Animal Behaviour 86 (2013) 475–482

Contents lists available at SciVerse ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Commentary

A method for automated individual, species and call type recognition in free-ranging animals

Alexander Mielke [a,b,*], Klaus Zuberbühler [a,c]

[a] School of Psychology & Neuroscience, University of St Andrews, St Andrews, U.K.
[b] Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[c] Department of Comparative Cognition, Institute of Biology, University of Neuchatel, Neuchatel, Switzerland

ARTICLE INFO

Article history:
Received 4 December 2012
Initial acceptance 11 January 2013
Final acceptance 8 April 2013
Available online 24 June 2013
MS. number: 12-00913R

Keywords:
artificial neural network
blue monkey
Cercopithecus mitis
mel frequency cepstral coefficient
voice recognition

The ability to identify individuals reliably is often a key prerequisite for animal behaviour studies in the wild. In primates, recognition of other group members can be based on individual differences in the voice, but these cues are typically too subtle for human observers. We applied a combined mechanism consisting of a call feature extraction (mel frequency cepstral coefficients) and pattern recognition algorithm (artificial neural networks) to investigate whether automated caller identification is possible in free-ranging primates. The mechanism was tested for its accuracy in recognizing species, call type and caller identity in a large population of free-ranging blue monkeys, *Cercopithecus mitis stuhlmanni*, in Budongo Forest, Uganda. Classification was highly accurate with 96% at the species, 98% at the call type and 73% at the caller level. It also outperformed conventional discriminant function analysis in the individual recognition task. We conclude that software based on this method will make a powerful tool for future animal behaviour research, as it allows for automatic, fast and objective classifications in different animal species.

For many questions in animal behaviour, individual recognition of subjects is essential. To this end, fieldworkers usually rely on individual differences in body shape, coloration and markings. Although voice and chemical cues are often also individually distinct, they are more difficult to observe directly (Tibbetts & Dale 2007). In most field studies, researchers thus go through a time- and resource-consuming learning process or need to rely on artificial markings, such as rings and radiotracking, which can be difficult to administer and are almost always invasive for the animal (Adi et al. 2010). While there has been a strong focus on identifying individuals visually, vocalizations have great potential to facilitate recognition, especially for species that are cryptic, arboreal or nocturnal (Bardeli et al. 2010).

In primates and other groups of animals that live in stable social groups, individual recognition is a well-documented, key aspect of social behaviour (Tibbetts & Dale 2007). Moreover, in species living in environments with restricted visual contact, individual recognition is often based on vocal communication. Accordingly, individual vocal signatures have been reported in a number of primate species (e.g. *Macaca mulatta*: Hammerschmidt et al. 2000; *Papio ursinus*: Fischer et al. 2001; *Pan troglodytes*: Kojima et al. 2003; *Macaca fascata*: Ceugniet & Izumi 2004). Playback studies have further shown that primates actively employ this information during social interactions (Lemasson et al. 2005; Seyfarth & Cheney 2008).

Despite this strong evidence for widespread vocal recognition in primates, as well as other animals, decisive steps have not yet been taken to develop software that can reliably replicate this ability, with its obvious benefits for fieldwork. For example, voice-based individual recognition would enable researchers to bypass the time-consuming process of learning to distinguish individuals and to carry out online identification of individuals that are not directly visible. Other potential benefits of automated caller identification are for research projects that involve large amounts of audio recordings, which are extremely time consuming to analyse. Here, individuals could be tracked using passive audio-recording equipment, which would be useful in helping to estimate home range size and use. New research questions could be asked while working with nonhabituated groups, such as estimating the length of tenure for males and other important demographic variables (Butynski et al. 1992). Finally, in habitats occupied by closely related species, automated caller recognition procedures may help in recognizing different species, which has census applications.

* Correspondence: A. Mielke, Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.
E-mail address: Alexander_Mielke@eva.mpg.de (A. Mielke).

0003-3472/$38.00
http://dx.doi.org/10.1016/j.anbehav.2013.04.017

In animal vocalization studies, information about a call is often extracted manually from its spectrogram, while the choice of parameters is often driven by the intuition of the researcher, potentially eliminating valuable information. The manual extraction makes the process unsuitable for online identification and analysis of large data sets. Call classification is usually executed using discriminant function analysis (DFA), based on the assumption that the extracted features are separable into subsets through the linear division of class patterns along linear planes in space (Bortz 2005). Possible nonlinear applications of DFA (Mika et al. 1999) have not yet been used in animal call recognition research. As the perception and classification by conspecifics might be nonlinear, the DFA might not be optimal to model recognition processes (Deecke & Janik 2006).

In this paper, we introduce a method for automated species, call type and caller identity identification consisting of a feature extraction mechanism and a classifier. Feature extraction is the process of transforming each call from a high-dimensional to a low-dimensional vector (Bahoura & Simard 2010), preserving enough information for further processing, while the classifier algorithm operates to assign each call to one of several predefined categories. In developing the method, we took advantage of recent developments in human speaker and speech recognition, which is based on instantaneous processing of speech samples, followed by a robust classification process (Anusuya & Katti 2009). A key attribute is the fast, automated and standardized extraction of features from a sound signal, to reduce the amount of information encoded to a low-dimensional feature space (Campbell 1997). The most widely used method in current human speech and speaker recognition systems involves mel frequency cepstral coefficients (MFCC; Beigi 2011). MFCCs, instead of focusing on certain spectral features, map the entire spectrum by slicing it along the time and frequency axes and assigning values to the resulting cells based on the amplitude of the signal in that cell. Various studies have shown that MFCCs can be employed to classify animal signals (African elephants, *Loxodonta africanus*: Clemins & Johnson 2003; Clemins et al. 2005; birds: Kogan & Margoliash 1998). In contrast to classification approaches based on manually extracted spectral features, the extraction process is fully automated, repeatable and standardized, which makes it particularly attractive for field applications (Cheng et al. 2010). Additionally, the low number of a priori assumptions about the features makes it possible to apply the same algorithm for different call types and species.

The second key component consists of a classification algorithm that is able to operate on the MFCC output. Here, we opted for an artificial neural networks (ANN) approach. ANNs are able to learn associatively, generalize and recognize patterns by using simple units ('neurons') with weighted connections, which enables them to respond to information in differentiated ways (Ghirlanda & Enquist 2007). Like other classifiers, ANNs extract a general pattern from a training set of vocalizations for which category membership is known, which is then used to classify unknown calls. In contrast to DFAs, ANNs do not make any assumption about underlying probability distributions of the input vector and they can map input nonlinearly (Reby et al. 1997). ANN-based algorithms have been used in combination with different feature extraction methods for recognition of callers, call types and species in marine mammals (Mercado & Kuh 1998; Deecke et al. 2000; Campbell et al. 2002; Bahoura & Simard 2010; Charrier et al. 2010; Marcoux et al. 2011), Gunnison's prairie dogs, *Cynomys gunnisoni* (Placer & Slobodchikoff 2004), bats (Armitage & Ober 2010), grasshoppers (Chesmore & Ohya 2004), tungara frogs, *Engystomops pustulosus* (Phelps & Ryan 2000) and various bird species (Chesmore 2001; Terry & McGregor 2002; Aubin et al. 2004). For primates, Pozzi et al. (2010, 2012) showed their potential use in call type recognition and species recognition in lemurs, with an ANN approach achieving recognition accuracies of 94% for seven different call types within one species (*Eulemur macaco*) and 89% between different species of *Eulemur*.

In this study, we investigated whether a combination of ANN and MFCC could identify blue monkey, *Cercopithecus mitis stuhlmanni*, males individually by one call type, the 'pyow' alarm call (Papworth et al. 2008) and whether the same algorithm could also be applied to discriminate different call types in blue monkeys, and between 'pyows' and the calls of sympatric primate species. Besides showing the impact that software based on this method could have on research with primates, which feature strongly in behavioural research and conservation, this study is the first to attempt all three recognition tasks (individual, call type and species) with one set of parameters and the same classification tool. Previous research has mostly focused on single recognition tasks. We tested whether MFCC and neural networks can be the basis for a more generalized recognition software. This is an important step towards an equivalent of human speech and speaker recognition software in animal research.

METHODS

Study Site and Species

The study was carried out between May and August 2012 in one of the last remaining fragments of the Albertine Rift Forest Complex, the Budongo Forest Reserve in Masindi District, western Uganda (between 1°35′–1°55′N and 31°18′–31°42′E). The reserve comprises 793 km^2 of forest reserve, 428 km^2 of which are continuous forest cover (Fairgrieve & Muhumuza 2003), classified as moist semideciduous tropical forest at medium altitude (Plumptre & Reynolds 1994). In an initial 2-week period (18 May to 2 June 2012), the territories of all blue monkey groups in the study area were mapped using GPS and the grid system of the study area, to permit individualized data collection. In the 9 km^2 study area (Schel & Zuberbühler 2012), a total of 52 groups were identified, indicating a high group density of approximately 5.77 groups/km^2.

Like other forest guenons, blue monkeys live in one-male, multifemale groups with male dispersal. Groups are territorial, even though there is overlap between home ranges (Cords 2007). In blue monkeys, the 'pyow' call, a long-distance vocalization used exclusively by resident males, has been shown to differ between individuals (see Fig. 1 for two examples; Marler 1973; Butynski et al. 1992; Price et al. 2009). Acoustically, 'pyows' can be described as loud, explosive calls lasting on average about 110 ms, given in repetitive bouts (Marler 1973). The calls are given in response to a variety of events, mainly general disturbances and dangers on the ground, but also without apparent external stimulus (Butynski et al. 1992; Papworth et al. 2008; Murphy et al. 2013).

Data Collection

To train an ANN classifier, the identity of every individual included in the training set must be known in advance ('label validity': Clemins & Johnson 2003). Owing to the study species' social structure, 'pyow' calls recorded in a particular part of the forest could reliably be assigned to one specific male, the single male of the resident group, which is useful when training the ANN classifier. If group identity was in doubt, for example if the recording was made in an overlapping region, the group was followed for a while and neighbouring groups were located, to prevent misclassification.

Nineteen blue monkey groups were selected in order to obtain repeated recordings of male calls. 'Pyows' were obtained both through playback experiments and opportunistically. All vocal behaviour was recorded using a Marantz PMD-660 (D&M Holdings

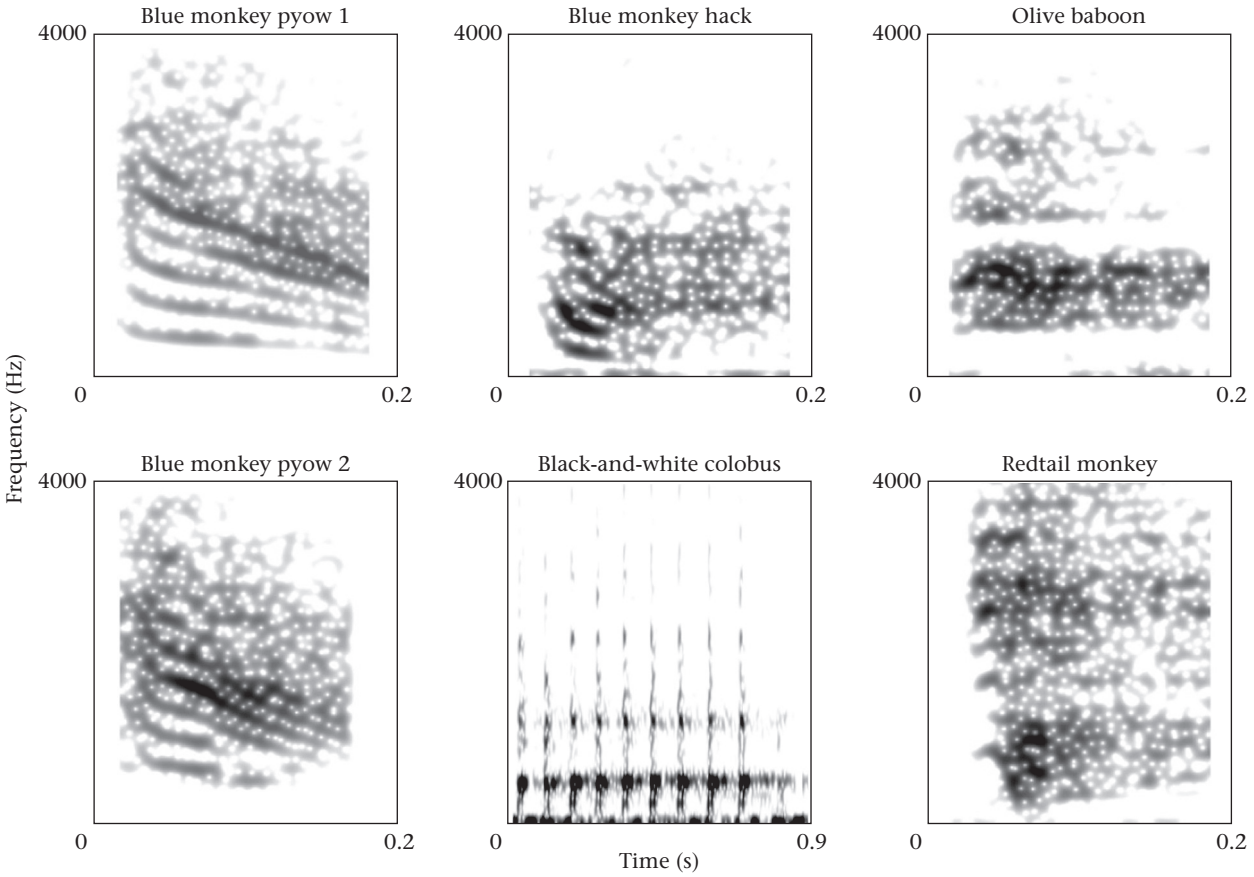

Figure 1. Spectrograms depicting two blue monkey 'pyow' calls, a blue monkey 'hack' call, an olive baboon alarm bark, a guereza colobus alarm roar, and a redtail monkey alarm call.

Inc., Tokyo, Japan) connected to a Sennheiser K6/ME66 directional microphone (Sennheiser Electronic GmbH & Co. KG, Wedemark, Germany). Playback experiments consisted of broadcasting one of three 'pyow' call sequences, recorded from three individuals that were not part of any of the study groups. Call sequences were played back to groups before they discovered the researcher, using an Apple iPod connected to a Nagra DSM speaker-amplifier (Nagra, Audio Technology, Romanel-sur-Lausanne, Switzerland) with an amplitude of approximately 90 dB to mimic the natural amplitude of blue monkey calls. The playback speaker was positioned 30–50 m from the group. If a group member discovered the researchers before a trial, the experiment was discontinued but the group was followed for up to 15 min, which often led to the male producing 'pyow' calls to the researchers, especially in the early morning and late afternoon. Although most blue monkey groups in the study area are used to human presence, they regularly produced 'pyows' when discovering humans. 'Pyows' given for other reasons (e.g. to chimpanzees, thunder) or for no apparent reason, were included as well.

'Pyows' were included in the data set if their frequency bands were clearly visible in the spectrogram. To create a data set that contained enough calls to detect and generalize underlying patterns, a sufficient number of single calls is needed. Therefore, we decided to include more than one 'pyow' per recording. A maximum of 10 'pyows' (referred to in the following as one sequence) were extracted from each recording, to avoid a single recording, containing a high number of 'pyows', being included in the training data set. This would potentially lead to problems of pseudoreplication, as idiosyncrasies of single recordings containing a large number of calls could have a disproportionate influence on the classifier, reducing the generalizability. To test the ability of the method to classify new recordings while avoiding pseudoreplication, a leave-one-out validation on the sequence level was used in the following way. A training set was constructed so that all but one sequences were included; the latter was then classified. This way, no calls with the same recording properties were present in both the training and test data sets. This procedure simulated their later use, that is, classifying unknown calls based on a known data set. Individuals were included in the final data set if more than 30 high-quality 'pyows' from at least five different recordings were available. The procedure for the other monkey species and for blue monkey 'hack' calls was the same as for blue monkey 'pyows'.

Recognition Tasks

(1) Individual recognition task. Fourteen blue monkey males fulfilled the inclusion criteria, resulting in a final full data set of $N = 630$ single 'pyow' calls, taken from 83 recordings. 'Pyows' were tested using the result both for single calls and for entire call sequences. The overall classification for the sequence was obtained by assessing which classification result the majority of the single calls in the sequence got.

(2) Call recognition task. Blue monkey 'hack' calls were recorded ad libitum over the course of the study, to test whether the method is able to distinguish between different call types (see Fig. 1). 'Hack' alarm calls are contextually much more specific in that they are almost exclusively given to aerial predators, especially crowned eagles, *Stephanoaetus coronatus* (Papworth et al. 2008). Encounters of monkeys with crowned eagles are common in the study area. Single 'hack' calls were extracted from six different recordings ($N = 56$) of

different individuals, and subsequently matched to six randomly chosen sequences of 'pyow' calls ($N = 56$) from six males. The resulting network was tested using a leave-one-out validation on the sequence level for the 'hack' calls, and one 'pyow' call sequence per individual that was not included in the training data set.

(3) Species recognition task. Recordings from the three other monkey species found in Budongo Forest (olive baboon, *Papio anubis*, barks; redtail monkey, *Cercopithecus ascanius schmidti*, alarm calls; guereza colobus, *Colobus guereza occidentalis*, alarm roars) were collected ad libitum over the course of the study, during situations in which they were associated with the blue monkeys (see Fig. 1 for examples). For colobus monkeys, we used recordings of ground-predator alarm roars ($N = 50$ phrases from recordings of five different individuals). For baboons, we used recordings of alarm barks ($N = 32$ single barks from recordings of five different individuals), and for redtail monkeys we used recordings of male redtail alarm calls ($N = 15$ from 15 different individuals). All 'pyows' for the training data set were taken from the same individual ($N = 48$) to test whether the classifier could identify new individuals from the same species. A leave-one-out validation method on the sequence level was used for baboons, colobus and redtail monkeys. One sequence per individual was used to classify 'pyows'.

(4) Comparison with DFA. To evaluate the abilities of ANNs as a classifier in recognition tasks, we tested them against the current standard in animal vocal research, the DFA, which has been used to classify blue monkey 'pyow' calls (Butynski et al. 1992). In DFA as it has been used in speaker recognition so far, linear combinations of continuous and categorical predictor variables are determined that allow for a maximum discrimination of compared groups (Bortz 2005), and the group membership of unknown individuals is predicted depending on the discriminant functions established in a training set. Possible nonlinear applications of DFA (Mika et al. 1999) have not featured widely in animal call recognition research. To compare both classifiers, the leave-one-out validation on the sequence level was performed with the DFA as well. Nonparametric exact Wilcoxon signed-ranks tests were used to test whether the recognition performance of one method was significantly higher, as both classifiers were run on the same data set.

Feature Extraction

Call preprocessing

All recordings were preprocessed using Audacity 2.0.0 (http://audacity.sourceforge.net). Both feature extraction and neural network analyses were carried out using Matlab R2012b (The MathWorks, Inc., Natick, MS, U.S.A.). The MFCCs in this study were computed using the 'melcepst'-routine available in the toolbox Voicebox (http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html) for MATLAB. As MFCCs are sensitive to additive background noise (Fox et al. 2006), recordings were first treated using the 'noise removal' option of Audacity. To this end, a brief segment of background noise was selected manually, and then automatically subtracted from the entire recording. Afterwards, a low-pass filter at 4000 Hz was applied to remove high-frequency noise, mainly cicadas.

Mel frequency cepstral coefficients

MFCC are based on the 'mel' scale, which better matches the pitch perception of the auditory system of terrestrial vertebrates than the more common and linear Hertz scale (Deecke & Janik 2006). For example, although humans perceive frequencies below 1000 Hz in a linear way, this is not true for acoustic energy above 1000 Hz, which is perceived in logarithmic spacing (Stevens et al. 1937). The Hertz scale therefore overemphasizes the high-pitched features of vocalizations (Cheng et al. 2010). The relationship between the Hertz and mel scales can be described as $F_{\mathrm{mel}} = 2595 \times \log(1 + F_{\mathrm{Hz}}/700)$.

MFCCs represent both static and dynamic features of vocalizations. This is achieved by slicing the power spectrum along the time and frequency axes and assigning values to the resulting cells based on the amplitude of the signal in that cell. A moving window cuts the call into frames, and filter banks are used to transform the spectrum into the mel scale (see Cheng et al. 2010). The resulting frames can be analysed as stationary signals (Beigi 2011). In human speaker recognition, the window size is typically chosen to be 30 ms; however, blue monkey 'pyows' have a relatively higher dominant frequency (1.68–2.15 kHz; Marler 1973) compared to the human voice, making it necessary to adjust the window size for the trade-off between frequency resolution and signal stationarity (Clemins et al. 2005). As the same number of parameters is needed for every call, even if they differ in duration, we chose seven frames to describe each call. The number of seven frames was chosen as the resulting windows have on average the adjusted size (between 15 and 20 ms), while the number of parameters for every call is standardized. To avoid the loss of information from edge effects of the cutting process and improve the temporal resolution, we ensured that the frames were overlapping by two-thirds, based on the recommendation of other authors (Clemins et al. 2005). Each frame was then multiplied with a Hamming window, the fast Fourier transform (FFT) was computed for each frame, and the frequency axis was warped to the mel scale by multiplying the transformed spectrum with a series of 32 mel-spaced triangular filters (Cheng et al. 2010). Finally, a discrete cosine transform, applied to the energy from the frequency band filters, was used to convert the mel spectrum into cepstral coefficients (Clemins et al. 2005).

As the MFCCs are static pictures of the frequency bands at a certain point in time, delta-cepstral coefficients have been proposed to capture the dynamics between the coefficients, and shown to increase the recognition rate significantly (Kumar et al. 2011). The delta-cepstral coefficients are the first-order derivative of the original cepstral coefficients, therefore depicting how the functions change over time (Beigi 2011). In our case, we had 32 MFCCs with their 32 delta-cepstral coefficients over seven frames, leaving us with a vector of 448 elements that describes each vocalization. The coefficients were automatically transformed into a suitable input vector for the ANN. The MFCCs as used here do not account for duration differences between call types or species, as every call, whether 2 s or 2 min long, is reduced to the same number of parameters. Still, information about the temporal course of the call is conserved, as every frame is a static picture of that specific area of the frequency curve. Furthermore, the delta-cepstral coefficients depict the gradient of the curve. Duration can be added as an additional element to the input vector of the classifier, which can be useful especially when two individuals, species or call types differ systematically in call duration. However, as there are indications that temporal variation exists even within the same individual and call type, depending on which position the call has in an entire call bout (Arnold & Zuberbühler 2006), and as the duration of calls overlapped strongly between the different species, call duration was not included.

Classification

ANNs have long been employed for their ability to detect patterns in different kinds of data, as they do not make assumptions about underlying probability distributions and can therefore be trained to solve complex, nonlinear problems (Anusuya & Katti 2009). The network selected for this study is a variation of a multilayer, feed-forward network (also called 'backpropagation

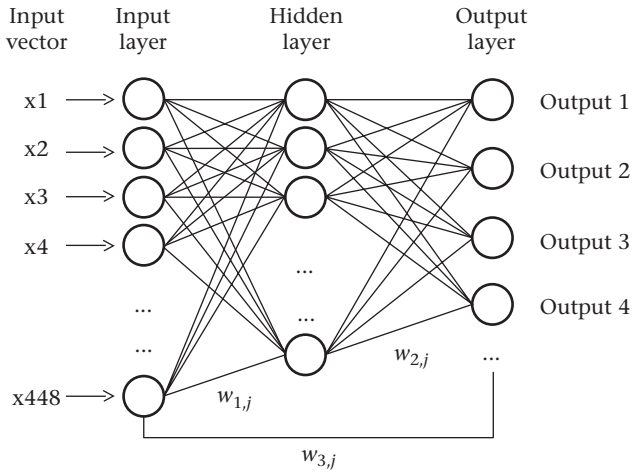

Figure 2. General architecture of a cascade feed-forward neural network with one hidden layer. Every neuron is connected to each neuron in the following layers with a connection of weight $w_{i,j}$, whose strength is determined during the training phase.

network', see Fig. 2; Rumelhart et al. 1986), the most commonly used ANN for vocal recognition (Bahoura & Simard 2010). It consists of three parts: (1) input layer: the input vector, created by the feature extraction process; (2) hidden layer: a predefined number of neurons, usually determined by trial and error with a preliminary data set and optimized for the specific problem (Bardeli et al. 2010); (3) output layer: the different classes in the training set (i.e. number of individuals, call types or species). All neurons of the input and hidden layers are connected to every neuron of the subsequent layer. In this case, a cascade feed-forward network was used, which varies from the original backpropagation by having a direct connection from the input to the output layer, therefore connecting every layer with all previous layers and potentially increasing the speed of learning of the network (Demuth et al. 2009).

The training data set is used to minimize the error between the predicted classification (the known identity of the calls in the training set) and actual classifications (the output of the network). The values of the input layer neurons of all calls in the training set are passed from one layer to the next, and they are changed according to the weights of the connections between the neurons. All weighted inputs that enter a neuron result in one value for that neuron, calculated using a tan-sigmoid transfer function. These values are then used to calculate the next layer. Weights were adjusted using a gradient descent function with momentum and adaptive learning rate.

If a change of weights leads to a better match between predicted and actual classification, the new weights are fixed, resembling the learning process of biological neurons, whose connections get stronger when reward is achieved. Over the course of numerous iterations with multiple calls of all individuals, the entire network will be shaped in a way that each individual's patterns are uniquely represented. For this process to be successful, the training data set has to be sufficiently large and representative for the calls of the individual or group. If not enough calls are used in the training set, or they represent only a small part of the spectrum found in this individual, type or species, the ANN will be unable to classify new calls correctly. A broad, high-quality training data set is therefore essential for this method (Anusuya & Katti 2009). The network and its weights can afterwards be saved and used to classify new calls.

The input layer size in all three conditions (caller, species, call type) consisted of 448 neurons, the number of MFCCs extracted per call. The optimal number of hidden layer neurons for the network has to be determined empirically, because an insufficient number of neurons stops the network from converging, whereas too high a number compromises computational efficiency. Similarly, the optimal number of iterations for which the training data set is run in order to fit the network is a compromise between underfitting (insufficient representation of individual patterns because of insufficient training) and overfitting of the training set (reduced robustness and generalization of the network by training it too much for this particular training data set; Reby et al. 1997). For the hidden layer, we determined the optimal size at 50, 40 and 40 neurons and 250, 60 and 150 iterations, for individual, call type and species recognition, respectively, by systematically varying the parameters on a partial data set. As a single neural network could potentially fail to converge on an optimal fit, 15 networks were computed for every classification, and the overall result was evaluated (Reby et al. 1997). This procedure is computationally more expensive, but increases the robustness of the software.

RESULTS

Recognition

For the 14 individuals tested, random assignment of caller identity would result in an expected classification result of 7.14%. The overall observed classification result for single 'pyows' was 57% (range 30–85%; Table 1) and 73% for entire sequences (range 40–100%; Table 2). An exact Wilcoxon signed-ranks test revealed that, when we combined the classification results of the 'pyows' of the entire sequence, the accuracy was significantly higher ($T = -3.297$, $P < 0.001$). One-sample Wilcoxon signed-ranks tests showed that both single calls ($T = 3.297$, $P < 0.001$) and sequences ($T = 3.297$, $P < 0.001$) performed significantly better than chance.

The calling species could be correctly assigned in 96% of cases for single calls (blue monkeys: 97%; black-and-white colobus: 100%; olive baboon: 100%, redtail monkey: 87%). The call type was identified correctly in 98% of cases for single calls ('hacks': 100%; 'pyows': 97%), and 100% when using whole sequences (see Table 3 for the overall results of all tests).

Comparison with Neural Networks: DFA

To reduce the number of parameters used in the DFA, a principal component analysis was conducted. Factors were included if they had an eigenvector exceeding 2, which led to 42 factors being extracted, with 86.62% of the variance explained. This factor solution was used as input for the DFA. The overall classification

Table 1
Confusion matrix: overlap between predicted (columns) and achieved classifications (rows) of single calls ($N = 630$) for all individuals ($N = 14$) using neural networks

Subject	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
A	34	0	0	0	0	0	0	0	1	0	0	5	0	0	0.85
B	0	26	0	0	1	3	3	1	0	0	4	3	0	7	0.54
C	0	0	38	0	0	1	0	6	0	0	3	2	1	0	0.75
D	0	0	0	27	0	0	1	0	0	2	4	2	0	3	0.69
E	0	1	0	1	20	0	1	2	0	0	4	3	0	11	0.47
F	2	3	0	0	0	30	6	1	3	1	0	4	0	1	0.58
G	0	3	4	1	5	5	13	0	3	1	3	1	4	0	0.30
H	0	2	4	0	0	0	0	26	2	0	2	0	0	0	0.72
I	0	0	0	0	0	1	2	0	24	1	4	4	0	6	0.57
J	2	0	2	3	1	4	1	0	1	33	2	1	1	1	0.63
K	0	4	2	1	4	3	4	3	3	3	14	0	1	2	0.32
L	1	3	0	0	0	4	0	0	4	2	1	38	2	0	0.69
M	0	0	0	0	0	0	1	2	0	0	5	0	16	4	0.57
N	0	2	4	0	8	5	2	0	5	8	2	0	3	18	0.32
Total															0.57

Table 2
Confusion matrix: overlap between predicted (columns) and achieved classifications (rows) of call sequences ($N = 83$) for all individuals ($N = 14$) using neural networks

Subject	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
A	8	0	0	0	0	0	0	0	0	0	0	1	0	0	0.89
B	0	7	0	0	0	0	0	0	0	0	0	0	0	1	0.88
C	0	0	5	0	0	0	0	1	0	0	0	0	0	0	0.83
D	0	0	0	5	0	0	0	0	0	0	0	0	0	0	1.00
E	0	0	0	0	4	0	0	0	0	0	0	0	0	1	0.80
F	0	1	0	0	0	5	0	0	0	0	0	1	0	0	0.71
G	0	0	0	0	0	0	2	0	1	0	0	1	1	0	0.40
H	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0.75
I	0	0	0	0	0	0	0	0	3	0	0	0	0	2	0.60
J	0	0	0	0	1	0	0	0	0	5	0	0	0	0	0.83
K	0	0	0	0	1	1	0	1	0	0	2	0	0	0	0.40
L	0	0	0	0	0	0	0	0	1	0	0	5	0	0	0.83
M	0	0	0	0	0	0	0	0	0	0	1	0	4	0	0.80
N	0	0	0	0	1	0	0	0	1	1	1	0	0	3	0.43
Total															0.73

accuracy for the DFA of single 'pyow' calls was 53% (range 20–80%), and 63% (range 20–100%) for the sequence level. The sequence-level classification was significantly better than the single-call level (Wilcoxon signed-ranks test: $T = -3.079$, $P = 0.001$). Although on the single-call level, ANN and DFA did not differ significantly (Wilcoxon signed-ranks test: $T = -1.669$, $P = 0.100$), the ANN outperformed the DFA on the sequence level (Wilcoxon signed-ranks test: $T = -2.252$, $P = 0.023$).

DISCUSSION

Classification Ability

Experienced fieldworkers can usually distinguish with some ease the calling species and call types in primates and, in some cases, the identity of a caller. More importantly, individual recognition by voice has been shown in a good number of primate species, suggesting that an appropriately designed, automated device should be able to achieve the same result. In this study, we combined an ANN and an MFCC algorithm to discriminate primate calls at the individual caller, call type and species level. The combination of MFCCs as feature vector and ANN as classifier was sufficient to distinguish correctly between four primate species with high accuracy (96% of cases). The main cause of error, redtail monkey vocalizations, was represented by a relatively small number of calls, making it likely that, with an improved training set, classification success would increase. Similarly, the neural network was able to distinguish correctly between blue monkey 'pyows' and 'hacks' with high accuracy (98% of the cases for single calls). Only two call types were compared; other call types, notably those given by females, should be included in future studies.

Our main goal, however, was to explore the method's potential for individual recognition, which yielded impressive results, especially considering how difficult this task generally is for human observers. All individuals had classification rates well above the level that would be expected by chance, while several individuals exceeded 80% correct identification.

Table 3
Classification results

	Individual ANN	Individual DFA	Species	Call type
Single call	0.57	0.53	0.96	0.98
Sequence	0.73	0.63	0.97	1

The table shows the amount of correct classification for individual recognition using both artificial neural networks (ANN) and discriminant function analysis (DFA) as classifier, as well as the species and call type recognition using neural networks.

Individuals with comparatively low classification success were individuals that had strong overlap with other blue monkey groups. It is therefore possible that we misclassified a small number of calls, despite our efforts to determine group identity. This finding illustrates that, even if using automated tools, some initial expertise in distinguishing call types, individuals and species is required to set up an efficient training set.

Our results thus suggest that neural networks combined with MFCCs provide an ideal basis for the development of individual recognition software for primate vocalizations. In species with calls longer than 'pyows', opportunities for individual differences and identification accuracy would potentially exceed the levels seen here. Further improvements can be achieved by increasing the number of recordings per individual, which will increase the ability of the neural network to generalize patterns to identify new calls, making it more robust. Given that many field sites work with known, habituated individuals and in some cases already have a significant number of recordings of their primates, these shortcomings can be easily addressed.

Another relevant finding was that the ANN classification algorithm outperformed DFA, which is classically used in animal communication research. ANNs have the advantage of not being restricted to linear patterns between parameters, increasing their suitability to model the actual recognition processes of animals (Deecke & Janik 2006).

For feature extraction we relied on MFCCs, which have a number of advantages over conventional parameter-based acoustic analysis. First, as they do not rely on a researcher-driven selection of spectral features, they are able to capture information that might seem irrelevant to a researcher, despite being important in the recognition process. MFCCs do not make assumptions about the relative importance of specific spectral features, making them very suitable for cross-call and cross-species comparisons. Another advantage of MFCCs is in the standardization of the duration while still incorporating temporal features. However, not being built on the linear Hertz scale normally used to describe sounds, MFCCs are difficult to interpret; the number of coefficients extracted makes it hard to assign importance to any one of them. Ultimately, however, spectral and cepstral approaches are complementary in nature. For scientific investigations into the acoustic basis of individual, species or call recognition, spectral features potentially yield more specific knowledge, whereas MFCCs allow for fast and automatic processing.

One prerequisite for both methods is the availability of a sufficient number of recordings with a high signal-to-noise ratio, as all classifiers available at the moment are sensitive to low recording quality (Brandes 2008). Here, the problem was solved by only using calls of sufficient quality, which was ensured by a manual procedure. For fully automated devices, a solution would have to be found to ensure sufficient acoustic quality of the training set, either by increasing the sensitivity of the recording equipment, using better filters, or having a sufficient number of microphones to minimize the recording distance. Otherwise, low-quality recordings in the data set can influence the training procedure, and make the classification process less accurate.

One drawback of the current method and a future area of research is its inability to classify new individuals, species or call types. This is because the backpropagation network is dependent on an exhaustive training set. New calls can be added to the training set only after they have been classified. One way to deal with this problem is to include a category 'unknown' (Chesmore 2001). Another potential option is the use of unsupervised neural networks, or self-organizing maps (Terry & McGregor 2002), which do not need initial training, but automatically establish the number of different patterns they can find in the data set. This approach has

repeatedly been used in animal bioacoustics (Mercado & Kuh 1998; Terry & McGregor 2002; Tanttu et al. 2003), and has been successfully used to identify lemur calls by Pozzi et al. (2012). It could be included as an initial step before the classification with a supervised neural network, to test whether the number of individuals in the data set changes because of the inclusion of the new call. Unsupervised neural networks might also have significant influence on our understanding of different call types in different species, as our perception of two calls as the same might not be shared by conspecific recipients, but be a result of our limited senses. An unsupervised network, not sharing our assumptions about a call, might be able to detect subtle differences (Deecke & Janik 2006).

Conclusion

Our results presented here demonstrate that automated recognition methods have considerable potential for the study of animals in the wild. Even though the use for nonhuman primates was highlighted here, it is by far not limited to this group, with possible applications ranging across a variety of taxa. The use of MFCCs is attractive because the same procedure can be applied to different species and call types, instead of having to establish customized sets of parameters for different occasions. So far, cepstral coefficients have not played a great role in the study of animal vocalizations, although their potential in complementing spectral-based approaches is considerable. Also, they enable researchers to use the same set of parameters for very different recognition tasks, potentially opening a road for the use of generalized rather than task- and species-specific software. We chose ANNs as a classifier owing to their robustness and good performance in previous studies, and they predictably achieved good results here as well. However, they require a rather large, representative training data set in order to converge. Other classifiers, such as hidden Markov models, Gaussian mixture models and support vector machines, have also been applied successfully to recognize species, call type and individuals in nonhuman animals (Clemins et al. 2005; Brandes 2008; Armitage & Ober 2010; Cheng et al. 2010). Overall, our results suggest that using these methods in animal communication is likely to provide a methodological breakthrough with implications for a range of disciplines.

We thank the Royal Zoological Society of Scotland for providing core funding for the Budongo Conservation Field Station (www.budongo.org). In Uganda, we gratefully acknowledge the National Forestry Authority, the Uganda Wildlife Authority, the Uganda National Council for Science and Technology, the President's Office, and the Jane Goodall Institute-Uganda for permission to conduct our research in the Budongo Forest Reserve. We thank the staff of the Budongo Conservation Field Station, especially Moses Lemi for helping with the data collection.

References

Adi, K., Johnson, M. T. & Osiejuk, T. S. 2010. Acoustic censusing using automatic vocalization classification and identity recognition. *Journal of the Acoustical Society of America*, **127**, 874–883.

Anusuya, M. A. & Katti, S. K. 2009. Speech recognition by machine: a review. *International Journal of Computer Science and Information Security*, **6**, 181–205.

Armitage, D. W. & Ober, H. K. 2010. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics*, **5**, 465–473.

Arnold, K. & Zuberbühler, K. 2006. The alarm-calling system of adult male putty-nosed monkeys, *Cercopithecus nictitans martini*. *Animal Behaviour*, **73**, 643–653.

Aubin, T., Mathevon, N., Luisa, M. & Silva, D. A. 2004. How a simple and stereotyped acoustic signal transmits individual information: the song of the white-browed warbler, *Basileuterus leucoblepharus*. *Anais da Academia Brasileira de Ciencias*, **76**, 335–344.

Bahoura, M. & Simard, Y. 2010. Blue whale calls classification using short-time Fourier and wavelet packet transforms and artificial neural network. *Digital Signal Processing*, **20**, 1256–1263.

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

Beigi, H. 2011. *Fundamentals of Speaker Recognition*. New York: Springer.

Bortz, J. 2005. *Statistik für Human- und Sozialwissenschaftler*. 5th edn. Berlin: Springer.

Brandes, T. S. 2008. Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, **18**, 163–173.

Butynski, T. M., Chapman, C. A., Chapman, L. J. & Weary, D. M. 1992. Use of male monkey pyow calls for long term individual identification. *American Journal of Primatology*, **28**, 183–189.

Campbell, G. S., Gisiner, R. C., Helweg, D. A. & Milette, L. L. 2002. Acoustic identification of female Steller sea lions (*Eumetopias jubatus*). *Journal of the Acoustical Society of America*, **111**, 2920–2928.

Campbell, J. P., Jr. 1997. Speaker recognition: a tutorial. *Proceedings of the Institute of Electrical and Electronics Engineers*, **85**, 1437–1462.

Charrier, I., Aubin, T. & Mathevon, N. 2010. Mother–calf vocal communication in Atlantic walrus: a first field experimental study. *Animal Cognition*, **13**, 471–482.

Cheng, J., Sun, Y. & Ji, L. 2010. A call-independent and automatic acoustic system for the individual recognition of animals: a novel model using four passerines. *Pattern Recognition*, **43**, 3846–3852.

Chesmore, E. D. 2001. Application of time domain signal coding and artificial neural networks to passive acoustical identification of animals. *Applied Acoustics*, **62**, 1359–1374.

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.

Clemins, P. & Johnson, M. 2003. Application of speech recognition to African elephant (*Loxodonta africana*) vocalizations. *Acoustics, Speech, and Signal Processing*, **1**, 484–487.

Clemins, P. J., Johnson, M., Leong, K. & Savage, A. 2005. Automatic classification and speaker identification of African elephant (*Loxodonta africanus*) vocalizations. *Journal of the Acoustical Society of America*, **117**, 956–963.

Ceugniet, M. & Izumi, A. 2004. Individual vocal differences of the coo calls in Japanese monkeys. *Comptes Rendus Biologies*, **327**, 149–157.

Cords, M. 2007. Variable participation in the defense of communal feeding territories by blue monkeys in the Kakamega Forest, Kenya. *Behaviour*, **144**, 1537–1550.

Deecke, V. & Janik, V. M. 2006. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. *Journal of the Acoustical Society of America*, **119**, 645–653.

Deecke, V. B., Ford, J. K. B. & Spong, P. 2000. Dialect change in resident killer whales: implications for vocal learning and cultural transmission. *Animal Behaviour*, **60**, 629–638.

Demuth, H., Beale, M. & Hagan, M. 2009. *Neural Network Toolbox 6 User's Guide*. Natick, Massachusetts: The MathWorks Inc.

Fairgrieve, C. & Muhumuza, G. 2003. Feeding ecology and dietary differences between blue monkey (*Cercopithecus mitis stuhlmanni* Matschie) groups in logged and unlogged forest, Budongo Forest Reserve, Uganda. *African Journal of Ecology*, **41**, 141–149.

Fischer, J., Hammerschmidt, K., Cheney, D. L. & Seyfarth, R. M. 2001. Acoustic features of female chacma baboon barks. *Ethology*, **107**, 33–54.

Fox, E. J. S., Roberts, J. D. & Bennamoun, M. 2006. Text-independent speaker identification in birds. *Proceedings of the Interspeech 2006 and Ninth International Conference on Spoken Language Processing*, **1–5**, 2122–2125.

Ghirlanda, S. & Enquist, M. 2007. How training and testing histories affect generalisation: a test of simple neural networks. *Philosophical Transaction of the Royal Society B*, **362**, 449–454.

Hammerschmidt, K., Newman, J. D., Champoux, M. M. & Suomi, S. J. 2000. Changes in rhesus macaque 'coo' vocalisations during early development. *Ethology*, **106**, 873–886.

Kogan, J. A. & Margoliash, D. 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. *Journal of the Acoustical Society of America*, **103**, 2185–2196.

Kojima, S., Izumi, A. & Ceugniet, M. 2003. Identification of vocalizers by pant hoots, pant grunts and screams in a chimpanzee. *Primates*, **44**, 225–230.

Kumar, K., Kim, C. & Stern, R. 2011. Delta-spectral cepstral coefficients for robust speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, **2011**, 4784–4787.

Lemasson, A., Hausberger, M. & Zuberbühler, K. 2005. Socially meaningful vocal plasticity in adult Campbell's monkeys (*Cercopithecus campbelli*). *Journal of Comparative Psychology*, **119**, 220–229.

Marcoux, M., Auger-Méthé, M. & Humphries, M. M. 2011. Variability and context specificity of narwhal (*Monodon monoceros*) whistles and pulsed calls. *Marine Mammal Science*, **28**, 649–665.

Marler, P. 1973. A comparison of vocalizations of redtail monkeys and blue monkeys, *Cercopithecus ascanius* and *C. mitis*, in Uganda. *Zeitschrift für Tierpsychologie*, **33**, 223–247.

Mercado, E., III & Kuh, A. 1998. Classification of humpback whale vocalizations using a self-organizing neural network. *International Joint Conference on Neural Networks, Proceedings*, **2**, 1584–1589.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B. & Müller, K.-R. 1999. Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing*, **IX**, 41–48.

Murphy, D., Lea, S. E. G. & Zuberbühler, K. 2013. Male blue monkey alarm calls encode predator type and distance. *Animal Behaviour*, **85**, 119–125.

Papworth, S., Boese, A.-S., Barker, J., Schel, A. M. & Zuberbühler, K. 2008. Male blue monkeys alarm call in response to danger experienced by others. *Biology Letters*, **4**, 472–475.

Phelps, S. M. & Ryan, M. J. 2000. History influences signal recognition: neural network models of tungara frogs. *Proceedings of the Royal Society B*, **267**, 1633–1639.

Placer, J. & Slobodchikoff, C. N. 2004. A method for identifying sounds used in the classification of alarm calls. *Behavioural Processes*, **67**, 87–98.

Plumptre, A. J. & Reynolds, V. 1994. The effects of selective logging on the primate populations in the Budongo Forest Reserve, Uganda. *Journal of Applied Ecology*, **31**, 631–641.

Pozzi, L., Gamba, M. & Giacoma, C. 2010. The use of Artificial Neural Networks to classify primate vocalizations: a pilot study on black lemurs. *American Journal of Primatology*, **72**, 337–348.

Pozzi, L., Gamba, M. & Giacoma, C. 2012. Artificial Neural Networks: a new tool for studying lemur vocal communication. In: Leaping Ahead (Ed. by J. Masters, M. Gamba & F. Génin), pp. 305–313. New York: Springer.

Price, T., Arnold, K., Zuberbühler, K. & Semple, S. 2009. Pyow but not hack calls of the male putty-nosed monkey (*Cercopithecus nictitans*) convey information about caller identity. *Behaviour*, **146**, 871–888.

Reby, D., Lek, S., Dimopoulos, I., Joachim, J., Lauga, J. & Aulagnier, S. 1997. Artificial neural networks as a classification method in the behavioural sciences. *Behavioural Processes*, **40**, 35–43.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. 1986. Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. *Vol. 1* (Ed. by D. E. Rumelhart & J. L. McClelland), pp. 675–695. Cambridge, Massachusetts: MIT Press.

Schel, A. M. & Zuberbühler, K. 2012. Predator and non-predator long-distance calls in Guereza colobus monkeys. *Behavioural Processes*, **91**, 41–49.

Seyfarth, R. & Cheney, D. 2008. *Baboon Metaphysics: The Evolution of a Social Mind*. Chicago: University of Chicago Press.

Stevens, S. S., Volkmann, J. & Newman, E. B. 1937. A scale for measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, **8**, 185–190.

Tanttu, J. T., Turunen, J. & Ojanen, M. 2003. Automatic classification of flight calls in crossbill species (*Loxia* spp.). In: Proceedings of the First International Conference on Communication by Animals, Maryland, U.S.A..

Terry, A. M. R. & McGregor, P. K. 2002. Census and monitoring based on individually identifiable vocalizations: the role of neural networks. *Animal Conservation*, **5**, 103–111.

Tibbetts, E. A. & Dale, J. 2007. Individual recognition: it is good to be different. *Trends in Ecology & Evolution*, **22**, 520–537.