Vocal complexity in the long calls of Bornean orangutans

*Erb, W.M.[1,2], Ross, W.[1], Kazanecki, H.[1], Mitra Setia, T.[3,4], Madhusudhana, S.[1,5], Clink, D.J.[1]

*Corresponding author

Email: erbivorous@gmail.com

[1] K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, New York, 14850, USA

[2] Department of Anthropology, Rutgers University, New Brunswick, New Jersey, 08901, USA

[3] Fakultas Biologi, Universitas Nasional Jakarta, Jakarta, Indonesia

[4] Primate Research Center, Universitas Nasional, Jakarta, Indonesia

[5] Centre for Marine Science and Technology, Curtin University, Perth, WA 6102, Australia

Short title: Orangutan Long Call Classification

ABSTRACT

Vocal complexity is central to many evolutionary hypotheses about animal communication. Yet, quantifying and comparing complexity remains a challenge, particularly when vocal types are highly graded. Male Bornean orangutans (*Pongo pygmaeus wurmbii*) produce complex and variable "long call" vocalizations comprising multiple sound types that vary within and among individuals. Previous studies described six distinct call (or pulse) types within these complex vocalizations, but none quantified their discreteness or the ability of human observers to reliably classify them. We studied the long calls of 13 individuals to: 1) evaluate and quantify the reliability of audio-visual classification by three well-trained observers, 2) distinguish among call types using supervised classification and unsupervised clustering, and 3) compare the performance of different feature sets. Using 46 acoustic features, we used machine learning (i.e., support vector machines, affinity propagation, and fuzzy c-means) to identify call types and assess their discreteness. We also used Uniform Manifold Approximation and Projection (UMAP) to visualize the separation of pulses using both extracted features and spectrograms. We found low inter-observer reliability and poor classification accuracy using supervised approaches, indicating that pulse types were not discrete. We propose a new pulse type classification scheme that is highly reproducible across observers and exhibits high classification accuracy using support vector machines. Although the low number of call types suggests long calls are fairly simple, the continuous gradation of sounds seems to greatly boost the complexity of this system. This work responds to calls for more quantitative research to define call types and measure the gradedness of animal vocal systems and highlights the need for a more comprehensive framework for studying vocal complexity vis-à-vis graded repertoires.

HIGHLIGHTS

- We used audio-visual (AV) analysis and machine-learning to discriminate pulse types.
- AV and support vector machines (SVM) did not support the six published pulse types.
- Hard and soft clustering algorithms showed a mixture of discrete and graded pulses.
- We propose three pulse types that show high reproducibility and classification accuracy.
- More work is needed to investigate the role of graded signals in vocal complexity.

KEYWORDS

acoustic communication; affinity propagation; fuzzy clustering; graded signals; machine learning; supervised classification; support vector machines; Uniform Manifold Approximation and Projection (UMAP); unsupervised clustering; vocal repertoire

INTRODUCTION

Vocal complexity, or the diversity of sounds in a species' repertoire, is central to many evolutionary hypotheses about animal communication (Bradbury & Vehrencamp, 2011; Fischer et al., 2016; Freeberg et al., 2012; McComb & Semple, 2005). This complexity has been hypothesized to be shaped by a range of factors including predation pressure, sexual selection, habitat structure, and social complexity (Bradbury & Vehrencamp, 2011; Fischer et al., 2016). Two common measures of vocal complexity are: 1) the diversity (or number) of call types as well as 2) their discreteness. For instance, within black-capped chickadee (*Poecile atricapillus*) groups, individuals flexibly increase the diversity of note types when they are in larger groups, presumably increasing the number of potential messages that can be conveyed (Freeberg et al., 2012). When comparing across species, similar themes emerge in rodents and primates. Sciurid species with a greater diversity of social roles have more alarm call types (Blumstein & Armitage, 1997) and primate species in larger groups with more intense social bonding have larger vocal repertoires (McComb & Semple, 2005). Further, it has been proposed that while discrete repertoires facilitate signal recognition in dense habitats, graded calls allow more complexity in open habitats where intermediate sounds communicate arousal and can be linked with visual signals (Marler et al., 1975).

Yet, quantifying vocal complexity in a standardized manner remains a challenge for comparative analyses. A primary aspect of this challenge is related to the identification and quantification of discrete call types, which is particularly vexing in repertoires comprising intermediate calls and in species that exhibit significant inter-individual variation (Fischer et al., 2016). The most common approaches to identifying call types are: 1) manual (visual or audio-visual) classification of spectrograms by a human observer and 2) automated (quantitative or algorithmic) using features that are either manually or automatically measured from spectrograms (Kershenbaum et al., 2016). Audio-visual classification involves one or more observers inspecting spectrograms visually while simultaneously listening to the sounds. This method has been applied to the vocalizations of numerous taxa (e.g., manatees, *Trichechus manatus latirostris*: Brady et al., 2020; spear-nosed bats, *Phyllostomus discolor*: Lattenkamp, 2019; humpback whales, *Megaptera novaeangliae*: Madhusudhana et al., 2019; New Zealand kea parrots, *Nestor notabilis*: Schwing et al., 2012). Audio-visual classification studies often rely on a single expert observer and only rarely quantify within- or between-observer reliability (reviewed in Jones et al., 2001). On one hand, when classification is done by a single observer, the study risks idiosyncratic or irreproducible results. On the other hand, when multiple observers are involved, the study risks inconsistent assessments among scorers. To assess the reproducibility of a human-based classification scheme, it is critical to evaluate the consistency of scores within and/or among the human observers using inter-rater reliability (IRR) statistics such as Cohen's kappa (Hallgren, 2012).

To compare and classify acoustic signals, researchers must make decisions about which features to estimate, as analyses of the waveform are generally too computationally costly. A commonly used approach for many classification problems is feature selection, in which a suite of selected time- and frequency-based characteristics of sounds are measured and compiled from manually annotated spectrograms (Odom et al., 2021). There is little standardization concerning the selection of acoustic variables across studies, which often include a combination of qualitative and quantitative measurements that are manually and/or automatically (i.e., using a sound analysis program, such as Raven Pro 1.6) extracted. As an alternative to feature selection, some researchers use automated approaches wherein the spectral content of sounds is measured using spectrograms, cepstra, multi-taper spectra, wavelets, or formants (reviewed in Kershenbaum et al., 2016).

Once features have been manually or automatically extracted, multivariate analyses can be used to classify or cluster sounds using supervised or unsupervised algorithms, respectively. In the case of supervised classification, users manually label a subset of representative sounds which are used to train the statistical model that will subsequently be used to automatically identify those sound types in an unlabeled set of data (Cunningham et al., 2008). In contrast to supervised classification, clustering is an unsupervised machine learning approach in which an algorithm divides a dataset into several groups or clusters such that observations in the same group are similar to each other and dissimilar to the observations in different groups (Greene et al., 2008). Thus, in the case of unsupervised clustering, the computer – rather than the human observer – learns the groupings and assigns labels to each value (Alloghani et al., 2020).

Enumeration of call types in a repertoire is especially challenging when there are intermediate forms that fall between categories. These so-called graded call types have been well documented across primate taxa (Fischer et al., 2016; Hammerschmidt & Fischer, 1998). An alternative to "hard clustering" of calls into discrete categories (e.g., k-means, k-medoids, affinity propagation), "soft clustering" (e.g., fuzzy c-means) allows for imperfect membership by assigning probability scores for membership in each cluster, thereby making it possible to identify call types with intermediate values (Cusano et al., 2021; Fischer et al., 2016). So-called fuzzy clustering can be used in tandem with hard clustering by also quantifying the degree of ambiguity (or gradedness) exhibited by particular sounds and continuities across call types. Thus, soft clustering provides a means of quantifying gradedness in repertoires and can enable the identification of intermediate members.

Across studies of animal vocal complexity, there is notable variation in the number and type of feature sets used, ranging from fewer than 10 to more than 100 parameters that are manually and/or automatically extracted. Table 1 provides a summary of 15 studies across mammalian and avian taxa that used supervised classification and unsupervised clustering approaches to identify call types across a range of mammalian and avian taxa. Though most studies paired audio-visual classification with an unsupervised clustering method, a few also included discriminant function analysis (DFA) to quantify the differences among the human-labeled call types and/or computer-identified clusters. Authors relied on a broad range of unsupervised clustering algorithms, though hierarchical agglomerative clustering was the most used method. Studies that aimed to provide an accurate classification of different call types often relied on a combination of supervised classification and unsupervised clustering methods to ensure results were robust and repeatable. However, those that compared feature sets or clustering methods often reported a lack of agreement on the number of clusters identified, highlighting the difficulty of the seemingly straightforward task of identifying and quantifying call types.

Table 1. Review of studies using supervised classification and unsupervised clustering approaches to identify vocal types.

Authors (Date)	**Taxon**	**Goals**	**N Features**	**Classification (N observers)**	**Clustering Method**
Wadewitz et al. (2015)	Chacma baboon *(Papio ursinus)*	Compare hard & soft clustering, evaluate influence of features	9, 38, 118 (+ 19 PCA factors)	A/V *	K-means, Hierarchical agglomerative (Ward's), Fuzzy c-means
Fuller (2014)	Blue monkey *(Cercopithecus mitis stulmanni)*	Catalog vocal signals	18 PCA factors	A/V (1), DFA **	Hierarchical agglomerative
Fournet et al., (2015)	Humpback whale *(Megaptera novaeangliae)*	Catalog non-song vocalizations	15	A/V (1), DFA	Hierarchical agglomerative
Brady et al. (2020)	Florida manatee *(Trichechus manatus latirostris)*	Catalog vocal repertoire	17	A/V (1)	Maximum likelihood, CART
Hammerschmidt & Fischer (2019**)**	Chacma *(Papio ursinus)*, olive *(P. anubis)*, and Guinea baboon *(P. papio)*	Catalog & compare vocal repertoires, Compare A/V to clustering	9	A/V (multiple), DFA **	Two-step cluster analysis
Sadhukhan et al. (2019)	Indian wolf *(Canis lupus pallipes)*	Catalog harmonic vocalizations	8	DFA	Hierarchical agglomerative
Hedwig et al. (2019)	African forest elephant *(Loxodonta cyclotis)*	Catalog vocal repertoire	23	DFA **	PCA
Huijser et al. (2020)	Sperm whale *(Physeter macrocephalus)*	Catalog coda repertoires	2	A/V (1)	K-means, Hierarchical agglomerative
Vester et al. (2017)	Long-finned pilot whale *(Globicephala melas)*	Catalog vocal repertoire	14	A/V (2), DFA	Two-step cluster analysis

Soltis et al. (2012)	Key Largo woodrat *(Neotoma floridana smalli)*	Catalog vocal repertoire	6	A/V*	Multidimensional scaling analysis (MDS)
Elie & Theunissen (2016)	Zebra finch *(Taeniopygia guttata)*	Catalog vocal repertoire, determine distinguishing features	22, 25 (MFCCs)	A/V (1), Fisher LDA, Random Forest	PCA, Gaussian mixture
Janik (1999)	Bottlenose dolphin *(Tursiops truncatus)*	Compare A/V to clustering	20	A/V (5)	K-means, Hierarchical agglomerative
Cusano et al. (2021)	Humpback whale *(Megaptera novaeangliae)*	Differentiate discrete vs. graded call types	25	A/V*	Fuzzy k-means
Garland et al. (2015)	Beluga whale *(Delphinapterus leucas)*	Catalog vocal repertoire	12	A/V*	CART, Random Forest
Thiebault et al. (2019)	Cape gannet *(Morus capensis)*	Catalog repertoire of foraging calls	12	A/V*	Random Forest

* Study did not report # of observers

** leave-one-out

In the present study, we examine vocal complexity in the long calls of Bornean orangutans (*Pongo pygmaeus wurmbii*) by evaluating how the choice of classification or clustering methods and feature inputs affect the number of call types we recognize. Orangutans are semi-solitary great apes who exhibit a promiscuous mating system in which solitary adult males range widely in search of fertile females (Spillmann et al., 2017). Flanged males (i.e., adult males who have fully developed cheek pads, throat sacs, and body size approximately twice that of adult females) emit loud vocalizations, or long calls, which travel up to a kilometer and serve to attract female mates and repel rival males (Mitra Setia & van Schaik, 2007) In this social setting, long calls thus hold an important function for coordination among widely dispersed individuals.

Long calls are complex and variable vocalizations comprising multiple call (or pulse) types that vary within and among individuals (Askew & Morrogh-Bernard, 2016; Spillmann et al., 2010). Long calls typically begin with a bubbly introduction of soft, short sounds that build into a climax of high-amplitude frequency-modulated pulses followed by a series of lower-amplitude and -frequency pulses that gradually transition to soft and short sounds, similar to the introduction (cf. MacKinnon, 1977, Table S1). Although Davilla Ross and Geissmann (2007) first attempted to classify and name the different elements of these calls, they noted a "wide variety of call elements do not belong to any of these note types" (Davila Ross & Geissmann, 2007 p. 309).

Spillmann and colleagues (2010) presented the most detailed description of orangutan long calls in which they identified six different pulse types (Table 2), but thus far there has been no attempt to systematically classify pulses or quantify how discrete these sounds are. Further, no studies have described the process for or the number of observers classifying sound types nor the reliability of classifications within or among observers. Thus, it is presently unclear how well pulse types can be discriminated by human observers or quantitative classification tools, thereby limiting our ability to repeat, reproduce, and replicate these studies.

Table 2. Names and descriptions of sound labels used in previous studies, using Spillmann et al. (2010) labels as reference.

Sound Type	MacKinnon 1974	Davila Ross & Geissmann 2007	Spillmann et al. 2010
Grumbles	bubbly introduction	bubbling	"preceding bubbling-like elements that are low in loudness"
Bubbles	n/a	bubbling	"low amplitude, looks like a cracked sigh"
Roar	"climax of full roars"	roar	"more rounded and lower in frequency"
Low Roar	n/a	n/a	"half the fundamental frequency at the highest point than roar"
Volcano Roar	n/a	n/a	"sharp tip and higher frequency than roar"
Huitus	n/a	huitus	"high amplitude with steeply ascending and descending part that are not connected"
Intermediary	n/a	intermediary	"low amplitude, frequency modulation starts with a rising part followed by a falling part that changes again into a rising and ends with a falling part"
Sigh	"tails off gradually into a series of sighs"	sigh	"low amplitude, starts with a short rising part and changes in a long falling part"

The present study aims to evaluate vocal complexity in orangutan long calls to compare different approaches to identifying the number of discrete calls and estimating the degree of gradedness in a model vocal system. Specifically, the objectives of our study are to: 1) evaluate and quantify the reliability of manual audio-visual (AV) classification by three well-trained observers, 2) classify and cluster call types using supervised classification (support vector machines) and unsupervised hard (affinity propagation) and soft (fuzzy c-means) clustering methods, and 3) compare the results using different feature sets (i.e., feature engineering, complete spectrographic representations). Based on these findings, we will make explore and assess alternative classification systems for identifying discrete and graded call types in this system.

METHODS

Ethical Note

This research was approved by the Institutional Animal Care and Use Committee of Rutgers, the State University of New Jersey (protocol number 11-030 granted to Erin Vogel). Permission to conduct the research was granted to WME by the Ministry of Research and Technology of the Republic of Indonesia (RISTEK Permit #137/SIP/FRP/SM/V/2013-2015). The data included in the present study comprise recordings collected during passive observations of wild habituated orangutans at distances typically exceeding 10 m. The population has been studied since 2003 and individual orangutans were not disturbed by observers in the execution of this study.

Study Site and Subjects

We conducted our research at the Tuanan Orangutan Research Station in Central Kalimantan, Indonesia (2^0 09' 06.1" S; 114^0 26' 26.3" E). Tuanan comprises approximately 900 hectares of secondary peat swamp forest that was selectively logged prior to the establishment of the study site in 2003 (see Erb et al., 2018 for details). For the present study, data were collected between June 2013 and May 2016 by WME and research assistants (see Acknowledgments) during focal observations of adult flanged males. Whenever flanged males were encountered, our field team followed them until they constructed a night nest and we returned to the nest before dawn the next morning to continue following the same individual. All subjects were individually recognized on the basis of unique facial features, scars, and broken or missing digits. Individuals were followed continuously for five days, unless they were lost or left the study area. During 316 partial- and full-day focal observations, we recorded 1,013 long calls from 23 known individuals.

Long Call Recording

During observations, we used all-occurrences sampling (Altmann, 1974) of long calls noting: time, GPS location, stimulus (preceded within 15 minutes by another long call, tree fall, approaching animal, or other loud sounds), and any accompanying movements or displays. Recordings of long calls were made opportunistically, using a Marantz PMD-660 solid-state recorder (44,100 Hz sampling frequency, 16 bits: Kanagawa, Japan) and a Sennheiser directional microphone (K6 power module and ME66 recording head: Wedemark, Germany). Observers made voice notes at the end of each recording noting the date and time, orangutan's name, height(s), distance(s), and movement(s), as well as the gain and microphone directionality (i.e., directly or obliquely oriented).

Long Call Analysis

For the present study, we selected a subset of recordings from 13 males from whom we had collected at least 10 high-quality long call recordings. When more than 10 long call recordings were available for a given individual, we randomly selected 10 of his recordings, stratified by study year, to balance our dataset across individuals and years. The final dataset comprised 130 long calls, 10 from each of 13 males.

Prior to annotating calls, we used Adobe Audition 14.4 to downsample recordings to 5,100 Hz (cf. Hammerschmidt & Fischer, 2019). We then generated spectrograms in Raven Pro 1.6 (K. Lisa Yang Center for Conservation Bioacoustics, 2019) with a 512-point (92.9 ms) Hann window (3 dB bandwidth = 15.5 Hz), with 90% overlap and a 512-point DFT, yielding time and frequency measurement precision of 9.25 ms and 10.8 Hz. Three observers (WME, WR, HK) annotated calls by drawing selections that tightly bounded the start and end of each pulse (Fig. S1) and assigned call type labels using the classification scheme outlined in Table 2. Except for huitus pulses (for which the rising and falling sounds are broken by silence), we operationally defined a pulse as the longest continuous sound produced on a single exhalation. Because most long calls are preceded and/or followed by a series of short bubbling sounds, we used a threshold duration of $\geq$ 0.2 seconds to differentiate pulses from these other sounds. Most selections were drawn with a fixed frequency range from 50 Hz to 1 kHz; however, in cases where the maximum fundamental frequency exceeded 1 kHz (e.g., huitus and volcano roars), selections were drawn from 50 Hz to 1.5 kHz. Occasionally, we manually reduced the frequency range of selections if there were disturbing background sounds, but only if this did not affect measures of the fundamental frequency contour or high-energy harmonics. We noted whether selections were tonal (i.e., the fundamental frequency contour was fully or partially visible) and whether they contained disturbing background noises such as birds, insects, or breaking branches.

Our selected feature set comprised 25 extracted measurements made in Raven (Table S1) as well as an additional 19 measurements estimated using the R package *warbleR* (Araya-Salas & Smith-Vidaurre, 2017). Prior to analyzing sounds in warbleR, we filtered out all pulse selections that were atonal or contained disturbing background noise, resulting in 2,270 clips. Two additional measurements (minimum and maximum) of the fundamental frequency (F0) were made using the "freq_ts" function in *warbleR* with the following settings: wavelength = 512, Hanning window, 70% overlap, 50 - 1,500 Hz, threshold = 85%. We then saved printed spectrograms depicting the F0 contours for each. One observer (WME) visually screened the minimum and maximum values of the F0 contours and scored them as accurate or inaccurate. After removing those pulses for which one or both F0 measures were inaccurate, the final full dataset comprised 1,033 pulses from 117 long calls for which all 46 parameters were measured.

Audio-Visual Analysis

To assess the inter-rater reliability (IRR) of the audio-visual analysis, we randomly selected 300 pulses (saved as individual .wav files). We included this step to remove any bias that may be introduced by information about the position or sequence of a pulse-type within a long call. Using the spectrograms and descriptions of pulse types published by Spillmann and colleagues (Spillmann et al., 2010), three observers (WME, WR, HK) labeled each sound as one of six pulse types (Fig. 1). Prior to completing this exercise, all observers had at least six months' experience classifying pulse types, which involved routine feedback and three-way discussion. We used the R package *irr* (Gamer et al., 2012) to calculate Cohen's kappa (a common statistic for assessing IRR for categorical variables) for each pair of observers, and averaged these values to provide an overall estimate of IRR (Light's kappa) across all pulse types (cf. Hallgren, 2012; Light, 1971).

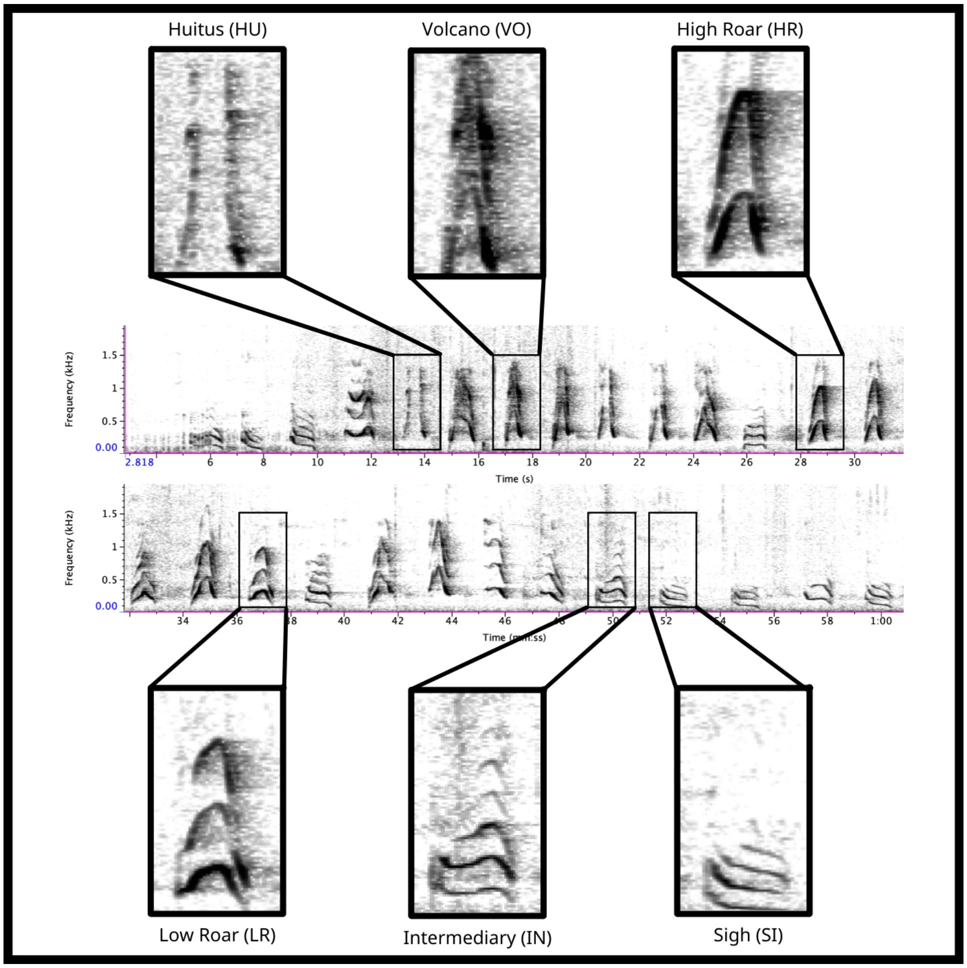


Figure 1. Spectrogram depicting long call pulse types. Pulses include HU = huitus, VO = volcano, HR = (high) roar, LR = low roar, IN = intermediary, SI = sigh. Spectrograms produced in Raven Pro 1.6.

Supervised Classification

For the supervised classification analysis, one observer (WME) manually classified all pulses (N=1,033). We then used support vector machines (SVM) in the R package *e1071* (Meyer et al., 2021) to evaluate how well pulse types could be discriminated using a supervised machine learning approach. SVMs are commonly used for supervised classification and have been successfully applied to the classification of primate calls (Clink & Klinck, 2020; Fedurek et al., 2016; Turesson et al., 2016). We used the sigmoidal kernel as previous research using SVM has found the most robust results using this kernel type (Clink & Klinck, 2020) and we estimated the best values for the gamma and cost parameters using the "tune" function. Following this, we calculated our classification accuracy using 10 iterations of leave-one-out cross-validation. Lastly, we used SVM recursive feature elimination to rank variables in order of their importance for classifying call types (cf. Clink et al., 2018). For each of the top five most influential variables identified by recursive feature elimination, we used Kruskal-Wallis nonparametric tests due to the non-normal distribution of the residuals when applying linear models. We followed these with Dunn's test of multiple comparisons to examine differences among pulse types and unsupervised clusters (described below) – applying the Benjamini-Hochberg adjustment to control the false discovery rate – using the R package *FSA* (Ogle et al., 2022).

Unsupervised Clustering

For the unsupervised analysis, we used both hard- and soft-clustering approaches. For hard clustering, we used affinity propagation, which has the advantage that it does not require the user to identify the number of clusters a priori; further, because all data points are considered simultaneously, the results are not influenced by the selection of an initial set of points (Frey & Dueck, 2007). Using the R package *apcluster* (Bodenhofer et al., 2011), we systematically varied the value of 'q' in 0.25 increments from 0 to 1. By comparing the mean silhouette coefficient for each of the cluster solutions (Wang et al., 2008), we found that $q = 0$ produced the optimal number of clusters and thus we report the results from this model. We used silhouette coefficients to quantify the stability of the resulting clusters (cf. Clink & Klinck, 2020).

For the soft clustering analysis, we used C-means fuzzy clustering. In this analysis, each pulse is assigned a membership value (m ranges from 0 = none to 1 = full accordance) for each of the clusters. We first determined the optimal number of clusters (c) by evaluating measures of internal validation and stability generated in the R package *clValid* (Brock et al., 2008) when c varied from 2 (the minimum) to 7 (one more than the previously described number of pulse types). We then systematically varied the fuzziness parameter μ from 1.1 to 5 (i.e., 1.1, 1.5, 2, 2.5, etc.: cf. Zhou et al., 2014)) using the R package 'cluster' (Maechler et al., 2021). When μ = 1, clusters are tight and membership values are binary; however, as μ increases, cases can show partial membership to multiple clusters, and the clusters themselves thereby become fuzzier and can eventually merge, leading to fewer clusters (Fischer et al., 2016). We used measures of internal validity (connectivity, silhouette width, and Dunn index) and stability (average proportion of non-overlap = APN, average distance = AD, average distance between means = ADM, and figure of merit = FOM) to evaluate the cluster solutions in the R package *clValid* (Brock et al., 2008). Once we had identified the best solution, we calculated typicality coefficients to assess the discreteness of each pulse, wherein higher values indicate pulses that are well separated from other clusters and lower values indicate pulses that are intermediate between classes (cf. Cusano et al., 2021; Wadewitz et al., 2015).

Non-linear dimensionality reduction techniques have recently emerged as fruitful alternatives to traditional linear techniques (e.g., principal component analysis) for classifying animal sounds (Sainburg et al., 2020). Uniform Manifold Approximation and Projection (UMAP) is a state-of-the-art unsupervised machine learning algorithm (McInnes et al., 2018) that has been applied to visualizing and quantifying structures in animal vocal repertoires (Sainburg et al., 2020). Like ISOMAP and t-SNE, UMAP constructs a topology of the data and projects that graph into a lower-dimensional embedding (McInnes et al., 2018; Sainburg et al., 2020) UMAP has been shown to preserve more global structure while achieving faster computation times (McInnes et al., 2018) and has been effectively applied to meaningful representations of acoustic diversity (reviewed in

Sainburg et al., 2020). This approach removes any a priori assumptions about which acoustical features are most salient or easily measured by humans.

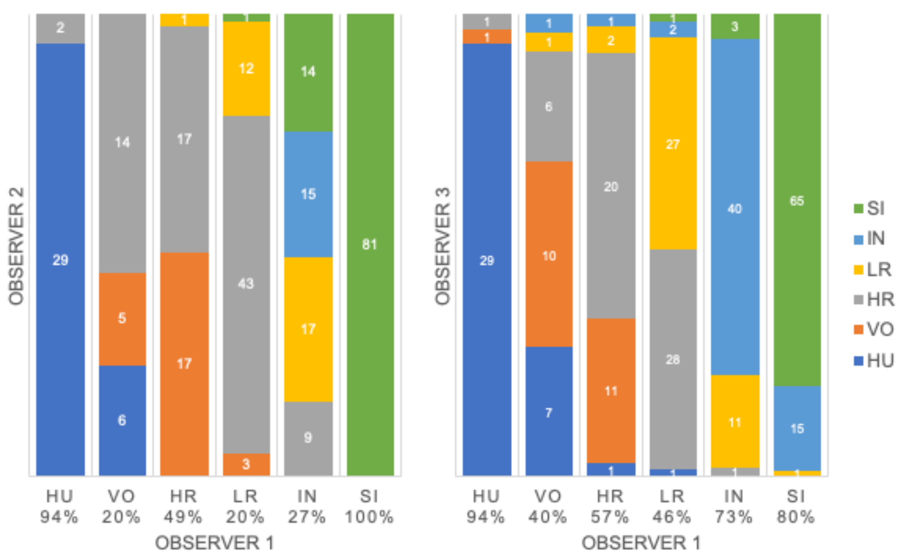
We applied UMAP separately to the 46-feature set and to time-frequency representations of extracted pulses. In the latter case, we used as inputs power density spectrograms of 0.9-s duration audio clips centered at the temporal midpoint of annotated pulses. The chosen duration was fixed irrespective of the selection duration. This means that, for short selections, the spectrograms also included sounds outside of the original selection. Short-time Fourier transforms of the clips were computed, using SciPy's (https://scipy.org/) *spectrogram* function, with a Hann window of 50 ms and 50% frame overlap (20 Hz frequency resolution, 25 ms time resolution). Spectral levels were converted to the decibel scale by applying $10 \times \log_{10}$. The bandwidth of the resulting spectrograms was limited to 50-1000 Hz prior to UMAP computation to suppress the influence of low-frequency noise on clustering. We used the *UMAP* function from the Python package *umap-learn* (McInnes et al., 2018) to compute the low-dimensional embeddings. Finally, we calculated Hopkin's statistic of clusterability on the resultant UMAP using the R package *factoextra* (Kassambara & Mundt, 2020).

Finally, we reviewed the outputs of our unsupervised clustering approaches to assess the putative number of pulses and graded variants. To identify a simple, data-driven, repeatable method for manually classifying pulse types, we began by pooling the typical pulses that belonged to each of the clusters identified by fuzzy clustering. Because F0 is a highly salient feature in long call spectrograms, our approach focused on the shape and height (or maximum frequency) of this feature. Using our revised definitions, we repeated the 1) audio-visual analysis and calculated IRR using manual labels from the same 300 pulses reviewed by the same three observers as before, and 2) SVM classification of 500 randomly selected pulses scored by a single observer (WME) following the methods described above.

RESULTS

Audio-visual analysis

Based on manual labels from three observers using audio-visual classification methods, we calculated Light's kappa $\varkappa = 0.599$ (i.e., the arithmetic mean of Cohen's Kappa for observers 1-2 = 0.48, 1-3 = 0.60, and 2-3 = 0.60), which indicated only moderate agreement among observers (Landis & Koch, 1977). Classification agreement varied widely by pulse type (Fig. 2, Table 3). Whereas huitus and sigh pulse types showed high agreement among observers (mean 2.88 and 2.77, respectively, where 3 indicates full agreement), low roar and volcano pulse types showed very low agreement (mean 2.08).



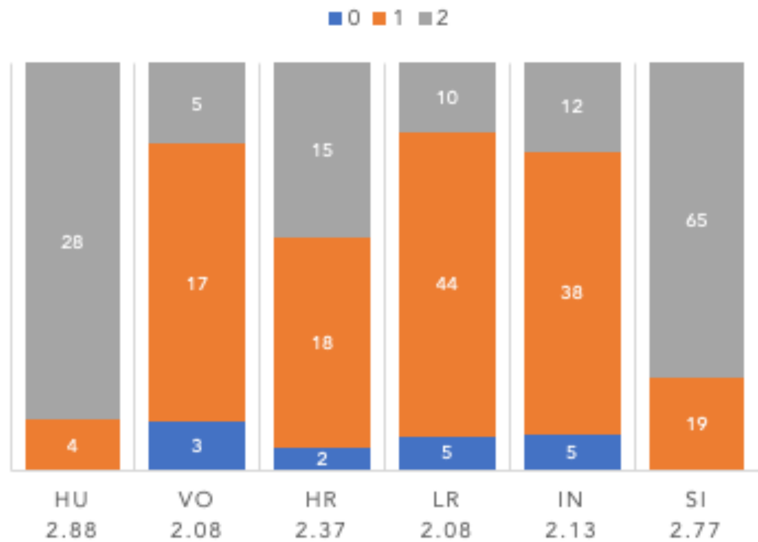


Figure 2. Audio-visual classification agreement across observers. Stacked barplots indicating (top) classification agreement by pulse type between observer 1-2 and observer 1-3 and (bottom) the number of observers who agreed on the pulse types assigned by observer 1; the average agreement index is indicated below each pulse type and demonstrates high agreement for HU and SI ($\geq$2.77), but low agreement for VO and LR (2.08).

Table 3. Mean values for A/V agreement index, SVM pulse classification accuracy, typicality coefficient, and frequency measures by pulse type.

Pulse	A/V index	SVM	Typicality	Center	Peak	Mean peak	3rd quart	1st quart
HU	2.88	77%	0.90	443.3	421.0	436.4	585.3	370.3
VO	2.08	41%	0.98	483.1	442.3	505.6	592.6	376.7
HR	2.37	61%	0.94	440.0	409.9	450.1	533.8	358.7
LR	2.08	54%	0.81	266.3	252.2	271.8	312.0	231.4
IN	2.13	52%	0.84	249.7	242.7	244.6	288.8	225.5
SI	2.77	81%	0.97	203.0	201.1	194.6	239.1	172.5

Supervised classification using extracted feature set: support vector machines

We tested the performance of SVM for the classification of orangutan long call pulse types using our full acoustic feature dataset. Using leave-one-out cross-validation, we found the average classification accuracy of pulse types was 64.8% (range: 64.28 – 65.44 $\pm$ 0.10 SE). SVM classification accuracy was higher than IRR agreement scores for most pulse types, though human observers were better at discriminating huitus and sigh pulses (Fig. 3). Classification accuracy was highly variable across pulse types. Whereas sighs and huituses were classified with the highest accuracy (81 and 77%, respectively), volcanoes were classified with the lowest accuracy (41%: Fig. 3, Table 3).

Recursive feature elimination revealed that center frequency, peak frequency, mean peak frequency, and third and first frequency quartiles were the most influential variables (Table 3). In all five influential features, high roars, huituses, and volcanoes overlapped, and in four of five features, intermediaries overlapped low roars (Fig. S2, Table S2). All other pairwise comparisons of pulse types showed significant differences in all features.

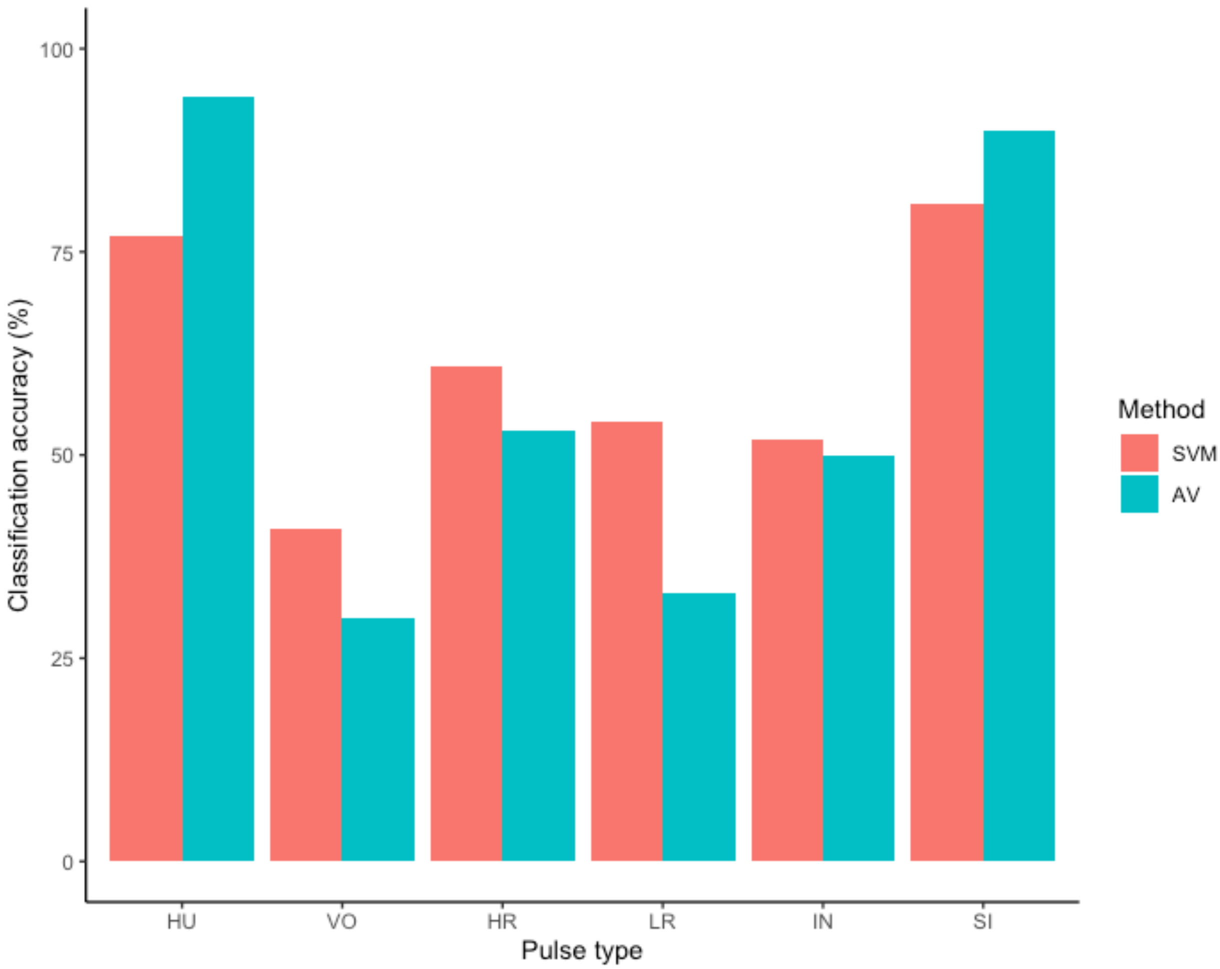


Figure 3. Barplot of classification accuracy for original pulse scheme. Comparison of classification accuracy of audio-visual classification (AV), calculated as the average agreement between three observer pairs compared to supervised machine learning classification (SVM).

Unsupervised clustering using extracted feature set: hard and soft clustering

Affinity propagation resulted in four clusters with an average silhouette coefficient of 0.32 (range: -0.22 – 0.61). Of these four clusters, two (clusters 616 and 152: Fig. 4) had relatively high silhouette coefficients (0.45 and 0.29, respectively) and separated the higher-frequency pulses (i.e., HU, VO, and HR pulses) from lower-frequency ones (i.e., LR, IN, and SI). The remaining two clusters had low silhouette coefficients (cluster 16 = 0.19, cluster 812 = 0.21) and both contained calls from all six pulse types (Fig. 4). We analyzed the separation of unsupervised clusters using the influential features identified from recursive feature elimination (Fig. S2). Two of the four clusters (16 and 152) overlapped in four of five features. These clusters primarily comprised high roars, volcanoes, and huituses.

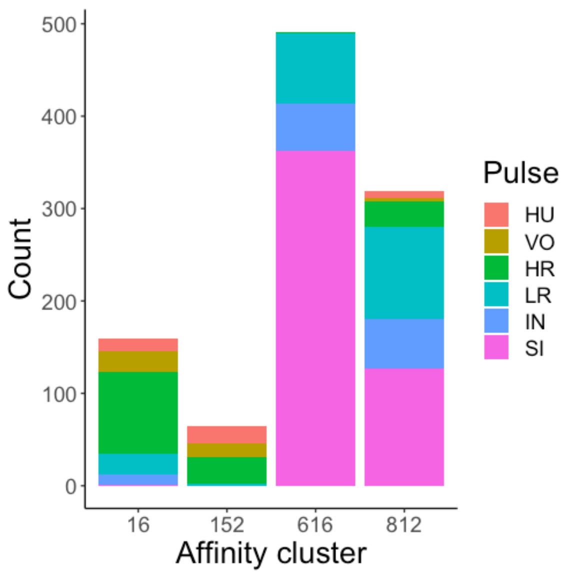


Figure 4. Stacked barplots of affinity propagation clusters showing the number of calls in each cluster classified by pulse type.

In a final approach to clustering our extracted feature set, we used c-means fuzzy clustering to provide another estimate of the number of clusters in our dataset and quantify the degree of gradation across pulse types. All three internal validity measures (connectivity, Dunn, and silhouette) and three of four stability measures (APN, AD, and ADM) indicated that the two-cluster solution was optimal. Only FOM indicated a 3-cluster solution was marginally more stable (0.855 for 2 vs. 0.860 for 3 clusters). We found that mu = 1.1 yielded the highest average silhouette width (0.312); silhouette widths decreased as mu increased.

Typicality coefficients were high overall (mean: 0.92 + 0.006 SE, Fig. 5) but varied widely by pulse type. Whereas volcanoes and sighs had the highest typicality coefficients (0.98 and 0.97, respectively) and intermediaries and low roars had the lowest coefficients (0.84 and 0.81, respectively, Table 3). Pairwise comparisons of typicality coefficients showed that typicality coefficients for low roars and intermediaries were significantly lower than those of all other pulse types but did not significantly differ between these two pulses (Fig. S2, Table S2).

We determined the thresholds for typical (>0.976) and atypical calls (<0.855) (cf. Wadewitz et al., 2015). Overall, 69% of calls were 'typical' for their cluster and 17% were 'atypical'; however, pulse types varied greatly (Fig. 6). Whereas sighs and volcanoes had a high proportion of typical calls (85% and 80% respectively), low roars and intermediaries had a high proportion of atypical calls (44% and 40% respectively).

Typical calls were found in both clusters (Fig. 6). Typical calls in cluster one included high roars, huituses, low roars, and volcanoes and those in cluster two included sighs, low roars, and intermediaries. Whereas typical sighs, huituses, and volcanoes were found in only one cluster (and only 1-2 intermediaries and high roars were typical for a secondary cluster), 24% of low roars belonged to a secondary cluster. Overall, cluster one comprised 189 typical and 99 atypical calls (53% and 28% of 353 calls, respectively) and cluster two comprised 526 typical and 75 atypical calls (77% and 11% of 680 calls, respectively), indicating that calls in cluster two were better separated from other call types than those in cluster one. We compared typical calls in each cluster and found that calls in different clusters significantly differed from each other in all five influential features (Fig. S2, Table S2a).

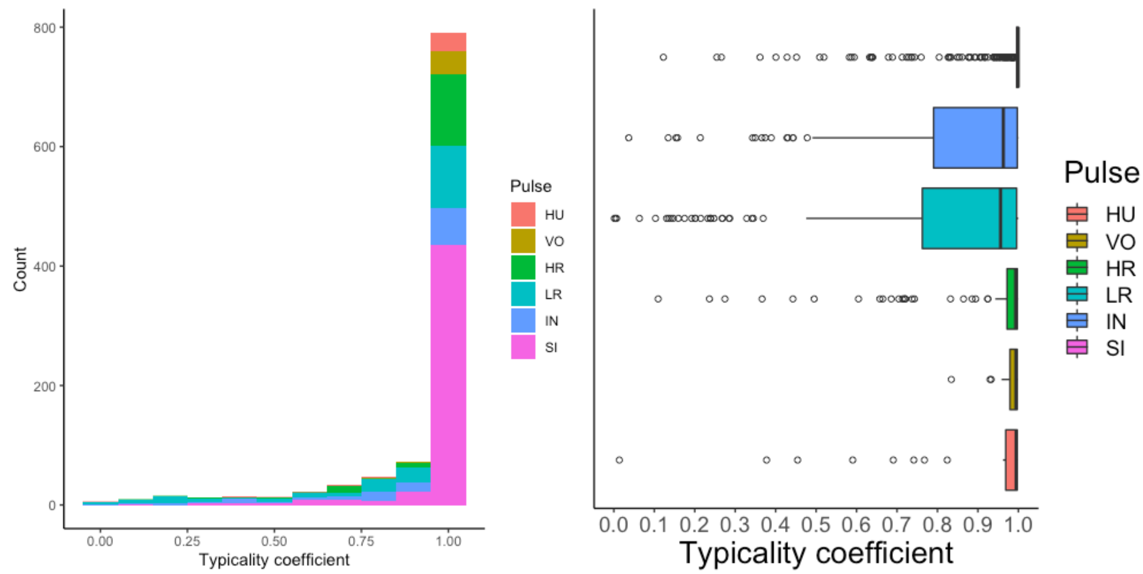


Figure 5. Typicality coefficients for each pulse type a) Histogram showing the distribution of coefficients and b) boxplot showing typicality values for each pulse type. Typicality thresholds were calculated following (Wadewitz et al., 2015). Typical calls were those whose typicality coefficients exceeded 0.976 and atypical calls were those below 0.855.

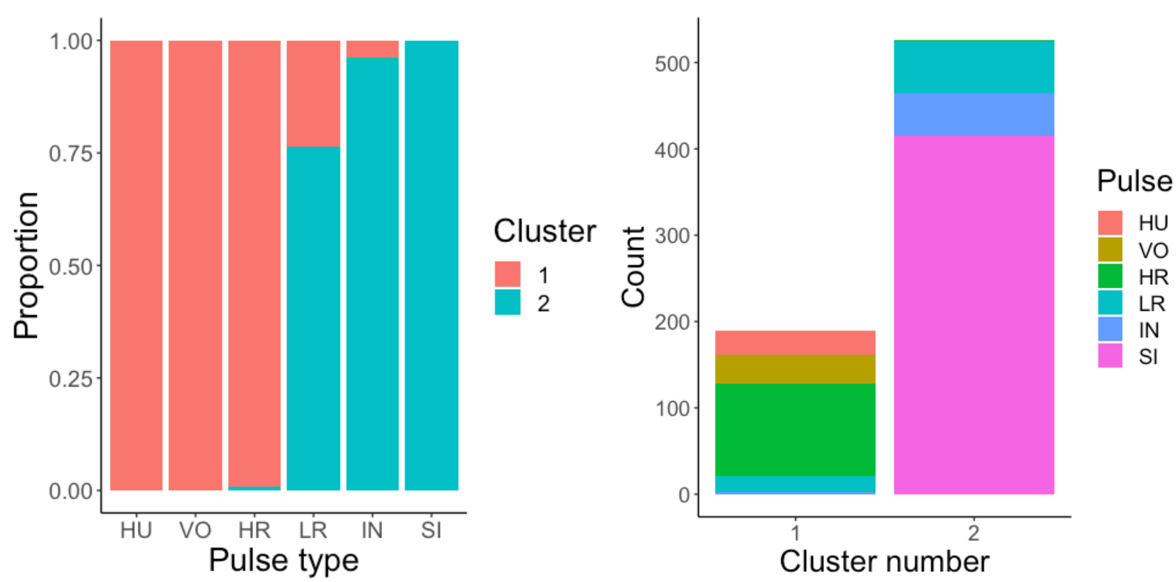


Figure 6. Stacked barplots of typical calls a) the proportion of each pulse type that was typical for each cluster and b) the number of typical calls in each cluster classified by pulse type.

UMAP visualization of extracted features and spectrograms

We used UMAP to visualize the separation of individual pulses using our extracted feature set, comparing the cluster results from affinity propagation and fuzzy clustering with manual classification (Fig. 7). We also used UMAP to visualize the separation of pulses based on the power density spectrograms (Fig. 7). For both datasets, it appears that there are two loose and incompletely separated clusters as well as a smaller number of pulses that grade continuously between the two clusters. The Hopkins statistic of clusterability for the extracted feature set was 0.940 and 0.957 for the power spectrograms, both of which indicate strong clusterability of calls.

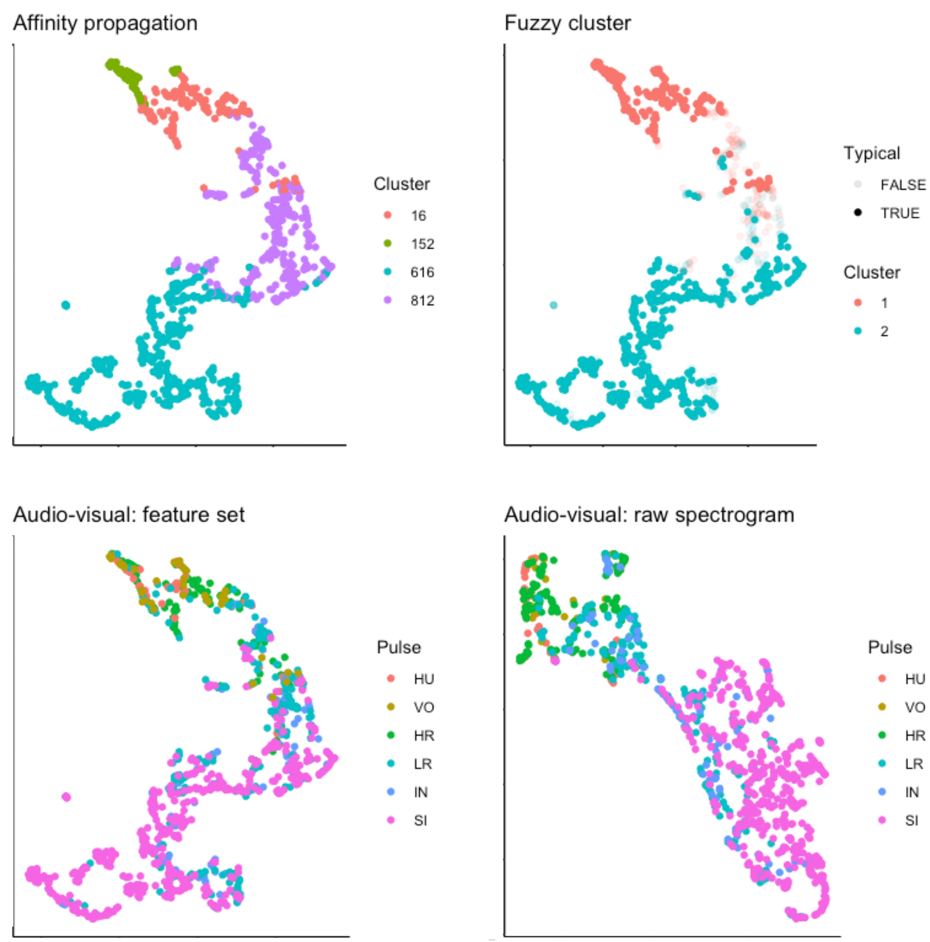


Figure 7. UMAP projection of 46-feature dataset. Colors indicate four clusters identified using unsupervised affinity propagation (upper left), two clusters and typical calls identified by fuzzy clustering (upper right), six pulse types labeled by human observer using the extracted feature set (lower left), and raw power density spectrograms (lower right).

Identification and evaluation of a new classification scheme

Collectively, our unsupervised clustering approaches showed broad agreement for a two-cluster solution with graded pulses occurring along a spectrum between the two classes. In fuzzy typical cluster 1, the mean value for F0 max was 764.3 Hz $\pm$ 351.5 SD Hz (range = 320-1,500); whereas for those pulses belonging to fuzzy typical cluster 2, the mean value of F0 max was 225.3 $\pm$

SD 67.9 SD Hz (range = 80-440). Pulses that were not typical for either cluster had a mean F0 max of 345.8 $\pm$ 159.9 SD Hz. Based on these patterns, as well as the shape of the F0 (a feature that was commonly used to distinguish among pulse types in previous studies), we distinguished among pulses as follows: **Roar (R)** = F0 ascends and reaches its maximum (>350 Hz) at or near the midpoint of the pulse before descending, **Sigh (S)** = F0 descends and reaches its maximum (typically, but not always < 350 Hz) at start of the pulse (i.e., no ascending portion of F0), and **Intermediate (I)** = either a) maximum F0 value occurs at the start of the pulse but with an ascending portion later in pulse, or b) F0 ascends and reaches its maximum (<350 Hz) at or near the midpoint of the pulse.

These revised definitions yielded Light's kappa $\varkappa = 0.838$ (i.e., the arithmetic mean of Cohen's Kappa for observers 1-2 = 0.84, 1-3 = 0.86, and 2-3 = 0.78), indicating near-perfect agreement among observers (Landis & Koch, 1977). Classification agreement varied only slightly by pulse type, with roars showing the highest agreement among observers (mean 2.92, where 3 indicates full agreement), and intermediaries and sighs showing slightly lower agreement (mean 2.79 and 2.72, respectively). Using leave-one-out cross-validation, we found the average classification accuracy of pulse types using SVM was 82.1% (range: 80.8 – 85.0 $\pm$ 0.47 SE). SVM classification accuracy was lower than IRR agreement scores for most pulse types (Fig. 8) but both roars and sighs were classified with high agreement using both methods.

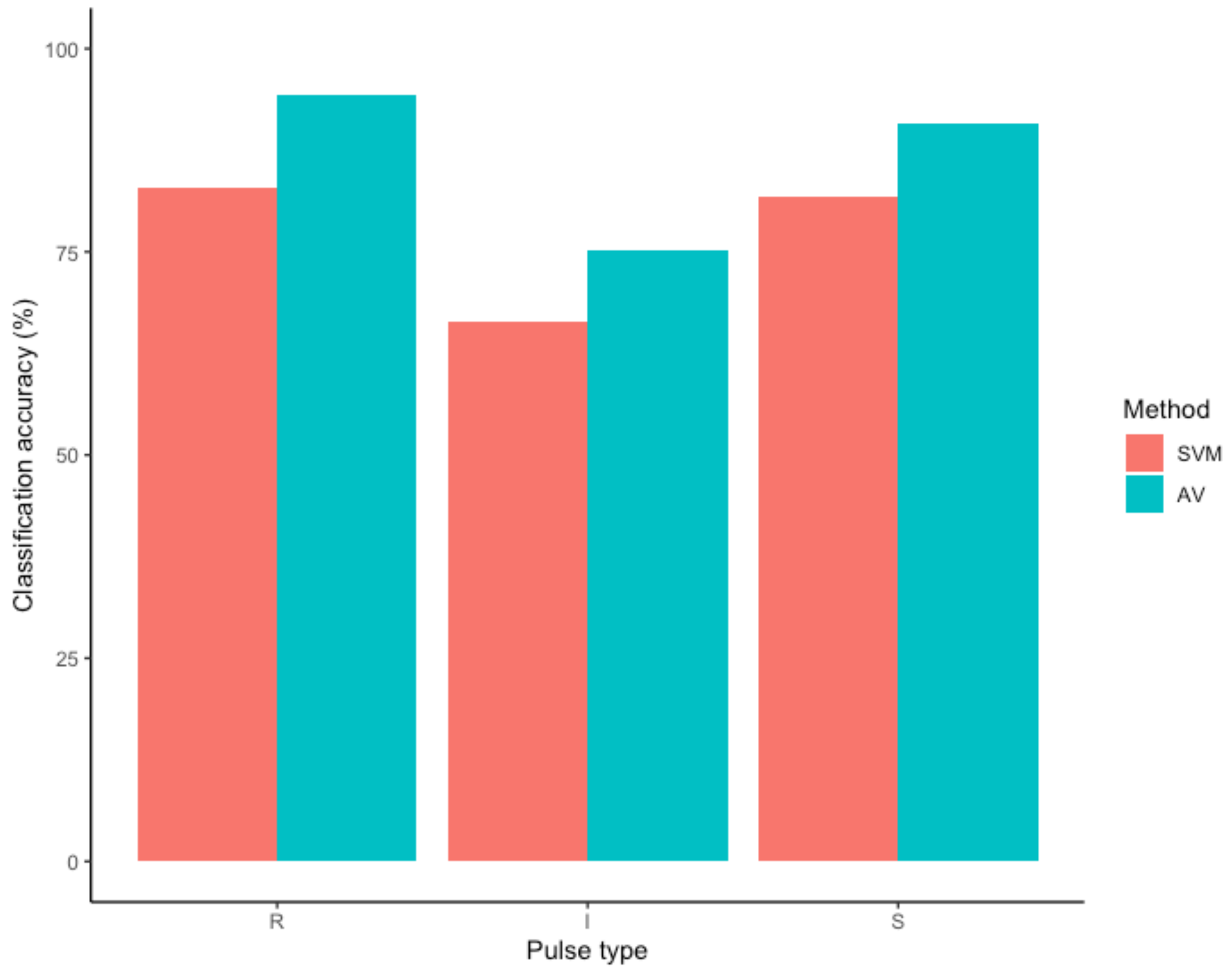


Figure 8. Barplot of classification accuracy for revised pulse scheme. Comparison of classification accuracy of audio-visual classification (AV), calculated as the average agreement between three observer pairs compared to supervised machine learning classification (SVM).

DISCUSSION

Here we present an extensive qualitative and quantitative assessment of the vocal complexity of the long-call vocalizations of Bornean orangutans. Relying on a large dataset comprising 46 acoustic measurements from 1,033 pulses from 117 long calls recorded from 13 males, we compared the ability of human observers and supervised and unsupervised machine-learning techniques to discriminate unique call (or pulse) types. Three human observers performed relatively well at discriminating two pulse types – huitus and sigh – but our inter-rater reliability score (i.e., Light's kappa) showed only moderate agreement across the six pulse types. Although support vector machines (SVM) performed better than human observers in classifying most pulse types (except for huitus and sigh pulses), the overall accuracy was less than 65%. Like humans, SVM's were best at discriminating huitus and sigh pulse types but performed relatively poorly for the others. Poor classification accuracy across audio-visual and supervised machine learning approaches indicates that these six pulse types are not discrete. This finding suggests that attempting comparisons of different pulse types (cf. Davila Ross & Geissmann, 2007; Spillmann et al., 2010) across observers or studies is not advisable, since these classes are not reliably reproduced by different observers or well separated by a robust set of acoustic features.

Having demonstrated that these six pulse types were not well discriminated, we turned to unsupervised clustering to characterize and classify the diversity of pulses comprising orangutan long calls. Whereas hard clustering, such as affinity propagation, seeks to identify a set of high-quality exemplars and corresponding clusters (Frey & Dueck, 2007), soft, or fuzzy, clustering is an alternative or complementary approach to evaluate and quantify the discreteness of call types within a graded repertoire (Cusano et al., 2021; Fischer et al., 2016; Wadewitz et al., 2015). Although the hard and soft unsupervised techniques yielded different clustering solutions – four clusters for affinity propagation and two for fuzzy c-means – both methods showed relatively poor separation across pulse types. Importantly, both hard and soft clustering solutions separated high-frequency pulses (i.e., HU, VO, and HR) from low-frequency ones (i.e., LR, IN, SI), but low roars and intermediaries showed low typicality coefficients and occurred in both fuzzy clusters. Together, the results of unsupervised clustering support our interpretation of the manual and supervised classification analysis in demonstrating that orangutan long calls contain a mixture of discrete and graded pulse types.

We used a final approach, UMAP, to visualize the separation and quantify the clusterability of call types. Because the number and type of features selected can have a strong influence on the cluster solutions and their interpretations (Fischer et al., 2016; Wadewitz et al., 2015), we compared the results of our extracted 46-feature dataset with raw power density spectrograms as inputs. Both datasets yielded similar and high Hopkin's statistic values, indicating strong clusterability of calls. At the same time, both datasets generated a V-shaped cloud of points showing two large loose clusters with a spectrum of points lying along a continuum between them.

Based on our comprehensive evaluation of orangutan pulse types, we have proposed a revised approach to the classification of orangutan pulses that we hope provides improved reproducibility for future researchers. We recommend using the following terms to categorize the range of pulse types comprising orangutan long calls: 1) 'Roar' for high-frequency pulses, 2) 'Sigh' for low-frequency pulses, and 3) 'Intermediate' for graded pulses that fall between these two extremes. We have provided detailed descriptions of each of these pulse types and demonstrated that they can be easily and reliably identified among different observers and exhibit high classification accuracy using SVM.

Thus, we find that orangutan calls can be clustered into three pulse types (two discrete and one graded). The low diversity of call types suggests that these vocalizations are not particularly complex. Like the long-calls of other apes (chimpanzees, *Pan troglodytes schweinfurthii*: Arcadi, 1996; Marler & Hobbett, 1975; gibbons, *Hylobates* spp: Marshall & Marshall, 1976), orangutan long calls typically comprise an intro and/or build-up phase (quiet, staccato grumbles, not analyzed in the present study), climax (high-energy, high-frequency roars), and a let-down phase (low-energy sighs). The low number of discrete pulse types could be interpreted as support for the hypothesis that long-distance signals have been selected to facilitate signal recognition in dense and noisy habitats (Marler et al., 1975). Yet, there is a spectrum of intermediate call types that yield a continuous gradation of sounds across phases and pulses, that seems to greatly boost the complexity of this signal.

Unfortunately, only a handful of studies have quantified the gradedness of animal vocal systems (but see Cusano et al., 2021; Fischer et al., 2017; Taylor et al., 2021; Wadewitz et al., 2015). Consequently, we are still lacking a comprehensive framework through which to quantify and interpret vocal complexity vis-à-vis graded repertoires (Fischer et al., 2017). Future research will explore the production of graded call types across individuals and call types to examine the sources of variation and the potential role of graded call types in orangutan communication.

In summary, we evaluated a range of supervised and unsupervised approaches to classifying and clustering sounds in animal vocal repertoires. We used a combination of traditional audio-visual methods and modern machine learning techniques that relied on human eyes and ears, a set of 46 features measured from spectrograms, and raw power density spectrograms to triangulate diverse datasets and methods to answer a simple question: how many pulse types exist within orangutan long calls, how can they be distinguished, and how graded are they? While each approach has its strengths and limitations, taken together, they can lead to a more holistic understanding of call types within graded repertoires and contribute to a growing body of literature documenting the graded nature of animal communication systems.

SUPPLEMENTARY MATERIALS

Table S1. Table describing features measured in Raven Pro and warbleR (Specan and freq_ts**)**

No	Program	Feature	Description
1	Raven	Delta.Time.s	difference between Begin Time and End Time for the selection (s)
2	Raven	Freq.5%.Hz	frequency at which summed energy exceeds 5% of total energy
3	Raven	Freq.95%.Hz	frequency at which summed energy exceeds 95% of total energy
4	Raven	Agg.Entropy.bits	aggregate entropy measures the disorder in a sound by analyzing the energy distribution (pure tone ~ 0)
5	Raven	Avg.Entropy.bits	average entropy measurement describes the amount of disorder for a typical spectrum within the selection
6	Raven	BW.50%	difference between the 25% and 75% frequencies (Hz)
7	Raven	BW.90%	difference between the 5% and 95% frequencies (Hz)
8	Raven	Center.Freq	frequency that divides the selection into two frequency intervals of equal energy (Hz)
9	Raven	Center.Time.Rel.	proportion of selection at which 50% of the sound energy has an earlier time
10	Raven	Dur.50%	difference between the 25% and 75% times (s)
11	Raven	Dur.90%	difference between the 5% and 95% times (s)
12	Raven	Freq.25%	frequency at which summed energy exceeds 25% of total energy (Hz)
13	Raven	Freq.75%	frequency at which summed energy exceeds 75% of total energy (Hz)
14	Raven	Peak.Freq	frequency at which Peak Power occurs within the selection (Hz)
15	Raven	PFC.Avg.Slope	Mean of the Peak Frequency Contour Slope Series of numbers (Hz/ms)
16	Raven	PFC.Max.Freq	Maximum of the Peak Frequency Contour Series of numbers (Hz)
17	Raven	PFC.Max.Slope	Maximum of the Peak Frequency Contour Slope Series of numbers (Hz/ms)
18	Raven	PFC.Min.Freq	Minimum of the Peak Frequency Contour Series of numbers (Hz)
19	Raven	PFC.Min.Slope	Minimum of the Peak Frequency Contour Slope Series of numbers (Hz/ms)
20	Raven	PFC.Num.Inf.Pts	Number of times the slope changes sign in Peak Frequency Contour Slope Series of numbers
21	Raven	Peak.Time.Rel.	proportion of selection at first time in a selection at which amplitude equal to Peak Amplitude occurs
22	Raven	Time.25%.Rel.	proportion of selection at which 25% of the sound energy has an earlier time
23	Raven	Time.5%.Rel.	proportion of selection at which 5% of the sound energy has an earlier time
24	Raven	Time.75%.Rel.	proportion of selection at which 75% of the sound energy has an earlier time

25	Raven	Time.95%.Rel.	proportion of selection at which 95% of the sound energy has an earlier time
26	specan	meanfreq	mean of frequency spectrum (kHz)
27	specan	sd	standard deviation of frequency (kHz)
28	specan	skew	skewness: asymmetry of the spectrum
29	specan	kurt	kurtosis: peakedness of the spectrum
30	specan	sp.ent	energy distribution of the frequency spectrum (pure tone ~ 0)
31	specan	time.ent	energy distribution on the time envelope (pure tone ~ 0)
32	specan	entropy	spectrographic entropy: product of time x spectral entropy
33	specan	sfm	spectral flatness (pure tone ~ 0)
34	specan	meandom	average of dominant frequency measured across the acoustic signal
35	specan	mindom	minimum of dominant frequency measured across the acoustic signal
36	specan	maxdom	maximum of dominant frequency measured across the acoustic signal
37	specan	dfrange	range of dominant frequency measured across the acoustic signal
38	specan	modindx	modulation index: cumulative difference between adjacent dominant frequencies / dominant frequency range
39	specan	startdom	dominant frequency measurement at the start of the signal
40	specan	enddom	dominant frequency measurement at the end of the signal
41	specan	dfslope	slope of the change in dominant frequency through time
42	specan	meanpeakf	frequency with highest energy from the mean frequency spectrum
43	specan	Freq_IQR	interquartile frequency range. Frequency range between 'freq.Q25' and 'freq.Q75' (kHz)
44	specan	Time_IQR	interquartile time range. Time range between 'time.Q25' and 'time.Q75' (s)
45	freq_ts	F0_min	frequency at which F0 contour is at its minimum value (kHz)
46	freq_ts	F0_max	frequency at which F0 contour reaches its maximum value (kHz)

Figure S1. Example of annotated spectrogram

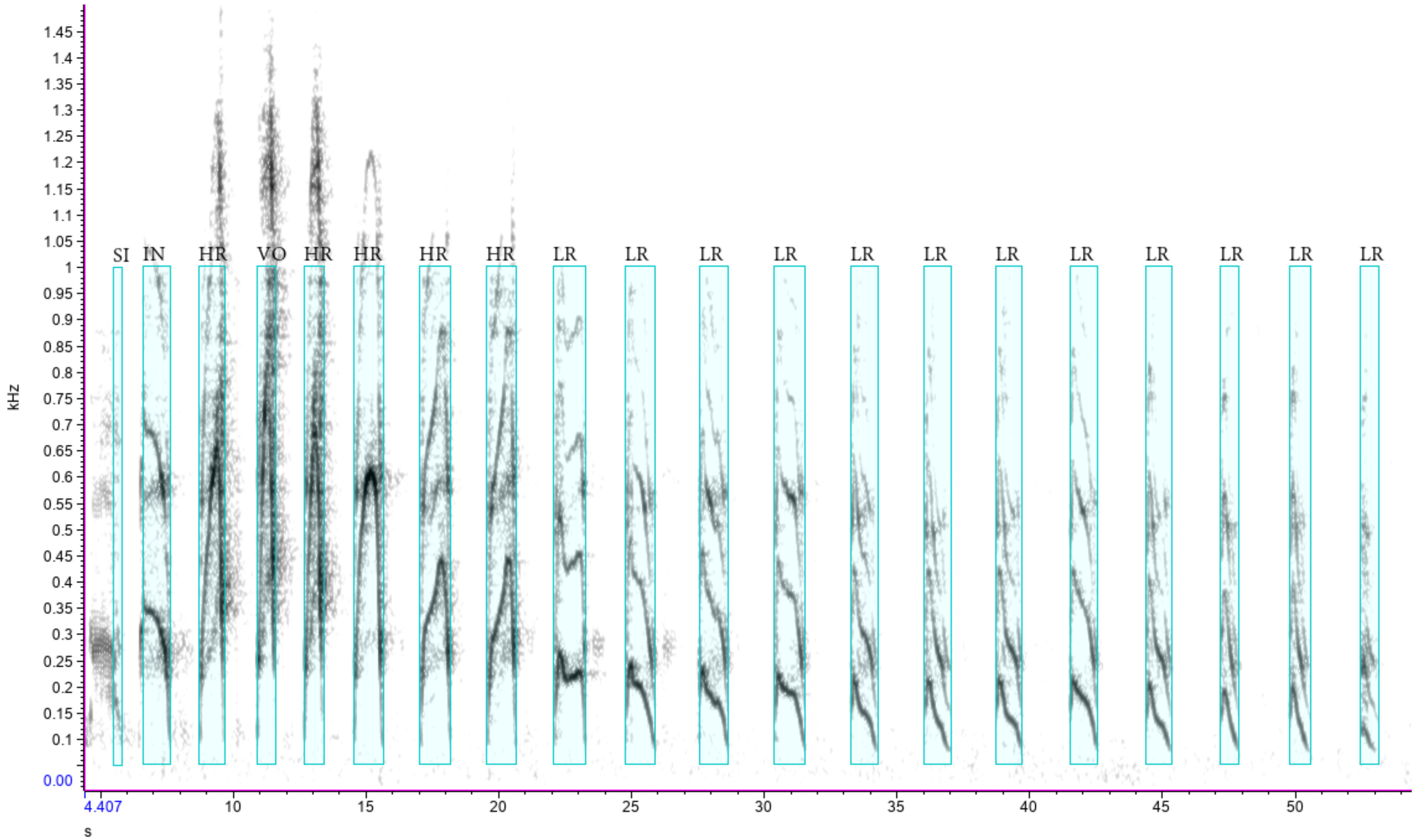
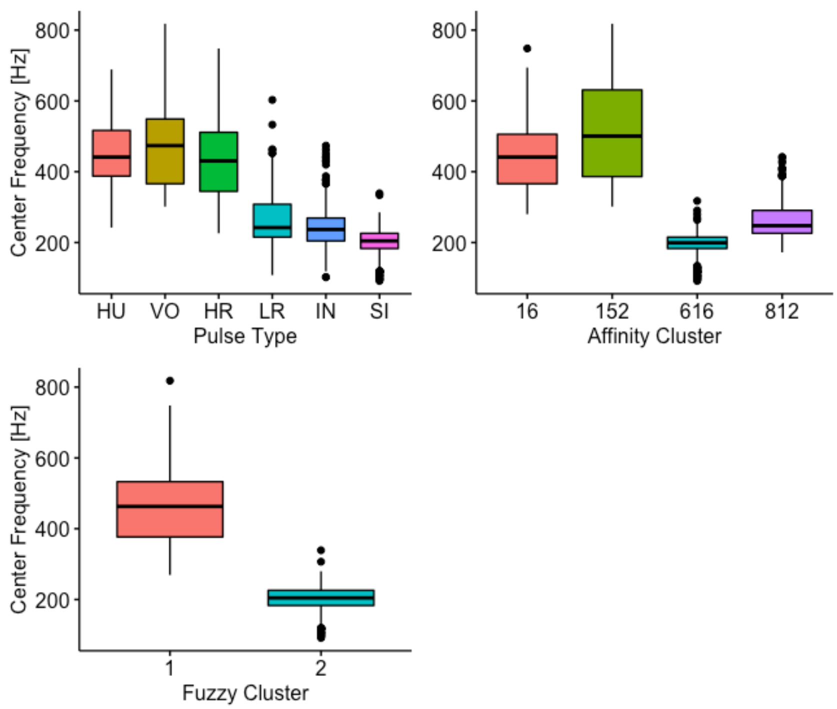
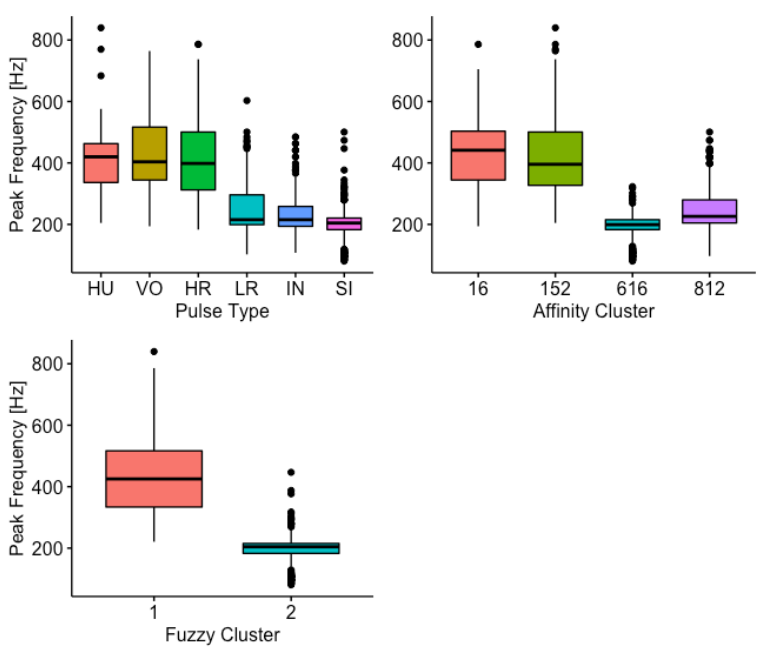
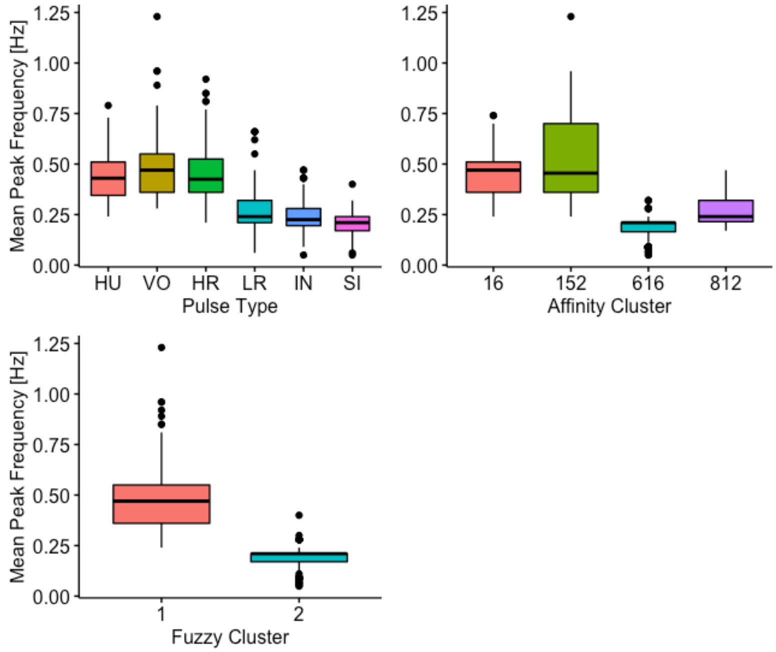
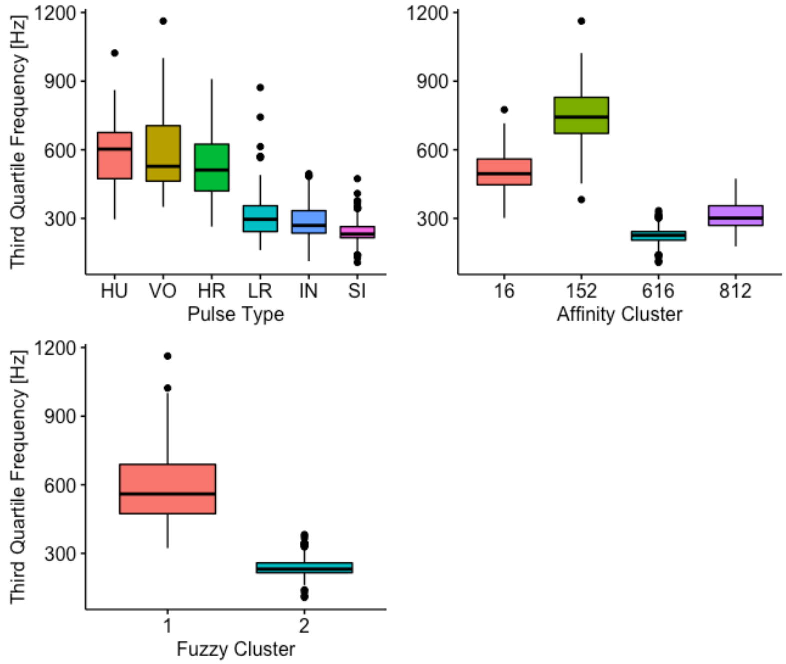


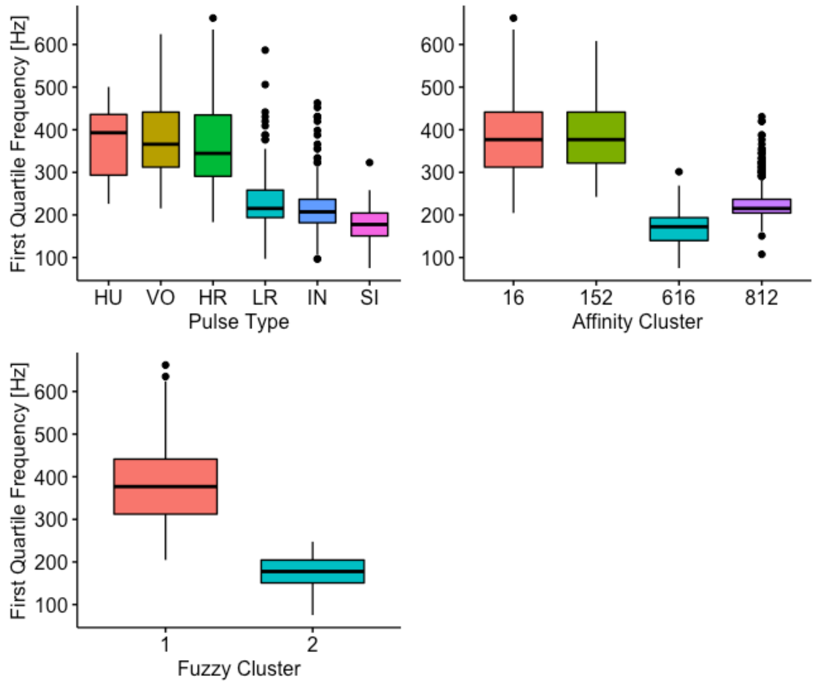
Figure S2. Boxplots of features that differed across human-labeled pulses (upper left), affinity propagation clusters (upper right), and typical calls in fuzzy clusters (lower left) for each of the following influential features: a) center frequency, b) peak frequency, c) mean peak frequency, d) third quartile frequency, e) first quartile frequency.






Mean Peak Frequency [Hz]
1.25
1.00
0.75
0.50
0.25
0.00
HU VO HR LR IN SI
Pulse Type
16 152 616 812
Affinity Cluster
1 2
Fuzzy Cluster


Third Quartile Frequency [Hz]
1200
900
600
300
HU VO HR LR IN SI
Pulse Type
16 152 616 812
Affinity Cluster
1 2
Fuzzy Cluster


First Quartile Frequency [Hz]
600
500
400
300
200
100
HU
VO
HR
LR
IN
SI
Pulse Type
16
152
616
812
Affinity Cluster
1
2
Fuzzy Cluster

552

Table S2a. Table summarizing results of Kruskal-Wallis tests for differences among pulses or clusters identified by human observers, affinity propagation, and fuzzy clustering for each of the top five influential variables.

	A/V			AFFINITY			FUZZY		
Variable	χ^2	**df**	**p**	χ^2	**df**	**p**	χ^2	**df**	**p**
Center	557.81	5.00	0.00	738.53	3.00	0.00	417.16	1.00	0.00
Peak	425.31	5.00	0.00	588.78	3.00	0.00	406.49	1.00	0.00
Mean peak	528.18	5.00	0.00	677.78	3.00	0.00	421.95	1.00	0.00
Third quart	570.30	5.00	0.00	777.79	3.00	0.00	416.69	1.00	0.00
First quart	536.72	5.00	0.00	684.50	3.00	0.00	414.47	1.00	0.00

Table S2b. Table summarizing results of Dunn tests for pair-wise differences among pulses identified by human observers for each of the top five influential variables.

A/V	Center			Peak			Mean peak			Third quart			First quart		
Pair	Z	P.unadj	P.adj	Z	P.unadj	P.adj	Z	P.unadj	P.adj	Z	P.unadj	P.adj	Z	P.unadj	P.adj
HR-HU	-0.21	0.84	0.84	-0.54	0.59	0.63	0.12	0.90	0.90	-0.60	0.55	0.64	-0.54	0.59	0.68
HR-IN	10.29	0.00	0.00	9.12	0.00	0.00	10.59	0.00	0.00	10.52	0.00	0.00	9.04	0.00	0.00
HR-LR	9.80	0.00	0.00	9.81	0.00	0.00	9.68	0.00	0.00	9.97	0.00	0.00	9.35	0.00	0.00
HR-SI	19.72	0.00	0.00	17.10	0.00	0.00	19.35	0.00	0.00	19.86	0.00	0.00	19.26	0.00	0.00
HR-VO	-0.67	0.50	0.58	-0.84	0.40	0.50	-0.44	0.66	0.71	-0.55	0.58	0.62	-0.50	0.62	0.66
HU-IN	7.12	0.00	0.00	6.65	0.00	0.00	7.00	0.00	0.00	7.65	0.00	0.00	6.60	0.00	0.00
HU-LR	6.31	0.00	0.00	6.66	0.00	0.00	5.90	0.00	0.00	6.81	0.00	0.00	6.37	0.00	0.00
HU-SI	11.40	0.00	0.00	10.27	0.00	0.00	10.83	0.00	0.00	11.90	0.00	0.00	11.49	0.00	0.00
HU-VO	-0.36	0.72	0.77	-0.23	0.82	0.82	-0.44	0.66	0.76	0.04	0.97	0.97	0.04	0.97	0.97
IN-LR	-1.83	0.07	0.08	-0.57	0.57	0.66	-2.26	0.02	0.03	-1.91	0.06	0.07	-0.91	0.36	0.45
IN-SI	5.60	0.00	0.00	4.63	0.00	0.00	4.91	0.00	0.00	5.46	0.00	0.00	6.70	0.00	0.00
IN-VO	-7.70	0.00	0.00	-7.06	0.00	0.00	-7.67	0.00	0.00	-7.74	0.00	0.00	-6.67	0.00	0.00
LR-SI	9.44	0.00	0.00	6.48	0.00	0.00	9.18	0.00	0.00	9.38	0.00	0.00	9.51	0.00	0.00
LR-VO	-6.92	0.00	0.00	-7.09	0.00	0.00	-6.60	0.00	0.00	-6.90	0.00	0.00	-6.45	0.00	0.00
SI-VO	-12.16	0.00	0.00	-10.83	0.00	0.00	-11.70	0.00	0.00	-12.12	0.00	0.00	-11.71	0.00	0.00

Table S2c. Table summarizing results of Dunn tests for pair-wise differences among clusters identified by affinity propagation for each of the top five influential variables.

AFFINITY	**Center**			**Peak**			**Mean peak**			**Third quart**			**First quart**		
Pair	Z	P.unadj	P.adj	Z	P.unadj	P.adj	Z	P.unadj	P.adj	Z	P.unadj	P.adj	Z	P.unadj	P.adj
152-16	0.84	0.40	0.40	-0.43	0.66	0.66	0.24	0.81	0.81	2.33	0.02	0.02	0.20	0.84	0.84
152-616	16.64	0.00	0.00	14.23	0.00	0.00	15.60	0.00	0.00	18.09	0.00	0.00	15.68	0.00	0.00
16-616	22.88	0.00	0.00	21.43	0.00	0.00	22.33	0.00	0.00	22.56	0.00	0.00	22.51	0.00	0.00
152-812	7.95	0.00	0.00	7.59	0.00	0.00	7.57	0.00	0.00	8.88	0.00	0.00	7.73	0.00	0.00
16-812	9.94	0.00	0.00	11.37	0.00	0.00	10.31	0.00	0.00	8.97	0.00	0.00	10.60	0.00	0.00
616-812	-15.61	0.00	0.00	-11.83	0.00	0.00	-14.41	0.00	0.00	-16.52	0.00	0.00	-14.26	0.00	0.00

ACKNOWLEDGEMENTS

Land Acknowledgment: Cornell University is located on the traditional homelands of the Gayogo̱hó꞉nǫ’, members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic and contemporary presence on this land. The Confederacy precedes the establishment of Cornell University, New York State, and the United States of America. We acknowledge the painful history of Gayogo̱hó꞉nǫ’ dispossession, and honor the ongoing connection of Gayogo̱hó꞉nǫ’ people, past and present, to these lands and waters. We likewise acknowledge the Dayak people who have tended and cared for the forests and waters of Kalimantan for millennia. We extend our deepest gratitude to the Tuanan community members who generously shared their deep knowledge of the forest and provided direct support in the field research, including Pak Nafisa, Pak Irwan, Pak Isman, Pak Rahmadt, Suga, Suwi, Pak Andre, and Awan. We thank the Tuanan Research Station’s directors Sri Suci Utami Atmoko, Maria van Noordwijk, and Carel van Schaik for access to facilities and resources and for maintaining the long-term demographic records on the study animals and Erin Vogel for intellectual, logistical, and financial support that made the three-year study possible. We are grateful to all Tuanan researchers and particularly Lynda Dunkel and Brigitte Spillmann for their habituation efforts. Our Universitas Nasional counterparts Tomi Ariyanto, Jito Sugardjito, Sri Suci Utami Atmoko, Didik Prasetyo, and Astri Zulfa provided essential administrative and logistical support; and Indonesian partners including RISTEK-DIKTI, BKSDA, LIPI, KLHK, and the Borneo Orangutan Survival Foundation were vital in issuing permits and support. We are grateful to Beth Barrow, Alli Hofner, and Yann Quenet for collecting some of the recordings included in this study as well as Emily Martines and Jessica Lecorchick for their contributions to long call annotations. Lastly, we thank our Yang Center colleagues Holger Klinck and Russ Charif for valuable conversations at the early stages of acoustic analyses. This project would not have been possible without funding (PI and co-PIs indicated) from Rutgers University

(Erin Vogel), The Center for Human Evolutionary Studies (Erin Vogel), USAID (No. AID-497-A-13-00005: Erin Vogel, Robert Scott, Jito Sugardjito), and the American Association of University Women (WME).

REFERENCES

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Unsupervised and Semi-Supervised Learning: Supervised and Unsupervised Learning for Data Science* (pp. 3–21). Springer International Publishing.

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, *49*, 227–267.

Araya-Salas, M., & Smith-Vidaurre, G. (2017). WarbleR: an r package to streamline analysis of animal acoustic signals. *Methods in Ecology and Evolution*, *8*(2), 184–191.

Arcadi, A. C. (1996). Phrase structure of wild chimpanzee pant hoots: Patterns of production and interpopulation variability. *American Journal of Primatology*, *39*(3), 159–178.

Askew, J. A., & Morrogh-Bernard, H. C. (2016). Acoustic characteristics of long calls produced by male orangutans (Pongo pygmaeus wurmbii): Advertising individual identity, context, and travel direction. *Folia Primatologica*, *87*(5), 305–319.

Blumstein, D. T., & Armitage, K. B. (1997). Does sociality drive the evolution of communicative complexity? A comparative test with ground-dwelling sciurid alarm calls. *The American Naturalist*, *150*(2), 179–200.

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: An R package for affinity propagation clustering. *Bioinformatics*, *27*, 2463–2464.

https://doi.org/10.1093/bioinformatics/btr406

Bradbury, J. W., & Vehrencamp, S. L. (2011). *Principles of animal communication (2nd ed.)*. Sinauer Associates.

Brady, B., Hedwig, D., Trygonis, V., & Gerstein, E. (2020). Classification of Florida manatee (*Trichechus manatus latirostris*) vocalizations. *The Journal of the Acoustical Society of America*, *147*(3), 1597–1606. https://doi.org/10.1121/10.0000849

Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for Cluster Validation. *Journal of Statistical Software*, *25*(4), 1–22.

Clink, D. J., Crofoot, M. C., & Marshall, A. J. (2018). Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia. *Bioacoustics*, 1–17.

Clink, D. J., & Klinck, H. (2020). Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. *Methods in Ecology and Evolution*.

Cunningham, P., Cord, M., & Delany, S. (2008). Supervised learning. In M. Cord & P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia*. Springer. https://doi.org/10.1007/978-3-540-75171-7_2

Cusano, D. A., Noad, M. J., & Dunlop, R. A. (2021). Fuzzy clustering as a tool to differentiate between discrete and graded call types. *JASA Express Letters*, *1*(6), 061201.

Davila Ross, M., & Geissmann, T. (2007). Call diversity of wild male orangutans: A phylogenetic approach. *American Journal of Primatology*, *69*(3), 305–324.

Elie, J., & Theunissen, F. (2016). The vocal repertoire of the domesticated zebra finch: A data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal Cognition*, *19*(2), 285–315.

Erb, W. M., Barrow, E. J., Hofner, A. N., Utami-Atmoko, S. S., & Vogel, E. R. (2018). Wildfire smoke impacts activity and energetics of wild Bornean orangutans. *Scientific Reports*, *8*(1), 7606. https://doi.org/10.1038/s41598-018-25847-1

Fedurek, P., Zuberbühler, K., & Dahl, C. D. (2016). Sequential information in a great ape utterance. *Scientific Reports*, *6*, 38226.

Fischer, J., Wadewitz, P., & Hammerschmidt, K. (2017). Structural variability and communicative complexity in acoustic communication. *Animal Behaviour*.

Fournet, M. E., Szabo, A., & Mellinger, D. K. (2015). Repertoire and classification of non-song calls in Southeast Alaskan humpback whales (Megaptera novaeangliae). *The Journal of the Acoustical Society of America*, *137*(1), 1–10. https://doi.org/10.1121/1.4904504

Freeberg, T. M., Dunbar, R. I. M., & Ord, T. J. (2012). Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society London Biological Sciences*, *367*(1597), 1785–1801.

Frey, B., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Fuller, J. L. (2014). The vocal repertoire of adult male blue monkeys (Cercopithecus mitis stulmanni): A quantitative analysis of acoustic structure. *American Journal of Primatology*, *76*(3), 203–216. https://doi.org/10.1002/ajp.22223

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). *irr: Various coefficients of interrater reliability and agreement*.

Garland, E., Castellote, M., & Berchok, C. (2015). Beluga whale (Delphinapterus leucas) vocalizations and call classification from the eastern Beaufort Sea population. *The Journal of the Acoustical Society of America*, *137*(6), 3054–3067.

Greene, D., Cunningham, P., & Mayer, R. (2008). Unsupervised learning and clustering. In M. Cord & P. Cunningham (Eds.), *Machine Learning Techniques for Multimedia*. Springer.

https://doi.org/10.1007/978-3-540-75171-7_3

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23.

Hammerschmidt, K., & Fischer, J. (1998). The vocal repertoire of Barbary macaques: A quantitative analysis of a graded signal system. *Ethology*, *104*(3), 203–216.

Hammerschmidt, K., & Fischer, J. (2019). Baboon vocal repertoires and the evolution of primate vocal diversity. *Journal of Human Evolution*, *126*, 1–13. https://doi.org/10.1016/j.jhevol.2018.10.010

Hedwig, D., Verahrami, A. K., & Wrege, P. H. (2019). Acoustic structure of forest elephant rumbles: A test of the ambiguity reduction hypothesis. *Animal Cognition*, *22*(6), 1115–1128. https://doi.org/10.1007/s10071-019-01304-y

Huijser, L. A. E., Estrade, V., Webster, I., Mouysset, L., Cadinouche, A., & Dulau-Drouot, V. (2020). Vocal repertoires and insights into social structure of sperm whales (*Physeter macrocephalus*) in Mauritius, southwestern Indian Ocean. *Marine Mammal Science*, *36*(2), 638–657. https://doi.org/10.1111/mms.12673

Janik, V. M. (1999). Pitfalls in the categorization of behaviour: A comparison of dolphin whistle classification methods. *Animal Behaviour*, *57*(1), 133–143.

Jones, A. E., ten Cate, C., & Bijleveld, C. C. J. H. (2001). The interobserver reliability of scoring sonagrams by eye: A study on methods, illustrated on zebra finch songs. *Animal Behaviour*, *62*(4), 791–801.

K. Lisa Yang Center for Conservation Bioacoustics. (2019). *Raven Pro: Interactive Sound Analysis Software (Version 1.6.1) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology.* Available from http://ravensoundsoftware.com

Kassambara, A., & Mundt, F. (2020). *factoextra: Extract and Visualize the Results of Multivariate Data*

Analyses. http://www.sthda.com/english/rpkgs/factoextra

Kershenbaum, A., Blumstein, D., Roch, M., Akçay, Ç., Backus, G., Bee, M., Bohn, K., Cao, Y., Carter, G., Cäsar, C., Coen, M., DeRuiter, S., Doyle, L., Edelman, S., Ferrer-i-Cancho, R., Freeberg, T., Garland, E., Gustison, M., Harley, H., … Zamora-Gutierrez, V. (2016). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biol Rev Camb Philos Soc*, *91*(1), 13–52.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lattenkamp, E. Z. (2019). The Vocal Repertoire of Pale Spear-Nosed Bats in a Social Roosting Context. *Frontiers in Ecology and Evolution*, *7*, 14.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365.

MacKinnon, J. (1977). A comparative ecology of Asian apes. *Primates*.

Madhusudhana, S. K., Chakraborty, B., & Latha, G. (2019). Humpback whale singing activity off the Goan coast in the Eastern Arabian Sea. *Bioacoustics*, *28*(4), 329–344. https://doi.org/10.1080/09524622.2018.1458248

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). *cluster: Cluster Analysis Basics and Extensions*. https://CRAN.R-project.org/package=cluster

Marler, P., & Hobbett, L. (1975). Individuality in a long-range vocalization of wild chimpanzees. *Zeitschrift Für Tierpsychologie*, *38*(1), 97–109.

Marler, P., Kavanaugh, J. F., & Cutting, J. E. (1975). On the origin of speech from animal sounds. In *On the origin of speech from animal sounds*. MIT Press.

Marshall, J., & Marshall, E. (1976). Gibbons and their territorial songs. *Science*, *193*(4249), 235–237.

McComb, K., & Semple, S. (2005). Coevolution of vocal communication and sociality in primates.

Biol Lett, *1*(4), 381–385.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, 1802.03426v3.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2021). *e1071: Misc functions of the Department of Statistics, probability*. https://CRAN.R-project.org/package=e1071

Mitra Setia, T., & van Schaik, C. P. (2007). The response of adult orang-utans to flanged male long calls: Inferences about their function. *Folia Primatologica*, *78*(4), 215–226.

Odom, K., Araya-Salas, M., Morano, J., Ligon, R., Leighton, G., Taff, C., Dalziell, A., Billings, A., Germain, R., Pardo, M., de Andrade, L., Hedwig, D., Keen, S., Shiu, Y., Charif, R., Webster, M., & Rice, A. (2021). Comparative bioacoustics: A roadmap for quantifying and comparing animal sounds across diverse taxa. *Biol Rev Camb Philos Soc*.

Ogle, D. H., Doll, J. C., Wheeler, P., & Dinno, A. (2022). *FSA: Fisheries Stock Analysis*. https://github.com/fishR-Core-Team/FSA

Sadhukhan, S., Hennelly, L., & Habib, B. (2019). Characterising the harmonic vocal repertoire of the Indian wolf (Canis lupus pallipes). *PLOS ONE*, *14*(10), e0216186. https://doi.org/10.1371/journal.pone.0216186

Sainburg, T., Thielk, M., & Gentner, T. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput Biol*, *16*(10), e1008228.

Schwing, R., Parsons, S., & Nelson, X. J. (2012). Vocal repertoire of the New Zealand kea parrot Nestor notabilis. *Current Zoology*.

Soltis, J., Alligood, C., Blowers, T., & Savage, A. (2012). The vocal repertoire of the Key Largo woodrat (Neotoma floridana smalli). *J Acoust Soc Am*, *132*(5), 3550–3558.

Spillmann, B., Dunkel, L. P., van Noordwijk, M. A., Amda, R. N. A., Lameira, A. R., Wich, S. A., & van Schaik, C. P. (2010). Acoustic properties of long calls given by flanged male orang-utans

(*Pongo pygmaeus wurmbii*) reflect both individual identity and context. *Ethology*, *116*(5), 385–395. https://doi.org/10.1111/j.1439-0310.2010.01744.x

Spillmann, B., Willems, E. P., van Noordwijk, M. A., Setia, T. M., & van Schaik, C. P. (2017). Confrontational assessment in the roving male promiscuity mating system of the Bornean orangutan. *Behavioral Ecology and Sociobiology*, *71*(1), 20.

Taylor, D., Dezecache, G., & Davila-Ross, M. (2021). Filling in the gaps: Acoustic gradation increases in the vocal ontogeny of chimpanzees (Pan troglodytes). *Am J Primatol*, *83*(5), e23249.

Thiebault, A., Charrier, I., Pistorius, P., & Aubin, T. (2019). At sea vocal repertoire of a foraging seabird. *Journal of Avian Biology*, *50*(5).

Turesson, H., Ribeiro, S., Pereira, D., Papa, J., & de Albuquerque, V. (2016). Machine learning algorithms for automatic classification of marmoset vocalizations. *PLoS One*, *11*(9), e0163041.

Vester, H., Hallerberg, S., Timme, M., & Hammerschmidt, K. (2017). Vocal repertoire of long-finned pilot whales (Globicephala melas) in northern Norway. *J Acoust Soc Am*, *141*(6), 4289.

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing Vocal Repertoires—Hard vs. Soft Classification Approaches. *PLOS ONE*, *10*(4), e0125785. https://doi.org/10.1371/journal.pone.0125785

Zhou, K., Fu, C., & Yang, S. (2014). Fuzziness parameter selection in fuzzy c-means: The perspective of cluster validation. *Science China Information Sciences*, *57*(11), 1–8.