Supporting Information

TRAINING OF THE AUTOMATED SYSTEM

The automated system was trained using a dataset of 72 h of annotated recordings, which were not used in the subsequent validation of the system. The general approach was to test the detection and classification of each type of target signal against all other sounds, i.e., true positive versus false positive detections.

For each algorithm setting a simulation was run in three-fold cross-validation scenarios. The training dataset was divided such that two thirds were used for training and one third was used for testing. This was done to avoid the album effect (Kim, Williamson & Pilli 2006), i.e., that a system trained and tested on the same 30-min ARU recording would deliver very high classification rates, but fail to classify signals from other ARU recordings.

The training dataset was selected from the recordings of different ARUs, different dates and times of day. This was done to ensure that different types and degrees of background noise would be included in the training dataset. Using training calls from different ARUs also reduced the chance of pseudoreplication because the three monkey species have small home ranges (Diana monkey: 0.63 km^2, King colobus: 0.78 km^2, red colobus: 0.58 km^2). This meant that training calls coming from recordings from different ARUs would very likely be from different groups and consequently from different individuals. Concerning the chimpanzee vocalization, very often individuals pant-hoot at the same time and chorus together (Fedurek, Schel & Slocombe 2013). We took training calls from such chorusing bouts to ensure that the calls used for the training dataset came from multiple individuals.

For the classification different algorithms were tested, namely Support Vector Machines (SVM) (Vapnik 1998) and Gaussian Mixture Models (GMM) (Hastie, Tibshirani &

Friedman 2001), each with different settings. For SVM we tested linear kernel, polynomial kernel and radial basis function kernel. For GMM we tested different number of Gaussians, i.e., number of components. The performance of these classifiers was compared using the area under the receiver operating characteristic curve (AUC) (Vapnik 1998).

We used threshold values to transform the output of the classification into a decision matrix. The threshold value chosen for one signal type also influenced the results for other signal types. To account for this interdependency of the threshold values we simulated a large number of possible combinations of threshold values (25,700) to determine the best possible combination. The final decision on which threshold values should be chosen was based on trying to reduce the proportion of false positive detections, while optimizing for true positive detections. We used confusion matrices for the quantification of misclassifications and accordingly adjusted the threshold values upward or downward.

At the end of each simulation confusion matrices were used to compare the performance of different settings and fine tune the system. We also used receiver operating characteristic curves to evaluate the trade-off between true positive and false positive detections.

EVALUATING THE PERFORMANCE OF DIFFERENT ALGORITHM SETTINGS

Methods

To investigate whether the algorithm setting had an effect on the recall rate, a Generalized Linear Mixed Model (GLMM) was used (Baayen 2008). As a large proportion of events were either detected by all the settings or by none, the model was only run on those events where the responses differed between the algorithm settings (Diana monkey: n=21, King colobus n=11, red colobus: n=5, chimpanzee drumming:

n=4). For the signal type 'chimpanzee vocalization', no model was run because the sample size was too small (n=1). The response variable was whether an event was detected or not, thus a binomial error structure was assumed. The output rate, the segment limit setting and the interaction between them were included as fixed effects. The event ID was included as a random effect. For the Diana monkey and King colobus events we also included the random slope for segment setting (Schielzeth & Forstmeier 2009; Barr *et al.* 2013). The random slope for the output rate could not be included because we only had two data points for each event. Random slopes were not included for chimpanzee drumming and red colobus because the dataset was too small to fit such a model. A likelihood-ratio test was used to test the significance of the full model as compared to the null model (Forstmeier & Schielzeth 2011), which included only the random effect and random slope where applicable. When the full model was significantly better than the null model, reduced models were run to determine whether the interaction was significant and/or which algorithm settings significantly differed from one another (Dobson & Barnett 2008). The analyses were done separately for each signal type.

As the predictor variable 'output rate' had four levels, the comparison between the full and reduced model showed whether the variable had an overall effect. Only when the overall effect was significant we applied Wilcoxon-tests as a post-hoc pair-wise comparison between settings. This implied doing six tests on the same dataset. To control for multiple testing the P-value was adjusted using the false discovery rate which controls the proportion of falsely rejected null hypothesis (Benjamini & Hochberg 1995).

To test for differences between the different algorithm settings regarding their precision another GLMM was run. The response variable was whether the detection was a true positive or a false positive detection, thus we assumed a binomial error structure. Output rate, segment limit and the interaction between the two were included as fixed effects.

We included signal type as a random effect and incorporated random slopes for segment setting. The selection of the model followed the same procedure as described above.

For all analyses, significance was established at an alpha level of 0.05. Analyses were done in R version 3.0.2 (R Core Team 2013) and we fitted GLMMs using the function 'lmer' provided by the package 'lme4' (Bates *et al.* 2013).

Results

Recall. The algorithm settings had different effects on the number of signals detected of each signal type. For the Diana monkey the algorithm setting had a clear impact on the number of events detected (likelihood ratio test comparing the full model with output rate, segment limit and their interaction with the null model: χ^2=36.94, df=7, P<0.001). After the removal of the non-significant interaction between output rate and segment limit (χ^2=3.03, df=3, P=0.34), it appeared that the segment setting had a significant effect on the recall. More events were detected for the 20-segment limit than the 10-segment limit (GLMM: estimate=2.57±0.40, z=6.41, P<0.001). The output rate also had a significant effect (full vs. reduced model: χ^2=11.14, df=3, P=0.011). More events were detected for the 5% and 20% than for the 2% and 10% output rates. The Wilcoxon-test indicated significant differences between the 2% and the 5% output rate (exact Wilcoxon signed rank test: T^+=47.5, N=10, P=0.041), but when controlling for multiple testing the P-value was not significant anymore (P=0.12).

The algorithm setting had no obvious effect on the number of detections for the King colobus calls (full vs. null model: χ^2=4.04, df=7, P=0.77).

For chimpanzee drummings the algorithm setting had an effect on the number of detected events (full vs. null model: χ^2=38.56, df=7, P<0.001) but the standard error of the estimates were very large. Upon further investigation we concluded that this was

obviously due to complete separation (Field 2005) and the small sample size (n=4). No further reasonable conclusions could be drawn from this output.

For the red colobus calls the algorithm setting also had an effect on the number of event detections (χ^2=29.66, df=7, P<0.001). The interaction between the two predictor variables was not significant (χ^2=0, df=3, P=1) nor was the segment setting (GLMM: estimate=$-7.14\times10^{-5}\pm1.32$, z=0, P=1). The output rate had a significant effect (χ^2=29.66, df=3, P<0.001) with more events being detected with the 5% and the 20% output rates as compared to the 2% and 10% output rates. The Wilcoxon-test was not applied, because it can only reach significance with at least six samples.

Precision. The GLMM did not reveal a significant influence of the algorithm setting on the precision (likelihood ratio test comparing the full model with output rate, segment limit and their interaction with the null model: χ^2=5.33, df=7, P=0.62).

DIFFERENT EVENT DEFINITIONS

In order to assess whether the use of different time periods for the event definition might change the results of the event detections, the analyses were rerun for one algorithm setting (10% output rate 20-segment limit) using several event definitions ranging from 5 seconds to 30 minutes. When changing the event definition using different minimum time gaps between events, the proportion of detected events changed markedly (Table S3). When events were defined more broadly, a higher proportion of events were detected. Notably, the one-minute definition used in this study was found to be in the mid-range of the proportions of events detected. Consequently, the grouping of signals has to be adjusted to the study objectives in order to correspond to the scale desired. For example, for the purpose of determining the presence of a species at a particular site, one event might span several hours or an entire recording day, while studies focussing on small-scale variations would need shorter event durations.

References

Baayen, R.H. (2008) *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge University Press, Cambridge.

Barr, D.J., Levy, R., Scheepers, C. & Tily, H.J. (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language,* **68,** 255-278.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2013) *lme4: Linear mixed-effects models using Eigen and S4.* URL http://cran.r-project.org/web/packages/lme4 [accessed 28 January 2014]

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate - A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological,* **57,** 289-300.

Dobson, A.J. & Barnett, A.G. (2008) *An Introduction to Generalized Linear Models,* 3rd edn. Chapman & Hall/ CRC, Boca Raton.

Fedurek, P., Schel, A.M. & Slocombe, K.E. (2013) The acoustic structure of chimpanzee pant-hooting facilitates chorusing. *Behavioral Ecology and Sociobiology,* **67,** 1781-1789.

Field, A. (2005) *Discovering Statistics using SPSS,* 2nd edn. Sage Publications, London.

Forstmeier, W. & Schielzeth, H. (2011) Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology,* **65,** 47-55.

Hastie, T., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York.

Kim, Y.E., Williamson, D.S. & Pilli, S. (2006) Towards Quantifying the "Album Effect" in Artist Identification. *Proceedings of the 7th International Conference on Music Information Retrieval***,** 393-394.

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]

Schielzeth, H. & Forstmeier, W. (2009) Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology,* **20,** 416-420.

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

Table S1. Comparison of automated and manual approaches to different stages of an acoustic monitoring system. Any combination of automated and manual approaches can be implemented.

	Automated approach	Manual approach
Timing and location of recording	not pre-selected	specifically selected, e.g., to avoid dawn chorus or specific species
Collection of recordings	passive (stationary recording device)	active (human observer uses a portable, handheld recording device)
Pre-processing	automated, e.g., using source separation or filters	manually selecting recordings with high signal-to-noise ratio
Signal detection	automated detection using, e.g., a segmentation algorithm	manually cutting targeted sounds from continuous recordings
Signal classification	automated classification using, e.g., a Support Vector Machine or decision tree	experienced human observer listens to and classifies signals

Table S2. Best performing classifier (Support Vector Machine SVM or Gaussian mixture model GMM) for each acoustic signal with the corresponding number of features that were used for the classification. Classifier performance was evaluated based on the area under the receiver operating characteristic curve (AUC).

Acoustic signal	Classifier	AUC	Final dimen-sionality of feature vector	Number of features					
				Loudness in logarithmic scale	Loudness normalized to one	Mel-Frequency Cepstral Coefficient (MFCC)/ modulation over 1st MFCC value/ modulation over 2nd MFCC value	Spectral crest factor	Spectral flatness measure (SFM)/ modulation of 2nd SFM dimension/ 5th dimension / 8th dimension / 14th dimension/ 15th dimension	Zero-crossing rate (ZCR) / temporal ZCR modulation
Chimpanzee drum	SVM with linear kernel	0.904	32	0	4	2 / 3 / 8	2	1 / 0 / 1 / 2 / 2 / 4	0 / 3
Chimpanzee vocal	GMM with one component	0.685	177	12	12	16 / 16 / 16	16	16 / 12 / 12 / 12 / 12 / 12	1 / 12
Diana monkey	SVM with polynomial kernel	0.885	32	5	3	6 / 3 / 0	6	3 / 0 / 0 / 4 / 1 / 0	0 / 1
King colobus	SVM with linear kernel	0.969	8	2	2	0 / 1 / 1	0	0 / 0 / 0 / 0 / 0 / 2	0 / 0
Red colobus	SVM with linear kernel	0.694	32	5	6	7 / 2 / 1	4	1 / 0 / 1 / 1 / 1 / 0	1 / 2

Table S3. Comparison of the performance of different classifiers using the area under the receiver operating characteristic curve (AUC). The following classifiers were tested: Gaussian Mixture Model (GMM) with one component (GMM-1), GMM with three components (GMM-3), GMM with five components (GMM-5), Support Vector Machine (SVM) with linear kernel (LINSVM), SVM with polynomial kernel (POLYSVM), SVM with radial basis function kernel (RBFSVM). For all classifiers the dimensionality of the feature vector was reduced using the feature selection algorithm 'Inertia Ratio Maximization using Feature Space Projection'. AUC for best performing classifier marked bold.

	GMM-1	GMM 3	GMM 5	LINSVM	POLYSVM	RBFSVM
Chimpanzee drum	0.841	0.875	0.875	**0.904**	0.879	0.849
Chimpanzee vocal	**0.685**	0.673	0.647	0.604	0.634	0.647
Diana monkey	0.834	0.876	0.876	0.876	**0.885**	0.876
King colobus	0.879	0.888	0.923	**0.969**	0.961	0.961
Red colobus	0.519	0.558	0.524	**0.694**	0.575	0.609

Table S4. Threshold values that were used to transform the output of the classification process (probability list showing the probability that a 30-ms frame belonged to a certain class) into a binary matrix specifying whether the frame was classified as a certain class or not. For chimpanzee vocalization, high threshold values had to be chosen because of the high rate of false positive detections. The threshold values differed between algorithm settings. The output rate corresponded to the proportion of frames that was not classified as 'background'.

	Algorithm setting			
Acoustic signal	2% output rate	5% output rate	10% output rate	20% output rate
Chimpanzee drum	0.5	0.5	0.5	0.5
Chimpanzee vocal	1	1	0.9999999999	0.9999999999
Diana monkey	0.5	0.5	0.5	0.5
King colobus	0.5	0.5	0.5	0.5
Red colobus	0.2	0.2	0.5	0.2

Table S5. Number of annotated events and true positive detections using different event definitions for the algorithm setting with 10% output rate and 20-segment limit.

Acoustic signal	Event definition	Annotated events	True positive detections	
		number	number	percent
Chimpanzee drum	30 min	58	10	17.24
	5 min	78	10	12.82
	2 min	89	10	11.24
	1 min	103	11	10.68
	30 s	107	11	10.28
	20 s	111	11	9.91
	10 s	112	11	9.82
	5 s	112	11	9.82
Chimpanzee vocal	30 min	16	1	6.25
	5 min	20	1	5.00
	2 min	26	1	3.85
	1 min	31	1	3.23
	30 s	35	1	2.86
	20 s	35	1	2.86
	10 s	39	1	2.56
	5 s	46	1	2.17
Diana monkey	30 min	70	40	57.14
	5 min	82	42	51.22
	2 min	84	41	48.81
	1 min	87	41	47.13
	30 s	94	42	44.68
	20 s	112	48	42.86
	10 s	168	64	38.10
	5 s	280	108	38.57
King colobus	30 min	21	10	47.62
	5 min	25	9	36.00
	2 min	30	11	36.67
	1 min	34	12	35.29
	30 s	42	13	30.95
	20 s	45	14	31.11
	10 s	59	16	27.12
	5 s	70	17	24.29
Red colobus	30 min	122	2	1.64
	5 min	181	2	1.10
	2 min	255	1	0.39
	1 min	322	1	0.31
	30 s	387	1	0.26
	20 s	421	2	0.48
	10 s	476	2	0.42
	5 s	522	2	0.38

Table S6 A. Confusion matrices for the segments detected by the four output rates with the 10-segment limit.

		Assigned class				
	2% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimpanzee drum	14	0	1	4	0
	Chimpanzee vocal	0	0	16	12	0
	Diana monkey	2	0	110	40	0
	King colobus	0	0	11	40	0
	Red colobus	0	0	29	8	0
	Background	356	0	1331	1266	0
	5% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimp drum	21	0	1	2	1
	Chimp vocal	0	0	19	8	2
	Diana monkey	4	0	133	30	10
	King colobus	0	0	5	36	0
	Red colobus	4	0	36	8	6
	Background	674	0	1892	769	386
	10% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimp drum	22	6	2	6	0
	Chimp vocal	2	0	31	11	3
	Diana monkey	2	19	112	37	1
	King colobus	1	31	7	62	0
	Red colobus	9	7	46	6	3
	Background	694	1340	1722	939	104
	20% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimp drum	21	18	0	1	0
	Chimp vocal	1	0	23	8	2
	Diana monkey	6	33	126	24	14
	King colobus	0	55	13	71	0
	Red colobus	5	17	45	12	7
	Background	756	2197	1885	1130	383

Table S6 B. Confusion matrices for the segments detected by the four output rates with the 20-segment limit.

		Assigned class				
	2% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimpanzee drum	14	0	1	11	0
	Chimpanzee vocal	0	0	19	12	0
	Diana monkey	2	0	223	46	0
	King colobus	0	0	12	45	0
	Red colobus	0	0	33	8	0
	Background	500	0	2723	2478	0
	5% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimpanzee drum	21	0	1	2	2
	Chimpanzee vocal	0	0	25	8	2
	Diana monkey	4	0	291	36	10
	King colobus	0	0	14	55	0
	Red colobus	5	0	50	8	6
	Background	1198	0	4801	1274	652
	10% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimpanzee drum	22	18	5	7	0
	Chimpanzee vocal	2	6	37	11	3
	Diana monkey	2	26	279	37	1
	King colobus	1	33	14	62	0
	Red colobus	9	68	84	6	3
	Background	1176	5511	4734	1288	116
	20% output rate	Chimpanzee drum	Chimpanzee vocal	Diana monkey	King colobus	Red colobus
Actual class	Chimpanzee drum	21	26	0	1	0
	Chimpanzee vocal	1	1	27	8	2
	Diana monkey	6	34	282	24	14
	King colobus	0	57	16	71	0
	Red colobus	5	94	66	12	8
	Background	1178	4524	4730	1313	475

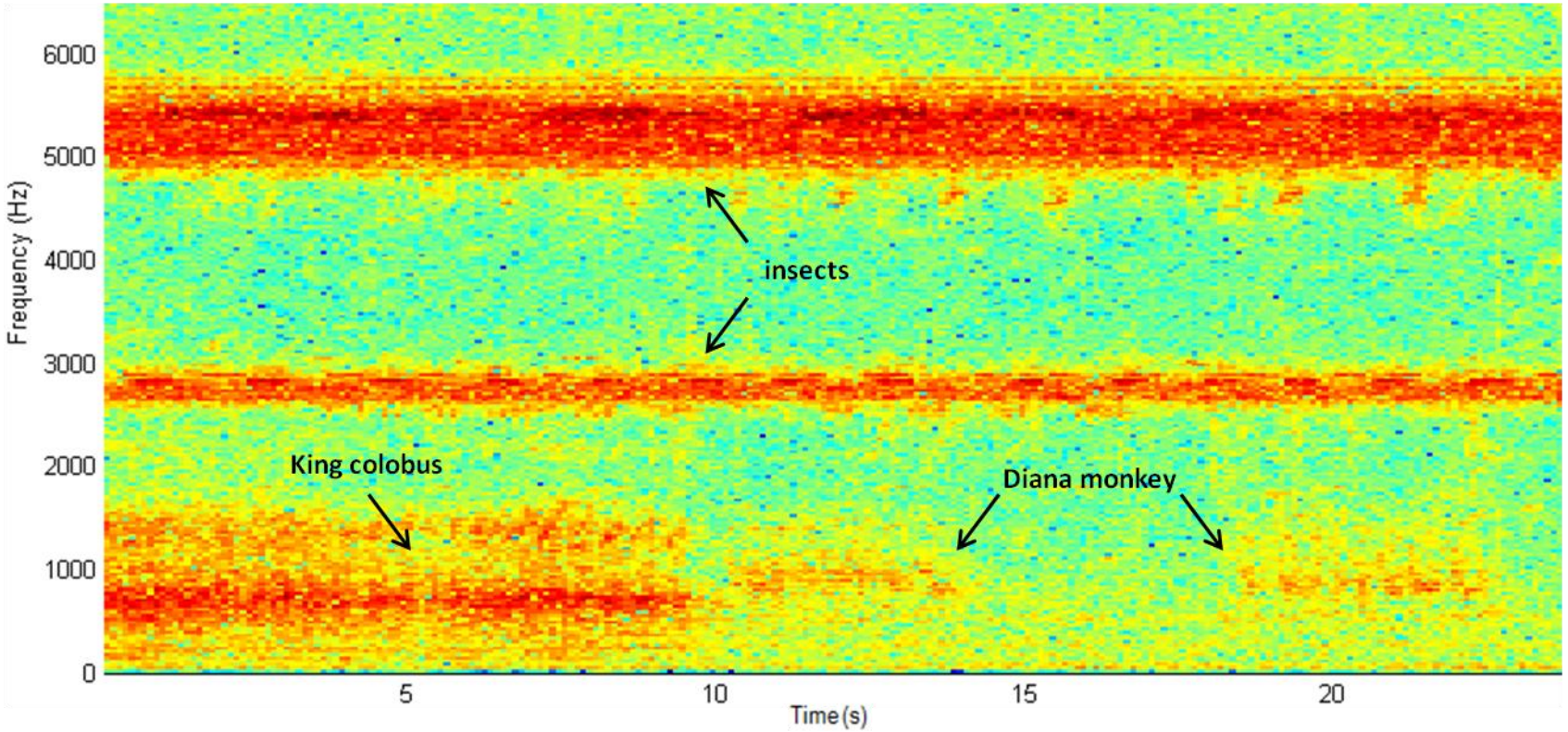

Fig. S1. Spectrogram for the loud calls of a male Diana monkey and male King colobus. The energy concentration across frequency and time is depicted by a colour scale which ranges from low energy concentration (blue) to high energy concentration (red). The background noise stems from insects.

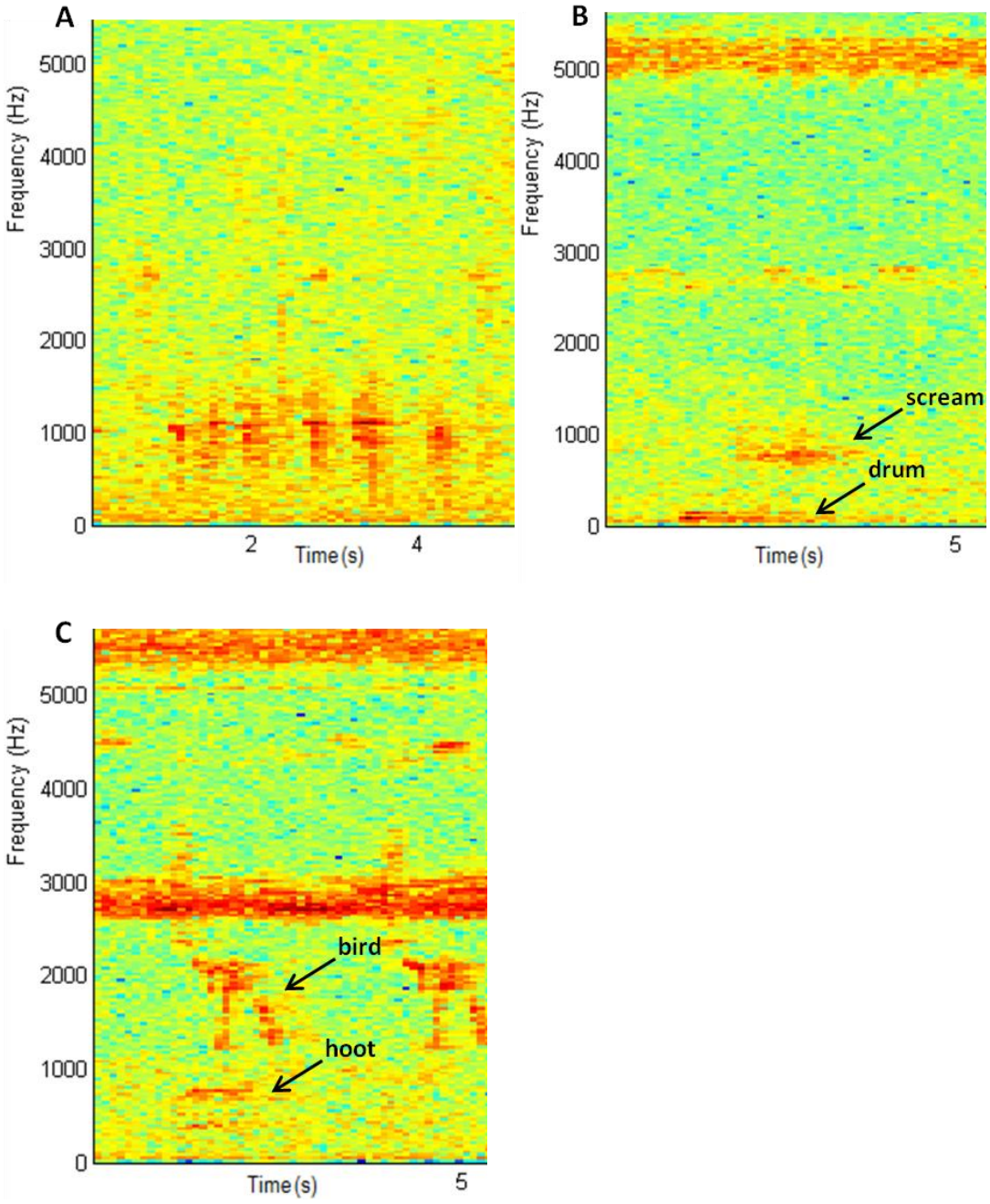

Fig. S2. Spectrograms for (A) red colobus contact call, (B) chimpanzee drumming and scream, and (C) chimpanzee hoot. The energy concentration across frequency and time is depicted by a colour scale which ranges from low energy concentration (blue) to high energy concentration (red).

Probability list (%)						
Frame number	1	2	3	4	5	6
Segment number	1	1	1	2	2	2
Chimpanzee drum	20	30	60	20	10	10
Chimpanzee vocal	0	0	0	0	0	0
Diana monkey	10	10	10	5	3	5
King colobus	40	85	90	85	30	40
Red colobus	0	0	0	0	0	0

Threshold mask (%)	
Chimpanzee drum	50
Chimpanzee vocal	100
Diana monkey	50
King colobus	50
Red colobus	20

Binary decision matrix						
Frame number	1	2	3	4	5	6
Segment number	1	1	1	2	2	2
Chimpanzee drum	0	0	1	0	0	0
Chimpanzee vocal	0	0	0	0	0	0
Diana monkey	0	0	0	0	0	0
King colobus	0	1	1	1	0	0
Red colobus	0	0	0	0	0	0
Background	1	0	0	0	1	1

Majority voting (%)		
Segment number	1	2
Chimpanzee drum	33	0
Chimpanzee vocal	0	0
Diana monkey	0	0
King colobus	67	33
Red colobus	0	0
Background	33	67

Final classification result		
Segment number	1	2
Classified as	King colobus	Background
with confidence of	67%	67%

Fig. S3. Exemplary representation of the process of signal classification (arrows depict the sequence in which steps were taken). The probability list is the first output of the classification procedure and shows with what probability each 30-ms frame belongs to a certain class. The threshold mask represents the threshold values to determine which class a frame is assigned to. This results in the binary decision matrix with 0 (does not belong to the class) and 1 (belongs to the class). When a frame cannot be assigned to a class it is determined to be 'Background'. Based on the segmentation, neighbouring frames were merged to segments. The class that the majority of frames belongs to was assigned to the entire segment.

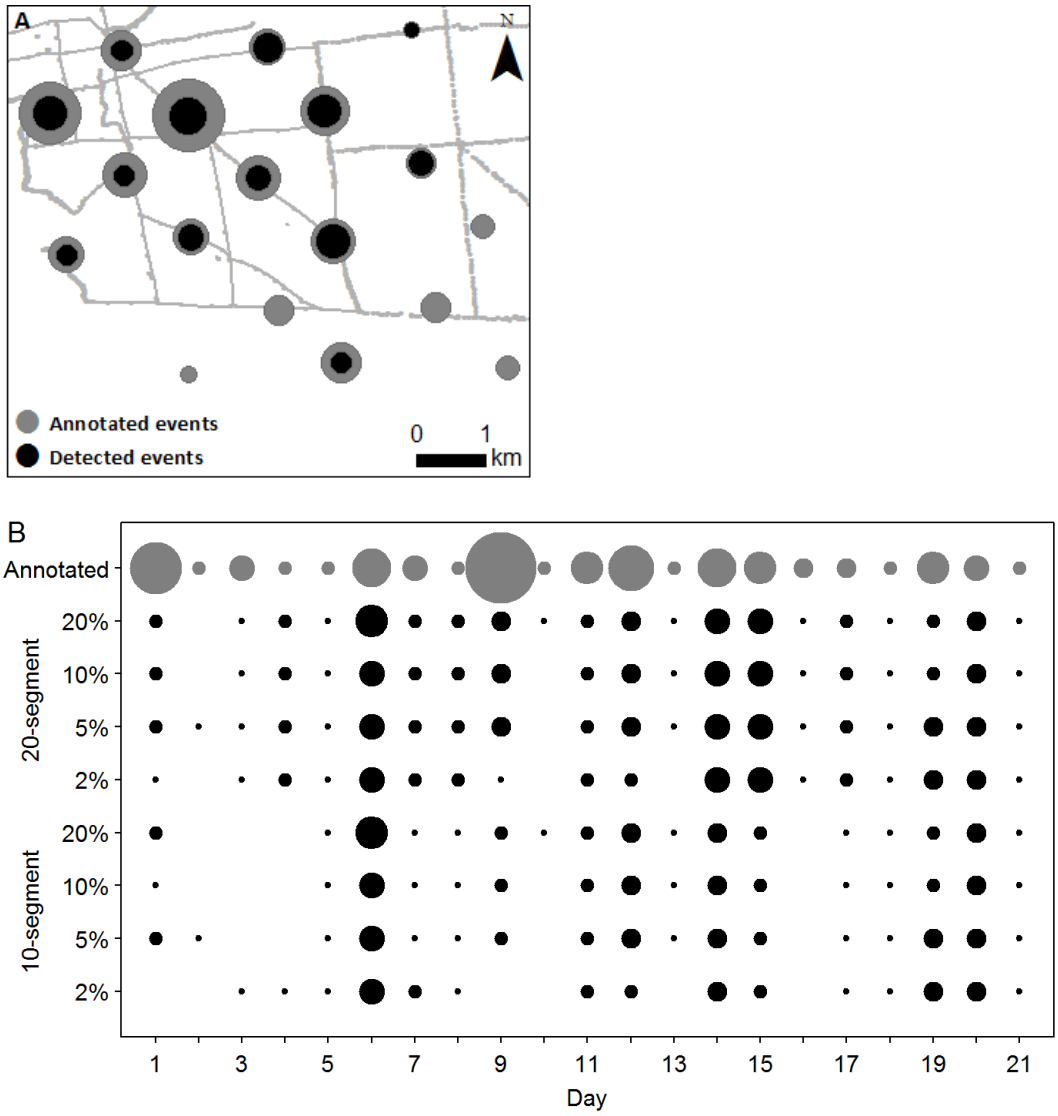

Fig. S4. Spatial (A) and temporal (B) distribution of Diana monkey call events annotated manually in the validation dataset (grey circles) and detected by the automated system (black circles). The relative area of the circles depicts the number of events annotated or detected. (A) The number of events at each ARU are shown for the algorithm setting 20% output rate 20-segment limit. (B) Number of events for each of the 21 recording days that were used in the validation dataset, shown for all eight algorithm settings.

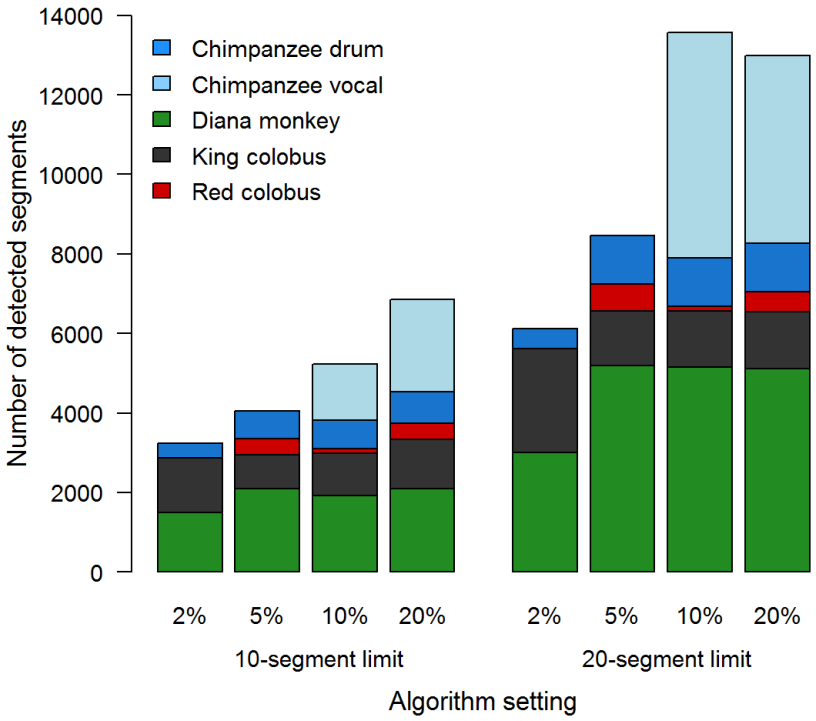

Fig. S5. Number of segments detected and classified for each of the eight algorithm settings (four output rates - 2%, 5%, 10% and 20% - with the 10-segment and the 20-segment limit).