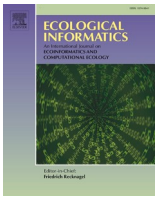
Ecological Informatics 72 (2022) 101909

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Method for passive acoustic monitoring of bird communities using UMAP and a deep neural network

Gabriel Morales [a], Víctor Vargas [b], Diego Espejo [b], Víctor Poblete [b,d,*], Jorge A. Tomasevic [c], Felipe Otondo [d], Juan G. Navedo [e,f]

[a] Master's Program of Acoustics and Vibrations, Graduate School of the Faculty of Engineering, Universidad Austral de Chile, Av. General Lagos 2086, Valdivia 5111187, Chile
[b] Audio Mining Laboratory (AuMiLab), Institute of Acoustics, Faculty of Engineering, Universidad Austral de Chile, Av. General Lagos 2086, Valdivia 5111187, Chile
[c] Centro de Humedales Río Cruces, Universidad Austral de Chile, Camino Cabo Blanco Alto s/n, Valdivia 5090000, Chile
[d] Institute of Acoustics, Faculty of Engineering, Universidad Austral de Chile, Av. General Lagos 2086, Valdivia 5111187, Chile
[e] Bird Ecology Lab, Instituto de Ciencias Marinas y Limnológicas, Universidad Austral de Chile, Campus Isla Teja, Valdivia 5090010, Chile
[f] Millennium Institute Biodiversity of Antarctic and Subantarctic Ecosystems (BASE), Chile

ARTICLE INFO

Keywords:
Passive acoustic monitoring
Bird community
Deep learning
Soundscape
Phenology

ABSTRACT

An effective practice for monitoring bird communities is the recognition and identification of their acoustic signals, whether simple, complex, fixed or variable. A method for the passive monitoring of diversity, activity and acoustic phenology of structural species of a bird community in an annual cycle is presented. The method includes the semi-automatic elaboration of a dataset of 22 vocal and instrumental forms of 16 species. To analyze bioacoustic richness, the UMAP algorithm was run on two parallel feature extraction channels. A convolutional neural network was trained using STFT-Mel spectrograms to perform the task of automatic identification of bird species. The predictive performance was evaluated by obtaining a minimum average precision of 0.79, a maximum equal to 1.0 and a mAP equal to 0.97. The model was applied to a huge set of passive recordings made in a network of urban wetlands for one year. The acoustic activity results were synchronized with climatological temperature data and sunlight hours. The results confirm that the proposed method allows for monitoring a taxonomically diverse group of birds that nourish the annual soundscape of an ecosystem, as well as detecting the presence of cryptic species that often go unnoticed.

1. Introduction

Birds acquire or learn a vocal repertoire to communicate (Marler, 2004; Miller et al., 2020). These signals can be very simple, with a single note or a cluster of notes (usually called syllables), or they can be very complex and variable, like oscine birdsong (Benitez Saldivar and Massoni, 2018; Moore et al., 2011; Wiley, 1991). Due to their neurobiological and sociobiological importance, acoustic signals have an essential role in different ecological processes (e.g., mating, defending territories, noticing potential danger) that ultimately affect individual bird fitness (see a review in Slater (2003)). Likewise, it is possible to find dialects in geographically separated groups of the same species (Otter et al., 2020; Toews, 2017; Wang et al., 2021). In short, the size of the characteristic vocal repertoire of a species is not always fixed, and the different signals that make it up are not always well stereotyped (Marler, 2004). However, the acoustic recognition of these signals—simple, complex, fixed and variable—is an effective practice for monitoring bird communities (LeBien et al., 2020; Luther, 2009; Morgan and Braasch, 2021; Zhong et al., 2021).

The ability to recognize and classify different bird calls permits the study of acoustic activity within a natural cycle (daily, seasonal or annual)—that is, the periodic occurrence of each call, repertoire and community soundscape. These temporal dynamics, analyzed with validated methods, offer answers about the phenology of birds (Demarée, 2011). For example, we can study migratory itineraries, the influence of the lunar phase or twilight synchronization, among other environmental correlations as a function of time. In fact, the study of acoustic phenology is providing excellent bioindicators that reflect the response

* Corresponding author at: Audio Mining Laboratory (AuMiLab), Institute of Acoustics, Faculty of Engineering, Universidad Austral de Chile, Av. General Lagos 2086, Valdivia 5111187, Chile.
E-mail address: vpoblete@uach.cl (V. Poblete).

https://doi.org/10.1016/j.ecoinf.2022.101909
Received 16 July 2022; Received in revised form 13 October 2022; Accepted 7 November 2022
Available online 18 November 2022

of certain species to climate change and other sources of anthropogenic disturbance (Buxton et al., 2016; Monczak et al., 2020; Saino et al., 2011).

Recently, an international team of 29 scientists reported a widespread decline in bird soundscape diversity and intensity at more than 200,000 sites in the Northern Hemisphere over the past 25 years due to significant declines in both species' richness and the abundance of individuals (Morrison et al., 2021). Therefore, it is imperative to redouble efforts and diversify the methods that contribute to the knowledge of the current state of diversity and acoustic activity in bird communities, especially in scarcely studied regions of the Southern Hemisphere.

Currently, non-invasive methods based on passive acoustic monitoring (PAM) have become relevant in ecological research, and their scope is increasingly wide and diverse (Clink and Klinck, 2021; Farina, 2014; Kowarski and Moors-Murphy, 2021; Sugai et al., 2019). However, an important aspect to consider is that this type of monitoring provides a large set of recordings, which in turn demands analysis processes based on automatic systems that are precise and efficient (LeBien et al., 2020).

Nowadays, unsupervised learning techniques, such as Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP), offer a suitable alternative for studying the topology of large datasets (McInnes et al., 2018). In the case of audio data, it is customary to previously perform a transformation of these temporal signals to a time-frequency feature space (i.e., spectrograms). The success of these algorithms lies mainly in their non-linear ability to reduce the high dimensionality of feature vectors, group them together if there is similarity, separate them if they show differences, and project them onto a low-dimensional graph space (Chepushtanova et al., 2020). Such capabilities complement the arduous task of auditory characterization carried out by experts, who are not free from perceptual biases (Sainburg et al., 2020).

On the other hand, supervised learning techniques based on deep artificial neural networks (deep learning) are suitable for identifying species through the automatic analysis of recordings that can come from a whole season in different simultaneous locations (Borowiec et al., 2021; Zhong et al., 2021).

Knowing the potential of these technologies and the context of climate emergency and biodiversity loss (Butt et al., 2021; Sueur et al., 2019; Tittensor et al., 2019), we wonder how we can precisely and exhaustively monitor the diversity and acoustic activity of birds. In addition, we are interested in examining some abiotic oscillations and their relationship with the annual dynamics of the calls of bird species as an approach to the field of acoustic phenology.

In this work, we document the detailed development of a method for passive monitoring of the activity, diversity and acoustic phenology of birds present in a scarcely studied ecoregion. The method combines advanced techniques in the construction of datasets, unsupervised learning and deep learning. We applied the UMAP algorithm to account for the inter- and intra-species acoustic diversity of a labeled dataset, then trained and put into production a deep learning model to extract the annual acoustic activity of each target species, and we then synchronized this data with environmental variables of climatological temperature and sunlight hours corresponding to the same annual study period.

2. Material and methods

During this investigation, we used the collected audio recordings of the associated project SoundLapse (Otondo and Poblete, 2020; Otondo and Rabello-Mestre, 2021). The SoundLapse project aims to investigate the natural and acoustic heritage of the urban wetlands in the city of Valdivia, Chile.

2.1. Areas, monitoring protocols, and period

We selected three acoustic monitoring areas within the urban wetland network that exists in the city: (1) Angachilla, (2) Miraflores, and (3) Parque Urbano El Bosque (Fig. 1).

The Angachilla urban wetland is located in the southern part of the city and is a tributary to and fed by the Angachilla River. The monitored segment corresponds to an area of the recently established Angachilla Urban Natural Reserve. Available eBird records, considering the hotspot closest to the monitoring point, indicate the presence of a total of 78 bird species (eBird, 2022). This area has experienced growing urbanization, and a variety of threats have been identified, including garbage dumping, deforestation, the introduction of exotic vegetation and ground use changes (Correa et al., 2018).

The second monitoring area covers the urban wetland Miraflores and a mixed residential-industrial zone, where shipyards and wood chipping facilities are in continuous operation (Poblete et al., 2021). Birds' diversity in this micro watershed is estimated at 34 species at the moment (eBird, 2022). However, given the known richness in the general area, this may be underestimated.

The monitoring radio in the wetland Parque Urbano El Bosque reaches one of the most residential areas of the city, including hospitals, schools and commercial activities. The recording unit was installed within the ecological reserve managed by the Lemu Lahuén Ecologic Committee, where frequent educational and cultural activities take place around this micro watershed. The observed birdlife in this hotspot reaches 51 species (eBird, 2022).

Autonomous recording units (ARU) were installed, and they have proven to be resistant to extreme environmental conditions (SM4 Wildlife Acoustics). The ARUs were attached to tree trunks at a height of 3 m. Sampling was automated to record the first five minutes of each hour for a year synchronically in these wetlands, achieving 2878.5 h of WAV-PCM stereo format audio with a depth of 16 bits. The chosen sample frequency was 44.1 kHz, allowing the possibility of capturing audio samples with a range of frequencies up to 22.05 kHz, well above prevailing frequencies as high as those produced by *Sicalis luteola*, *Aphrastura spinicauda*, or *Sephanoides sephaniodes*, which do not exceed 10 kHz (e.g., ML289668941 and ML371194791, macaulaylibrary.org). An annual monitoring period of 406 days was achieved, extending from October 19, 2019, to November 28, 2020.

2.2. Preliminary exploration

An auditive-spectrographic exploration of the obtained data was performed in collaboration with ornithologist experts from Rio Cruces Wetlands Center and Bird Ecology Lab. This preliminary approach was oriented towards the audiovisual recognition of the stereotypical vocal forms present in the birds' repertoire, with a special focus on the syntax and spectral structure of these forms (Wu et al., 2008), and sought to obtain a valid identification by species of each observed vocal form. Visiting the wetlands helped clarify minor disagreements regarding

Fig. 1. Map of the three monitoring areas: Angachilla, Miraflores, and Parque Urbano El Bosque in the city of Valdivia, Chile (−39.839, −73.243).

identification in the presence of on-site evidence of the sound event under discussion.

We systematized this exploration by creating lists of the species present in every five-minute audio recording. A total of 480 complete audios (corresponding to 1.6% of the entire annual audio recordings) were reviewed. These lists allowed us to observe the trends of the daily and seasonal vocal activity of 36 species, which were identified in the acoustic-spectrographic dimension. With this notion about the presence of species in each wetland, we selected a set of 16 target species (Table 1) based on the following criteria: (1) the quality and quantity of the available signals; (2) migratory species; (3) the balance between diurnal, crepuscular and nocturnal species; (4) representativity in the stational vocal activity; (5) vocal form diversity; and (6) the size of the repertoire. Each species was selected for at least one of the six criteria, and each criteria supports at least one target species.

2.3. Semi-automatic labeling

The material, which we obtained from one year of recordings in the three monitoring areas, was stored on the RFCx-ARBIMON platform (Aide et al., 2013). All the audio recordings were divided into one-minute chunks and compressed into a monophonic FLAC (free lossless audio codec) format. To construct a dataset of the vocal repertoires, we used the pattern matching algorithm to detect signals based on the creation of the spectrographic reference samples (templates) (LeBien et al., 2020). With this algorithm, recording lists (playlists) were analyzed encompassing the three areas and the four seasons.

This process required manual validation of each detection. If the template signal was totally or partially present, we labeled it as presence, while, if the signal was not related to the template and did not correspond to any other characteristic sign of the species, we labeled it as absence. Labeled absences corresponded mainly to human-related sounds, also called anthropophony, and abiotic sounds, also named geophony, or other species and taxa. Particularly, an effort was made to include vocal forms of other birds and of great similarity to the target signals in this category of absences. Doing so helped minimize fine mistakes in the training of the neural network and improved the learning.

With this semi-automatic technique, it was possible to register strong labeling of 22,243 presences and 125,257 absences from 22 templates associated with 16 target species (Table 1). Specific details of this method are described in Subsection 2.2 of LeBien et al. (2020).

Each label included the name of the audio file, area, year, month, day, hour, minute, start second, end second, species, vocal form (template) and validation (presence or absence), among other metadata. Regarding the time interval in which the target signal occurred (start and end), this data was determined by the duration d of each template, which varied between $d = [0.51s, 5.30s]$. For example, three vocal forms were selected from the *Agelasticus thilius* repertoire and labeled as a, b and c. For those repertoires without fixed forms or with many vocal forms, the single label multi was assigned (for example, the multiplicitous trills of the vocal forms of *Cistothorus platensis*). In some cases, the stereotypical forms were contained in more general sequences (i.e., syllables within a phrase); in these cases, they were also single-labeled (for example, the vocal form c of *Elaenia albiceps* was contained in a more general sequence) (Fig. 2).

This fine labeling allowed us to monitor the intraspecies dataset balance.

Table 1
Number of labeled samples of presences p and absences a.

	Vocal form	p	a
1.Mareca sibilatrix	a	1052	1728
2.Pardirallus sanguinolentus	a	1318	12,146
3.Porzana spiloptera	a	1413	17,616
4.Gallinago magellanica	a	1010	14,993
5.Glaucidium nana	a and b	718	21,813
6.Eugralla paradoxa	a	459	904
7.Scytalopus magellanicus	a	1018	1117
8.Phleocryptes melanops	a	639	1078
9.Anairetes parulus	a	302	1061
10.Elaenia albiceps	a, b and c	4137	19,549
11.Colorhamphus parvirostris	a	1256	17,411
12.Troglodytes aedon	multi	710	1000
13.Cistothorus platensis	multi	1434	8515
14.Curaeus curaeus	multi	603	1003
15.Agelasticus thilius	a, b and c	5014	9540
16.Sicalis luteola	a and b	1160	1168
Total	22	22,243	130,642
Media	–	1390	8165

2.4. Feature extraction and visualization

We adjusted the start and end data of all labels to obtain a fixed duration of $d = 2s$. This value represented an attempt to reach a minimum signal within extended signals and not exceed facing short signals (Fig. 2).

2.4.1. Spectrograms

The audio data of these fragments (presence and absence) were extracted with a sampling rate of 44,100 samples per second. Later, they were transformed through the time-frequency space using the short-time Fourier transform (STFT) (Oppenheim et al., 1999) with a Hann window function of 1024 samples (~ 23 ms) and 50% overlap (512 samples), and the Fourier transform was applied using 2048 bins. The quadratic magnitude spectrums were mapped through the Mel frequency scale (Davis and Mermelstein, 1980) using a filter bank of 128 bands and converted to normalized units of decibels with reference to the maximum quadratic magnitude of each spectrogram. We computed a bidimensional array for each fragment to represent spectrograms, and a color palette was assigned to represent the energy values.

2.4.2. Acoustic feature visualization

To obtain a visual perspective of the inter- and intra-species variety of the labeled vocal forms, we used the nonlinear dimensionality-reduction technique UMAP. UMAP analyzes the topologies formed by high-dimensional data and projects an approximation in 2D or 3D without modifying the original topologies. In this case, the data were hyper-vectors of acoustic features extracted from each temporal signal labeled as presence. The result was a map in the feature space through which it was possible to appreciate clouds of points that represent the clusterings of the vocal forms.

The feature extraction stage was conducted with two parallel channels. The temporal signals were transformed to the time-frequency domain using the STFT through one channel with the same values used to generate the spectrograms. Another channel was used to apply the chirplet transform (Mann and Haykin, 1991), which consists of a signal expansion over the modulated transient waves in amplitude and frequency called a chirplet. These modulated transient waves model the extremely fast frequency sweeps that are characteristic of bird vocalizations, providing a suitable framework for their representation (Xie et al., 2018). A fast chirplet transform (FCT) was implemented and validated using recordings of whales and birds (Glotin et al., 2017).

In both channels, we computed a time-frequency matrix for the 22,243 presence samples. We then flattened these matrices were flattened to obtain hyper-vectors with 11,136 and 22,000 coefficients to the STFT and FCT respectively. Once the high-dimensional feature extraction was stored, we executed the UMAP technique was in both channels. In order to avoid losing the globality of the total set and to maintain a uniform dispersion of the plotted points, the following three basic UMAP parameters were adjusted: the random state was equal to 42, the nearest points for each data point were equal to 20 and the effective minimum distance between embedded points was equal to 0.5. These values were achieved by experimentation after an aesthetic appreciation of the

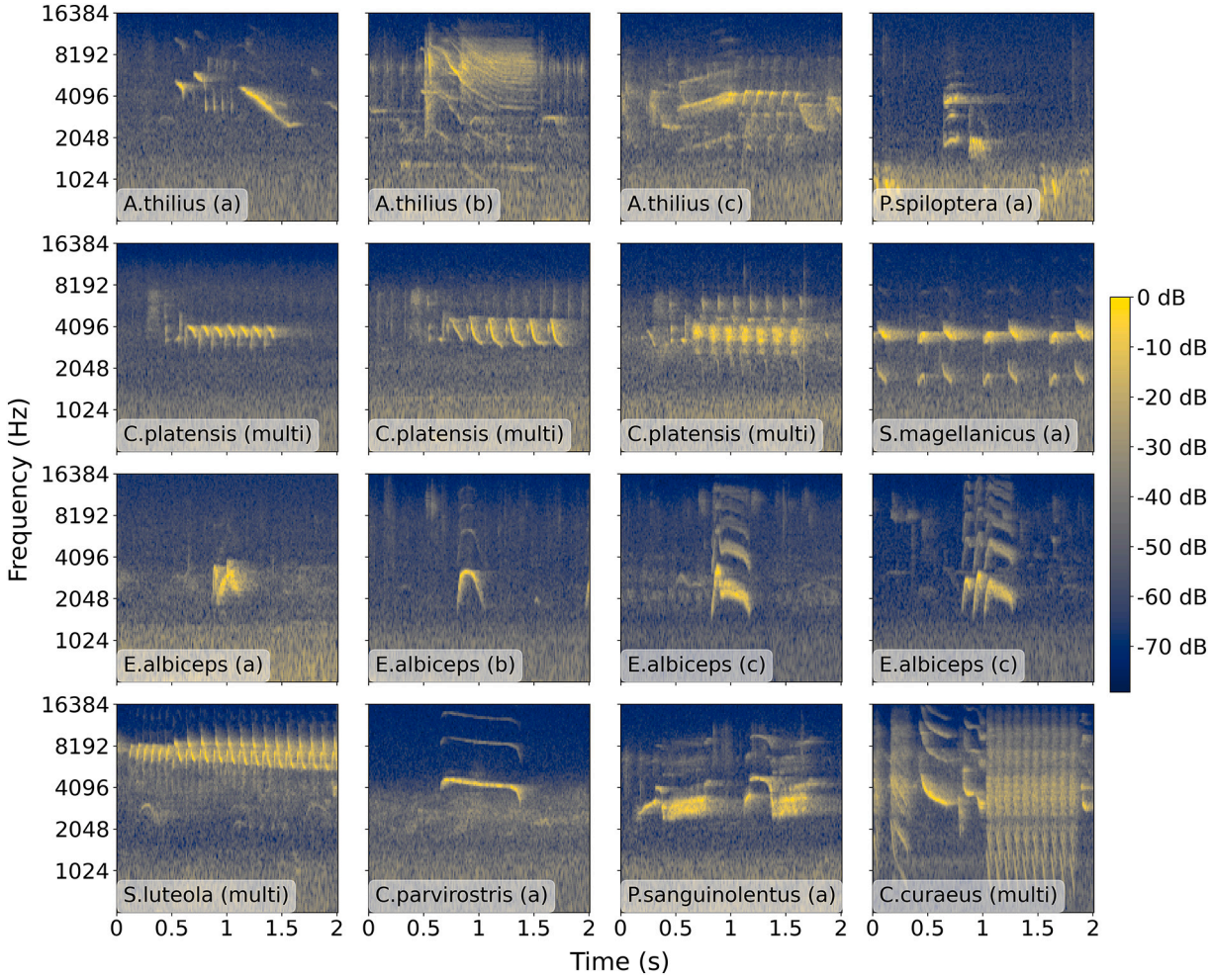


Fig. 2. The STFT-Mel spectrograms of some vocal forms associated with specific labels.

resultant projection (see Fig. 4 in Section 3).

2.5. Findings, simulation and data augmentation

The labels of the vocal form (a) of *Porzana spiloptera* corresponded to the territorial singing of this species, which has been catalogued in a state of vulnerability by the Red List of Threatened Species of the International Union for Conservation of Nature (IUCN). These labels represent an important finding as a result of what is described in Subsections 2.2 and 2.3. Its presence at this latitude of the Valdivian ecoregion was expected to be confirmed due to a recent bycatch of a juvenile individual in the Río Cruces wetland 6 km north of the Miraflores wetland, (Ruiz et al., 2022). Thanks to these acoustic inspections, it was possible to confirm an important vocal activity of at least two individuals. However, manual efforts to find the other known calls of *P. spiloptera* were insufficient. Motivated to detect the contact calls of this vulnerable species, we designed a simulation and data augmentation from external recordings of the call described as *pw call*, which is available on the platform Xeno-canto. Thirty-four samples of this call were downloaded, and they were recorded in the Otamendi Nature Reserve, Argentina (for example, XC64503, xeno-canto.org).

This method consisted of applying known augmentation techniques (Lasseck, 2019) that slightly deformed the signal to represent changes in the distinctive patterns of individuals within the same species. These individual differences may be the product of anatomical variations of the vocal tract or inexact matches of the innate or learned auditory template, and they may even occur through improvisation (Stowell et al., 2019).

We processed these 34 audio signals to isolate them from the original acoustic environment. The following chain was applied with specific parameters for each of the 34 samples: (1) segmentation, (2) frequency filtering and (3) spectral subtraction of stationary noise (Boll, 1979), which was captured from some period of absence of the target signal within the same sample.

Subsequently, a Python function was built that, from a *pw call* sample, delivered 12 slightly transformed samples with energy reduction, stretching and contraction in the time domain (time stretch); in frequency (pitch shift); and mixtures between them. The next step was to collect samples without the presence of *P. spiloptera*, which were randomly extracted from the three local wetlands, and then mix the local soundscapes and the external *pw calls*. In total, 1224 simulated *pw call* samples were collected.

We employed the same method to simulate and augment samples of only one additional species, *Glaucidium nana*, which was of particular interest for the representation of nocturnal activity.

2.6. Automatic identification task layout

We carried out the construction of classes for training, validation and testing of the neural network by grouping vocal forms corresponding to the same species, except those that were subjected to simulation and data augmentation, such as *P. spiloptera* and *G. nana*, since these were sub-divided into two classes each—*P. spiloptera (a)*, *P. spiloptera (b)*, *G. nana (a)* and *G. nana (b)*—reconfiguring the dataset into 18 classes

associated with 16 species (Table 2). In this way, the automatic species identification task was designed with an additional differentiation in two of them.

2.7. Convolutional neural network (CNN) and transfer learning-based model

We implemented a convolutional neural network (CNN) (LeCun and Bengio, 1995) to perform the automatic task of detection and identification of 18 classes. A CNN is a type of network inspired by the visual cortex that shows good performance in recognizing objects in images, and its use has been extended to the acoustic domain (Aloysius and Geetha, 2017; Stowell, 2022). The architecture is organized according to a hierarchy of layers designed for feature extraction based on a digital image. The first layers can detect edges, then simple shapes, and deeper layers can learn more complex shapes with a high level of abstraction (Albawi et al., 2017; Krizhevsky et al., 2012; LeCun et al., 2015).

We applied the transfer learning and fine-tuning technique with the deep learning model ResNet50 (He et al., 2016; Tan et al., 2018), which was already pre-trained on ImageNet data (Deng et al., 2009). Transfer learning is a technique that enabled us to overcome the issue of shortage of training data and construct a model efficiently by transferring knowledge from a similar task to, in this case, our target task. The implemented architecture received RGB images of $224\times224\times3$ as inputs—in this case, color spectrograms of two-second durations—and included only the feature extraction layers from the ResNet50, discarding the superior classification layers (known as the top network). The newly created model reconfigured the top network with two fully connected layers (FC) that could learn new features. To reduce overfitting and imitate the training of a set of different models, a dropout layer was intercalated between both FCs. The final layer was compounded by 18 nodes, which delivered an 18-punctuation vector representing the presence probabilities for each of the 18 classes. For more details, we have left a repository with the architecture implemented in CNN-Bioacoustic-monitoring.

Following the approach proposed by LeBien et al. (2020), we adopted a custom training loss function that took advantage of the samples of both presences and absences in the training stage and permitted multiclass learning from the labeled samples with a single class. The learning process setup was completed with an Adam optimizer (Kingma and Ba, 2017), setting its learning rate and decay to 1×10^{-4} and 1×10^{-7} respectively.

We randomly split the 18 classes of the labeled and augmented dataset into training and testing datasets; specifically, we used 80% for training and 20% for testing with the presence and absence samples. In the training stage, 10% of the dataset was used as a validation subset. It was likely that samples extracted from the same audio file were separated, both for training and for testing and validation. However, this bias was reduced, in the semi-automatic labeling stage, by limiting the extraction of samples to a maximum of three per file.

The iterative learning process of the model was executed by proposing 50 training epochs. The early stopping method was used with a patience equal to 5, to avoid overfitting. In this way, optimal training was achieved at the end of the ninth epoch, reaching a maximum loss of 4.5×10^{-4} and 1.1×10^{-3} for the training and validation subsets respectively.

2.8. Evaluation

Once the model was trained, we assessed the model's performance using the split sample subset (spectrograms of two seconds) for testing. Thus, we did not use those samples in the training stage. The multiclass predictions of the model were assessed with the typical indicator series: true positives (*TP*), false positives (*FP*), true negatives (*TN*) and false negatives (*FN*). Because the model gave probability values in the $[0,1]$ range, the prediction depended on a decision threshold θ to define it as positive (presence) or negative (absence). With these θ-dependent elemental indicators, two useful relationships were obtained, and these helped assess different facets of the model performance. One of them was the precision *P*:

$$P_{(\theta)} = \frac{TP}{TP + FP} \tag{1}$$

Eq. 1 represents the success number of all presence predictions. On the other hand, the recall *R* expresses the proportion of well-detected presence—that is:

$$R_{(\theta)} = \frac{TP}{TP + FN} \tag{2}$$

Precision and recall are inversely related metrics as we increase or decrease the threshold θ. To achieve a balance between both metrics, it is common to draw a precision-recall curve ($P-R$ curve) as a result of the assessment of all possible thresholds θ. This plot allowed us to analyze from which recall we had a degradation of precision and vice versa. To summarize this curve in a single value, the average precision (*AP*) was computed:

Table 2
Dataset for artificial neural network training, validation and testing.

Class	Vocal form	Labeled presences	Simulated presences	Total presences	Absences
1.Mareca sibilatrix	a	1052	–	1052	1728
2.Pardirallus sanguinolentus	a	1318	–	1318	12,146
3.Porzana spiloptera	a	1413	–	1413	17,616
4.Porzana spiloptera	b (*pw call*)	–	1224	1224	7891
5.Gallinago magellanica	a	1010	–	1010	14,993
6.Glaucidium nana	a	160	420	580	11,202
7.Glaucidium nana	b	558	500	1058	10,601
8.Eugralla paradoxa	a	459	–	459	904
9.Scytalopus magellanicus	a	1018	–	1018	1117
10.Phleocryptes melanops	a	639	–	639	1078
11.Anairetes parulus	a	302	–	302	1061
12.Elaenia albiceps	a, b and c	4137	–	4137	19,549
13.Colorhamphus parvirostris	a	1256	–	1256	17,411
14.Troglodytes aedon	multi	710	–	710	1000
15.Cistothorus platensis	multi	1434	–	1434	8515
16.Curaeus curaeus	multi	603	–	603	1003
17.Agelasticus thilius	a, b and c	5014	–	5014	9540
18.Sicalis luteola	a and b	1160	–	1160	1168
Total	23	22,243	2144	24,387	138,523
Media	–	1390	–	1355	7696

$$AP = \sum_{i=2}^{n} \left(R_{(\theta_i)} - R_{(\theta_i - 1)}\right) P_{(\theta_i)} \tag{3}$$

where n represents the total number of thresholds to assess, which are organized in descending form. This metric corresponded to the weighted metrics of the achieved precisions in each threshold, with the increase in the previously used threshold recall being weighted, i.e., an approximation of the area under the $P - R$ curve.

As a measure to evaluate and compare the performance of the multiclass classification models, we used the average of all the *AP* obtained by each class (*mAP*):

$$mAP = \frac{1}{N_c} \sum_{k=1}^{N_c} AP_k \tag{4}$$

where the N_c is the number of classes. The advantage of using *mAP* is that it is independent of the selected threshold for the implementation of the model. The closer its value is to one, the better the model will be.

We searched a θ-threshold by class to provide an optimum balance between precision and recall metrics. This was done by observing the highest measure of the harmonic mean ($F_1 score$) and the most general measure variations of the *Fscore*:

$$F_\beta = \left(1 + \beta^2\right) \frac{P_{(\theta)} R_{(\theta)}}{P_{(\theta)} \beta^2 + R_{(\theta)}} \tag{5}$$

This decreased the β coefficient up to 0.5 to provide a higher ponderation to the precision than to the recall (Kahl et al., 2021).

2.9. Non-acoustic environmental variables

We studied data from non-acoustic domains provided by the climatic service of the Chilean Meteorological Directorate. The data were collected at the Pichoy Station (−39.657, −73.087) located 22 km north of the city of Valdivia at an altitude of 18 m.a.s.l. Records of mean daily climatological temperature (°C), total monthly precipitation (mm of accumulated precipitation) and sunrise and sunset times were acquired from October 19, 2019, to November 27, 2020 (406 days). We adjusted the data to the UTC-3 time standard, which corresponds to summertime in Chile. We did not consider the change to wintertime so as to provide continuity in the graphs and to not disturb the interpretation.

2.10. Production

We designed an algorithm to apply the model over a year of unlabeled recordings. A Python code block was written based on a function that offered the following results: (1) a daily detected occurrence vector (positive predictions of a class) in the first five minutes of each hour of the day (24 audio recordings of five minutes), (2) a vector whose components represent the probability value associated with each occurrence and (3) a location vector with a timestamp for each occurrence.

With this detection, quantification and localization instrument, which collected daily information, 406 days of recordings were processed to obtain the annual dynamic in two axes: hours and days. These bioacoustic data separated by species, together with the variables of sunlight and temperature, were plotted on graphs known as heatmaps to observe activity and environmental correlations. We tested the visualization with data separated by area representing local activity (local heatmap), and we also tested it with detection data from the three recording areas, representing the joint activity of the entire monitored zone (zonal heatmap).

On the other hand, we wrote Python scripts to extract the samples of each located audio, which was classified by species and area without losing the specific timestamp of each sample. All the mentioned codes are available in the GitHub repository.

The thresholds obtained according to *GMean* and *Fscore* in the evaluation process led to insufficient data to analyze the annual acoustic activity of species that naturally have a lower occurrence rate (for example, *P. spiloptera* and *C. parvirostris*). We solved this dilemma by lowering the thresholds associated with these species. In this way, the sensitivity of the model against infrequent signals was increased, finding an operating point without adding too many false positives. Since we applied the model to unlabeled recordings, we manually reviewed the detections located in atypical times and seasons to identify false positives that would cause greater distortions in the interpretation. This allowed us to empirically adjust the best threshold value.

Because the entire year needed to be exhaustively analyzed and the spatial monitoring coverage was minimal (i.e., only three areas), the recordings containing training samples were not excluded.

The workflow is summarized in Fig. 3 in order to facilitate understanding of the inputs and outputs of the techniques used throughout the process.

3. Results

The UMAP technique (Fig. 4) elegantly and effectively revealed the variety and clustering of the representative samples given in both feature spaces (STFT and FCT). This allowed us to predict an encouraging forecast of separability in a supervised learning process. Considering the high dimensionality of the hyper-vectors extracted with FCT (22,000 components versus 11,136 with STFT) and the longer processing time (160 ms per sample versus 3 ms with STFT), the creation of spectrograms for the training stage was calculated with STFT only, as described in Subsection 2.4.1.

CNN training was run successfully, and the optimized weights of the model were stored. To evaluate the predictive performance in the testing dataset, we plotted the $P - R$ curves of all classes on the same graph (Fig. 5). The points of each curve closest to the coordinate $(1, 1)$ correspond to the best performance in the prediction, with the threshold θ associated with that point. The observed performance was ideal in the *Scytalopus magellanicus* class and very close to the ideal in a majority group of classes. The curves that were separated from this group represented classes far from the ideal: *Glaucidium nana* (a) and *Glaucidium nana* (b), which in turn were augmented with simulated samples. This suggests three likely sources of conflict: (1) inconsistency between real and simulated samples, (2) less relative differentiation between classes belonging to the same species and (3), in the case of *G. nana* (a), a gap of insufficient time (two seconds) to cover the distinctive gestures of this vocal form that can easily be extended to four seconds.

The $P - R$ curves were summarized in the usual values of *AP* shown in the legend of the figure. As a general result of the evaluation, a mAP

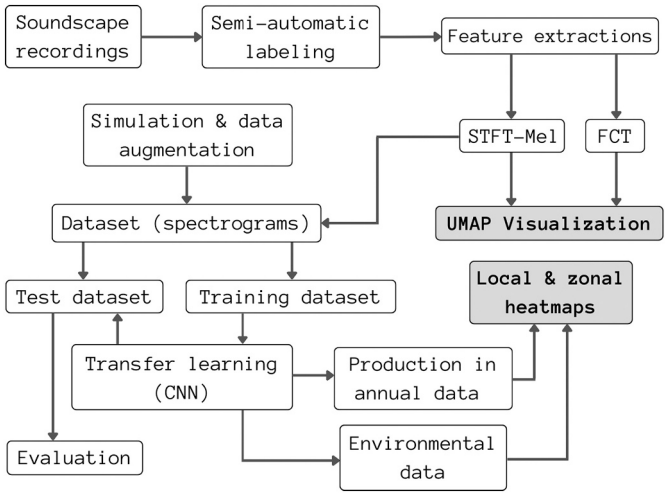


Fig. 3. Workflow diagram of the designed method.

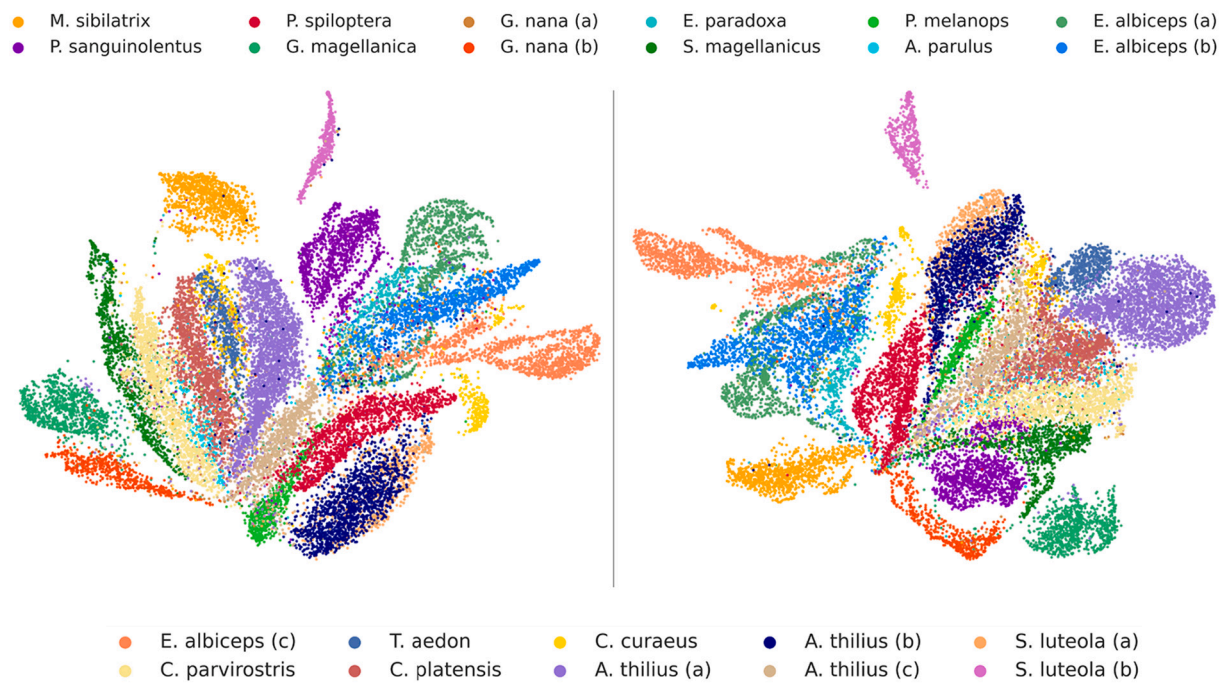


Fig. 4. UMAP visualization from the hyper-vectors of acoustic features, corresponding to the 22,243 samples labeled as presence. Left: UMAP fed with STFT features. Right: UMAP fed with FCT features.

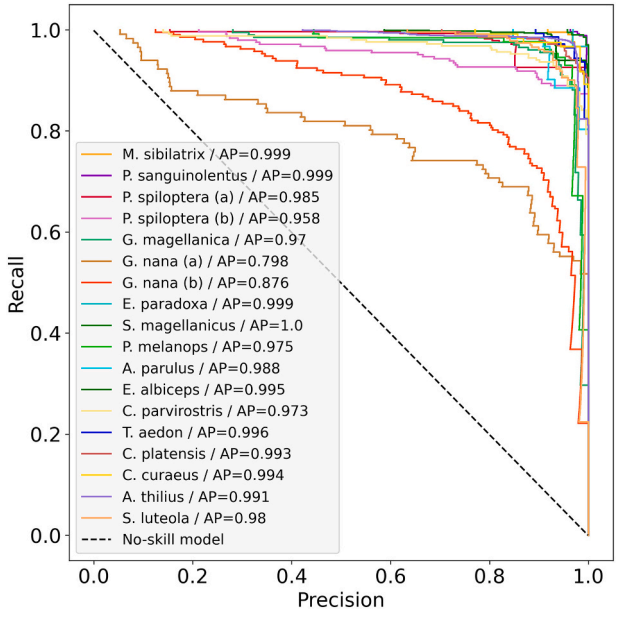


Fig. 5. Predictive performance of 18 classes according to precision-recall curves and *AP* per class.

(mean of *AP*) equal to 0.97 was obtained.

The application of the model in unlabeled recordings showed, through a non-exhaustive manual review, a better performance in the classes that were trained with a large number of absences (>8000 samples).

We made six zonal heatmaps with those classes that obtained a consistent recovery of information (Fig. 6).

With a threshold of $\theta = 0.99$, detections of *P. sanguinolentus* indicate a concentration of evening crepuscular activity in the four seasons. A high density of occurrences was detected on spring days and nights, with a higher concentration around sunrise and sunset, decreasing towards the December solstice. Detections during daylight hours were minimal on summer days with higher recorded temperatures. The abundant bright green colored lines, indicating more than 20 detections in five minutes, confirmed the persistent behavior of territorial song containing the vocal form chosen for CNN training (e.g., XC721627, xeno-canto.org).

The predictions of *G. magellanica* that exceeded the threshold $\theta = 0.57$ show the activity associated with the instrumental (non-vocal) call described as «winnow» (Miller et al., 2020). It initiated its instrumental repertoire in perfect synchronicity with the June solstice on the shortest and coldest days of the year. The season reached a high concentration of occurrences in October, culminating during the first days of November. Its acoustic display occurred markedly from twilight to night throughout the season.

E. albiceps, a Neotropical austral migrant that plays a key role in the regeneration of the Patagonian forests (Bravo et al., 2017), was detected from three vocal forms that were taken to training, achieving practically complete learning of its repertoire. This allowed for highly accurate and comprehensive results to determine vocal activity in their breeding season at this latitude. With a threshold of $\theta = 0.99$, the first occurrences detected upon arrival of their spring migration were found starting the third week of October, while their gradual withdrawal began to be reflected once the summer days with the highest recorded temperature had passed, during which the highest concentration of occurrences was produced. Detections completely disappeared at the March equinox. Their vocal routine occurred strictly during the daytime. Atypical detections were reviewed, finding only two false positives registered in winter. The error was caused by two signals with a very similar spectrographic shape: a note produced by *Troglodytes aedon* and an emergency vehicle siren.

With a threshold of $\theta = 0.20$, the vocal presence of *C. parvirostris*, another migratory flycatcher, was detected. Vocal activity was detected only in the autumn and winter seasons, with occurrences during daylight hours. Nighttime detections (outliers) were reviewed and attributed to an unidentified whistle, a girl's scream, a distant grinding noise and a security alarm—anthropogenic sounds similar in pitch and duration to the spectrographic form learned by the CNN. During the reproductive season (November to February) (McGehee et al., 2004), no

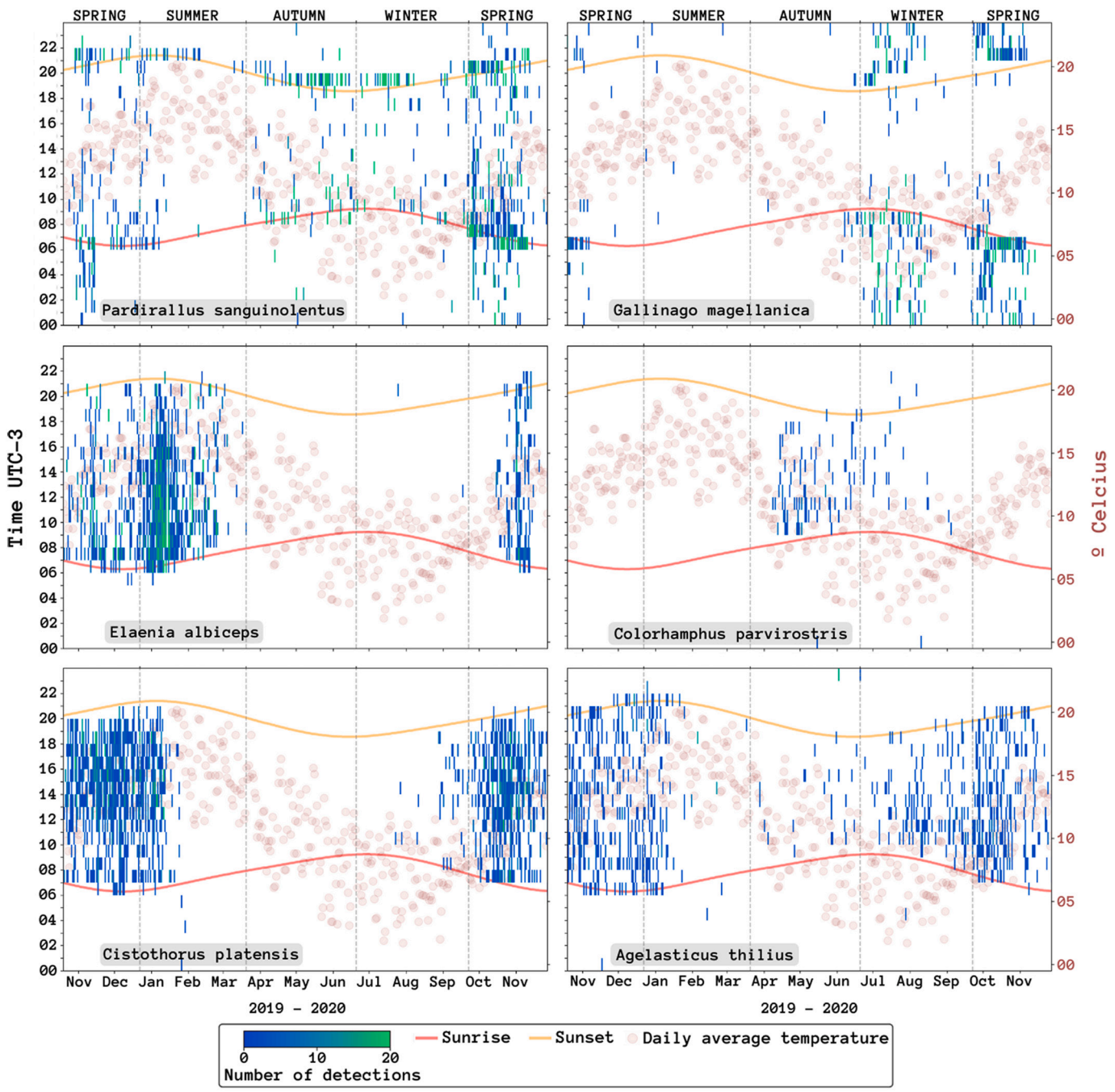


Fig. 6. Zonal heatmaps with bioacoustic activity detections of 6 bird species, synchronized with data on hours of sunlight and average daily temperature, between October 2019 and November 2020. Each colored line from blue to green represents the number of detections in the first 5 min of a given hour and day. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

acoustic activity was detected in the three areas.

The multiform trills of *C. platensis* were detected with predictions above a threshold of $\theta = 0.95$. Marked diurnal activity was detected in the spring season—from the September equinox until the second week of January, after the solstice and before the highest temperatures recorded. Detections at atypical times and seasons revealed false positives associated with a species of amphibian (*Batrachyla leptopus*) that is very common in these urban wetlands (Nuñez et al., 2020). Its spectrographic mark shares a relatively similar shape to that of *C. platensis* but is one octave lower. Although the nocturnal vocal activity of this anuran was numerous between February and April, only three occurrences confused the model.

In the same way that *E. albiceps* was treated, three vocal forms of *A. thilius* were trained. However, as it is a species of the *oscine* group, its repertoire is more abundant and heterogeneous. Considering the nature of *A. thilius*, the most observed forms were chosen in the preliminary exploration described in Subsection 2.2. This allowed learning by the CNN that was appropriate for the studied acoustic environment. The predictions that exceeded a threshold $\theta = 0.10$ formed the zonal heatmap, in which a large diurnal vocal display was observed in the spring months and part of the summer until the second week of January. Significant winter activity was also detected, along with activity to a lesser extent during autumn days. Atypical nocturnal detections were associated with false positives due to overlapping sounds in two similar sound signals: (1) vocalizations of *Mareca sibilatrix* and (2) anthropogenic emissions from residential alarms.

Regarding the representation of local heatmaps, in addition to the environmental correlations mentioned, the detections of *Pardirallus sanguinolentus* indicated itinerant behavior in two monitored areas (Fig. 7). Thanks to the fact that the monitoring period covered the last two weeks of October and the first of November, both in 2019 and 2020, we could make a comparison between the same spring season of two consecutive years. Although, in some areas, we collected more recordings, these three weeks were the minimum common period for comparison. In Miraflores, during November 2020, the high activity detected in November 2019 did not occur again, and, to the contrary, in the wetland Parque Urbano El Bosque, during the minimum period mentioned in 2020, important vocal activity was detected that was totally absent the previous year. Observations in the Angachilla wetland remained relatively consistent between both seasons.

Another relevant local heatmap of this research was the one dedicated to *Porzana spiloptera*, since the questions regarding the spatial and temporal location of this vulnerable species were clarified. With a threshold of $\theta = 0.17$, it was possible to retrieve minimal and sufficient information regarding the territorial song associated with the form (a) learned by the CNN only in the wetland Miraflores (Fig. 8). As for the temporal activity, it was located in high concentration during the nights from the second week of January to the second week of February, in synchrony with the records of the highest average daily temperature. The review of atypical detections revealed false positives due to confusion with short calls of *Turdus falcklandii* in winter, a flight call of *Theristicus melanopis* and the call of the endemic Chilean amphibian *Eupsophus roseus*.

4. Discussion

With only three units and a minimum sampling of five minutes every hour, our method was able to build a representative dataset of the bird community in an intricate soundscape (i.e., high probability of overlap and low signal-to-noise ratio).

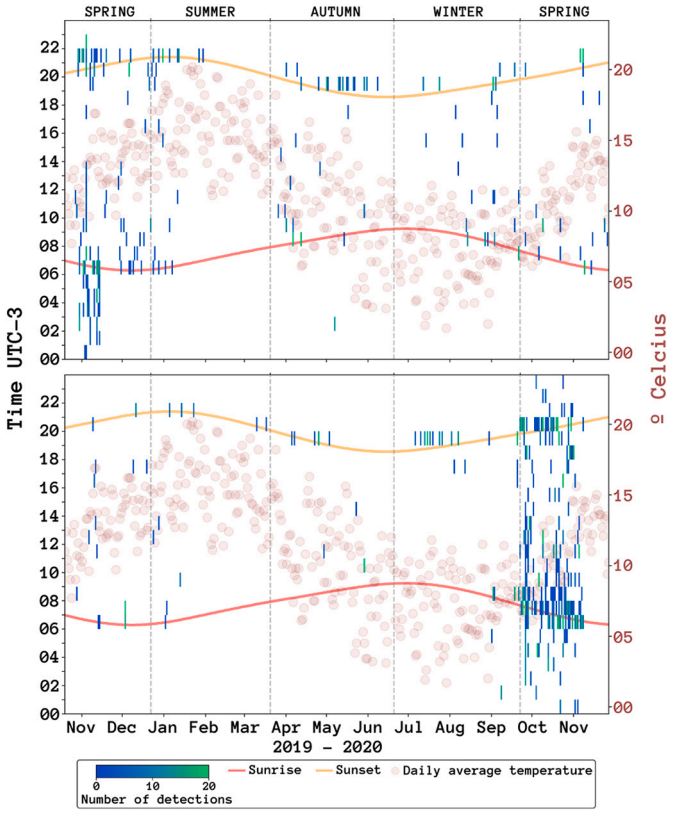


Fig. 7. Local heatmaps with detections of bioacoustic activity of *Pardirallus sanguinolentus*, suggesting itinerant behavior in the wetlands: Miraflores (top) and Parque Urbano El Bosque (bottom).

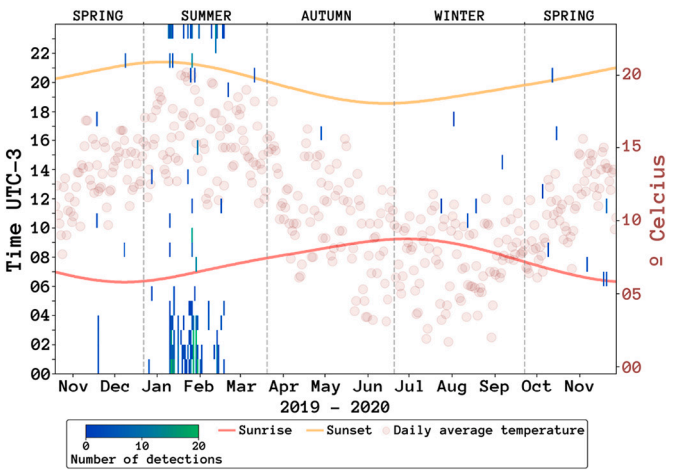


Fig. 8. Local heatmaps with detections in the Miraflores wetland, the only monitored area that revealed the presence of *Porzana spiloptera* from the vocal form (a) that was taken to CNN training.

The task of building a dataset was carried out in an agile and friendly way thanks to the pattern matching algorithm integrated in the platform RFCx-ARBIMON, although it presented difficulties when searching for signals that were weak or infrequent or had unstable spectrographic forms. For these cases, a good solution could be to increase the spatial coverage with more recording units, increase the duration of the time windows sampled by these units and/or oversample the moments of greatest bioacoustic activity, such as sunrise and sunset. However, this is subject to the technical and human resources of the investigation (Wood et al., 2021).

Both STFT-Mel and FCT provided equally useful features to feed the UMAP algorithm and obtain a graphical representation (Fig. 4) of the diversity of inter- and intra-species vocal forms. The consistent clustering achieved by UMAP confirms the usefulness of this technique for creating bioacoustic diversity maps. No statistical comparisons were made between the three monitored areas due to the imbalance of data between them, but the door remains open for a future comparative community analysis to account for potential differences in bioacoustic diversity. This could be done based on an ecoacoustic approach, applying the UMAP algorithm on labeled data and making comparisons with well-documented indices based on energy and frequency (Farina, 2018).

The decision to train the CNN with STFT spectrograms and discard FCT derived solely from reasons surrounding computational cost. It is possible that, with an optimization of parameters in FCT, the computation times could have been reduced and a better spectrographic representation could have been obtained to train the CNN, as suggested in Glotin et al. (2017) and in Xie et al. (2018).

The training and test datasets presented a significant imbalance between presences and absences, with a tendency towards a greater number of absences in various classes (Table 2). However, this is consistent with the natural composition in passive recordings, where there are often long periods of inactivity (absence) of the target species.

The impact of the imbalance between the number of tagged absences was only noticed once we applied the model to a large number of recordings, revealing an insufficient test subset for an assessment of generalizability. Therefore, we advise those who follow this approach of the importance of creating training and test sets with numerous and diverse absence samples, thus striving to include similar signals from other species not included in the classifier design, as they are the main source of confusion. Consideration should also be given to expanding the number of classes with those species that are more abundant and have

important acoustic activity, even though they are not a species of interest, since estimation methods based on deep learning allow for a large number of classes and benefit from numerous and diverse data.

It is very likely that the high performance shown in the test subset is partially biased by the individuals' vocalizations that were recorded for both training and testing. Consequently, we recommend reserving monitoring areas for the creation of validation and test sets, avoiding a link with the training data and allowing the generalization of the model to be correctly estimated. This should speed up threshold calibration during production.

The *pw call* sample simulation of *P. spiloptera* could not be properly evaluated due to the unavoidable link between training, validation and testing. It is unclear whether the model managed to learn to recognize this simple and highly stereotyped shape. The continuation of this research should improve the evaluation methodology in training cases with completely simulated data and/or search for innovative alternatives to face scenarios with minimal data, which is known as the few-shot learning method (Morfi et al., 2021).

The implemented CNN corresponded to a modified version of ResNet50, one of the dominant architectures in bioacoustic tasks, and, although other authors have applied previous ImageNet training to the bioacoustic domain (LeBien et al., 2020; Zhong et al., 2021), other datasets such as Audio Set (Gemmeke et al., 2017) or VGG-Sound (Chen et al., 2020) can be just as good as ImageNet for pre-training, either on ResNet or on other architectures, such as VGGish, Inception or MobileNet. Another viable option is to pretrain with synthetic clicks or chirps (Glotin et al., 2017; Yang et al., 2021). Models already available in mobile apps that perform this same spectrogram-based identification task are advancing rapidly. To date (October 2022), the BirdNet application (Kahl et al., 2021) allows for the identification of more than 3000 bird species (Wood et al., 2022). In the short term, this particular model is expected to encompass a large majority of acoustically active species worldwide, including the Valdivian ecoregion. However, these applications are optimized for recordings with target signals in the foreground, i.e., performance decreases drastically in passive recordings of soundscapes (Kahl et al., 2021). The latter is of consideration in a site−/time-specific investigation like ours, where high performance is required in noise conditions and optimized learning in local repertoires (LeBien et al., 2020), which justifies model training with local data. For a comprehensive view of the state of the art in bioacoustic monitoring from 2016 onwards, we suggest the reader review the recent publication by Stowell (2022).

By making an annual estimate, the model achieved potentially relevant results, automatically performing a second-by-second, class-by-class analysis on a large set of recordings. The results demonstrated the importance of a zonal approach (i.e., adding the detections of the three areas in the same graph) to visualize a greater number of detections, without losing sight of the local activity of each monitored area, allowing us to report the location of the little-known species *Porzana spiloptera* and suggesting population displacement of *Pardirallus sanguinolentus*.

The observed synchrony of the vocal activity of certain species with the environmental oscillations of temperature and sunlight hours provides motivation for future work focused on the creation of a bioindicator of environmental correlation. However, from an integral point of view regarding the ecology of these wetlands, the observed correlations must be analyzed with caution, since the non-linear interactions of the measured variables have not been considered. For this, it would be ideal to analyze information from at least two complete annual cycles to capture interannual environmental variability. In this way, we could study the effects of environmental variables, reducing the uncertainty associated with the use of data from a particular year with certain environmental conditions.

5. Conclusions

The visualization provided by the UMAP algorithm and the zonal and local heatmaps account for the implementation and application of the proposed method, which, through passive recordings, permits monitoring the diversity, activity and acoustic phenology of structural species of a community of birds throughout the annual cycle, as well as detecting the presence of cryptic species that often go unnoticed.

Future efforts should continue with the calibration and scaling of the CNN in terms of repertoires, species and acoustically active taxa. In addition, new registration campaigns must be designed with greater coverage of the network of urban and peri-urban wetlands. In this way, a line of bioacoustic research based on a permanent monitoring program would be promoted, which would address the urgent need to preserve the integrity of both species and wetlands, as well as soundscapes, in a place that has been little studied and is a priority for global conservation, such as the Valdivian ecoregion (Mittermeier et al., 2011; Myers et al., 2000).

Moreover, the proposed monitoring method, being based on passive recordings, facilitates the study of bird communities in remote areas and/or areas subject to different conflicts that prevent other types of monitoring.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Victor Poblete reports financial support and article publishing charges were provided by Austral University of Chile - Campus Isla Teja. Victor Poblete reports a relationship with Austral University of Chile - Campus Isla Teja that includes: employment.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank Sebastián Arriagada for the diagramming of the map of the monitoring areas (Fig. 1), to Jorge Ruiz for his contributions in the auditory identification of the analyzed species, the students of the Audio Mining Laboratory (AuMiLab) for the valuable discussions around deep learning tools and the scholarship from the National Agency for Research and Development of Chile ANID awarded to Gabriel Morales to complete his postgraduate program.

The research that led to this article was funded by the Chilean National Fund for Scientific and Technological Development (FONDECYT) under grants 1190722 and 1220320.

References

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

Albawi, S., Mohammed, T.A., Al-Zawi, S., 2017. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), 1–6. IEEE. https://doi.org/10.1109/ICEngTechnol.2017.8308186.

Aloysius, N., Geetha, M., 2017. A review on deep convolutional neural networks. In: 2017 International Conference on Communication and Signal Processing (ICCSP). IEEE, pp. 0588–0592. https://doi.org/10.1109/ICCSP.2017.8286426.

Benitez Saldivar, M.J., Massoni, V., 2018. Song structure and syllable and song repertoires of the saffron finch (Sicalis flaveola pelzelni) breeding in Argentinean pampas. Bioacoustics 27, 327–340. https://doi.org/10.1080/09524622.2017.1344932.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120. https://doi.org/10.1109/TASSP.1979.1163209.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2021. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13901.

Bravo, S.P., Cueto, V.R., Gorosito, C.A., 2017. Migratory timing, rate, routes and wintering areas of White-crested Elaenia (Elaenia albiceps chilensis), a key seed disperser for Patagonian forest regeneration. PLoS One 12, e0170188. https://doi.org/10.1371/journal.pone.0170188.

Butt, N., Chauvenet, A.L., Adams, V.M., Beger, M., Gallagher, R.V., Shanahan, D.F., Ward, M., Watson, J.E., Possingham, H.P., 2021. Importance of species translocations under rapid climate change. Conserv. Biol. 35, 775–783. https://doi.org/10.1111/cobi.13643.

Buxton, R.T., Brown, E., Sharman, L., Gabriele, C.M., McKenna, M.F., 2016. Using bioacoustics to examine shifts in songbird phenology. Ecol. Evol. 6, 4697–4710 doi: 10.1002/ece3.2242, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.2242.

Chen, H., Xie, W., Vedaldi, A., Zisserman, A., 2020. Vggsound: a large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 721–725. https://doi.org/10.48550/arXiv.2004.14368.

Chepushtanova, S., Farnell, E., Kehoe, E., Kirby, M., Kvinge, H., 2020. Dimensionality reduction. In: Data Science for Mathematicians. Chapman and Hall/CRC, pp. 291–337.

Clink, D.J., Klinck, H., 2021. Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. Methods Ecol. Evol. 12, 328–341. https://doi.org/10.1111/2041-210X.13520.

Correa, H., Blanco-Wells, G., Barrena, J., Tacón, A., 2018. Self-organizing processes in urban green commons. The case of the Angachilla wetland, Valdivia-Chile. Int. J. Commons 12, 573–595. https://doi.org/10.18352/ijc.856.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 357–366 https://doi.org/10.1109/TASSP.1980.1163420.

Demarée, G.R., 2011. From "periodical observations" to "Anthochronology" and "Phenology" –the scientific debate between Adolphe Quetelet and Charles Morren on the origin of the word "phenology". Int. J. Biometeorol. 55, 753–761. https://doi.org/10.1007/s00484-011-0442-5.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

eBird, 2022. eBird: An Online Database of Bird Distribution and Abundance [Web Application]. Cornell Lab of Ornithology, Ithaca, New York. Available: http://www.ebird.org. Accessed: 2022-01-09.

Farina, A., 2014. Soundscape and landscape ecology. In: Soundscape ecology. Springer, pp. 1–28. https://doi.org/10.1007/978-94-007-7374-5_1.

Farina, A., 2018. Ecoacoustics: a quantitative approach to investigate the ecological role of environmental sounds. Mathematics 7, 21.

Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio set: an ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 776–780. https://doi.org/10.1109/ICASSP.2017.7952261.

Glotin, H., Ricard, J., Balestriero, R., 2017. Fast chirplet transform to enhance cnn machine listening - validation on animal calls and speech. arXiv. Doi: 10.48550/arXiv.1611.08749, arXiv:1611.08749.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. Birdnet: a deep learning solution for avian diversity monitoring. Ecol. Inform. 61, 101236 https://doi.org/10.1016/j.ecoinf.2021.101236.

Kingma, D.P., Ba, J., . Adam: A Method for Stochastic Optimization. arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980.

Kowarski, K.A., Moors-Murphy, H., 2021. A review of big data analysis methods for baleen whale passive acoustic monitoring. Marine Mammal Sci. 37, 652–673. https://doi.org/10.1111/mms.12758.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. 25, 1097–1105.

Lasseck, M., 2019. Bird species identification in soundscapes. CLEF (Working Notes) 2380.

LeBien, J., Zhong, M., Campos-Cerqueira, M., Velev, J.P., Dodhia, R., Ferres, J.L., Aide, T. M., 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. Ecol. Inform. 59, 101113 https://doi.org/10.1016/j.ecoinf.2020.101113.

LeCun, Y., Bengio, Y., 1995. Convolutional Networks for Images, Speech and Time Series. The MIT Press, pp. 255–258.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Luther, D., 2009. The influence of the acoustic community on songs of birds in a neotropical rain forest. Behav. Ecol. 20, 864–871. https://doi.org/10.1093/beheco/arp074.

Mann, S., Haykin, S., 1991. The chirplet transform: A generalization of Gabor's logon transform. In: Vision Interface '91. Canadian Image Processing and Patern Recognition Society, Citeseer, pp. 205–212.

Marler, P., 2004. Bird calls: their potential for behavioral neurobiology. Ann. N. Y. Acad. Sci. 1016, 31–44. https://doi.org/10.1196/annals.1298.034.

McGehee, S.M., Rozzi, R., Anderson, C., Ippi, S., Vásquez, R., Woodland, S., 2004. Presencia de la viudita, *Colorhamphus parvirostris* (Darwin) a fines de verano en Isla Navarino, comuna Cabo de Hornos, Chile., in: Anales del Instituto de la Patagonia, pp. 25–33.

McInnes, L., Healy, J., Melville, J., . UMAP: Uniform Manifold Approximation and Projection for dimension reduction. arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426.

Miller, E.H., Areta, J.I., Jaramillo, A., Imberti, S., Matus, R., 2020. Snipe taxonomy based on vocal and non-vocal sound displays: the south american snipe is two species. IBIS 162, 968–990. https://doi.org/10.1111/ibi.12795.

Mittermeier, R.A., Turner, W.R., Larsen, F.W., Brooks, T.M., Gascon, C., 2011. Global biodiversity conservation: the critical role of hotspots. In: Biodiversity Hotspots. Springer, pp. 3–22. https://doi.org/10.1007/978-3-642-20992-5_1.

Monczak, A., McKinney, B., Mueller, C., Montie, E.W., 2020. What's all that racket! Soundscapes, phenology, and biodiversity in estuaries. PLoS One 15, e0236874. https://doi.org/10.1371/journal.pone.0236874.

Moore, J.M., Székely, T., Büki, J., DeVoogd, T.J., 2011. Motor pathway convergence predicts syllable repertoire size in oscine birds. Proc. Natl. Acad. Sci. 108, 16440–16445. https://doi.org/10.1073/pnas.1102077108.

Morfi, V., Nolasco, I., Lostanlen, V., Singh, S., Strandburg-Peshkin, A., Gill, L., Pamuła, H., Benvent, D., Stowell, D., 2021. Few-shot bioacoustic event detection: a new task at the dcase 2021 challenge. In: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), pp. 145–149.

Morgan, M., Braasch, J., 2021. Long-term deep learning-facilitated environmental acoustic monitoring in the capital region of New York State. Ecol. Inform. 61, 101242 https://doi.org/10.1016/j.ecoinf.2021.101242.

Morrison, C.A., Auniņš, A., Benkő, Z., et al., 2021. Bird population declines and species turnover are changing the acoustic properties of spring soundscapes. Nat. Commun. 12, 6217. https://doi.org/10.1038/s41467-021-26488-1.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., Da Fonseca, G.A., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858. https://doi.org/10.1038/35002501.

Nuñez, J.J., Suárez-Villota, E.Y., Quercia, C.A., Olivares, A.P., Sites Jr., J.W., 2020. Phylogeographic analysis and species distribution modelling of the wood frog *Batrachyla leptopus* (Batrachylidae) reveal interglacial diversification in south western Patagonia. PeerJ 8, e9980. https://doi.org/10.7717/peerj.9980.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing, Second ed. Prentice-hall Englewood Cliffs.

Otondo, F., Poblete, V., 2020. Using a sonic time-lapse method as a compositional tool. Org. Sound 25, 198–204. https://doi.org/10.1017/S1355771820000102.

Otondo, F., Rabello-Mestre, A., 2021. The soundlapse project: exploring spatiotemporal features of wetland soundscapes. Leonardo 1–9. https://doi.org/10.1162/leon_a_02142.

Otter, K.A., Mckenna, A., LaZerte, S.E., Ramsay, S.M., 2020. Continent-wide shifts in song dialects of white-throated sparrows. Curr. Biol. 30, 3231–3235. https://doi.org/10.1016/j.cub.2020.05.084.

Poblete, V., Espejo, D., Vargas, V., Otondo, F., Huijse, P., 2021. Characterization of sonic events present in natural-urban hybrid habitats using UMAP and SEDnet: the case of the urban wetlands. Appl. Sci. 11, 8175. https://doi.org/10.3390/app11178175.

Ruiz, J., Biscarra, G., Flores, M., Morales, G., Tomasevic, J.A., Otondo, F., Poblete, V., Navedo, J.G., 2022. Dot-Winged Crake (*Porzana spiloptera*) Durnford, 1877 (Rallidae) in Chile: new records and a review of the status of Pacific populations. Ornitol. Neotrop. (In press).

Sainburg, T., Thielk, M., Gentner, T., 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLoS Comput. Biol. 16 (10), e1008228 https://doi.org/10.1371/journal.pcbi.1008228.

Saino, N., Ambrosini, R., Rubolini, D., von Hardenberg, J., Provenzale, A., Hueppop, K., Hueppop, O., Lehikoinen, A., Lehikoinen, E., Rainio, K., Romano, M., Sokolov, L., 2011. Climate warming, ecological mismatch at arrival and population decline in migratory birds. Proc. Royal Soc. B. Biol. Sci. 278, 835–842. https://doi.org/10.1098/rspb.2010.1778.

Slater, P., 2003. Fifty years of bird song research: a case study in animal behaviour. In: Essays in Animal Behaviour: Celebrating 50 Years of Animal Behaviour, pp. 633–639.

Stowell, D., 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10, e13152. https://doi.org/10.7717/peerj.13152.

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

Sueur, J., Krause, B., Farina, A., 2019. Climate change is breaking earth's beat. Trends Ecol. Evol. 34, 971–973. https://doi.org/10.1016/j.tree.2019.07.014.

Sugai, L.S.M., Silva, T.S.F., Ribeiro Jr., J.W., Llusia, D., 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69, 15–25. https://doi.org/10.1093/biosci/biy147.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Springer, pp. 270–279. https://doi.org/10.1007/978-3-030-01424-7_27.

Tittensor, D.P., Beger, M., Boerder, K., Boyce, D.G., Cavanagh, R.D., Cosandey-Godin, A., Crespo, G.O., Dunn, D.C., Ghiffary, W., Grant, S.M., et al., 2019. Integrating climate adaptation and biodiversity conservation in the global ocean. Science. Advances 5, eaay9969. https://doi.org/10.1126/sciadv.aay9969.

Toews, D.P.L., 2017. From song dialects to speciation in white-crowned sparrows. Mol. Ecol. 26, 2842–2844. https://doi.org/10.1111/mec.14104.

Wang, D., Forstmeier, W., Farine, D., Maldonado-Chaparro, A.A., Martin, K., Pei, Y., Alarcón-Nieto, G., Klarevas-Irby, J.A., Ma, S., Aplin, L.M., Kempenaers, B., 2021.

Machine learning reveals cryptic dialects that guide mate choice in a songbird. bioRxiv. https://doi.org/10.1101/2021.02.08.430277.
Wiley, R., 1991. Associations of song properties with habitats for territorial oscine birds of eastern North-America. Am. Nat. 138, 973–993. https://doi.org/10.1086/285263.
Wood, C.M., Kahl, S., Chaon, P., Peery, M.Z., Klinck, H., 2021. Survey coverage, recording duration and community composition affect observed species richness in passive acoustic surveys. Methods Ecol. Evol. 12, 885–896. https://doi.org/10.1111/2041-210X.13571.
Wood, C.M., Kahl, S., Rahaman, A., Klinck, H., 2022. The machine learning–powered birdnet app reduces barriers to global bird research by enabling citizen science participation. PLoS Biol. 20, e3001670 https://doi.org/10.1371/journal.pbio.3001670.
Wu, W., Thompson, J.A., Bertram, R., Johnson, F., 2008. A statistical method for quantifying songbird phonology and syntax. J. Neurosci. Methods 174, 147–154. https://doi.org/10.1016/j.jneumeth.2008.06.033.
Xie, J.J., Ding, C.Q., Li, W.B., Cai, C.H., . Audio-Only Bird Species Automated Identification Method with Limited Training Data Based on Multi-Channel Deep Convolutional Neural Networks. arXiv:1803.01107. https://doi.org/10.48550/arXiv.1803.01107.
Yang, W., Chang, W., Song, Z., Zhang, Y., Wang, X., 2021. Transfer learning for denoising the echolocation clicks of finless porpoise (Neophocaena phocaenoides sunameri) using deep convolutional autoencoders. J. Acoust. Soc. Am. 150, 1243–1250. https://doi.org/10.1121/10.0005887.
Zhong, M., Taylor, R., Bates, N., Christey, D., Basnet, H., Flippin, J., Palkovitz, S., Dodhia, R., Ferres, J.L., 2021. Acoustic detection of regionally rare bird species through deep convolutional neural networks. Ecol. Inform. 101333 https://doi.org/10.1016/j.ecoinf.2021.101333.