BIOACOUSTICS, 2018
https://doi.org/10.1080/09524622.2018.1426042

Taylor & Francis
Taylor & Francis Group

Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia

Dena J. Clink[a], Margaret C. Crofoot[a,b] and Andrew J. Marshall[c]

[a]Department of Anthropology, University of California, Davis, Davis, CA, USA; [b]Smithsonian Tropical Research Institute, Balboa Ancon, Republic of Panama; [c]Department of Anthropology, Program in the Environment, and School for Natural Resources and Environment, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Vocal individuality has been documented in a variety of mammalian species and it has been proposed that this individuality can be used as a vocal fingerprint to monitor individuals. Here we provide and test a classification method using Mel-frequency cepstral coefficients (MFCCs) to extract features from Bornean gibbon female calls. Our method is semi-automated as it requires manual pre-processing to identify and extract calls from the original recordings. We compared two methods of MFCC feature extraction: (1) averaging across all time windows and (2) creating a standardized number of time windows for each call. We analysed 376 calls from 33 gibbon females and, using linear discriminant analysis, found that we were able to improve female identification accuracy from 95.7% with spectrogram features to 98.4% accuracy when averaging MFCCs across time windows, and 98.9% accuracy when using a standardized number of windows. We divided our data randomly into training and test data-sets, and tested the accuracy of support vector machine (SVM) predictions over 100 iterations. We found that we could predict female identity in the test data-set with a 98.8% accuracy. Using SVM on our entire data-set, we were able to predict female identity with 99.5% accuracy (validated by leave-one-out cross-validation). Lastly, we used the method presented here to classify four females recorded during three or more recording seasons using SVM with limited success. We provide evidence that MFCC feature extraction is effective for distinguishing between female Bornean gibbons, and make suggestions for future vocal fingerprinting applications.

ARTICLE HISTORY
Received 9 July 2017
Accepted 19 December 2017

KEYWORDS
Support vector machine; Mel-frequency cepstral coefficients; *Hylobates*; Stability of Altered Forest Ecosystems Project; vocal fingerprinting; acoustic monitoring

Introduction

Passive acoustic monitoring is a non-invasive technique that utilizes sound recording devices to monitor vocal animals (Merchant et al. 2015). The potential for passive acoustic monitoring to improve conservation efforts for terrestrial animals is widely recognized (Blumstein et al. 2011; Wrege et al. 2017), and bioacoustics techniques are being used to

CONTACT Dena J. Clink ✉ djc426@cornell.edu

Supplemental data for this article can be accessed https://doi.org/10.1080/09524622.2018.1426042

identify individuals in a wide variety of taxa including owls (Grava et al. 2008), orangutans (Spillmann et al. 2017) and tigers (Ji et al. 2013), as well as for occupancy detection of primates (Heinicke et al. 2015; Kalan et al. 2015) and monitoring of primate group ranging and territory use (Kalan et al. 2016). The use of bioacoustical methods to address ecological and conservation questions has become increasingly popular due to the increase in data storage capabilities and battery life, a decrease in size and cost of recording devices and the development of new methods for automating acoustic analyses (Blumstein et al. 2011).

Acoustic identification of individuals requires two main steps: feature extraction and subsequent classification using an algorithm. The traditional method of feature extraction in many bioacoustics applications, particularly in studies of primates, is to convert the waveform obtained from recorded vocalizations to a spectrogram, and then manually estimate features, such as note duration and frequency, from the spectrogram (Marler and Hobbett 1975; Haimoff and Tilson 1985; Feng et al. 2014; Terleph et al. 2015). This technique is often subjective and highly labour-intensive, and it is not clear which features most accurately permit individual discriminability (Kirschel et al. 2009). The use of Mel-frequency cepstral coefficients (MFCCs) provides a fully automated method of feature extraction that is standardized and reproducible (Mielke and Zuberbühler 2013). MFCCs have been used successfully in human speech recognition (Picone 1993; Han et al. 2006), bird song classification (Chou et al. 2008; Lee et al. 2006), classification of species, call type and caller identity in blue monkeys (Mielke and Zuberbühler 2013) and discrimination between male orangutan individuals (Spillmann et al. 2017).

The main goal of classification in bioacoustics is to predict predefined class membership (e.g. individual identity) based on extracted call features (Lee 2010). One of the most commonly used methods for classification of primate vocalizations is linear discriminant function analysis (DFA) (Delgado 2007; Wich et al. 2008; Heller et al. 2010; Santorelli et al. 2013). Linear DFA is a supervised, multivariate technique that tests whether different classes of objects (e.g. calls) can be distinguished by a set of parameters (or features) estimated from each of those objects (Venables and Ripley 2002; Mundry and Sommer 2007). A major limitation of linear DFA in bioacoustics applications is that it only permits consideration of a single factor at a time. For example, most studies of animal calls include multiple calls per subject, and studies that investigate differences in calls between species, sex or social context exhibit a two-factorial design, with 'subject' being one factor and species, sex or social context being the second factor (Mundry and Sommer 2007). Linear discriminant analysis is therefore appropriate for distinguishing between individuals, but not between sexes or sites when several replicate calls are included for each individual, or if the same individual is recorded at two separate times, as this violates the assumptions of statistical independence (Venables and Ripley 2002).

Support vector machines (SVMs) are supervised classification techniques that are recognized as the leading approach for many discriminative problems. They make fewer assumptions about the underlying data structure (Roma et al. 2010) and consequently, they are more flexible than linear DFA. SVMs were originally developed for binary classification problems, and perform classification by maximizing the margin between two classes (Cortes and Vapnik 1995). Multi-class classification can be done using SVM with a variety of different strategies; one of the more successful approaches is 'one-against-one', where a binary classifier is trained for each pair of classes in the data-set (Hsu and Lin 2002). SVMs have been used to effectively classify bird songs (Cheng et al. 2010; Dufour et al. 2014), dolphin whistles (Esfahanian et al. 2014), 88 different insect species (Noda et al. 2016) and primates (chimpanzees, Fedurek et al. 2016; marmosets, Turesson et al. 2016).

Gibbons (family Hylobatidae) provide a model system for using passive acoustic monitoring, as they are highly territorial species with relatively long territory tenure (Bartlett et al. 2016), and most gibbons regularly engage in duetting (Geissmann 2002). Vocal individuality has been documented in many gibbon species including the white-handed gibbon (*Hylobates lar*; Terleph et al. 2015), the agile gibbon (*H. agilis;* Oyakawa et al. 2007) and the Bornean gibbon (*H. muelleri*; Clink et al. 2017). Gibbon researchers have proposed that this vocal individuality may be useful for tracking and monitoring individuals, a technique known as vocal fingerprinting (Sun et al. 2011). Wide-scale acoustic monitoring of gibbons has yet to be adopted, presumably due in part to the lack of automated, easy-to-use acoustic identification methods. In fact, the use of vocal individuality in census or monitoring roles has been under-utilized across taxa (Terry et al. 2005), and more work is needed to collect and analyze acoustic data in ways that will be relevant to conservation (Pimm et al. 2015).

Here, we present and test a semi-automated vocal fingerprinting method to identify Bornean gibbon females within the landscape of a large-scale habitat fragmentation experiment at the Stability of Altered Forest Ecosystems site in Sabah, Malaysia. First, we compare the effectiveness of two distinct methods of MFCC feature extraction for identification of individual gibbon females. Second, to compare the influence of different SVM kernel types on our prediction accuracy, we test the ability of SVMs to predict female identity using a multi-class SVM and four different kernel types over multiple iterations of randomly selected sets of training and test calls. Third, using the SVM kernel that yielded the highest classification accuracy, we compare the results of SVM classification with linear discriminant analysis using our entire data-set (validated using leave-one-out cross-validation). Lastly, we use a multi-class SVM to predict identity of calls recorded at the same recording location but during different recording seasons (which are assumed to have been produced by the same female). We used a reduced data-set with only four females for which we had recordings taken over at least three separate recording seasons.

Methods

Study site

We conducted our study at the Stability of Altered Forest Ecosystems (SAFE) project located within the Kalabakan Forest Reserve (N04°422367′, E117°3559′), in Sabah, Malaysia. The SAFE project covers approximately 7200 ha, with a significant portion of the land allotted for conversion to oil palm plantation, and 800 ha of cultivable land to be left as forest fragments. A complete description of the study site can be found in Ewers et al. (2011). The SAFE site is divided into six replicate blocks (denoted by the letters A–F), each containing plots with samples of four 1 ha fragments, two 10 ha fragments and one 100 ha fragment. A control plot of 2200 ha of old growth forest is located within the reserve, and there are 3 control plots located in 1 million ha of continuous old growth forest located approximately 60 km away from SAFE, in the Maliau Basin Conservation Area. DJC visited the SAFE site in January 2013, July–August 2013, December 2015, August 2015 and September 2016. As of September 2016, most of the site had been cleared for conversion to oil palm.

Study subjects and data collection

A detailed explanation of data collection methods can be found in Clink et al. (2017). All analyses presented here focused on the female contribution to the duet, known as the great call (Geissmann 2002). Briefly, recordings were made using a Marantz PMD 660 flash recorder equipped with a RODE NTG-2 directional condenser microphone at a distance of ~250 m or less, at a sample rate of 44.1 kHz and 16-bit. Data collection was augmented using a Roland CUBE Street EX 4-Channel 50-Watt Battery Powered Amplifier to broadcast a recorded duet in assumed territories of gibbon groups to elicit vocal responses. We considered females that were recorded >500 m apart as separate females, as this is the approximate width of a gibbon territory (Brockelman and Srikosamatara 1993; Bartlett et al. 2016). Previously, we found there were no substantial differences between call features collected during playbacks vs. spontaneous calling bouts (Clink et al. 2017), so we lumped all calls together for the present analysis.

Female identification

We identified groups based on recording location, group composition and unique behaviours (e.g. long calling bouts, unique male contribution, co-singing daughters). As we were working with wild, unhabituated gibbons it was difficult to distinguish among individuals via unique markings, which is a commonly used method to identify habituated primates. To identify groups across recording season we relied mostly on recording location along with group composition. Gibbons are highly territorial, and maintain relatively long territory tenure. For example, lar gibbons (*H. lar*) were found to have highly stable territories over the course of 10 years (Bartlett et al. 2016). Therefore, it seems unlikely that gibbons at our study site would have shifted their territories substantially over the course of our study period.

Call pre-processing

Our method required pre-processing to identify and extract individual great calls from field recordings. To do this, we created spectrograms using the program Raven Pro 1.5 Sound Analysis Software (Cornell Lab of Ornithology, Bioacoustics Research Program, Ithaca, New York). We made spectrograms with a 512-point (11.6 ms) Hann window (3 dB bandwidth = 124 Hz), with 75% overlap, and a 1024-point DFT, yielding time and frequency measurement precision of 2.9 ms and 43.1 Hz. Gibbon female great calls follow a highly standardized structure with a series of longer, frequency modulated notes leading into rapidly repeated shorter notes. Bornean gibbon great calls recorded at our site range in duration from 9.1 to 27.3 s, with the total number of notes ranging from 42 to 124 (Clink et al. 2017). We defined the start of the great call as the portion of the call when notes first reach a duration greater than 0.20 s, the start of the trill as the point when introductory note duration decreases to 0.135 s or less, and the end of the call to be when the trill ends (representative spectrogram in Figure 1). We identified each instance of a female great call and then saved each individual great call as a Waveform Audio file.

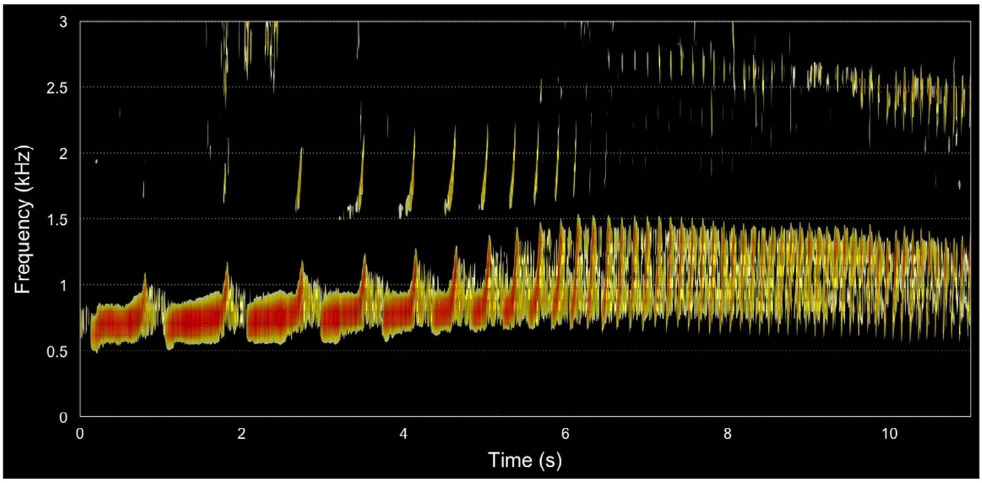
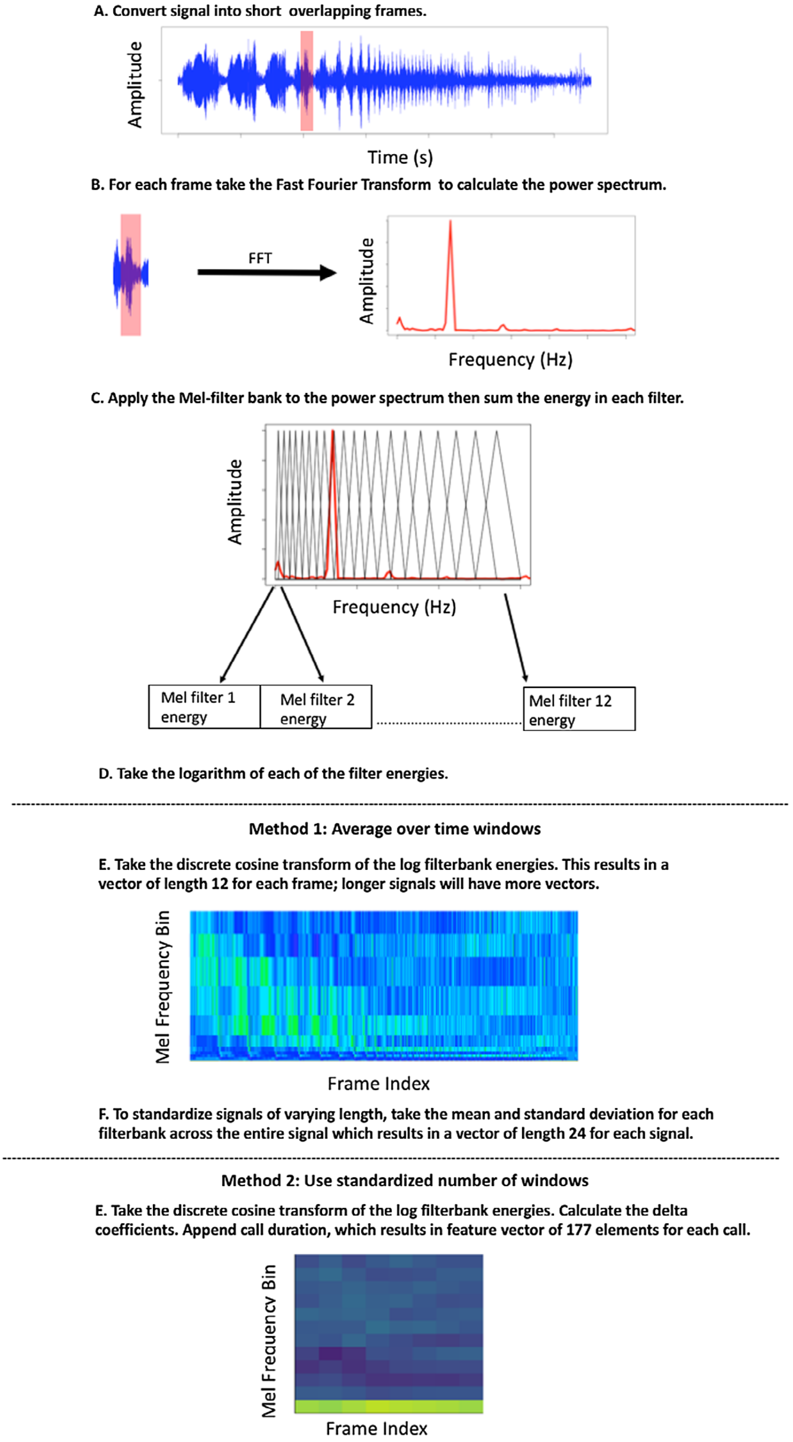


Figure 1. Representative spectrogram of Bornean gibbon female great call.

MFCCs: averaging over time windows

For the first method of MFCC feature extraction we followed the steps outlined in Figure 2. First, we converted each Waveform Audio file to a waveform in the time domain (Figure 2(A)). Then we extracted overlapping frames of 0.25 s at 0.01 s intervals. For each 0.25 s frame we applied the Fast Fourier Transform to convert the signal into a power spectrum (Figure 2(B)). We calculated 12 Mel-filters (or band pass filters) between 400 and 2000 Hz (the frequency range of Bornean gibbon female great calls), applied the Mel-filters to the power spectrum and calculated the energy in each filter (Figure 2(C)). The Mel-filter is based on the 'mel' scale, which more closely aligns with pitch perception in terrestrial vertebrates (Deecke and Janik 2006), with smaller filters at lower frequencies and larger filters at higher frequencies. Animals perceive changes in frequency below 1000 Hz linearly, but that is not the case at frequencies above 1000 Hz, which means that the linear scale tends to overemphasize high-frequency components of vocalizations (Cheng et al. 2010).

We took the logarithm of each of the Mel-filter energies, which is motivated by human hearing, as humans don't hear loudness on a linear scale (Stevens and Guirao 1962). We then took the discrete cosine transform of each of the log Mel-filter energies, which de-correlates the values as they tend to be highly correlated due to their overlapping time windows. This results in a vector with 12 elements for each time frame; longer signals will have more vectors (Figure 2(E)). The first MFCC for each window corresponds to the power of the signal, and represents signal loudness (Han et al. 2006; Muda et al. 2010), which will vary depending on distance to calling animal, ambient background noise and many other factors, therefore we did not include the first MFCC for our discriminative tasks, as it may improve discrimination based characteristics of the particular recording and not the calling animal. To create a feature vector of standardized length to be used in subsequent analyses, we took the mean and the standard deviation of the values for each Mel-filter across the entire signal (Guo and Li 2003), which resulted in a vector containing 24 features for each great call.



A. Convert signal into short overlapping frames.

B. For each frame take the Fast Fourier Transform to calculate the power spectrum.

C. Apply the Mel-filter bank to the power spectrum then sum the energy in each filter.

D. Take the logarithm of each of the filter energies.

Method 1: Average over time windows

E. Take the discrete cosine transform of the log filterbank energies. This results in a vector of length 12 for each frame; longer signals will have more vectors.

F. To standardize signals of varying length, take the mean and standard deviation for each filterbank across the entire signal which results in a vector of length 24 for each signal.

Method 2: Use standardized number of windows

E. Take the discrete cosine transform of the log filterbank energies. Calculate the delta coefficients. Append call duration, which results in feature vector of 177 elements for each call.

Figure 2. Steps outlining the calculation of Mel-frequency cepstral coefficients and the two MFCC feature extraction methods presented. (A) Representative waveform of a Bornean gibbon great call; (B) Representative power spectrum from a single time frame; (C) 12 Mel-filters applied to the power spectrum; (E) A plot of the 12 MFCC coefficients calculated for each time frame of the great call.

MFCCs: using a standardized number of time windows and recursive feature elimination

For the second method of MFCC feature extraction, we created a standardized number of windows for each great call (Mielke and Zuberbühler 2013), and calculated the MFCCs as outlined above for each of the windows (Figure 2; steps A–D). We used eight windows for this feature extraction method, and gibbon great calls range in duration from 9.1 to 27.3 s, so our window length ranged from 1.13 to 3.4 s. Although this window length is substantially longer than those used by other authors (e.g. Mielke and Zuberbühler 2013), we experimented with shorter window lengths and did not find that this improved our classification ability. We also calculated the delta-cepstral coefficients, which have been proposed to capture the dynamics of the MFCCs over the course of a call (Kumar et al. 2011); delta coefficients are the first-order derivative of the original cepstral coefficient (Beigi 2011). Including the delta coefficients allowed us to incorporate information about how the MFCCs change over the course of a call. We included a standardized number of windows for each call, and this did not provide information about call duration, so we included call duration as an additional element, which resulted in a vector with 177 elements for each call.

Our second method of MFCC feature extraction included many more features than the previous method of averaging over time windows, and many of these features were likely to be redundant. Therefore, to identify which of the features were most important we used recursive feature selection (Guyon et al. 2002). Recursive feature selection is an iterative process wherein predictors are ranked, and the less important features are subsequently eliminated, until a subset of predictors is identified that can produce an accurate model (Kuhn 2008). We implemented recursive feature selection using the ‘mSVM-rfe’ package (Colby 2011) in the R programming environment (R Development Core Team 2017). The SVM-RFE fits a simple linear SVM, ranks the features, then eliminates the feature with the lowest rank, which results in a list of features ranked from most to least important (Guyon et al. 2002; Colby 2011).

Linear discriminant function analysis

Linear DFA is a supervised technique wherein classes are defined prior to analysis (Venables and Ripley 2002). The use of DFA was appropriate for distinguishing between females when recordings were taken during a single recording session (i.e. there were no repeat recordings at the same location on different days), as the data-set structure consisted of replicate calls within females, resembling a one-way analysis of variance with ‘individual’ as the factor (Mundry and Sommer 2007). We used DFA to compare the two methods of MFCC feature extraction presented here with previously published features estimated from the spectrogram using the same data-set (Clink et al. 2017), and we used leave-one-out cross-validation to estimate accuracy. To identify the ideal number of features to include for our second MFCC feature extraction method we calculated a linear DFA (validated using leave-one out cross-validation), and sequentially added the highest ranked features, until we identified the minimum number of features needed to obtain a high classification accuracy. We also compared the results of DFA with SVM using both MFCC feature extraction techniques (methods outlined below).

Support vector machines

SVM is a supervised machine learning algorithm where each data point is plotted in *n*-dimensional space (where *n* is the number of features) and the value of each feature determines the coordinate; classification is then performed by finding the hyperplane that best differentiates between classes (Cortes and Vapnik 1995; Lee 2010). We used SVMs to distinguish between females using the 'one-against-one' approach wherein a binary classifier was created for each pair of females. To improve the performance of SVM, the cost parameter and the kernel function parameters need to be chosen carefully, as the SVM can be highly sensitive to variation; the method of choosing the values for these parameters that minimizes the identification error is known as tuning (Duan et al. 2003).

Tuning the cost and gamma parameters for SVM

The cost parameter controls the penalty of misclassified instances, and when the cost is small the margins will be very wide and there will be many support vectors (Karatzoglou et al. 2004; Roma et al. 2010; James et al. 2013; Figure 3). We tested the effect of varying cost parameters on our prediction accuracies, using the 'tune' function in the 'e1071' R package (Meyer et al. 2017), and running the SVM using the following cost parameter values: 0.001, 0.01, 0.1, 1, 2, 10, 100, 1000. The gamma parameter controls the curvature of the hyperplane for non-linear kernels; if gamma is too small then the model is too constrained and cannot capture the complexity of the data, whereas if gamma is too large this can lead to overfitting and low generalizability (Pedregosa et al. 2011). For the non-linear kernels we tested the following gamma parameters: 0.001, 0.01, 0.1, 0.5, 1.0, 2.0.

SVM training and validation

Data are often not linearly separable (Roma et al. 2010), so we experimentally tested which of the four different kernel types: 'linear', 'polynomial', 'radial basis' or 'sigmoid', resulted in the best ability to predict the test data from the training data. We randomly assigned approximately 80% of our original data-set as training data, and the remaining 20% of our data-set to the test data-set (Murphy 2012; Fedurek et al. 2016; Turesson et al. 2016). The 'svm' function in the 'e1071' R package allows the user to set the k-fold cross-validation parameter, wherein the training data is divided into k-folds, with one fold being used to validate the model and the rest used to train the model (Meyer et al. 2017). We set the k-fold parameter equal to 5, which means that 80% of our training data was used to train the model, and 20% of the training data was used for validation. Therefore, our data were effectively divided into three sets: training (64%), validation (16%) and test (20%). We then calculated the accuracy of the SVM model predictions of female identity by creating a confusion matrix of actual versus predicted female identity in the test data-set, and then averaging over the classification accuracies for each female. We did this over 100 iterations for each of the kernel types, running the 'tune' function on the test data-set over each iteration. After identifying which kernel yielded the best performance, we then tested the performance of the SVM classification using our whole data-set, and estimated classification accuracy using leave-one out cross-validation.

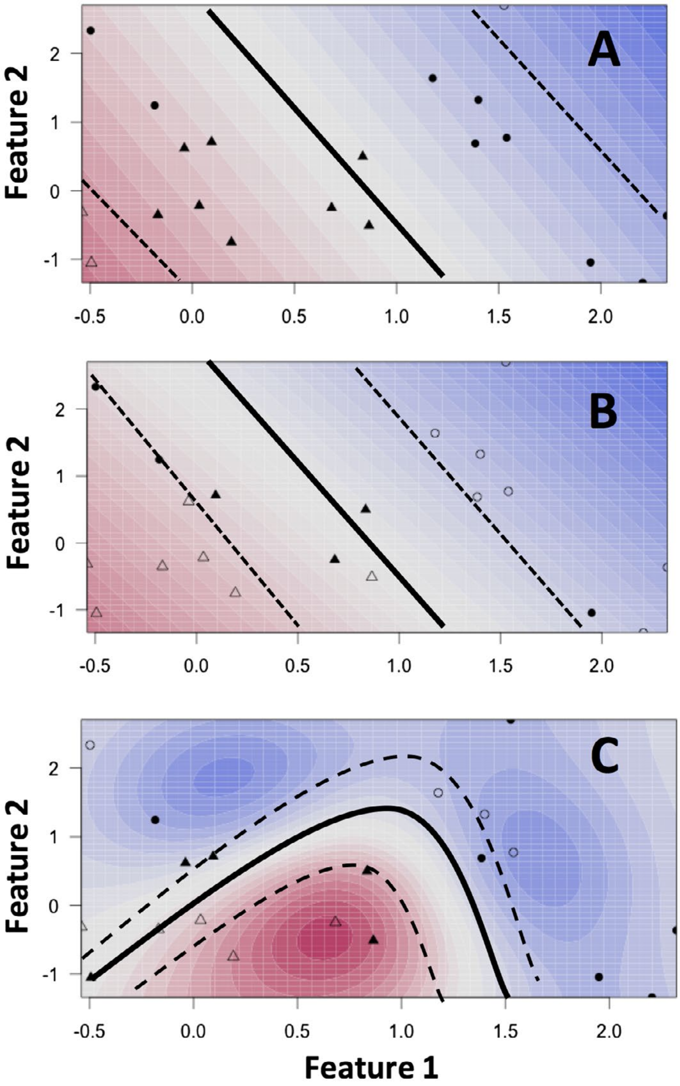


Figure 3. The influence of cost parameter and kernel type on SVM decision boundaries between two females (red and blue) in a highly simplified two-dimensional feature space showing: (A) a linear kernel when the cost parameter is low; (B) a linear kernel when the cost parameter is high; and (C) a radial kernel when the cost parameter is high.

Notes: The solid line represents the decision boundary (i.e. optimal hyperplane) and the dotted lines represent the margins; observations that lie within the margin or on the wrong side of the margin are called the support vectors (represented by the bold points) and influence the classifier. When the cost parameter is small (A), the margins will be large and there will be more support vectors, whereas if the cost parameter is high (B) there will be fewer support vectors and smaller margins. Figures made using the "kvsm" package (Karatzoglou et al. 2004).

Females with repeat recording sessions

There were four females that we believe were recorded in at least three separate recording seasons, so we assigned the two recording seasons for each female to the training data-set and the third recording to the test data-set, and rotated the training and test seasons so that all seasons were contained in the test and training data-set. This type of analysis would not be appropriate for classification using DFA, as it is a two-factorial data-set that includes 'individual' and 'recording season' as two distinct factors (Mundry and Sommer 2007). We used the 'one-against-one' approach, creating a binary classifier for each pair of females in the training set, and then calculated the accuracy of SVM model predictions for the test data-set. All SVM models were created using the 'e1071' package in the R programming environment (Meyer et al. 2017; R Development Core Team 2017).

Data availability

All data and R code for the present analyses are available as online supporting material and on GitHub (https://github.com/DenaJGibbon/MFCC-Vocal-Fingerprinting).

Results

MFCC feature extraction

We analyzed 376 calls from 33 females (median number of calls: 12; range: 3–43; recording locations shown in Figure 4). We compared two methods of MFCC feature extraction, one that averages MFCCs over all time windows, and another which divides each call into a standardized number of windows and incorporates delta-cepstral coefficients, providing information about how the calls change over time. We found that both MFCC feature extraction methods substantially improved classification accuracy when using DFA (98.4% correct identification when averaging over time windows and 98.9% when using a standardized number of time windows), compared to DFA using 23 features estimated from the spectrogram (95.7% accuracy; Clink et al. 2017).

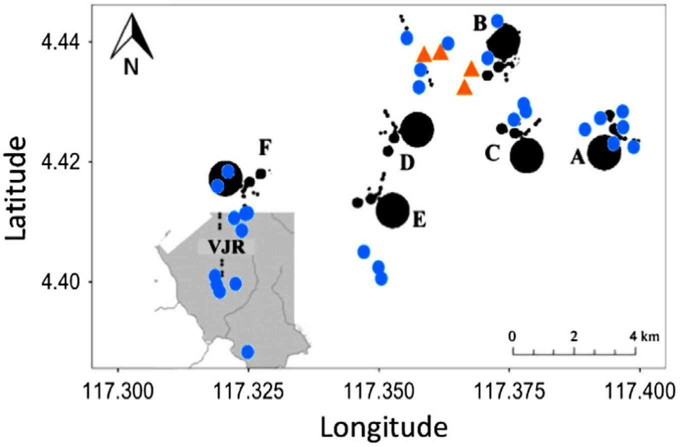


Figure 4. Approximate recording location of Bornean gibbon females in the Stability of Altered Forest Ecosystems landscape in Sabah, Malaysia. Letters A–F denote replicate fragment blocks within the SAFE landscape and VJR denotes the virgin jungle reserve (see Ewers et al. 2011 for a complete explanation of the study design).
Note: Blue circles denote recording locations for which we have a recording taken during a single season, and red triangles denote recording locations where we recorded gibbons over multiple seasons.

For our second MFCC feature extraction method, we used recursive feature elimination to identify the features with the most predictive power, and found that including the 45 highest ranked features resulted in the highest classification accuracy, but including more features did not improve our classification ability (Figure 5). The most important features for distinguishing between individuals were a combination of MFCCs and the delta coefficients. Loadings of the first five discriminant functions are available as online supporting material (Online Supporting Material Table 1).

Training and test data with random subsets of calls

We randomly assigned approximately 20% of calls to the test data-set and the remaining 80% of calls to the training data-set (which was further divided into 80% training and 20% validation) over 100 iterations for each kernel type ('linear', 'polynomial', 'radial basis' or 'sigmoid'). We found that using MFCC features averaged over time windows and an SVM with a 'sigmoid' kernel yielded the most accurate predictions (mean = 98.8% accuracy, SD = 1.2, Figure 6), but the 'radial basis' (mean = 98.7%, SD = 1.5) and 'linear' (mean = 98.6%, SD = 1.7) kernel yielded only slightly lower mean prediction accuracies. The 'polynomial' kernel yielded substantially lower prediction accuracies for our data-set (mean = 92.1%, SD = 3.5). Using MFCC features estimated for a standardized number of time windows yielded slightly lower mean classification accuracies (90.1–96.8% accuracy).

SVM classification using our full data-set

We tested the performance of SVM for classification of gibbon females using our full data-set. We found that the 'sigmoidal' kernel yielded the highest classification accuracies so we used this kernel. Using the time averaged MFCC feature extraction method we found that classification with SVM accuracy (validated using leave-one-out cross-validation) increased to 99.5% accuracy, and with the second MFCC feature extraction method (using a standardized number of time windows) the classification accuracy was 96.8%.

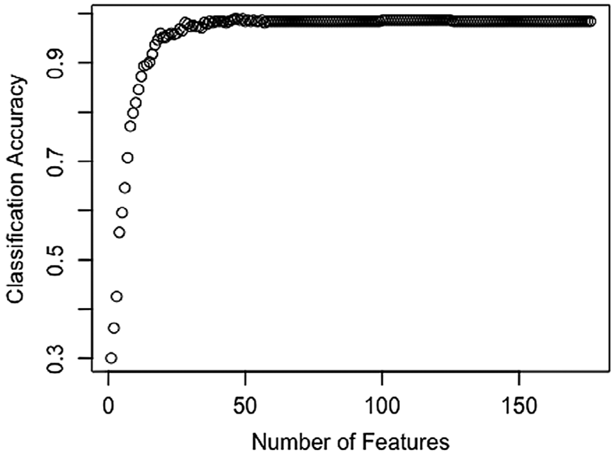


Figure 5. Plot of classification accuracy using linear discriminant analysis (validated via leave-one-out cross-validation) iteratively adding the highest to lowest ranked MFCC features.

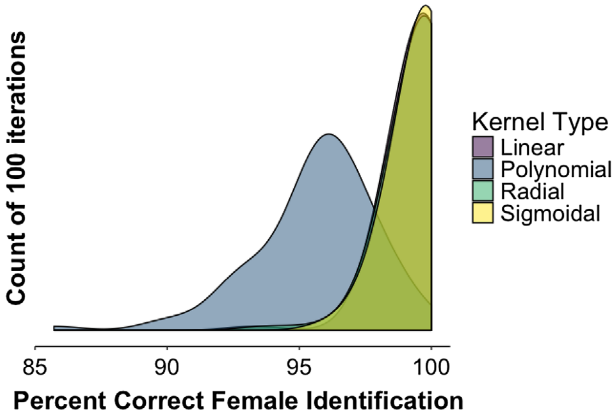


Figure 6. Density plots of the percent of correct female identification over 100 iterations using randomly selected calls to train and test the SVM for four different kernel types.

Females with repeat recording sessions

There were four females in our data-set that had repeated recording sessions over three or more recording seasons. To account for potential changes in calls over time, we trained the SVM using two recording seasons, and used the third recording season as the test data-set. We found that when we used the first and second recording seasons as the training set, and the third recording season as a test set, our classification accuracy was 94.8%. However, when we rotated the recording seasons used for training our classification accuracy ranged from 61.7 to 78.5%.

Discussion

We tested and validated a semi-automated vocal fingerprinting method to monitor female Bornean gibbons. We show that MFCC feature extraction is highly effective for distinguishing between individual Bornean gibbon females, improving accuracy of linear DFA from 95.7% with features estimated from the spectrogram to 98.9% accuracy using MFCC features. Importantly, the use of MFCC feature extraction reduces call processing time substantially compared to spectrogram feature extraction. Using a multi-class SVM and 'one-against-one' approach wherein binary classifiers were created for each pair of females with a randomly chosen subset of calls to train the SVM, we were able to identify females in the test data-set with a mean 98.9% accuracy. Using the MFCC features averaged over all the time windows, the accuracy of female identification for our entire data-set using SVM increased to 99.5% (validated using leave-one-out cross-validation). Lastly, we used our method to identify females recorded at the same recording location but during different recording seasons with limited success (mean identification accuracy of 78%), which may be related to our inability to effectively identify females recorded over subsequent years. With continually improving technology, and decreasing costs of recording devices, automated identification and classification systems for gibbon individuals have the potential to revolutionize monitoring efforts of gibbon species across SE Asia.

MFCCs

MFCCs have been used extensively as a feature extraction method in human speech recognition (Han et al. 2006; Chou et al. 2008; Muda et al. 2010; Dahake and Shaw 2016), and their success in these applications is in part because they were designed with human hearing and perception of sound in mind. Humans do not perceive changes in frequency on a linear scale; as frequencies increase (above 1000 Hz), a human's ability to detect relatively small changes in frequencies decreases (Stevens and Guirao 1962), and the Mel-scale more accurately reflects human frequency perception. Although gibbon great calls serve a different function than human speech, gibbons are primates and it seems likely that we share many similarities in sound production and perception (Belin 2006). We show here that MFCCs are more efficient and effective for classifying individual Bornean gibbon females than spectrogram feature extraction methods.

Somewhat surprisingly, averaging over the time windows yielded slightly better classification results, compared to calculating MFFCs over a standardized number of time windows, which incorporated information about changes in MFCCs over the course of a call. Regardless of the MFCC feature extraction method used, interpreting the results is somewhat challenging. MFCCs have high potential to be useful for classification problems, but since MFCCs are not on the typical linear scale used to study sounds, they are difficult to interpret (Mielke and Zuberbühler 2013), and spectrogram feature extraction methods are likely to be more informative for questions regarding the mechanism and function of call variation. For example, when calculating MFFCs over a standardized number of time windows we found that the most important feature for discriminating among females was the fourth MFCC calculated for the fourth time window. Whether variation in this feature has any biological relevance, or if gibbon individuals can detect variation in this feature remains to be determined. Nonetheless, we urge other researchers to expand and adapt the methods presented here for classification problems, particularly in field sites where long-term acoustic monitoring is feasible and individual identities are known.

Limitations

A major limitation of the present study was our inability to continually monitor groups, and our resulting inability to identify females with a high degree of certainty. Our only method to identify groups across years was based on group composition, location and characteristic behaviours of the group (i.e. long calling bouts). Therefore, the potential for misidentifying females from year to year was quite high. It is possible that our low classification accuracy of re-recorded females is due to our error in female identification, as opposed to limitations of the classification algorithm. In addition, the SAFE site around the B fragment, where we have many repeat recordings, was logged between subsequent recording seasons. Although some of the groups were still present in the same location, it is possible that there was movement of groups around the area that could have influenced our results. Also, the distance between the caller and the microphone has been shown to influence the accuracy of caller identification in orangutans (Spillmann et al. 2017), and it is possible that differences in recording distances of the focal females between subsequent seasons could have influenced our results. Despite these limitations, our mean classification accuracy across seasons (78%) was substantially better than chance (25%) and it seems likely that with a larger data-set our classification accuracy would improve. We provide these results as a test case for a new and promising method of analysis, but would not feel comfortable making management recommendations based on our results.

Future directions

Although we were able to successfully identify gibbon females using MFCC feature extraction and SVM classification algorithms, there is still much that needs to be done before vocal fingerprinting can be used to effectively monitor gibbon populations. First, our method still requires a substantial amount of pre-processing, as it requires the individual calls to be extracted from the longer recordings. There are many existing methods that have been successful in call detection, such as spectrogram cross-correlation (Munger et al. 2005), Hidden Markov Models (Spillmann et al. 2015) and SVMs (Zeppelzauer et al. 2015), and it is seems likely that these methods can be successfully adapted for detection of gibbon calls. Second, consistent, long-term monitoring is necessary to establish group territories or core areas, and more accurate methods to distinguish individuals are needed (e.g. focal follows, genetic data and/or continuous acoustic monitoring) to 'ground-truth' acoustic classification results. Third, relatively little is known about the stability of acoustic signatures over time. Feng et al. (2014) documented vocal stability in four male Cao Vit Gibbons (*Nomascus nasutus*) over 2–4 years, but more research is needed to determine if these results are generalizable to both sexes and all species of gibbons. Finally, more research needs to be done on the similarity between the calls of parents and offspring, as individual turnover may not be detected if offspring replace their parents, and there is a strong family signature in call structure.

Acknowledgements

We gratefully acknowledge Mark N. Grote, Yu Shiu and Amin Jalali for their suggestions on the analysis and fruitful conversations that resulted in this manuscript. We also gratefully acknowledge Rob Ewers of Imperial College London for providing logistical support for the completion of this project, and Henry Bernard of the Universiti of Malaysia, Sabah, for acting as our local collaborator. Permission to conduct research at the Stability of Altered Forest Ecosystems site was kindly granted by the Sabah Biodiversity Centre JKM/MBS 1000-2/2(90). We also gratefully acknowledge the input of two anonymous reviewers on earlier versions of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Funding to conduct this research was kindly provided by Primate Conservation, Inc., Cleveland Metroparks Zoo, the American Primatological Society and the Fulbright U.S. Student Program.

ORCID

Dena J. Clink http://orcid.org/0000-0003-0363-5581

References

Bartlett TQ, Light LEO, Brockelman WY. 2016. Long-term home range use in white-handed gibbons (*Hylobates lar)* in Khao Yai National Park, Thailand. Am J Primatol. 78(2):192–203. DOI:10.1002/ajp.22492.

Beigi H. 2011. Fundamentals of speaker recognition. New York, NY: Springer Science & Business Media.

Belin P. 2006. Voice processing in human and non-human primates. Phil Trans R Soc London B. 361(1476):2091–2107.

Blumstein DT, Mennill DJ, Clemins P, Girod L, Yao K, Patricelli G, Deppe JL, Krakauer AH, Clark C, Cortopassi KA, et al. 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J Appl Ecol. 48(3):758–767. DOI:10.1111/j.1365-2664.2011.01993.x.

Brockelman WY, Srikosamatara S. 1993. Estimation of density of gibbon groups by use of loud songs. Am J of Primatol. 29(2):93–108. DOI:10.1002/ajp.1350290203.

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 Recogn. 43:3846–3852. DOI:10.1016/j.patcog.2010.04.026.

Chou CH, Liu PH, Cai B. 2008. On the studies of syllable segmentation and improving MFCCs for automatic birdsong recognition. Proceedings of the 3rd IEEE Asia-Pacific Services Computing Conference, Yilan, Taiwan: APSCC 2008. p. 745–750. DOI:10.1109/APSCC.2008.6.

Clink DJ, Bernard H, Crofoot MC, Marshall AJ. 2017. Investigating individual vocal signatures and small-scale patterns of geographic variation in female Bornean gibbon (*Hylobates muelleri*) Great Calls. Int J Primatol. 38(4):656–671. DOI:10.1007/s10764-017-9972-y.

Colby J. 2011. (multiple) Support vector machine recursive feature elimination – mSVM-rfe. http://github.com/johncolby/SVM-RFE.

Cortes C, Vapnik V. 1995. Support-vector networks. Mach Learn. 20(3):273–297.

Dahake PP, Shaw K. 2016. Speaker dependent speech emotion recognition using MFCC and support vector machine. In International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT); Pune, India. p. 1080–1084.

Deecke VB, Janik VM. 2006. Automated categorization of bioacoustic signals: avoiding perceptual pitfalls. J Acoust Soc Am. 119(1):645–653. DOI:10.1121/1.2139067.

Delgado RA. 2007. Geographic variation in the long calls of male orangutans (*Pongo spp.*). Ethology. 113(5):487–498. DOI:10.1111/j.1439-0310.2007.01345.x.

Duan K, Keerthi SS, Poo AN. 2003. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing. 51:41–59. DOI:10.1016/S0925-2312(02)00601-X.

Dufour O, Artieres T, Glotin H, Giraudet P. 2014. Clusterized Mel filter cepstral coefficients and support vector machines for bird song identification. In Soundscape Semiotics - Localisation and Categorisation. InTech; Rijeka, Croatia. DOI:10.5772/56872.

Esfahanian M, Zhuang H, Erdol N. 2014. On countour-based classificaiton of dolphin whistles by type. Appl Acoust. 76:274–279. DOI:10.1121/1.4881320.

Ewers RM, Didham RK, Fahrig L, Ferraz G, Hector A, Holt RD, Kapos V, Reynolds G, Sinun W, Snaddon JL, et al. 2011. A large-scale forest fragmentation experiment: the Stability of Altered Forest Ecosystems Project. Philos Trans R Soc London Ser B, Biol Sci. 366(1582):3292–3302. DOI:10.1098/rstb.2011.0049.

Fedurek P, Zuberbühler K, Dahl CD. 2016. Sequential information in a great ape utterance. Sci Rep. 6:38226. DOI:10.1038/srep38226.

Feng J-J, Cui L-W, Ma C-Y, Fei H-L, Fan P-F. 2014. Individuality and stability in male songs of cao vit gibbons (*Nomascus nasutus)* with potential to monitor population dynamics. PLoS ONE. 9(5):e96317. DOI:10.1371/journal.pone.0096317.

Geissmann T. 2002. Duet-splitting and the evolution of gibbon songs. Biol Rev. 77(1):57–76. DOI:10.1017/S1464793101005826.

Grava T, Mathevon N, Place E, Balluet P. 2008. Individual acoustic monitoring of the European Eagle Owl *Bubo bubo*. Ibis. 150(2):279–287. DOI:10.1111/j.1474-919X.2007.00776.x.

Guo G, Li SZ. 2003. Content-based audio classification and retrieval by support vector machines. IEEE Trans Neural Network. 14(1):209–215. DOI:10.1109/TNN.2002.806626.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. Mach Learn. 46(1/3):389–422. DOI:10.1023/A:1012487302797.

Haimoff E, Tilson R. 1985. Individuality in the female songs of Wild Kloss' Gibbons (*Hylobates klossii)* on Siberut Island, Indonesia. Folia Primatol. 44(3–4):129–137. DOI:10.1159/000156207.

Han W, Chan C-F, Choy C-S, Pun K-P. 2006. An efficient MFCC extraction method in speech recognition. In 2006 IEEE International Symposium on Circuits and Systems. p. 4. Island of Kos, Greece: IEEE. DOI:10.1109/ISCAS.2006.1692543.

Heinicke S, Kalan AK, Wagner OJJ, Mundry R, Lukashevich H, Kühl HS. 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods Ecol Evol. 6(7):753–763. DOI:10.1111/2041-210X.12384.

Heller RÃ, Sander AF, Wang CW, Usman F, Dabelsteen T. 2010. Macrogeographical variability in the great call of *Hylobates agilis*: assessing the applicability of vocal analysis in studies of fine-scale taxonomy of gibbons. Am J Primatol. 72(2):142–151. DOI:10.1002/ajp.20762.

Hsu C, Lin C. 2002. A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on. 13(2):415–425. DOI:10.1109/TNN.2002.1021904.

James G, Witten D, Hastie T, Tibshirani R. 2013. Support vector machines. In An introduction to statistical learning with applications in R. New York (NY): Springer. p. 337–372.

Ji A, Johnson MT, Walsh EJ, Mcgee J, Armstrong DL. 2013. Discrimination of individual tigers (*Panthera tigris*) from long distance roars discrimination of individual tigers (*Panthera tigris*) from long distance roars. J Acoust Soc Am. 133(3):1762–1769.

Kalan AK, Mundry R, Wagner OJJ, Heinicke S, Boesch C, Kühl HS. 2015 Jul. Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. Ecol Ind. 54:217–226. DOI:10.1016/j.ecolind.2015.02.023.

Kalan AK, Piel AK, Mundry R, Wittig RM, Boesch C, Kühl HS. 2016. Passive acoustic monitoring reveals group ranging and territory use: a case study of wild chimpanzees (*Pan troglodytes*). Front Zool. 13(1):34. DOI:10.1186/s12983-016-0167-8.

Karatzoglou A, Smola A, Hornik K, Zeileis A. 2004. kernlab – An S4 Package for Kernel Methods in R. J Stat Softw. 11(9):1–20. DOI:10.1016/j.csda.2009.09.023.

Kirschel ANG, Earl DA, Yao Y, Escobar IA, Vilches E, Vallejo EE, Taylor CE. 2009. Using songs to indentify individual Mexican antthrush *Formicarius moniliger*: comparison of four classification methods. Bioacoustics 19:1–20.

Kuhn M. 2008. Caret package. J Stat Softw. 28(5):1–26.

Kumar K, Kim C, Stern RM. 2011. Delta-spectral cepstral coefficients for robust speech recognition. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings. p. 4784–4787. Prague, Czech Republic: IEEE. DOI:10.1109/ICASSP.2011.5947425.

Lee C, Lee Y, Huang R. 2006. Automatic recognition of bird songs using cepstral coefficients. J Inf Technol. 1:17–23.

Lee Y. 2010. Support vector machines for classification: a statistical portrait. In: Bang H, Zhou XK, van Epps HL, Mazumdar M, editors. Statistical methods in molecular biology. Totowa, NJ: Humana Press; p. 347–368. DOI:10.1007/978-1-60761-580-4_11.

Marler P, Hobbett L. 1975. Individuality in a long-range vocalization of wild chimpanzees. Zeitschrift für Tierpsychologie. 38(1):97–109. DOI:10.1111/j.1439-0310.1975.tb01994.x.

Merchant ND, Fristrup KM, Johnson MP, Tyack PL, Witt MJ, Blondel P, Parks SE. 2015. Measuring acoustic habitats. Methods Ecol Evol. 6:257–265. DOI:10.1111/2041-210X.12330.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. 2017. e1071: Misc functions of the department of statistics. Probability Theory Group. (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071

Mielke A, Zuberbühler K. 2013. A method for automated individual, species and call type recognition in free-ranging animals. Anim Behav 8:475–482. DOI:10.1016/j.anbehav.2013.04.017.

Muda L, Begam M, Elamvazuthi I. 2010. Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. J Comput. 2(3):2151–9617.

Mundry R, Sommer C. 2007. Discriminant function analysis with nonindependent data: consequences and an alternative. Anim Behav. 74(4):965–976. DOI:10.1016/j.anbehav.2006.12.028.

Munger LM, Mellinger DK, Wiggins SM, Moore SE, Hildebrand JA. 2005. Performance of spectrogram cross-correlation in detecting right whale calls in long-term recordings from the Bering Sea. Can Acoust. 33(2):25–34.

Murphy KP. 2012. Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press.

Noda JJ, Travieso CM, Sanchez-Rodriguez D, Dutta MK, Singh A. 2016. Using bioacoustic signals and support vector machine for automatic classification of insects. In 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). p. 656–659. Noida, India: IEEE. DOI:10.1109/SPIN.2016.7566778.

Oyakawa C, Koda H, Sugiura H. 2007. Acoustic features contributing to the individuality of wild agile gibbon (*Hylobates agilis agilis*) songs. Am J Primatol. 69(7):777–790. DOI:10.1002/ajp.20390.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: machine learning in python. J Mach. 12:2825–2830.

Picone J. 1993. 1011898294pdf. Proceedings of the IEEE. DOI:10.1520/JTE12271J.

Pimm SL, Alibhai S, Bergl R, Dehgan A, Giri C, Jewell Z, Joppa L, Kays R, Loarie S, et al. 2015. Emerging technologies to conserve biodiversity. Trends Ecol Evol. 30(11):685–696. DOI:10.1016/j.tree.2015.08.008.

R Development Core Team. 2017. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Roma G, Janer J, Kersten S, Schirosa M, Herrera P, Serra X. 2010. Ecological acoustics perspective for content-based retrieval of environmental sounds. EURASIP Journal on Audio, Speech, and Music Processing. 11:1–11. DOI:10.1155/2010/960863.

Santorelli CJ, Aureli F, Ramos-Fernández G, Schaffner CM. 2013. Individual variation of whinnies reflects differences in membership between spider monkey (*Ateles geoffroyi)* communities. Int J Primatol. 34(6):1172–1189. DOI:10.1007/s10764-013-9736-2.

Spillmann B, van Noordwijk MA, Willems EP, Mitra Setia T, Wipfli U, van Schaik CP. 2015. Validation of an acoustic location system to monitor Bornean orangutan (*Pongo pygmaeus wurmbii*) long calls. Am J Primatol. 77(7):767–776. DOI:10.1002/ajp.22398.

Spillmann B, van Schaik CP, Setia TM, Sadjadi SO. 2017. Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (*Pongo pygmaeus wurmbii*) long calls. Bioacoustics. 26(2):109–120. DOI:10.1080/09524622.2016.1216802.

Stevens SS, Guirao M. 1962. Loudness, reciprocality, and partition scales. J Acoust Soc Am. 34(9B):1466–1471. DOI:10.1121/1.1918370.

Sun G-Z, Huang B, Guan Z-H, Geissmann T, Jiang X-L. 2011. Individuality in male songs of wild black crested gibbons (*Nomascus concolor*). Am J Primatol. 73(5):431–438. DOI:10.1002/ajp.20917.

Terleph TA, Malaivijitnond S, Reichard UH. 2015Oct. Lar gibbon (*Hylobates lar*) great call reveals individual caller identity. Am J Primatol. 77:811–821. DOI:10.1002/ajp.22406.

Terry AM, Peake TM, McGregor PK, Baptista L, Gaunt S, McGregor PPK, et al. 2005. The role of vocal individuality in conservation. Front Zool. 2(1):10. DOI:10.1186/1742-9994-2-10.

Turesson HK, Ribeiro S, Pereira DR, Papa JP, De Albuquerque VHC. 2016. Machine learning algorithms for automatic classification of marmoset vocalizations. PLoS ONE. 11(9):e0163041. DOI:10.1371/journal.pone.0163041.

Venables WN, Ripley BD. 2002. Modern applied statistics with S. 4th ed. New York (NY): Springer.

Wich SA, Schel AM, de Vries H. 2008. Geographic Variation in Thomas Langur (*Presbytis thomasi*) Loud Calls. Am J Primatol. 70(6):566–574. DOI:10.1002/ajp.20527.

Wrege PH, Rowland ED, Keen S, Shiu Y. 2017. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. Methods Ecol Evol. 8:1292–1301. DOI:10.1111/2041-210X.12730.

Zeppelzauer M, Hensman S, Stoeger AS. 2015. Towards an automated acoustic detection system for free-ranging elephants. Bioacoustics. 24(1):13–29. DOI:10.1080/09524622.2014.906321.