AVIAN CONSERVATION
& ECOLOGY

VOLUME 12, ISSUE 2, ARTICLE 14

Knight, E. C., K. C. Hannah, G. Foley, C. Scott, R. Mark Brigham, and E. Bayne. 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology* 12(2):14. https://doi.org/10.5751/ACE-01114-120214

Methodology, part of a Special Feature on Advancing bird population monitoring with acoustic recording technologies

Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs

Elly C. Knight [1,2], *Kevin C. Hannah* [3], *Gabriel J. Foley* [4], *Chris D. Scott*, *R. Mark Brigham* [4] and *Erin Bayne* [1]

[1]Bioacoustic Unit, Department of Biological Sciences, University of Alberta, [2]WildResearch Nightjar Survey, WildResearch, [3]Canadian Wildlife Service, Environment and Climate Change Canada, [4]Department of Biology, University of Regina

ABSTRACT. Automated signal recognition software is increasingly used to extract species detection data from acoustic recordings collected using autonomous recording units (ARUs), but there is little practical guidance available for ecologists on the application of this technology. Performance evaluation is an important part of employing automated acoustic recognition technology because the resulting data quality can vary with a variety of factors. We reviewed the bioacoustic literature to summarize performance evaluation and found little consistency in evaluation, metrics employed, or terminology used. We also found that few studies examined how score threshold, i.e., cut-off for the level of confidence in target species classification, affected performance, but those that did showed a strong impact of score threshold on performance. We used the lessons learned from our literature review and best practices from the field of machine learning to evaluate the performance of five readily-available automated signal recognition programs. We used the Common Nighthawk (*Chordeiles minor*) as our model species because it has simple, consistent, and frequent vocalizations. We found that automated signal recognition was effective for determining Common Nighthawk presence-absence and call rate, particularly at low score thresholds, but that occupancy estimates from the data processed with recognizers were consistently lower than from data generated by human listening and became unstable at high score thresholds. Of the five programs evaluated, our convolutional neural network (CNN) recognizer performed best, with recognizers built in Song Scope and MonitoR also performing well. The RavenPro and Kaleidoscope recognizers were moderately effective, but produced more false positives than the other recognizers. Finally, we synthesized six general recommendations for ecologists who employ automated signal recognition software, including what to use as a test benchmark, how to incorporate score threshold, what metrics to use, and how to evaluate efficiency. Future studies should consider our recommendations to build a body of literature on the effectiveness of this technology for avian research and monitoring.

Recommandations pour l'évaluation des performances de reconnaissance acoustique et application à cinq programmes courants de reconnaissance automatisée de signaux sonores

RÉSUMÉ. Les logiciels de reconnaissance automatisée de signaux sonores sont de plus en plus utilisés pour extraire les données de détection des espèces d'enregistrements acoustiques récoltés au moyen d'unités d'enregistrement autonomes (ARU en anglais), mais il existe peu d'instructions pratiques sur l'utilisation de cette technologie pour les écologistes. L'évaluation de la performance est une étape importante dans l'utilisation d'une technologie de reconnaissance acoustique automatisée parce que la qualité des résultats peut varier en fonction de divers facteurs. Nous avons passé en revue la littérature sur la bioacoustique afin de résumer les critères d'évaluation de la performance, et avons trouvé que l'évaluation, les paramètres choisis et la terminologie utilisée étaient inconsistants. Nous avons aussi constaté que peu d'études examinaient dans quelle mesure le seuil du score, c'est-à-dire la limite du niveau de confiance de la classification de l'espèce cible, influait sur la performance; toutefois, les chercheurs qui l'ont fait ont observé que le seuil du score avait un fort effet sur la performance. Nous avons appliqué les leçons apprises de notre revue de la littérature et les meilleures pratiques dans le domaine de l'apprentissage automatique pour évaluer la performance de cinq programmes de reconnaissance acoustique automatisée rapidement et facilement utilisables. Nous avons choisi l'Engoulevent d'Amérique (*Chordeiles minor*) comme espèce-modèle, parce que ses vocalisations sont simples, invariables et fréquentes. Nous avons réalisé que la reconnaissance automatisée était efficace pour déterminer la présence-absence de l'engoulevent et sa fréquence de chant, particulièrement à des seuils de score bas. Par contre, l'occurrence calculée à partir des données traitées par reconnaissance automatisée était systématiquement plus faible que celle calculée à partir des résultats issus d'experts ayant écouté les enregistrements, et elle devenait instable à des seuils de score élevés. Des cinq programmes évalués, notre reconnaisseur « Convolutional neural network » (CNN) est celui qui a le mieux performé; les reconnaisseurs intégrés dans Song Scope et MonitoR ont aussi bien performé. Les reconnaisseurs RavenPro et Kaleidoscope ont été moyennement performants et ont produit plus de faux positifs que les autres reconnaisseurs. Enfin, nous proposons six recommandations générales destinées aux écologistes qui utilisent les logiciels de reconnaissance acoustique automatisée, y compris quoi faire comme test de performances, comment incorporer un seuil de score, quels paramètres utiliser et comment en évaluer l'efficacité. Les recherches à venir devraient prendre en compte notre recommandation à l'effet de concevoir un corpus sur l'efficacité de cette technologie pour la recherche et les suivis aviaires.

Key Words: *automated signal recognition; autonomous recording unit; bioacoustics; Common Nighthawk; recognizer; signal processing*

Address of Correspondent: Elly C. Knight, Department of Biological Sciences, CW 405 Biological Sciences Bldg., University of Alberta, Edmonton, AB,, T6G 2E9 Canada, ecknight@ualberta.ca

INTRODUCTION

Autonomous acoustic sampling is a popular method of data collection for ecological research and monitoring because many species use sound as a primary method of communication (Catchpole and Slater 2008, Shonfield and Bayne 2017). In avian research, autonomous recording units (ARUs) are used to collect acoustic recordings, which can then be used for monitoring population trends (Furnas and Callas 2015), behavioral studies (Ehnes and Foote 2014), modeling habitat associations (Campos-Cerqueira and Aide 2016), and detecting rare or inconspicuous species (Homes et al. 2014, Sidie-Slettedahl et al. 2015). ARUs provide a variety of benefits over traditional human point counts, including the ability to collect data over repeat visits (Drake et al. 2016) and the flexibility to collect data at any time of day or year (Goyette et al. 2011). Additionally, recordings provide a permanent record that can reduce observer bias (Haselmayer and Quinn 2000, Campbell and Francis 2012), be used to verify identification of rare species (Swiston and Mennill 2009, Holmes et al. 2015), and analyzed later for other objectives (Luther and Derryberry 2012). ARU technology has also been widely used to study marine mammals, bats, insects, and frogs.

One of the challenges of using ARUs for ecological research and monitoring is the time required to extract target species detections from recordings (Shonfield and Bayne 2017). In response, automated signal recognition programs have been developed (e.g., de Oliveira et al. 2015, Katz et al. 2016, Nicholson 2016). Automated acoustic species recognition is the process of training a computer to detect, recognize, and evaluate the acoustic signature of a target species' vocalization. The computer runs the resultant detection algorithm (hereafter "recognizer") on recordings and evaluates the fit of the acoustic signal in the recording using a moving window. Some programs employ a single step process that runs the algorithm against every window (hereafter "moving window recognizer") while others use a two-step process that first conducts signal detection with a moving window, and then runs the algorithm only on detected signals (hereafter "signal detection recognizer"). For each window or signal evaluated, the recognizer assigns a score value, which can be interpreted as a measure of confidence that a target vocalization match has been found. The recognizer then registers a "hit" for each signal with a score above a user-defined threshold. Choosing a high score threshold will minimize false positives, i.e., false identifications, but also results in false negatives, i.e., missed detections. If the score threshold is set low by the user, this will minimize false negatives, but create many false positives. Choosing a score threshold is generally a subjective process based on the priorities of the user (Katz et al. 2016). The results of automated signal recognition are often manually validated by the user to separate true positives from false positives. Many approaches to automated acoustic species recognition or classification have been employed including random forest (Aide et al. 2013, Campos-Cerqueira and Aide 2016), Hidden Markov models (HMM; Skowronski and Harris 2006, Potamitis et al. 2014, de Oliveira et al. 2015) and/or Gaussian mixture models (GMM; Ganchev et al. 2015, Heinicke et al. 2015), binary point matching (Katz et al. 2016), spectrogram cross-correlation (Katz et al. 2016), artificial neural networks (Jennings et al. 2008, Tachibana et al. 2014, Nicholson 2016), decision trees (Digby et al. 2013), and band-pass filters (Charif et al. 2010). There are annual and one-time machine learning competitions that drive the development of new birdsong recognizer methods (Stowell et al. 2016, Goëau et al. 2017), with current state-of-the-art approaches using deep machine learning models such as convolutional neural networks to recognize multiple species from soundscape recordings (Koops et al. 2014, Joly et al. 2016, Salamon and Bello 2017). Some of these approaches are commercially or freely available, while others have been custom-built for specific research projects.

The number of tools available for automated signal recognition are rapidly increasing, yet there remains a need for a set of general recommendations for recognizer development and performance evaluation in ecology (Blumstein et al. 2011). Many authors have compared individual automated signal recognition programs to human processing to substantiate their use in ecological monitoring and research; however, authors have used a variety of metrics for evaluation, making it difficult to compare across studies. In other acoustic signal processing disciplines such as music analysis, speech classification, and machine learning, there are established best practices that ecologists can draw on to develop standardized evaluation methods (Salzberg 1997, Sokolova and Lapalme 2009, Raffel et al. 2014, Mesaros et al. 2016). Recognizer evaluation is particularly important because the quality of the species detection data produced can depend on a variety of factors including score threshold (Brauer et al. 2016), signal complexity of target species, quality of training data, spectrogram conversion, e.g., FFT size (Crump and Houlahan 2017), and recognition approach (Stowell et al. 2016). Ultimately, the appropriateness of automated acoustic species recognition will depend on the objective of the research or monitoring.

In response to this need for guidance, our goal was to provide general recommendations for recognizer performance comparison and evaluation. First, we review the literature for bioacoustic recognizer evaluation studies to confirm the need for such recommendations and identify the most commonly used metrics. Next, we conduct a recognizer evaluation based on the different approaches used in the literature to compare five Common Nighthawk (*Chordeiles minor*) recognizers: MonitoR (Katz et al. 2016), convolutional neural networks (CNN; Abadi et al. 2015), Song Scope (Wildlife Acoustics 2011), Kaleidoscope (Wildlife Acoustics 2016), and RavenPro (Charif et al. 2010). Finally, we use our literature review, results from our evaluation, and best practices from other disciplines to synthesize general evaluation recommendations for ecologists who want to use automated acoustic recognition for data processing.

LITERATURE REVIEW OF EVALUATION TOOLS

Methods

We searched for ecological journal articles, technical reports, and conference proceedings that have evaluated the performance of automated signal recognition software to scan audio recordings for species detections. We searched the literature using Web of Science and combinations of the keywords "acoustic," "classif*," "recogn*," "autom*," and "song." We found and reviewed 68 papers that used computers to automatically scan audio recordings and identify detections of target species, including birds, frogs, and mammals (Appendix 1). We performed an initial review of these papers to determine recognizer type (single or multiple species), and evaluation data type (clip or recording;

Table 1. Recognizer performance metrics used in single-species recognizer studies that assessed recognizer performance on real-field recordings. TP = true positive; FP = false positive; TN = true negative; FN = false negative; β = weighting factor used to balance the weighted average of precision and recall.

Metric	Equation	Synonyms	Papers used
Accuracy	(TP-FP)/(TP+FN)		3
F-score	$(\beta^2+1)TP/((\beta^2+1)TP+\beta^2FN+FP)$		1
False negative rate	FN/(TP+FP+TN+FN)	“missed”	3
False positive rate	FP/(TP+FP+TN+FN)		4
Negative predictive value	TN/(TN+FP)		1
Precision	TP/(TP+FP)	“positive predictive value”; “accuracy”	7
Recall	TP/(TP+FN)	“correct”; “sensitivity”; “scanning comprehensiveness”	9
ROC curve			3
Total error	(FP+FN)/(TP+FP+TN+FN)		1
True negative rate	TN/(TP+FP+TN+FN)	“specificity”	2
True positive rate	TP/(TP+FP+TN+FN)		3

Table 1). We excluded multispecies recognizers from further review because multiclass evaluation generally employs a different set of metrics than single species evaluation (Sokolova and Lapalme 2009). We also excluded papers that did not use a test dataset of unedited field recordings (see Potamitis et al. 2014) to evaluate their recognizer. The final subset included 12 single-species recognizer papers with a real-world evaluation (Appendix 1, Table 1).

Results

Benchmark

Eleven papers used human data processing as the benchmark for recognizer evaluation, and one was unclear about the benchmark used. Of the 11 that specified the benchmark, 8 used detections that had been annotated during human listening, 2 used events that had been annotated during visual spectrogram scanning, and 1 used events that had been annotated during listening and visual spectrogram scanning, i.e., two benchmarks. One paper also included a decibel level threshold as part of their benchmark (Katz et al. 2016).

Score threshold

Score threshold is a user-selected parameter that is the minimum score of any given hit reported by the recognizer. Of the 12 papers reviewed, 7 described the score threshold selected. Of those seven, four papers reported selecting a single score threshold after tests such as Youden’s J statistic (Youden 1950, Swiston and Mennill 2009, Ganchev et al. 2015, Ulloa et al. 2016, Crump and Houlahan 2017), two reported choosing low thresholds that allowed for analysis of metrics across score values (Digby et al. 2013, Katz et al. 2016), and one reported a comparison of three score thresholds (Brauer et al. 2016). Two of those seven papers also reported receiver operating characteristic (ROC) metrics (Katz et al. 2016, Ulloa et al. 2016), which incorporate scores from 0 to 1 implicitly. Of the other five papers that did not report score threshold, four mentioned score but did not report threshold used (Waddle et al. 2009, Bardeli et al. 2010, Potamitis et al. 2014, Jahn et al. 2017) and one did not mention score at all (Duan et al. 2013).

All papers that examined the performance of the recognizer across score values reported that the performance improved with increasing score. Digby et al. (2013) found that recall (true negative rate) varied from nearly 100% at high scores to 0% at low scores. Similarly, Katz et al. (2016) showed that recall and specificity (the proportion of true negatives) ranged from 0 to 1 depending on the chosen score threshold. Brauer et al. (2016) compared three different score thresholds, “low” (minimized false negatives), “medium” (balanced false negatives and positives), and “high” (minimized false positives), and found that the total error of the recognizer ranged from 30% for the low threshold to 18% for the high threshold.

Metrics

In total, 11 different metrics were used across the 12 papers reviewed (Table 1). The most frequently used metrics were recall and precision. Among the metrics used, we found a lack of standardization and clarity in the 12 papers reviewed. There was variation in the terminology used for the metrics, with synonyms for 4 of the 12 metrics, and up to 4 synonyms per metric. In particular, the term “accuracy” was used to describe precision and accuracy; however, the formula for accuracy used in the papers we reviewed differs from the formula defined in the classifier evaluation literature (Sokolova et al. 2006, Sokolova and Lapalme 2009). Furthermore, “accuracy” was undefined in one of the papers reviewed (Duan et al. 2013), so we assigned it the same mathematical formula as the other two papers that did define accuracy. Two of the papers reviewed (Bardeli et al. 2010, Brauer et al. 2016) did not cite or define the metrics used, including “total error,” which is not a widely used classifier metric, so we back-calculated the mathematical formula or assigned the metric to the common name used in the paper. The remaining nine papers either provided the mathematical formula for the metrics used, explained the metric in plain language, or provided a citation for the metric formula.

RECOGNIZER COMPARISON USING COMMON NIGHTHAWK

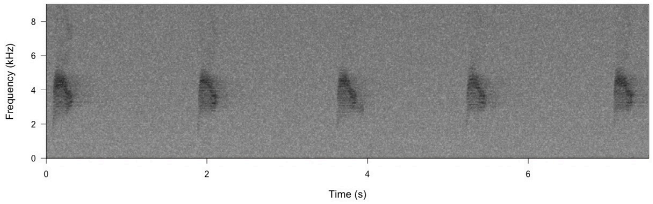
Methods

We used a standardized training dataset to allow for a comparison of four commercially or freely available recognizer programs. We also included one custom recognizer program to compare the other programs to the current state-of-the-art. To make this comparison useful to ecologists with minimal bioacoustic experience, we used an “out-of-the-box” approach by relying on the advice given by the program developer for recognizer construction and allowed ourselves 8–12 hours of learning time for each program. The exception was the custom CNN recognizer, which required us to write a Python script to carry out model training and evaluation.

Species

We used the Common Nighthawk as a model species to test single-species automated acoustic recognition software because this species has simple and consistent calls that have minimal acoustic masking from other species because nighthawks vocalize primarily at dusk and before dawn (Fig. 1). Further, the Common Nighthawk vocalizes frequently, making it an ideal candidate with which to evaluate recognizer error rates in detectability and calling rate. The development of a high quality Common Nighthawk recognizer is also a conservation priority because this species is listed as Threatened under Canada’s Species at Risk Act, and there are limited data for the species because of its crepuscular nature (Environment Canada 2016).

Fig. 1. Specotrogram of Common Nighthawk (*Chordeiles minor*) vocalizations. Spectrogram constructed with a 2048 FFT window size and Blackman-Harris window type.



Training dataset

We built Common Nighthawk recognizers for five automated signal recognition programs using vocalizations from a standardized training dataset. The standardized training dataset consisted of 400 minutes of audio data processed by human listeners: 200 minutes of audio data with Common Nighthawk detections and 200 minutes of audio data with no Common Nighthawks. The data were collected from 11 locations in south central British Columbia, Canada during the breeding season from 12 June to 14 July 2014 and 2015 at dawn or dusk. The absence data were collected from the same locations, but during times of year and day when Common Nighthawks are not active. Although Common Nighthawks produce relatively simple and consistent calls, there is variation between individuals (Armstrong 1965), so we hand-selected recordings to incorporate variation in call frequency, duration, and strength. All recordings were made using SM2+ or SM3 recorders (Wildlife Acoustics Inc.) with a bit depth of 16 bits, and a 16 kHz (2014) or 48 kHz (2015) sampling rate.

Song Scope recognizer

Song Scope is a signal detection recognizer that uses Hidden Markov models (HMMs) to maximize the probability of the arrangement of individual syllables, based on the spectral feature vectors of those syllables. We built the Song Scope recognizer iteratively, following advice available in the software manual (Wildlife Acoustics 2011). First, we extracted 100 “high-quality” calls evenly distributed across 11 locations (9–10 calls from each location). We defined “high-quality” calls as calls that were produced near the microphone, i.e., had little attenuation, and were not masked by any other acoustic signals, e.g., other birds or weather. We included approximately 0.1 seconds of silence preceding and following the vocalization. We then converted the clips to Song Scope annotations and loaded them into the Song Scope software as a single class. Common Nighthawk calls have frequencies below 8 kHz, so we set the sample rate at 20 kHz to exceed the Nyquist frequency (double the highest frequency of interest in the signal) with some headroom. We set the frequency minimum, range, max syllable, max syllable gap, max song, and dynamic range at values that maximized the detection of the 100 training annotations in the logarithmic scale with signal detection view (Appendix 2 Table A2.1). All other settings were left at default values. We reviewed each of the 100 training annotations to determine how much of each annotation was detected by Song Scope and removed any annotations where the full call was not completely detected. We replaced annotations with new annotations from the same location and reviewed those for detection completeness without adjusting the settings. We repeated this process until all 100 calls were completely detected in the logarithmic scale with signal detection view, and then generated the recognizer with the Song Scope software. The resultant recognizer had a cross training value of 77.32 +/- 5.87% (mean +/-SD) and a total training value of 77.22 ± 4.87% (Wildlife Acoustics 2011).

Kaleidoscope recognizer

Similar to Song Scope, Kaleidoscope is a signal detection recognizer that builds a classification algorithm by running individual call syllables through HMMs that maximize the probability of detecting the entire call structure. Kaleidoscope differs from Song Scope in that it uses K-means clustering of Fisher scores from a 12-state HMM to cluster all the signals detected into different classes, as opposed to only identifying the signals that match the algorithm above a user-set score threshold. We built the Kaleidoscope recognizer using the cluster analysis function following the tutorial video available from the software manufacturer for “Converting Song Scope Recognizers to Kaleidoscope Cluster-based Classifiers” (Wildlife Acoustics 2016). We exported the annotation information from the 100 Song Scope annotations into a text file as presence training data. Because Kaleidoscope performs cluster analysis, it requires at least two classes to build a recognizer, so we created an absence training class by scanning our 200-minute absence dataset with Song Scope and exporting the highest scored 100 detections into the same text file. As per the training video, we then used the Kaleidoscope software to rescan the training dataset with the training clips to create a Kaleidoscope recognizer. We set maximum cluster distance to the maximum allowable value to simulate a minimum score threshold (Appendix 2 Table A2.2). We adjusted the clustering parameters to create a two-cluster recognizer with a presence class and an absence class (Appendix 2 Table A2.2). We then processed the test dataset with the Kaleidoscope recognizer using similar signal detection parameters to the Song Scope recognizer (Appendix 2 Table A2.2). We validated only those detections that were classified as presence by the Kaleidoscope recognizer and used only hits from channel 1 to prevent duplicate hits.

MonitoR recognizer

We used the binary-point matching function in MonitoR instead of the cross-correlation approach because our initial tests suggested it was more effective for Common Nighthawk calls. The binary-point matching function in MonitoR is a template-based approach, where the program converts each cell of the spectrogram of a clip to a 1 or 0 using an amplitude cut-off. As a moving window recognizer, MonitoR then processes audio data by comparing this single-call template to each moving window of the data and scores how many cells the window has in common with the template. Multiple calls can be used to train MonitoR recognizers, but the program creates a template for each training call and scans the data once with each template, as opposed to other programs that aggregate the training calls and scan the data only once. We built the MonitoR recognizer following the training vignette (Hafner and Katz 2017). We used the MakeBinTemplate function to inspect each of the 100 training clips from the Song Scope training dataset, and adjusted the time limit, frequency limits, and amplitude cut-off manually for each template to ensure each call was completely detected (Appendix 2 Table A2.3).

CNN recognizer

Convolutional neural networks (CNNs) are a class of machine learning models that have been successfully applied in a range of domains including speech recognition and visual object recognition (LeCun et al. 2015). CNNs are a type of artificial neural network (ANN) that use moving window convolutional layers to extract features from their inputs, which makes CNNs particularly suited to acoustic detection as they can be applied directly to variable length raw audio, spectrogram inputs, or other representations of sound. ANNs have previously been used for automated acoustic signal recognition, but require that call parameters are first extracted from each acoustic signal before being passed to the ANNs for classification (e.g., Jennings et al. 2008), whereas CNNs can scan and classify the spectrograms directly. In general, the filters in convolutional layers are used to detect acoustic features while sliding over the spectrogram, or other visual input. To train a CNN as a moving window recognizer, we used a simple architecture that had multiple convolutional layers, but output a single convolutional feature map (detection function) in the final layer (Appendix 2 Table A2.4). During model training we presented short clips to the network, typically with a single Common Nighthawk call either present or absent. We used the maximum value of the detection function to classify presence/absence, which forced the model to learn a discriminative detection function. We used the TensorFlow framework and the Python API to define and train our CNN model (Abadi et al. 2015). As input to our model, we used log-power mel-scaled spectrograms calculated using librosa (McFee et al. 2017). We used rectified linear units (ReLUs) as the activation function in all layers of the network except the last, which used a sigmoid function. We trained the network for 100 epochs with a cross-entropy cost function, using minibatch stochastic gradient descent with batch size 64 and Adam optimization (Kingma and Ba 2014) with learning rate of 0.001. During model evaluation on continuous recordings, the full time-series output of the detection function was used as the recognizer score. A simple threshold-based peak-picking method was then used to extract a list of discrete detections. The CNN model required fixed length inputs during training, so we created a dataset by manually extracting 100 clips of 2-s duration from across the presence dataset and the same number from the absence dataset.

RavenPro recognizer

RavenPro uses band-pass filters, a band-limited energy detector, and an amplitude detector, to perform signal detection and identify calls of the appropriate duration within the frequency range of the target species. We followed the RavenPro 1.4 manual to configure our RavenPro recognizer (Charif et al. 2010). We extracted 100 high-quality calls (defined as above) and measured target signal parameters, i.e., frequency, duration, and separation, for each Common Nighthawk vocalization. We used the default setting for most noise estimation parameters, with adjustments made to those that increased the true positive rate (Appendix 2 Table A2.5).

Test dataset

To test the generalizability of our recognizers, we used a test dataset from a different geographic region than the training dataset. Our test dataset comprised 117 recordings of 5-min duration (2.28 GB) from 45 study sites in northwestern Ontario, Canada. The recordings were made on 13 June and 25 June 2012 at 21:00 and 22:00 when Common Nighthawks are acoustically active, and there were between 1 and 4 recordings for each of the 45 study sites. The individual recordings within the test dataset were selected randomly from a larger pool of samples, though the resulting dataset represented a gradient of low to high Common Nighthawk call density. All recordings were collected as 16-bit 16 kHz WAV files using SM2+ Songmeters (Wildlife Acoustics Inc.).

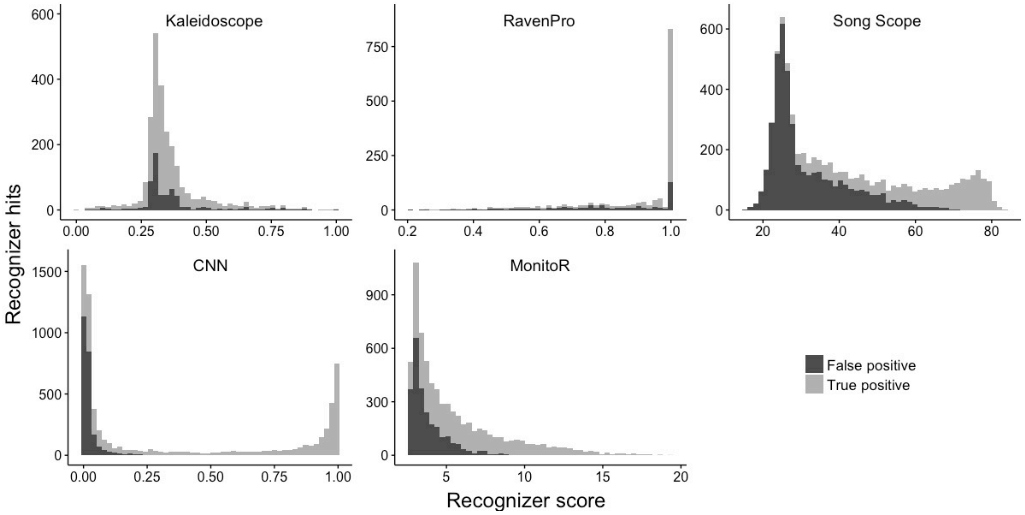
Automated processing

The test dataset was processed with each recognizer. We chose low score thresholds for each of the recognizers so that we could evaluate performance across a gradient of score thresholds (Appendix 2). We set the score threshold at 0 for the signal detection recognizers (Song Scope, Kaleidoscope, RavenPro) to allow for full analysis of the score threshold gradient. We then ran the moving window recognizers (MonitoR and CNN) with a similarly low threshold and selected the highest scored 6750 hits, which was the maximum number of hits detected by any of the signal detection recognizers (Song Scope). Without this hit threshold, both moving window recognizers would have produced as many hits as moving windows, i.e., hundreds of thousands (Fig. 2) because they have no signal detection process. We ran each recognizer with the same MacBook Pro (late 2013) with a 2.3 GHz Intel Core i7 and 16 GB 1600 MHz DDR3 of RAM. We timed the processing duration of the test dataset while no other software was running.

Benchmark development

We compared our recognizers to human listening and used the maximum number of true detections by any method as our benchmark because the recognizers detected the presence of Common Nighthawks in several recordings that human listeners had missed. Using a human listening benchmark would have decreased the presence-absence recall of those recognizers because the comparison would have been to a benchmark that included false negatives. To develop the human listening dataset, two human observers viewed and simultaneously listened to each 5-min recording in its entirety using sound visualization software

Fig. 2. Distribution of true positive and false positive recognizer hits relative to score for Common Nighthawk (*Chordeiles minor*) recognizers in five different programs. The top row programs are signal detection recognizers and the bottom row programs are moving window recognizers. Recognizer scores are the raw scores reported by the programs and are unstandardized. Kaleidoscope score is the inverse of the distance metric.



and time-stamped each Common Nighthawk vocalization using a Microsoft Access data entry form.

Statistical analysis

We referred to existing best practices in the machine learning literature and other acoustic signal detection disciplines to develop our evaluation approach (Davis and Goadrich 2006, Sokolova and Lapalme 2009, Raffel et al. 2014). We evaluated the overall performance of each of the five Common Nighthawk recognizers relative to the benchmark. We also evaluated the applied performance of each of the recognizers including presence-absence recall, occupancy modeling, and call rate correlation. All analyses were conducted in R version 3.3.1 (R Core Team 2016) with the base package, the PRROC package (Grau et al. 2015), and the ROCR package (Sing et al. 2005).

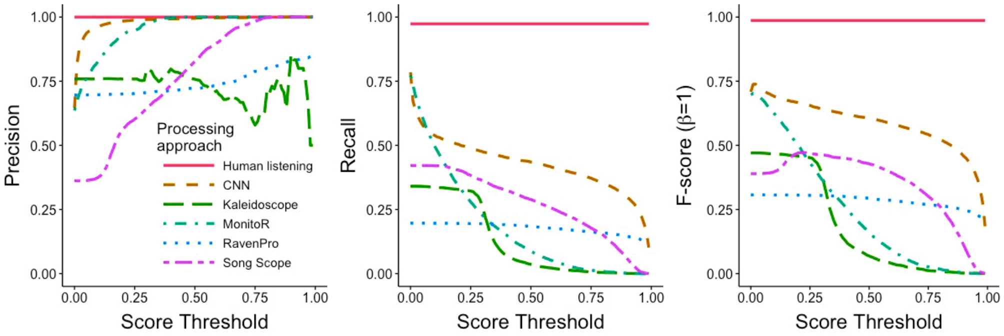
Prior to analysis, we standardized the score of each hit for each recognizer on a scale from 0 (lowest score) to 1 (highest score) to enable comparison between recognizers. We standardized the score of each hit by dividing it by the maximum score for that recognizer minus the minimum score for that recognizer. Kaleidoscope does not directly report a score, but instead uses a clustering approach to report distance between detections, so we used the inverse of the distance to cluster center as a surrogate for score. We included score threshold in our evaluation by applying a score threshold in 0.01 increments to the dataset for each recognizer before calculating each metric.

To evaluate overall performance of the recognizers, we calculated precision, recall, F-score, and area under the curve (AUC) because these metrics are suitable for one-class classifiers (recognizers trained only with examples of the target species, e.g., Song Scope, MonitoR, RavenPro) and binary classifiers (recognizers trained with examples of both the target species and nontarget species, e.g., CNN, Kaleidoscope; Sokolova et al. 2006). Precision is the proportion of recognizer hits that are true detections of the target species (Table 1). Recall is the proportion of target species vocalizations detected as hits by a recognizer (Table 1). F-score incorporates precision and recall, and allows the user to weight the relative importance of precision versus recall by setting the β value (Table 1). For AUC, we plotted precision-recall as well as ROC curves for each of the recognizers because some authors suggest precision-recall is more appropriate for recognizer performance evaluation (Davis and Goadrich 2006). We did not apply a score threshold for this evaluation because AUC incorporates score implicitly. We did not include human listening in AUC calculation because human listening detections do not have score values.

We then evaluated the applied performance of the recognizers and human listening in a presence-absence study because presence-absence data are used for a variety of applications in ecological research and monitoring. To simulate a presence-absence study and to balance sampling effort across study sites, we subsampled our test recording dataset to the first recording for each of the 45 study sites. We then determined whether the recognizer or listener accurately determined the presence or absence of a Common Nighthawk for each score threshold increment of 0.01, and then modeled this presence-absence recall with a binomial logistic regression for each processing approach. For each approach, we constructed null, first-order, second-order, and third-order polynomial models with score threshold as the covariate. We compared the four models for each approach using Akaike Information Criteria (AIC; Burnham and Anderson 2002) and selected the model with the lowest AIC score.

We also evaluated the performance of the recognizers and human listening for occupancy modeling. Occupancy modeling is a widely used application of presence/absence data that uses repeated visits to account for imperfect detection of the target species (MacKenzie et al. 2002). ARU data are particularly well-suited for occupancy modeling because they collect multiple time-

Fig. 3. Precision, recall, and F-score of Common Nighthawk (*Chordeiles minor*) call detection for automated acoustic recognition programs at varying score thresholds. Precision, recall, and F-score of human listening is provided for comparison. Precision is the proportion of recognizer hits that are true detections of the target species. Recall is the proportion of target species vocalizations detected by the recognizer. F-score combines precision and recall into a single evaluation metric.



series recordings that can be used as repeat-visit data (Shonfield and Bayne 2017). We modeled Common Nighthawk detection and occupancy for each of the recognizers and human listening using a single season occupancy model framework (MacKenzie et al. 2002) with each 5-min recording used as a separate “sampling occasion.” Prior to modeling, we removed seven study sites from the dataset for which there was only one recording because occupancy models require at least two recordings, i.e., visits, to estimate the detectability parameter. The resultant dataset comprised 38 sites. We then ran a null occupancy model with the validated recognizer data for each 0.01 score threshold for each recognizer to examine how detectability and occupancy changed with score threshold.

We also evaluated the performance of each recognizer and human listening for measuring call rate. Call rate ARU data have been used for behavioral studies (Ehnes and Foote 2014), and can be used as a proxy for abundance of some species if baseline patterns in call rates or song frequency are well known, which can in turn be used for monitoring population trends (Jeliazkov et al. 2016). We calculated the Spearman correlation coefficient between the benchmark and the call rate for each score threshold increment using the individual recording as the sampling unit.

Finally, we compared the efficiency of each of the five automated acoustic recognition programs and human listening as the time required to learn the software, build the recognizer, scan the test audio dataset, and validate the recognizer results as true or false positives. We limited learning time to 8–12 hours to develop a functional aptitude for each of the programs using our “out-of-the-box” approach. We quantified the time spent to build each recognizer, including a standardized four hours of training dataset compilation time because we used a single compiled training dataset for all five recognizers. We quantified the time required to scan by timing the computer processing of our test dataset. We quantified the time to validate by timing the validation of each of the recognizer hits and taking the mean validation time per hit. To compare the efficiency of the five recognition programs to human listening, we calculated processing time in hours per hour of audio data for a 10 hour audio dataset and a 1000 hour audio dataset. We calculated processing time as the time required to learn and build the recognizer plus time to validate the recognizer results. We did not include scanning time in our efficiency calculation because this part of the process does not require human supervision. For time to validate, we calculated the time it would take to validate the recognizer when run with a score threshold for the peak of the precision-recall curve, i.e., the maximum value of precision + recall. Finally, we calculated the audio dataset size at which the efficiency of recognizer processing becomes faster than human listening, assuming 1 hour of listening per 1 hour of audio data and 1 hour of initial learning.

Results

A total of 5556 Common Nighthawk calls were detected across the 117 five-minute recordings (mean = 152 per recording, SD = 196), which was used as the benchmark for recognizer evaluation. Common Nighthawks were detected in 85 of the 117 recordings, and at 38 of 45 sites from northwestern Ontario, Canada.

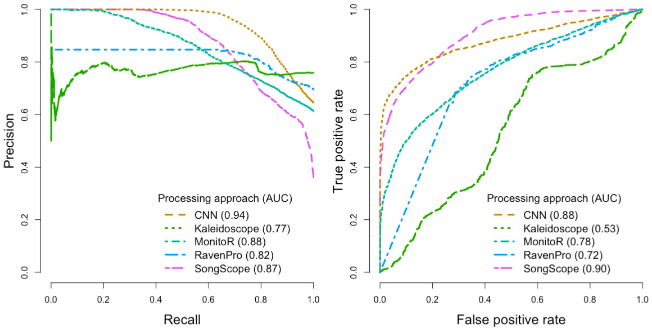
Precision, recall, and F-score

As expected, recall and F-score decreased and precision increased with increasing score threshold for all recognizers (Fig. 3). Score threshold had a minimal impact on precision and recall of the RavenPro recognizer, with impacts seen only at score thresholds above 0.7. Precision for the CNN, MonitoR, and Song Scope recognizers neared 1.0 at high score thresholds, with few false positives reported by the two moving window recognizers (CNN and MonitoR) except at low thresholds. The Kaleidoscope and RavenPro recognizers both had a precision of approximately 0.7 across most score thresholds. The CNN had the highest recall across all score thresholds, with the MonitoR recognizer also reaching a high recall of 0.75 at low score thresholds. The Song Scope recognizer recall decreased steadily from 0.42 at the lowest threshold, while the Kaleidoscope recognizer recall of 0.34 dropped off rapidly above a score threshold of 0.3. The RavenPro recognizer had relatively low recall of approximately 0.2 across all score thresholds. The F-scores of the five recognizers were similar to the recall values, with the exception of a lower F-score for the Song Scope recognizer below a score of 0.2. Human listening precision was 1.0 because we assumed that every human listener detection was a true positive; however, human listening recall was 0.97 because human listeners missed 146 of the 5556 Common Nighthawk calls detected in the test dataset.

Area under the curve

The CNN recognizer had the highest precision-recall curve AUC (0.94), followed by MonitoR (0.88), Song Scope (0.87), RavenPro (0.82), and Kaleidoscope (0.77; Fig. 4). The ranking of the top two recognizers from the ROC curve AUC was different than the precision-recall curve AUC; the SongScope recognizer had an AUC of 0.90, while the CNN had an AUC of 0.88. The ranking of the other recognizers was the same between the two AUC measures; however, the ROC AUC of the Kaleidoscope recognizer (0.53) was much lower than the precision-recall AUC (0.77).

Fig. 4. Precision-recall curve (left) and receiver operating characteristic (ROC; right) curve of Common Nighthawk (*Chordeiles minor*) call detection for automated acoustic recognition programs. AUC is area under the curve for each program.



Presence-absence

At low score thresholds, the CNN, Song Scope, and MonitoR recognizers determined Common Nighthawk presence-absence with similar recall as a human listener (95.4%; Fig. 5). At high score thresholds, only the CNN and RavenPro recognizers detected Common Nighthawk presence-absence with greater than 50% recall. As with precision and recall, score threshold had little impact on the presence-absence recall of the RavenPro recognizer. The CNN recognizer had the highest presence-absence recall of the five programs across the score threshold gradient. The CNN ($w_i = 0.95$), Kaleidoscope ($w_i = 0.92$), and Song Scope ($w_i = 0.97$) recognizers were modeled as third-order polynomials, and the MonitoR recognizer ($w_i = 0.69$) was modeled as a second-order polynomial (Appendix 3 Table A3.1). The null model with the lowest AIC score for the RavenPro recognizer was the null model ($w_i = 0.43$), suggesting that score threshold had no effect on presence-absence recall.

Occupancy

Naive occupancy of the 110 visits, i.e., recordings, included in occupancy modeling was 0.89 (34 of 38 sites). The occupancy estimate from human listening was 0.87 (SE = 0.06; Fig. 6). In

Fig. 5. Recall of five automated acoustic recognition programs for detecting Common Nighthawk (*Chordeiles minor*) presence per recording at varying score thresholds. Recall of human listening is provided for comparison. Shaded areas indicate 95% confidence intervals.

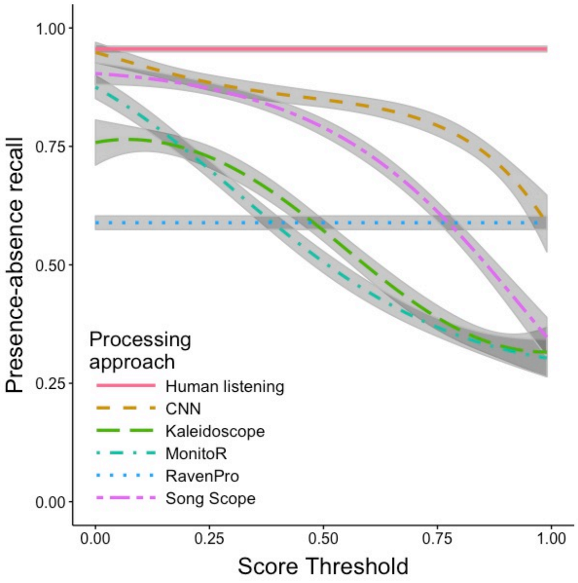


Fig. 6. Common Nighthawk (*Chordeiles minor*) occupancy and detection in null occupancy models for automated acoustic recognition programs at varying score thresholds. Occupancy and detection of human listening is provided for comparison. Shaded areas indicate 95% confidence intervals.

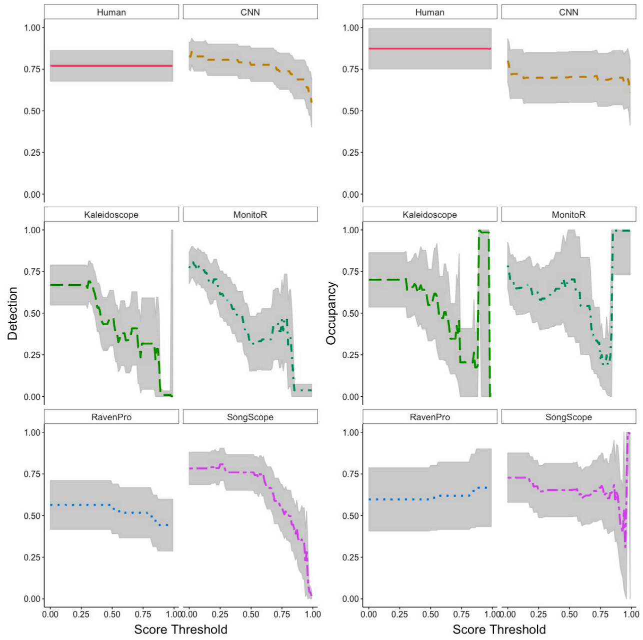


Table 2. Time in hours spent to learn each of the automated acoustic recognition programs, build a recognizer, scan audio recordings with the recognizer, and validate the recognizer output. Total times and dataset size were calculated using the number of hits produced by each recognizer when the score threshold is set to maximize accuracy.

Recognizer	Learn time	Build time	Scan time per hr audio	Validate time per hr audio	Total time per hr audio (10 hr dataset)	Total time per hr audio (1000 hr dataset)	Dataset size (hr) where recognizer is faster than human listening
Human listening	1	0	0	1	1.1	1.00	NA
CNN	24	8	0.003	0.11	3.31	0.14	36
Kaleidoscope	8	4	0.001	0.16	1.76	0.17	19
MonitoR	8	8	0.32	0.52	2.20	0.22	25
RavenPro	8	2	0.03	0.13	1.50	0.12	16
Song Scope	12	8	0.03	0.11	2.48	0.11	26

general, the occupancy estimates from recognizer data were lower than the estimate from human listening, although the occupancy estimate from the CNN recognizer (0.80) was not significantly so. The occupancy estimates from the Kaleidoscope, MonitoR, and Song Scope recognizers decreased with increasing score threshold as detection also decreased, and at high score thresholds, the estimates became unstable, varying between 0 and 1. The occupancy estimates from the CNN and the RavenPro recognizers were more stable across score thresholds, although the RavenPro estimate was much lower (0.60).

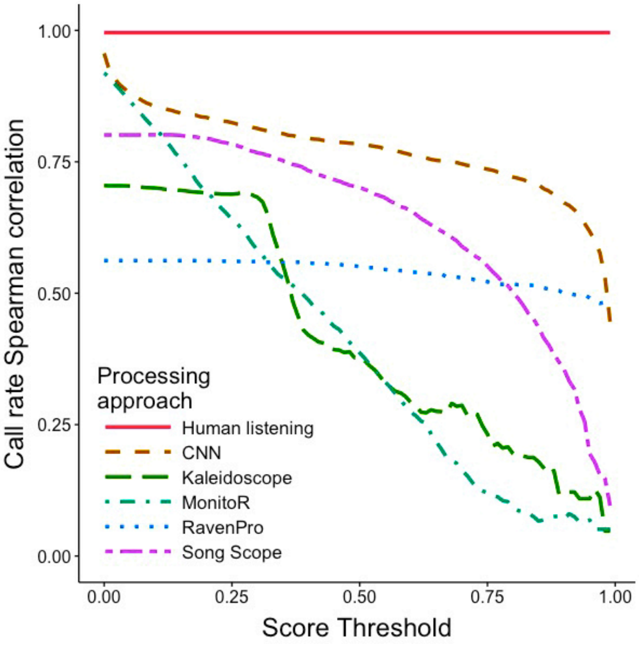
Call rate

At low score thresholds, the CNN and MonitoR call rate correlation was similar to human listening (0.96 and 0.91, respectively); however, call rate correlation of the MonitoR recognizer decreased rapidly and linearly to near 0 with increasing score threshold, while the CNN recognizer call rate correlation decreased slowly before dropping steeply at a score threshold of 0.9 (Fig. 7). The Song Scope recognizer call rate correlation was between 0.7 and 0.8 at moderate score thresholds. Call rate correlation for the RavenPro recognizer varied minimally across score thresholds (max = 0.56, min = 0.48). The Kaleidoscope call rate correlation was 0.7 and decreased steadily but irregularly after a score threshold of approximately 0.3.

Efficiency

All five of the automated signal recognition programs became faster than human listening for datasets larger than 36 hours of audio (Table 2). The CNN recognizer had the largest initial time investment, and thus had the highest processing time per hr of audio data for a small dataset (10 hours audio). For a large audio dataset (1000 hours audio) the differences between the recognizers were due primarily to differences in the number of hits at maximum precision-recall between recognizers. The Song Scope recognizer was the most efficient, while the Kaleidoscope recognizer was the slowest. Although not included in the processing time calculations, scanning time should also be included in efficiency considerations. The CNN and Kaleidoscope recognizers were the fastest to scan our test dataset, while the MonitoR recognizer was two orders of magnitude slower because this program scanned the audio dataset separately through each of the 100 templates.

Fig. 7. Spearman correlation of Common Nighthawk (*Chordeiles minor*) call rate between automated acoustic recognition programs across varying score thresholds. Correlation of call rate from human listening is provided for comparison.



EVALUATION RECOMMENDATIONS

Based on our analysis, we suggest that ecologists who use automated acoustic recognition for processing acoustic recordings follow six general recommendations. These suggestions are drawn largely from best practices in machine learning and other acoustic signal processing disciplines (Salzberg 1997, Sokolova et al. 2006, Sokolova and Lapalme 2009, Raffel et al. 2014), as well as our literature review of evaluation methods in ecology and lessons learned during our Common Nighthawk recognizer evaluation. We also suggest that ecologists familiarize themselves with general machine learning practices because there is great potential for interdisciplinary research, but a known lack of communication between the two disciplines (Thessen 2016).

Recommendation 1: Benchmark

Recognizer evaluation should employ a test dataset that differs from the training dataset to avoid "overly optimistic" results (Salzberg 1997). Within the test dataset, it is important to establish a benchmark of known target species detections to evaluate recognizer performance. We recommend human listening as a comparison benchmark; however, we remind readers that human listening is also subject to error (Bart and Schoultz 1984, McClintock et al. 2010, Brauer et al. 2016). If any false negatives in human detections are discovered during the process of reviewing recognizer detections, we recommend instead using the maximum number of target species detections detected by any method, i.e., human processing or a recognizer, as the benchmark. In our performance evaluation, there were 146 Common Nighthawk calls (2.63% of total) detected by a recognizer that were missed by human listeners. Brauer et al. (2016) also reported a 2% error rate in human identification of anuran calls, while Rydell et al. (2017) found error rates ranging from 9–22% for bat species identified by human listeners. If the target species vocalizations are susceptible to false positive identification by human observers, we recommend using a dependent double observer method when developing the benchmark to reduce the probability of misidentification (Forcey et al. 2006). Acoustic signals at farther distances (Skowronski and Brock Fenton 2009), lower sound pressure (Jahn et al. 2017), or with low signal-to-noise ratios, i.e., high levels of background noise, will be difficult to detect for both humans and recognizers, and therefore should not be excluded when preparing a benchmark (Skowronski and Harris 2006). Human listening can also be subject to observer bias (Sauer et al. 1994). Jennings et al. (2008) found that human observers with less than a single year of experience performed worse at classification than recognizers. Human annotation error can also be reduced by using the consensus from multiple observers as the benchmark dataset (e.g., Drake et al. 2016).

Recommendation 2: Score threshold

We strongly recommend that the influence of score be included in recognizer evaluation because our review showed it has a fundamental impact on recognizer performance, no matter what metric was used. Following Katz et al. (2016), we further recommend the use of score threshold instead of the reported raw scores of each detection in recognizer evaluation so that ecologists can use their evaluation results to select an optimal score threshold for data processing. We found in both our own recognizer evaluation and in our review of the literature that performance varied widely with score threshold. Furthermore, not all papers that used recognizers reported how they selected their score threshold despite the importance of this decision. Factors such as project objective, recording quality, call complexity, and signal clarity influence the choice of score threshold and the subsequent performance metrics. In our evaluation, the exception was the RavenPro recognizer, whose performance was largely unaffected by score threshold, perhaps because RavenPro is a band limited energy detector that identifies signals based only on a frequency range specification. It is possible that score threshold may be particularly important for programs with more complex classification approaches. Inclusion of a gradient of score thresholds in evaluation will facilitate selection of an appropriate score threshold for further analysis, which can be chosen based on the objectives of the project (Katz et al. 2016). We also found that some papers did not report score threshold, and we argue that it is crucial that score thresholds are explicitly reported within papers that use automated signal recognition.

Recommendation 3: Metrics

We suggest ecologists use metrics that are considered best practice in other signal processing disciplines (Sokolova and Lapalme 2009). Specifically, we suggest that four metrics always be reported for single species recognizer evaluation: (1) precision, (2) recall, (3) F-score, and (4) area under the curve (AUC). These metrics are regularly reported during classifier evaluation in other disciplines and will also allow ecologists to compare evaluation results with state-of-the-art studies in machine learning and elsewhere. Ecologists can also calculate these statistics across multiple datasets or partitioned datasets so that variance in metrics can be evaluated (Salzberg 1997) and statistical tests to compare recognizer performance can be applied (Dietterich 1998, Demšar 2006).

Precision and Recall

Precision is the proportion of recognizer hits that are true detections of the target species and is calculated as

$$Precision = \frac{tp}{tp + fp} \quad (1)$$

where *tp* is the number of true positives (detections of target species) and *fp* is the number of false positives (recognizer hits that were mislabelled as the target species).

Recall is the proportion of target species vocalizations detected as hits by a recognizer and is calculated as

$$Recall = \frac{tp}{tp + fn} \quad (2)$$

where *fn* is the number of false negatives (detections of the target species in the benchmark dataset that the recognizer missed). Precision and recall were the most commonly used metrics in our literature review and in the classification literature (Raghavan et al. 1989, Provost et al. 1998, Davis et al. 2006). Precision and recall are appropriate for signal recognition evaluation because unlike some metrics, they do not require quantification of true negatives, i.e., other species, which are not reported in single-class recognizers such as Song Scope and MonitoR. In contrast, accuracy focuses on true and false negatives and assumes that false negative and positive errors are equally likely and consequential, which is often a poor assumption in signal recognition (Provost et al. 1998). Precision and recall are also particularly appropriate when the target species is rare because a recognizer can have a high accuracy by simply predicting the target species is always absent, and the accuracy of a recognizer can be inflated by adding more negative examples to the dataset. Using precision and recall allows for direct comparison of recognizer performance with other published studies. Across the studies we reviewed, the mean recall was 0.60 and the mean precision was 0.71 (Swiston and Mennill 2009, Bardeli et al. 2010, Digby et al. 2013, Duan et al. 2013, Potamitis et al. 2014, Ganchev et al. 2015, Jahn et al. 2017). With the exception of the Kaleidoscope recognizer and the Song Scope recognizer at low score thresholds, the precision of our Common Nighthawk recognizers was above 0.71. The recall of our MonitoR and CNN recognizers reached 0.60 at low score thresholds, but the other recognizers did not.

F-score

F-score combines precision and recall into a single metric and is calculated as

$$\frac{(\beta^2+1)*precision*recall}{\beta^2*precision+recall} \qquad (3)$$

where β is a user-defined metric that allows for prioritization of precision over recall, or vice-versa. Precision and recall are evenly balanced when $\beta = 1$, precision is favored when $\beta > 1$, and recall is favored when $\beta < 1$ (Sokolova et al. 2006). We recommend that if ecologists choose to use a value for β other than 1, that they also report F-score with $\beta = 1$ to allow for comparison across studies. Situations where ecologists might consider using $\beta < 1$ include detection of rare species or situations with legal implications.

Area under the curve (AUC)

Following other acoustic signal processing disciplines, we recommend reporting the AUC of a precision-recall curve as a univariate method for comparing recognizers. Receiver operating characteristic (ROC) curve AUC is more commonly used in the classifier evaluation literature; however, precision-recall curves are more appropriate for cases with class imbalance such as recognizer evaluation (Davis and Goadrich 2006). In other words, a large quantity of false positives, as is the case for many recognizers at low score thresholds, is more accurately reflected in the AUC of a precision-recall curve than an ROC curve, and our comparison of the two approaches supports this. We therefore recommend a precision-recall AUC; however, ecologists may also want to calculate an ROC AUC for comparison with other published studies.

Recommendation 4: Application evaluation

Although overall recognizer evaluation is important, the influence of the metrics chosen can depend on the intended application for the data (Stowell et al. 2016). We therefore also recommend evaluation be done for the intended application of the resultant species detection data. Recognizer evaluation for occupancy modeling purposes is particularly important, and as our results suggest this approach becomes unreliable for recognizer data with low recall because species detection probability is too low for reliable occupancy estimates (MacKenzie et al. 2002). We also found that the shape of the curve across the score threshold gradient for all three response variables we examined (presence-absence recall, occupancy estimate, and call rate correlation) was similar to the shape of the recall curve. Future work should investigate whether the relationship between the shape of the score-recall curve is an adequate proxy for all response variables, or whether it varies depending on the detectability, call rate, and occupancy of the target species.

Recommendation 5: Regional generalizability

Geographic variation in acoustic signal is demonstrated in many bird species (Slabbekoorn and Smith 2002) and other animals that produce sound (Pröhl et al. 2006, Campbell et al. 2010, Sun et al. 2013), which is important to consider during recognizer evaluation (Gillespie et al. 2013, Russo and Voigt 2016). For simplicity, we evaluated the regional generalizability of our Common Nighthawk recognizer with a test dataset from a different region than the training data; however, in best practice, ecologists should test recognizers across multiple geographic regions. Evaluating with multiple test datasets will help ecologists determine whether a single recognizer is effective or whether regionally specific recognizers are required for their target species. For example, marine mammal classifiers have been shown to be 14.4% less accurate when tested with data from a different region than the training data (Erbs et al. 2017). For ecologists that plan to use recognizers for a single region, training and test data should be sourced from the region of interest.

Recommendation 6: Efficiency evaluation

For many ecologists, the purpose of employing an automated signal recognition approach is to increase the efficiency of audio data processing; therefore, we recommend collecting data on time spent to build and run a recognizer and validate the output. The time per hour of audio data can then be compared to other data processing approaches, including human listening. For our recognizers, we found that human listening became less efficient with datasets larger than 36 hours of audio; however, we note that using a visual scanning approach, i.e., viewing the spectrogram, instead of listening may have improved the efficiency of our human processing approach. If the automated recognizer used performs poorly, however, the manual postprocessing time required may outweigh the advantages of automation because of the time required to validate the results (Stowell et al. 2016). Digby et al. (2013) found that automated recognition (2 minutes per hour of recording) could be at least as or more efficient than manual scanning (2–5 minutes per hour of recording). Joshi et al. (2017) found that manual scanning was more time-efficient than automated signal recognition for four species of forest birds, but noted that the efficiency of a recognizer will depend on the species' vocalization characteristics, call rate, and the quality of recognizer. Indeed, human listening may be more efficient than single-species recognizers if multiple species data are needed from audio recordings; however, there are also many multispecies recognizer approaches currently under development (Stowell et al. 2016, Goëau et al. 2017). Ultimately, relative efficiency will depend on a variety of factors including score threshold, with more time required to validate recognizer output if a low score threshold is chosen to prioritize recall over precision.

DISCUSSION

Autonomous recording units (ARUs) are important tools for ecological monitoring and research because they are portable, collect data over extended periods, can be used in remote locations, are not restricted to a particular season, and the data they collect can be archived as a permanent record (Shonfield and Bayne 2017). The use of automated signal recognition for processing ARU data is growing because it can reduce the time required to process the large amounts of data; however, best practices are needed (Blumstein et al. 2011). In particular, recognizer performance evaluation is a critical step for projects that employ automated signal recognition. All recognizers misclassify detections to some extent, which can have implications for study results and may lead to poor management decisions if the results are not validated (Russo and Voigt 2016, Rydell et al. 2017). In our review of the bioacoustics literature, we found little similarity in recognizer performance evaluation between studies. Some studies reported minimal performance evaluation results, which renders the ecological results of these studies difficult to interpret. In papers that did report performance evaluation, we found an inconsistency in the evaluation terminology used and a lack of reference to the classification literature (Salzberg 1997, Davis and Goadrich 2006, Sokolova and Lapalme 2009). Given the increasing use of recognizers by ecologists, these deficiencies suggest a need for guidance on performance evaluation. We used best practices from other acoustic signal processing disciplines and our own evaluation of automated signal recognition software to provide recommendations for recognizer evaluation.

Using the Common Nighthawk as a model species, we found that a convolutional neural network (CNN) recognizer outperformed the other recognizers across all evaluations. The Song Scope and MonitoR recognizers had similar precision and recall rates to the CNN recognizer at some score thresholds. Currently, the construction of CNN recognizers requires programming expertise, but an increasing number of authors have reported success with this method for automated signal recognition (Koops et al. 2014, Salamon and Bello 2017, Salamon et al. 2017). Using our "out-of-the-box" approach, we found MonitoR and Song Scope had similar learning curves, assuming the operator is already familiar with the R programming language. At the time of writing, however, Song Scope was no longer under development or supported by the manufacturer. As the simplest automated signal recognition program, RavenPro was the easiest to learn, but the simplicity of its band-width delimitation classification approach limited its performance. Duan et al. (2013) also compared Raven Pro and Song Scope, and similarly reported a more intuitive user interface. Duan et al. (2013) also found that RavenPro had higher recall but lower precision than Song Scope. The Kaleidoscope recognizer also had low precision and recall relative to the other recognizers, with precision varying erratically across score threshold, likely because we used distance to cluster center as a surrogate for score. Rydell et al. (2017) similarly found that Kaleidoscope performed worse than other recognizers for bat call classification. We caution that our performance and efficiency evaluation of these five programs was based on a single model species with a simple, diagnostic call and little ambient masking noise and that ecologists should compare these programs for other species before choosing which program to use for audio data processing.

Overall, automated signal recognition was effective for determining Common Nighthawk presence-absence and call rate, particularly at lower score thresholds, but the occupancy estimates from the data processed with recognizers were consistently lower than derived from human listening, with the exception of the CNN recognizer. Other authors have successfully derived occupancy estimates from recognizer data that are comparable to naive occupancy (Kalan et al. 2015, Campos-Cerqueira and Aide 2016). Although ARUs can be as effective as human surveyors at detecting occurrences (Holmes et al. 2014, Kalan et al. 2015), the greater number of false negatives from an automated analysis (Brauer et al. 2016) reduces the apparent occupancy estimate for an organism at a location (MacKenzie et al. 2002). It has been suggested that the difference in recall between automated signal recognition and human listening is caused by a smaller detection radius of the recognizer relative to the human listener (Jahn et al. 2017; Knight and Bayne, *unpublished data*), which could be due to both the signal detection and classification components of the recognizer and would explain our reduced occupancy estimates. This may not be an error per se but may instead reflect the fact that more standardization is needed when using ARUs to determine the effective area being sampled (Yip et al. 2017). We also found that occupancy estimates became unstable at high score thresholds with low recall, and therefore caution against the use of occupancy models produced from recognizer data with low recall recognizers because low recall contributes to low detectability, which biases occupancy estimates (MacKenzie et a. 2002). Future research should investigate the sensitivity of occupancy modeling to this new data type.

Although automated signal recognition is effective for Common Nighthawks, there is little consensus to date on the overall effectiveness of the existing technology for avian ecological research and monitoring. Future application of our recommendations would be most useful for taxa with more complex acoustic signals, different calling rates, and in environments with varying levels of ambient noise. Thorough performance evaluation in recognizer studies following our general recommendations will contribute to building a body of literature for future meta-analysis on the overall effectiveness of automated signal recognition for wildlife monitoring and research.

Responses to this article can be read online at:
http://www.ace-eco.org/issues/responses.php/1114

Acknowledgments:

Our sincere thanks to the Subject Editor and three reviewers for their insightful comments, which greatly improved the clarity and accuracy of the manuscript. We gratefully acknowledge funding provided for the bioacoustic component of this work from the Natural Sciences and Engineering Research Council of Canada, the Alberta Biodiversity Monitoring Institute, the Ecological Monitoring Committee for the Lower Athabasca, and the Joint Oil Sands Monitoring Program. Funding for Common Nighthawk research was provided by the Natural Sciences and Engineering Research Council of Canada, Environment and Climate Change Canada, Mitacs, the Baillie Fund, Science Horizons, the Public Conservation Assistance Fund, and the TD Friends of the Environment Foundation.

LITERATURE CITED

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz,

L. Kaiser, M. Kudler, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. chuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2015. Tensorflow: large-scale machine learning on heterogeneous systems. [online] URL: https://www.tensorflow.org/

Aide, T. M., C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez. 2013. Real-time bioacoustics monitoring and automated species identification. *PeerJ* 1:e103-19. http://dx.doi.org/10.7717/peerj.103

Armstrong, J. T. 1965. Breeding home range in the nighthawk and other birds: its evolutionary and ecological significance. *Ecology* 46(5):619-629. http://dx.doi.org/10.2307/1935001

Bardeli, R., D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters* 31(12):1524-1534. http://dx.doi.org/10.1016/j.patrec.2009.09.014

Bart, J., and J. D. Schoultz. 1984. Reliability of singing bird surveys: changes in observer efficiency with avian density. *Auk* 101(2):307-318.

Blumstein, D. T., D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi, S. F. Hanser, B. McCowan, A. M. Ali, and A. N. G. Kirschel. 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology* 48:758-767. http://dx.doi.org/10.1111/j.1365-2664.2011.01993.x

Brauer, C. L., T. M. Donovan, R. M. Mickey, J. Katz, and B. R. Mitchell. 2016. A comparison of acoustic monitoring methods for common anurans of the northeastern United States. *Wildlife Society Bulletin* 40(1):140-149. http://dx.doi.org/10.1002/wsb.619

Burnham, K. P., and D. R. Anderson. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, New York, NY, USA. http://dx.doi.org/10.1007/b97636

Campbell, M., and C. M. Francis. 2012. Using microphone arrays to examine effects of observers on birds during point count surveys. *Journal of Field Ornithology* 83(4):391-402. http://dx.doi.org/10.1111/j.1557-9263.2012.00389.x

Campbell, P., B. Pasch, J. L. Pino, O. L. Crino, M. Phillips, and S. M. Phelps. 2010. Geographic variation in the songs of neotropical singing mice: testing the relative importance of drift and local adaptation. *Evolution* 64(7):1955-1972. http://dx.doi.org/10.1111/j.1558-5646.2010.00962.x

Campos-Cerqueira, M., and T. M. Aide. 2016. Improving distribution data of threatened species by combining acoustic monitoring and occupancy modelling. *Methods in Ecology and Evolution* 7(11):1340-1348. http://dx.doi.org/10.1111/2041-210X.12599

Catchpole, C. K., and P. J. B. Slater. 2008. *Bird song: biological themes and variations.* Cambridge University Press, Cambridge, UK. http://dx.doi.org/10.1017/CBO9780511754791

Charif, R. A., A. M. Waack, and L. M. Strickman. 2010. *Raven Pro 1.4 User's manual.* Cornell Lab of Ornithology, Ithaca, New York, USA. [online] URL: http://www.birds.cornell.edu/brp/raven/Raven14UsersManual.pdf

Crump, P. S., and J. Houlahan. 2017. Designing better frog call recognition models. *Ecology and Evolution* 7:3087-3099. http://dx.doi.org/10.1002/ece3.2730

Davis, J., E. S. Burnside, I. de Castro Dutra, and D. Page. 2006. View learning for statistical relational learning: with an application to mammography. Pages 233-240 *in* D. Page and J. Shavlik, editors. *Knowledge-intensive, interactive and efficient relational pattern learning*. Report prepared for AFRL/IFTD by the University of Wisconsin-Madison, USA.

Davis, J., and M. Goadrich. 2006. The relationship between Precision-Recall and ROC curves. Pages 432-437 in *Proceedings of the 23rd International Conference on Machine Learning.* 25–29 June, Pittsburgh, Pennsylvania, USA. ACM, New York, New York, USA. http://dx.doi.org/10.1145/1143844.1143874

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7:1-30.

de Oliveira, A. G., T. M. Ventura, T. D. Ganchev, J. M. de Figueiredo, O. Jahn, M. I. Marques, and K.-L. Schuchmann. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics* 98:34-42. http://dx.doi.org/10.1016/j.apacoust.2015.04.014

Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* 10(7):1895-1923. http://dx.doi.org/10.1162/089976698300017197

Digby, A., M. Towsey, B. D. Bell, and P. D. Teal. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution* 4(7):675-683. http://dx.doi.org/10.1111/2041-210X.12060

Drake, K. L., M. Frey, D. Hogan, and R. Hedley. 2016. Using digital recordings and sonogram analysis to obtain counts of Yellow Rails. *Wildlife Society Bulletin* 40(2):346-354. http://dx.doi.org/10.1002/wsb.658

Duan, S., J. Zhang, P. Roe, J. Wimmer, and X. Dong. 2013. Timed probabilistic automaton: a bridge between automatic species recognition. Pages 1519-1524 in *Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference*. 14–18 July, Bellevue, Washington, USA. AAAI, Palo Alto, California, USA.

Ehnes, M., and J. R. Foote. 2014. Comparison of autonomous and manual recording methods for discrimination of individually distinctive Ovenbird songs. *Bioacoustics* 24(2):111-121. http://dx.doi.org/10.1080/09524622.2014.994228

Environment Canada. 2016. Recovery strategy for the Common Nighthawk (*Chordeiles minor*) in Canada. *Species at Risk Act Recovery Strategy Series.* Environment Canada, Ottawa.

Erbs, F., S. H. Elwen, and T. Gridley. 2017. Automatic classification of whistles from coastal dolphins of the southern African subregion. *Journal of the Acoustical Society of America* 141:2489-2500. http://dx.doi.org/10.1121/1.4978000

Forcey, G. M., J. T. Anderson, F. K. Ammer, and R. C. Whitmore. 2006. Comparison of two double-observer point-count approaches for estimating breeding bird abundance. *Journal of Wildlife Management* 70(6):1674-1681. http://dx.doi.org/10.2193/0022-541X(2006)70[1674:COTDPA]2.0.CO;2

Furnas, B. J., and R. L. Callas. 2015. Using automated recorders and occupancy models to monitor common forest birds across a large geographic region. *Journal of Wildlife Management* 79 (2):325-337. http://dx.doi.org/10.1002/jwmg.821

Ganchev, T. D., O. Jahn, M. I. Marques, J. M. de Figueiredo, and K-L. Schuchmann. 2015. Automated acoustic detection of *Vanellus chilensis lampronotus*. *Expert Systems with Applications* 42(15-16):6098-6111. http://dx.doi.org/10.1016/j.eswa.2015.03.036

Gillespie, D., M. Caillat, J. Gordon, and P. White. 2013. Automatic detection and classification of odontocete whistles. *Journal of the Acoustical Society of America* 134:2427-2437. http://dx.doi.org/10.1121/1.4816555

Goëau, H., H. Glotin, W. P. Vellinga, R. Planqué, and A. Joly. 2017. LifeCLEF Bird Identification Task 2017. *LifeCLEF 2017 Working Notes* 1866:8.

Goyette, J. L., R. W. Howe, A. T. Wolf, and W. D. Robinson. 2011. Detecting tropical nocturnal birds using automated audio recordings. *Journal of Field Ornithology* 82(3):279-287. http://dx.doi.org/10.1111/j.1557-9263.2011.00331.x

Grau, J., I. Grosse, and J. Keilwagen. 2015. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* 31(15):2595-2597. http://dx.doi.org/10.1093/bioinformatics/btv153

Hafner, S. D., and J. Katz. 2017. A short introduction to acoustic template matching with monitoR. [online] URL: https://cran.r-project.org/web/packages/monitoR/vignettes/monitoR_QuickStart.pdf

Haselmayer, J., and J. S. Quinn. 2000. A comparison of point counts and sound recording as bird survey methods in Amazonian southeast Peru. *Condor* 102(4):887-893. http://dx.doi.org/10.1650/0010-5422(2000)102[0887:ACOPCA]2.0.CO;2

Heinicke, S., A. K. Kalan, O. J. J. Wagner, R. Mundry, H. Lukashevich, and H. S. Kühl. 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods in Ecology and Evolution* 6(7):753-763. http://dx.doi.org/10.1111/2041-210X.12384

Holmes, S. B., K. A. McIlwrick, and L. A. Venier. 2014. Using automated sound recording and analysis to detect bird species-at-risk in southwestern Ontario woodlands. *Wildlife Society Bulletin* 38(3):591-598. http://dx.doi.org/10.1002/wsb.421

Holmes, S. B., K. Tuininga, K. A. McIlwrick, M. Carruthers, and E. Cobb. 2015. Using an integrated recording and sound analysis system to search for Kirtland's Warbler (*Setophaga kirtlandii*) in Ontario. *Canadian Field-Naturalist* 129(2):115-120. http://dx.doi.org/10.22621/cfn.v129i2.1688

Jahn, O., T. D. Ganchev, M. I. Marques, and K-L. Schuchmann. 2017. Automated sound recognition provides insights into the behavioral ecology of a tropical bird. *PLoS ONE* 12:e0169041-29. http://dx.doi.org/10.1371/journal.pone.0169041

Jeliazkov, A., Y. Bas, C. Kerbiriou, J.-F. Julien, C. Penone, and I. Le Viol. 2016. Large-scale semi-automated acoustic monitoring allows to detect temporal decline of bush-crickets. *Global Ecology and Conservation* 6:208-218. http://dx.doi.org/10.1016/j.gecco.2016.02.008

Jennings, N., S. Parsons, and M. J. O. Pocock. 2008. Human vs. machine: identification of bat species from their echolocation calls by humans and by artificial neural networks. *Canadian Journal of Zoology* 86(5):371-377. http://dx.doi.org/10.1139/Z08-009

Joly, A., H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W. P. Vellinga, J. Champ, R. Planqué, S. Palazzo, and H. Müller. 2016. LifeCLEF 2016: multimedia life species identification challenges. Pages 286-310 in *International Conference of the Cross-Language Evaluation Forum for European Languages*. 5–8 September, Évora, Portugal. Springer Nature Group.

Joshi, K. A., R. A. Mulder, and K. M. C. Rowe. 2017. Comparing manual and automated species recognition in the detection of four common south-east Australian forest birds from digital field recordings. *Emu-Austral Ornithology* 117(1):1-14.

Kalan, A. K., R. Mundry, O. J. J. Wagner, S. Heinicke, C. Boesch, and H. S. Kühl. 2015. Towards the automated detection and occupancy estimation of primates using passive acoustic monitoring. *Ecological Indicators* 54:217-226. http://dx.doi.org/10.1016/j.ecolind.2015.02.023

Katz, J., S. D. Hafner, and T. Donovan. 2016. Assessment of error rates in acoustic monitoring with the R package monitoR. *Bioacoustics* 25(2):177-196. http://dx.doi.org/10.1080/09524622-.2015.1133320

Kingma, D., and J. Ba. 2014. Adam: a method for stochastic optimization. *arXiv*:1412.6980.

Koops, H. V., J. Van Balen, and F. Wiering. 2014. A deep neural network approach to the lifeclef 2014 bird task. *CLEF2014 Working Notes* 1180:634-642.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521:436-444. http://dx.doi.org/10.1038/nature14539

Luther, D. A., and E. P. Derryberry. 2012. Birdsongs keep pace with city life: changes in song over time in an urban songbird affects communication. *Animal Behaviour* 83(4):1059-1066. http://dx.doi.org/10.1016/j.anbehav.2012.01.034

MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83:2248-2255. http://dx.doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010. Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology* 91:2446-2454. http://dx.doi.org/10.1890/09-1287.1

McFee, B., M. McVicar, O. Nieto, S. Balke, C. Thome, D Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, D. Ellis, F.-R. Stoter, D. Repetto, S. Walschek, C. J. Carr, S. Ranzler, C. Keunwoo, P. Viktorin, J. F. Santos, A. Holovaty, W. Pimenta, and H. Lee. 2017. *librosa 0.5.0.* [online] URL: https://librosa.github.io/librosa/

Mesaros, A., T. Heittola, and T. Virtanen. 2016. Metrics for polyphonic sound event detection. *Applied Sciences* 6(6):162. http://dx.doi.org/10.3390/app6060162

Nicholson, D. 2016. Comparison of machine learning methods applied to birdsong element classification. Pages 57-61 in *Proceedings of the 15th Python in Science Conference.* 11–17 July, Austin, Texas, USA. SciPy.

Potamitis, I., S. Ntalampiras, O. Jahn, and K. Riede. 2014. Automatic bird sound detection in long real-field recordings: applications and tools. *Applied Acoustics* 80:1-9. http://dx.doi.org/10.1016/j.apacoust.2014.01.001

Pröhl, H., R. A. Koshy, U. Mueller, A. S. Rand, and M. J. Ryan. 2006. Geographic variation of genetic and behavioral traits in northern and southern Túngara frogs. *Evolution* 60(8):1669-1679. http://dx.doi.org/10.1111/j.0014-3820.2006.tb00511.x

Provost, F. J., T. Fawcett, and R. Kohavi. 1998. The case against accuracy estimation for comparing induction algorithms. Pages 445-453 in *Proceedings of the Fifteenth International Conference on Machine Learning.* 24–27 July, San Francisco, California, USA. Morgan Kaufmann, San Francisco, California, USA.

R Core Team. 2016. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. [online] URL: https://www.R-project.org/

Raffel, C., B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis. 2014. mir_eval: Transparent implementation of mir metrics. Pages 367-372 *in* Proceedings of the 15th International Society for Music Information Retrieval Conference. 27–31 October, Taipei, Taiwan.

Raghavan, V., P. Bollmann, and G. S. Jung. 1989. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems* 7(3):205-229. http://dx.doi.org/10.1145/65943.65945

Russo, D., and C. C. Voigt. 2016. The use of automated identification of bat echolocation calls in acoustic monitoring: a cautionary note for a sound analysis. *Ecological Indicators* 66:598-602. http://dx.doi.org/10.1016/j.ecolind.2016.02.036

Rydell, J., S. Nyman, J. Eklöf, G. Jones, and D. Russo. 2017. Testing the performances of automated identification of bat echolocation calls: a request for prudence. *Ecological Indicators* 78:416-420. http://dx.doi.org/10.1016/j.ecolind.2017.03.023

Salamon, J., and J. P. Bello. 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters* 24(3):279-283. http://dx.doi.org/10.1109/LSP.2017.2657381

Salamon, J., J. P. Bello, A. Farnsworth, and S. Kelling. 2017. Fusing shallow and deep learning for bioacoustic bird species classification. Pages 141-145 *in* 2017 IEEE International Conference on *Acoustics, Speech and Signal Processing (ICASSP).* 5-9 March, New Orleans, Louisiana, USA. http://dx.doi.org/10.1109/ICASSP.2017.7952134

Salzberg, S. L. 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery* 1:317-328. http://dx.doi.org/10.1023/A:1009752403260

Sauer, J. R., B. G. Peterjohn, and W. A. Link. 1994. Observer differences in the North American breeding bird survey. *Auk* 111 (1):50-62. http://dx.doi.org/10.2307/4088504

Shonfield, J., and E. M. Bayne. 2017. Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology* 12(1):14. http://dx.doi.org/10.5751/ACE-00974-120114

Sidie-Slettedahl, A. M., K. C. Jensen, R. R Johnson, T. W. Arnold, J. E. Austin, and J. D. Stafford. 2015. Evaluation of autonomous recording units for detecting 3 species of secretive marsh birds. *Wildlife Society Bulletin* 39(3):626-634. http://dx.doi.org/10.1002/wsb.569

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics* 21 (20):3940-3941. http://dx.doi.org/10.1093/bioinformatics/bti623

Skowronski, M. D., and M. Brock Fenton. 2009. Detecting bat calls: an analysis of automated methods. *Acta Chiropterologica* 11(1):191-203. http://dx.doi.org/10.3161/150811009X465811

Skowronski, M. D., and J. G. Harris. 2006. Acoustic detection and classification of microchiroptera using machine learning: lessons learned from automatic speech recognition. *Journal of the Acoustical Society of America* 119(3):1817-1833. http://dx.doi.org/10.1121/1.2166948

Slabbekoorn, H., and T. B. Smith. 2002. Bird song, ecology and speciation. *Philosophical Transactions of the Royal Society of London B* 257:493-503. http://dx.doi.org/10.1098/rstb.2001.1056

Sokolova, M., N. Japkowicz, and S. Szpakowicz. 2006. Beyond accuracy, F-score, and ROC: a family of discriminant measures for performance evaluation. Pages 1015-1021 *in* A. Sattar and B. Kang, editors. *AI 2006: Advances in Artificial Intelligence.* Lecture Notes in Computer Science, vol 4304. Springer, Berlin, Germany. http://dx.doi.org/10.1007/11941439_114

Sokolova, M., and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45(4):427-437. http://dx.doi.org/10.1016/j.ipm.2009.03.002

Stowell, D., M. Wood, Y. Stylianou, and H. Glotin. 2016. Bird detection in audio: a survey and a challenge. *2016 IEEE International Workshop on Machine Learning for Signal Processing.* 13–16 Sept, Salerno, Italy. arXiv:1608.03417 http://dx.doi.org/10.1109/MLSP.2016.7738875

Swiston, K. A., and D. J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-Billed, and putative Ivory-billed Woodpeckers. *Journal of Field Ornithology* 80(1):42-50. http://dx.doi.org/10.1111/j.1557-9263.2009.00204.x

Sun, K., L. Luo, R. T. Kimball, X. Wei, L. Jin, T. Jiang, G. Li, and J. Feng. 2016. Geographic variation in the acoustic traits of greater horseshoe bats: testing the importance of drift and ecological selection in evolutionary processes. *PLoS ONE* 8(8): e70368. http://dx.doi.org/10.1371/journal.pone.0070368

Tachibana, R. O., N. Oosugi, and K. Okanoya. 2014. Semi-automatic classification of birdsong elements using a linear support vector machine. *PLoS ONE* 9:e92584. http://dx.doi.org/10.1371/journal.pone.0092584

Thessen, A. 2016. Adoption of machine learning techniques in ecology and earth science. *One Ecosystem* 1:e8621. http://dx.doi.org/10.3897/oneeco.1.e8621

Ulloa, J. S., A. Gasc, P. Gaucher, T. Aubin, M. Réjou-Méchain, and J. Sueur. 2016. Screening large audio datasets to determine the time and space distribution of Screaming Piha birds in a tropical forest. *Ecological Informatics* 31:91-99. http://dx.doi.org/10.1016/j.ecoinf.2015.11.012

Waddle, J. H., T. F. Thigpen, and B. M. Glorioso. 2009. Efficacy of automatic vocalization recognition software for anuran monitoring. *Herpetological Conservation and Biology* 4 (3):384-388.

Wildlife Acoustics. 2011. *Song Scope Bioacoustics Software Version 4.0 Documentation.* Wildlife Acoustics, Maynard, Massachusetts, USA. [online] URL: https://songsleuth.com/images/documentation/Song-Scope-Users-Manual.pdf

Wildlife Acoustics. 2016. *Converting Song Scope recognizer to Kaleidoscope Cluster-based Classifiers*. Wildlife Acoustics, Maynard, Massachusetts, USA. [online] URL: https://www.wildlifeacoustics.com/products/kaleidoscope-software-acoustic/tutorial-videos/714-converting-song-scope-recognizers-to-kaleidoscope-cluster-based-classifiers

Yip, D. A., L. Leston, E. M. Bayne, P. Sólymos, and A. Grover. 2017. Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data. *Avian Conservation and Ecology* 12(1):11. http://dx.doi.org/10.5751/ACE-00997-120111

Youden, W. J. 1950. Index for rating diagnostic tests. *Cancer* 3:32-35. http://dx.doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Editor-in-Chief: Keith A.Hobson
Subject Editor: Steven L.Van Wilgenburg

Sponsored by the Society of Canadian Ornithologists and Bird Studies Canada
Parrainée par la Société des ornithologistes du Canada et Études d`oiseaux Canada

SCO - SOC

BIRD STUDIES ÉTUDES D'OISEAUX CANADA

Table A1.1. Acoustic classification or recognition articles for a review of automated signal recognition assessment methods. Articles in bold were selected for detailed review because they used single species recognizers and assessed recognizer performance with a field recording test dataset.

Reference	Single or multi species recognizer	Test dataset type
Acevedo, M. A., C. J. Corrada-Bravo, H. Corrada-Bravo, L. J. Villanueva-Rivera, and T. M. Aide. 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics* 4:206–214. http://dx.doi.org/10.1016/j.ecoinf.2009.06.005	Multi	Field recording
Agranat, I. 2009. *Automatically identifying animal species from their vocalizations*. Wildlife Acoustics, Concord, MA, USA. [online] URL: https://www.wildlifeacoustics.com/images/documentation/Automatically-Identifying-Animal-Species-from-their-Vocalizations.pdf	Single	Clip
Aide, T. M., C. Corrada-Bravo, M. Campos-Cerqueira, C. Milan, G. Vega, and R. Alvarez. 2013. Real-time bioacoustics monitoring and automated species identification. *PeerJ* 1:e103. http://dx.doi.org/10.7717/peerj.103	Multi	Field recording
Anderson, S. E., A. S. Dave, and D. Margoliash. 1996. Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America* 100:1209–1219. http://dx.doi.org/10.1121/1.415968	Multi	Captive recording
Arencibia, J. J. N, C. M. Travieso, and D. Sánchez-Rodríguez. 2015. Automatic classification of frogs calls based on fusion of features and SVM. Pages 59-63 *in 2015 Eighth International Conference on Contemporary Computing (IC3)*. 20-22 August, Noida, India. http://dx.doi.org/10.1109/ic3.2015.7346653	Single	Clip
Armitage, D. W., and H. K. Ober. 2010. A comparison of supervised learning techniques in the classification of bat echolocation calls. *Ecological Informatics* 5:465–473. http://dx.doi.org/10.1016/j.ecoinf.2010.08.001	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Bang, A. V., and P. P. Rege. 2014. Classification of bird species based on bioacoustics. *International Journal of Image Processing Techniques* 1(1):6-10.	Multi	Clip
Bardeli, R., D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters* 31(12):1524-1534. http://dx.doi.org/10.1016/j.patrec.2009.09.014	**Single**	**Field recording**
Bedoya, C., C. Isaza, J. M. Daza, and J. D. López. 2014. Automatic recognition of anuran species based on syllable identification. *Ecological Informatics* 24:1–11. http://dx.doi.org/10.1016/j.ecoinf.2014.08.009	Multi	Clip
Belyaeva, N. A., K. L. Yash-yee, K. M. Smartc, and E. Ribeirod. (n.d.). WhatFrog: A Comparison of Classification Algorithms for Automated Anuran Recognition. [online] URL: http://www.amalthea-reu.org/pubs/amalthea_tr_2014_02.pdf	Single	Clip
Brauer, C. L., T. M. Donovan, R. M. Mickey, J. Katz, and B. R. Mitchell. 2016. A comparison of acoustic monitoring methods for common anurans of the northeastern United States. *Wildlife Society Bulletin* 40(1):140-149. http://dx.doi.org/10.1002/wsb.619	**Single**	**Field recording**
Briggs, F., B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts. 2012. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America* 131:4640–4650. http://dx.doi.org/10.1121/1.4707424	Multi	Clip
Cakir, E., S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen. 2017. Convolutional recurrent neural networks for bird audio detection. Pages 1744-1748 *in 2017 25th European Signal Processing Conferences (EUPISCO)*. 28 August – 2 September, Kos Island, Greece. http://dx.doi.org/10.23919/EUSIPCO.2017.8081508	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Chesmore, E. D., and E. Ohya. 2007. Automated identification of field-recorded songs of four British grasshoppers using bioacoustic signal recognition. *Bulletin of Entomological Research* 94:1–13. http://dx.doi.org/10.1079/BER2004306	Multi	Clip
Chu, W., and D. T. Blumstein. 2011. Noise robust bird song detection using syllable pattern-based hidden Markov models. Pages 345–348 *in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 22-27 May, Prague, Czech Republic. http://dx.doi.org/10.1109/ICASSP.2011.5946411	Single	Clip
Colbert, D. S., J. A. Ruttinger, M. Streich, M. Chamberlain, L. M. Conner, and R. J. Warren. 2015. Application of autonomous recording units to monitor gobbling activity by Wild Turkey. *Wildlife Society Bulletin* 39:757–763. http://dx.doi.org/10.1002/wsb.577	Single	Field recording: no benchmark
Colonna, J. G., M. Cristo, M. S. Júnior, and E. F. Nakamura. 2015. An incremental technique for real-time bioacoustic signal segmentation. *Expert Systems with Applications* 42:7367–7374. http://dx.doi.org/10.1016/j.eswa.2015.05.030	N/A	Field recording
Crump, P. S., and J. Houlahan. 2017. Designing better frog call recognition models. *Ecology and Evolution* 7:3087-3099. http://dx.doi.org/10.1002/ece3.2730	**Single**	**Field recording**
Digby, A., M. Towsey, B. D. Bell, and P. D. Teal. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution* 4(7):675-683. http://dx.doi.org/10.1111/2041-210X.12060	**Single**	**Field recording**
Doležel, P., M. Mariška, and I. Taufer. 2015. Possibilities of feedforward multilayer neural network classifier as a detector of pest birds in vineyards. *International Journal of Engineering Research in Africa* 18:184–191. http://dx.doi.org/10.4028/www.scientific.net/JERA.18.184	Multi	Clip
Dong, X., M. Towsey, A. Truskinger, M. Cottman-Fields, J. Zhang, and P. Roe. 2015. Similarity-based birdcall retrieval from environmental audio. *Ecological Informatics* 29:66–76. http://dx.doi.org/10.1016/j.ecoinf.2015.07.007	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Duan, S., J. Zhang, P. Roe, J. Wimmer, and X. Dong. 2013. Timed probabilistic automaton: a bridge between automatic species recognition. Pages 1519-1524 *in Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference*. 14–18 July, Bellevue, Washington, USA. AAAI, Palo Alto, California, USA. [online] URL: https://www.aaai.org/ocs/index.php/IAAI/IAAI13/paper/view/6092/6429	**Single**	**Field recording**
Dufour, O., T. Artieres, H. Glotin, and P. Giraudet. 2014. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. Pages 83-95 *in Soundscape Semiotics - Localisation and Categorisation.* InTech, London, UK. http://dx.doi.org/10.5772/56872	Multi	Field recording
Ferroudj, M. 2015. Detection of rain in acoustic recordings of the environment using machine learning techniques. Master of Information Technology thesis, Queensland University of Technology, Queensland, Australia.	N/A	Clip
Ganchev, T. D., O. Jahn, M. I. Marques, J. M. de Figueiredo. 2007. Automatic acoustic identification of singing insects. *Bioacoustics* 16:281–328. http://dx.doi.org/10.1080/09524622.2007.9753582	Multi	Clip
Ganchev, T. D., O. Jahn, M. I. Marques, J. M. de Figueiredo, and K-L. Schuchmann. 2015. Automated acoustic detection of *Vanellus chilensis lampronotus*. *Expert Systems with Applications* 42(15-16):6098-6111. http://dx.doi.org/10.1016/j.eswa.2015.03.036	**Single**	**Field recording**
Gonzalez, R. 2010. Effects of compression and window size on remote acoustic identification using sensor networks. Pages 1-10 *in 2010 4th International Conference in Signal Processing and Communication Systems (ICSPCS).*13-15 December, Gold Coast, Australia. http://dx.doi.org/10.1109/ICSPCS.2010.5709762	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Gradišek, A., G. Slapničar, J. Šorn, M. Luštrek, M. Gams, and J. Grad. 2016. Predicting species identity of bumblebees through analysis of flight buzzing sounds. *Bioacoustics* 26(1):63-76. http://dx.doi.org/10.1080/09524622.2016.1190946	Multi	Clip
Harma, A. 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. Pages V-545–548 *in Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. 6-10 April, Hong Kong. http://dx.doi.org/10.1109/ICASSP.2003.1200027	Multi	Clip
Heinicke, S., A. K. Kalan, O. J. J. Wagner, R. Mundry, H. Lukashevich, and H. S. Kühl. 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods in Ecology and Evolution* 6:753–763. http://dx.doi.org/10.1111/2041-210X.12384	Multi	Field recording
Hollowood, K., O. Kular, and E. Ribeiro. 2015. Classifying frog calls using gaussian mixture model and locality sensitive hashing. AMALTHEA REU Technical Report No. 2015-2. [online] URL: http://www.amathea-reu.org/	Multi	Clip
Holmes, S. B., K. A. McIlwrick, and L. A. Venier. 2014. Using automated sound recording and analysis to detect bird species-at-risk in southwestern Ontario woodlands. *Wildlife Society* Bulletin 38(3):591-598. http://dx.doi.org/10.1002/wsb.421	Single	Field recording: no benchmark
Huang, C.-J., Y.-J. Yang, D.-X. Yang, and Y.-J. Chen. 2009. Frog classification using machine learning techniques. *Expert Systems with Applications* 36:3737–3743. http://dx.doi.org/10.1016/j.eswa.2008.02.059	Multi	Clip
Jaafar, H., D. A. Ramli, and B. A. Rosdi. 2014. Comparative study on different classifiers for frog identification system based on bioacoustic signal analysis. Pages 172-176 *in Proceedings of the 2014 International Conference on Communications, Signal Processing and Computers.* 3-5 April in Melmaruvathur, India	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Jahn, O., T. D. Ganchev, M. I. Marques, and K-L. Schuchmann. 2017. Automated sound recognition provides insights into the behavioral ecology of a tropical bird. *PLoS ONE* 12:e0169041-29. http://dx.doi.org/10.1371/journal.pone.0169041	**Single**	**Field recording**
Jaiswara, R., D. Nandi, and R. Balakrishnan. 2013. Examining the effectiveness of discriminant function analysis and cluster analysis in species identification of male field crickets based on their calling songs. *PLoS ONE* 8:e75930–11. http://dx.doi.org/10.1371/journal.pone.0075930	Multi	Clip
Kaewtip, K., L. N. Tan, and A. Alwan. 2013. A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification. Pages 768-772 in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 26-31 May in Vancouver, BC, Canada. http://dx.doi.org/10.1109/ICASSP.2013.6637752	Multi	Clip
Kasten, E. P., P. K. McKinley, and S. H. Gage. 2010. Ensemble extraction for classification and detection of bird species. *Ecological Informatics* 5:153–166. http://dx.doi.org/10.1016/j.ecoinf.2010.02.003	Multi	Clip, field recording
Katz, J., S. D. Hafner, and T. Donovan. 2016. Assessment of error rates in acoustic monitoring with the R package monitoR. *Bioacoustics* 25(2):177-196. http://dx.doi.org/10.1080/09524622.2015.1133320	**Single**	**Field recording**
Lee, C.-H., C.-H. Chou, C.-C. Han, and R.-Z. Huang. 2006. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. *Pattern Recognition* Letters 27:93–101. http://dx.doi.org/10.1016/j.patrec.2005.07.004	Multi	Clip
Mencıa, E. L., J. Nam, and D. H. Lee. 2013. Learning multi-labeled bioacoustic samples with an unsupervised feature learning approach. Pages 1-6 *in Proceedings of the International Symposium for Neural Information Scaled for Bioacoustics.* Dec in Nevada, USA.	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Nicholson, D. 2016. Comparison of machine learning methods applied to birdsong element classification. Pages 57-61 *in Proceedings of the 15th Python in Science Conference (SCIPY).* July 10-16, Austin, TX, USA. [online] URL: http://conference.scipy.org/proceedings/scipy2016/david_nicholson.html	Multi	Clip
Noda, J. J., C. M. Travieso, and D. Sánchez-Rodríguez. 2016a. Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Systems with Applications* 50:100–106. http://dx.doi.org/10.1016/j.eswa.2015.12.020	Multi	Clip
Noda, J., C. Travieso, and D. Sánchez-Rodríguez. 2016b. Automatic taxonomic classification of fish based on their acoustic signals. *Applied Sciences* 6:443–12. http://dx.doi.org/10.3390/app6120443	Multi	Clip
Potamitis, I. 2014. Automatic classification of a taxon-rich community recorded in the wild. *PLoS ONE* 9:e96936–11. http://dx.doi.org/10.1371/journal.pone.0096936	Multi	Clip
Potamitis, I., S. Ntalampiras, O. Jahn, and K. Riede. 2014. Automatic bird sound detection in long real-field recordings: applications and tools. *Applied Acoustics* 80:1-9. http://dx.doi.org/10.1016/j.apacoust.2014.01.001	**Single**	**Clip, field recording**
Ptacek, L., L. Machlica, P. Linhart, P. Jaska, and L. Muller. 2015. Automatic recognition of bird individuals on an open set using as-is recordings. *Bioacoustics* 25(1):55-73. http://dx.doi.org/10.1080/09524622.2015.1089524	Multi	Field recording
Qian, K., Z. Zhang, F. Ringeval, and B. Schuller. 2015. Bird sounds classification by large scale acoustic features and extreme learning machine. Pages 1317–1321 *in 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP).* 14-16 December, Orlando, FL, USA. http://dx.doi.org/10.1109/GlobalSIP.2015.7418412	Multi	Clip
Ranjard, L., and H. A. Ross. 2008. Unsupervised bird song syllable classification using evolving neural networks. *The Journal of the Acoustical Society of America* 123(6):4358–4368. http://dx.doi.org/10.1121/1.2903861	Multi	Clip

Reference	Single or multi species recognizer	Test dataset type
Ranjard, L., S. J. Withers, D. H. Brunton, H. A. Ross, and S. Parsons. 2015. Integration over song classification replicates: Song variant analysis in the hihi. *The Journal of the Acoustical Society of America* 137(5):2542–2551. http://dx.doi.org/10.1121/1.4919329	Multi	Clip
Ruiz-Muñoz, J. F., G. Castellanos-Dominguez, and M. Orozco-Alzate. 2016. Enhancing the dissimilarity-based classification of birdsong recordings. *Ecological Informatics* 33:75–84. http://dx.doi.org/10.1016/j.ecoinf.2016.04.001	Multi	Clip
Salamon, J., J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling. 2016. Towards the automatic classification of avian flight calls for bioacoustic monitoring. *PLoS ONE* 11:e0166866–26. http://dx.doi.org/10.1371/journal.pone.0166866	Multi	Clip, field recording
Souza Filho, N. E., B. C. Oliveira, M. L. D. Silva, and J. Vielliard. 2014. Automatic classification of *Turdus rufiventris* song notes by spectrographic image template matching. *Ciência e Natura* 36:1–9. http://dx.doi.org/10.5902/2179460X11303	Multi	Clip
Stowell, D., and M. D. Plumbley. 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ* 2:e488–31. http://dx.doi.org/10.7717/peerj.488	Multi	Clip
Swiston, K. A., and D. J. Mennill. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-Billed, and putative Ivory-billed Woodpeckers. *Journal of Field Ornithology* 80(1):42-50. http://dx.doi.org/10.1111/j.1557-9263.2009.00204.x	**Single**	**Field recording**
Tachibana, R. O., N. Oosugi, and K. Okanoya. 2014. Semi-automatic classification of birdsong elements using a linear support vector machine. *PLoS ONE* 9:e92584. http://dx.doi.org/10.1371/journal.pone.0092584	Multi	Clip, field recording

Reference	Single or multi species recognizer	Test dataset type
Tan, L. N., A. Alwan, G. Kossan, M. L. Cody, and C. E. Taylor. 2015. Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data. *The Journal of the Acoustical Society of America* 137(3):1069–1080. http://dx.doi.org/10.1121/1.4906168	Multi	Clip
Tan, L. N., K. Kaewtip, M. L. Cody, C. E. Taylor, and A. Alwan. 2012. Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions. *In Interspeech.* [online] URL: http://www.seas.ucla.edu/spapl/paper/Tan_Interspeech2012.pdf	Multi	Clip
Thakur, A., J. Jain, P. Rajan, and A. D. Dileep. 2017. Bird audio detection using probability sequence kernels. [online] URL: http://machine-listening.eecs.qmul.ac.uk/wp-content/uploads/sites/26/2017/02/badChallenge_iitMandi.pdf	Multi	Clip
Tsai, W. H., Y. Y. Xu, and W. C. Lin. 2014. Bird species identification based on timbre and pitch features of their vocalization. *Journal of Information Science and Engineering* 30:1927-1944.	Multi	Clip
Turesson, H. K., S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque. 2016. Machine learning algorithms for automatic classification of marmoset vocalizations. *PLoS* ONE 11:e0163041–14. http://dx.doi.org/10.1371/journal.pone.0163041	Single	Clip
Ulloa, J. S., A. Gasc, P. Gaucher, T. Aubin, M. Réjou-Méchain, and J. Sueur. 2016. Screening large audio datasets to determine the time and space distribution of Screaming Piha birds in a tropical forest. *Ecological Informatics* 31:91-99. http://dx.doi.org/10.1016/j.ecoinf.2015.11.012	**Single**	**Field recording**
Vega, G., C. J. Corrada-Bravo, and T. M. Aide. 2016. Audio segmentation using Flattened Local Trimmed Range for ecological acoustic space analysis. *PeerJ Computer Science* 2:e70. http://dx.doi.org/10.7717/peerj-cs.70	N/A	Field recording

Reference	Single or multi species recognizer	Test dataset type
Ventura, T. M., A. G. de Oliveira, T. D. Ganchev, J. M. de Figueiredo, O. Jahn, M. I. Marques, and K.-L. Schuchmann. 2015. Audio parameterization with robust frame selection for improved bird identification. *Expert Systems with Applications* 42:8463–8471. http://dx.doi.org/10.1016/j.eswa.2015.07.002	Multi	Clip
Vignolo, L., J. A. Sarquis, E. León, and E. Albornoz. 2016. Furnariidae species recognition using speech-related features and machine learning. Pages 53-61 *in 17° Simposio Argentino de Inteligencia Artificial (ASAI)*. 5-9 September, Buenos Aires, Argentina. [online] URL: http://45jaiio.sadio.org.ar/sites/default/files/ASAI-15_0.pdf	Multi	Clip
Waddle, J. H., T. F. Thigpen, and B. M. Glorioso. 2009. Efficacy of automatic vocalization recognition software for anuran monitoring. *Herpetological Conservation and Biology* 4(3):384-388.	**Single**	**Field recording**
Walters, C. L., R. Freeman, A. Collen, C. Dietz, M. Brock Fenton, G. Jones, M. K. Obrist, S. J. Puechmaille, T. Sattler, B. M. Siemers, S. Parsons, and K. E. Jones. 2012. A continental-scale tool for acoustic identification of European bats. *Journal of Applied Ecology* 49:1064–1074. http://dx.doi.org/10.1111/j.1365-2664.2012.02182.x	Multi	Clip
Xie, J., J. Zhang, and P. Roe. 2016. Acoustic features for multi-level classification of Australian frogs. Pages 1-5 *in Proceedings of the 2015 International Conference on Information, Communications and Signal Processing (ICICS*). 2-4 April, Melmaruvathur, India. http://dx.doi.org/10.1109/ICICS.2015.7459891	Multi	Clip
Xie, J., M. Towsey, J. Zhang, and X. Dong. 2015. Application of image processing techniques for frog call classification. Pages 4190–4194 *in Proceedings of the 2015 International Conference on Image Processing (ICIP)*. 27-30 September, Québec City, QC, Canada. http://dx.doi.org/10.1109/icip.2015.7351595	Single	Clip

Table A2.1. Parameter settings used for a Common Nighthawk (*Chordeiles minor*) acoustic recognizer built in Song Scope software.

Parameter	Setting
FFT size	256
FFT overlap	½
Frequency minimum	30
Frequency range	80
Amplitude gain (dB)	0
Background filter (s)	1
Max syllable (ms)	723
Max syllable gap (ms)	0
Max song (ms)	723
Dynamic range (dB)	26
Algorithm	2.0
Maximum complexity	32
Maximum resolution	8
Score threshold	0
Quality threshold	20

Table A2.2. Parameter settings used for a Common Nighthawk (*Chordeiles minor*) acoustic recognizer built in Kaleidoscope software.

Parameter	Setting
FFT size	256
Max distance from cluster centre to include outputs in cluster.csv	2.0
Max states	12
Max distance to cluster centre for building clusters	1.0
Max clusters	2
Frequency minimum (kHz)	1.0
Frequency maximum (kHz)	7.0
Min song (ms)	100
Max song (ms)	700
Max syllable gap (ms)	0

Table A2.3. Parameter settings used for a Common Nighthawk (*Chordeiles minor*) acoustic recognizer built with the binary point template function in the MonitoR package in R software. Frequency minimum, frequency maximum, and amplitude cutoff were adjusted by hand within the indicated ranges for each of the 100 templates made.

Parameter	Setting
FFT size	512
FFT transformation	Hanning window
FFT overlap	None
Frequency minimum (kHz)	2.1 to 2.8
Frequency maximum (kHz)	5.0 to 5.8
Amplitude cutoff (db)	-53 to -17
Buffer	0
Score threshold	0.1
Min gap between hits (s)	0.1

Table A2.4. Parameter settings used for a Common Nighthawk (*Chordeiles minor*) acoustic recognizer built using a convolutional neural network (CNN) in Tensorflow software.

Parameter	Setting
Spectrogram	Input mel-scaled (96 mel filters)
FFT size	512
FFT overlap	75%
Sample rate (kHz)	16
Layer 1	7x7 conv stride 2. 8 ReLU units
Layer 2	3x3 max-pooling stride 2
Layer 3	24x9 conv. 32 ReLU units
Layer 4	1x1 conv. 1 sigmoid unit
Layer 5	Global max-pooling
Loss	Cross-entropy
Optimizer	Adam
Batch size	64
Learning rate	0.001
Score threshold	0.001
Min gap between hits (s)	0.1

Table A2.5. Parameter settings used for a Common Nighthawk (*Chordeiles minor*) acoustic recognizer built in RavenPro software.

Parameter	Setting
FFT Size	512
Minimum frequency (kHz)	1.8
Maximum frequency (kHz)	6
Minimum duration (s)	0.2
Maximum duration (s)	0.6
Minimum separation (s)	0.096
Minimum occupancy (%)	15
SNR threshold (dB)	10
Block size (s)	0.8
Hop size (s)	0.4
Percentile	20

Table A3.1 AIC ranking of polynomial models for Common Nighthawk (*Chordeiles minor*) presence-absence recall from acoustic data processed with automated acoustic recognition programs. Bold indicates the model selected.

Recognizer	Model	df	logLik	AIC	∆AIC	AICw
CNN	Presence = null	1	-2048.3	4098.6	200.3	0.00
CNN	Presence = score	2	-1951.0	3905.9	7.6	0.02
CNN	Presence = score + I(score^2)	3	-1949.5	3905.0	6.7	0.03
CNN	**Presence = score + I(score^2) + I(score^3)**	**4**	**-1945.2**	**3898.3**	**0.00**	**0.95**
Kaleidoscope	Presence = null	1	-3087.4	6176.8	502.3	0.00
Kaleidoscope	Presence = score	2	-2838.1	5680.2	5.7	0.05
Kaleidoscope	Presence = score + I(score^2)	3	-2837.7	5681.3	6.8	0.03
Kaleidoscope	**Presence = score + I(score^2) + I(score^3)**	**4**	**-2833.2**	**5674.5**	**0.0**	**0.92**
MonitoR	Presence = null	1	-3103.4	6208.9	634.3	0.00
MonitoR	Presence = score	2	-2795.2	5594.3	19.7	0.00
MonitoR	**Presence = score + I(score^2)**	**3**	**-2784.3**	**5574.6**	**0.0**	**0.69**
MonitoR	Presence = score + I(score^2) + I(score^3)	4	-2784.1	5576.2	1.6	0.31
RavenPro	**Presence = null**	**1**	**-3047.7**	**6097.3**	**0.0**	**0.43**
RavenPro	Presence = score	2	-3046.8	6097.6	0.3	0.37
RavenPro	Presence = score + I(score^2)	3	-3046.8	6099.6	2.3	0.14
RavenPro	Presence = score + I(score^2) + I(score^3)	4	-3046.7	6101.4	4.0	0.06
Song Scope	Presence = null	1	-2615.7	5233.3	610.6	0.00
Song Scope	Presence = score	2	-2320.8	4645.7	23.0	0.00
Song Scope	Presence = score + I(score^2)	3	-2311.9	4629.9	7.16	0.03
Song Scope	**Presence = score + I(score^2) + I(score^3)**	**4**	**--2307.4**	**4622.7**	**0.0**	**0.97**