1 Introduction

Threats to biodiversity in highly biodiverse tropical regions are increasing (Alroy 2017). Understanding the full implications of these requires frequent monitoring on large enough scale (Underwood et al. 2005, Porter et al. 2009) and over a sufficiently long period of time (Porter et al. 2005). However, at present there is a lack of sufficient effective monitoring systems (Proença et al. 2017). It is imperative that cost-effective monitoring techniques that have the potential to be implemented at large scales and over long time periods are developed, especially for crucially important highly biodiverse regions. One such approach, emerging due to technological and theoretical advancements, is the combination of machine learning and passive acoustic monitoring.

Passive acoustic monitoring (hereafter PAM) is the process of collecting acoustic data in the field using sensors such as microphones which can then be analysed at a later date. The acoustic data collected can be used to answer a number of questions relating to the ecology and distribution of species (Mellinger et al. 2011, Bader et al. 2015, Davies et al. 2016, Campos-Cerqueira & Aide 2016). This information can be used to aid in the design - location and habitat type - of protected areas (Rayment et al. 2009).

It holds promise as an efficient surveying tool for a number of reasons. When compared with traditional point count surveys, acoustic monitoring approaches reduce biases such as observer bias, detection bias (initial data collection is independent of observer skill level (Klingbeil & Willig 2015)), and temporal bias, the latter having been shown in point count studies to result in missed behaviours and underestimated population sizes (Bridges & Dorcas 2000). They also reduce levels of disturbance caused by field surveys (Kühl & Burghardt 2013), avoiding the biases that this can introduce in surveying efforts (Alldredge et al. 2007). The area being surveyed can be increased in scale for a comparatively smaller increase in cost, allowing for ecological questions to be tested on large scales (Wrege et al. 2017). The data collected forms a permanent record, which enables survey analyses to be repeatable, and any number of further analyses can be carried out (for example, testing of different ecological questions (**?**)e.g. Newson et al. 2017) or investigations into changes of community composition). Furthermore, the data can be contributed to global repositories of biodiversity information, increasing the potential for widescale monitoring and modelling (Honrado et al. 2016).It also provides the ability to monitor cryptic species/behaviours (Wrege et al. 2017), and to locate target species for e.g. pest control (Kiskin et al. 2017). Where suitable, due to its comparatively low-cost, it could be used to evaluate the effectiveness of conservation actions (Wrege et al. 2017), a crucial stage which is too often overlooked (Ferraro & Pattanayak 2006).

There are further reasons why PAM could be particularly beneficial for the surveying of highly

biodiverse regions, such as rainforests. The advantage of associated reduction in observer effort (Digby et al. 2013) is accentuated, as these areas are often remote and potentially difficult to access, allowing for data to be collected over longer time frames more easily. Acoustic monitoring approaches are much less seasonally restricted (Shonfield & Bayne 2017), which is important in tropical biomes which often have prohibitive seasonal weather, and it becomes possible to survey in areas where direct observation of species may not be feasible. Significant technological improvements in recording devices - such as increasing the storage capabilities and weatherproofing while becoming much more affordable (Fanioudakis & Potamitis 2017) - have also increased the potential and scale of PAM analyses. However, these improvements, while allowing for a lot of data to be collected very efficiently, can cause significant challenges when it comes to the use of the data.

With terrabytes of audio being generated, managing and analysying PAM data has been a significant problem (Villanueva-Rivera & Pijanowski 2012, Shonfield & Bayne 2017). Extraction of the sounds of interest requires an expert to spend a large amount of time listening to the recordings, and unquantifiable souces of bias can be introduced at this stage. As a result, it is common that only a fraction of data collected is able to be used to investigate biological hypotheses (Kobayasi & Riquimaroux 2012). There has therefore been a strong incentive to incorporate techniques from the field of machine learning, in which algorithms can be designed that are capable of automating the processing element of the task.

TOPIC SENTENCE...?

Although there is a documented lack of communication between the two fields of research (Thessen 2016), traditional machine learning (hereafter ML) techniques such as support vector machines (REF REF REF), random forest (REF REF REF), and naive bayes (REF REF REF) have been applied to bioacoustic datasets for wide a variety of taxa (see comprehensive recent review by Knight et al. (2017)). These algorithms compare key hand-designed features of sounds of interest - temporal (e.g. duration) or spectral (e.g. peak frequency) - with those in a learned sound library created using a training data set, and return the most likely match via the use of probabilistic scores (Reason et al. 2016).

Several different metrics can be then used to report the performance of an ML detection algorithm, each placing varying levels of importance on the numbers of the following: true positives (positives correctly classified), true negatives (negatives correctly classified), false positives (negatives incorrectly considered to be positive), and false negatives (positives incorrectly considered to be negative). There is a reported inconsistency between the ecology literature and the ML literature in what ecologists are referring to with certain metric names; however, the recent best-practice paper by Knight et al.

(2017), uniting standards in ML and ecology, recommends that all single-species detectors report precision (the proportion of clips reported as being positive that were actually true positives), recall (the proportion of true positives detected compared with the total number of positives to be found in the test dataset), F1 score (the harmonic mean of precision and recall), and the area under the precision-recall curve, to allow for unambiguous comparision between investigations. While reporting all of these give an overall picture of the performance of the system, different metrics may be more important for different tasks; for example, it may be more important to maximise recall value for a system that is surveying for rare species and it will be possible to spend more time manually processing the results to seperate out true positives from false positives.

While excellent metric results have been reported using these traditional methods (although classification algorithms coming even close to expert observer accuracy rates are not common (Ovaskainen et al. 2018)) few attempts have been made to combine automated detection systems with PAM data in hyperbiodiverse regions such as rainforests (Browning et al. 2017). This bias also extends availability of suitable training datasets, and has been reported as a major gap in the field at present (Browning et al. 2017).

CHALLENGES: There are a number of aspects that make ML in rainforest environments challenging: — NOISE

ATTEMPTS, HOW THEIR PROBLEM DIFFERS FROM MINE: While some precedents — THIS PAPER DID IT, GOT THESE RESULTS — THIS PAPER DID IT, GOT THESE RESULTS

- A number of these have been species-specific bespoke algorithms, but it is also possible to get off-the-shelf software.

- reason being: CHALLENGING

- While the hand-designed features are often selected intelligently using expert domain knowledge (Humphrey et al. 2013), leading to very good classification results

PROBLEMS WITH RAINFORESTS.

- - very challenging - .

difficulties: - complexity of real data. if signal overlaps with noise it can be difficult to detect, as it is a complex task to separate signal from noise (Ovaskainen et al. 2018) - variability in background levels of noise is thought to have limited the effectiveness of fully automated approaches (Heinicke et al. 2015), and rainforests are known to experience high levels of variation in noise throughout the days (Waser & Waser 1977) - traditional machine learning techniques can be very affected by poor weather conditions, with data containing wind and rain often having to be discarded (Stowell et al. 2018). clearly problematic if wanting to survey in rainforests - different species in tropics can have

very similar calls (Zamora-Gutierrez et al. 2016)

However, recent methods that automatically As this can avoid the process of dimensionality-reduction (and therefore information loss) preceding selection of hand-designed features, this technique has been shown to substantially increase the accuracy and robustness of automated detection systems GOTTOCITE...(Stowell & Plumbley 2014, Browning et al. 2017). - therefore, there it is a big deal that emerging methods which automatically select features ('feature learning') are having a great deal of success. one such method is...

- advantages of deep learning over other methods (could perhaps be very few references to recent reviews or papers that used a couple). - examples of deep learning methods used in ecology - ***find key examples of it being more successful - difficulty has been in getting labelled datasets big enough - people have been working on this in recent years, lots of work on **data augmentation** (only specific augmentation methods may be effective given the classes of this problem - see Salamon and Bello 2016 for variable effectiveness of augmentation) and **transfer learning**

This is the process of applying various transformations to the data to artificially create new samples with which to train the network. Many common augmentation techniques used in image classification cannot be applied when visually representating audio data for classification, such as the mirrored flipping of the sample (it is semantically valid to say a mirrored image of a cat is still a cat, whereas the same cannot be said for a (non-symmetric) visualisation of a sound).

To knowledge, little applied in rainforests, for example to survey primates.

- monitoring primates - some ML techniques have been applied (ref, ref, ref), but maybe couldn't work here - to my knowledge, nobody has used deep learning for primates - suitability of species for this type of monitoring approach: heavily reliant on acoustic communication due to being almost entirely arboreal, frugivorous (patchily-distributed food) with complex fission-fusion societies - have an array of calls of known meaning (but fundamentally we'll first be working with presence/absence only) - high impact, very important species for ecosytem functioning (spread loads of seeds) in incredibly biodiverse areas - conservation status?

— keep this to be like 1 sentence. spider monkey is only model.

good example of all of the above coming together for high impact conservation research is...

- I will investigate deep learning techniques - architecture design, audio preprocessing techniques and data augmentation - while developing an automated detection system for spider monkey whinnys.

following similar work done previously (e.g. link with Kahl et al. 2017, but more on this in methods e.g. setting best parameters for CNNs etc) - this type of research has been done before for a small number of other species - wider project - whole of Osa Peninsula of Costa Rica will be

acoustically-monitored, and deep learning will enable the collection of a huge amount of ecological data on the spider monkeys - (possibly even for multiple calls) - reason being for design of wildlife corridors suitable for A. geoffroyi - as habitat of suitable for this target species is known to then be of sufficient quality for a number of other threatened species - to connect the populations currently isolated on the peninsula with unnocupied suitable habitat further inland)

- sentence describing technological and theoretical advancements leading to this being an exciting time in ecology, an incredibly powerful tool for species monitoring at a critical time

CONCLUDING PART OF INTRO - clearly define aims of research project and any hypotheses tested

aim - create an automated detection and classification system for spider monkey calls, accurate enough (or with the potential to be accurate enough, if subsequently labelled examples of calls are used to supplement training) to be used as a tool in species occupancy modelling the distribution on the Osa Peninsula of Costa Rica

2 Methods

2.1 Data: collection and labelling

The original data was continuous recordings of rainforest sounds on the Osa Peninsula, a portion of which was collected in December 2017, which was then supplemented by data I collected during one-month of fieldwork in May 2018. We used AudioMoth recording devices (CITE X HILL?), which create minute-long files '.wav' files named with the time and date of recording. These devices are examples of technological advancement in recent years, as rather than costing hundreds of pounds like many devices on the market, these small programable units can be produced for around 30. We made the recordings by fastening the devices to trees for periods of approximately three days, orienting the omnidirectional microphone upwards and angled into unsheltered areas of the forest so as to give the best chance of recording clear spider monkey calls. The possibility of water damamge influenced how they were placed (for example slightly sheltered by vegetation); however, some water damamge did cause some data loss. Nonetheless, over the two recording periods we collected X hours of data, which goes towards filling the data-gap in tropical biomes as reported by (Browning et al. 2017)

A primatologist with four years of experience listening to spider monkey calls listened to X hours of the recordings, seperating out minute-long clips containing the signal of interest (a spider monkey 'whinny'). She then created label files in the software PRAAT (cite X) containing the start and end times of periods with and without the call. Using Python, I wrote functions capable of reading the

label files to then clip the audio files into three second 'positive' sections containing a call. As calls were one second long on average, with a standard deviation of X, the longest of which was 2.X seconds, I decided that a three second window was suitable. X et al reported that it was most beneficial to train their neural network detector using positives from as many different locations within the study region as possible. Due to a lot of the data labelling being done before the start of the project, most of the positives (X/124) used to train the network were from one location. However, a portion of positives added in the later stages were from different locations, X/124 from SHADY LANE, X/124 from OSA, and X/124 from Corcovado National Park (the latter not being collected by us but taken from CORNELL WEBSITE with permission of X). As this detector is intended to only be used in one region I only used positive examples from individuals in the region to train the detector (as recommended by Knight et al. (2017)), forgoing the opportunity to add positive clips from ITALIC Ateles geoffroyi recorded in Mexico.

I created the 'negative' training examples (three second clips known to not contain the signal of interest) in a three ways: (1) random sampling from regions of call-containing minute-long clips known to not contain calls; (2) carrying out a process of 'hard-negative mining' (as done by Mac Aodha et al. (2018)) in which an early-stage trained version of the detector was ran on minute-long clips that have labelled call and non-call regions, and any three-second sections classified as being postive (but known to be negative) are used as supplementary negative training examples, and (3): early-stage versions of the detector were also run on entire folders (one recording location, over a period of days, constituted one folder of files), and the same individual that originally labelled the calls listened to a large number of the positively-classified clips, separating out any false positives (clips not contain the signal of interest). I applied the hard-negative mining technique as Mac Aodha et al. (2018) reported significant improvements as a result of this training on more challenging examples.

In total, the original dataset with which to train the network consisted of 124 positive clips and an equal number of negative examples (classes balanced as done by Mac Aodha et al. (2018) and (Kiskin et al. 2017) for similar neural network binary detection problems) Where possible, for a given location, I balanced the number of negatives with the number of positives so as to not introduce any biases by overrepresenting certain locations (which may have had different levels/combinations of background noise). For locations with both hard-negative mined negatives and randomly sampled negatives, I added an equal ratio of both. As I had no negatives from one of the sites, CONSULT SOMEONE ABOUT WHETHER I SHOULD ADMIT TO THIS.

2.2 Data: preprocessing, augmentation

I applied several preprocessing stages to the positive and negative audio clips before using them to train the network, as well as experimenting with various augmentation methods. When leveraging the state-of-the-art performance of deep learning for audio classification, it is most common to convert each audio sample into a spectrogram, a visual representation of the sound enabling the problem to then be treated as an image classification task. This is done using a mathematical process, a Fourier transform, which decomposes the audio signal into its composite frequencies. The resulting image has frequencies (grouped into bins) on the y-axis, the amplitude (loudness) of the values in these bins shown as a heat level, and the x-axis representing time steps. Early machine learning research on speech processing commonly used logarithmically-scaled frequency bins, mimicking how human ears process sounds of differing frequencies, to create what is known as a mel-frequency spectrogram. This has been shown to continue to be a succesful transformation for preprocessing for deep learning methods, with many of the state-of-the-art classification approaches using it in a recent machine learning audio detection competiton (Stowell et al. 2018). Therefore I carried out this transformation on all input data to the neural network.

It is common practice in machine learning to standardise all inputs to the network, as this often helps with the learning process (expanded on below). Standardising approaches differ depending on the task. A commonly used technique in image classification is to divide all values by 255 - the maximum pixel value - to bound all values between 0 and 1. However, although I was performing an image-classification task, the amplitude values were not bounded by any definitive value. Therefore, the method I chose to investigate was to divide all amplitude values in a spectrogram by the maximum amplitude present, achieving the goal of bounding all inputs to be between 0 and 1 (pers. comm X Sethi).

A further preprocessing technique that has been shown to increase performance in tasks of this sort is to apply a function capable of 'denoising' spectrograms to increase the signal-to-noise ratio. To investigate the effectiveness of this on my problem, I chose the denoising function of Aide et al. (2013), which was used subsequently in the competition-winning binary detection system of Kahl et al. (2017). This function works by subtracting the mean amplitude of each frequency bin from all values in that bin, keeping only particularly loud signals present in the spectrogram.

As the initial training dataset size for the neural net was very small, I experimented with data augmentation. This stage was crucial as CNNs are known to require a large amount of data, e.g. even thousands of examples) to produce truly state-of-the-art results. I implemented several augmentation methods used by Kahl et al. (2017) known to be valid for use on audio data (i.e. retaining the semantic

validity of the label).

These augmentations were adding a varying amount of Gaussian noise (slight distortion) to the mel-spectrograms once generated, slightly decreasing the signal-to-noise ratio, and also producing spectrograms representing articial situations in which the spider monkey whinny overlapped with noise, such as a chirping bird or a howler monkey calling in the distance. To do the latter, I selected a number of three-second 'noise' clips, and the augmentation function would randomly select from these, add together mel-frequency spectrograms of the 'signal' and 'noise' clips, and renormalise to ensure the background noise levels had not been articially doubled.

I also developed a random crop augmentation function, in which the full positive-containing clips and the labels are used to reclip three-second positive clips, but the calls are repositioned within the window of the clip with a high probability that they are at least part-way cut off (retaining a minimum of 20% of the call within the window). This was to ensure the that network was trained on calls that had been interupted part-way through, a stage I belied to be important as the full system splits minute-long files into three-second clips for testing, increasing the possibility that any calls present will span seperate input clips.

A further key benefit of the augmentations is to increase the generalisability of the system: applying augmentations that mimic a wide variety of conditions possible in the test data (such as overlapping of signal and a prominent noise) that may well have been absent from the training data. I investigated the performance of the system with and without the augmentation stages added.

Although rarely done in the ecology machine learning literature, I applied statistical tests to determine whether the metrics for each modification (preprocessing/augmentation) were significantly different to the metric values of the control in which nothing was changed. Due to using stratified 10-fold cross-validation, each method investigated had ten of each metric value (precision, recall, and F1), reporting the performance for each separately trained model. Therefore, as the assumption of independent samples was violated (each sample was used as training data in nine of the models), I used Wilcoxon signed ranks tests over Student's t-tests as they have more statistical power when this assumption is violated (**?**).

2.3 Detector: design, optimisation and training

The deep learning classifier at the center of the overall detection system was a convolutional neural network (CNN, or ConvNet). It is known to be very challenging to build effective neural networks entirely from the beginning, and so I followed common practise and chose an architecture from a recent paper (Salamon & Bello 2017) that had achieved very good results on a similar detection

problem (sound classification using small datasets with data augmentation (PERHAPS SEE FIGURE ?? X). However, neural networks contain a number of alterable hyperparameters (see GLOSSARY for a detailed explanation of these terms) and configuring these can optimise the performance for a given problem. The hyperparameters that commonly vary between papers tackling similar detection problems include the optimiser chosen, the learning rate, batch size, type of activation layer, inclusion or exclusion of batch-normalisation layers, dropout percentage, and number of neurons in the fully-connected layers at the end of the network. Therefore using the Python package hyperas (CITE X) I wrote and implimented a script enabling a so-termed 'grid-search' of parameters whereby all possible choices to be tested are entered, and all possible combinations are tested in order to find the optimum set. The type of data (preprocessed/augmented) I used for this optimisation step was that which resulted in the best performance of the network prior to optimisation.

I trained the neural net for 40 epochs (a measure of machine learning training time, where one epoch is the number of training steps required for the network to have trained on every sample in the training set), as preliminary investigations showed that this was enough time to record the optimal perfomance of the models before overfitting occured (a common issue with small datasets in which the network begins to learn the exact patterns of the data rather than the general trends, reducing its ability to generalise to unseen data).

2.4 Detector: evaluation

To evaluate the performance of the detector, I followed the best-practise recommendations of Knight et al. (2017), and recorded the metrics recall, precision, and F1. For each of the elements I was investigating, I used stratified 10-fold cross-validation, which is regarded as the most comprehensive method of evaluating the performance of a neural net. In this method the complete dataset is split into 10 'folds' (maintaining the proportion of positives and negatives in each fold). Ten CNNs are then created, with each being trained on a different withheld portion of the overall dataset, allowing me to obtain average metric values with an indication of measure of spread.

As much more labelled positive data will be collected over the next few years as part of the wider project, I carried out a 'learning curve' investigation to see whether training on a larger number of samples (without artificially creating them using augmentation) had a noticeable increase on network performance. I ran stratified 10-fold cross-validations, randomly selecting 50, 75, 100, and then all 124 positives (balanced by the same number of randomly selected negatives), and recorded the average detector performance at each level.

- very small dataset, investigated whether increasing data increased performance - 'learning curve'

2.5 Overall system design and functionality

The overall detection system iterates over folder of one-minute long files (as produced by the AudioMoth recording devices). For each file, it splits it into twenty three-second clips, which are sequentially inputted into the trained CNN. Each resulting activation value of the last (output) layer of the network (a value between 0 and 1) is checked, and if this value is above a threshold activation value (e.g. 0.5) - specified by the user at runtime - the clip is considered to contain the signal. The metrics reported for the CNN peformance were tested with a threshold of 0.5, but I have included the option to alter the threshold when running the system as a whole to allow for the altering of the sensitivity of the system - increasing the threshold would decrease the number of false positives, but also increase the number of false negatives (calls that are more difficult to detect).

I programmed the system to give two outputs: (1) a folder containing all detected-positive three second clips (informatively labelled with the original file name of the 60-second clip, the time location within the clip they came from i.e. the start and end of three-second interval, a number representing which of the total number of detected clips from their file they were, and the activation value multiplied by 100 acting as a proxy of confidence); and (2): a summary CSV file of all detected clips, with headings file name, approx. position in recording (secs), the time and date of recording of the original file, and the confidence of the CNN's classification (again, calculated using activation value of output layer for each clip).

3 Results

Comparing effectiveness of different preprocessing techniques: - none are significant

Comparing effectiveness of augmenting vs non-augmenting: - recall was significantly, p value of

RUNTIME OF SYSTEM AS A WHOLE: (or perhaps this should be in Results): approx. one second per file, so a folder of four thousand 60-second recordings takes just over an hour. TEST THIS AGAIN WHEN I DO OVERNIGHT RUNS.

4 Discussion

SEMI-SUPERVISED LEARNING TO OBTAIN MORE DATAPOINTS:

- (Stowell et al. 2018) reported difficulties in their deep learning system detecting signals that were faint, with interference from masking noise being their second concern - for mine, highly likely monkeys will call multiple times (pers comm. Jenna)

5 Glossary

backpropagation - the progressive altering the weights (strengths) of connections in the network to lessen how incorrect the predictions of the network are optimiser - the mathematical method determining how the network should carry out batch size - how many samples are shown to the network before an instance of backpropagation occurs activation layers - a step determining, based on the input to each neuron in the neural network layer, whether those neurons should in turn fire dropout -

5.1 Why Do I Think It Didn't Work

5.2 What Would I Do To Improve It With More Time

- Other Algorithms

- genetic algorithms: more efficient way to search for hyperparameters

- I kept the classes (positive and negative) balanced, as is often done in machine learning. Ratio would be very biased. Maybe someone should alter.

- McMara **?**

I experimented with a biased training ratio which was more reflective of the true ratio of positive to negative clips in the original data

DISCUSSION POINTS

- HIGH RECALL IN AUGMENTED MEANS TIME-SAVER

- why not standardising? - why not denoising? - why no improvement in learning curve? - threshold?

Next steps to try: - certainly transfer learning - more sophisticated methods for optimising architecture. grid search only goes so far (e.g. kernels) - exploring 1D kernels - CNN for feature extraction, then added into SVM (reference) - other augmentations (pitch roll) - ran out of time - learning curve including augmented data, as it had a significantly beneficial effect on the data - unabalanced datasets - early stopping during training to prevent overfitting - enables more robust statistical testing - look into literature for RSED - increase run-time speed of system by rewriting in e.g. Cython such as

However, once networks overfit, the metric performance on the validation set will begin to fall. Therefore the preferable alternative would be to only save the weights of the model if they increase

- CHECK DISCUSSION POINTS PAGE FOR THINGS TO ADD

References

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013), 'Real-time bioacoustics monitoring and automated species identification', *PeerJ* **1**, e103.

Alldredge, M. W., Pollock, K. H., Simons, T. R., Collazo, J. A. & Shriner, S. A. (2007), 'Time-of-detection method for estimating abundance from point-count surveys', *The Auk* **124**(2), 653–664.

Bader, E., Jung, K., Kalko, E. K., Page, R. A., Rodriguez, R. & Sattler, T. (2015), 'Mobility explains the response of aerial insectivorous bats to anthropogenic habitat change in the neotropics', *Biological Conservation* **186**, 97–106.

Bridges, A. S. & Dorcas, M. E. (2000), 'Temporal variation in anuran calling behavior: implications for surveys and monitoring programs', *Copeia* **2000**(2), 587–592.

Browning, E., Gibb, R., Glover-Kapfer, P. & Jones, K. E. (2017), 'Passive acoustic monitoring in ecology and conservation'.

Campos-Cerqueira, M. & Aide, T. M. (2016), 'Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling', *Methods in Ecology and Evolution* **7**(11), 1340–1348.

Collen, A. (2012), The evolution of echolocation in bats: a comparative approach, PhD thesis, UCL (University College London).

Davies, T. E., Ruzicka, F., Lavery, T., Walters, C. L. & Pettorelli, N. (2016), 'Ultrasonic monitoring to assess the impacts of forest conversion on solomon island bats', *Remote Sensing in Ecology and Conservation* **2**(2), 107–118.

Digby, A., Towsey, M., Bell, B. D. & Teal, P. D. (2013), 'A practical comparison of manual and autonomous methods for acoustic monitoring', *Methods in Ecology and Evolution* **4**(7), 675–683.

Fanioudakis, L. & Potamitis, I. (2017), 'Deep networks tag the location of bird vocalisations on audio spectrograms', *arXiv preprint arXiv:1711.04347* .

Ferraro, P. J. & Pattanayak, S. K. (2006), 'Money for nothing? a call for empirical evaluation of biodiversity conservation investments', *PLoS biology* **4**(4), e105.

Heinicke, S., Kalan, A. K., Wagner, O. J., Mundry, R., Lukashevich, H. & Kühl, H. S. (2015), 'Assessing the performance of a semi-automated acoustic monitoring system for primates', *Methods in Ecology and Evolution* **6**(7), 753–763.

Honrado, J. P., Pereira, H. M. & Guisan, A. (2016), 'Fostering integration between biodiversity monitoring and modelling', *Journal of Applied Ecology* **53**(5), 1299–1304.

Humphrey, E. J., Bello, J. P. & LeCun, Y. (2013), 'Feature learning and deep architectures: New directions for music informatics', *Journal of Intelligent Information Systems* **41**(3), 461–481.

Kahl, S., Wilhelm-Stein, T., Hussein, H., Klinck, H., Kowerko, D., Ritter, M. & Eibl, M. (2017), 'Large-scale bird sound classification using convolutional neural networks', *Working notes of CLEF* .

Kiskin, I., Orozco, B. P., Windebank, T., Zilli, D., Sinka, M., Willis, K. & Roberts, S. (2017), 'Mosquito detection with neural networks: the buzz of deep learning', *arXiv preprint arXiv:1705.05180* .

Klingbeil, B. T. & Willig, M. R. (2015), 'Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols', *PeerJ* **3**, e973.

Knight, E., Hannah, K., Foley, G., Scott, C., Brigham, R. & Bayne, E. (2017), 'Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs', *Avian Conservation and Ecology* **12**(2).

Kobayasi, K. I. & Riquimaroux, H. (2012), 'Classification of vocalizations in the mongolian gerbil, meriones unguiculatus', *The Journal of the Acoustical Society of America* **131**(2), 1622–1631.

Kühl, H. S. & Burghardt, T. (2013), 'Animal biometrics: quantifying and detecting phenotypic appearance', *Trends in ecology & evolution* **28**(7), 432–441.

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E. et al. (2018), 'Bat detectivedeep learning tools for bat acoustic signal detection', *PLoS computational biology* **14**(3), e1005995.

Mellinger, D. K., Nieukirk, S. L., Klinck, K., Klinck, H., Dziak, R. P., Clapham, P. J. & Brandsdóttir, B. (2011), 'Confirmation of right whales near a nineteenth-century whaling ground east of southern greenland', *Biology letters* p. rsbl20101191.

Ovaskainen, O., Moliterno de Camargo, U. & Somervuo, P. (2018), 'Animal sound identifier (asi): software for automated identification of vocal animals', *Ecology letters* **21**(8), 1244–1254.

Porter, J., Arzberger, P., Braun, H.-W., Bryant, P., Gage, S., Hansen, T., Hanson, P., Lin, C.-C., Lin, F.-P., Kratz, T. et al. (2005), 'Wireless sensor networks for ecology', *AIBS Bulletin* **55**(7), 561–572.

Porter, J. H., Nagy, E., Kratz, T. K., Hanson, P., Collins, S. L. & Arzberger, P. (2009), 'New eyes on the world: advanced sensors for ecology', *BioScience* **59**(5), 385–397.

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., Böhm, M., Brummitt, N., García-Moreno, J., Gregory, R. D. et al. (2017), 'Global biodiversity monitoring: from data sources to essential biodiversity variables', *Biological Conservation* **213**, 256–263.

Rayment, W., Dawson, S. & Slooten, L. (2009), 'Use of t-pods for acoustic monitoring of cephalorhynchus dolphins: a case study with hectors dolphins in a marine protected area', *Endangered Species Research* **10**, 333–339.

Reason, P. F., Newson, S. E. & Jones, K. E. (2016), 'Recommendations for using automatic bat identification software with full spectrum recordings'. [Online; accessed 8-April-2018].
URL: *http://www.bats.org.uk/data/files/AutomaticID$_R$ecommendations$_V$ersion$_d$ate$_2$10416.pdf*

Salamon, J. & Bello, J. P. (2017), 'Deep convolutional neural networks and data augmentation for environmental sound classification', IEEE Signal Processing Letters **24***(3), 279–283.*

Shonfield, J. & Bayne, E. (2017), 'Autonomous recording units in avian ecological research: current use and future applications', Avian Conservation and Ecology **12***(1).*

Stowell, D., Stylianou, Y., Wood, M., Pamuła, H. & Glotin, H. (2018), 'Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge', arXiv preprint arXiv:1807.05812 .

Thessen, A. (2016), 'Adoption of machine learning techniques in ecology and earth science', One Ecosystem **1***, e8621.*

Underwood, N., Hambäck, P. & Inouye, B. (2005), 'Large-scale questions and small-scale data: empirical and theoretical methods for scaling up in ecology', Oecologia **145***(2), 176–177.*

Villanueva-Rivera, L. J. & Pijanowski, B. C. (2012), 'Pumilio: a web-based management system for ecological recordings', The Bulletin of the Ecological Society of America **93***(1), 71–81.*

Waser, P. M. & Waser, M. S. (1977), 'Experimental studies of primate vocalization: Specializations for long-distance propagation', Zeitschrift für Tierpsychologie **43***(3), 239–263.*

Wrege, P. H., Rowland, E. D., Keen, S. & Shiu, Y. (2017), 'Acoustic monitoring for conservation in tropical forests: examples from forest elephants', Methods in Ecology and Evolution **8***(10), 1292–1301.*

Zamora-Gutierrez, V., Lopez-Gonzalez, C., Gonzalez, M. C. M., Fenton, B., Jones, G., Kalko, E. K., Puechmaille, S. J., Stathopoulos, V. & Jones, K. E. (2016), 'Acoustic identification of mexican bats based on taxonomic and ecological constraints on call design', Methods in Ecology and Evolution **7***(9), 1082–1091.*