Classification and automatic transcription of primate calls

Maarten Versteegh, Jeremy Kuhn, Gabriel Synnaeve, Lucie Ravaux, Emmanuel Chemla

Laboratoire de Sciences Cognitives et Psycholinguistique and Institut Jean-Nicod (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 29 rue d'Ulm, Pavillon jardin, Paris 75004, France
maarten.versteegh@gmail.com, kuhn@nyu.edu, gabriel.synnaeve@gmail.com, lucie.ravaux@gmail.com, chemla@ens.fr

Cristiane Cäsar

Instituto de Ciências da Natureza, Universidade Federal de Alfenas, Rua Gabriel Monteiro da Silva, 700, Centro, Alfenas, MG, Brazil
criscasar@gmail.com

James Fuller

Department of Ecology, Evolution, and Environmental Biology, Columbia University, 1200 Amsterdam Ave, New York, NY 10027, New York Consortium in Evolutionary Primatology (NYCEP), and Bronx Community College, City University of New York, 2155 University Avenue, Bronx, NY 10453
jlf2140@columbia.edu

Derek Murphy

School of Biological Sciences, University of Aberdeen, Tillydrone Avenue, Aberdeen, AB24 2TZ, United Kingdom
d.murphy@abdn.ac.uk

Anne Schel

Animal Ecology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
a.m.schel@uu.nl

Ewan Dunbar

Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 29 rue d'Ulm, Pavillon jardin, Paris 75004, France
emd@umd.edu

Abstract: This paper reports on a new automated and openly available tool for automatic acoustic analysis and transcription of primate calls, which takes raw field recordings and outputs call labels time-aligned with the audio. The system's output predicts a majority of the start times of calls accurately within 200 milliseconds. The tools do not require any manual acoustic analysis or selection of spectral features by the researcher.

PACS numbers: 2016 Acoustical Society of America
Date received date accepted etc etc

1. Introduction

A central topic in bioacoustics is the description of animal call repertoires, including what the calls are and how they are combined and used. However, traditional acoustic analysis of calls requires a significant amount of manual work, which means that only

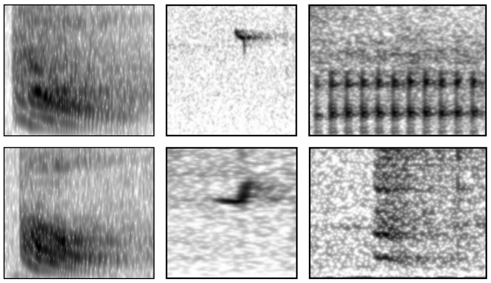

Fig. 1. Spectrograms of calls. Left: Blue monkey Hack (top) and Pyow (bottom) calls; center: Titi A (top) and B (bottom); right: Colobus Roar sequence (top) and Snort (bottom).

a fraction of the data collected in the field is actually used, and the majority of the otherwise useful data does not serve its role in answering scientific questions (Kobayasi and Riquimaroux, 2012). Recently, techniques from speech processing have been applied to animal vocalizations. The key advance they offer is to bypass a step where researchers extract preselected acoustic features, such as durations or peak frequencies. Standard speech processing tools represent signals using rich, general purpose spectral representations, with no hand selection of acoustic features. Previous analyses automatically classified isolated calls by call type, species, and caller using such representations (Mielke and Zuberbühler, 2013). Our system, in addition to labeling isolated calls, detects and labels calls in raw field recordings. We apply it to three primate species with acoustically diverse calls (see Figure 1): Blue monkeys (*Cercopithecus mitis*), Titi monkeys (*Callicebus nigrifrons*), and Colobus monkeys (*Colobus guereza*).

2. Data sets

Recordings of three species were taken from several field researchers for a total of 5.58 hours of audio. A trained primatologist marked the start and end times (calls typically do not overlap) and labeled each Blue monkey call as either Hack or Pyow (Arnold and Zuberbühler, 2006), Colobus calls as Roar or Snort (Marler, 1972), and Titi calls as A or B (Cäsar et al., 2012). Table 1 documents the length of the audio recordings for each data set, the percentage of that time taken up by calls, and the token count for each type of call. Estimated signal-to-noise ratios for these data sets (Vondrášek and Pollák, 2005) were low (between 0.5 and 5.3), typical of field recordings in primatology.

Species	Source (location)	Recorder	Microphone	Dur (% calls)	Types	N
Blue	Murphy (Budongo Reserve, Uganda)	Marantz PMD660	Sennheiser ME66-K6	1:56:45 (0.33%)	Hack Pyow	145 108
Blue	Fuller (Kakamega Forest, Kenya)	Marantz PMD660	Sennheiser ME67	0:59:15 (4.31%)	Hack Pyow	510 364
Titi	Cäsar (Serra do Caraça, Brazil)	Marantz PMD660	Sennheiser ME66-K6	0:11:58 (3.24%)	A B	125 539
Colobus	Schel (Budongo Reserve, Uganda)	Sony TCD D8	Sennheiser ME66-K6	2:27:02 (5.09%)	Roar Snort	739 141

Table 1. Total length of audio recordings, information about collection, percentage of the signal where calls were present, and count of each labeled call type.

Acoustic features were automatically extracted from the audio recordings using a standard speech feature extraction pipeline, adapted minimally. Since the recordings had non-zero mean and varying average amplitudes within each recording due to recording conditions and manual adjustment of the gain levels by field scientists, we removed the DC component with a notched high pass filter. We increased the ratio between calls and noise with a five-point temporal median filter (i.e., averaging windows of five consecutive samples in the time domain) followed by a two-dimensional three-point median filter pass in the spectral domain. The first enhances the ratio of the amplitude of the calls to noise, and the second flattens the spectrum for low transient noise passages and enhances the contrast with calls. We then estimated a noise signature based on the spectral components of the first half second of each audio file (which never included a call) and subtracted this noise signature from the rest of the audio stream. We calculated spectral representations of the signal using short-term Fourier transforms on overlapping windows of 25ms shifted by 10ms, and transformed the frequency components through a set of 40 filters evenly spaced on the Mel scale. This filter distribution is common in speech processing and is copied here for generality. Finally, each filter was mean-variance normalized independently.

3. Classification system

In this section we describe three experiments classifying isolated calls using the generic acoustic features just described. Each call was represented by concatenating the first 50 frames from the call onset (a $40 \times 50 = 2000$-dimensional vector, corresponding to 515 ms), capturing the full length of 84% of calls. In Experiment 1, we assessed the ability to classify call types within each species based on these representations. In Experiment 2, we assessed classification of species. In Experiment 3, we assessed the six-way labeling of species and call type required when all three species are pooled.

To predict the calls, we used a sparse radial basis function support vector machine (SVM) trained with block coordinate descent with squared hinge loss and L1 regularization. This is a standard statistical approach to classification problems that may not be amenable to classification using a linear model. Instead of computing the full Gram matrix of the kernel, we employ the Nyström approximation to significantly speed up the training time of our classifiers (Williams and Seeger, 2001). The approximation computes the eigendecomposition on a random small subset of the Gram matrix and scales the results up to the original number of dimensions (the number of samples). We achieved good results with a 500-component approximation. Experiments 2 and 3 involve more than two classes, so we employed a one-versus-rest strategy (training N individual binary classifiers, where N is the number of classes). Training was on 80% of the data, with evaluation on the remaining, unseen, 20%. Three hyperparameters (weight of the loss term, C, weight of the penalty term, λ, and kernel coefficient, γ) were optimized using the sequential model-based algorithm configuration (SMAC) technique (Hutter et al., 2011) by 5-fold cross-validation within the training set.

Table 2 shows the results of Experiments 1–3. We give precision (positive predictive value: among the calls the classifier gives label x, the fraction that are actually x and not false positives) and recall (sensitivity: among the calls that should be labelled x, the fraction that are labelled x and not false negatives), and F_1 ($2 \cdot \text{precision} \cdot \text{recall}/(\text{precision} + \text{recall})$). Classification was good, with average F_1 of between 0.91 and 0.99. Experiment 1 extends previous findings using different methodology and new species (Mielke and Zuberbühler, 2013). Experiments 1 and 3 were repeated with subsets of increasing sizes of the full (i.e., 80%) training set. Figure 2 shows the F_1 score on the test set as a function of the number of annotated calls given for training.

4. Automatic transcription system

In Experiment 4, we trained a call transcription system whose input is raw, unsegmented field recordings. It predicts call labels using a support vector machine and uses

Labels	Precision	Recall	F_1 score	Test Support
Expt 1				
Blue Hack	0.97	0.99	0.98	131
Blue Pyow	0.99	0.96	0.97	95
Average	0.98	0.98	0.98	
Colobus Roar	0.94	1.00	0.97	148
Colobus Snort	1.00	0.68	0.81	28
Average	0.95	0.95	0.94	
Titi A	0.89	0.68	0.77	25
Titi B	0.93	0.98	0.95	108
Average	0.92	0.92	0.92	
Expt 2				
Blue	0.99	0.98	0.98	226
Titi	0.99	0.98	0.98	176
Colobus	0.98	1.00	0.99	133
Average	0.99	0.99	0.99	
Expt 3				
Blue Hack	0.99	0.95	0.97	131
Blue Pyow	0.95	0.95	0.95	95
Colobus Roar	0.86	0.97	0.91	148
Colobus Snort	0.92	0.43	0.59	28
Titi A	0.85	0.68	0.76	25
Titi B	0.88	0.94	0.91	108
Average	0.91	0.91	0.91	

Table 2. Classification results for Experiments 1 (call type, within species), 2 (species only), and 3 (species and call type).

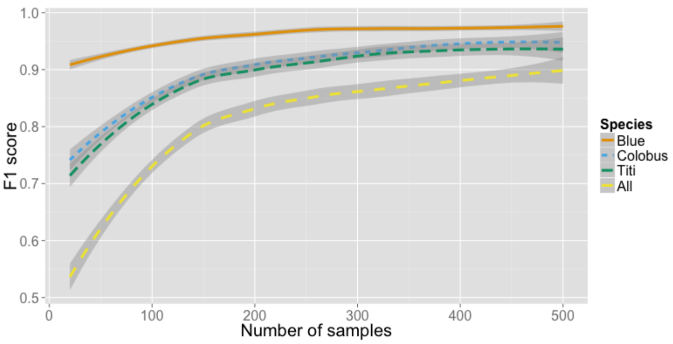

Fig. 2. (color online) Classification performance (y) by species, as a function of the number of annotated calls provided in the training set (x).

a conditional random field (CRF) to correct unlikely sequences.

The SVM was trained on annotated data to predict call labels from individual frames. Input features consisted of a concatenation of MFCC features (13 cepstral coefficients with first and second derivatives) with activations from a voice activity detection (VAD) system (Lee and Hasegawa-Johnson, 2007). The classifier was trained within species to predict one of the two call types or a third class indicating the absence of a call. The sequence of Platt-calibrated predictions of the SVM were used as input to a linear chain CRF. The CRF's predictions are also sequences of frame labels, but the CRF takes into account statistical dependencies between adjacent frames and smoothes the predictions in the time domain. The hyperparameters of the SVM and the CRF were optimized using SMAC. We evaluated on a 10% held out test set. The third label (absence of any call) is removed from the output for evaluation.

The system outputs call sequences, time-aligned with an audio file. We evaluate these transcriptions for the held-out test data. Considering the sequences of calls (not the alignment with the audio), we evaluate using word error rate (WER) and match error rate (MER), used in speech recognition (Morris et al., 2004). Results are in Table 3. The majority of calls are correctly identified. Most errors are deletions (missing calls) for Blue and Colobus monkeys and insertions (noise identified as calls) for Titis, perhaps because Titi calls are high frequency, similar to the noise.

Species	WER	MER	H	D	S	I	N
Blue	35.1%	32.1%	213	69	6	26	288
Colobus	34.4%	33.8%	106	47	4	3	157
Titi	32.9%	28.1%	68	8	6	14	82

Table 3. Evaluation of transcriber: word and match error rate (WER, MER), number of hits (H), deletions (D), substitutions (S), insertions (I) and number of calls (N).

To evaluate how well the predicted calls are time-aligned, we match each call in the gold transcription to the nearest predicted call whose onset and offset are within a 200ms tolerance of the real onset and offset, and count a gold call as having a true positive only if it has such a match, and that match is correctly labelled; otherwise, it counts as a false negative. Similarly, for each predicted call, we look for the nearest such match among the calls in the gold transcription, and count a false positive if there is no match or if the match is mislabelled. Since it is likely easier to accurately mark the onsets of calls than their offsets, both for our human annotator and for the transcription system, we also compute an alternative scoring in which only call onsets need to be matched within the 200ms tolerance. For both scorings, we compute precision, recall, and F_1, as shown in Table 4. The results show that call onsets are indeed much easier to match to the annotation than offsets, particularly for Colobus monkeys, where performance is relatively poor when offsets are required to be correctly marked.

5. Conclusions

General purpose acoustic features and voice activity detection techniques, as used in speech recognition, can automate the labeling of primate calls, both in isolation and in unannotated recordings, using data representative of field recordings. The system needs to be bootstrapped by a set of annotated examples. We showed that good isolated call labelling requires less than 200 labelled examples. It accurately transcribes around 90 percent of the frames in an audio file, vastly reducing the amount of manual work.

The results also imply that generic acoustic features, rather than specialized acoustic measurements taken manually by the researcher, can be used for detailed

	Call detection			Onset detection		
Species	Precision	Recall	F1	Precision	Recall	F1
Blue	0.76	0.65	0.70	0.85	0.72	0.78
Colobus	0.46	0.33	0.38	0.74	0.54	0.62
Titi	0.63	0.68	0.66	0.71	0.77	0.74

Table 4. Evaluation of predicted calls versus the nearest gold transcribed call with both its onset and offset (left) or just its onset (right) within 200 ms.

analysis. For example, there are competing descriptions of the call repertoires of certain species. Previous analyses have appealed to clustering analyses on hand-selected acoustic features as evidence (Fuller, 2014; Keenan et al., 2013). The results here validate an automated process of feature extraction that may be used as the input to these analyses. Both results allow much larger data sets from the field to be used than are currently being used for research and make it easier to create shared databases between researchers. Our tools can be downloaded at http://github.com/mwv/mcr.

Acknowledgments

This research received funding from the European Research Council under the EU's Seventh Framework Programme (FP/2007–2013), grant 324115-FRONTSEM (Schlenker) and ERC-2011-AdG-295810 BOOTPHON (Dupoux), and from the Agence Nationale pour la Recherche ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC.

References and links

Kate Arnold and Klaus Zuberbühler. The alarm-calling system of adult male putty-nosed monkeys, *Cercopithecus nictitans martini*. *Animal behaviour*, 72:643–653, 2006.

Cristiane Cäsar, Richard W Byrne, Robert J Young, and Klaus Zuberbühler. The alarm call system of wild black-fronted titi monkeys, *Callicebus nigrifrons*. *Behav Ecol Sociobiol*, 66:653–667, 2012.

James Lewis Fuller. The vocal repertoire of adult male blue monkeys (*Cercopithecus mitis stulmanni*): A quantitative analysis of acoustic structure. *American journal of primatology*, 76:203–216, 2014.

Jean-Pierre Gautier. Redrawn phylogeny of guenons based on their calls. *Bioacoustics*, 2:11–21, 1989.

F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proc. of LION-5*, page 507523, 2011.

Sumir Keenan, Alban Lemasson, and Klaus Zuberbhler. Graded or discrete? A quantitative analysis of Campbell's monkey alarm calls. *Animal Behaviour*, 85(1):109–118, 2013.

Kohta I Kobayasi and Hiroshi Riquimaroux. Classification of vocalizations in the Mongolian gerbil, *Meriones unguiculatus*. *The Journal of the Acoustical Society of America*, (2):1622–1631, 2012.

Bowon Lee and Mark Hasegawa-Johnson. Minimum mean-squared error a posteriori estimation of high variance vehicular noise. *Biennial on DSP for In-Vehicle and Mobile Systems*, 2007.

Peter Marler. Vocalizations of East African monkeys II. *Behaviour*, 42:175–197, 1972.

A. Mielke and K. Zuberbühler. A method for automated individual, species and call type recognition in free-ranging animals. *Animal Behavior*, 86(2):475–482, 2013.

M. Vondrášek and Petr Pollák. Methods for speech SNR estimation: Evaluation tool and analysis of vad dependency. *Radioengineering*, 14(1):6–11, 2005.

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, number EPFL-CONF-161322, pages 682–688, 2001.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *INTERSPEECH*, 2004.

Melvyn Hunt and Claude Lefebvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), pages 262–265, 1989.