Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015, **6,** 753–763 doi: 10.1111/2041-210X.12384

Assessing the performance of a semi-automated acoustic monitoring system for primates

Stefanie Heinicke[1]*, Ammie K. Kalan[1], Oliver J.J. Wagner[1], Roger Mundry[1], Hanna Lukashevich[2] and Hjalmar S. Kühl[1,3]

[1]*Max Planck Institute for Evolutionary Anthropology, Department of Primatology, Deutscher Platz 6, 04103 Leipzig, Germany;* [2]*Fraunhofer Institute for Digital Media Technology, Ehrenbergstr. 31, 98693 Ilmenau, Germany; and* [3]*German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany*

Summary

1. Passive acoustic monitoring is frequently used for marine mammals, and more recently it has also become popular for terrestrial species. Key advantages are the monitoring of (1) elusive species, (2) different taxa simultaneously, (3) large temporal and spatial scales, (4) with reduced human presence and (5) with considerable time saving for data processing. However, terrestrial sound environments can be highly complex; they are very challenging when trying to automatically detect and classify vocalizations because of low signal-to-noise ratios. Therefore, most studies have used manual preselection of high-quality sounds to achieve better classification rates. Consequently, most systems have never been validated under realistic field conditions.

2. In this study, we evaluated the performance of a passive acoustic monitoring system for four primate species in the highly noisy rain forest environment of the Taï National Park, Côte d'Ivoire. We collected 12 851 h of recordings with 20 autonomous recording units and did not preselect high-quality sounds manually. To automatically detect and classify the sounds of interest, we used an automated system built on speaker segmentation, support vector machines and Gaussian mixture models. One hundred and seventy-nine hours of recordings were used for validating the system.

3. The system performed well in detecting the loud calls of *Cercopithecus diana* and *Colobus polykomos* with a recall of 50% and 42%, respectively. Recall rates were lower for *Pan troglodytes* and *Procolobus badius*. To determine the presence of *Cercopithecus diana* and *Colobus polykomos* with a certainty of $P > 0{\cdot}999$, 2 and 7 h of recordings were needed, respectively. For these two species, our automated approach reflected the spatio-temporal distribution of vocalization events well. Despite the seemingly low precision, time investment for the manual removal of false positives in the system's output was only 3·5% compared to a human collecting and processing the primate vocalization data.

4. The proposed monitoring system is already fully applicable for *Cercopithecus diana* and *Colobus polykomos*, whereas it needs further improvement for the other species tested. In principle, it can be applied to any distinctive animal sound and can be implemented for the collection of acoustic data for behavioural, ecological and conservation studies.

Key-words: automated signal recognition, bioacoustics, chimpanzee drumming, Gaussian mixture model, primate vocalization, speaker segmentation, species identification algorithm, support vector machine

Introduction

Considering the extent of anthropogenic threats to wild animal populations, there is a growing need for effective approaches to biodiversity monitoring (Nichols & Williams 2006). Successful strategies to biodiversity conservation have yet to be developed and their effectiveness evaluated by long-term monitoring. The most widely applied monitoring methods for terrestrial animals are transect and plot sampling, as well as various types of capture–mark–recapture methods (Krebs 1998; Borchers, Buckland & Zucchini 2002; Buckland *et al.* 2004; Borchers & Efford 2008). These methods have been very successful in estimating species abundances and population trends at various spatial and temporal scales. In turn, the increasing availability of data recording sensors and processing algorithms in recent years has paved the way for automated data collection and processing, contributing to improved spatial and longitudinal survey coverage and enhanced comparability of results.

PASSIVE ACOUSTIC MONITORING

In contrast to active acoustic surveys in which humans observing animals record their vocalizations with hand-held

*Correspondence author. E-mail: stefanie_heinicke@eva.mpg.de devices, in passive acoustic monitoring, autonomous recording units (ARUs) are used to capture animal sounds. It is often implemented as an alternative, or complement, to existing monitoring methods to collect site-specific data on the distribution and diversity of species (Blumstein *et al.* 2011). It has been widely applied for monitoring marine mammals (Mellinger *et al.* 2007) and fish (Gannon 2008), since acoustic signals propagate efficiently in water and visual surveys are problematic or expensive (Marques *et al.* 2013). The identification of acoustic signals in the terrestrial environment is often more difficult because of the complexity of the soundscape. Still, passive acoustic monitoring has been applied to a range of taxa, including anurans (Pellet & Schmidt 2005), bats (MacSwiney *et al.* 2008; Walters *et al.* 2012), birds (Swiston & Mennill 2009; Digby *et al.* 2013), elephants (Payne, Thompson & Kramer 2003) and insects (Chesmore & Ohya 2004).

The main advantages of a passive acoustic monitoring approach are that it (1) enables detection of target animals in areas and situations where visual detection is greatly limited, for example dense rain forest or nocturnal animals (Marques *et al.* 2013), (2) allows for the study of different taxa simultaneously (Farnsworth & Russell 2007), (3) enables sampling over large temporal and spatial scales because ARU-based surveys can be upscaled at relatively low costs (Blumstein *et al.* 2011), (4) requires human presence only for the installation and maintenance of recording devices, thereby minimizing disturbances and costs (Mennill *et al.* 2012) and (5) leads to a reduction in time needed for data processing (Swiston & Mennill 2009).

To analyse the large amounts of data collected with passive recorders, techniques for automated signal recognition can be used. An automated approach provides an objective measure of detection because false positives and false negatives can be quantified so that bias introduced by humans collecting or analysing data can be avoided (Kühl & Burghardt 2013). The replicability and transparency of results is thereby increased, leading to a greater degree of standardization in data collection (Celis-Murillo, Deppe & Allen 2009).

Different algorithms have been applied for automatic signal recognition, which includes signal detection and classification, such as decision trees (Acevedo *et al.* 2009; Digby *et al.* 2013), spectrogram correlation (Mellinger & Clark 2000; Swiston & Mennill 2009) and support vector machines (Fagerlund 2007; Acevedo *et al.* 2009). Feature extraction for such approaches is most commonly based on Mel-frequency cepstral coefficients or descriptive spectral and temporal features (Blumstein *et al.* 2011). However, few studies have used an automated approach for both signal detection and classification (Swiston & Mennill 2009; Aide *et al.* 2013; Digby *et al.* 2013).

Results of automated analyses are probabilistic, that is they determine the most probable match. Consequently, there remains a degree of uncertainty due to false-positive and false-negative detections (Kühl & Burghardt 2013). With an automated approach, the likelihood of these two types of errors can be quantified when validating results. On the other hand, traditional survey methods are also subject to the risk of false-positive and false-negative detections (Miller *et al.* 2012), but this is usually not accounted for statistically, although it can considerably bias results (Guschanski *et al.* 2009).

Traditionally, automated systems achieved high recognition rates by using recordings actively collected by human observers (Anderson, Dave & Margoliash 1996) or by manually editing recorded sounds to obtain a higher signal-to-noise ratio (Herr, Klomp & Atkinson 1997). Recordings are often preselected by excluding particularly noisy time periods, such as the dawn chorus (Bardeli *et al.* 2010), or excluding sounds from non-target species (Niezrecki *et al.* 2003) (Table S1). A common problem is the use of the same data set for training and validating a system, because it fails to provide an independent validation and inflates the estimation of the system's accuracy. For an acoustic monitoring system to be convincing for field practitioners, its performance has to be evaluated using data representative of real field conditions which requires an independent validation data set. Recently, systems were developed that performed well in realistic environmental settings using continuous recordings, but these are often species specific (Digby *et al.* 2013).

THIS STUDY

The objective of this study was to evaluate the performance of a passive acoustic monitoring system to automatically detect and classify signals from continuous rain forest recordings. The tropical rain forest represents a challenging environment for such a system because a large number of taxa vocalize simultaneously in overlapping frequency ranges (Slabbekoorn 2004). We applied an automated procedure to identify five different acoustic signals from four diurnal primate species, including drumming and vocalizations of chimpanzees *Pan troglodytes ssp. verus* and vocalizations of Diana monkey *Cercopithecus diana*, King colobus *Colobus polykomos* and Western red colobus *Procolobus badius*. We then assessed recall and precision of different algorithm settings and how well the output reflected spatial and temporal patterns of vocalization events. This study thus provides an assessment of the applicability of an automated system for the monitoring of multiple species under realistic field conditions.

Materials and methods

STUDY AREA

The study site was located in the western section of the Taï National Park, Côte d'Ivoire (Fig. 1). It partially covered the territories of two chimpanzee communities, and eight diurnal monkey species inhabit the area (McGraw & Zuberbühler 2007). Given the difficulty of visually detecting primates in dense tropical forest, this was an ideal setting for the implementation of an acoustic monitoring system. The soundscape in Taï Forest is very complex, featuring a wide variety of biogenic sounds including those of birds and insects, anthropogenic sounds (mainly from airplanes) and geophysical ambient noise from wind, rain and thunder.

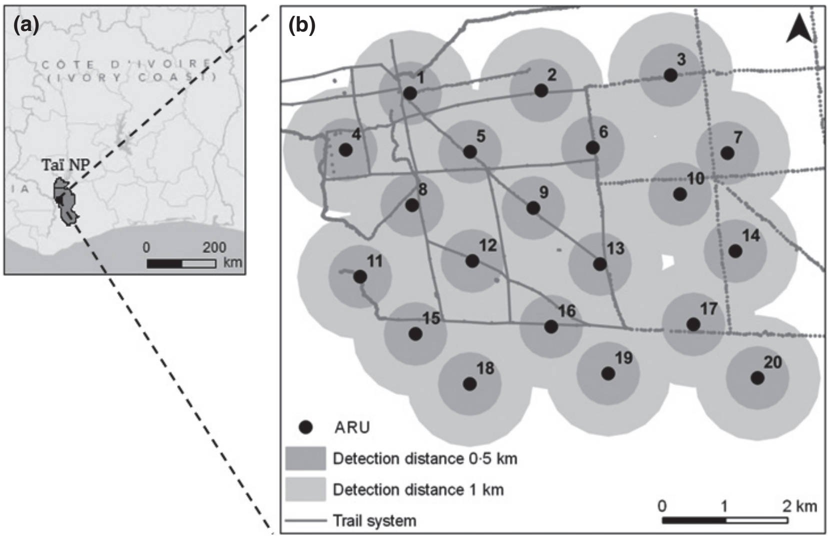

Fig. 1. Location of (a) the study area in the Taï National Park, Côte d'Ivoire and (b) the placement of autonomous recording units (ARU). Grey circles represent the area covered by the system when a sound is detected within a distance of 0·5 km (dark grey) or 1 km (light grey).

DATA COLLECTION

The acoustic data were collected between November 2011 and May 2012 using 20 ARUs (Songmeter SM2, Wildlife Acoustics Inc., Concord, MA, USA). ARUs were systematically placed in the field across an area of 35 km^2 and were capable of detecting sounds within a total area of ca. 45 km^2 (Fig. 1b). The stereo recordings were made with microphones covered by windshields. Prior to field deployment, microphones were calibrated in a sound laboratory to ensure that they recorded with equal sensitivity. Data were digitized (16 kHz sampling rate, 16 bits per second) and saved onto 32-gigabyte SD memory cards in uncompressed wave format. The ARUs recorded from 6 am to 6 pm, always on the full hour for 30 min, over a period of seven consecutive months. In total, 12 889 h of recordings were obtained.

ACOUSTIC SIGNALS

We selected several acoustic signals to determine how well the automated system performed for different signal types (Table 1, Fig. S1, Fig. S2). First, the drumming of chimpanzees was used, which is produced when chimpanzees hit different substrates, usually the buttress roots of trees, with their hands or feet (Arcadi, Robert & Boesch 1998). This is a chimpanzee-specific signal as no other species in the region produces similar sounds. Secondly, the pant-hoot, the long-distance call of chimpanzees, was targeted. It consists of hoots and screams and is often accompanied by drumming (Crockford & Boesch 2005). The other targeted signals included the loud call of the male Diana monkey, a low-frequency, uniform series of vocalizations (Zuberbühler, Noë & Seyfarth 1997); the loud call of the male King colobus, characterized by a series of high-intensity roars (Schel, Tranquilli & Zuberbühler 2009); and the very short contact call '*nyow*' by the Western red colobus (Struhsaker 2010). While the two chimpanzee sounds occur as clumped or overlapping but isolated events, the monkey calls are uttered repeatedly as part of long, stereotyped vocalization sequences.

DESIGN OF THE AUTOMATED SYSTEM FOR SIGNAL DETECTION AND CLASSIFICATION

The automated system was implemented using Fraunhofer IDMT feature extraction and pattern recognition methods. It was trained using a data set of 72 h of annotated recordings, which were not used in the subsequent validation of the system. We aimed for a similar total duration of training calls for each signal type. Due to the varying durations of target calls, a different number of training calls was used for each signal type (chimpanzee drumming: 351, chimpanzee vocalization: 228, Diana monkey: 223, King colobus: 61, red colobus: 158). We also included a background class with signals that were not of interest such as insects, birds, other monkey species, rain and thunder. More details on the training and testing are provided in the Supporting Information. The automated system was implemented in MATLAB (MathWorks 2011) and encompassed three main stages: preprocessing, signal detection and classification, and postprocessing (Fig. 2).

Table 1. Acoustic signals targeted in this study with the approximate fundamental frequency, duration and detection range. The detection range for chimpanzee signals was determined using ARU cross-referencing while following habituated chimpanzees in the Taï Forest. Those of the three monkey species are approximations (A. Kalan, unpublished data)

Species	Acoustic signal	Frequency (Hz)	Duration (s)	Detection range (km)
Chimpanzee	Drumming	<20[*]	<3·5[†]	<1
	Hoot	200–700[*]	0·17–1·2[*]	0·23
	Scream	800–2000[*]	0·5–1[*]	0·5
Diana monkey	Loud call	50–70[‡]	<5[‡]	0·75
King colobus	Loud call	25–30[§]	<10[§]	<1·5
Red colobus	Contact call	<1000[¶]	<0·5[¶]	0·5

[*] Crockford & Boesch (2005).
[†] Arcadi, Robert & Boesch (1998).
[‡] Zuberbühler, Noë & Seyfarth (1997).
[§] Schel, Tranquilli & Zuberbühler (2009).
[¶] Struhsaker (2010).

Preprocessing

Quality scan. The 30-min files were first scanned for hardware and data transfer errors to remove empty and duplicated files. 12 851 h of recordings remained. In some files, one of the two channels contained a high degree of noise either because of heavy rain or due to microphone damage caused by insects. In files where the power spectrum correlated

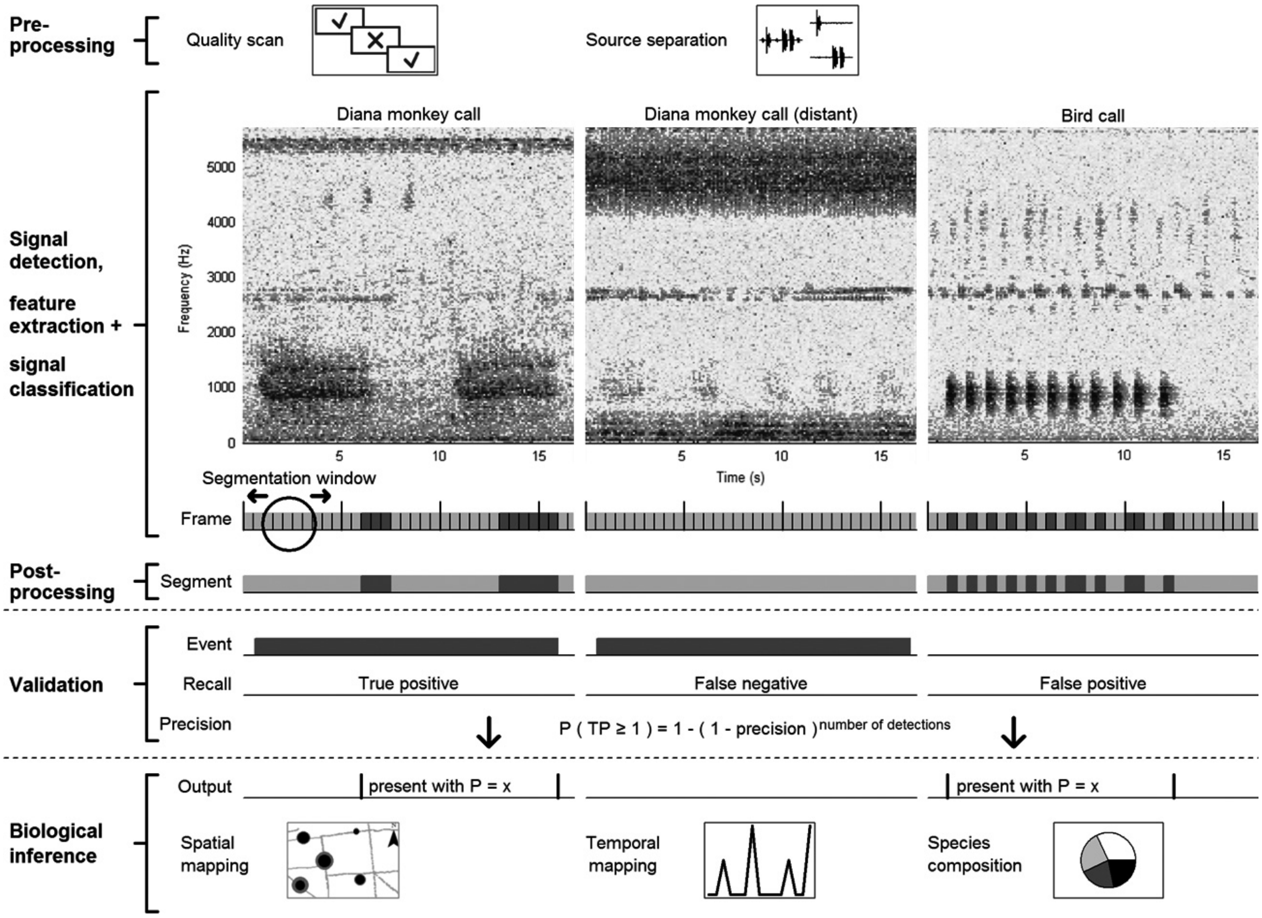

Fig. 2. Schematic representation of the three main stages of the automated system, followed by the validation of the output and biological inference for three exemplary signals. The first spectrogram shows two consecutive Diana monkey calls that were detected by the system (true positive). The second spectrogram shows distant Diana monkey calls that were not detected by the system (false negative), and the third example is a bird call that was erroneously classified as Diana monkey (false positive). Classified frames were merged into segments as part of the automated postprocessing. The system output (Segment) was then compared to the events detected during manual validation. After having determined the precision, the system's output can be interpreted and used for different applications.

with the power spectrum of noise and rain, only the channel with the high-quality recording was used for further analysis (4456 out of 25 701 files).

Source separation. To improve the detection of chimpanzee signals, chimpanzee drummings and vocalizations were separated from the original sound using non-negative matrix factorization (Schmidt, Larsen & Hsiao 2007) with 32 basis vectors and 50 iterations. Here the original sound was decomposed into the so called dictionary, which was constructed from training calls, and the code, a weighted matrix that determined for which frequency the amplitude was increased. Thereby, the target signal was amplified and the background noise reduced. This resulted in two separated sounds: chimpanzee drummings and chimpanzee vocalizations. The original recordings were used to identify the targeted monkey vocalizations. Since chimpanzee drumming is a low-frequency sound, we also applied a low-pass filter (cut-off frequency 400 Hz).

Signal detection and classification

Feature extraction. During the training of the automated system, we extracted 177 features, including Mel-frequency cepstral coefficients (Fagerlund 2007), loudness (Kim, Moreau & Sikora 2005), spectral crest factor (Allamanche *et al.* 2001), spectral flatness measure (Allamanche *et al.* 2001) and zero-crossing rate (Kim, Moreau & Sikora 2005). We aimed to create a generic system that could be easily applied to other animal sounds. Therefore, we did not preselect acoustic features manually. Instead, we used the feature selection algorithm 'Inertia Ratio Maximization using Feature Space Projection' (Peeters & Rodet 2003) with the modifications proposed by Essid (2005) to reduce the dimensionality of the feature space. This algorithm has been shown to reduce redundancy in the feature vector while increasing recognition rates (Peeters & Rodet 2003). The final dimensionality of the feature vector was: chimpanzee drumming: 32 features, Diana monkey: 32, King colobus: 8, red colobus: 32. For chimpanzee vocalization, the best classification results were achieved with the full set of 177 features.

Segmentation. To divide the continuous recordings into segments that contain only one signal type, thereby demarcating a call of interest, a segmentation algorithm was used. The continuous recordings were first divided into 2-min units. Based on the approach applied by Delacourt & Wellekens (2000) for speaker segmentation, several sliding windows were used to calculate spectral and timbre characteristics for each 30-ms frame. Adjacent windows were compared using the generalized likelihood ratio (GLR), and local minima of the GLR measure identified possible segment borders. Subsequently, segments were grouped according to signal type using the Bayesian information criterion (for details, see Delacourt & Wellekens 2000). The resulting segments therefore had varying durations.

Classification. Signals were classified using support vector machines (Vapnik 1998) with linear or polynomial kernels, or Gaussian mixture models (GMMs) (Table S2). The parameters of the GMMs were estimated using the expectation-maximization algorithm (Dempster, Laird & Rubin 1977). For each signal, the best performing classifier was used based on the area under the receiver operating characteristic curve (AUC) (Fawcett 2006) (Table S3). The result of the classification was a probability list for each 30-ms frame which showed with what probability a frame belonged to a certain class. When a frame did not match a signal class, it was assigned to 'background' (Fig. S3).

Decision matrix. The probability list for each frame was transformed into a binary decision matrix which consisted of 0s (does not belong to the class) and 1s (belongs to the class). This transformation was based on threshold values (Table S4) which determined for which class the frame would be assigned a 0 or 1. Low threshold values resulted in a higher number of false positives, while high threshold values increased the probability of false negatives.

Output rate. The output rate corresponded to the proportion of frames that was not classified as 'background'. This meant that a higher rate resulted in a higher number of total detections, consequently a higher false-positive rate but also a lower false-negative rate. Using receiver operating characteristic curves, four output rates – 2%, 5%, 10% and 20% – were chosen to evaluate which setting performed the best.

Postprocessing

Majority voting. Based on the segmentation, neighbouring frames were merged into segments. The class to which most frames were assigned was then assigned to the entire segment (hard majority voting, Fig. S3).

Segment limit. The segment limit was the maximum number of consecutive segments that can be classified as the same signal type. This was used to reflect the maximum duration of the targeted primate vocalizations and to exclude long continuous signals such as airplane noise or thunderstorms. Two segment limits were chosen, a 10-segment limit and a 20-segment limit, to test how much this setting influenced the performance of the automated system.

We ran the automated system eight times, each with a different setting (four output rates – 2%, 5%, 10% and 20% – with the 10-segment and the 20-segment limit), across the validation data set and then compared the performance of each setting.

VALIDATION OF SYSTEM OUTPUT

As the output of an automated system is probabilistic, a validation was necessary to determine which proportion of the detections were true positives and how many signals were missed by the system. As a validation data set, we randomly selected a total of 358 sound files, each of 30 min duration. The selection was balanced across ARUs (per ARU one file every seven days) and across three daytime blocks (6·00–9·00, 10·00–13·00 and 14·00–17·00). This ensured that the validation data set covered different types and degrees of background noise, as well as seasonal variation.

Recall

To determine how many signals were missed by the system, we listened to the validation data set and manually annotated all targeted acoustic signals using a customized program developed in MATLAB (MathWorks 2011). Primate vocalizations often occurred clumped in space and time so that vocalizations occurring together should be considered as part of the same vocalization event. Consequently, the annotated signals were grouped into events to allow for a comparison at a biologically meaningful scale. An event was defined as a group of consecutive signals of the same signal type when the time lag between consecutive signals was shorter than 1 min (see also Table S5). The number of events detected by the automated system in relation to the total number of events in the recording was calculated to determine the recall rate.

Precision

The segments detected by the automated system were verified by listening to the classified segments and determining whether those were true-positive, false-positive or misclassified detections. A misclassification implies that a sound of interest was detected but assigned to the wrong signal type. Interobserver reliability between the two annotators was checked and revealed 100% agreement on the classification of signals (N = 40 signals, all signal types included). The precision was calculated as the proportion of true-positive detections among the total number of detections registered by the automated system for each signal type. Accuracy was calculated as the sum of the duration of true-positive and true-negative segments divided by the total duration of the recording (Aide *et al.* 2013).

Algorithm settings

We also analysed whether different algorithm settings had an effect on recall and precision using Generalized Linear Mixed Models (GLMM) (Baayen 2008). For further details, see Supporting Information.

BIOLOGICAL INFERENCE

Determining presence

Based on the precision, the probability that at least one of the detections registered by the automated system in N detections was a true-positive detection was calculated with the equation: $P(\mathrm{TP} \geq 1) = 1 - (1 - \mathrm{precision})^{\mathrm{number\ of\ detections}}$.

We used a corresponding approach to determine for how long ARUs would have to record to confirm the presence of the targeted species. We calculated the number of hours needed to obtain at least one true-positive detection with a probability of 0·999. In addition, we used Spearman correlations to determine how well the automated system represented the spatial and temporal pattern of signal events.

Determining absence

Determining the true absence of a species required a manual check of system output. If none of the verified segments were true positives, the maximum occurrence probability was calculated for a range of efforts, that is, number of segments to be listened to manually. We accounted for false-negative detections by extending the duration of recordings to be verified manually, that is, the output of the automated system, by 1/recall rate. For example, if the recall is 50%, on average every second call is missed by the system. The 50% missed detections are accounted for by doubling the duration of recordings to be verified manually (1/0·5 = 2).

All analyses and graphs were done in R version 3.0.2 (R Core Team 2013). Maps were created in ArcGIS 10.0 (ESRI 2010).

Results

VALIDATION OF SYSTEM OUTPUT

Recall

In total, 577 events were annotated in the validation data set, most of which were red colobus contact calls (322), followed by chimpanzee drumming (103), Diana monkey (87), King colobus (34) and chimpanzee vocalization (31). The recall was

Table 2. Performance of the system for the five targeted signal types and eight algorithm settings. For settings with no detection, some parameters could not be determined (n/a)

Signal type and algorithm setting		Precision (%)	Recall (%)	Accuracy (%)	Detections by automated system (min)	Time needed for validating system output (% of manual time)	No. automated detections needed for ≥1 true-positive detection ($P > 0·999$)	No. recorded hours needed for ≥1 true-positive detection ($P > 0·999$)
Chimpanzee drum								
10-segment	2%	3·76	6·80	99·81	10·19	0·095	181	71
	5%	2·99	9·71	99·69	23·43	0·218	228	42
	10%	3·01	10·68	99·66	26·81	0·250	226	38
	20%	2·66	9·71	99·63	30·53	0·284	257	43
20-segment	2%	2·71	6·80	99·77	13·98	0·130	252	71
	5%	1·71	9·71	99·46	48·94	0·456	401	42
	10%	1·82	10·68	99·45	49·92	0·465	378	38
	20%	1·73	9·71	99·45	49·35	0·284	395	43
Chimpanzee vocal								
10-segment	2%	n/a	0	99·89	0	0	n/a	n/a
	5%	n/a	0	99·89	0	0	n/a	n/a
	10%	0	0	99·01	94·08	0·877	n/a	n/a
	20%	0	0	98·40	160·10	1·492	n/a	n/a
20-segment	2%	n/a	0	99·89	0	0	n/a	n/a
	5%	n/a	0	99·89	0	0	n/a	n/a
	10%	0·11	3·23	96·41	373·74	3·482	6516	45
	20%	0·02	3·23	96·63	349·74	3·259	>20 000	1030
Diana monkey								
10-segment	2%	7·34	32·18	99·07	41·29	0·385	91	5
	5%	6·38	36·78	98·91	60·47	0·563	105	4
	10%	5·83	33·33	98·95	51·42	0·479	115	4
	20%	6·02	36·78	98·88	61·42	0·572	112	4
20-segment	2%	7·41	42·53	98·73	99·74	0·929	90	3
	5%	5·62	49·43	97·58	233·71	2·178	120	2
	10%	5·41	47·13	97·57	231·32	2·155	125	2
	20%	5·51	49·43	97·58	232·50	2·166	122	2
King colobus								
10-segment	2%	2·92	32·35	99·38	50·20	0·468	234	11
	5%	4·22	32·35	99·56	30·46	0·284	161	11
	10%	5·84	35·29	99·49	40·90	0·381	115	7
	20%	5·70	41·18	99·39	54·33	0·506	118	7
20-segment	2%	1·73	35·29	98·70	123·93	1·155	396	9
	5%	3·98	38·24	99·32	59·85	0·558	171	7
	10%	4·39	35·29	99·26	65·36	0·609	154	7
	20%	4·97	41·18	99·27	66·32	0·618	136	7
Red colobus								
10-segment	2%	n/a	0	98·85	0	0	n/a	n/a
	5%	1·48	1·24	98·71	19·25	0·179	463	28
	10%	2·70	0·31	98·82	4·59	0·043	253	41
	20%	1·72	1·24	98·71	19·63	0·183	398	27
20-segment	2%	n/a	0	98·85	0	0	n/a	n/a
	5%	0·89	1·24	98·57	34·77	0·324	771	28
	10%	2·44	0·31	98·82	5·04	0·047	280	41
	20%	1·60	1·24	98·62	30·03	0·280	428	21

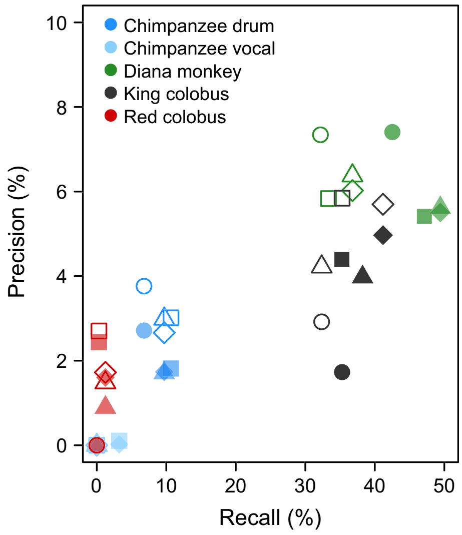

Fig. 3. Recall and precision for each signal type and each of the eight system settings (○ 10-segment limit 2% output rate, Δ 10-segment 5%, □ 10-segment 10%, ◇ 10-segment 20%, ● 20-segment 2%, ▲ 20-segment 5%, ■ 20-segment 10%, ♦ 20-segment 20%).

highest for Diana monkey with up to 50%, followed by King colobus. The recall was noticeably lower for the other signal types (Table 2).

For Diana monkey, events were detected at most ARUs and the variation in the number of detected events corresponded to the variation in the number of annotated events [Spearman's rank correlation (*r*s) range across algorithm settings: 0·66–0·83, mean = 0·78] (Fig. S4). The same applied to the number of events detected on different recording days (*r*s range: 0·39–0·75, mean = 0·66) (Fig. S4). The spatial (*r*s range: 0·44–0·89, mean = 0·62) and temporal (*r*s range: 0·42–0·72, mean = 0·53) patterns in King colobus signal events were similarly well reflected in the number of detected events. For the other signal types, the automated detections did not correspond well to the patterns of manually detected events, due to the lower recall (chimpanzee drumming: spatial: *r*s range: 0·56–0·59, mean = 0·57; temporal: *r*s range: 0·56–0·57, mean = 0·57; red colobus: spatial: *r*s range: 0·17–0·19, mean = 0·19; temporal: *r*s range: –0·10–(–0·04), mean = –0·08).

Precision

The rate of true-positive detections differed between algorithm settings and especially between signal types. Overall, less than 5% of detected segments were true-positive detections and an additional 6% of detections were assigned to the wrong signal type. The precision ranged between 0 and 7·5% and was highest for Diana monkey, followed by King colobus and chimpanzee drumming (Table 2). Comparing recall and precision between signal types, the system performed the best for Diana monkey and King colobus calls (Fig. 3).

Time saving

Listening to and verifying the 179 h validation data set took approximately 360 h. For the automated analysis of this data set, one computer with 8 CPUs needed approximately 17 h. Validating the system output required less than 3·5% of the time needed for listening to the entire recording (Table 2, Fig. 4).

Algorithm settings

In general, higher output rates resulted in a higher number of total detections. For the 2% output rate, there were no red colobus detections, and for the 2% and 5% output rate, there were no chimpanzee vocalization detections (Fig. S5). The 2% output rate had the lowest recall.

The evaluation of the segment limit showed that the 20-segment limit led to almost twice as many detections as the 10-segment limit (Fig. S5). At the same time, the precision was lower for the 20-segment limit because of the high proportion of

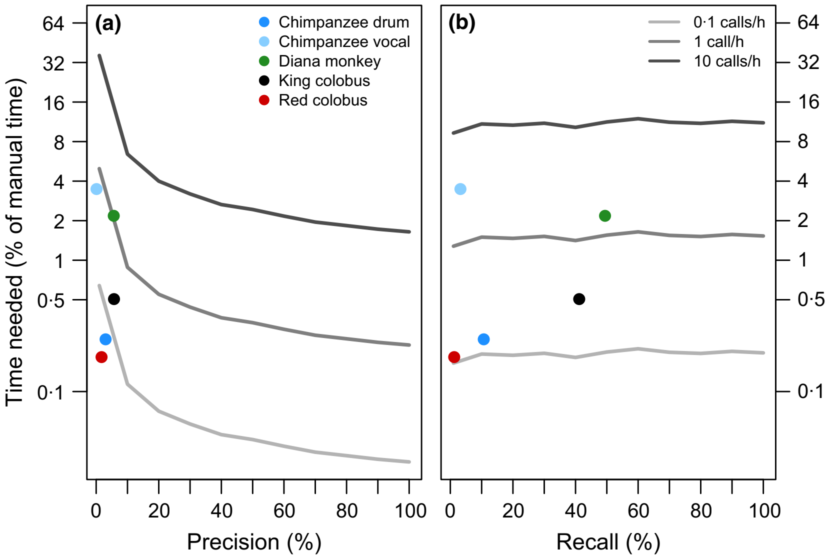

Fig. 4. Time needed for validating the output of an automated system to confirm primate presence, compared to the time needed for listening to the recordings until the first call event. The response was simulated (a) for different precision rates with a recall of 50% and (b) for different recall rates with a precision of 5% for three call densities. Time needed for validating system output for signal types targeted in this study (coloured points) is shown for the best performing algorithm setting; chimpanzee drum: recall = 10·68%, precision = 3·01%, density = 0·58 calls h^{-1}; chimpanzee vocal: *r* = 3·23%, p = 0·11%, *d* = 0·17 calls h^{-1}; Diana monkey: *r* = 49·43%, p = 5·62%, *d* = 0·49 calls h^{-1}; King colobus: *r* = 41·18%, p = 5·70%, *d* = 0·19 calls h^{-1}; red colobus: *r* = 1·24%, p = 1·72%, *d* = 1·80 calls h^{-1}.

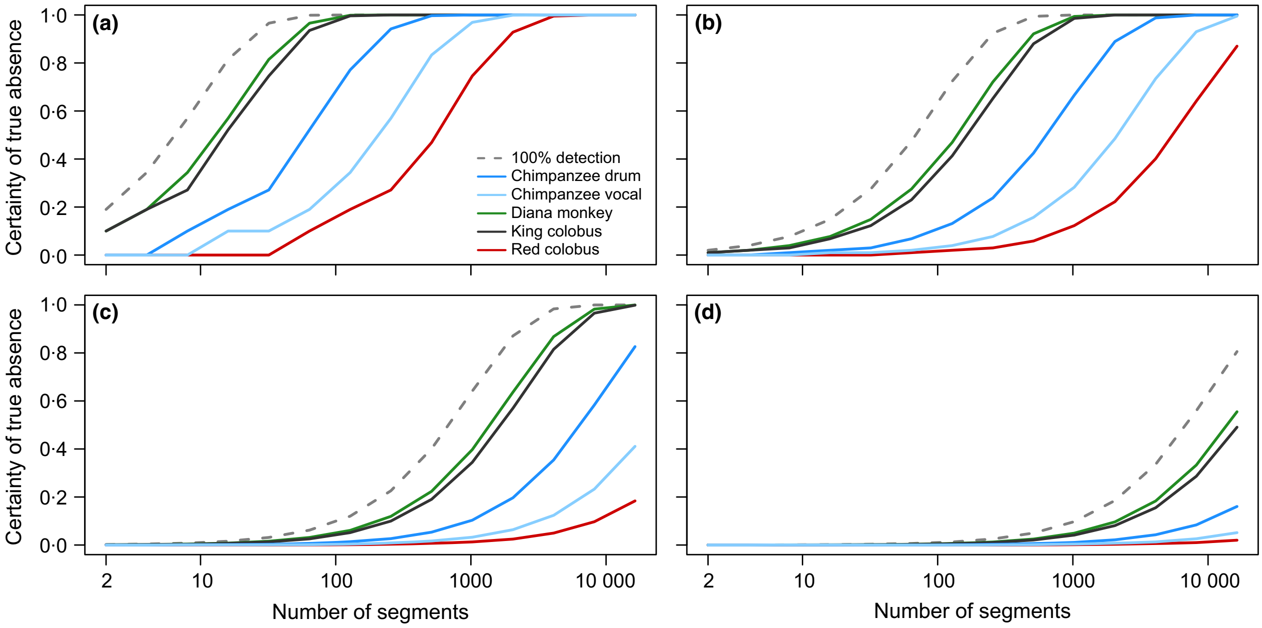

Fig. 5. Number of detected signals needed to be screened manually to ascertain the true absence of each species, assuming different occurrence probabilities of (a) 0·1, (b) 0·01, (c) 0·001 and (d) 0·0001. For each signal type, the algorithm setting with the highest recall was used: chimpanzee drumming 10·68% and vocalization 3·23%, Diana monkey 49·43%, King colobus 41·18% and red colobus 1·24%.

false-positive detections. These results highlight the trade-off between false-positive and false-negative detections. The GLMM analyses revealed that the segment limit only had a significant effect on the recall of Diana monkey calls (see Supporting Information for details).

Source separation improved recall rates for chimpanzee drummings by up to 4% and for chimpanzee vocalizations by about 2%. Precision increased only slightly.

BIOLOGICAL INFERENCE

Determining presence

Depending on the precision, between 90 and 6500 detections by the system were needed to ascertain that there was at least one true-positive detection with a probability of 0·999. Because of its high precision, the smallest number was required for Diana monkey (90) and King colobus (115) (Table 2). The recording time needed to confirm true presence with a probability of 0·999 was less than eight days for all signal types and shortest for Diana monkey (2 h), followed by King colobus (7 h) (Table 2).

Determining absence

The number of segments that had to be checked manually to ascertain the absence of a species was noticeably smaller for signal types with a high recall (Fig. 5). For example, when checking 100 detected segments for the Diana monkey call (mean segment length = 0·048 min) and all of them were false positives, it was concluded that the occurrence probability of a vocalizing individual at this site was less than 0·1 with a certainty >0·9 (Fig. 5a), and when checking 100 segments for chimpanzee drumming without detecting a true positive, the occurrence probability of a drumming individual at this site was less than 0·1 with a certainty >0·6. For a species with a very low occurrence probability (0·0001), a large number of segments would have to be checked to be fairly certain of its absence at a site (Fig. 5d).

Discussion

VALIDATION OF SYSTEM OUTPUT

Recall

The system performed especially well for detecting and classifying the long-distance calls of Diana monkeys and King colobus, but detected a considerably smaller proportion of chimpanzee drumming events. The recall was similar to other studies, such as Swiston & Mennill (2009), where the recall for different woodpecker calls ranged between 8% and 56%, or a recent study on kiwis with a recall of 40% (Digby *et al.* 2013).

The low recall for the red colobus call is likely a consequence of its call characteristics having a short duration and low sound intensity. Chimpanzee vocalizations, on the other hand, are very variable, even within individuals (Mitani, Gros-Louis & Macedonia 1996). This makes species detection by an automated system more difficult. Chimpanzee calls can reach frequencies of up to 10 kHz and these high frequencies attenuate quickly as the distance between the caller and the recording device increases. Moreover, our 16 kHz sampling rate may not have been sufficient to capture vocalizations of a high quality. More high-quality training calls might have been necessary to account for the high intraspecies variability.

Precision

The precision of the system was rather low. The high proportion of false-positive detections (Table S6) can be explained by the rich soundscape that typifies a tropical rain forest. Similar false-positive rates were produced by the automated approach of Swiston & Mennill (2009) whose study was located in the tropical forest of Costa Rica and in Florida. In contrast, other studies had a much higher precision, which might be due to the low level of background noise, a high signal-to-noise ratio for the targeted signal or the exclusion of vocalizations from other animals (Niezrecki *et al.* 2003; Bardeli *et al.* 2010; Digby *et al.* 2013). In many studies, it is not clear whether the same data set was used for training and validating a system, which would inflate precision rates. When implementing an automated system, a high false-positive rate results in more data to be stored and analysed, but it has no other detrimental effects on the performance of an automated system (Blumstein *et al.* 2011). Since continuous recordings contain relatively few and short target calls, the proportion of true-negative detections is generally very high. Here this led to high accuracy values across all settings (Table 2). Therefore, recall and precision might be more appropriate parameters for evaluating the performance of different algorithm settings.

Time saving

The large time saving of an automated system enables the analysis of very large data sets for which there would not be enough manpower. Beyond monitoring, this approach can also be used for the collection of large data sets of acoustic signals for ecological or behavioural studies.

Algorithm settings

Which setting is considered the most suitable depends on the type of signal targeted, the desired outcome of the study and the costs associated with false-positive and false-negative detections. If a maximization of detected events is desired, higher output rates are more suitable. Furthermore, additional biological information may be incorporated in the system design, such as separate maximum signal durations for each target signal.

APPLICATION IN THE FIELD

Having selected sound signals with different characteristics enabled us to show that signals used for long-distance communication are especially suitable target signals for a passive acoustic monitoring system. They are audible over large distances, due to a high sound intensity and efficient signal transduction. Ideally, the targeted vocalization should consist of a repeatedly uttered sound as part of a sequence, as this repetition greatly increases the chance of detection. The signal should also be easily distinguishable from background noise and other target sounds. Signals that are relatively variable within and between individuals are more difficult to be detected.

An acoustic bio-monitoring system allows for the collection of presence and absence data of one or more species. The presence data can be used to estimate the area occupied by a species or the species richness at a site by using occupancy modelling (Kalan *et al.* 2015). Thereby, imperfect detections of species are incorporated in the estimate.

Another area of application is to cross-reference the output of a fully automated system with a well-established method, such as transect sampling, and adjust the automated output accordingly. The automated system could be used as a predictor for the distribution of species which would facilitate modelling the distribution of species on large spatial scales.

An automated system can also serve as a key element for real-time monitoring within the framework of an early warning system. Monitoring could be conducted almost in real-time with continuous data collection and analysis. Acoustic monitoring has, proven to be useful in detecting disturbances occurring on a short-term basis, for example the effects of oil exploration on forest elephants (Wrege *et al.* 2010).

Conclusion

We have shown that passive acoustics can be used to monitor terrestrial animals in a high noise environment. Accepting certain limitations at its current stage of development, this system can be directly applied in the field to monitor the distribution of Diana monkeys and King colobus. Due to its generic design, it could be trained to recognize other animal sounds. This approach enables a timely analysis of large data sets and provides repeatable results. It can be used to collect data for different applications, such as monitoring for conservation, ecological and behavioural studies.

Acknowledgements

This study was part of the SAISBECO project, a collaboration between the Fraunhofer Institute for Digital Media Technology (Ilmenau), the Fraunhofer Institute for Integrated Circuits (Erlangen) and the Max Planck Institute for Evolutionary Anthropology (Leipzig), funded within the framework of the Pact for Research and Innovation between the Max Planck Society and the Fraunhofer-Gesellschaft. We thank the Ministère de la Recherche Scientifique, the Ministère de l'Environnement et des Eaux et Forêts of Côte d'Ivoire and the Office Ivorien des Parcs et Reserves for permissions to work in the country and Taï National Park. We gratefully acknowledge the Centre Suisse de Recherche Scientifique, Taï Chimpanzee Project, and the Wild Chimpanzee Foundation, especially Christophe Boesch and Roman Wittig for logistical support. We also thank Karin Bahrke for helping to process the data and the Robert Bosch Foundation. Special thanks to the reviewers for their helpful comments that improved this manuscript.

Data accessibility

The verified output of the automated system for the validation data set was uploaded as Supporting Information. The acoustic raw data can be accessed via the IUCN SSC A.P.E.S data base (http://apesportal.eva.mpg.de).

References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009) Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, **4**, 206–214.

Aide, T.M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013) Real-time bioacoustics monitoring and automated species identification. *PeerJ*, **1**, e103.

Allamanche, E., Herre, J., Hellmuth, O., Fröba, B., Kastner, T. & Cremer, M. (2001) Content-based identification of audio material using MPEG-7 low level description. *Proceedings of the 2nd International Symposium on Music Information Retrieval*, 197–204.

Anderson, S.E., Dave, A.S. & Margoliash, D. (1996) Template-based automatic recognition of birdsong syllables from continuous recordings. *Journal of the Acoustical Society of America*, **100**, 1209–1219.

Arcadi, A.C., Robert, D. & Boesch, C. (1998) Buttress drumming by wild chimpanzees: temporal patterning, phrase integration into loud calls, and preliminary evidence for individual distinctiveness. *Primates*, **39**, 505–518.

Baayen, R.H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.H. & Frommolt, K.H. (2010) Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, **31**, 1524–1534.

Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G. *et al.* (2011) Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, **48**, 758–767.

Borchers, D.L., Buckland, S.T. & Zucchini, W. (2002) *Estimating Animal Abundance: Closed Populations*. Springer-Verlag, London.

Borchers, D.L. & Efford, M.G. (2008) Spatially explicit maximum likelihood methods for capture-recapture studies. *Biometrics*, **64**, 377–385.

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., Borchers, D.L. & Thomas, L. (2004) *Advanced Distance Sampling: Estimating Abundance of Biological Populations*. Oxford University Press, Oxford.

Celis-Murillo, A., Deppe, J.L. & Allen, M.F. (2009) Using soundscape recordings to estimate bird species abundance, richness, and composition. *Journal of Field Ornithology*, **80**, 64–78.

Chesmore, E.D. & Ohya, E. (2004) Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. *Bulletin of Entomological Research*, **94**, 319–330.

Crockford, C. & Boesch, C. (2005) Call combinations in wild chimpanzees. *Behaviour*, **142**, 397–421.

Delacourt, P. & Wellekens, C.J. (2000) DISTBIC: A speaker-based segmentation for audio data indexing. *Speech Communication*, **32**, 111–126.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, **39**, 1–38.

Digby, A., Towsey, M., Bell, B.D. & Teal, P.D. (2013) A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, **4**, 675–683.

ESRI (2010) *ArcGIS Desktop: Release 10*. Environmental Systems Research Institute, Redlands.

Essid, S. (2005) *Classification Automatique des Signaux Audio-Fréquences. Reconnaissance des Instruments des Musique*. PhD thesis, l'Université Pierre et Marie Curie, Paris.

Fagerlund, S. (2007) Bird species recognition using support vector machines. *Eurasip Journal on Advances in Signal Processing*, **2007**, 1–8.

Farnsworth, A. & Russell, R.W. (2007) Monitoring flight calls of migrating birds from an oil platform in the northern Gulf of Mexico. *Journal of Field Ornithology*, **78**, 279–289.

Fawcett, K. (2006) An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874.

Gannon, D.P. (2008) Passive acoustic techniques in fisheries science: a review and prospectus. *Transactions of the American Fisheries Society*, **137**, 638–656.

Guschanski, K., Vigilant, L., McNeilage, A., Gray, M., Kagoda, E. & Robbins, M.M. (2009) Counting elusive animals: comparing field and genetic census of the entire mountain gorilla population of Bwindi Impenetrable National Park, Uganda. *Biological Conservation*, **142**, 290–300.

Herr, A., Klomp, N.I. & Atkinson, J.S. (1997) Identification of bat echolocation calls using a decision tree classification system. *Complexity International*, **4**, 1–9.

Kalan, A.K., Mundry, R., Wagner, O.J.J., Heinicke, S., Boesch, C. & Kühl, H.S. (2015) Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecological Indicators*, **54**, 217–226.

Kim, H.-G., Moreau, N. & Sikora, T. (2005) *MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval*. John Wiley & Sons, Chichester.

Krebs, C. (1998) *Ecological Methodology*, 2nd edn. Addison-Wesley, Menlo Park.

Kühl, H.S. & Burghardt, T. (2013) Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in Ecology & Evolution*, **28**, 432–441.

MacSwiney, M.C., Clarke, F.M. & Racey, P.A. (2008) What you see is not what you get: the role of ultrasonic detectors in increasing inventory completeness in neotropical bat assemblages. *Journal of Applied Ecology*, **45**, 1364–1371.

Marques, T.A., Thomas, L., Martin, S.W., Mellinger, D.K., Ward, J.A., Moretti, D.J., Harris, D. & Tyack, P.L. (2013) Estimating animal population density using passive acoustics. *Biological Reviews*, **88**, 287–309.

MathWorks (2011) *MATLAB 7.13*. The MathWorks Inc., Natick.

McGraw, W.S. & Zuberbühler, K. (2007) The Monkeys of the Taï Forest: an introduction. *Monkeys of the Taï Forest: An African Primate Community* (eds W.S. McGraw, K. Zuberbühler & R. Noë), pp. 1–48. Cambridge University Press, Cambridge.

Mellinger, D.K. & Clark, C.W. (2000) Recognizing transient low-frequency whale sounds by spectrogram correlation. *Journal of the Acoustical Society of America*, **107**, 3518–3529.

Mellinger, D.K., Stafford, K.M., Moore, S.E., Dziak, R.P. & Matsumoto, H. (2007) An overview of fixed passive acoustic observation methods for Cetaceans. *Oceanography*, **20**, 36–45.

Mennill, D.J., Battiston, M., Wilson, D.R., Foote, J.R. & Doucet, S.M. (2012) Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. *Methods in Ecology and Evolution*, **3**, 704–712.

Miller, D.A.W., Weir, L.A., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Simons, T.R. (2012) Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications*, **22**, 1665–1674.

Mitani, J.C., Gros-Louis, J. & Macedonia, J.M. (1996) Selection for acoustic individuality within the vocal repertoire of Wild Chimpanzees. *International Journal of Primatology*, **17**, 569–583.

Nichols, J.D. & Williams, B.K. (2006) Monitoring for conservation. *Trends in Ecology & Evolution*, **21**, 668–673.

Niezrecki, C., Phillips, R., Meyer, M. & Beusse, D.O. (2003) Acoustic detection of manatee vocalizations. *Journal of the Acoustical Society of America*, **114**, 1640–1647.

Payne, K.B., Thompson, M. & Kramer, L. (2003) Elephant calling patterns as indicators of group size and composition: the basis for an acoustic monitoring system. *African Journal of Ecology*, **41**, 99–107.

Peeters, G. & Rodet, X. (2003) Hierarchical Gaussian tree with inertia ratio maximization for the classification of large musical instruments databases. *Proceedings of the 6th International Conference on Digital Audio Effects*, 1–6.

Pellet, J. & Schmidt, B.R. (2005) Monitoring distributions using call surveys: estimating site occupancy, detection probabilities and inferring absence. *Biological Conservation*, **123**, 27–35.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL http://www.R-project.org [accessed 8 February 2014]

Schel, A.M., Tranquilli, S. & Zuberbühler, K. (2009) The alarm call system of two species of Black-and-White Colobus Monkeys (*Colobus polykomos* and *Colobus guereza*). *Journal of Comparative Psychology*, **123**, 136–150.

Schmidt, M., Larsen, J. & Hsiao, F. (2007) Wind noise reduction using non-negative sparse coding. *IEEE Workshop on Machine Learning for Signal Processing*, 431–436.

Slabbekoorn, H. (2004) Habitat-dependent ambient noise: consistent spectral profiles in two African forest types. *Journal of the Acoustical Society of America*, **116**, 3727–3733.

Struhsaker, T.T. (2010) *The Red Colobus Monkeys: Variation in Demography, Behavior, and Ecology of Endangered Species*. pp. 18–44. Oxford University Press, Oxford.

Swiston, K.A. & Mennill, D.J. (2009) Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology*, **80**, 42–50.

Vapnik, V. (1998) *Statistical Learning Theory*. Wiley-Interscience, New York.

Walters, C.L., Freeman, R., Collen, A., Dietz, C., Brock Fenton, M., Jones, G. *et al.* (2012) A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology*, **49**, 1064–1074.

Wrege, P.H., Rowland, E.D., Thompson, B.G. & Batruch, N. (2010) Use of acoustic tools to reveal otherwise cryptic responses of forest elephants to oil exploration. *Conservation Biology*, **24**, 1578–1585.

Zuberbühler, K., Noë, R. & Seyfarth, R.M. (1997) Diana monkey long-distance calls: messages for conspecifics and predators. *Animal Behaviour*, **53**, 589–604.

Received 11 October 2014; accepted 24 March 2015
Handling Editor: Kate Jones

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Details on training of the automated system and on the evaluation of different algorithm settings.

Table S1. Comparison of automated and manual approaches to different stages of an acoustic monitoring system.

Table S2. Best performing classifier for each acoustic signal with the corresponding number of features that were used for the classification.

Table S3. Comparison of the performance of different classifiers using the area under the receiver operating characteristic curve.

Table S4. Threshold values that were used to transform the output of the classification process into a binary matrix.

Table S5. Number of annotated events and true positive detections using different event definitions.

Table S6. Confusion matrices.

Fig. S1. Spectrogram for the loud calls of a male Diana monkey and male King colobus.

Fig. S2. Spectrograms for red colobus contact call, and for chimpanzee drumming, scream and hoot.

Fig. S3. Exemplary representation of the process of signal classification.

Fig. S4. Spatial and temporal distribution of Diana monkey call events annotated manually in the validation dataset and detected by the automated system.

Fig. S5. Number of segments detected and classified for each of the eight algorithm settings.

Data S1. Output of the automated system for the validation dataset.