

4. Discussion

We show that the information about sex and source identity encoded in marmoset calls can be harnessed for constructing hierarchical ML classifiers. Such hierarchical classifiers exhibit higher precisions and recalls than their non-hierarchical counterparts. Our findings have implications for marmoset communication research and will benefit understanding group-level communication dynamics. The same methods can also be extended for efficient source identification in other species.

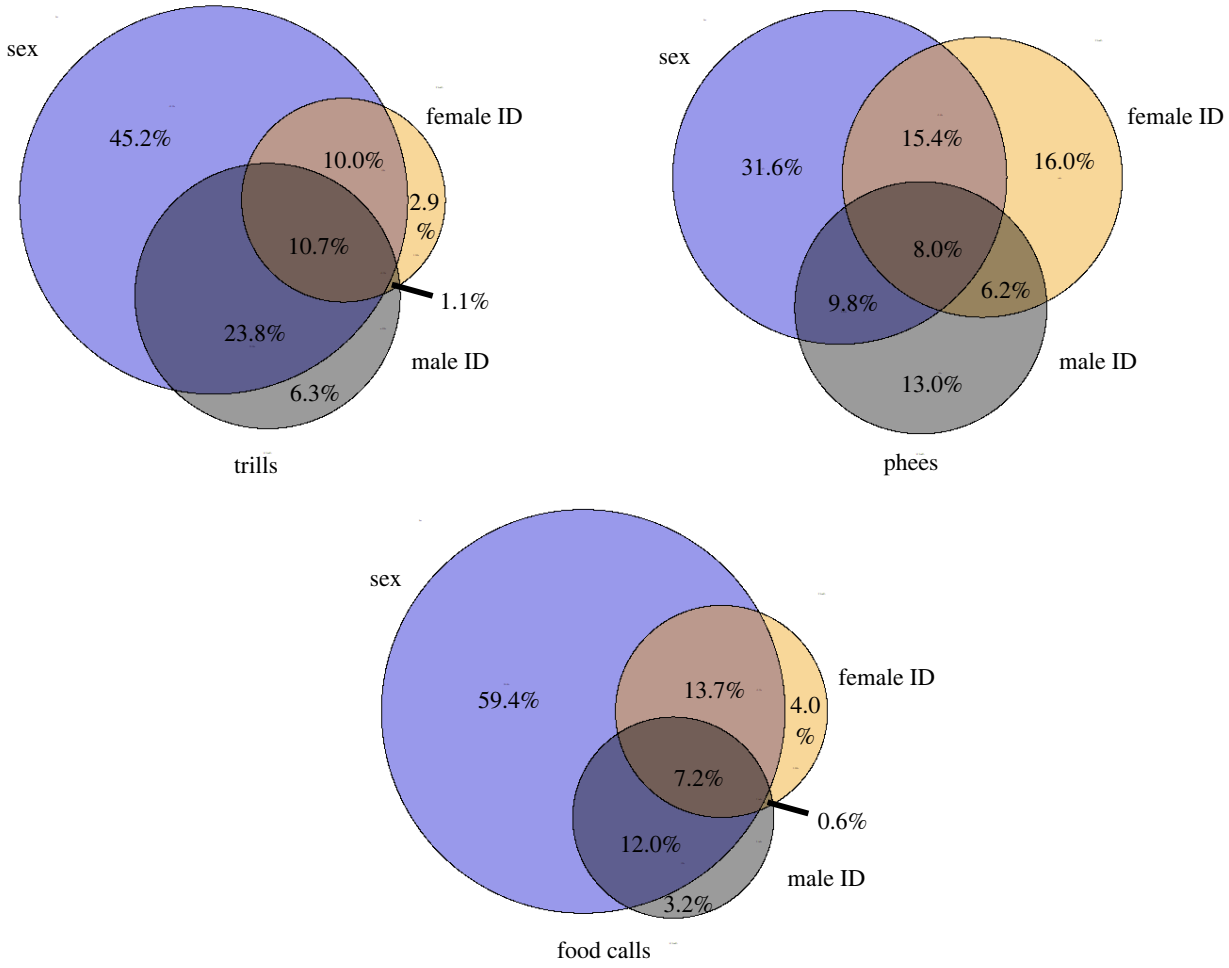

Figure 6. Mutual and distinct features used by hierarchical classifiers at various levels of classification. Venn diagrams denote the set of features used for determining the sex, source identity among females (female ID), and source identity among males (male ID). The area of the circle is scaled to the number of features at that level. The percentage of total features used by the hierarchical classifier belonging to each area within the Venn diagram is denoted.

4.1. Optimizing source identification

Of the large number of features extracted with HCTSA, we found that the AdaBoost reliably selected the most important features that would best cluster the data for its classification task (figure 3, electronic supplementary material, tables S1–S3). Intriguingly, the selected features varied with the task of the classifier (figure 6). Secker *et al*. [41] have previously used a hierarchical classifier with independent feature selection by the components to classify proteins successfully. They suggest that independent feature selection maintains high predictive performance while improving computational efficiency. In our case, the independent feature selection enabled the hierarchical classifier to efficiently use broader-category cues (i.e. sex) for classification, boosting performance over its non-hierarchical counterpart. AdaBoost thus provides task-specific and flexible feature extraction with a customized set of features for every dataset and can be applied to a wide range of animal vocalization datasets.

Even with the substantial number of features provided by HCTSA, AdaBoost performed poorly on the unbalanced dataset, with accuracies below 70% (table 2). A large body of ML literature points out the problem of class-unbalanced datasets and its solutions [42,43]. Recent reviews have emphasized using data augmentation and balancing techniques to improve ML accuracy when handling acoustic data [44,45]. Consistent with other studies [46,47], balancing the datasets significantly improved the performance of AdaBoost across call types. Therefore, data balancing using tools like SMOTE combined with random undersampling is important before running MLAs on any dataset.

Classifying data points from a noisy dataset to multiple classes (10–18 individuals in our case) is often demanding for an MLA. Here, we broke the problem of classifying calls to over 10 sources into the problem of first assigning the sex of the source and, only then, given the sex, classifying source identity in a second step (figure 2). With this hierarchical approach, we showed that at large sample sizes, mean precisions and recalls across datasets increased by more than eight percentage points (table 3). We found the accuracies of the hierarchical classifier on the largest balanced datasets to remain satisfactory and higher than most recent studies classifying animal vocalizations using MLAs [48–50], despite the number of samples per class of data being lower than those studies. The same was reflected in the performance of the hierarchical classifier on the originally collected calls (electronic supplementary material, figure S2).

The difference in precisions and recalls between the hierarchical and non-hierarchical approaches diminished as the sample sizes decreased, suggesting that the hierarchical classifier requires exposure to enough data to perform significantly better than its non-hierarchical counterpart (figure 5, electronic supplementary material, figure S3). Lower sample sizes pose a

higher risk of error cascading in the hierarchical classifier, and this has been identified in similar classifiers developed for text classification [51,52]. However, this limitation only arises at the level of the training dataset.

The requirement of a large training dataset and the presence of a small collected dataset can almost always be bridged. The median sample sizes per individual in our Imbalanced-X datasets were 47 for trills, 83 for phees and 173 for food calls (electronic supplementary material, table S1). Using this small, highly imbalanced dataset, we could generate larger balanced datasets and train and test our classifiers on them (see Methods). Therefore, to obtain optimal performance from the hierarchical classifier, one need not necessarily acquire a large number of calls to begin with (see performance of the hierarchical classifier on the Imbalanced-X dataset in electronic supplementary material, figure S2).

While we find that source identification from vocalizations can be optimized by using broader-category cues with the help of hierarchical classifiers for marmoset calls, the same methods can be extended to other tree-based classifiers and vocalizations from other animals due to two reasons. First, as tree-based classifiers inherently follow a hierarchical decision-making process, we predict that embedding them in a larger hierarchical framework and providing more information about the vocalizations will improve their performance. Second, the ability to select customized features for every dataset and task, combined with the supervised nature of learning, makes our pipeline highly flexible and extendable for analysing vocalizations of diverse animal species. However, because our pipeline uses a supervised MLA, a major requirement is that the calls of all individuals need to be represented in the training dataset. The source identity options available for the classifier to return as the output for any source determination task is simply the set of all the individuals it has encountered in the training dataset. Our method will not be useful, for example, for estimating the number of individuals in a large number of animal vocal recordings, as this information is required as input for the pipeline to work. In such cases, unsupervised MLAs such as Gaussian mixture models and hidden Markov models (see [53,54]) are suitable alternatives. However, the pipeline can classify calls of all individuals it has 'seen' during the training step. In the future, we hope to extend the idea to larger, more complex animal vocalization datasets with a greater number of individuals that would require multiple levels in the hierarchy and use other cues, such as the age and social status of the animal, for efficient source identification.

4.2. Implications for understanding the marmoset communication system

The classification precisions and recalls for food calls were lower than trills and phees, despite over three times higher sample sizes (table 3). In particular, food calls required a higher sample size per class to be classified as accurately as trills and phees (figure 5). This could be due to three possible reasons, or a combination of them: (i) food calls are a highly heterogeneous group of call types [55]. As the calls function in signalling other individuals about the availability of food and in inducing food sharing, it is highly likely that their acoustic structure is influenced by the food type, the internal state of the marmoset, its motivation to share food, and the social bond strength between the signaller and the surrounding individuals. Our method may therefore show limited performance in cases where the acoustic structure of calls change significantly with time (e.g. through seasons or during development) or context. This can be especially problematic if the variation in calls due to the aforementioned reasons surpass the levels due to inter-individual variability. Nevertheless, a prudent approach would involve uniformly sampling calls across various temporal and contextual settings for training the classifier. (ii) As the primary purpose of food calls is to alert other individuals about a food source, it may be under lower selection pressure to encode source identity information as compared with contact calls like trills and phees. (iii) Food calls are predominantly produced in bouts of multiple repeated call units [56]. Furthermore, each call unit is much shorter than a trill or a phee [56,57], providing reduced information for time series analysis. The poor classification results could thus arise because some of the source identity information may well be encoded at the level of the bout. This is testable in the future by repeating the classification procedure on bouts of food calls.

Intriguingly, the mean precision and recall for determining source identity from trills when considering all except two individuals were as high as 98.38% and 97.65%, respectively (figure 3). The two marmosets with low scores happened to be twins of the same sex. Thus, the low classification scores were probably due to the high vocal similarity between the twins' calls, which is also probably why the female identity classifier performed poorer than the male identity classifier (figure 4). It is known that in multi-class classifiers, decision boundaries for classes are not independent; therefore, poor performance on one class may negatively affect the performance of other classes [58]. Similar patterns were present to a lesser extent for phee calls but not for food calls (electronic supplementary material, figure S1). Whereas the food calls may require further scrutiny at different levels of analyses (see above), the contrast between trills and phees is interpretable with regard to their biological function. Signalling identity is essential for phee calls that individuals typically use to establish acoustic contact when visual contact is not possible [59]. In contrast, trill calls are given in close proximity to a social partner, and the caller's identity is thus redundant (marmosets may also use visual or olfactory cues to identify the partner) [60]. It is, therefore, possible that the twins actively diverged from each other in their phee calls but not in their trill calls. This is consistent with a recent study [33] on newly paired marmosets that found that partners would converge in the structure of their phee calls. However, newly formed pairs that had initially similar phee calls diverged rather than converged in their call structure, supposedly to make themselves better distinguishable.

Callitrichids can differentiate between contact calls originating from cage-mates versus foreign individuals and from males versus females [29,30]. An intriguing question is how they achieve that and whether their decision-making process may likewise be hierarchically structured with broader-category cues used as a first distinction. Multiple studies on humans allude to the hierarchical nature of decision-making in various contexts [61–65]. Some frog species seem to employ hierarchical decision-making for prey capture [66]; a hierarchical decision-making model appears to explain best the strategies used by rhesus macaques while playing a slightly modified, semi-controlled

adaptation of a video game [67]; sea lions use a hierarchy of multimodal cues during mother–offspring recognition [68,69]; and evidence from fruit flies, locusts and zebrafish suggest that they break down a complex problem of deciding between spatially distributed options into a series of smaller problems [70]. Hierarchical decision-making may thus be a widespread feature of cognitive processing. Given that sex could be attributed with extremely high precision by our classifier, we hypothesize that marmosets, too, use these sex-based cues for efficient source identification from calls, and they are doing so in a hierarchical manner, similar to the hierarchical classifier. Cognitive and psychological experiments will be required to test this hypothesis.

Even though recent studies have turned towards collecting large amounts of group-level acoustic data from animals [71–74], they lack information about the identity of the callers, which is important for understanding the ecology and behaviour of the species. With few manually labelled vocalizations, one can train our algorithm to determine source identities of a larger number of unlabelled vocalizations. This can be done by passing the unlabelled vocalizations through feature extraction and the same hierarchy of classifiers the labelled dataset was passed through. In this case, the sex labels provided by the first level of the hierarchical classifier are to be used to split the data into male and female vocalizations. These have to be passed through separate individual-level classifiers trained on the vocalizations of the respective sex. Our pipeline is thus an easy yet powerful tool to add source identity information to non-individualized datasets. It will support the analysis of vocal signals from groups of animals simultaneously, in contrast to the traditional method of focusing on one focal individual in a group or isolating individuals for recordings. Such a step is essential for understanding group-level communication dynamics of highly social species like marmosets and shedding light onto the communicative interactions that help in group-level coordination for raising young, which is thought to be a driver for the evolution of language.

Ethics. All experiments were approved by Zürich's cantonal veterinary office (licence ZH223/16, degree of severity: 0).

Data accessibility. All datasets and codes used in the study can be accessed from the Zenodo repository: https://doi.org/10.5281/zenodo.8367132 [75].

Supplementary figures and tables are provided in electronic supplementary material [76].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' Contributions. N.P.: conceptualization, formal analysis, investigation, methodology, writing—original draft; K.W.: methodology, supervision, writing—review and editing; Y.Z.: data curation, investigation, methodology; J.B.: conceptualization, funding acquisition, project administration, resources, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the Swiss National Science Foundation (grant number 31003A_149796, the NCCR Evolving Language, agreement number 51NF40_180888) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 101001295). N.P. was a recipient of a grant from the A. H. Schultz Foundation.

References

1. Fedurek P, Slocombe KE. 2011 Primate vocal communication: a useful tool for understanding human speech and language evolution? *Hum. Biol.* **83**, 153–173. (doi:10.1353/hub.2011.a438018)
2. Zuberbühler K. 2017 The primate roots of human language. In *Primate hearing and communication* (eds RM Quam, MA Ramsier, RR Fry, AN Popper), pp. 175–200. Berlin, Germany: Springer.
3. Fitch WT. 2017 Empirical approaches to the study of language evolution. *Psychon. Bull. Rev.* **24**, 3–33. (doi:10.3758/s13423-017-1236-5)
4. Seyfarth R, Cheney D. 2018 Pragmatic flexibility in primate vocal production. *Curr. Opin. Behav. Sci.* **21**, 56–61. (doi:10.1016/j.cobeha.2018.02.005)
5. Zhao L, Roy S, Wang X. 2019 Rapid modulations of the vocal structure in marmoset monkeys. *Hear. Res.* **384**, 107811. (doi:10.1016/j.heares.2019.107811)
6. Pomberger T, Risueno-Segovia C, Löschner J, Hage SR. 2018 Precise motor control enables rapid flexibility in vocal behavior of marmoset monkeys. *Curr. Biol.* **28**, 788–794. (doi:10.1016/j.cub.2018.01.070)
7. Zürcher Y, Willems EP, Burkart JM. 2019 Are dialects socially learned in marmoset monkeys? Evidence from translocation experiments. *PLoS ONE* **14**, e0222486. (doi:10.1371/journal.pone.0222486)
8. Miller CT, Wang X. 2006 Sensory-motor interactions modulate a primate vocal behavior: antiphonal calling in common marmosets. *J. Comp. Physiol. A* **192**, 27–38. (doi:10.1007/s00359-005-0043-z)
9. Pistorio AL, Vintch B, Wang X. 2006 Acoustic analysis of vocal development in a New World primate, the common marmoset (*Callithrix jacchus*). *J. Acoust. Soc. Am.* **120**, 1655–1670. (doi:10.1121/1.2225899)
10. Takahashi DY, Liao DA, Ghazanfar AA. 2017 Vocal learning via social reinforcement by infant marmoset monkeys. *Curr. Biol.* **27**, 1844–1852. (doi:10.1016/j.cub.2017.05.004)
11. Takahashi DY, Fenley AR, Teramoto Y, Narayanan DZ, Borjon JI, Holmes P, Ghazanfar AA. 2015 The developmental dynamics of marmoset monkey vocal production. *Science* **349**, 734–738. (doi:10.1126/science.aab1058)
12. Burkart JM, Adriaense JEC, Brügger RK, Miss FM, Wierucka K, van Schaik CP. 2022 A convergent interaction engine: vocal communication among marmoset monkeys. *Phil. Trans. R. Soc. B* **377**, 20210098. (doi:10.1098/rstb.2021.0098)
13. Jovanovic V, Miller CT. 2021 Mechanisms for communicating in a marmoset 'cocktail party'. *bioRxiv* **9**, 416693. (doi:10.1101/2020.12.08.416693)
14. Clink DJ, Klinck H. 2021 Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. *Methods Ecol. Evol.* **12**, 328–341. (doi:10.1111/2041-210X.13520)
15. Martinez JG, Bohn KM, Carroll RJ, Morris JS. 2013 A study of Mexican free-tailed bat chirp syllables: Bayesian functional mixed models for nonstationary acoustic time series. *J. Am. Stat. Assoc.* **108**, 514–526. (doi:10.1080/01621459.2013.793118)
16. Wijers M, Trethowan P, Du Preez B, Chamaillé-Jammes S, Loveridge AJ, Macdonald DW, Markham A. 2021 Vocal discrimination of African lions and its potential for collar-free tracking. *Bioacoustics* **30**, 575–593. (doi:10.1080/09524622.2020.1829050)
17. Sauvé CC, Beauplet G, Hammill MO, Charrier I. 2015 Acoustic analysis of airborne, underwater, and amphibious mother attraction calls by wild harbor seal pups (*Phoca vitulina*). *J. Mammal.* **96**, 591–602. (doi:10.1093/jmammal/gyv064)
18. Clink DJ, Tasirin JS, Klinck H. 2020 Vocal individuality and rhythm in male and female duet contributions of a nonhuman primate. *Curr. Zool.* **66**, 173–186. (doi:10.1093/cz/zoz035)
19. Opzeeland ICV, Parijs SMV, Frickenhaus S, Kreiss CM, Boebel O. 2012 Individual variation in pup vocalizations and absence of behavioral signs of

maternal vocal recognition in Weddell seals (*Leptonychotes weddellii*). *Mar. Mammal. Sci.* **28**, E158–E172. (doi:10.1111/j.1748-7692.2011.00505.x)

20. Coye C, Zuberbühler K, Lemasson A. 2022 The evolution of vocal communication: inertia and divergence in two closely related primates. *Int. J. Primatol.* **43**, 712–732. (doi:10.1007/s10764-022-00294-y)
21. Yen S-C, Shieh B-S, Wang Y-T, Wang Y. 2013 Rutting vocalizations of Formosan sika deer *Cervus nippon taiouanus*—acoustic structure, seasonal and diurnal variations, and individuality. *Zool. Sci.* **30**, 1025–1031. (doi:10.2108/zsj.30.1025)
22. Oyakawa C, Koda H, Sugiura H. 2007 Acoustic features contributing to the individuality of wild agile gibbon (*Hylobates agilis agilis*) songs. *Am. J. Primatol.* **69**, 777–790. (doi:10.1002/ajp.20390)
23. Sanchez FJ B, Hossain MR, English NB, Moore ST. 2021 Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Sci. Rep.* **11**, 15733. (doi:10.1038/s41598-021-95076-6)
24. Huang C-J, Chen Y-J, Chen H-M, Jian J-J, Tseng S-C, Yang Y-J, Hsu P-A. 2014 Intelligent feature extraction and classification of anuran vocalizations. *Appl. Soft Comput.* **19**, 1–7. (doi:10.1016/j.asoc.2014.01.030)
25. Nanni L, Maguolo G, Paci M. 2020 Data augmentation approaches for improving animal audio classification. *Ecol. Inform.* **57**, 101084. (doi:10.1016/j.ecoinf.2020.101084)
26. Fulcher BD, Jones NS. 2014 Highly comparative feature-based time-series classification. *IEEE Trans. Knowl. Data Eng.* **26**, 3026–3037. (doi:10.1109/TKDE.2014.2316504)
27. Fulcher BD, Jones NS. 2017 hctsa: a computational framework for automated time-series phenotyping using massive feature extraction. *Cell Syst.* **5**, 527–531. (doi:10.1016/j.cels.2017.10.001)
28. Norcross JL, Newman JD. 1993 Context and gender-specific differences in the acoustic structure of common marmoset (*Callithrix jacchus*) phee calls. *Am. J. Primatol.* **30**, 37–54. (doi:10.1002/ajp.1350300104)
29. Miller CT, Scarl J, Hauser MD. 2004 Sensory biases underlie sex differences in tamarin long call structure. *Anim. Behav.* **68**, 713–720. (doi:10.1016/j.anbehav.2003.10.028)
30. Miller J, Hauser M, Miller C, Gil-Da-Costa R. 2001 Selective phonotaxis by cotton-top tamarins (*Saguinus oedipus*). *Behaviour* **138**, 811–826. (doi:10.1163/156853901753172665)
31. Agamaite JA, Chang C-J, Osmanski MS, Wang X. 2015 A quantitative acoustic analysis of the vocal repertoire of the common marmoset (*Callithrix jacchus*). *J. Acoust. Soc. Am.* **138**, 2906–2928. (doi:10.1121/1.4934268)
32. Brown GR, Almond RE, Bergen YV. 2004 Begging, stealing, and offering: food transfer in nonhuman primates. *Adv. Stud. Behav.* **34**, e295.
33. Zürcher Y, Willems EP, Burkart JM. 2021 Trade-offs between vocal accommodation and individual recognisability in common marmoset vocalizations. *Sci. Rep.* **11**, 1–10. (doi:10.1038/s41598-021-95101-8)
34. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002 SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357. (doi:10.1613/jair.953)
35. Larsen BS. 2020 Synthetic minority over-sampling technique (SMOTE). See https://github.com/dkbsl/matlab_smote (accessed on 7 May 2020).
36. Van der Maaten L, Hinton G. 2008 Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605.
37. Schapire RE. 1990 The strength of weak learnability. *Mach. Learn.* **5**, 197–227. (doi:10.1007/BF00116037)
38. Freund Y, Schapire RE. 1997 A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55**, 119–139. (doi:10.1006/jcss.1997.1504)
39. Kégl B. 2013 The return of AdaBoost.MH: multi-class Hamming trees. *ArXiv*. (https://arxiv.org/abs/1312.6086)
40. Kobak D, Berens P. 2019 The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416. (doi:10.1038/s41467-019-13056-x)
41. Secker A, Davies MN, Freitas AA, Clark E, Timmis J, Flower DR. 2010 Hierarchical classification of G-protein-coupled receptors with data-driven selection of attributes and classifiers. *Int. J. Data Min. Bioinform.* **4**, 191–210. (doi:10.1504/IJDMB.2010.032150)
42. Fernandez A, Garcia S, Herrera F, Chawla NV. 2018 SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res* **61**, 863–905. (doi:10.1613/jair.1.11192)
43. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. 2018 *Learning from imbalanced data sets*. Berlin, Germany: Springer.
44. Sun Y, Midori Maeda T, Solís-Lemus C, Pimentel-Alarcón D, Buřivalová Z. 2022 Classification of animal sounds in a hyperdiverse rainforest using convolutional neural networks with data augmentation. *Ecol. Indic.* **145**, 109621. (doi:10.1016/j.ecolind.2022.109621)
45. Mao A *et al.* 2022 Automated identification of chicken distress vocalizations using deep learning models. *J. R. Soc. Interface* **19**, 20210921. (doi:10.1098/rsif.2021.0921)
46. Batista GE, Prati RC, Monard MC. 2004 A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **6**, 20–29. (doi:10.1145/1007730.1007735)
47. Laurikkala J. 2001 Improving identification of difficult small classes by balancing class distribution. In *Conf. on Artificial Intelligence in Medicine in Europe*, pp. 63–66. Berlin, Germany: Springer.
48. Barker AJ, Veviurko G, Bennett NC, Hart DW, Mograby L, Lewin GR. 2021 Cultural transmission of vocal dialect in the naked mole-rat. *Science* **371**, 503–507. (doi:10.1126/science.abc6588)
49. Ivanenko A, Watkins P, van Gerven MAJ, Hammerschmidt K, Englitz B. 2020 Classifying sex and strain from mouse ultrasonic vocalizations using deep learning. *PLOS Comput. Biol.* **16**, e1007918. (doi:10.1371/journal.pcbi.1007918)
50. Lehmann KDS, Jensen FH, Gersick AS, Strandburg-Peshkin A, Holekamp KE. 2022 Long-distance vocalizations of spotted hyenas contain individual, but not group, signatures. *Proc. R. Soc. B* **289**, 20220548. (doi:10.1098/rspb.2022.0548)
51. Babbar R, Partalas I, Gaussier E, Amini MR. 2013 On flat versus hierarchical classification in large-scale taxonomies. *Adv. Neural Inf. Process. Syst.* **26**, 1824–1832.
52. Dumais S, Chen H. 2000 Hierarchical classification of web content. In *Proc. of the 23rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Athens, Greece, 24–28 July*, pp. 256–263. ACM.
53. Cheng J, Sun Y, Ji L. 2010 A call-independent and automatic acoustic system for the individual recognition of animals: a novel model using four passerines. *Pattern Recognit.* **43**, 3846–3852. (doi:10.1016/j.patcog.2010.04.026)
54. Peso Parada P, Cardenal-López A. 2014 Using Gaussian mixture models to detect and classify dolphin whistles and pulses. *J. Acoust. Soc. Am.* **135**, 3371–3380. (doi:10.1121/1.4876439)
55. Burkart J, Martins EG, Miss F, Zürcher Y. 2018 From sharing food to sharing information: cooperative breeding and language evolution. *Interact. Stud.* **19**, 136–150. (doi:10.1075/is.17026.bur)
56. Vitale A, Zanzoni M, Queyras A, Chiarotti F. 2003 Degree of social contact affects the emission of food calls in the common marmoset (*Callithrix jacchus*). *Am. J. Primatol.* **59**, 21–28. (doi:10.1002/ajp.10060)
57. Rogers LJ, Stewart L, Kaplan G. 2018 Food calls in common marmosets, *Callithrix jacchus*, and evidence that one is functionally referential. *Animals* **8**, 99. (doi:10.3390/ani8070099)
58. Gupta MR, Bengio S, Weston J. 2014 Training highly multiclass classifiers. *J. Mach. Learn. Res.* **15**, 1461–1492.
59. Jones BS, Harris DHR, Catchpole CK. 1993 The stability of the vocal signature in phee calls of the common marmoset, *Callithrix jacchus*. *Am. J. Primatol.* **31**, 67–75. (doi:10.1002/ajp.1350310107)
60. Smith T. 2006 Individual olfactory signatures in common marmosets (*Callithrix jacchus*). *Am. J. Primatol.* **68**, 585–604. (doi:10.1002/ajp.20254)
61. Lashley KS. 1951 The problem of serial order in behavior. In *Cerebral mechanisms in behavior; the Hixon Symposium*, pp. 112–146. Oxford, UK: Wiley.
62. Miller GA, Eugene G, Pribram KH. 2017 Plans and the Structure of Behaviour. In *Systems research for behavioral sciencesystems research* (ed. W Buckley), pp. 369–382. Routledge.
63. Newell A, Simon HA. 1961 GPS, a program that simulates human thought. In *Lernende Automaten* (ed. H Billing), pp. 109–112. Munich, Germany: R.Oldenbourg.
64. Schneider DW, Logan GD. 2006 Hierarchical control of cognitive processes: switching tasks in sequences. *J. Exp. Psychol. Gen.* **135**, 623. (doi:10.1037/0096-3445.135.4.623)
65. Saaty TL. 2008 Decision making with the analytic hierarchy process. *Int. J. Serv. Sci.* **1**, 83–98.

66. Monroy JA, Nishikawa K. 2011 Prey capture in frogs: alternative strategies, biomechanical trade-offs, and hierarchical decision making. *J. Exp. Zool. Part Ecol. Genet. Physiol* **315**, 61–71. (doi:10.1002/jez.601)
67. Yang Q, Lin Z, Zhang W, Li J, Chen X, Zhang J, Yang T. 2022 Monkey plays Pac-Man with compositional strategies and hierarchical decision-making. *Elife* **11**, e74500. (doi:10.7554/eLife.74500)
68. Wierucka K, Pitcher BJ, Harcourt R, Charrier I. 2018 Multimodal mother–offspring recognition: the relative importance of sensory cues in a colonial mammal. *Anim. Behav.* **146**, 135–142. (doi:10.1016/j.anbehav.2018.10.019)
69. Wierucka K, Pitcher BJ, Harcourt R, Charrier I. 2017 The role of visual cues in mother–pup reunions in a colonially breeding mammal. *Biol. Lett.* **13**, 20170444. (doi:10.1098/rsbl.2017.0444)
70. Sridhar VH, Li L, Gorbonos D, Nagy M, Schell BR, Sorochkin T, Gov NS, Couzin ID. 2021 The geometry of decision-making in individuals and collectives. *Proc. Natl Acad. Sci. USA* **118**, e2102157118. (doi:10.1073/pnas.2102157118)
71. Heinicke S, Kalan AK, Wagner OJJ, Mundry R, Lukashevich H, Kühl HS. 2015 Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods Ecol. Evol.* **6**, 753–763. (doi:10.1111/2041-210X.12384)
72. Spillmann B, van Noordwijk MA, Willems EP, Mitra Setia T, Wipfli U, van Schaik CP. 2015 Validation of an acoustic location system to monitor Bornean orangutan (*Pongo pygmaeus wurmbii*) long calls. *Am. J. Primatol.* **77**, 767–776. (doi:10.1002/ajp.22398)
73. Kalan AK, Mundry R, Wagner OJJ, Heinicke S, Boesch C, Kühl HS. 2015 Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecol. Indic.* **54**, 217–226. (doi:10.1016/j.ecolind.2015.02.023)
74. Kalan AK, Piel AK, Mundry R, Wittig RM, Boesch C, Kühl HS. 2016 Passive acoustic monitoring reveals group ranging and territory use: a case study of wild chimpanzees (*Pan troglodytes*). *Front. Zool.* **13**, 34. (doi:10.1186/s12983-016-0167-8)
75. Phaniraj N, Wierucka K, Zürcher Y, Maria Burkart J. 2023 Code for: Who is calling? Optimizing source identification from marmoset vocalisations with hierarchical machine learning classifiers. *Zenodo*. (doi:10.5281/zenodo.8367132)
76. Phaniraj N, Wierucka K, Zürcher Y, Maria Burkart J. 2023 Who is calling? Optimizing source identification from marmoset vocalisations with hierarchical machine learning classifiers. Figshare.