INTERFACE

royalsocietypublishing.org/journal/rsif

Research

**Cite this article:** Stowell D, Petrusková T, Šálek M, Linhart P. 2019 Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *J. R. Soc. Interface* **16**: 20180940. http://dx.doi.org/10.1098/rsif.2018.0940

Received: 13 December 2018
Accepted: 21 March 2019

**Subject Category:**
Life Sciences–Mathematics interface

**Subject Areas:**
environmental science

**Keywords:**
animal communication, individuality, acoustic monitoring, data augmentation, song repertoire, vocalization

**Author for correspondence:**
Dan Stowell
e-mail: dan.stowell@qmul.ac.uk

THE ROYAL SOCIETY
PUBLISHING

# Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions

Dan Stowell[1], Tereza Petrusková[2], Martin Šálek[3,4] and Pavel Linhart[5]

[1]Machine Listening Lab, Centre for Digital Music, Queen Mary University of London, London, UK
[2]Department of Ecology, Faculty of Science, Charles University, Prague, Czech Republic
[3]Institute of Vertebrate Biology, The Czech Academy of Sciences, Brno, Czech Republic
[4]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Prague, Czech Republic
[5]Department of Behavioural Ecology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

DS, 0000-0001-8068-3769; TP, 0000-0002-8046-1581; MŠ, 0000-0003-2639-9745; PL, 0000-0002-6097-5892

Many animals emit vocal sounds which, independently from the sounds' function, contain some individually distinctive signature. Thus the automatic recognition of individuals by sound is a potentially powerful tool for zoology and ecology research and practical monitoring. Here, we present a general automatic identification method that can work across multiple animal species with various levels of complexity in their communication systems. We further introduce new analysis techniques based on dataset manipulations that can evaluate the robustness and generality of a classifier. By using these techniques, we confirmed the presence of experimental confounds in situations resembling those from past studies. We introduce data manipulations that can reduce the impact of these confounds, compatible with any classifier. We suggest that assessment of confounds should become a standard part of future studies to ensure they do not report over-optimistic results. We provide annotated recordings used for analyses along with this study and we call for dataset sharing to be a common practice to enhance the development of methods and comparisons of results.

## 1. Introduction

Animal vocalizations exhibit consistent individually distinctive patterns, often referred to as acoustic signatures. Individual differences in acoustic signals have been reported widely across all classes of vertebrates (e.g. fish [1], amphibians [2], birds [3], mammals [4]). Individual differences may arise from various sources, for example distinctive fundamental frequency and harmonic structure between acoustic signals can result from individual vocal tract anatomy [4,5]; distinct temporal or frequency modulation patterns of vocal elements may result from inaccurate matching of innate or learned template or can occur de novo through improvisation [6]. Such individual signatures provide individual recognition cues for other conspecific animals, and individual recognition based on acoustic signals is widespread among animals [7]. Long-lasting individual recognition spanning over one or more years has also been often demonstrated [8–10]. External and internal factors such as, for example, sound degradation during transmission [11,12], variable ambient temperature [13], inner motivation state [14,15], acquisition of new sounds during life [16], may potentially increase variation of acoustic signals. Despite these potential complications, robust individual signatures have been found in many taxa.

Besides being studied for their crucial importance in social interactions [17–19], individual signatures can become a valuable tool for monitoring animals. Acoustic monitoring of individuals of various species based on vocal cues could become a powerful tool in conservation (reviewed in [3,20,21]).

Classical capture-mark methods of individual monitoring involve physically disturbing the animals of interest and might have a negative impact on the health of studied animals or their behaviour (e.g. [22–25]). Also, concerns have been raised about possible biases in demographic and behavioural studies resulting from trap boldness or shyness of specific individuals [26]. Individual acoustic monitoring offers a great advantage of being non-invasive, and thus can be deployed across species with fewer concerns about effect on behaviour [3]. It also may reveal complementary or more detailed information about species behaviour than classical methods [27–30].

Despite many pilot studies [28,31–33], automatic acoustic individual identification is still not routinely applied. It is usually restricted to a particular research team or even to a single research project, and, eventually, might be abandoned altogether for a particular species. Part of the problem probably lies in the fact that methods of acoustic individual identification were closely tailored to a single species (software platform, acoustic features used, etc.). This is good in order to obtain the best possible results for a particular species but it also hinders general, widespread application because methods need to be developed from scratch for each new species or even project. Little attention has been paid to developing general methods of automatic acoustic individual identification (henceforth 'AAII') which could be used across different species.

A few studies in the past have proposed to develop a general, call-type-independent acoustic identification, working towards approaches that could be used across different species, having simple as well as complex vocalizations [34]. Despite promising results, most of the published papers included vocalizations recorded within very limited periods of time (a few hours in a day) [34–37]. Hence, these studies might have failed to separate effects of target signal and potentially confounding effects of particular recording conditions and background sound, which have been reported as notable problems in case of other machine learning tasks [38,39]. For territorial animals, the acoustic confounds will be clear in some cases: if one territory is close to a river and another is not, then a black-box classifier might use the sounds of the river itself or of the fauna living near the river to distinguish the two individuals, rather than the sounds that the individuals themselves make. Typical confounds are more insidious, less obvious than this, although a human observer might not note subtle distinctions between acoustic environments, such as the relative amount of energy coming from specific sources or their distance from the receiver, but an automated algorithm may pick up on them and give them undue weight.

Reducing such confounds directly, by recording an animal in different backgrounds, may not be achievable in field conditions since animals typically live within limited home ranges and territories. However, acoustic background can change during the breeding season due to vegetation changes or cycles in activity of different bird species. Also, song birds may change territories in subsequent years or even within a single season [27]. Some other studies of individual acoustic identification, on the other hand, provided evidence that machine learning acoustic identification can be robust in respect to possible long-term changes in the acoustic background but did not provide evidence of being generally usable for multiple species [30,32]. Therefore, the challenge of reliable generalization of the machine learning approach in acoustic individual identification across different conditions and different species has not yet been satisfactorily demonstrated.

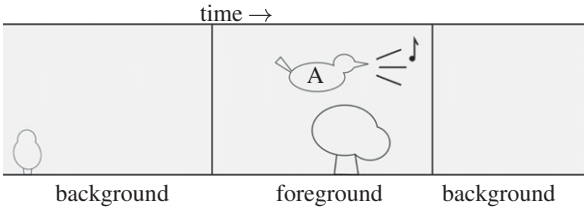


**Figure 1.** Most data items used in bird sound recognition are clipped from longer recordings, whether automatically or manually. We refer to these as 'foreground' audio clips, and we also create corresponding 'background' audio clips from the other audio segments that are typically discarded.

In the work reported in this paper, we tested the generalization of machine learning across species and across recording conditions in the context of individual acoustic identification. We used extensive data for three different bird species, including repeated recordings of the same individuals within and across two breeding seasons. As well as directly evaluating across seasons, we also introduced ways to modify the evaluation data to probe the generalization properties of the classifier. We then improved on the baseline approach by developing novel methods which help to improve generalization performance, again by modifying the data used. Although tested with selected species and classifiers, our approach of modifying the data rather than the classification algorithm was designed to be compatible with a wide variety of automatic identification workflows.

## 1.1. Previous methods for AAII

We briefly review studies representing methods for automatic classification of individuals. Note that in the present work, as in many of the cited works, we set aside questions of automating the prior steps of recording focal birds and isolating the recording segments in which they are active. It is common, in preparing data sets, for recordists to collate recordings and manually trim them to the regions containing the 'foreground' individual of interest (often with some background noise), discarding the regions containing only background sound (figure 1). In the present work, we will make use of both the foreground and background clips, and our method will be applicable whether such segmentation is done manually or automatically.

Matching a signal against a library of templates is a well-known bioacoustic technique, most commonly using spectrogram (sonogram) representations of the sound, via spectrogram cross-correlation [40]. For identifying individuals, template matching will work in principle when the individuals' vocalizations are strongly stereotyped with stable individual differences—and, in practice, this can give good recognition results for some species [41]. However, template matching is only applicable to a minority of species. It is strongly call-type dependent and requires a library covering all of the vocalization units that are to be identified. It is unlikely to be useful for species which have a large vocabulary, high variability, or whose vocabulary changes substantially across seasons.

Classical approaches that are more flexible include Gaussian mixture models (GMMs) and hidden Markov models (HMMs), previously used extensively in human

speech technology [30,42]. These do not rely on a strongly fixed template but rather build a statistical model summarizing the spectral data that are likely to be produced from each individual. GMM-based methods have been used in songbirds, although without testing across multiple seasons [42], and for orangutan including across-season evaluation [30]. Adi *et al.* used HMMs for recognizing individual songbirds, in this case ortolan buntings, with a pragmatic approach to the call-type dependence intrinsic to HMM sequence models [32].

Other computational approaches have been studied. Cheng *et al.* compared four classifier methods, aiming to develop call-type-independent recognition across three passerine species [37]. They found HMM and support vector machines to be favourable among the methods they tested. However, the data used in this study were relatively limited: it was based on single recording sessions per individual, and thus could not test across-year performance; and the authors deliberately curated the data to select clean recordings with minimal noise, acknowledging that this would not be representative of realistic recordings. Fox *et al.* also focused on the challenge of call-independent identification, across three other passerine species [34,35]. They used a neural network classifier, and achieved good performance for their species. However, again the data for this study were based on a single session per individual, which makes it unclear how far the findings generalize across days and years, and also does not fully test whether the results may be affected by confounding factors such as recording conditions.

### 1.2. Automatic classification and data augmentation

More generally, computational methods for various automatic recognition tasks have recently been dominated and dramatically improved by new trends in machine learning, including deep learning. In bioacoustic tasks, machine learning has enabled extremely strong performance in automatic detection of animal sounds [43], automatic species classification from sound [44,45] and other tasks [46].

These gains come partly from the machine learning methods but also notably from the use of very large annotated datasets for training. Applying the same methods to small datasets, like those typically available from specific wild individuals, is known to be an open research challenge [47]. In fact the challenge of reliable generalization even with large datasets is far from solved, and is an active research topic within the broad field of machine learning. Within bioacoustics, this has recently been studied for detection of bird sounds [43]. In deep learning, it was discovered that even the best-performing deep neural networks might be surprisingly non-robust, and could be forced to change their decisions by the addition of tiny imperceptible amounts of background noise to an image [38].

Note that deep learning systems also typically require very large amounts of data to train, meaning they may currently be infeasible for tasks such as acoustic individual ID in which the number of recordings per individual is necessarily limited. For deep learning, 'data augmentation' has been used to expand dataset sizes. Data augmentation refers to the practice of synthetically creating additional data items by modifying or recombining existing items. In the audio domain, this could be done, for example, by adding noise, filtering or mixing audio clips together [48,49]. Data augmentation has become common for enlarging datasets to train deep learning, and some of the highest-performing automatic species recognition systems rely in part on such data augmentations to attain their strongest results [49]. The basic principle is to encourage a classifier to learn the correct associations, by making use of expert knowledge, for example that adding a small amount of background noise in most cases does not alter the correct labelling of a data item. This therefore should typically encourage a classifier to specialize in the phenomena of interest rather than irrelevances. However, simple unprincipled data augmentation does not reduce issues such as undersampling (e.g. some vocalizations unrepresented in data set) or confounding factors, since the expanded data sets typically contain the same limitations but repeated.

There thus remains a gap in applying machine learning for automatic individual identification as a general-purpose tool that can be shown to be reliable for multiple species and can generalize correctly across recording conditions. In the present work, we will address specific confound difficulties that are present in AAII, by developing structured data augmentation schemes which can expose and then reduce generalization problems.

## 2. Material and methods

### 2.1. Data collection

For this study, we chose three bird species of varying vocal complexity (figure 2), in order to explore how a single method might apply to the same task at differing levels of difficulty and variation. Little owl (*Athene noctua*) represents a species with simple vocalization (figure 2*a*): the territorial call is a single syllable which is individually unique and it is held to be stable over time (P Linhart and M Šálek 2018, unpublished data) as was shown in several other owl species (e.g. [31,50]). Then, we selected two passerine species, which exhibit vocal learning: chiffchaff (*Phylloscopus collybita*) and tree pipit (*Anthus trivialis*). Tree pipit songs are also individually unique and stable over time [27]; but the male on average uses 11 syllable types (6–18) which are repeated in phrases that can be variably combined to create a song ([51], figure 2*b*). Chiffchaff song, when visualized, may seem simpler than that of the pipit. However, the syllable repertoire size might actually be higher—9 to 24 types—and, contrary to the other species considered, chiffchaff males may change syllable composition of their songs over time ([52], (figure 2*c*). Selected species also differ in their ecology. While little owls are sedentary and extremely faithful to their territories [53], tree pipits and chiffchaffs belong to migratory species with high fidelity to their localities. Annual returning rates for both are 25% to 30% ([27], P Linhart 2012, unpublished data).

For each of these species, we used targeted recordings of single vocally active individuals. Distance to the recorded individual varied across individuals and species according to their tolerance towards people. We tried to get the best recording and minimize distance to each singing individual without disturbing its activities. Recordings were always done under favourable weather conditions (no rain, no strong wind). All three species were recorded with the following equipment: Sennheiser ME67 microphone, Marantz PMD660 or 661 solid-state recorder (sampling frequency 44.1 kHz, 16 bit, PCM). In general, the signal-to-noise ratio is very good in all of our recordings (not rigorously assessed), but there are also environmental sounds, sounds from other animals or conspecifics in the recording background.

**Little owl (Linhart & Šálek [54]):** Little owls were recorded in two Central European farmland areas: northern Bohemia, Czech Republic (50°23′ N, 13°40′ E), and eastern Hungary (47°33′ N, 20°54′ E). Recordings were made from sunset until midnight between March and April of 2013—2014. Territorial

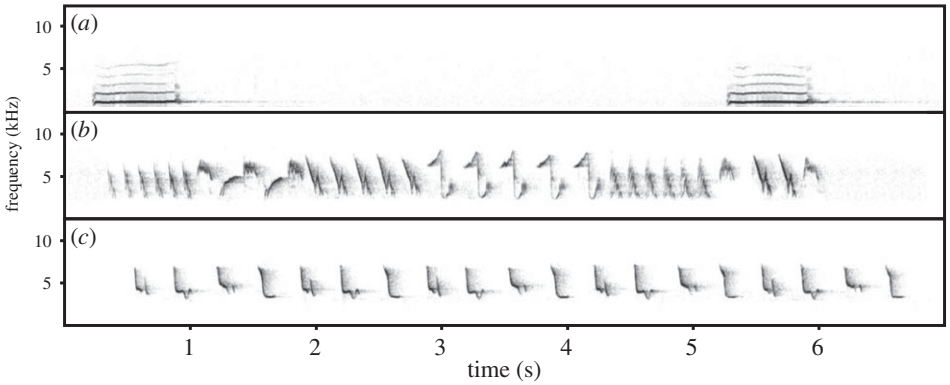


**Figure 2.** Example spectrograms representing our three study species: (*a*) little owl, (*b*) tree pipit and (*c*) chiffchaff.

**Table 1.** Details of the audio recording datasets used.

| evaluation scenario | num. inds | foreground # audio files (train : eval) | foreground total minutes (train : eval) | background # audio files (train : eval) | Background total minutes (train : eval) |
|---|---|---|---|---|---|
| chiffchaff within-year | 13 | 5107 : 1131 | 451 : 99 | 5011 : 1100 | 453 : 92 |
| chiffchaff only-15 | 13 | 195 : 1131 | 18 : 99 | 195 : 1100 | 21 : 92 |
| chiffchaff across-year | 10 | 324 : 201 | 32 : 20 | 304 : 197 | 31 : 24 |
| little owl across-year | 16 | 545 : 407 | 11 : 8 | 546 : 409 | 34 : 27 |
| pipit within-year | 10 | 409 : 303 | 27 : 21 | 398 : 293 | 49 : 47 |
| pipit across-year | 10 | 409 : 313 | 27 : 19 | 398 : 306 | 49 : 37 |

calls of each male were recorded for up to 3 min after a short playback provocation (1 min) inside their territories from up to 50 m distance from the individuals. The identities of the males could not be explicitly checked because only a small proportion of males were ringed. Therefore, we inferred identity by the territory location combined with the call frequency modulation pattern which is distinctive per individual.

**Chiffchaff (Průchová *et al.* [42,52]):** Chiffchaff males were recorded in a former military training area on the outer boundary of České Budějovice town, the Czech Republic (48°59.5′ N, 14°26.5′ E). Males were recorded for the purposes of various studies from 2008 up to and including 2011. Recordings were done from 05.30 to 11.00 h in the morning. Only spontaneously singing males were recorded from within about 5–15 m distance. The identities of the males were confirmed by colour rings.

**Tree Pipit (Petrusková *et al.* [27]):** Tree pipit males were recorded at the locality Brdská vrchovina, the Czech Republic (49°84′ N, 14°10′ E) where the population has been continuously studied since 2011. Spontaneously singing males were recorded throughout whole day according to the natural singing activity of Tree pipits from mid-April to mid-July. Males were identified either based on colour ring observations or their song structure [27].

All audio files were divided into separate sound files during which the focal individual was vocally active (foreground) and inactive (background). These sound files formed basic units for the whole recognition process. In the case of pipits and chiffchaffs, one bout of territorial song was treated as the basic recognition unit, whereas for little owl a unit was a single territorial call, since these can occur in isolation (figure 2). The total numbers of individuals and sound files in each dataset are summarized in table 1.

## 2.2. Structured data augmentation

'Data augmentation' in machine learning refers to creating artificially large or diverse datasets by synthetically manipulating items in datasets to create new items—for example, by adding noise or performing mild distortions. These artificially enriched datasets, used for training, often lead to improved automatic classification results, helping to mitigate the effects of limited data availability [55,56]. Data augmentation is increasingly used in machine learning applied to audio. Audio-specific manipulations used might include filtering or pitch-shifting, or the mixing together of audio files (i.e. summing their signals together) [48,49,57]. This last option is somewhat related to an idea called 'mixup' data augmentation, which is based on linearly interpolating between pairs of data items [58].

In this work, we describe two augmentation methods used specifically to evaluate and to reduce the confounding effect of background sound. These *structured* data augmentations are based on audio mixing but with an explicit strategy for the choices of files to mix, selected based on foreground and background identity metadata. We make use of the fact that when recording audio from focal individuals in the wild, it is common to obtain recording clips in which the focal individual is vocalizing (figure 3*a*), as well as 'background' recordings in which the vocal individual is silent (figure 3*b*). The latter are commonly discarded. We used them as follows:

> **Adversarial data augmentation:** To evaluate the extent to which confounding from background information is an issue, we created datasets in which each foreground recording has been mixed with one background recording from some other individual (figure 3*c*). In the best case, this should make no difference, since the resulting sound clip is acoustically equivalent to a recording of the foreground individual, but with a little extra irrelevant background noise. In fact, it could be considered a synthetic test of the case in which an individual is recorded having travelled out of their home range. In the worst case, a classifier that has learnt undesirable correlations between foreground and background will be misled by the modification, either increasing the probability of classifying as the individual whose territory provided the extra background, or simply confusing the classifier and reducing its general ability to classify well. In our implementation, each foreground item was used once, each mixed with a different background item. Thus the evaluation set remains the same size

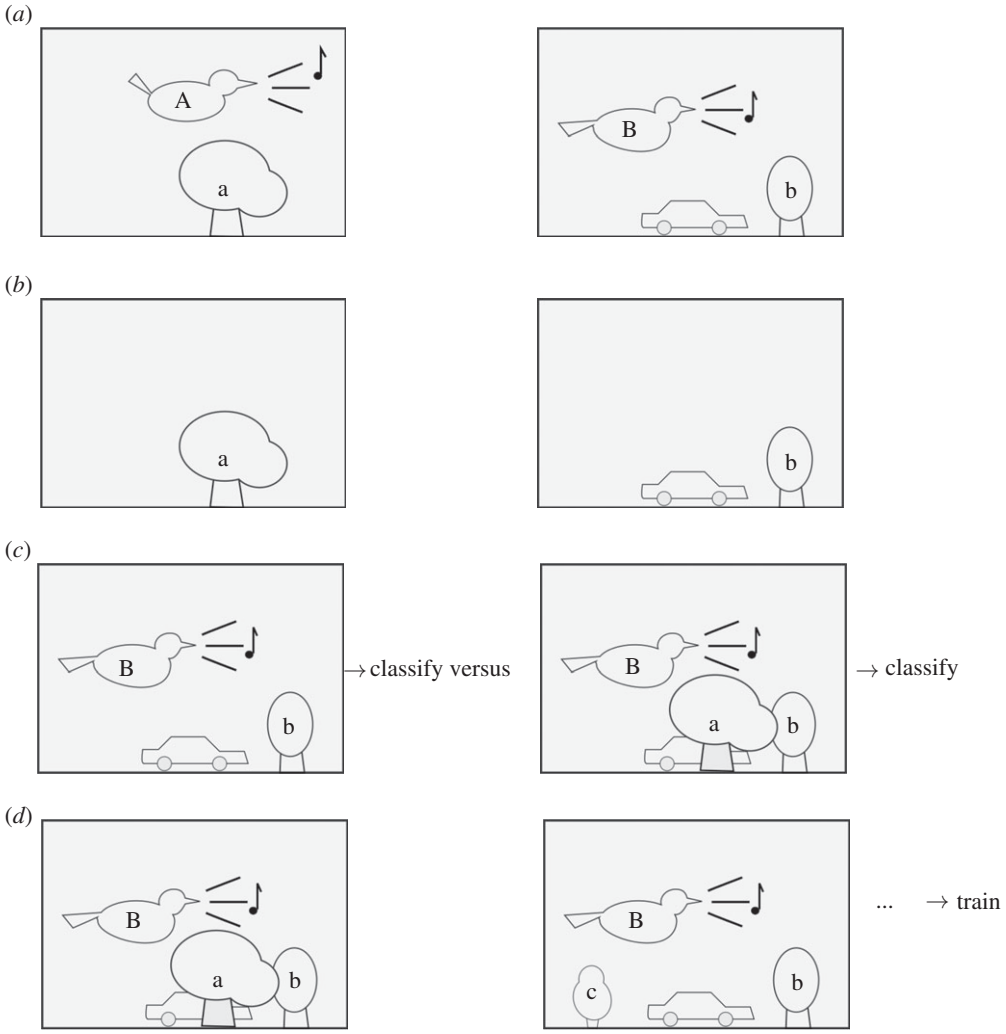


**Figure 3.** Explanatory illustration of our data augmentation interventions. (*a*) 'Foreground' recordings, which also contain some signal content coming from the background habitat. The foreground and background might not vary independently, especially in the case of territorial animals. (*b*) 'Background' recordings, recorded when the focal animal is not vocalizing (*c*) In *adversarial* data augmentation, we mix each foreground recording with a background recording from another individual, and measure the extent to which this alters the classifier's decision. (*d*) In *stratified* data augmentation, each foreground recording is mixed with a background recording *from each of the other classes*. This creates an enlarged data set with reduced confounding correlations.

as the unmodified set. We evaluated the robustness of a classifier by looking at any changes in the overall correctness of classification, or in more detail via the extent to which the classifier outputs are modified by the adversarial augmentation.

**Stratified data augmentation:** We can use a similar principle during the training process to create an enlarged and improved training data set. We created training datasets in which each training item had been mixed with an example of background sound from each other individual (figure 3*d*). If there are $K$ individuals this means that each item is converted into $K$ synthetic items, and the dataset size increases by a factor of $K$. Stratifying the mixing in this way, rather than selecting background samples purely at random, is intended to expose a classifier to training data with reduced correlation between foreground and background, and thus reduce the chance that it uses confounding information in making decisions.

To implement the foreground and background audio file mixing, we used the `sox` processing tool v.14.4.1 to perform simple additive mixing between the foreground and background recordings.[1]

## 2.3. Using background items directly

Alongside our data augmentation, we can also consider simple interventions in which the background sound recordings are used alone without modification.

One way of diagnosing confounding-factor issues in AAII is to apply the classifier to *background-only* sound recordings. If there are no confounds in the trained classifier, trained on foreground sounds, then it should be *unable* to identify the corresponding individual for any given background-only sound (identifying 'a' or 'b' in figure 3*b*). Automatic identification (AAII) for background-only sounds should yield results at around chance level.

A second use of using the background-only recordings is to create an explicit 'wastebasket' class during training. As well as training the classifier to recognize individual labels A, B, C, . . ., we created an additional 'wastebasket' class which should be recognized as 'none of the above', or in this case, explicitly as 'background'. The explicit-background class may or may not be used in the eventual deployment of the system. Either way, its inclusion in the training process could help to ensure that the classifier learns not to make mistaken associations with the other classes.

This latter approach is related to the universal background model (UBM) used in open-set recognition methods [42]. Note that the 'background' class is likely to be different in kind from the other classes, having very diverse sounds. In methods with an explicit UBM, the background class can be handled differently than the others [42]. Here, we chose to use methods that can work with any classifier, and so the background class was simply treated analogously to the classes of interest.

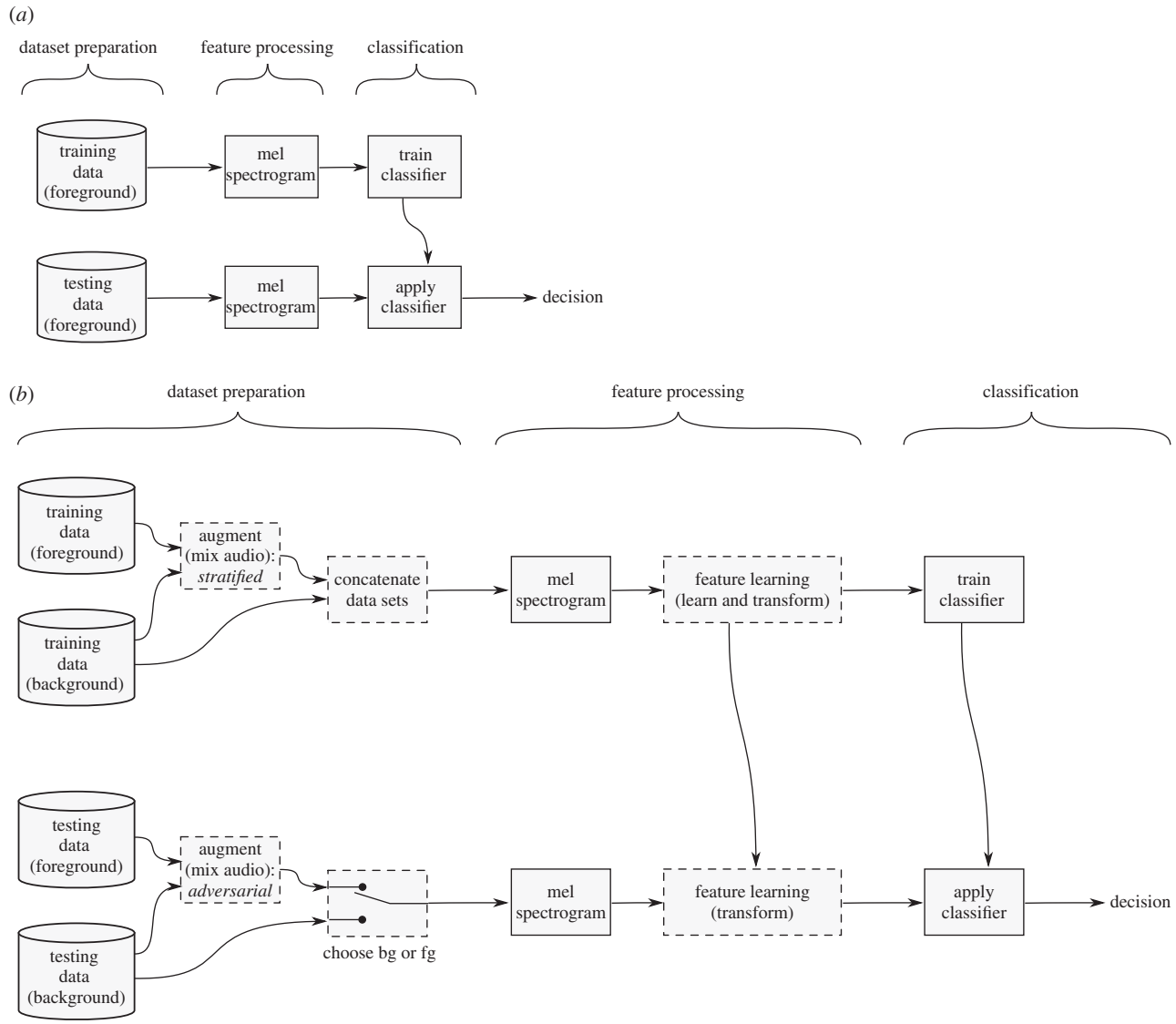


**Figure 4.** Classification workflows. (*a*) A standard workflow for automatic audio classification. The upper portion shows the training procedure, and the lower shows the application or evaluation procedure. (*b*) Workflow for our automatic classification experiments. Dashed boxes represent steps which we enable/disable as part of our experiment. The upper portion shows the training procedure, and the lower shows the evaluation procedure. The two portions are very similar. However, note that the purpose and method of augmentation is different in each, as is the use of background-only audio: in the training phase, the 'concatenation' block creates an enlarged training set as the union of the background items and the foreground items, while in the evaluation phase, the 'choose' block select only one of the two, for the system to make predictions about.

## 2.4. Automatic classification

In this work, we started with a standard automatic classification processing workflow (figure 4*a*), and then experimented with inserting our proposed improvements. We modified the feature processing stage, but our main innovations in fact came during the dataset preparation stage, using the foreground and/or background datasets in various combinations to create different varieties of training and testing data (figure 4*b*).

As in many other works, the audio files—which in this case may be the originals or their augmented versions—were not analysed in their raw waveform format, but were converted to a mel spectrogram representation: 'mel' referring to a perceptually motivated compression of the frequency axis of a standard spectrogram. We used audio files (44.1 kHz mono) converted into spectrograms using frames of length 1024 (23 ms), with Hamming windows, 50% frame overlap, and 40 mel bands. We applied median-filtering noise reduction to the spectrogram data, which helps to remove unchanging or slowly changing background noise—i.e. broadly similar effects as inverse Fourier transform denoising.

Following the findings of [59], we also applied *unsupervised feature learning* to the mel spectrogram data as a preprocessing step. This procedure scans through the training data in unsupervised fashion (i.e. neglecting the data labels), finding a linear projection that provides an informative transformation of the data. The features used are then not the mel spectral energies, but their transformed versions.[2] We evaluated the audio feature data with and without this feature learning step, to evaluate whether the data representation had an impact on the robustness and generalizability of automatic classification. In other words, as input to the classifier we used either the mel spectrograms, or the learned representation obtained by transforming the mel spectrogram data. In all cases, the features, which vary over time, are summarized by their mean and standard deviation over time so they can be processed as fixed-length feature vectors.

The automatic classifier we used was one based on a random forest classifier that was previously tested successfully for bird species classification, but had not been tested for AAII [59].[3] The classifier is a multi-class classifier and outputs scores indicating, for a given audio clip, which of the known individuals is singing.

## 2.5. Evaluation

As is standard in automatic classification evaluation, we divided our datasets into portions used for training the system, and

portions used for evaluating system performance. Items used in training were not used in evaluation, and the allocation of items to the training or evaluation sets was done to create a partitioning through time: evaluation data came from different days within the breeding season, or subsequent years, than the training data. This corresponds to a plausible use-case in which a system is trained with existing recordings and then deployed; the partitioning also helps to reduce the probability of over-estimating performance.

To quantify performance we used receiver operating curve (ROC) analysis, and as a summary statistic the area under the ROC curve (AUC). The AUC summarizes classifier performance and has various desirable properties for evaluating classification [60].

We evaluated the classifiers following the standard paradigm used in machine learning. Note that during evaluation, we optionally modified the evaluation data sets in two possible ways, as already described: adversarial data augmentation, and background-only classification. In all cases, we used AUC as the primary evaluation measure. However, we also wished to probe the effect of adversarial data augmentation in finer detail: even when the overall decisions made by a classifier are not changed by modifying the input data, there may be small changes in the full set of prediction scores it outputs. A classifier that is robust to adversarial augmentation should be one for which the scores it outputs change little if at all. Hence for the adversarial augmentation test, we also took the scores output from the classifier and compared them against their equivalent scores from the same classifier in the non-adversarial case. We measured the difference between these sets of scores simply by their root-mean-square error (RMS error).

### 2.6. Phase one: testing with chiffchaff

For our first phase of testing, we wished to compare the effectiveness of the different proposed interventions, and their relative effectiveness on data tested within-year or across-year. We chose to use the chiffchaff datasets for these tests, since the chiffchaff song has an appropriate level of complexity to elucidate the differences between classifier performance, in particular, the possible change of syllable composition across years. The chiffchaff dataset is also by far the largest.

We wanted to explore the difference in estimated performance when evaluating a system with recordings from the same year, separated by days from the training data, versus recordings from a subsequent year. In the latter case, the background sounds may have changed intrinsically, or the individual may have moved to a different territory; and of course, the individual's own vocalization patterns may change across years. This latter effect may be an issue for AAII with a species such as the chiffchaff, and also impose limits to the application of previous approaches such as template-based matching. Hence, we wanted to test whether this more flexible machine learning approach could detect individual signature in the chiffchaff even when applied to data from a different field season. We thus evaluated performance on 'within-year' data—recordings from the same season—and 'across-year' data—recordings from the subsequent year, or a later year.

Since the size of data available is often a practical constraint in AAII, and since dataset size can have a strong influence on classifier performance, we further performed a version of the 'within-year' test in which the training data had been restricted to only 15 items per individual. The evaluation data was not restricted.

To evaluate formally the effect of the different interventions, we applied generalized linear mixed models (GLMM) to our evaluation statistics, using the `glmmadmb` package within R v.3.4.4 [61,62]. Since AUC is a continuous value constrained to the range [0, 1], we used a beta link function. Since RMSE is a non-negative error measure, we used a gamma family with a logarithmic link function. For each of these two evaluation measures, we applied a GLMM, using the data from all three evaluation scenarios (within-year, cross-year, only-15). The evaluation scenario was included as a random effect. Since the same evaluation-set items were reused in differing conditions, this was a repeated-measures model with respect to the individual song recordings.

We tested the GLMM residuals for the two evaluation measures (AUC, RMSE) and found no evidence for overdispersion. We also tested all possible reduced models with factors removed, comparing among models using AIC. In both cases, the full model as well as a model with 'exbg' (explicit-background training) removed gave the best fit, with the full model less than 2 units above the exbg-reduced model and leading to no difference in significance estimates. We therefore report results from the full models.

### 2.7. Phase two: testing multiple species

In the second phase of our investigations, we evaluated the selected approach across the three species separately: chiffchaff, pipit and little owl. For each of these, we compared the most basic version of the classifier (using mel features, no augmentation and no explicit-background) against the improved version that was selected from phase one of the investigation. For each species separately, and using within-year and across-year data according to availability, we evaluated the basic and the improved AAII system for the overall performance (AUC measured on foreground sounds). We also evaluated their performance on background-only sounds, and on the adversarial data augmentation test, both of which checked the relationship between improved classification performance and improvements or degradations in the handling of confounding factors.

For both of these tests (background-only testing and adversarial augmentation), we applied GLMM tests similar to those already stated. In these cases, we entered separate factors for the testing condition and for whether the improved AAII system was in use, as well as an interaction term between the two factors. This therefore tested for an effect of whether our improved AAII system indeed mitigated the problems that the tests were designed to expose.

## 3. Results

### 3.1. Phase one: chiffchaff

AAII performance over the 13 chiffchaff individuals was strong, above 85% AUC in all variants of the within-year scenario (figure 5). For interpretation, note that this corresponds to over 85% probability that a random true-positive item is ranked higher than a random true-negative item by the system [60]. This reduced to around 70–80% when the training set was limited to 15 items per individual, and reduced even further to around 60% in the across-year evaluation scenario. Recognizing chiffchaff individuals across years remains a challenging task even under the studied interventions.

The focus of our study is on discriminating between recorded individuals, not on the prior step of detecting the presence of bird sounds. However, our 'explicit-background' configuration gave some insight into the potential for automation of this prior step. Across all three of the conditions mentioned above, foreground-versus-background discrimination (aka 'detection' of any individual) for chiffchaff was strong at over 95% AUC. Mel spectral features performed slightly better for this (range 96.6–98.6%) than learnt features (range 95.3–96.7%). Having considered this, in the remainder of the results, we focus on our main question of discriminating between individuals.

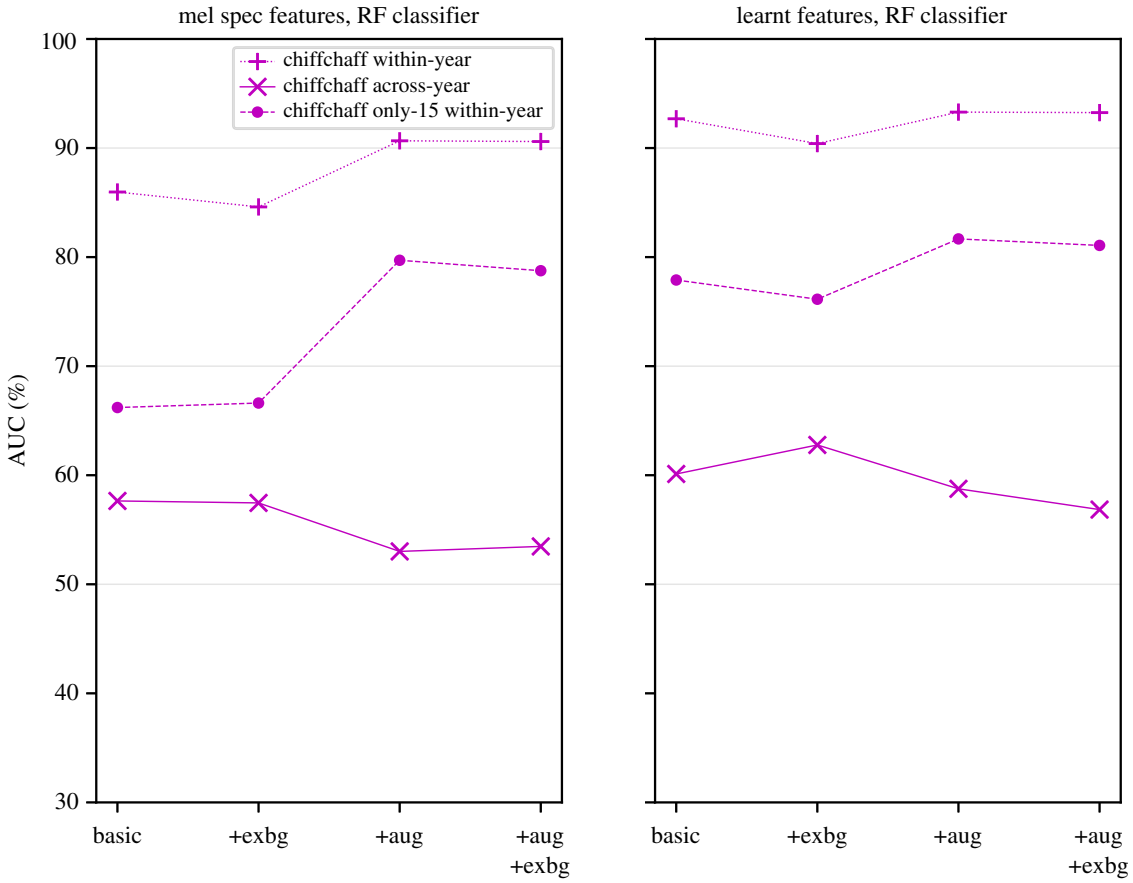


**Figure 5.** Performance of classifier (AUC) across the three chiffchaff evaluation scenarios, and with various combinations of configuration: with/without augmentation (aug), learnt features, and explicit-background (exbg) training. (Online version in colour.)

**Table 2.** Results of GLMM test for AUC, across the three chiffchaff evaluation scenarios.

| factor | estimate | *p*-value |
|---|---|---|
| (intercept) | 0.8199 | 0.041* |
| feature learning | 0.3093 | 0.014* |
| augmentation | 0.2509 | 0.048* |
| explicit-bg class | 0.0626 | 0.621 |

*$p < 0.05$.

**Table 3.** Results of GLMM fit for RMSE in the adversarial data augmentation test, across the three chiffchaff evaluation scenarios.

| factor | estimate | *p*-value |
|---|---|---|
| (intercept) | 1.8543 | $1.9 \times 10^{-05***}$ |
| feature-learning | −0.5044 | $1.9 \times 10^{-08***}$ |
| augmentation | −0.8734 | $< 2 \times 10^{-16***}$ |
| explicit-bg class | −0.0141 | 0.87 |

***$p < 0.001$.

Feature learning and structured data augmentation were both found to significantly improve classification performance (table 2) as well as robustness to adversarial data augmentation (table 3). Explicit-background training was found to lead to mild improvement but this was a long way below significance.

There were two possible interpretations for the benefit conferred by data augmentation: it could be due to our stratified data augmentation having the intended effect of reducing the foreground–background correlations in the data, or more simply due to the mere fact of training with a larger volume of data items. We expected both aspects to be implicated. To examine this *post hoc*, we created a smaller-yet-augmented training set for the chiffchaff within-year scenario: we took 50% of the items from the primary dataset, plus an equal number of items sampled (without replacement) from the augmented dataset, selected in such a way that this hybrid training set contained the same number of items per each individual as in the primary training set. When the AAII system was trained with this data, the AUC results for the learnt features gave the same strong performance as with full augmentation (92.6%). For the mel features, the AUC score of 89.9% indicated mild impairment relative to full augmentation, but stronger performance than the base unaugmented scenario.

## 3.2. Phase two: multiple species

Based on the results of our first study, we took forward an improved version of the classifier (using stratified data augmentation, and learnt features, but not explicit-background training) to test across multiple species.

Applying this classifier to the different species and conditions, we found that it led in most cases to a dramatic improvement in recognition performance of foreground recordings, and little change in the recognition of background recordings (figure 6 and table 4). This unchanged response to background recordings serves as evidence that the

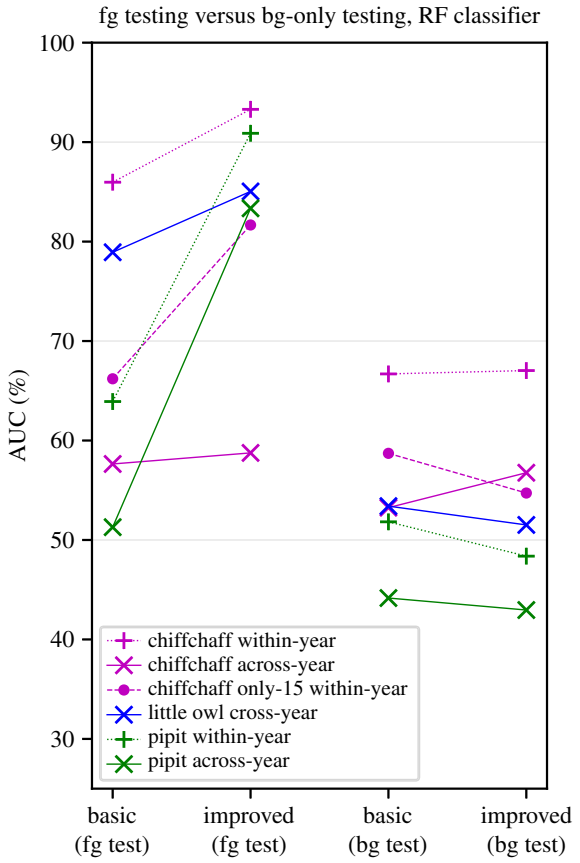


**Figure 6.** Our selected interventions—data augmentation and feature-learning—improve classification performance, in some cases dramatically (left-hand pairs of points), without any concomitant increase in the background-only classification (right-hand pairs of points) which would be an indication of counfounding. (Online version in colour).

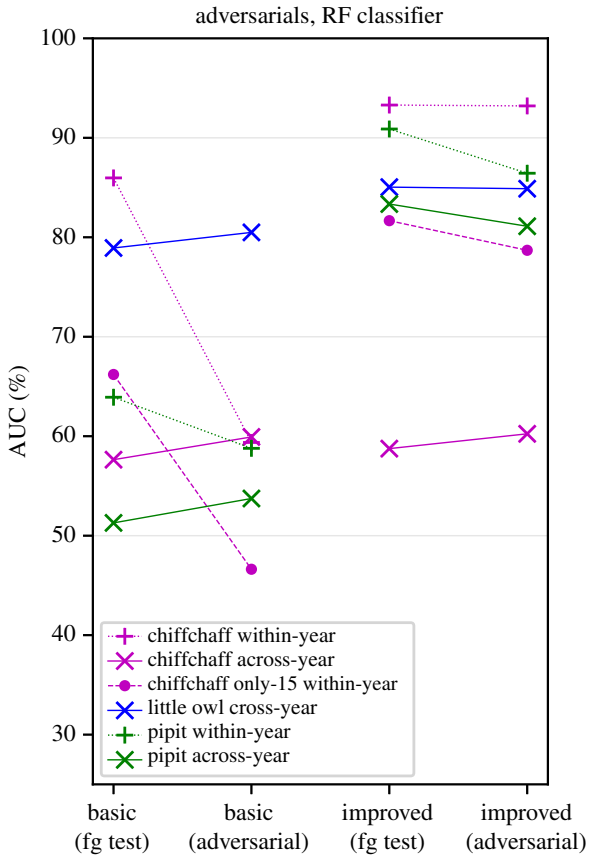


**Figure 7.** Adversarial augmentation has a varied impact on classifier performance (left-hand pairs of points), in some cases giving a large decline. Our selected interventions vastly reduce the impact of this adversarial test, while also generally improving classification performance (right-hand pairs of points). (Online version in colour)

**Table 4.** Results of GLMM test for AUC, across all three species, to quantify the general effect of our improvements on the foreground test and the background test (cf. figure 6).

| | estimate | *p*-value |
|---|---|---|
| (intercept) | 0.792 | 0.00150** |
| use of improved AAII system | 0.852 | 0.00032*** |
| background-only testing | −0.562 | 0.00624** |
| interaction term | −0.896 | 0.00391** |

**$p < 0.01$; ***$p < 0.001$.

**Table 5.** Results of GLMM test for AUC, across all three species, to quantify the general effect of our improvements on the adversarial test (cf. figure 7).

| | estimate | *p*-value |
|---|---|---|
| (intercept) | 0.873 | 0.0121* |
| use of improved AAII system | 0.820 | 0.0027** |
| adversarial data augmentation | −0.333 | 0.1713 |
| interaction term | 0.225 | 0.5520 |

*$p < 0.05$; **$p < 0.01$.

improvement is based on the individuals' signal characteristics and not confounding factors.

Our adversarial augmentation, intended as a diagnostic test to adversarially reduce classification performance, did not have strong overall effects on the headline performance indicated by the AUC scores (figure 7 and table 4). Half of the cases examined—the across-year cases—were not adversely impacted, in fact showing a very small increase in AUC score. The chiffchaff within-year tests were the only to show a strong negative impact of adversarial augmentation, and this negative impact was removed by our improved AAII system.

We also conducted a more fine-grained analysis of the effect of augmentation, by measuring the amount of deviation induced in the probabilities output from the classifier. On this measure, we observed a consistent effect, with our improvements reducing the RMS error by ratios of approximately 2–6, while the overall magnitude of the error differed across species (figure 8).

## 4. Discussion

We have demonstrated that a single approach to AAII can be successfully used across different species with different complexity of vocalizations. One exception to this is the hardest case, chiffchaff tested across years, in which automatic classification performance remains modest. The chiffchaff case (complex song, variable song content), in particular, highlights the need for proper assessment of identification performance. Without proper assessment, we cannot be sure if promising results reflect the real potential of proposed

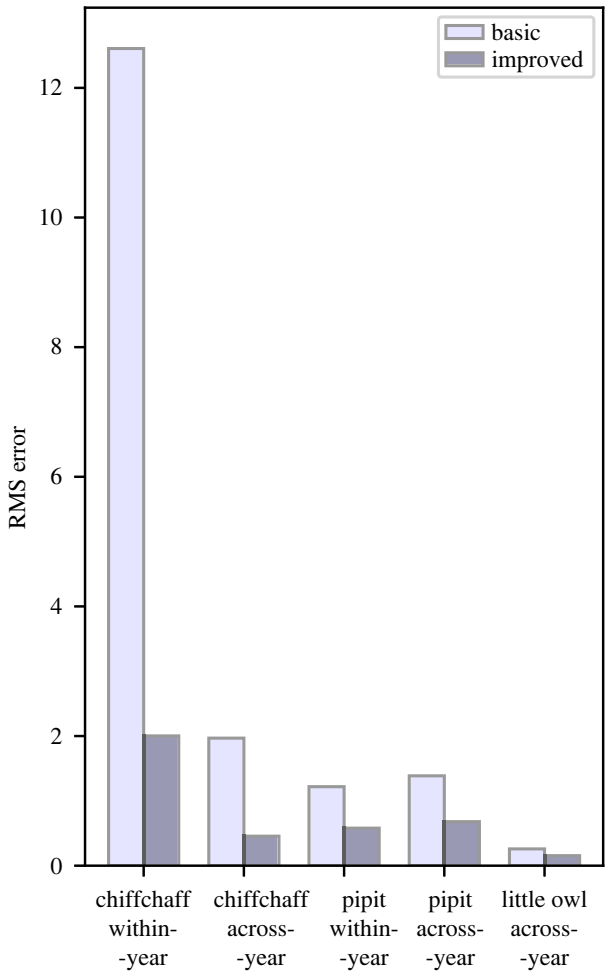


**Figure 8.** Measuring in detail how much effect the adversarial augmentation has on classifier decisions: RMS error of classifier output, in each case applying adversarial augmentation and then measuring the differences compared against the non-adversarial equivalent applied to the exact same data. In all five scenarios, our selected interventions lead to a large decrease in the RMS error. (Online version in colour.)

identification method. We document that our proposed improvements to the classifier training process are able, in some cases, to improve the generalization performance dramatically and, on the other hand, reveal confounds causing over-optimistic results.

We evaluated *spherical k-means* feature learning as previously used for species classification [59]. We found that for individual identification it provides an improvement over plain mel spectral features, not just in accuracy (as previously reported) but also in resistance to confounding factors. We believe this is due to the feature learning having been tailored to reflect fine temporal details of bird sound [59]; if so, this lesson would carry across to related systems such as convolutional neural networks. Our machine learning approach may be particularly useful for automatic identification of individuals in species with more complex songs, such as pipits (note huge increase in performance over mel features in figure 6), or chiffchaffs (on short-time scale though).

Using silence-regions from focal individuals to create an 'explicit-background' training category provided only a mild improvement in the behaviour of the classifier, under various evaluations. Also, we found that the best-performing configuration used for detecting the presence/absence of a focal individual was not the same as the best-performing configuration for discriminating between individuals. Hence, it seems generally preferable not to combine the detection and AAII tasks into one classifier.

By contrast, using silence-regions to perform dataset augmentation of the foreground sounds was found to give a strong boost to performance as well as resistance against confounding factors. This benefit was not universal—it was not the case for the difficult case of chiffchaff across-years—but was, in general, a strong factor in improved performance. Furthermore, the benefit was not eliminated when we reduced the augmented dataset back to its original size, indicating that the effect is indeed due to improved invariance to noise/confound, and not merely to increased sample size. Background sounds are useful in training a system for AAII, through data augmentation (rather than explicit-background training).

We found that adversarial augmentation provided a useful tool to diagnose concerns about the robustness of an AAII system. In the present work, we found that the classifier was robust against this augmentation (and thus we can infer that it was largely not using background confounds to make its decision), except for the case of chiffchaff with the simple mel features (figure 7). This latter case exhorts us to be cautious, and suggests that results from previous call-type independent methods may have been over-optimistic in assessing performance [34–37,42]. Our adversarial augmentation method can help to test for this even in the absence of across-year data.

Background-only testing was useful to confirm that when the performance of a classifier was improved, the confounding factors were not aggravated in parallel, i.e. that the improvement was due to signal and not confound (figure 6). However, the performance on background sound recordings was not reduced to chance, but remained at some level reflecting the foreground–background correlations in each case, so results need to interpreted comparatively against the foreground improvement, rather than in isolation. This individual specificity of the background may be related to the time interval between recordings. This is clear from the across-year outcomes; within-year, we note that there was one day of temporal separation for chiffchaffs (close to 70 per cent AUC on background-only sound), while an interval of weeks for pipits (chance-level classification of background). These effects surely depend on characteristics of the habitat.

Our improved AAII system performs much more reliably than the standard one; however, the most crucial factor still seems to be a targeted species. For the little owl we found good performance, and least affected by modifications in methods—consistent with the fact that it is the species with the simplest vocalizations. Little owl represents a species well suited for template matching individual identification methods which have been used in past for many species with similar simple, fixed vocalizations (discriminant analysis, cross-correlation). For these cases, it seems that our automatic identification method does not bring advantage regarding improved classification performance. However, a general classifier such as ours, automatically adjusting a set of features for each species, would allow common users to start individual identification right away without the need to choose an appropriate template-matching method (e.g. [54]).

We found that feature learning gave the best improvement in case of pipits (figure 6). Pipits have more complex song, where simple template matching cannot be used to identify individuals. In pipits, each song may have different duration

and may be composed of different subsets of syllable repertoire, and so any single song cannot be used as template for template matching approach. This singing variation likely also prevents good identification performance based on mel features in pipits. Nevertheless, a singing pipit male will cycle through the whole syllable repertoire within a relatively low number of songs and individual males can be identified based on their unique syllable repertoires ([27]). We think that our improvements to the automatic identification might allow the system to pick up correct features associated with stable repertoire of each male. This extends the use of the same automatic identification method to the large part of songbird species that organize songs into several song types and, at the same time, are so-called closed-ended learners ([63]).

Our automatic identification, however, cannot be considered fully independent of song content in a sense defined earlier (e.g. [34,36]). Such content-independent identification method should be able to classify across-year recordings of chiffchaffs in which syllable repertoires of males differ almost completely between the two years [52]. Owing to vulnerability of mel feature classification to confounds reported here and because the performance of content independent identification has been only tested on short-term recordings, we believe that the concept of fully content-independent individual identification has yet to be reliably demonstrated.

Our approach seems certainly to be suitable for species with individual vocalization stable over time, even if that vocalization is complex—i.e. for a very wide range of species. In future work, these approaches should also be tested with 'open-set' classifiers allowing for the possibility that new unknown individuals might appear in data. This is well-developed in the 'UBM' developed in GMM-based speaker recognition [42], and future work in machine learning is needed to develop this for the case of more powerful classifiers.

Important for further work in this topic is open sharing of data in standard formats. Only this way can diverse datasets from individuals be used to develop/evaluate automatic recognition that works across many taxa and recording conditions.

We conclude by listing the recommendations that emerge from this work for users of automatic classifiers, in particular for acoustic recognition of individuals:

(1) Record 'background' segments, for each individual (class), and publish background-only audio samples alongside the trimmed individual audio samples. Standard data repositories can be used for these purposes (e.g. Dryad, Zenodo).
(2) Improve robustness by:
   (a) suitable choice of input features;
   (b) structured data augmentation, using background sound recordings.
(3) Probe your classifier for robustness by:
   (a) background-only recognition: higher-than-chance recognition strongly implies confound;
   (b) adversarial distraction with background: a large change in classifier outputs implies confound;
   (c) across-year testing (if such data are available): a stronger test than within-year.
(4) Be aware of how species characteristics will affect recognition. The vocalization characteristics of the species will influence the ease with which automatic classifiers can identify individuals. Songbirds whose song changes within and between seasons will always be harder to identify reliably—as is also the case in manual identification.
(5) Best practice is to test manual features and learned features since the generalization and performance characteristics are rather different. In the present work, we compare basic features against learned features; for a different example see [12]. Manual features are usually of lower accuracy, but with learned features more care must be taken with respect to confounds and generalization.

Ethics. Our study primarily involved only non-invasive recording of vocalizing individuals. In the case of ringed individuals (all chiffchaffs and some tree pipits and little owls), ringing was done by experienced ringers (P.L., M.Š., T.P.) who all held ringing licences at the time of the study. Tree pipits and chiffchaff males were recorded during spontaneous singing. Only for little owls, short playback recording (1 min) was used to provoke calling. Playback provocations as well as handling during ringing were kept as short as possible and we are not aware of any consequences for subjects' breeding or welfare.

Data accessibility. Our audio data and the associated metadata files are available online under the Creative Commons Attribution licence (CC BY 4.0) at http://doi.org/10.5281/zenodo.1413495

Authors' contributions. D.S. and P.L. conceived and designed the study. P.L., T.P. and M.Š. recorded audio. P.L. processed the audio recordings into data sets. D.S. carried out the classification experiments and performed data analysis. D.S., P.L. and T.P. wrote the manuscript. All authors gave final approval for publication.

Competing interests. We declare we have no competing interests.

Funding. D.S. was supported by EPSRC Early Career research fellowship EP/L020505/1. P.L. was supported by the National Science Centre, Poland, under Polonez fellowship reg. no UMO-2015/19/P/NZ8/02507 funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 665778. T.P. was supported by the Czech Science Foundation (project P505/11/P572). M.Š. was supported by the research aim of the Czech Academy of Sciences (RVO 68081766).

## Endnotes

[1]Command used: `sox -m [fgfile] [bgfile] [outfile] trim 0 [fgduration]`

[2]This feature learning increases the feature dimensionality from 40 to 500. High-dimensional data can lead to difficulties in some analysis, but are not intrinsically a problem for modern classification algorithms. Note that [59] evaluated whether the difference in dimensionality alone was sufficient to explain the strong performance of feature learning, by performing an additional experiment projecting all data to fixed dimension; they found that it explained only a small component of the effect.

[3]Implemented using scikit-learn version 0.20.2. Parameter settings for the random forest used 200 trees and the 'entropy' criterion for splitting branches, as was chosen in the previous research. In the present work, we did not perform hyperparameter tuning, since the purpose was not to optimize one particular classifier but to evaluate interventions that could assist classifiers in general.

## References

1. Amorim MCP, Vasconcelos RO. 2008 Variability in the mating calls of the Lusitanian toadfish *Halobatrachus didactylus*: cues for potential individual recognition. *J. Fish Biol.* **73**, 1267–1283. (doi:10.1111/jfb.2008.73.issue-6)

2. Bee MA, Gerhardt HC. 2001 Neighbour-stranger discrimination by territorial male bullfrogs (*Rana catesbeiana*): I. Acoustic basis. *Anim. Behav.* **62**, 1129–1140. (doi:10.1006/anbe.2001.1851)
3. Terry AM, Peake TM, McGregor PK. 2005 The role of vocal individuality in conservation. *Front. Zool.* **2**, 10. (doi:10.1186/1742-9994-2-10)
4. Taylor AM, Reby D. 2010 The contribution of source-filter theory to mammal vocal communication research. *J. Zool.* **280**, 221–236. (doi:10.1111/j.1469-7998.2009.00661.x)
5. Gamba M, Favaro L, Araldi A, Matteucci V, Giacoma C, Friard O. 2017 Modeling individual vocal differences in group-living lemurs using vocal tract morphology. *Curr. Zool.* **63**, 467–475. (doi:10.1093/cz/zox023)
6. Janik V, Slater PB. 1997 *Vocal learning in mammals*, vol. 26, pp. 59–99. New York, NY: Academic Press.
7. Wiley RH. 2013 Specificity and multiplicity in the recognition of individuals: implications for the evolution of social behaviour. *Biol. Rev.* **88**, 179–195. (doi:10.1111/brv.2013.88.issue-1)
8. Boeckle M, Bugnyar T. 2012 Long-term memory for affiliates in ravens. *Curr. Biol.* **22**, 801–806. (doi:10.1016/j.cub.2012.03.023)
9. Insley SJ. 2000 Long-term vocal recognition in the northern fur seal. *Nature* **406**, 404–405. (doi:10.1038/35019064)
10. Briefer EF, de la Torre MP, McElligott AG. 2012 Mother goats do not forget their kids' calls. *Proc. R. Soc. B* **279**, 3749–3755. (doi:10.1098/rspb.2012.0986)
11. Slabbekoorn H. 2004 Singing in the wild: the ecology of birdsong. In *Nature's music: the science of birdsong* (eds P Marler, H Slabbekoorn), pp. 178–205. San Diego, CA: Elsevier Academic Press.
12. Mouterde SC, Elie JE, Theunissen FE, Mathevon N. 2014 Learning to cope with degraded sounds: female zebra finches can improve their expertise at discriminating between male voices at long distance. *J. Exp. Biol.* **217**, 3169–3177. (doi:10.1242/jeb.104463)
13. Gambale PG, Signorelli L, Bastos RP. 2014 Individual variation in the advertisement calls of a Neotropical treefrog (*Scinax constrictus*). *Amphibia-Reptilia* **35**, 271–281. (doi:10.1163/15685381-00002949)
14. Collins SA. 2004 Vocal fighting and flirting: the functions of birdsong. In *Nature's music: the science of birdsong* (eds PR Marler, H Slabbekoorn), pp. 39–79. Elsevier Academic Press.
15. Linhart P, Jaška P, Petrusková T, Petrusek A, Fuchs R. 2013 Being angry, singing fast? Signalling of aggressive motivation by syllable rate in a songbird with slow song. *Behav. Processes* **100**, 139–145. (doi:10.1016/j.beproc.2013.06.012)
16. Kroodsma DE. 2004 The diversity and plasticity of bird song. In *Nature's music: the science of birdsong* (eds PR Marler, H Slabbekoorn), pp. 108–131. Elsevier Academic Press.
17. Thom MDF, Dytham C. 2012 Female choosiness leads to the evolution of individually distinctive males. *Evolution* **66**, 3736–3742. (doi:10.1111/evo.2012.66.issue-12)
18. Bradbury JW, Vehrencamp SL. 1998 *Principles of animal communication*, 1st edn. Sunderland, MA: Sinauer Associates.
19. Crowley PH, Provencher L, Sloane S, Dugatkin LA, Spohn B, Rogers L, Alfieri M. 1996 Evolving cooperation: the role of individual recognition. *Biosystems* **37**, 49–66. (doi:10.1016/0303-2647(95)01546-9)
20. Mennill DJ. 2011 Individual distinctiveness in avian vocalizations and the spatial monitoring of behaviour. *Ibis* **153**, 235–238. (doi:10.1111/j.1474-919X.2011.01119.x)
21. Blumstein DT *et al.* 2011 Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *J. Appl. Ecol.* **48**, 758–767. (doi:10.1111/j.1365-2664.2011.01993.x)
22. Johnsen A, Lifjeld J, Rohde PA. 1997 Coloured leg bands affect male mate-guarding behaviour in the bluethroat. *Anim. Behav.* **54**, 121–130. (doi:10.1006/anbe.1996.0437)
23. Gervais JA, Catlin DH, Chelgren ND, Rosenberg DK. 2006 Radiotransmitter mount type affects burrowing owl survival. *J. Wildl. Manag.* **70**, 872–876. (doi:10.2193/0022-541X(2006)70[872:RMTABO]2.0.CO;2)
24. Linhart P, Fuchs R, Poláková S, Slabbekoorn H. 2012 Once bitten twice shy: long-term behavioural changes caused by trapping experience in willow warblers *Phylloscopus trochilus*. *J. Avian Biol.* **43**, 186–192. (doi:10.1111/j.1600-048X.2012.05580.x)
25. Rivera-Gutierrez HF, Pinxten R, Eens M. 2015 Songbirds never forget: long-lasting behavioural change triggered by a single playback event. *Behaviour* **152**, 1277–1290. (doi:10.1163/1568539X-00003278)
26. Camacho C, Canal D, Potti J. 2017 Lifelong effects of trapping experience lead to age-biased sampling: lessons from a wild bird population. *Anim. Behav.* **130**, 133–139. (doi:10.1016/j.anbehav.2017.06.018)
27. Petrusková T, Pišvejcová I, Kinštová A, Brinke T, Petrusek A. 2015 Repertoire-based individual acoustic monitoring of a migratory passerine bird with complex song as an efficient tool for tracking territorial dynamics and annual return rates. *Methods Ecol. Evol.* **7**, 274–284. (doi:10.1111/2041-210X.12496)
28. Laiolo P, Vögeli M, Serrano D, Tella JL. 2007 Testing acoustic versus physical marking: two complementary methods for individual-based monitoring of elusive species. *J. Avian Biol.* **38**, 672–681. (doi:10.1111/jav.2007.38.issue-6)
29. Kirschel ANG, Cody ML, Harlow ZT, Promponas VJ, Vallejo EE, Taylor CE. 2011 Territorial dynamics of Mexican Ant-thrushes *Formicarius moniliger* revealed by individual recognition of their songs. *Ibis* **153**, 255–268. (doi:10.1111/j.1474-919X.2011.01102.x)
30. Spillmann B, van Schaik CP, Setia TM, Sadjadi SO. 2017 Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (*Pongo pygmaeus wurmbii*) long calls. *Bioacoustics* **26**, 109–120. (doi:10.1080/09524622.2016.1216802)
31. Delport W, Kemp AC, Ferguson JWH. 2002 Vocal identification of individual African Wood Owls *Strix woodfordii*: a technique to monitor long-term adult turnover and residency. *Ibis* **144**, 30–39. (doi:10.1046/j.0019-1019.2001.00019.x)
32. Adi K, Johnson MT, Osiejuk TS. 2010 Acoustic censusing using automatic vocalization classification and identity recognition. *J. Acoust. Soc. Am.* **127**, 874–883. (doi:10.1121/1.3273887)
33. Terry AMR, McGregor PK. 2002 Census and monitoring based on individually identifiable vocalizations: the role of neural networks. *Anim. Conserv.* **5**, 103–111. (doi:10.1017/S1367943002002147)
34. Fox EJS. 2008 A new perspective on acoustic individual recognition in animals with limited call sharing or changing repertoires. *Anim. Behav.* **75**, 1187–1194. (doi:10.1016/j.anbehav.2007.11.003)
35. Fox EJS, Roberts JD, Bennamoun M. 2008 Call-independent individual identification in birds. *Bioacoustics* **18**, 51–67. (doi:10.1080/09524622.2008.9753590)
36. Cheng J, Sun Y, Ji L. 2010 A call-independent and automatic acoustic system for the individual recognition of animals: a novel model using four passerines. *Pattern Recognit.* **43**, 3846–3852. (doi:10.1016/j.patcog.2010.04.026)
37. Cheng J, Xie B, Lin C, Ji L. 2012 A comparative study in birds: call-type-independent species and individual recognition using four machine-learning methods and two acoustic features. *Bioacoustics* **21**, 157–171. (doi:10.1080/09524622.2012.669664)
38. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. 2013 Intriguing properties of neural networks. (http://arxiv.org/abs/13126199).
39. Mesaros A, Heittola T, Virtanen T. 2018 Acoustic Scene Classification: an Overview of DCASE 2017 Challenge Entries. In *16th Int. Workshop on Acoustic Signal Enhancement (IWAENC)*. Tokyo, Japan.
40. Khanna H, Gaunt S, McCallum D. 1997 Digital spectrographic cross-correlation: tests of sensitivity. *Bioacoustics* **7**, 209–234. (doi:10.1080/09524622.1997.9753332)
41. Foote JR, Palazzi E, Mennill DJ. 2013 Songs of the Eastern Phoebe, a suboscine songbird, are individually distinctive but do not vary geographically. *Bioacoustics* **22**, 137–151. (doi:10.1080/09524622.2012.740174)
42. Ptáček L, Machlica L, Linhart P, Jaška P, Muller L. 2016 Automatic recognition of bird individuals on an open set using as-is recordings. *Bioacoustics* **25**, 55–73. (doi:10.1080/09524622.2015.1089524)
43. Stowell D, Stylianou Y, Wood M, Pamuła H, Glotin H. 2019 Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. *Methods Ecol. Evol.* **10**, 363–380. (doi:10.1111/2041-210X.13103)
44. Knight EC, Hannah KC, Foley GJ, Scott CD, Brigham RM, Bayne E. 2017 Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conserv. Ecol.* **12**, 14. (doi:10.5751/ACE-01114-120214)

45. Joly A, Goëau H, Botella C, Glotin H, Bonnet P, Vellinga WP *et al.* 2018 Overview of LifeCLEF 2018: a large-scale evaluation of species identification and recommendation algorithms in the Era of AI. In *Int. Conf. of the Cross-Language Evaluation Forum for European Languages*, pp. 247–266. Berlin, Germany: Springer.
46. Stowell D. 2018 Computational bioacoustic scene analysis. In *Computational analysis of sound scenes and events*, pp. 303–333. Berlin, Germany: Springer.
47. Morfi V, Stowell D. 2018 Deep learning for audio event detection and tagging on low-resource datasets. *Appl. Sci.* **8**, 1397. (doi:10.3390/app8081397)
48. Salamon J, Bello JP. 2017 Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process Lett.* **24**, 279–283. (doi:10.1109/LSP.2017.2657381)
49. Lasseck M. 2018 Audio-based bird species identification with deep convolutional neural networks. Working Notes of CLEF. 2018.
50. Grava T, Mathevon N, Place E, Balluet P. 2008 Individual acoustic monitoring of the European Eagle Owl *Bubo bubo*. *Ibis* **150**, 279–287. (doi:10.1111/j.1474-919X.2007.00776.x)
51. Petrusková T, Osiejuk TS, Linhart P, Petrusek A. 2008 Structure and complexity of perched and flight songs of the tree pipit (*Anthus trivialis*). *Ann. Zool. Fennici* **45**, 135–148. (doi:10.5735/086.045.0205)
52. Průchová A, Jaška P, Linhart P. 2017 Cues to individual identity in songs of songbirds: testing general song characteristics in Chiffchaffs *Phylloscopus collybita*. *J. Ornithol.* **158**, 911–924. (doi:10.1007/s10336-017-1455-6)
53. Nieuwenhuyse DV, Génot JC, Johnson DH. 2008 *The little owl: conservation, ecology and behavior of Athene noctua*. Cambridge, UK: Cambridge University Press.
54. Linhart P, Šálek M. 2017 The assessment of biases in the acoustic discrimination of individuals. *PLoS ONE* **12**, e0177206. (doi:10.1371/journal.pone.0177206)
55. Krizhevsky A, Sutskever I, Hinton GE. 2012 ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, pp. 1097–1105. See http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
56. Cireşan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. (http://arxiv.org/abs/12022745).
57. Schlüter J, Grill T. 2015 Exploring data augmentation for improved singing voice detection with neural networks. In *Proc. of the Int. Conf. on Music Information Retrieval (ISMIR)*, pp. 121–126.
58. Liang D, Yang F, Zhang T, Yang P. 2018 Understanding mixup training methods. *IEEE Access* **6**, 58 774–58 783. (doi:10.1109/ACCESS.2018.2872698)
59. Stowell D, Plumbley MD. 2014 Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ* **2**, e488. (doi:10.7717/peerj.488)
60. Fawcett T. 2006 An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874. (doi:10.1016/j.patrec.2005.10.010)
61. Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, Nielsen A, Sibert J. 2012 AD model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optim. Methods Softw.* **27**, 233–249. (doi:10.1080/10556788.2011.597854)
62. Skaug H, Fournier D, Bolker B, Magnusson A, Nielsen A. Generalized linear mixed models using 'AD Model Builder'; 2016-01-19. R package version 0.8.3.3.
63. Beecher MD, Brenowitz EA. 2005 Functional aspects of song learning in songbirds. *Trends Ecol. Evol.* **20**, 143–149. (doi:10.1016/j.tree.2005.01.004)