Robust sound event detection in bioacoustic sensor networks

Vincent Lostanlen[1,2,3*], Justin Salamon[2,3], Andrew Farnsworth[1], Steve Kelling[1], and Juan Pablo Bello[2,3]

1 Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA
2 Music and Audio Research Laboratory, New York University, New York, NY, USA
3 Center for Urban Science and Progress, New York University, New York, NY, USA

* vincent.lostanlen@nyu.edu

Abstract

Bioacoustic sensors, sometimes known as autonomous recording units (ARUs), can record sounds of wildlife over long periods of time in scalable and minimally invasive ways. Deriving per-species abundance estimates from these sensors requires detection, classification, and quantification of animal vocalizations as individual acoustic events. Yet, variability in ambient noise, both over time and across sensors, hinders the reliability of current automated systems for sound event detection (SED), such as convolutional neural networks (CNN) in the time-frequency domain. In this article, we develop, benchmark, and combine several machine listening techniques to improve the generalizability of SED models across heterogeneous acoustic environments. As a case study, we consider the problem of detecting avian flight calls from a ten-hour recording of nocturnal bird migration, recorded by a network of six ARUs in the presence of heterogeneous background noise. Starting from a CNN yielding state-of-the-art accuracy on this task, we introduce two noise adaptation techniques, respectively integrating short-term (60 ms) and long-term (30 min) context. First, we apply per-channel energy normalization (PCEN) in the time-frequency domain, which applies short-term automatic gain control to every subband in the mel-frequency spectrogram. Secondly, we replace the last dense layer in the network by a context-adaptive neural network (CA-NN) layer, i.e. an affine layer whose weights are dynamically adapted at prediction time by an auxiliary network taking long-term summary statistics of spectrotemporal features as input. We show that both techniques are helpful and complementary: while PCEN reduces temporal overfitting across dawn vs. dusk audio clips, context adaptation reduces spatial overfitting across sensor locations. Moreover, combining them yields state-of-the-art results that are unmatched by artificial data augmentation alone. We release a pre-trained version of our best performing system under the name of BirdVoxDetect, a ready-to-use detector of avian flight calls in field recordings.

Introduction

Machine listening for large-scale bioacoustic monitoring

The past decades have witnessed a steady decrease in the hardware costs of sound acquisition [1], processing [2], transmission [3], and storage [4]. As a result, the application domain of digital audio technologies has extended far beyond the scope of interhuman communication to encompass the development of new cyberphysical systems [5]. In particular, passive acoustic sensor networks, either terrestrial or underwater, contribute to meet certain challenges of industrialized societies, including wildlife conservation [6], urban planning [7], and the risk assessment of meteorological disasters [8].

Biodiversity monitoring is one of the most fruitful applications of passive acoustics. Indeed, in comparison with optical sensors, acoustic sensors are minimally invasive [9], have a longer detection range — from decameters for a flock of migratory birds to thousands of kilometers for an oil exploration airgun [10] — and their reliability is independent of the amount of daylight [11]. In this context, one emerging application is the species-specific inventory of vocalizing animals [12], such as birds [13], primates [14], and marine mammals [15], whose occurrence in time and space reflects the magnitude of population movements [16], and can be correlated with other environmental variables, such as local weather [17].

The principal motivation for this article is to monitor bird migration by means of a bioacoustic sensor network [18]. From one year to the next, each species is susceptible to alter its migratory onset and route as a function of both intrinsic factors [19] and extrinsic (e.g. human-caused) environmental pressures [20, 21]. Mapping in real time [22], and even forecasting [23, 24], the presence and quantity of birds near hazardous sites (e.g. airports [25], windfarms [26], and dense urban areas [27]) would enable appropriate preventive measures for avian wildlife conservation, such as temporary reduction of light pollution [28]. In addition, it could benefit civil aviation safety as well as agricultural planning [29].

At present, monitoring nocturnal bird migration is a challenge of integrating complementary methods to try to produce the most comprehensive understanding of migrants' movements. The two most readily available sources of information for tracking the movements of avian populations at large (e.g. continental) scales are weather surveillance radar data [30] and crowdsourced observations of birders [31]. Both of these information sources are valuable but imperfect. In particular, the former does not distinguish different species, rather providing data only on bird biomass aloft. Conversely, the latter is dominated by diurnal information, which does not describe spatial and temporal distribution of species when they are actively migrating at night, and is sparse, requiring state-of-the-art computational approaches to produce distribution models. In contrast, flight calls can provide species information, at the least for vocal species; and may, in principle, be detected in real time [32]. Supplementing spatiotemporal exploratory models [33], currently restricted to radar and observational modalities [34], with the output of a bioacoustic sensor network, could improve our ability to detect species flying over the same area simultaneously, and offer new insights in behavioral ecology and conservation science.

Despite the aforementioned assets of using bioacoustic monitoring, rather than other sources of data, in order to monitor bird migration, the lack of highly accurate

In a large-scale setting of bioacoustic monitoring at the continental scale and over multi-month migration seasons, the task of counting individual vocalizations in continuous recordings by human annotators to achieve these ends is impractical, unsustainable, and unscalable. Rather, there is dire need for a fully automated solution to avian flight call detection [35], that would rely on machine listening, i.e. the auditory analogue of computer vision [36]. We propose that, in the future, each sensor could run autonomously [37], by sending hourly digests of bird vocalization activity to the central server, which in turn would aggregate information from all sensors, ultimately resulting in a spatiotemporal forecast of nocturnal migration [38].

The detection of far-field signals despite the presence of background noise constitutes a fundamental challenge for bioacoustic sensor networks [39]. Whereas, in a typical fieldwork setting, a human recordist would use a directional microphone and point it towards a source of interest, thus minimizing background noise or other interference [40], autonomous recording units (ARUs) are most often equipped with single omnidirectional microphones [41]. As a result, instead of tracking the sources of interest, they capture a global *soundscape* (sonic landscape) of their environment [42], which also comprises spurious sources of noise [43]. Furthermore, in the context of avian flight calls, migratory birds move rapidly, vocalize intermittently, and may simultaneously be present at multiple azimuths around a sensor [44]. Consequently, none of the well-established methods for beamforming animal vocalizations — which assume that each sensor combines multiple directional microphones [45] — would apply to the use case of flight

call monitoring. On the contrary, we formulate a scenario in which sound event detection occurs in natural soundcapes without prior localization of sources. This formulation represents a potential use case for the deployment of a large-scale bioacoustic sensor network consisting of low-cost, single-microphone hardware [46].

Because migratory birds appear to vocalize at a relatively low acoustic intensity and at a relatively high distance to the sensor [47], simple energy-based detection functions [48] or spectrotemporal template matching [49] may be inadequate for solving problems of retrieving avian flight calls in continuous recordings. Instead, machine learning appears necessary for detecting acoustic events in noisy, highly reverberant environments [50]. Yet, one fundamental assumption behind conventional machine learning methods is that samples from the training set and samples from the test set are drawn from the same high-dimensional probability distribution.

In the specific case of bioacoustic sensor networks, a training set may consist of audio clips from a limited number of recordings that are manually annotated a priori, whereas the test set will encompass a broader variety of recording conditions, including days, sensor locations, and seasons that are unreviewed or unlabeled [51]. Although it is plausible to assume that, from one recording condition to another, the statistical properties of the flight calls themselves — hereafter denoted as foreground — are identically or almost identically distributed, the same cannot be said of background sources of noise. Rather, natural soundscapes, even at the spatial scale of a few square kilometers and at the temporal scale of a few hours, may exhibit large variations in background noise spectra [52]. Therefore, state-of-the-art machine learning systems for sound event detection, once trained on the far-field recordings originating from a limited number of sensors, might fail to generalize once deployed on a different sensor [53].

The current lack of robust methods for sound event detection in heterogeneous environments have caused past bioacoustic studies to focus on relatively few acoustic sensors in close proximity [54, 55]. Nevertheless, the goal of deploying a large-scale network of acoustic sensors for avian migration monitoring requires sound event detection to adapt to nonstationarities (i.e. variations in time) and nonuniformities (i.e. variations in space) of background noise. In this article, we propose a combination of novel methods, not only to improve the accuracy of state-of-the-art detectors on average, but also to make these detectors more reliable across recording conditions, such as those arising at dawn vs. dusk or across different sensor locations.

Evidence of technical bias in state-of-the-art bioacoustic detection systems

For example, Figure 1 illustrates the challenges of a state-of-the-art sound event detector of nocturnal flight calls, namely the convolutional neural network architecture of [56], hereafter called “CNN baseline” in this paper. In the top plot, which is replicated from a previous study [57], the authors measured the evolution of recall of the CNN baseline over a publicly available machine listening benchmark for avian flight call detection, named BirdVox-full-night. This benchmark consists of six continuous recordings of flight calls, corresponding to six different autonomous recording units; it will be described in further detail in the Methods section. Over the course of ten hours, the CNN baseline system exhibits large variabilities in recall, i.e. fraction of detected events that are true positives, through time: both within the vocal ranges of thrushes (from 0 to 5 kHz) and of warblers and sparrows (from 5 to 10 kHz), recall oscillates between 5% and 35% during dusk and night before soaring rapidly up to 75%.

One explanation for these variations lies in the unequal amount of available training data in function of recording conditions: as shown in the middle plot of Figure 1, the average number of flight calls per minute increases with time. Because its loss function assigns the same importance to every misclassified example, the baseline CNN model overfits dawn audio clips and underfits dusk audio clips.

The nonstationarity of background noise, at the time scale of a full night, aggravates the phenomenon of overfitting of this machine learning system. In the bottom plot of Figure 1, we extract the evolution of sound pressure level (SPL) within a narrow subband of interest (between

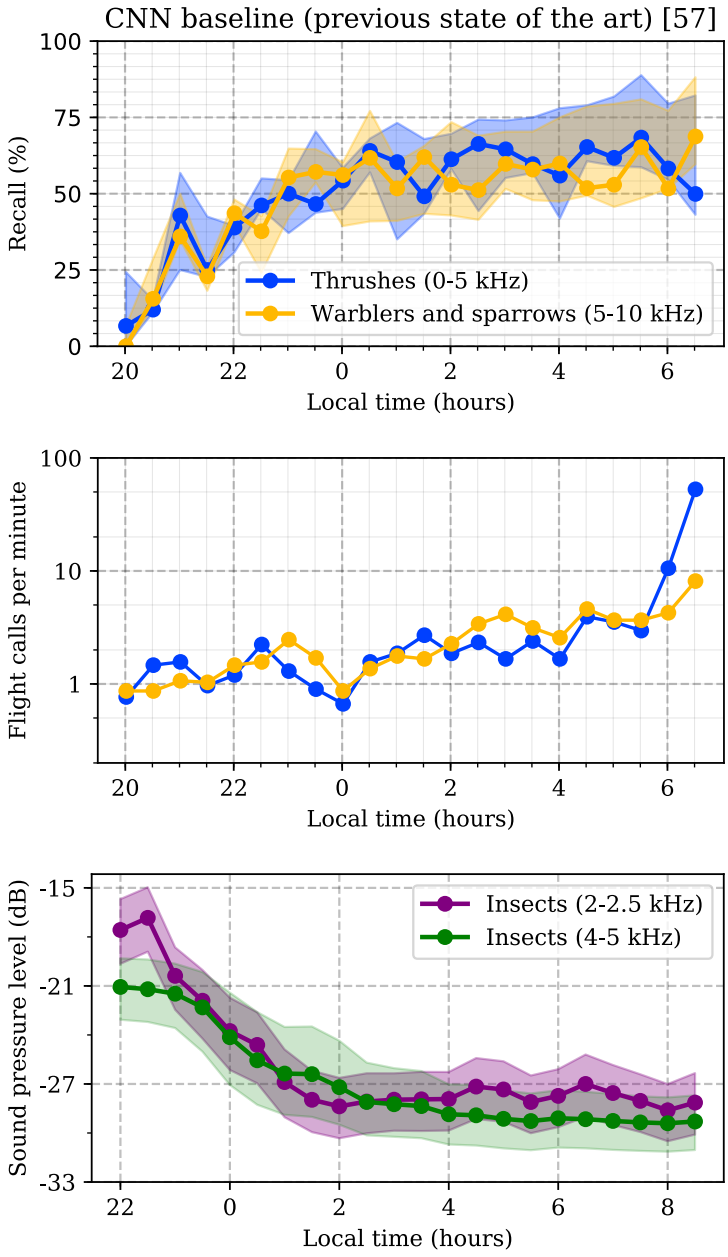

Fig 1. On BirdVox-full-night, the recall of the baseline CNN increases with time, as the density of flight calls increases and the noise level decreases. Shaded areas denote interquartile variations across sensors.

2 and 2.5 kHz), as well as within the subband corresponding to its second harmonic (i.e. between 4 and 5 kHz). Both correspond to the frequency range of stridulating insects in the background. In both subbands, we find that the median SPL, as estimated over 30-minute temporal windows, decreases by about 10 dB between 8 p.m. and midnight. This is because insect stridulations are most active at dusk, before fading out gradually.

The large variations in accuracy through time exhibited above are particularly detrimental when applying the baseline CNN detector for bird migration monitoring. Indeed, deploying this baseline CNN detector over a bioacoustic sensor network will likely lead to a systematic underestimation of vocal activity of migratory birds at dusk and an overestimation at dawn. This is a form of technical bias that, if left unchecked, might lead to wrong conclusions about the behavioral ecology and species composition of nocturnally migrating birds. Furthermore, and perhaps worse, such bias could create a foundation for conservation science that, contrary to original ambitions, is not based on the actual distribution and attributes of the target species of concern.

Related work

To the best of our knowledge, the only computational system for long-term bird migration monitoring that currently relies on acoustic sensor data is Vesper [58]. In order to detect thrushes, warblers, and sparrows, Vesper implements algorithms originally described in [48] that do not adapt dynamically to the changes in background noise described above. Instead, these detectors employ a measure of spectral flux [59] within manually defined passbands, associated with some *ad hoc* constraints on the minimal and maximal duration of a flight call. This straightforward and computationally elegant approach has been a standard for many in the amateur, academic, and professional migration monitoring communities. Yet, despite its simplicity and computational efficiency, such algorithms suffer from considerable shortcomings in detection accuracy, and may not be a reliable replacement for human inspection. In particular, a previous evaluation campaign showed that these detectors can exhibit precision and recall metrics both below 10% in a multi-sensor setting [57].

Another line of research that is related to this article is that of "bird detection in audio" [60], i.e. a yearly challenge during which machine listening researchers train systems for general-purpose detection of vocalizations over a public development dataset, and then compete for maximal accuracy over a private evaluation dataset. In recent years, the organizers of this challenge have managed to attract researchers from the machine learning and music information retrieval (MIR) communities [61]. This had led to the publication of new applications of existing machine learning methods to the domain of avian bioacoustics: these include multiple instance learning [62], convolutional recurrent neural networks [63], and densely connected convolutional networks [64]. Despite its undeniable merit of having gathered several data collection initiatives into a single cross-collection evaluation campaign, the methodology of the "bird detection in audio" challenge suffers from a lack of interpretability in the discussion of results *post hoc*. Indeed, because bird vocalizations are not annotated at the time scale of individual acoustic events but at the time scale of acoustic scenes, it is impossible to draw a relationship of proportionality between the average miss rates of competing systems and their respective technical biases, in terms of robustness to nonstationarity and nonuniformity of background noise. Furthermore, because these acoustic scenes are presented to the competitors under the forms of ten-second audio segments, rather than continuous recordings of several hours, the development and evaluation of some context-adaptive machine listening methods, such as the ones relying on spectrotemporal summary statistics for modeling background noise, remain out of the scope of practical applicability.

Contributions

The aim of this article is to improve the reliability of state-of-the-art sound event detection algorithms across acoustic environments, thus mitigating the technical bias incurred by nonstationarity and nonuniformity in background noise. We present four contributions to address this problem.

First, we develop a new family of neural network architectures for sound event detection in heterogeneous environments. The commonality among these architectures is that they comprise an auxiliary subnetwork that extracts a low-dimensional representation of background noise and incorporates it into the decision function of the main subnetwork. As such, they resemble context-adaptive neural networks (CA-NNs), i.e. an existing line of research in automatic speech recognition from multichannel audio input [65]. Yet, our CA-NN architectures differ from the current literature, both in the choice of auxiliary features and in the choice of mathematical formulation of the context-adaptive layer. We introduce long-term spectral summary statistics as auxiliary features for representing acoustic environments, whereas previous publications [66] relied on short-term spatial diffuseness features [67]. Furthermore, we generalize the mathematical formulation of context adaptation — initially described as a mixture-of-experts multiplicative gate [68] — within the broader topic of dynamic filter networks [69], and especially discuss the cases of context-adaptive dense layers with dynamic weights or with dynamic biases.

Second, we apply a new time-frequency representation to bioacoustic signal detection. Known as per-channel energy normalization (PCEN), this representation was recently proposed with the aim of improving robustness to channel distortion in a task of keyword spotting [70]. In this article, we demonstrate that, after we reconfigure its intrinsic parameters appropriately, PCEN also proves to be useful in outdoor acoustic environments. Indeed, we find that it enhances trasient events (e.g. bird vocalizations) while discarding stationary noise (e.g. insect stridulations). To the best of our knowledge, this article is the first in successfully applying PCEN to the analysis of non-speech data.

Third, we conduct a thorough evaluation of the respective effects of each component in the development of a deep convolutional network for robust sound event detection: presence of artificial data augmentation; choice of time-frequency representation (PCEN vs. logarithm of the mel-frequency spectrogram); and formulation of context adaptation. The overall computational budget that is incurred by this thorough evaluation is of the order of 10 CPU-years. After summarizing the results of our benchmark, we provide conclusive evidence to support the claim that CA-CNN and PCEN, far from interchangeable, are in fact complementary. Furthermore, they lead to improvements in robustness that are umatched by artificial data augmentation alone.

Finally, we combine all our findings into a deep learning system for avian flight call detection in continuous recordings. This system is named BirdVoxDetect, is written in the Python language, and is released under the MIT free software license[1]. This open source initiative is directed towards the machine listening community, in order to allow the extension of our research beyond its current application setting. In addition, we release our best performing BirdVoxDetect model under the form of a command-line interface, which segments and exports all detected sound events as separate audio clips, thus faciliating further inspection or automatic processing. This interface is directed towards the avian bioacoustics community, in order to allow the large-scale deployment of autonomous recording units for flight call monitoring. In an effort of conducting transparent, sustainable, and reproducible audio research [71], BirdVoxDetect also comprises documentation, a test suite, a Python package indexation, and an interoperable application programming interface (API).

[1] Open source repository for downloading BirdVoxDetect: https://github.com/BirdVox/birdvoxdetect

Methods

Overview

All methods presented herein rely on machine learning. Therefore, their comparison entails a training stage followed by a prediction stage. Figure 2 illustrates both stages schematically.

First, we formulate the training stage as binary classification of presence vs. absence of a sound event. In this setting, the input to the system is a short audio clip, whose duration is equal to 150 ms. We represent this audio clip by a time-frequency representation $\mathbf{E}(t, f)$. In the state-of-the-art model of [57], the matrix $\mathbf{E}(t, f)$ contains the magnitudes in the mel-frequency spectrogram near time t and mel frequency f. The output to the system is a number y between 0 and 1, denoting the probability of presence of a sound event of interest. In the case of this paper, this sound event is a nocturnal flight call.

Next, we formulate the prediction stage as sound event detection. In this setting, the input to the system is an acoustic scene of arbitrarily long duration. The output of the system is an event detection function $y(t)$, sampled at a rate of 20 frames per second. For every t, we compute $y(t)$ by sliding a window of duration equal to 150 ms and hop size equal to 50 ms over the time-frequency representation $\mathbf{E}(t, f)$ of the acoustic scene. We turn the event detection function $y(t)$ into a list of predicted timestamps by a procedure of thresholding and peak extraction. The total number of predicted timestamps is a computer-generated estimate of the vocal activity of migratory birds near the sensor location at hand. In the realm of avian ecology, this number could potentially be used as a proxy for the density of birds over the course of an entire migration season. Furthermore, the short audio clips corresponding to detected flight flights could be subsequently passed to an automatic species classifier [56] to obtain the distribution of species in the vicinity of each sensor.

We shall describe the procedures of training and evaluating our proposed system in greater detail in the Experimental Design section of this article.

Context-adaptive neural network

Related work

There is a growing body of literature on the topic of filter-generating networks [69], which are deep learning systems of relatively low complexity that generate the synaptic weights in another deep learning system of greater complexity. The association between the filter-generating network, hereafter denoted as auxiliary network, and the high-complexity network, hereafter denoted as main branch, constitutes an acyclic computation graph named dynamic filter network. Like any other deep learning system, a dynamic fiter network is trained by gradient backpropagation, with both the main branch and the auxiliary branch being updated to minimize the same loss function. In the computation graph, the two branches merge into a single output branch. Several mathematical formulations to this merging procedure coexist in the machine learning literature [72–74]. This article compares three of the most straightforward ones, namely adaptive threshold (AT), adaptive weights (AW), and mixture of experts (MoE).

In the application setting of automatic speech recognition, one prominent instance of dynamic filter network is known as context-adaptive neural network (CA-NN) [68]. In a CA-NN for sound event detection, the purpose of the auxiliary branch is to learn a feature representation that would characterize the intrinsic properties of the acoustic environment, while remaining invariant to whether a sound event is present or not in the environment. Therefore, the auxiliary branch does not act upon the audio clip itself; but rather, onto some engineered transformation thereof, hereafter known as a vector of *auxiliary features*.

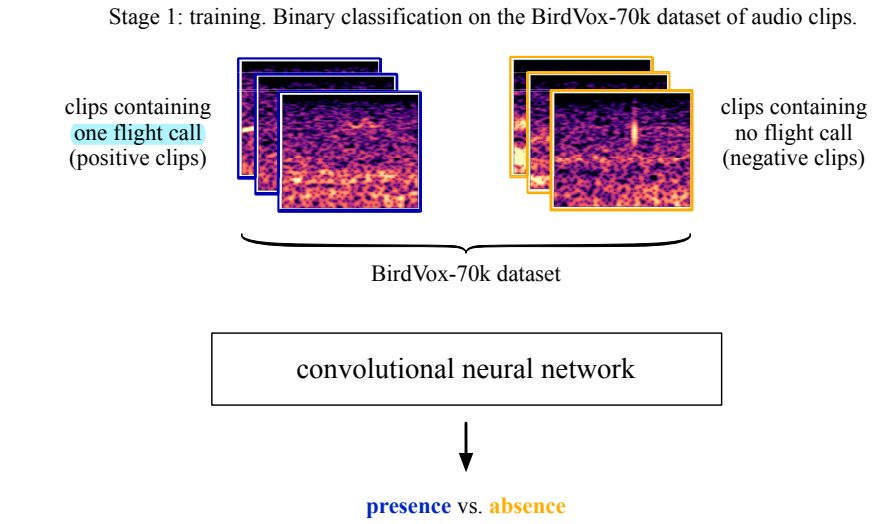

Fig 2. Overview of the presented baseline. After training a deep learning model to identify the presence of a sound event within short audio cips (150 ms), we run this model on a continuous recording by a sliding window procedure. We compare the peaks in the resulting event detection function (EDF) with a fixed threshold τ in order to obtain a list of predicted timestamps for the sound event of interest. In the case of the presented baseline, these sound events of interest are avian flight calls; the input representation is a mel-frequency spectrogram; and the deep learning model is a convolutional network.

Percentile summary statistics as auxiliary features

Original implementations of CA-NN aim at improving robustness of far-field speech recognition systems to reverberation properties of indoor acoustic environments. To this effect, they rely on auxiliary features that characterize spatial diffuseness [67], and are derived from a stereophonic audio input. In contrast, in the application setting of bioacoustic sound event detection, we argue that the leading spurious factor of variability is not reverberation, but rather, background noise. One distinctive property of background noise, as opposed to foreground events, is that it is locally stationary: although bird calls modulate rapidly in amplitude and frequency, a swarm of insects produce a buzzing noise that remains unchanged at the time scale of several minutes. Likewise, a vehicle approaching the sensor will typically grow progressively in acoustic intensity, yet without changing much of its short-term spectrum. We denote by context adaptation (CA) the integration of a sensor-specific, long-term trend into a rapidly changing event detection function, by means of a learned representation of acoustic noise.

It stems from the two observations above that, coarsening the temporal resolution of the time-frequency representation $\mathbf{E}(t,f)$ provides a rough description of the acoustic environment, yet is unaffected by the presence or absence of a short sound event in the short-term vicinity of the time instant t. Hence, we design auxiliary features as nine long-term order statistics (median, quartiles, deciles, percentiles, and permille) summarizing the spectral envelope in $\mathbf{E}(t,f)$ over windows of duration T_{CA}. In the following, we denote by $\mu(t,q,f)$ the three-way tensor of auxiliary features, where the indices q and f correspond to quantile and mel-frequency respectively. After cross-validating the parameter T_{CA} as a geometric progression ranging between one second and two hours, a preliminary experiment revealed that all values above five minutes led to a background estimator of sufficiently low variance to avoid overfitting. We set T_{CA} to 30 minutes in the following, and sample $\mu(t,q,f)$ at a rate of 8 frames per hour.

Computational architecture of a context-adaptive neural network

Figure 3 is a block diagram of our proposed context-adaptive neural network (CA-NN) for avian flight call detection in continuous recordings. The main branch is a convolutional neural network with three convolutional layers followed by two dense layers. The main branch takes the time-frequency representation $\mathbf{E}(t,f)$ of a short audio clip as input, and learns a 64-dimensional representation $\mathbf{z}(t,n)$ as output, where n is an integer between 0 and 63. At prediction time, the value taken by $\mathbf{z}(t,n)$ solely depends on the content of the audio clip, and is not context-adaptive. As regards the auxiliary branch, it is a convolutional neural network with one convolutional layer followed by one dense layer. The auxiliary branch takes a slice of the tensor of quantile summary statistics $\mu(t,q,f)$ as input, and learns some context-adaptive parameters of arbitrary dimension. Because the temporal sampling of μ (4 frames per hour) is coarser than the temporal sampling of y (20 frames per second), the slice in $\mu(t,q,f)$ that is fed to the network consists of a single temporal frame. More precisely, it is a matrix of 9 quantiles q and 128 mel-frequency bins f.

Main branch of the context-adaptive neural network

The main branch has exactly the same architecture as the one that reported state-of-the-art results in urban sound classification [75] (Urban-8K dataset [76]) and species classification from clips of avian flight calls [56] (CLO-43SD dataset [77]). Its first layer consists of 24 convolutional kernels of size 5x5, followed by a rectified linear unit (ReLU) and a strided max-pooling operation whose receptive field has a size of 4x2, that is, 4 logmelspec frames (i.e. 6 ms) and 2 subbands (i.e. about a musical quartertone). Likewise, the second layer consists of 24 convolutional kernels of size 5x5, followed by a ReLU and a strided max-pooling operation whose receptive field has a size of 4x2, that is, 16 logmelspec frames (i.e. 24 ms) and 4 mel-frequency subbands (i.e. about a musical semitone). The third layer consists of 48 convolutional kernels of size 5x5, followed by a ReLU. There is no pooling after the third layer.

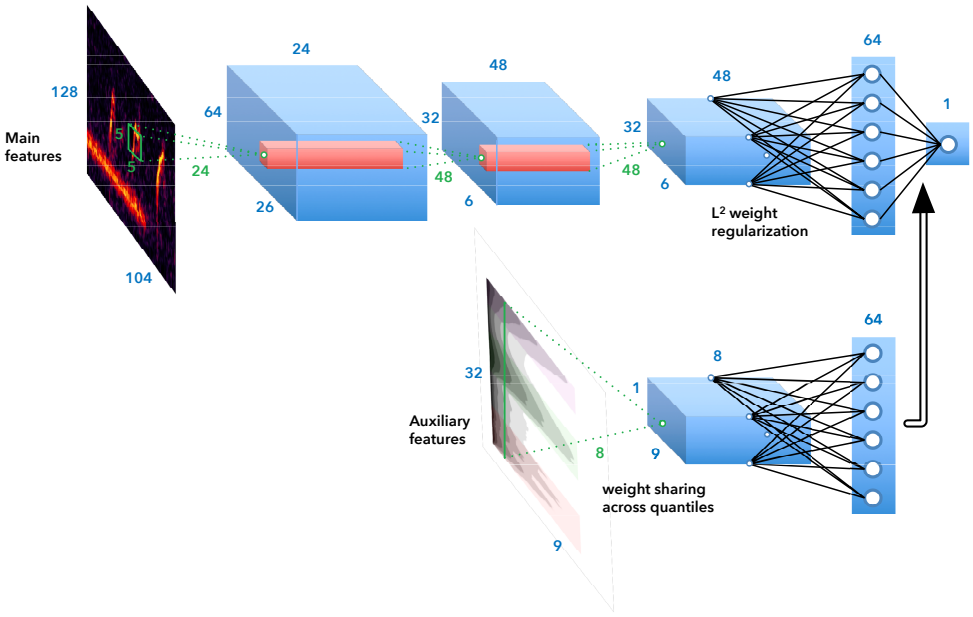

Fig 3. Architecture of our context-adaptive neural network (CA-CNN) with spectral summary statistics as auxiliary features. The double arrow depicts an operation of merging between the main branch and the auxiliary branch.

The fourth layer is a fully connected layer with 64 hidden units, and whose weights are regularized in L^2 norm with a multiplicative factor set to 10^{-3}, followed by a ReLU. The fifth layer is a fully connected layer with a single output unit, followed by a sigmoid nonlinearity. We train the whole deep learning architecture to minimize binary cross-entropy by means of the Adam optimizer [78]. We use the Keras [79] and pescador [80] Python libraries, respectively to build the model and stream trainining data efficiently under a fixed memory budget.

Auxiliary branch of the context-adaptive neural network

In the absence of any context adaptation, the last layer of the convolutional neural network for absence vs. presence classification of a flight call in the short audio clip $\mathbf{E}(t,f)$ is an affine transformation of the vector $\mathbf{z}$ followed by a sigmoid nonlinearity; that is,

$$y(t) = \sigma\left(b + \sum_n \mathbf{w}(n)\mathbf{z}(t,n)\right) \tag{1}$$

where $\mathbf{w}(n)$ is a 64-dimensional vector of synaptic weights and the scalar b is a synaptic bias. Both parameters $\mathbf{w}(n)$ and b are optimized by Adam at training time, yet remain unchanged at prediction time.

The convolutional layer in the auxiliary branch consists of 8 kernels of size 32x1, followed by a ReLU. Observe that, because the height of the kernels is equal to the number of mel-frequency bins in the auxiliary features (i.e. 32) and does not involve any input padding, this convolutional layer performs weight sharing only across quantiles q, and not across neighboring frequency bins f. Each of the learned kernels can be interpreted *post hoc* as a spectral template of background noise, onto which auxiliary features are projected. The dense layer in the auxiliary branch is an affine transformation from the $9 \times 8 = 72$ output activations of the first layer onto 64 nodes, followed by a ReLU. In all three cases, we denote by $\mathbf{z}_{\text{aux}}(t,n)$ the 64-dimensional output of this dense layer. Because it directly proceeds from the auxiliary features $\mu(t,q,f)$ and not from the main features $\mathbf{E}(t,f)$, $\mathbf{z}_{\text{aux}}(t,n)$ has a coarse sampling rate of 8 frames per hour; that is, one context slice every 450 seconds.

In this article, we compare exerimentally three formulations of such a feature map: adaptive weights (AW), adaptive threshold (AT), and mixture of experts (MoE). These formulations correspond to different equations connecting the output $\mathbf{z}(t,n)$ of the main branch with the output $\mathbf{z}_{\text{aux}}(t,n)$ of the auxiliary branch into a predicted probability of presence $y(t)$ at time t, described below.

Adaptive weights

In its adaptive weights formulation (AW), context adaptation replaces $\mathbf{w}(n)$ by $\mathbf{z}_{\text{aux}}(t,n)$ verbatim in Equation 1, resulting in an event detection function of the form

$$y(t) = \sigma\left(b + \sum_n \mathbf{z}_{\text{aux}}(t,n)\mathbf{z}(t,n)\right). \tag{2}$$

Observe that the CNN baseline is a particular case of this formulation, in which the vector $\mathbf{z}_{\text{aux}}(t,n)$ is constant through time. This is made possible by setting the synaptic weights of the dense layer in the auxiliary branch to zero, and keeping only nonnegative biases for each of the 64 nodes. Therefore, a CA-CNN with adaptive weights has an optimal training accuracy that is, in theory, at least as good as that of a conventional CNN with static weights. However, because the loss surface of a deep neural network is nonconvex, an iterative stochastic optimizer such as Adam reaches a local optimum rather than the global optimum in the space of neural network parameters. Consequently, a CA-CNN with adaptive weights may in practice underperform a conventional CNN.

Adaptive threshold

In its adaptive threshold formulation (AT), context adaptation learns a 64-dimensional static vector $\mathbf{w}_{\text{aux}}(n)$, onto which is projected the auxiliary representation $\mathbf{z}_{\text{aux}}(t,n)$ by canonical inner product. This inner product replaces the static scalar bias in Equation 1, resulting in an event detection function of the form

$$y(t) = \sigma\left(\sum_n \mathbf{w}_{\text{aux}}(n)\mathbf{z}_{\text{aux}}(t,n) + \sum_n \mathbf{w}(n)\mathbf{z}(t,n)\right). \tag{3}$$

Again, the CNN baseline is a particular case of the AT formulation. Indeed, setting the vector $\mathbf{z}_{\text{aux}}(t,n)$ to a constant and the weights $\mathbf{w}_{\text{aux}}(n)$ such that the product $\mathbf{w}_{\text{aux}}(n)\mathbf{z}_{\text{aux}}(t,n)$ is equal to $\mathbf{w}(n)$ for every n is equivalent to discarding context adaptation altogether.

Furthermore, this formulation can also be interpreted as the application of a slowly varying threshold onto a static event detection function. This is because, by monotonicity of the inverse sigmoid function σ^{-1}, and given some fixed threshold τ, the inequality $y(t) > \tau$ is equivalent to

$$\sigma\left(\sum_n \mathbf{w}(n)\mathbf{z}(t,n)\right) > \sigma\left(\sigma^{-1}(\tau) - \sum_n \mathbf{w}_{\text{aux}}(n)\mathbf{z}_{\text{aux}}(t,n)\right). \tag{4}$$

The interpretation of the right-hand side as a time-varying threshold is all the more insightful given that $\mathbf{z}_{\text{aux}}(t)$ has much slower variations than $\mathbf{z}(t)$, i.e. 8 frames per hour vs. 20 frames per second. Under this framework, we may draw a connection between context adaptation in neural networks and a long-lasting line of research on engineering adaptive thresholds for sound onset detection [81].

Mixture of experts

Under the adaptive weights formulation, each scalar weight in $\mathbf{w}_{\text{aux}}(n)$ is an independent output of the auxiliary network. In contrast, the mixture of experts formulation (MoE) reduces this requirement by learning a fixed weight vector $\mathbf{w}(n)$ and having a much smaller number of adaptive weights (e.g. $K = 4$) that are applied to subsets of $\mathbf{w}(n)$. Each of these subsets comprises $\frac{N}{K}$ nodes and can be regarded as an "expert". Therefore, the small number K of outputs from the auxiliary branch no longer corresponds to the number of node weights in the main branch, but to the number of mixture weights across expert subsets, hence the name of "mixture of experts" (MoE) formulation.

In practice, the output of the main branch $\mathbf{z}(t,n)$ is reshaped into a tensor $\widetilde{\mathbf{z}}(t,m,k)$, with the integer indices m and k respectively being the quotient and remainder of the Euclidean division of the integer n by the constant K. Likewise, we reshape $\mathbf{w}(t,n)$ into $\widetilde{\mathbf{w}}(t,m,k)$, $\mathbf{z}_{\text{aux}}(t,n)$ into $\widetilde{\mathbf{z}}_{\text{aux}}(t,m,k)$, and $\mathbf{w}_{\text{aux}}(t,n)$ into $\widetilde{\mathbf{w}}_{\text{aux}}(t,m,k)$, where $n = K \times m + k$ for every $0 \leq n < 64$.

The integer k, known as expert index, ranges from 0 to $(K-1)$, and is a hyperparameter of the chose context-adaptive architecture. In accordance with [68], we manually set $K = 4$ in all of the following. On the other hand, the integer m, known as mixture index, ranges from 0 to $M = \frac{N}{K}$, i.e. from 0 to $M = 16$ for $N = 64$ nodes and $K = 4$ experts.

First, the auxiliary branch converts $\widetilde{\mathbf{z}}_{\text{aux}}(t,m,k)$ into a K-dimensional time series $\alpha_{\text{aux}}t,k$, by means of an affine transformation over the mixture index m:

$$\alpha_{\text{aux}}(t,k) = b_{\text{aux}}(k) + \sum_m \widetilde{\mathbf{w}}_{\text{aux}}(t,m,k)\widetilde{\mathbf{z}}_{\text{aux}}(t,m,k). \tag{5}$$

Secondly, a softmax transformation maps $\alpha_{\text{aux}}(t,k)$ onto a discrete probability distribution over the experts k. Each softmax coefficient then serves as a multiplicative gate to the static inner product between $\widetilde{\mathbf{w}}(t,m,k)$ and $\widetilde{\mathbf{z}}(t,m,k)$ over the mixture index m in the main branch. This leads to the following definition for the event detection function $y(t)$:

$$y(t) = \sigma\left(b + \sum_k \frac{e^{\alpha_{\text{aux}}(t,k)}}{\sum_{k'} e^{\alpha_{\text{aux}}(t,k')}} \left(\sum_m \widetilde{\mathbf{w}}(t,m,k)\widetilde{\mathbf{z}}(t,m,k)\right)\right). \tag{6}$$

Like the AW and AT formulations, the MoE formulation is a generalization of the CNN baseline. Indeed, setting the learned representation $\widetilde{\mathbf{w}}_{\text{aux}}(t,m,k)$ to zero and the static vector of auxiliary biases $b_{\text{aux}}(k)$ to an arbitrary constant will cause the probability distribution over experts k to be a flat histogram.

Per-channel energy normalization

Definition

Per-channel energy normalization (PCEN) [70] has recently been proposed as an alternative to the logarithmic transformation of the mel-spectrogram (logmelspec), with the aim of combining dynamic range compression (DRC, also present in logmelspec) and adaptive gain control (AGC) with temporal integration. AGC is a prior stage to DRC involving a low-pass filter ϕ_T of support T, thus yielding

$$\mathbf{PCEN}(t,f) = \left(\frac{\mathbf{E}(t,f)}{(\varepsilon + (\mathbf{E} \overset{t}{*} \phi_T)(t,f))^{\alpha}} + \delta \right)^{r} - \delta^{r} \tag{7}$$

where $\alpha, \varepsilon, \delta$, and r are positive constants. While DRC reduces the variance of foreground loudness, AGC is intended to suppress stationary background noise. The resulting representation has shown to improve performance in far-field ASR [82], keyword spotting [70], and speech-to-text systems [83].

There is practical evidence that, over a large class of real-world recording conditions, PCEN decorrelates and Gaussianizes the background while enhancing the contrast between the foreground and the background [84]. From the standpoint of machine learning theory, this Gaussianization property appears to play a key role in avoiding statistical overfitting. Indeed, deep neural networks are optimally robust to adversarial additive perturbations if the background in the training set is a realization of additive, white, and Gaussian noise (AWGN) [85]. This theoretical argument is all the more crucial to the success of sound event detection systems given that, in the case of bioacoustic sensor networks, background noise is nonuniform, and thus will typically vary in terms of spectral envelope between training set and test set. Therefore, in our study, PCEN serves the double purpose of, first, disentangling foregorund and background as independent sources of variability and, second, facilitating the transferability of learned audio representations between one recording condition and another.

Parameter settings

As Equation 7 shows, the instantiation of PCEN depends upon six parameters: T, α, ε, δ, and r. The effect of these parameters relate to respective properties of the foreground and background noise, as well as the underlying choice of time-frequency representation $\mathbf{E}(t,f)$. Yet, the motivation for developing PCEN initially arose in the context of far-field automatic speech recognition in domestic environments [86]. In contrast, the detection of avian flight calls in rural outdoor areas with autonomous recording units is a starkly different application setting, thus requiring adjustments in the choice of parameters.

One previous publication [84] has conducted an asymptotic analysis of PCEN components, and concluded with some practical recommendation for making such adjustments according to the task at hand. It appears that, in comparison with indoor applications (e.g. ASR in the smart home), bioacoustic event detection distinguishes itself by faster modulations of foreground, higher skewness of background magnitudes, a louder background, and more distant sources. Such idiosyncrasies respectively call for a lower T, a lower α, a higher δ, and a lower r. More precisely, we decode each audio signal as a sequence of floating-point numbers in the range $[-2^{31}; 2^{31}[$ with a sample rate of 22.050 Hz, and apply a short-term Fourier transform (STFT) with window size 256 (12 ms) and hop size 32 (1.5 ms). Then, we map the frequency bins of the STFT squared modulus to a mel scale, with 128 mel-frequency subbands ranging from 2 kHz to

11.025 kHz. Lastly, we apply PCEN according to Equation 7 with $\varepsilon = 10^{-6}$; $\alpha = 0.8$; $\delta = 10$; $r = \frac{1}{4}$; and $T_{\mathrm{PCEN}} = 60\,\mathrm{ms}$ after following the recommendations of [84]. Replacing these ad hoc constants by trainable, frequency-dependent parameters $\alpha(f)$, $\delta(f)$, and so forth, is a promising line of research, but is beyond the scope of this paper, as it does not fundamentally change its overall narrative [70].

Baseline: convolutional neural network

The baseline model of our study is a CNN in the logmelspec domain for avian flight call detection, whose architecture is replicated from a previous study [57]. In spite of its simplicity, this deep learning model has shown to significantly outperform other algorithms for avian flight call detection in the BirdVox-full-night dataset, including spectral flux [81], the Vesper library reimplementation of the "Old Bird" energy-based detection function [58], and the PCA-SKM-SVM shallow learning pipeline [75].

In a preliminary stage, we explored over 100 common variations in the architecture of the baseline, including changes in kernel size, layer width, number of layers, mel scale discretization, multiresolution input [87], choice of nonlinearity, use of dropout, use of batch normalization, and choice of learning rate schedule. Yet, none of these general-purpose variations, unrelated in their design to the question of robutness to background noise, led to systematic improvements upon the baseline. Therefore, although the baseline architecture is by no means optimal, there are grounds to believe that the following improvements brought by CA and PCEN woud not easily be matched by applying other, more well-established variations.

The logmelspec consists of 128 bands between 2 kHz and 11.025 kHz (i.e. the Nyquist frequency), and is extracted over short-term Hann windows of duration 12 ms (256 samples at a sampling rate of 22.050 kHz) and a hop length of 1.5 ms (32 samples). The choice of minimal frequency at 2 kHz corresponds to a lower bound on the vocal range of avian flight calls of thrushes. With the librosa Python library [88], the computation of logmelspec is about 20 times faster than real time on a dual-core Intel Xeon E-2690v2 3.0 GHz central processing unit (CPU).

Artificial data augmentation

Applying randomized digital audio effects to every sample in a dataset at training time often reduces overfitting without any extra computational cost at prediction time [89]. Hence, many deep machine listening systems are trained on augmented data: related applications to this study include bird species classification [56], singing voice detection [90], and urban sound classification [75]. Yet, one difficulty of artificial data augmentation is that the chosen distribution of parameters needs to reflect the underlying variability of the data. In the context of avian flight calls, we use domain-specific knowledge in animal behavior so as to find an appropriate range of parameters for each perturbation.

We distinguish two kinds of data augmentation: geometrical and adaptive. Geometrical data augmentation (GDA) includes all digital audio effects whose parameters are independent of the probability distribution of samples in the training set, such as pitch shifting and time stretching. On the contrary, adaptive data augmentation (ADA) takes into account the whole training data, and in some cases also the corresponding labels, to transform each sample. For instance, mixing each audio clip in the sensor at hand with a negative (noisy) audio clip belonging to a different sensor in the training set leads to greater generalizability [91]. However, this adaptive procedure causes the number of augmented samples to scale quadratically with the number of sensors. Furthermore, it cannot be easily combined with CA because the addition of extraneous noise to the front end would require to also re-compute the corresponding auxiliary features for the mixture of signal and noise at a long temporal scale ($T_{\mathrm{CA}} = 30$ minutes), which is intractable for large T_{CA}. Therefore, we apply geometrical data augmentation to all models, but adaptive data augmentation only to the models that do not include context adaptation.

We use the muda Python package (MUsical Data Augmentation [89]) to apply 20 randomized digital audio effects to each audio clip: four pitch shifts; four time stretchings; and four additions of background noise originating from each of all three training sensors in the cross-validation fold at hand. The choices of probability distributions and hyperparameters underlying these augmentations are identical to those of [57], and are chosen in accordance with expert knowledge about the typical vocal ranges of thrushes, warblers, and sparrows.

All transformations are independent from each other in the probabilistic sense, and never applied in combination. In the case of the addition of background noise, we restrict the set of augmentations to those in which the background noise and the original audio clip belong to recordings that are both in the training set, or both in the validation set.

Experimental design

Stage 1: binary classifier

Dataset: BirdVox-70k

We train all sound event detection models presented in this article as binary classifiers of presence vs. absence of a flight call, at the time scale of audio clips of duration 150 ms. To this end, we rely on the BirdVox-70k dataset, which contains 35k positive clips and 35k negative clips, originating from a network of 6 bioacoustic sensors. We refer to [57] for more details on the curation of the BirdVox-70k dataset.

Evaluation: leave-one-sensor-out cross-validation

Because our study focuses on the comparative generalizability of automated systems for flight call detection, we split the BirdVox-70k dataset according to a stratified, "leave-one-sensor-out" evaluation procedure. After training all systems on the audio recordings originating from three sensors (training set), we use two of the remaining sensors to identify the optimal combination of hyperparameters (validation set), and leave the last sensor out for reporting final results (test set). From one fold to the next, all boundaries between subsets shift by one sensor, in a periodic fashion.

The loss function for training the system is binary cross-entropy $\mathscr{L}(y) = \log|y - y_{\text{true}}|$, where y_{true} is set to 1 if a flight call is present in the audio clip at hand, and 0 otherwise. To evaluate the system in its validation stage, we measure a classification accuracy metric; that is, the proportion of clips in which the absolute difference $|y - y_{\text{true}}|$ is below 0.5 over a hold-out validation set.

Stage 2: detection in continuous audio

Dataset: BirdVox-full-night

We evaluate all sound event detection models presented in this article on a task of species-agnostic avian flight call detection. To this end, we rely on the BirdVox-full-night dataset, which contains recordings originating from one ten-hour night of fall migration, as recorded from 6 different sensors. Each of these sensors is located in rural areas near Ithaca, NY, USA, and is equipped with one omnidirectional microphone of moderate cost. The resulting bioacoustic sensor network covers a total land area of approximately 1000 km^2. The 6 recordings in BirdVox-full-night amount to 62 hours of monaural audio data. The split between training set, validation set, and test set follows the same "leave-one-sensor-out" evaluation procedure as presented in the previous section. Therefore, all models are tested on recording conditions that are extraneous to the training and validation subsets. We refer to [77] for more details on sensor hardware and to [57] for more details on BirdVox-full-night.

Evaluation: precision and recall metrics

We formulate the task of avian flight call detection as follows: given a continuous audio recording from dusk to dawn, the system should produce a list of timestamps, each of them denoting the temporal center of a different flight call. Then, we may evaluate the effectiveness of the system by comparing this list of timestamps against an expert annotation. To this aim, we begin by extracting local peaks in the event detection function according to a fixed threshold τ. The baseline CNN model of [57] constrains consecutive detections to be spaced in time by a minimum lag of at least 150 ms. This constraint improved precision in the baseline CNN model without much detriment to recall, and for consistency we kept this constraint throughout our evaluation. However, as we will see in the Results section, this constraint becomes unnecessary in our state-of-the-art combined model, named BirdVoxDetect. Therefore, BirdVoxDetect may produce predicted timestamps as close to each other as 100 ms (i.e. two discrete hops of duration 50 ms) as it does not induce any constraint on the minimum duration between adjacent peaks in the event detection function.

Once the procedure of thresholding and peak-picking is complete, the detected peaks (flight calls) are evaluated by matching them to the manually labeled calls — the "reference", sometimes called "ground truth" — and computing the number of true positives (TP), false positives (FP) and false negatives (FN). This process is repeated for varying peak detection threshold values τ between 0 and 1 to obtain the standard information retrieval metric of Area Under the Precision Recall Curve (AUPRC) with 0 being the worst value and 1 being the best. A detected peak and a reference peak are considered to be a matching pair if they are within 500 ms of each other. To ensure optimum matching of detected peaks to reference peaks while ensuring each reference peak can only be matched to a single estimated peak, we treat the problem as a maximum bipartite graph matching problem [92] and use the implementation provided in the `mir_eval` Python library for efficiency and transparency [93].

Results

Stage 1: training of a binary classifier

Exhaustive benchmark on validation set

Figure 4 summarizes the validation error rates on BirdVox-70k of twelve different models. These models represent different combinations between three design choices: choice of time-frequency representation, choice of formulation in the context adaptation, and use of artificial data augmentation. In order to mitigate the influence of random initialization on these validation error rates, we train and evaluate each of the twelve different models ten different times on each of the six folds, and report the median validation error rate only. The cumulative computational budget for training all models is of the order of 180 GPU-days for training, and 180 CPU-days for prediction. In both cases, we parallelize massively across models, folds, and trials, resulting in 720 different jobs in total, each running independently for approximately six hours on a high-performance computing cluster.

We find that, across all models, some folds consistently lead to a greater error rate than others. In the case of the logmelspec-CNN baseline, the typical error rate is of the order of 5%, but varies between 2% and 20% between folds. Individual variations of that baseline are not equally beneficial. First, replacing the logmelspec acoustic frontend by PCEN improves validation accuracy on five folds out of six, and GDA improves it on four folds out of six. Secondly, adding context adaptation to the baseline, by means of a mixture of experts (MoE), is detrimental to validation accuracy in four folds, while using an adaptive threshold (AT) instead of MoE essentially leaves the baseline unchanged, as it improves and degrades per-fold performance in comparable measures. Therefore, it appears that context adaptation alone fails to improve the generalizability of a logmelspec-based deep learning model for avian flight call

detection. In what follows, we focus on analyzing the effects of context adaptation on models that are either trained with PCEN, GDA, or both.

Figure 4 also shows that applying GDA to a PCEN-based model consistently improves validation accuracy over all six folds, whether an AT is present or not. We hypothesize that this consistent improvement is the relational effect of artificial pitch shifts in GDA and background noise reduction caused by PCEN. Indeed, one shortcoming of pitch shifting in GDA is that it affects foreground and background simultaneously. Yet, natural factors of variability in avian flight call detection, such as those arising due to animal behavior, will typically affect the absolute fundamental frequency of the foreground while leaving the background — i.e. the spectral envelope of insects or passing cars — unchanged. Consequently, artificial pitch shifts, even as small as a musical semitone, may lead to a plausible foreground, yet mixed with an implausible background in logmelspec domain. On the contrary, as described in the Methods section, PCEN tends to bring the distribution of background time-frequency magnitudes closer to additive white Gaussian noise (AWGN). Because AWGN has a flat spectrum, transposing a polyphonic mixture containing a nonstationary foreground and an AWGN background has the same effect as transposing the foreground only while leaving the background unchanged. Therefore, not only does replacing the logmelspec acoustic frontend by PCEN improve the robustness of a classifier to background noise, it also helps disentangling pitch transpositions of background and foreground, thus allowing for more extensive geometrical data augmentation by pitch shifts and time stretchings.

Because the six folds in the leave-one-sensor-out cross-validation procedure are of unequal size and acoustic diversity, it is not straightforward to rank all twelve models according to a single global evaluation metric. However, we may induce a structure of partial ordering between models by the following definition: a model A is regarded as superior to model B if and only if switching from A to B degrades accuracy on half of the folds or more. According to this definition, the last model (GDA-PCEN-AT) is the only one that is superior to all others. Moreover, we find that PCEN is superior to the logmelspec baseline; that GDA-PCEN is superior to PCEN; and that GDA-PCEN-AT is superior to PCEN. We also find that GDA-logmelspec is superior to logmelspec, and that GDA-PCEN is superior to GDA-logmelspec. However, we do not find either logmelspec-AT or logmelspec-MoE to be superior to logmelspec. In addition, GDA-PCEN-MoE is superior to GDA-PCEN, yet inferior to GDA-PCEN-AT.

Because GDA-PCEN-AT and GDA-PCEN-MoE perform almost equally across the board, one supplementary question that arises from this benchmark is whether AT and MoE could somehow be combined into a hybrid form of context adaptation. To challenge this hypothesis, we trained a thirteenth model, named GDA-PCEN-AT-MoE, ten times on each fold of BirdVox-70k, and measure median validation accuracies. We found that this model performs below GDA-PCEN-AT on the majority of folds, and failed to train at all on many trials. Threfore, we do not pursue this line of research further. Rather, we adopt the adaptive threshold (AT) formulation as a simple, yet effective, method. We postulate that the overall degradation in accuracy from GDA-PCEN-AT to GDA-PCEN-AT-MoE is caused by an excessive number of degrees of freedom in the design of the context-adaptive neural network.

Two conclusions arise from all the observations above. First, the best performing model, in terms of validation accuracy on BirdVox-70k, appears to be GDA-PCEN-AT. Therefore, in the following, GDA-PCEN-AT is the model that we will choose to report results on the test set. Second, because context adaptation does not improve the baseline, but only models that feature PCEN, we deduce that an ablation study from GDA-PCEN-AT should begin by removing AT before removing PCEN. Therefore, in the following, we discuss and compare the evolution of test set recall through time for GDA-PCEN-AT and GDA-PCEN, but do not report test set results on GDA-logmelspec-AT because this model is excluded by cross-validation.

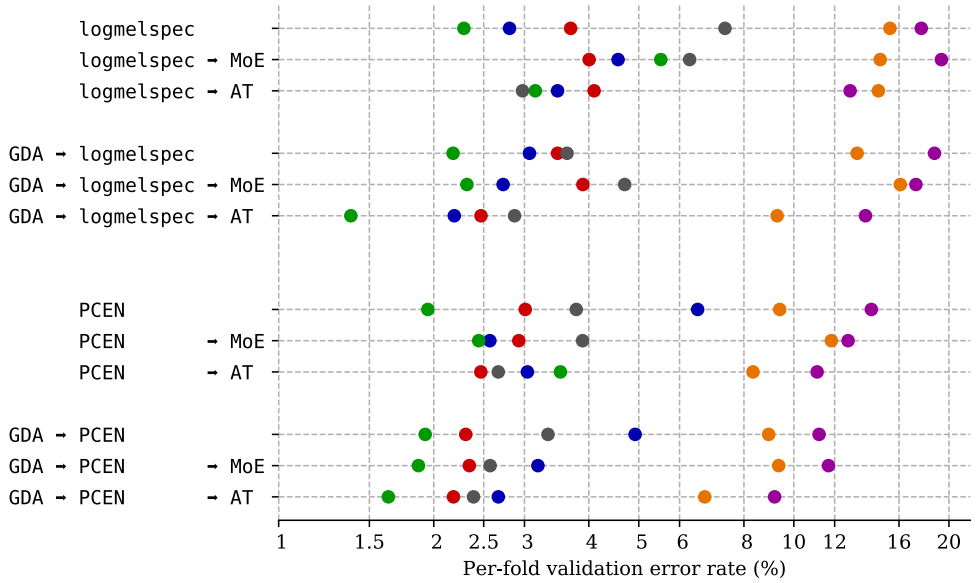

Fig 4. Exhaustive benchmark of architectural variations with respect to the logmelspec-CNN baseline, on a task of binary classification of presence vs. absence of a bird in audio clips. Dot colors represent folds in BirdVox-70k. GDA: geometrical data augmentation. logmelspec: log-mel-spectrogram. PCEN: per-channel energy normalization. MoE: mixture of experts. AT: adaptive threshold.

Ablation study

Once the exhaustive benchmark has identified one reference model — namely, GDA-PCEN-AT — we may measure the relative difference in error rate between the reference and some other model in the benchmark for each fold, and compute quantiles across folds. The reason why we opt for averaging relative differences rather than absolute differences is that the former, unlike the latter, tends to follow a symmetric distribution across folds, and thus can be represented on a box-and-whisker plot. Then, we may compare and rank the respective positions of the boxes for different ablations of the reference.

Figure 5 summarizes the results of our ablation study. Replacing AT by MoE hardly affects our results. Therefore, it is more likely the presence of any form of context adaptation at all, rather than specific architectural choices in the side-channel neural network, that enables a greater generalization across folds. However, removing geometrical data augmentation, and training the PCEN-AT model on original audio clips from BirdVox-70k only, does hinder acccuracy consistently, though less so than other improvements upon the baseline (i.e. PCEN and context adaptation). This supports our hypothesis that shortcomings of the baseline are mainly attributable to its lack of robustness to background noise, more so than its lack of robustness to the geometrical variability in time-frequency patterns of avian flight calls.

We also found that ADA-PCEN-AT and GDA-PCEN bring comparable differences in miss rate with respect to the reference model GDA-PCEN-AT. In other words, the addition of noise to the main branch of the network without reflecting it in the auxiliary features is, quite unsurprisingly, about as detrimental as not having auxiliary features at all.

Finally, replacing PCEN by logmelspec in the reference model increases miss rates by about 60% on average, and over 100% in two out of the six folds. Thus, there are grounds to believe that PCEN is the predominant contributor to validation accuracy in the GDA-PCEN-AT reference model.

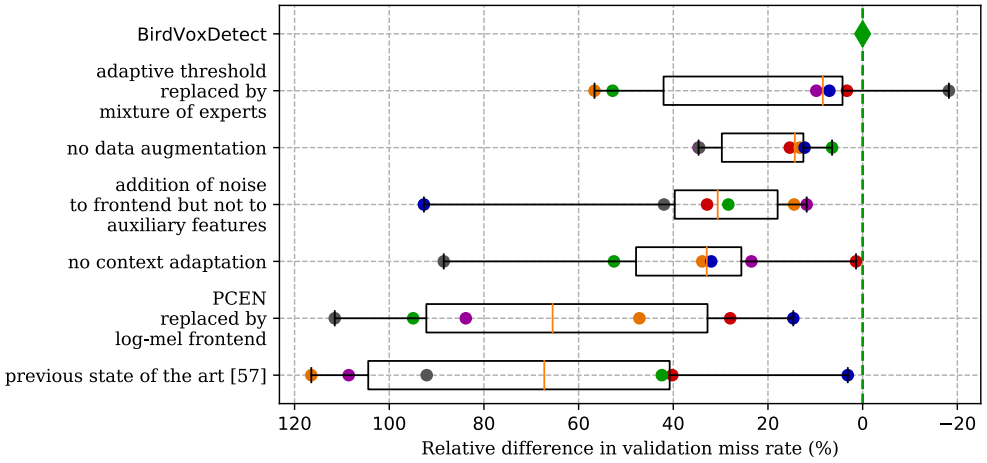

Fig 5. Ablation study of best model (CNN+PCEN+CA+GDA) on the BirdVox-full-night dataset. Boxes (resp. whiskers) denote interquartile (resp. extremal) variations between sensors.

Stage 2: detection in continuous audio

Precision-recall curves

Although the BirdVox-70k dataset is particularly well suited for training machine listening systems for avian flight call detection, it does not reflect the practical use case of flight call monitoring in continuous recordings. Indeed, as described in [57], BirdVox-70k is curated in a semi-automatic fashion: while the positive clips proceed from human annotations, the negative clips correspond to the false alarms of an off-the-shelf shallow learning model. Specifically, BirdVox-70k contains a larger proportion of challenging confounding factors — such as siren horns and electronic beeps — and, conversely, a smaller proportion of quasi-silent sound clips, than a full night of bird migration. Therefore, whereas the previous subsection used validation accuracy on BirdVox-70k as a proxy for singling out an optimal model, it is, from the perspective of applied bioacoustics, less insightful to report test set accuracy on BirdVox-70k than it is to plot a precision-recall curve on BirdVox-full-night.

Figure 6 illustrates the combined effects of PCEN, CA, and GDA on the area under the precision-recall curve (AUPRC) when applying CNN to flight call event detection on the BirdVox-full-night test set recordings. In agreement with the ablation study, the best model (CNN+PCEN+CA+GDA) reaches a test AUPRC of 72.0%, thus outperforming models lacking either PCEN, CA, or GDA. In addition to the precision-recall curves that are shown in Figure 6, we computed predictions over BirdVox-full-night for each of the twelve models presented in the exhaustive benchmark (Figure 4), over 6 folds and 10 randomized trials. This last procedure represents about 10 CPU-years of computation in total. From it, we can confirm that BirdVoxDetect does not overfit the validation set more than any of its counterparts.

Error analysis

We opened this article by pointing out that many state-of-the-art systems for bioacoustic event detection lack robustness to spatiotemporal variations in background noise, thus preventing their reliability at the scale of distributed sensor networks. In particular, we had shown in Figure 1 that the CNN baseline of [57] exhibits a poor recall (below 50%) in the early hours of BirdVox-full-night, while only achieving a satisfying recall towards the end of each full night continuous recording. We hypothesized that such drastic variation in performance was attributable to the scarcity of training examples at dusk in comparison to dawn, in conjunction with more intense levels of background noise at dusk than at dawn. Now, Figure 7 offers

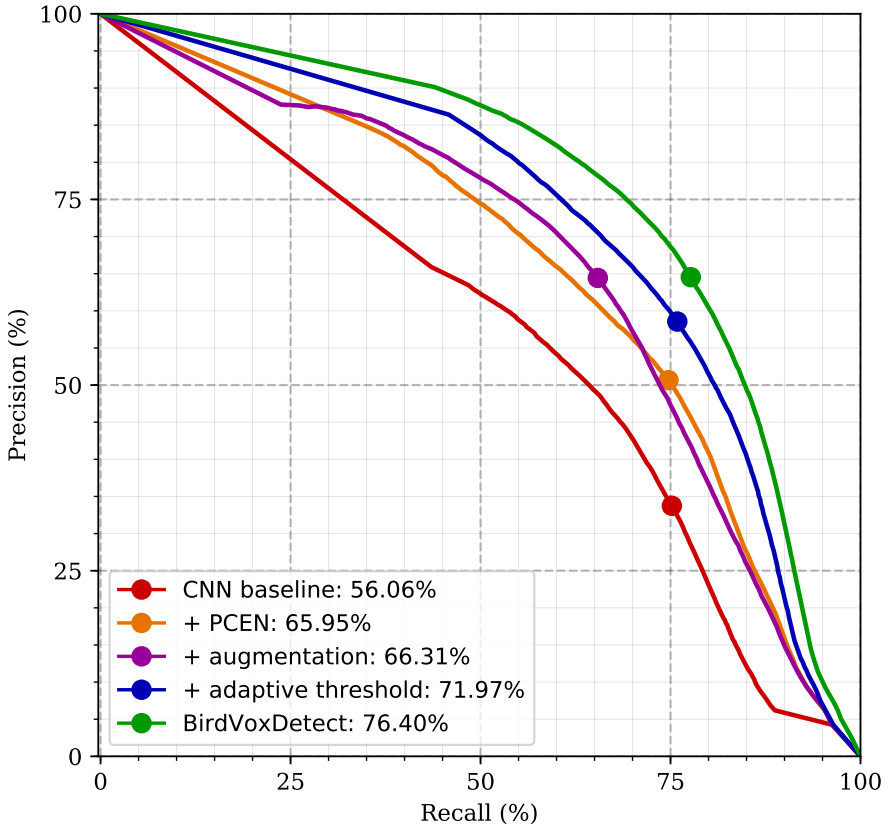

Fig 6. Precision-recall detection curves in avian flight call detection (BirdVox-full-night dataset). The area under each precision-recall curve (AUPRC) is shown in the legend of the plot. The red line (76.40%) is the CNN baseline (previous state of the art of [57] without context adaptation). The blue line (76.40%) is our best performing model on the validation set, and is released under the name of BirdVoxDetect. The thick dot on each curve denotes the optimal tradeoff between precision and recall, corresponding to a maximal F_1-score. CNN: convolutional neural network; PCEN: per-channel energy normalization.

evidence to support this initial hypothesis. While the top subfigure shows the evolution of recall of the CNN baseline on the BirdVox-full-night dataset, the other two subfigures in Figure 7 show the evolution of recall from two models presented in this paper, both of which are designed to be more robust to noise than the baseline.

First, Figure 7 (middle) shows that the PCEN model, comprising per-channel energy normalization, is not only useful at dawn, but also earlier in the night: at certain sensor locations, the recall rate is above 70% from 10 p.m. onwards, as opposed to 2 a.m. for the CNN baseline. This qualitative finding confirms that replacing the logarithmic compression of the mel-frequency spectrogram (logmelspec) by per-channel energy normalization (PCEN) may turn out to be greatly beneficial to the practical usefulness of deep machine listening models for sound event detection. Indeed, not only does PCEN significantly improve the tradeoff between precision and recall over the global test set (as was demonstrated in Figure 6), but it also considerably reduces the probability of missed detection within time slots in which sound events are very rare, such as dusk in the case of avian flight calls. It is striking to note that our end-to-end learning system, once endowed with a PCEN acoustic frontend, manages to perform almost as well on these time slots, despite the fact that, having fewer events, they contribute marginally to the global precision-recall curve of Figure 6.

However, it should be noted that the increase in robustness to nonstationary noise that is afforded by the introduction of PCEN is not accompanied by an increase in robustness to nonuniform noise, as one could have hoped. Rather, as illustrated by the shaded areas surrounding the line plots, and which denote interquartile variations across sensors, the GDA-PCEN model suffers from large variations in recall between sensor locations at any given time of the full night recording. For example, near 2 a.m., the median recall for warblers and sparrows (frequency subband above 5 kHz) is about 60%, but as low as 30% for one of the sensors. From the standpoint of the practitioner in the life sciences, the fact that such variations are both large and difficult to anticipate indicates that the use of PCEN alone is insufficient to offer any guarantees of reliability in the realm of automated bioacoustic event detection.

Secondly, Figure 7 (bottom) shows the evolution of recall of the GDA-PCEN-AT model, also known as BirdVoxDetect, over the course of the BirdVox-full-night dataset. It appears that this model, which combines a PCEN-based convolutional neural network and an auxiliary branch learning an adaptive threshold (AT), exhibits narrower interquartile differences between sensor locations than any of its counterparts. In other words, even though context adaptation leaves the amount of robustness to nonstationarity in background noise essentially unchanged, it noticeably improves the robustness to nonuniformity in background noise of the sound event detection system at hand. This observation suggests that the deployment of distributed machine listening software for flight call monitoring in a bioacoustic sensor network of autonomous recording units requires the resort to deep, data-driven methods for context adaptation, in addition to a shallow procedure of adaptive gain control in the time-frequency domain.

Conclusion

Spatial and temporal variability in background noise and the inability to generalize automatic detectors in such conditions are major obstacles to the large-scale deployment of bioacoustic sensor networks. In this article, we have developed, benchmarked, and combined several machine listening techniques to improve the generalizability of SED models across heterogeneous acoustic environments.

Our main finding is that, although both per-channel energy normalization (PCEN) and context adaptation (CA) improve the generalizability of deep learning models for sound event detection, these two methods are not interchangeable, but instead complementary: whereas PCEN is best suited for mitigating the temporal variations of background noise in a single sensor, CA is best suited for mitigating spatial variations in background noise across sensor locations, whether the acoustic environment surrounding each sensor varies through time or not.

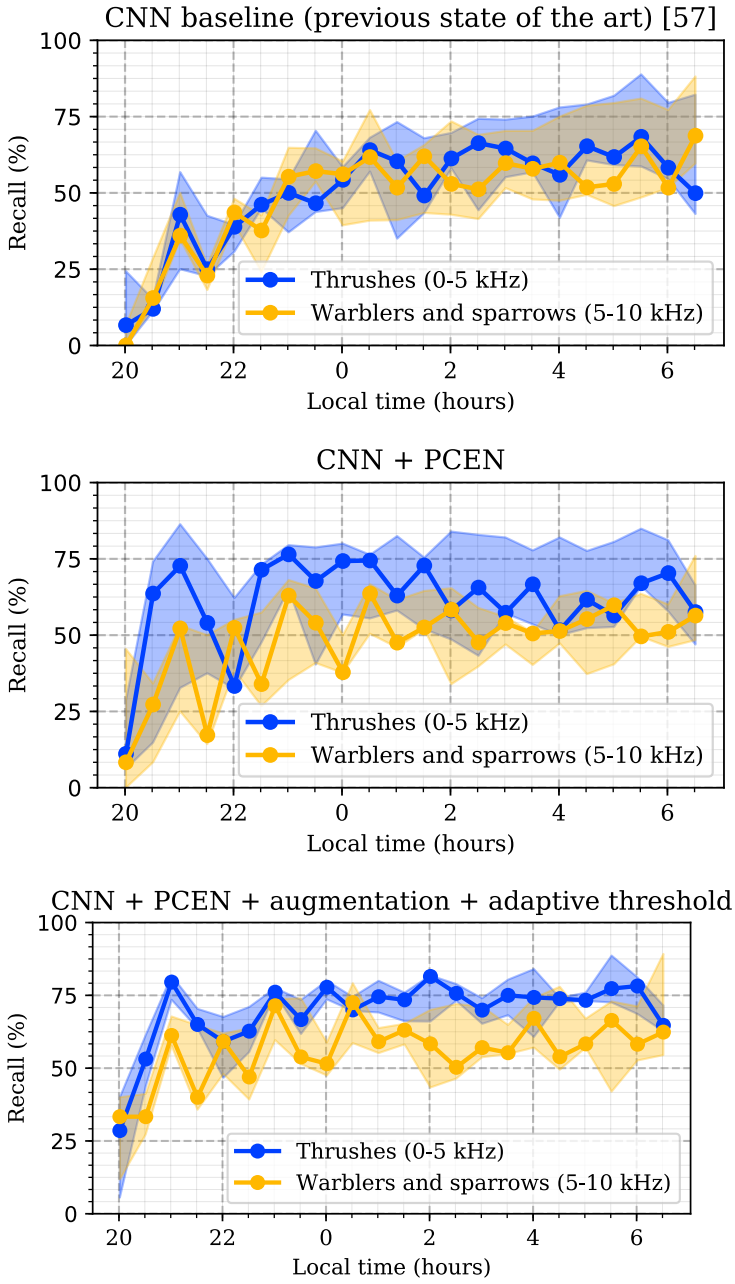

Fig 7. Evolution of recall in the automatic detection of avian flight calls over 30-minute segments in BirdVox-full-night, for two taxa of migratory birds: thrushes (blue curve, 0-5 kHz frequency range) and warblers and sparrows (orange curve, 5-10 kHz frequency range). CNN: convolutional neural network. PCEN: per-channel energy normalization. Shaded areas denote interquartile variations across sensors. We find that PCEN improves robustness to noise nonstationarity, while context adaptation improves robustness to noise nonuniformity.

Indeed, PCEN relies on the assumption that background noise is stationary at a short time scale ($T_{\text{PCEN}} = 60\,\text{ms}$), of the order of the duration of the acoustic events of interest; whereas CA computes auxiliary features at a longer temporal scale ($T_{\text{CA}} = 30$ m). Consequently, PCEN compensates intermittent changes in the loudness of background sources, such as a passing vehicle or the stridulation of an insect; however, it assumes statistical independence between background and foreground, and is thus inadequate to model how different habitats might trigger different vocalization behaviors in the species of interest. For its part, the CA-CNN draws on the variety of sensors in the training set to learn a joint model of both background and foreground; however, this joint model needs to be regularized by integrating long-term context into auxiliary features of relatively low dimensionality, which are, by design, invariant to rapid changes in environmental noise.

After a comprehensive benchmark of architectural variations between convolutional neural networks, we obtain statistically significant evidence to suggest that a combination of PCEN, adaptive threshold, and artificial data augmentation (pitch shifts and time stretchings) provides a consistent and interpretable improvement over the logmelspec-CNN baseline. Reductions in miss rates with respect to the state of the art range between 10% and 50% depending on the location of the sensor, and bring the area under the precision-recall curve (AUPRC) of the BirdVox-full-night benchmark [57] from 61% to 76%. In addition, the recall of our selected model for sound event detection, named BirdVoxDetect, remains relatively high even in recording conditions where less training data is available, e.g. at dusk or in sensors with peculiar characteristics in background noise.

Alongside this article, we release BirdVoxDetect as a pre-trained model on BirdVox-full-night[2]. We encourage bioacoustics researchers to download it and run it on their own recordings of nocturnal flight calls in the wild, especially if these recordings also contain high levels of background noise and/or spurious sound events. BirdVoxDetect can detect the nocturnal flight calls of warblers, thrushes, and sparrows, with a high level of generality in terms of target species as well as sensor locations. Indeed, as demonstrated by our benchmark, the procedures of PCEN and unsupervised context adaptation allow BirdVoxDetect to be deployed in a broad variety of recording conditions, exceeding those that are present in the BirdVox-full-night dataset.

Deriving computer-generated estimates of migratory activity at ranges of spatiotemporal scales from a decentralized network of low-cost bioacoustic sensors is a promising avenue for new insights in avian ecology and the conservation of biodiversity. Future work will apply the BirdVoxDetect machine listening system to large-scale bioacoustic migration monitoring.

Acknowledgment

The authors wish to thank Marc Delcroix, Holger Klinck, Peter Li, Richard F. Lyon, and Brian McFee for fruitful discussions. This work is partially supported by NSF awards 1633259 and 1633206, the Leon Levy Foundation, and a Google faculty award.

Author contributions

Conceptualization VL, JS, AF, SK, JPB

Data curation AF, SK

Formal analysis VL

Funding acquisition AF, SK, JPB

[2]Link to download BirdVoxDetect: https://github.com/BirdVox/birdvoxdetect

Investigation VL, JS

Methodology VL, JS, JPB

Project administration AF, SK, JPB

Resources VL, JS

Software VL, JS

Supervision SK, JPB

Visualization VL, JS

Writing — original draft VL, JS, AF

References

1. Segura-Garcia J, Felici-Castell S, Perez-Solano JJ, Cobos M, Navarro JM. Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks. Sensors Journal. 2015;15(2):836–844.
2. Mack C. The multiple lives of Moore's law. IEEE Spectrum. 2015;52(4):31–31.
3. Hecht J. Is Keck's law coming to an end? IEEE Spectrum. 2016; p. 11–23.
4. McCallum JC. Graph of Memory Prices Decreasing with Time; 2017. http://jcmit.net/memoryprice.htm.
5. Stowell D, Giannoulis D, Benetos E, Lagrange M, Plumbley MD. Detection and classification of acoustic scenes and events. Transactions on Multimedia. 2015;17(10):1733–1746.
6. Laiolo P. The emerging significance of bioacoustics in animal species conservation. Biological Conservation. 2010;143(7):1635–1645.
7. Bello JP, Mydlarz C, Salamon J. Sound Analysis in Smart Cities. In: Virtanen T, Plumbley MD, Ellis D, editors. Computational Analysis of Sound Scenes and Events. Springer; 2018. p. 373–397.
8. Zhao Z, D'Asaro EA, Nystuen JA. The sound of tropical cyclones. Journal of Physical Oceanography. 2014;44(10):2763–2778.
9. Merchant ND, Fristrup KM, Johnson MP, Tyack PL, Witt MJ, Blondel P, et al. Measuring acoustic habitats. Methods in Ecology and Evolution. 2015;6(3):257–265.
10. Nieukirk SL, Mellinger DK, Moore SE, Klinck K, Dziak RP, Goslin J. Sounds from airguns and fin whales recorded in the mid-Atlantic Ocean, 1999–2009. Journal of the Acoustical Society of America. 2012;131(2):1102–1112.
11. Blumstein DT, Mennill DJ, Clemins P, Girod L, Yao K, Patricelli G, et al. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. Journal of Applied Ecology. 2011;48(3):758–767.
12. Marques TA, Thomas L, Martin SW, Mellinger DK, Ward JA, Moretti DJ, et al. Estimating animal population density using passive acoustics. Biological Reviews. 2013;88(2):287–309.

13. Shonfield J, Bayne E. Autonomous recording units in avian ecological research: current use and future applications. Avian Conservation and Ecology. 2017;12(1):42–54.

14. Heinicke S, Kalan AK, Wagner OJ, Mundry R, Lukashevich H, Kühl HS. Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods in Ecology and Evolution. 2015;6(7):753–763.

15. Baumgartner MF, Fratantoni DM, Hurst TP, Brown MW, Cole TVN, Van Parijs SM, et al. Real-time reporting of baleen whale passive acoustic detections from ocean gliders. Journal of the Acoustical Society of America. 2013;134(3):1814–1823.

16. Stewart FEC, Fisher JT, Burton AC, Volpe JP. Species occurrence data reflect the magnitude of animal movements better than the proximity of animal space use. Ecosphere. 2018;9(2):e02112.

17. Oliver RY, Ellis DP, Chmura HE, Krause JS, Pérez JH, Sweet SK, et al. Eavesdropping on the Arctic: Automated bioacoustics reveal dynamics in songbird breeding phenology. Science Advances. 2018;4(6):eaaq1084.

18. Fiedler W. New technologies for monitoring bird migration and behaviour. Ringing and Migration. 2009;24(3):175–179.

19. Gordo O. Why are bird migration dates shifting? A review of weather and climate effects on avian migratory phenology. Climate Research. 2007;35(1-2):37–58.

20. Bairlein F. Migratory birds under threat. Science. 2016;354(6312):547–548.

21. Loss SR, Will T, Marra PP. Direct mortality of birds from anthropogenic causes. Annual Review of Ecology, Evolution, and Systematics. 2015;46:99–120.

22. Dokter AM, Farnsworth A, Fink D, Ruiz-Gutierrez V, Hochachka WM, La Sorte FA, et al. Seasonal abundance and survival of North America's migratory avifauna determined by weather radar. Nature ecology & evolution. 2018;2(10):1603–1609.

23. Farnsworth A, Sheldon D, Geevarghese J, Irvine J, Van Doren B, Webb K, et al. Reconstructing velocities of migrating birds from weather radar — a case study in computational sustainability. AI Magazine. 2014;35(2):31–48.

24. Van Doren BM, Horton KG. A continental system for forecasting bird migration. Science. 2018;361(6407):115–1118.

25. DeVault TL, Belant JL, Blackwell BF, Seamans TW. Interspecific variation in wildlife hazards to aircraft: implications for airport wildlife management. Wildlife Society Bulletin. 2011;35(4):394–402.

26. Drewitt AL, Langston RH. Assessing the impacts of wind farms on birds. Ibis. 2006;148:29–42.

27. Blair RB. Land use and avian species diversity along an urban gradient. Ecological Applications. 1996;6(2):506–519.

28. Van Doren BM, Horton KG, Dokter AM, Klinck H, Elbin SB, Farnsworth A. High-intensity urban light installation dramatically alters nocturnal bird migration. Proceedings of the National Academy of Sciences. 2017;114(42):11175–11180.

29. Bauer S, Chapman JW, Reynolds DR, Alves JA, Dokter AM, Menz MM, et al. From agricultural benefits to aviation safety: realizing the potential of continent-wide radar networks. BioScience. 2017;67(10):912–918.

30. Farnsworth A, Van Doren BM, Hochachka WM, Sheldon D, Winner K, Irvine J, et al. A characterization of autumn nocturnal migration detected by weather surveillance radars in the northeastern USA. Ecological Applications. 2016;26(3):752–770.

31. Sullivan BL, Aycrigg JL, Barry JH, Bonney RE, Bruns N, Cooper CB, et al. The eBird enterprise: an integrated approach to development and application of citizen science. Biological Conservation. 2014;169:31–40.

32. Farnsworth A. Flight calls and their value for future ornithological studies and conservation research. The Auk. 2005;122(3):733–746.

33. Fink D, Hochachka WM, Zuckerberg B, Winkler DW, Shaby B, Munson MA, et al. Spatiotemporal exploratory models for broad-scale survey data. Ecological Applications. 2010;20(8):2131–2147.

34. Fink D, Damoulas T, Bruns NE, La Sorte FA, Hochachka WM, Gomes CP, et al. Crowdsourcing meets ecology: hemisphere-wide spatiotemporal species distribution models. AI magazine. 2014;35(2):19–30.

35. Pamuła H, Kłaczyński M, Remisiewicz M, Wszołek W, Stowell D. Adaptation of deep learning methods to nocturnal bird audio monitoring. In: Postępy akustyki. Polskie Towarzystwo Akustyczne, Oddziałl Górnośląski; 2017. p. 149–158.

36. Stowell D. Computational bioacoustic scene analysis. In: Virtanen T, Plumbley MD, Ellis D, editors. Computational Analysis of Sound Scenes and Events. Springer; 2018. p. 303–333.

37. Ross SRPJ, Friedman NR, Dudley KL, Yoshimura M, Yoshida T, Economo EP. Listening to ecosystems: data-rich acoustic monitoring through landscape-scale sensor networks. Ecological Research. 2018;33(1):135–147.

38. Shamoun-Baranes J, Farnsworth A, Aelterman B, Alves JA, Azijn K, Bernstein G, et al. Innovative visualizations shed light on avian nocturnal migration. PLOS One. 2016;11(8):e0160106.

39. Warren PS, Katti M, Ermann M, Brazel A. Urban bioacoustics: it's not just noise. Animal Behavior. 2006;71(3):491–502.

40. Lanzone M, Deleon E, Grove L, Farnsworth A. Revealing undocumented or poorly known flight calls of warblers (Parulidae) using a novel method of recording birds in captivity. The Auk. 2009;126(3):511–519.

41. Hobson KA, Rempel RS, Greenwood H, Turnbull B, Wilgenburg SLV. Acoustic surveys of birds using electronic recordings: new potential from an omnidirectional microphone system. Wildlife Society Bulletin. 2002;30(3):709–720.

42. Pijanowski BC, Villanueva-Rivera LJ, Dumyahn SL, Farina A, Krause BL, Napoletano BM, et al. Soundscape ecology: the science of sound in the landscape. BioScience. 2011;61(3):203–216.

43. Naguib M. Reverberation of rapid and slow trills: implications for signal adaptations to long-range communication. The Journal of the Acoustical Society of America. 2003;113(3):1749–1756.

44. Krim H, Viberg M. Two decades of array signal processing research: the parametric approach. IEEE Signal Processing Magazine. 1996;13(4):67–94.

45. Wilson S, Bayne E. Use of an acoustic location system to understand how presence of conspecifics and canopy cover influence Ovenbird (Seiurus aurocapilla) space use near reclaimed wellsites in the boreal forest of Alberta. Avian Conservation and Ecology. 2018;13(2).

46. Mydlarz C, Salamon J, Bello JP. The implementation of low-cost urban acoustic monitoring devices. Applied Acoustics. 2017;117:207–218.

47. Knight EC, Bayne EM. Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. Bioacoustics. 2018;To appear.

48. Evans WR. Monitoring avian night flight calls — The new century ahead. The Passenger Pigeon. 2005;67:15–27.

49. Kaewtip K, Alwan A, O'Reilly C, Taylor CE. A robust automatic birdsong phrase classification: a template-based approach. The Journal of the Acoustical Society of America. 2016;140(5):3691–3701.

50. Heittola T, Çakir E, Virtanen T. The machine learning approach for analysis of sound scenes and events. In: Virtanen T, Plumbley MD, Ellis D, editors. Computational Analysis of Sound Scenes and Events. Springer; 2018. p. 13–40.

51. Joly A, Goëau H, Glotin H, Spampinato C, Bonnet P, Vellinga WP, et al. LifeCLEF 2017 Lab Overview: Multimedia Species Identification Challenges. In: Jones GJF, Lawless S, Gonzalo J, Kelly L, Goeuriot L, Mandl T, et al., editors. Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer International Publishing; 2017. p. 255–274.

52. Ulloa JS, Aubin T, Llusia D, Bouveyron C, Sueur J. Estimating animal acoustic diversity in tropical environments using unsupervised multiresolution analysis. Ecological Indicators. 2018;90:346–355.

53. Brumm H, Zollinger SA, Niemelä PT, Sprau P. Measurement artefacts lead to false positives in the study of birdsong in noise. Methods in Ecology and Evolution. 2017;8(11):1617–1625.

54. Marcarini M, Williamson GA, de Sisternes Garcia L. Comparison of methods for automated recognition of avian nocturnal flight calls. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2008. p. 2029—-2032.

55. Efford MG, Dawson DK, Borchers DL. Population density estimated from locations of individuals on a passive detector array. Ecology. 2009;90(10):2676–2682.

56. Salamon J, Bello JP, Farnsworth A, Kelling S. Fusing shallow and deep learning for bioacoustic bird species classification. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2017. p. 141–145.

57. Lostanlen V, Salamon J, Farnsworth A, Kelling S, Bello JP. BirdVox-full-night: a dataset and benchmark for avian flight call detection. In: Proceedings of the Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2017. p. 266–270.

58. Mills H. HaroldMills/Vesper-Old-Bird-Detector-Eval: v1.0.2; 2018. Available from: https://doi.org/10.5281/zenodo.1306879.

59. Klapuri A. Sound onset detection by applying psychoacoustic knowledge. In: Procedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 6. IEEE; 1999. p. 3089–3092.

60. Stowell D, Wood M, Stylianou Y, Glotin H. Bird detection in audio: a survey and a challenge. In: Proceedings of the International Conference on Machine Learning for Signal Processing (MLSP). IEEE; 2016. p. 1–7.

61. Stowell D, Wood MD, Pamuła H, Stylianou Y, Glotin H. Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge. Methods in Ecology and Evolution. 2018;.

62. Grill T, Schlüter J. Two convolutional neural networks for bird detection in audio signals. In: Procedings of the European Signal Processing Conference. IEEE; 2017. p. 1764–1768.

63. Cakir E, Adavanne S, Parascandolo G, Drossos K, Virtanen T. Convolutional recurrent neural networks for bird audio detection. In: Proceedings of the European Signal Processing Conference (EUSIPCO). IEEE; 2017. p. 1744–1748.

64. Pellegrini T. Densely connected CNNs for bird audio detection. In: Proceedings of the European Signal Processing Conference (EUSIPCO). IEEE; 2017. p. 1734–1738.

65. Delcroix M, Kinoshita K, Yu C, Ogawa A, Yoshioka T, Nakatani T. Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2015. p. 5270–5274.

66. Huemmer C, Delcroix M, Ogawa A, Kinoshita K, Nakatani T, Kellermann W. Online environmental adaptation of CNN-based acoustic models using spatial diffuseness features. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2017. p. 4875–4879.

67. Schwarz A, Huemmer C, Maas R, Kellermann W. Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments. In: Proceedings of theInternational Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2015. p. 4380–4384.

68. Delcroix M, Kinoshita K, Ogawa A, Huemmer C, Nakatani T. Context adaptive neural network-based acoustic models for rapid adaptation. Transactions on Audio, Speech, and Language Processing. 2018;26(5):895–908.

69. Jia X, De Brabandere B, Tuytelaars T, Gool LV. Dynamic Filter Networks. In: Proceedings of the Conference on Neural Information Processing Systems. NeurIPS; 2016. p. 667–675.

70. Wang Y, Getreuer P, Hughes T, Lyon RF, Saurous RA. Trainable frontend for robust and far-field keyword spotting. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE; 2017. p. 5670–5674.

71. McFee B, Kim JW, Cartwright M, Salamon J, Bittner RM, Bello JP. Open-Source Practices for Music Signal Processing Research: Recommendations for Transparent, Sustainable, and Reproducible Audio Research. Signal Processing Magazine. 2019;36(1):128–137. doi:10.1109/MSP.2018.2875349.

72. Dai J, Qi H, Xiong Y, Li Y, Zhang G. Deformable convolutional networks. In: Procedings of the International Conference on Computer Vision (ICCV). IEEE; 2017. p. 764–773.

73. Ha D, Dai A, Le QV. HyperNetworks. In: Proceedings of the International Conference on Learnining Represention (ICLR); 2017. p. 1–29.

74. Li D, Chen X, Zhang Z, Huang K. Learning deep context-aware features over body and latent parts for person re-identification. In: Procedings of the Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017. p. 384–393.

75. Salamon J, Bello JP. Deep convolutional neural networks and data augmentation for environmental sound classification. Signal Processing Letters. 2017;24(3):279–283.

76. Salamon J, Jacoby C, Bello JP. A Dataset and Taxonomy for Urban Sound Research. In: International Conference on Multimedia. Association for Computing Machinery; 2014. p. 1041–1044.

77. Salamon J, Bello JP, Farnsworth A, Robbins M, Keen S, Klinck H, et al. Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLOS One. 2016;11(11).

78. Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In: Proceedings of the International Conference on Learning Representations (ICLR); 2015. p. 1–15.

79. Chollet F. Keras v2.0.0; 2018. https://github.com/fchollet/keras.

80. McFee B, Jacoby C, Humphrey E. pescador; 2017. Available from: https://doi.org/10.5281/zenodo.400700.

81. Bello JP, Daudet L, Abdallah S, Duxbury C, Davies M, Sandler MB. A tutorial on onset detection in music signals. Transactions on Speech and Audio Processing. 2005;13(5):1035–1047.

82. Battenberg E, Child R, Coates A, Fougner C, Gaur Y, Huang J, et al. Reducing bias in production speech models. arXiv preprint 170504400. 2017;.

83. Shan C, Zhang J, Wang Y, Xie L. Attention-based End-to-End Models for Small-Footprint Keyword Spotting. arXiv preprint arXiv:180310916. 2018;.

84. Lostanlen V, Salamon J, Cartwright M, McFee B, Farnsworth A, Kelling S, et al. Per-Channel Energy Normalization: Why and How. Signal Processing Letters. 2019;26(1):39–43. doi:10.1109/LSP.2018.2878620.

85. Franceschi JY, Fawzi A, Fawzi O. Robustness of classifiers to uniform ℓ^p and Gaussian noise. In: Procedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR; 2018. p. 1280–1288.

86. Krstulović S. Audio Event Recognition in the Smart Home. In: Virtanen T, Plumbley MD, Ellis D, editors. Computational Analysis of Sound Scenes and Events. Springer; 2018. p. 335–371.

87. Andén J, Lostanlen V, Mallat S. Joint time-frequency scattering for audio classification. In: Proceedings of the International Conference on Machine Learning for Signal Processing (MLSP). IEEE; 2015. p. 1–6.

88. McFee B, McVicar M, Balke S, Thomé C, Raffel C, Lee D, et al.. librosa/librosa: 0.6.1; 2018. Available from: https://doi.org/10.5281/zenodo.1252297.

89. McFee B, Humphrey EJ, Bello JP. A software framework for musical data augmentation. In: Procedings of the Conference of the International Society on Music Information Retrieval; 2015. p. 248–254.

90. Schlüter J, Grill T. Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks. In: Proceedings of the Conference of the International Society for Music Information Retrieval (ISMIR); 2015. p. 121–126.

91. Salamon J, MacConnell D, Cartwright M, Li P, Bello JP. Scaper: A library for soundscape synthesis and augmentation. In: Proceedings of the Workshop on Applications of Signal Processing to Acoustics and Audio (WASPAA). IEEE; 2017. p. 344–348.

92. Hopcroft JE, Karp RM. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. Journal on Computing. 1973;2(4):225–231.

93. Raffel C, McFee B, Humphrey EJ, Salamon J, Nieto O, Liang D, et al. mir_eval: a transparent implementation of common MIR metrics. In: Procedings of the Conference of the International Society for Music Information Retrieval (ISMIR); 2014. p. 367–372.