Ecological Informatics 55 (2020) 101023

Contents lists available at ScienceDirect

Ecological Informatics

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

Detection and identification of European woodpeckers with deep convolutional neural networks

Juliette Florentin[a,*], Thierry Dutoit[b], Olivier Verlinden[a]

[a] Theoretical Mechanics, Dynamics and Vibrations, Université de Mons, Place du Parc 20, Mons, Belgium
[b] Information, Signal and Artificial Intelligence, Université de Mons, Place du Parc 20, Mons, Belgium

ARTICLE INFO

Keywords:
Bird call detection
Bird sound classification
Deep convolutional neural networks
Drumming
Ecoacoustics
Woodpecker calls
Woodpeckers

ABSTRACT

Every spring, European forest soundscapes fill up with the drums and calls of woodpeckers as they draw territories and pair up. Each drum or call is species-specific and easily picked up by a trained ear. In this study, we worked toward automating this process and thus toward making the continuous acoustic monitoring of woodpeckers practical. We recorded from March to May successively in Belgium, Luxemburg and France, collecting hundreds of gigabytes of data. We shed 50–80% of these recordings using the Acoustic Complexity Index (ACI). Then, for both the detection of the target signals in the audio stream and the identification of the different species, we implemented transfer learning from computer vision to audio analysis. This meant transforming sounds into images via spectrograms and retraining legacy deep image networks that have been made public (e.g. Inception) to work with such data. The visual patterns produced by drums (vertical lines) and call syllables (hats, straight lines, waves, etc.) in spectrograms are characteristic and allow an identification of the signals. We retrained using data from Xeno-Canto, Tierstimmen and a private collection. In the subsequent analysis of the field recordings, the repurposed networks gave outstanding results for the detection of drums (either 0.2–9.9% of false positives, or for the toughest dataset, a reduction from 28,601 images to 1000 images left for manual review) and for the detection and identification of calls (73.5–100.0% accuracy; in the toughest case, dataset reduction from 643,901 images to 14,667 images). However, they performed less well for the identification of drums than a simpler method using handcrafted features and the k-Nearest Neighbor (k-NN) classifier. The species character in drums does not lie in shapes but in temporal patterns: speed, acceleration, number of strikes and duration of the drums. These features are secondary information in spectrograms, and the image networks that have learned invariance toward object size tend to disregard them. At locations where they drummed abundantly, the accuracy was 83.0% for *Picus canus* (93.1% for k-NN) and 36.1% for *Dryocopus martius* (81.5% for k-NN). For the three field locations we produced time lines of the encountered woodpecker activity (6 species, 11 signals).

1. Introduction

Acoustic monitoring is now a front-row tool to study bird populations (Blumstein et al., 2011; Sueur and Farina, 2015). Audio scene recordings accumulate to terabytes and call for efficient algorithms to fulfill two critical functions: 1) detecting bird sounds in audio streams and 2) identifying the species emitting these sounds. Recently, major performance gains were obtained in these two tasks by using deep convolutional neural networks. In the 2017 Bird Audio Detection challenge (BAD) (Stowell et al., 2019), most participants used Deep Convolutional Neural Networks (DCNN) to separate bird calls from other noises. Adavanne et al. (2017) obtained 16% of false positives, 8% of false negatives, and estimated that 42% of all false identifications had incorrect labels. Pellegrini (2017) obtained 13–22% of false positives and 4–8% of false negatives (accuracy 88.3–90.7%). DCNNs also dominated the 2018 BirdCLEF competition (Joly et al., 2018).

By design, neural networks work from a raw signal, build up their own features in their lower layers and then classify in their upper layers. An early network such as the one used by Fox et al. (2008), which used 12–15 Mel-Frequency Cepstral Coefficients (MFCC) as inputs and had 4 layers, only deployed the classification capacity. Grill and Schlüter (2017) upgraded to using full spectrograms as inputs and a seven layer network (4 convolutional layers, 3 dense layers) to deliver the strongest performance in the BAD challenge. Salamon et al. (2017) used a similar solution (spectrograms as inputs, 3 convolutional layers, 2 dense layers) to classify flight calls. In effect, these works replaced

* Corresponding author.
E-mail addresses: juliette.florentin.be@gmail.com, juliette.florentin@umons.ac (J. Florentin).

https://doi.org/10.1016/j.ecoinf.2019.101023
Received 22 August 2019; Received in revised form 21 October 2019; Accepted 22 October 2019
Available online 11 November 2019

traditional acoustic features (e.g. the MFCC) by a visual representation of sound (spectrograms), which then allowed the use of image analysis tools (convolutional networks). Spectrograms are compact and comprehensive representations of sounds, unlike the MFCC that strongly synthesize the acoustic signal and furthermore require that the time dimension is averaged or represented through various statistics. Most importantly, in the spectrograms, bird vocalizations produce visual patterns from which species are often readily identifiable. This is the basis for the switch to image analysis. Convolutional layers extend the concept of spectrogram cross-correlation by convoluting the input image with a large number of small template patterns ("filters"). The training of a convolutional network consists in optimizing these filters, as well as the weights in the dense upper layers where the classification is done. The depth of a network, i.e. the number of layers, embodies its analytical power. The first convolutional layers detect visual patterns, the next ones study how they are arranged with respect to each other.

Work on the ImageNet database of 1.28 million images has led to deeper networks: AlexNet has 8 layers (Krizhevsky et al., 2012), VGG has 19 (Simonyan and Zisserman, 2014), Inception has 22 (Szegedy et al., 2015), ResNet up to 152 (He et al., 2016) and DenseNet up to 264 (Huang et al., 2017). Most of the layers are convolutional layers, e.g. DenseNet has only one classification layer. Because of their depth and of the wealth of data on which they were trained, these legacy networks are powerful image analysis tools. They are also all publicly available as Pytorch implementations.[1] For a multitude of problems, instead of training a model from scratch to analyze specific images, video, spectrograms, etc., it is now more efficient to restart from the legacy models, who know how to decipher images, and to further their training to have them learn the current problem specifically. This means replacing the final 1000 classes of the ImageNet problem by the current targets in the last dense layer. The retraining mostly addresses these final connections; the construction of features in the convolutional layers is unchanged, whereas the classification using these features is revised. This approach is particularly convenient for ecoacousticians, as that they do not have training sets available that could compare in size to ImageNet. In the past few years, such transfer learning has become widespread in machine learning works (Laraba et al., 2017) and as Joly et al. (2018) report, in ecoacoustics as well. Sevilla and Glotin (2017) used Inception v4. Lasseck (2018) obtained his best results with Inception v3 but tried others including ResNet 152 and DenseNet.

Because DCNNs demand vast amounts of labeled data, data augmentation is their key companion. This technique consists in artificially increasing the size of the training set by modifying the original images in a way that does not compromise their meaning. For example, an image of a cat flipped left-right still represents a cat. Inception in its original training used 144 transformations per image, which combined the use of subparts of images ("crops") with rotations, flips, enlargements, etc. Working on bird calls, Lasseck (2018) appended segments of background noise to the images, appended segments from other recordings of the same class, skewed or stretched the images in time and in frequency, applied a cyclic time shift, randomly dropped time intervals, etc. The addition of segments of background noise was the most successful action, followed by deformed spectrograms and incomplete spectrograms with time intervals missing. We note that some image transformations such as rotations and flips are not permissible for sounds; they alter the meaning of the image.

In the BirdClef competition, the soundscape analysis task ultimately yielded 80–85% of false identifications on average (Joly et al., 2018; Lasseck, 2018). This shows that despite their formidable potential, further investigations are required to better understand the limitations of deep networks in addressing bird vocalizations. The contribution of the present work comes in the form of an in-depth look at species detection and identification for a manageable subset of species, the European woodpeckers. Particular attention is devoted to the design choices and to the mechanisms that condition the performance of DCNNs. By restricting the number of species, we are able to connect the shortcomings of DCNNs with the peculiarities of woodpecker sounds, and gain insight that resonates beyond the woodpecker case. This being said, woodpecker monitoring is an end in itself (Mikusiński and Angelstam, 1998) and the phylogenetics and sounds of woodpeckers remain popular research topics (Florentin et al., 2017; Fuchs and Pons, 2015; Miles et al., 2018). In that regard, the second contribution of our work is that it treats the European woodpecker problem from end to end, i.e. from the recording campaigns in the wild to the identification of species in the audio archives.

Woodpeckers have simple, innate calls, but their most famous acoustic signal is their drumming on trees. Few works have specifically targeted the identification of woodpecker sounds in the past. Swiston and Mennill (2009) searched for the double-knocks of two species of woodpeckers (*Campephilus guatemalensis* and *Campephilus principalis*) using spectrogram cross-correlation. Respectively 24% and 8% of double-knocks were detected, with respectively 97.0% and 98.5% of false positives. Because knocks are a simple and nondescript acoustic signal, they were confused with rain, wind, microphone static and with the calls of *Momotus coeruliceps*, which bear little resemblance but share the same frequency range. The proportion of false positives was a strong inconvenience: the results had to be reviewed by a human in a time-consuming process. Florentin et al. (2016) classified the drums of the European species. The accuracy for the classes with sufficient data available was in the range 64.4–90.0% (full validation set: 87.2%). This was achieved using handcrafted acoustic features and the simple k-Nearest Neighbor (k-NN) classifier. Indeed, the design space for drumming is so restricted that its analysis does not warrant using complex algorithms. Drumming is foremost a time signal and as it will turn out, some of its characteristics are not necessarily well rendered on spectrograms. The calls are better candidates for DCNNs.

In the present paper, we will discuss three problems: the detection of drums, the identification of drums, and the combined detection and identification of calls. Finally, we will apply the developed techniques to fully analyze 3 years of field recordings. The paper is organized as follows: in Section 2 (Materials), we introduce the sounds of European woodpeckers that we intend to classify and our audio collections; in Section 3 (Methods), we describe our various processing steps (audio selection and segmentation), a few reference methods used in comparisons, and our implementation of DCNNs; Section 3.3 in particular details how the three different problems were addressed with different images; in Section 4 (Results), we present the outcome of our calculations and finally in Section 5 (Discussion), we comment on the performance of DCNNs and on the qualities of woodpecker sounds that complicate their identification.

2. Materials

2.1. Woodpecker sounds

European woodpeckers use a variety of acoustical signals, some rare, some frequent, used alone or in combinations (Blume, 1996; Blume and Tiefenbach, 1997; Gorman, 2014; Winkler and Short, 1978). The loud ones that travel long distances to claim a territory or to attract a mate are the easiest to detect. Depending on the species, these functions are filled by drums, calls or both. For this reason, monitoring woodpeckers is a two-sided problem: on the one hand the calls, on the other hand the drums. In both contexts of territorial declaration and reproduction, the species information has to be conveyed to the other party; a corollary is that it should be possible to decode the species from these signals.

In the present work, we considered the drums of 10 species and nine calls from seven species (Table 1). In Table 1, the 11 European

[1] https://pytorch.org/

Table 1
Woodpecker signals considered in the present work.

Species	Drums	Calls
Drummers		
Picoides tridactylus	✓	–
Dendrocopos syriacus	✓	–
Dendrocopos major	✓	–
Dendrocopos leucotos	✓	–
Versatile		
Dryobates minor[a]	✓	Rattle
Dryocopus martius	✓	Rattle Flight Contact
Picus canus	✓	Rattle
Vocal		
Jynx torquilla	✓	Rattle
Dendrocoptes medius[a]	–	Kweek
Picus viridis	✓	Rattle
Picus sharpei[b]	✓	Rattle

✓ or text: considered; –: not considered
[a] Formerly *Dendrocopos*, renamed following Fuchs and Pons (2015).
[b] Formerly a subspecies of *P. viridis*, renamed following Perktas et al. (2011).

woodpecker species are listed under three groups: the species that primarily use drumming for long-distance advertising, the versatile species that both drum and possess an advertising call, and finally the vocal species, which do not produce loud territorial drums. These last four species, *Jynx torquilla*, *Dendrocoptes medius*, *Picus viridis* and *Picus sharpei*, only drum in rare occasions. Their drums are soft drums, which have a function pertaining to pair communication at close range (Florentin et al., 2017).

Spectrogram bandwidth 0–3 kHz. Spectrogram frame 43 ms, except *D. major*, *D. syriacus*, *D. minor*: 21 ms. The species-specific time structures are visible on the 'time interval between strikes' versus 'elapsed time' plots. Recordings by Kyle Turner.

Recordings from Xeno-Canto. Bandwidth 0–12 kHz. The full *D. martius* rattle call in XC110355 has 64 syllables, which is unconventionally high. A common number would be 10–20 syllables (Gorman, 2014).

Drumming produces a distinctive succession of vertical lines on spectrograms. Each strike of the bill excites a range of frequencies and thus draws a line. Fig. 1 shows examples for the most frequent drummers. After propagation through the forest, the vertical lines are reduced to smaller batons. The time structure of the drums (acceleration, speed), the number of strikes and the drum duration are species-specific (Florentin et al., 2016; Zabka, 1980). Some species accelerate through the drum, others maintain a constant speed or slightly decelerate (Fig. 1). Acceleration and speed are constrained by morphology, while the number of strikes and thus the drum duration can be adapted to an extent for differentiation (Miles et al., 2018). The number of design parameters in drumming is too low for the different species to fully singularize (Stark et al., 1998) and without further context information, it can be problematic to distinguish the drums of *Dendrocopos major* and *Dendrocopos syriacus*, or *Dryobates minor* and *Picus canus* (Florentin et al., 2016). The soft drums of rare drummers are signals that did not develop to the point of decisively encoding species identity, the desired functions being picked up by advertising calls instead. In a narrow design space, soft drums introduce further confusion when found in recordings (Florentin et al., 2017).

The common woodpecker advertising call is a rattle call (Winkler and Short, 1978), which consists in a series of near-identical syllables. The kweek call of *D. medius* has a similar form but with longer syllables, and is more characteristic of the species than its rattle call in the

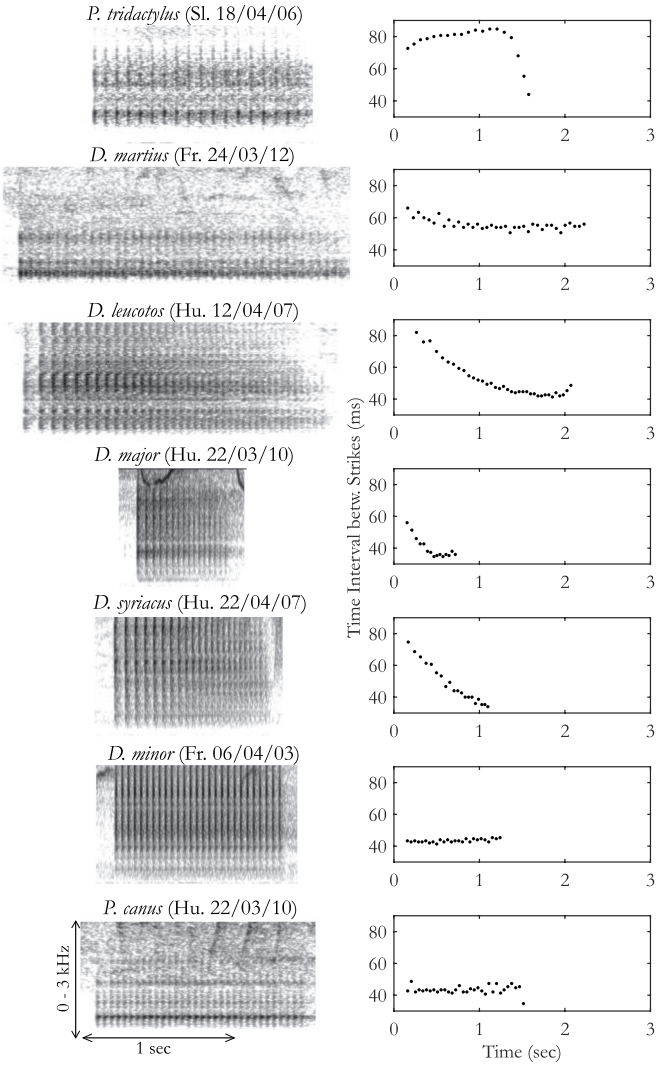

Fig. 1. Most frequent European woodpecker drums.

reproduction season. Because *Dryocopus martius* has a large territory, all its calls are far-carrying (Blume, 1996); in addition to its rattle, we included its iconic flight call (kru-kru-kru), which has a structure similar to the rattle, and the kleee, a long one-syllable wail. For Gorman (2014), the kleee is a territorial and contact call; for Blume (1996), it is an excitation, territorial and disturbance call. The species that primarily drum employ call notes (brief and high-pitched kik, chik, kyuk or kip) that are not as distinctive and as far-carrying as the rattles discussed above, and were thus not considered in this work. There are more calls that could have been included (the scolding keyak of *D. martius*, the rattle of *D. medius*) but data availability constrained our scope. Drums and rattle calls (or the kweek for *D. medius*) are the closest comparison in abundance and function to the songs of passerine birds. In every call listed in Table 1, the unit syllable is characteristic and therefore reveals the species. Each produces a given shape on spectrograms (Fig. 2). The call structure is also species-specific in some aspects, e.g. the variation in pitch or the number of syllables.

Drums all have spectral content below 1500 Hz (Florentin et al., 2016), and the fundamental note of calls typically lies within 1500–2500 Hz. For both types of signal the limiting taxon is *D. minor*. The main spectral peak for its drums can reach 2000 Hz (third quartile of the distribution in Florentin et al. (2016)) and for the calls, 2700 Hz.

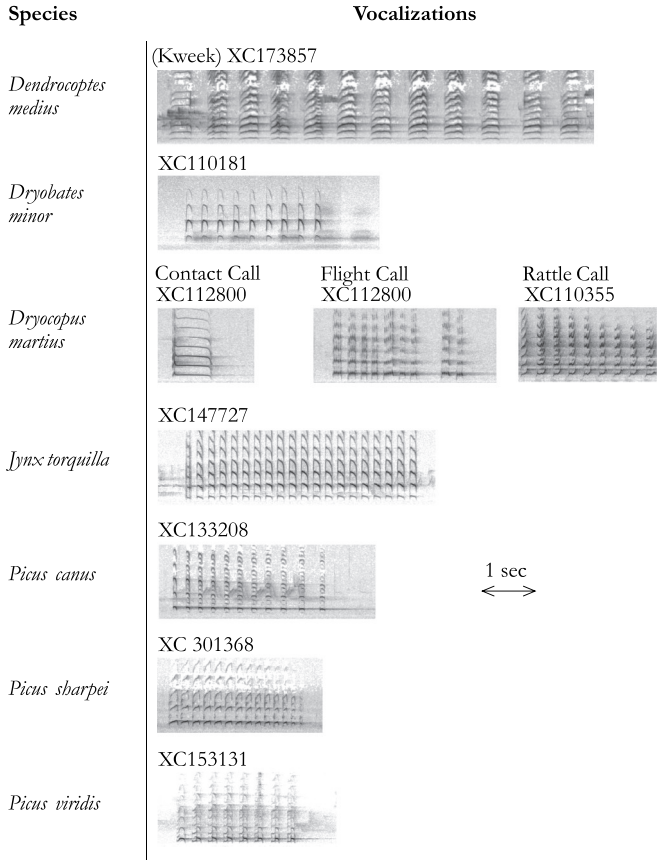

Fig. 2. Target vocalizations of European woodpeckers.

Table 2
Datasets from Xeno-Canto (XC)/Tierstimmen (TS)/K. Turner (KT).

Data type→		Drums		Calls
Source→				
Species↓		XC/TS	KT	XC/KT
D. leucotos		248	288	
D. major		818	589	
D. martius	Drums	84	388	
	Rattle			120
	Flight			89
	Contact			154
D. medius		8[a]		168
D. minor		832	528	171
D. syriacus		8	308	
J. torquilla			4	628
P. canus		104	130	208
P. sharpei			16	35
P. tridactylus		547	418	
P. viridis		16	100	263
Total		2665	2769	1836

[a] Dubious *D. medius* labels (Turner, 2011). These drums were eventually discarded.

2.2. Recordings

We assembled training sets from Xeno-Canto (XC),[2] Tierstimmen (TS)[3] and from the private collection of ornithologist Kyle Turner (KT). Table 2 describes the content of these datasets. The recordings are most often of a high quality and acquired at close range. The files last up to a few minutes, with a sampling frequency of either 44.1 kHz or 48 kHz. Being a collaborative archive, Xeno-Canto is not devoid of the odd labeling error, but the datasets compiled from this source are the most diverse in that they were gathered from multiple recordists. Kyle Turner, on the other hand, might have had access to fewer birds. His collection was recorded in Hungary, Slovakia, France, Spain and in the United Kingdom. The numbers in Table 2 reflect the abundance of the various signals in the wild, to an extent. *D. syriacus* is only present in Eastern Europe where fewer recordings are available. Recordings of the *P. sharpei* call are scarce. The high number of *J. torquilla* calls is due to XC177894 yielding 276 calls from a single pair.

We also acquired 44.1 kHz continuous recordings in woodpecker habitats using an autonomous station (Florentin and Verlinden, 2017). The first campaign (2016) took place in Tenneville (TN), Belgium, where a single *P. canus* had been spotted. In 2017, the station was deployed in the nature reserve at Remerschen (RM), Luxembourg, which is known to host 3–4 *P. canus* territories, including breeding pairs. The Remerschen wetlands are an important stop for migratory birds. In 2018, we installed the station in La Petite Raon (LPR), France, which is located in the Vosges mountains, i.e. the northernmost stronghold of *P. canus* in France (Sordello, 2012). Belgium is on the distribution edge of *P. canus*, which makes this species a local rarity (Schmitz and Dumoulin, 2004) of particular interest to us. The collected datasets are presented in Table 3. A schematic map with the positions of the station and the *P. canus* distribution area is shown in Fig. 3. *J. torquilla*, *D. minor*, *D. medius*, *D. major*, *D. martius*, *P. viridis* and *P. canus* are present in the region.

Table 3
Field datasets.

Location	Code	Months	Size	Mean SNR[a]	Habitat
Tenneville (BE)	TN	Mar./Apr.	8 GB 96 h	32.1 dB	Deciduous forest
Remerschen (Lux.)	RM	Apr./May	128 GB 435 h	17.7 dB	Wetlands
La Petite Raon (FR)	LPR		118 GB 397 h		Coniferous forest
	LPR1	March		24.2 dB	
	LPR2	April		22.1 dB	
	LPR3	May		26.1 dB	

[a] Evaluated on the detected calls.

2.3. Soft- and hardware for calculations

The image generation from audio was performed using MATLAB. Downloading models from the Pytorch libraries, retraining them and making new predictions was managed through a set of Python/Pytorch

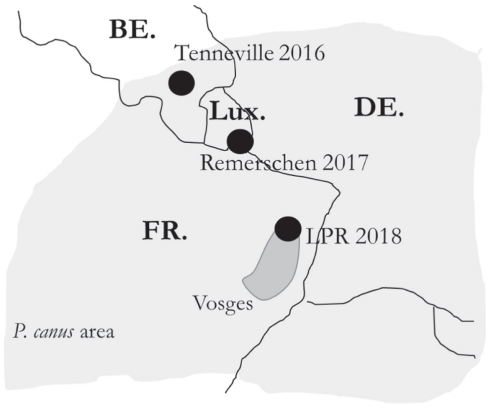

Fig. 3. Area map with station positions.

[2] The Xeno-Canto Foundation, https://www.xeno-canto.org/. The identification number for XC files in our figures is the one used by the website. A file with an additional trailing index is a segment from the original recording.
[3] Museum für Naturkunde Berlin, http://www.animalsoundarchive.org/.

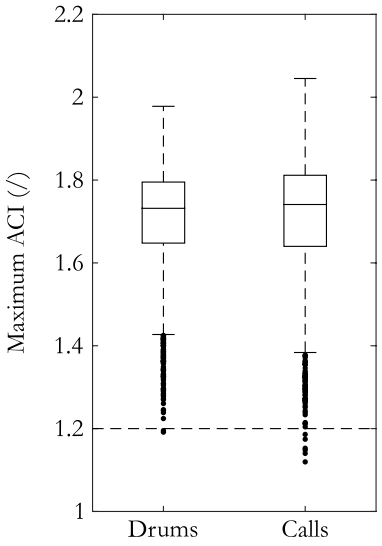

Fig. 4. Maximum ACI in XC/TS/KT drums and calls.

scripts, documented as supplementary material. We used Python 3.6.8, Pytorch 0.3.1 and CUDA 10.0. Calculations involving deep networks were performed using a Graphical Processing Unit (GPU), the NVIDIA GEFORCE GTX 1080 Ti (GP102 processor, 11 GB in RAM) and a SAMSUNG 850 EVO solid-state drive (SSD).

3. Methods

For the selection of candidate sounds (see Paragraph 3.1 below), we used frames of 46.4 ms (2048 bins, 44.1 kHz) with 50% overlap to generate spectrograms. For all subsequent analyses, we resampled the recordings to 12 kHz and used frames of 21.3 ms (256 bins, 12 kHz).

3.1. Selection of candidate sounds

The recording station continuously collected 30-s WAV files, that were examined on board using the Acoustic Complexity Index (ACI). The ACI, first developed by Pieretti et al. (2011), reacts to acoustic content with significant time variations, such as birdsong. Tests performed on the XC/TS/KT datasets revealed that almost all woodpecker drums and calls clear an ACI threshold of 1.2 (Fig. 4). The ACI is calculated from a spectrogram and per frequency band; what is shown in Fig. 4 is the maximum value of the ACI spectrum in the 500–1500 Hz bandwidth for drums and in the 500–2100 Hz bandwidth for calls. The 2100 Hz upper bound for calls is a restriction to minimize the interference from passerine songs. The 500 Hz lower bound removes most of the background noise. The 1.2 threshold value is dependent on the duration of the complete segment being processed (in our case, 30 s[4]) and on the frame duration used in the spectrogram calculation. Here, wide frames of 46.4 ms were used, in the spirit of a quick assessment in the field. Our setup is not always favorable to *D. minor*; the taxon has fast drums, 40 ms intervals between strikes being common, and high-pitched calls. The problem remains modest: 2 *D. minor* drums and 3 *D. minor* calls miss the cut in Fig. 4. For all species, the ACI does not offer much separation from some of the other sounds in the same bandwidth: other bird calls, rain drops, any sort of shock. It sets aside silence and sustained sound events (cars, planes) but the resulting datasets are far from specific. Still, 50–80% of the field audio was discarded in this manner.

The ACI is calculated in the interval 0.5–1.5 kHz (drums) and 0.5–2.1 kHz (calls) with 46.4 ms frames. The dashed line indicates the 1.2 ACI threshold. The distribution outliers are indicated with dots.

In a second step, the 30-s files selected above, along with all XC/TS/KT files, were segmented into unique sound events using median-based thresholding, a common method in ecoacoustics (Lasseck, 2015; Potamitis, 2014; Stowell and Plumbley, 2014). We imposed a minimum length on the output segments (0.4 s for drums, 1 s for calls), allowed gaps of up to 0.3 s to connect smaller segments together (interruptions and longer inter-syllable intervals occur) and included 0.15 s of sound before and after the target signal (early and late strikes or syllables might be significantly quieter). We used the 300–1500 Hz bandwidth for drums and the 1000–2700 Hz bandwidth for calls. Such targeted segmentation is not strictly necessary, but provides a considerable reduction of the datasets. For example, in the TN dataset, segmenting reduced 47 h of audio to 4.3 h (6760 extracts; mean duration 2.3 s; maximum duration 80.7 s, long files being caused by increased background noise from cars, planes, chain saws, etc.). Alternatively, one may chop up the audio in successive files of a given duration and post-process all, at a greater computational cost. However, smart segmentation reduces the number of false detections in that it limits the analysis to segments where there are potential target signals. Less noise passes the barriers.

3.2. Methods used in comparisons

We benchmarked DCNNs against competing methods for the detection and the identification of drums, not for the calls. The characteristics of woodpecker calls, namely the peculiar syllable sounds, are best captured by the visual description provided by spectrograms. In this case an image-based technique such as a convolutional neural network is the natural approach.

For the detection of drums we tested out two other methods: spectrogram cross-correlation and an analysis of repetitions in the signal. For cross-correlation, we used a single template image of a *P. canus* drum. This was experimented with on the TN dataset for which the station was located next to a *P. canus* drumming tree. The TN dataset is the most modest of the three (8 GB); we bypassed the segmentation step and we used the fast Matlab normxcorr2 function. We focused on content in the 300–1500 Hz bandwidth and initially detected instants at which the correlation exceeded a threshold. However correlation values do not offer much contrast; the method yields as many false positives as false negatives in this primitive form. We improved the performance by noticing that drums also generate a dip in correlation before and after the drum peak; we reoriented our code to search for such a pattern in the correlation time series.

The repetitions analysis is inspired by music-retrieval techniques and uses a similarity matrix and a beat curve (Foote et al., 2002; Lartillot and Toiviainen, 2007). Without getting into too many details, considering that this approach gives satisfaction only for the clearest signals, the method exploits the fact that the different strikes in a drum have a nearly-identical spectral content and are repeated at intervals in the range 40–90 ms in average over a drum. First, identical frames of signal are searched for (one strike is contained in one frame and appears in 1–3 frames with the overlap), then the method controls whether the succession of similar frames follows the desirable rhythm.

As mentioned earlier, the time parameters are critical in identifying drumming species. On all drums, the acceleration, speed, number of strikes and drum duration were evaluated. The acceleration and speed were derived from respectively the slope and the y-intercept of a line fit through the series of time intervals between strikes versus time (such series are presented in Fig. 1). With these simple parameters, the drums from the field campaigns were confronted to the reference datasets XC/TS/KT using the k-NN classifier. We used $k = 5$, which matches a sample to the most frequent class among the five nearest neighbors. This straightforward approach is well suited to the simplicity and to the temporal nature of drums (Florentin et al., 2016). The footnotes below Fig. 1 prefigure how spectrograms might not render the subtleties of

[4] Shorter files from the XC/TS/KT datasets were padded with silence.

Table 4
DCNN *re*-training: networks, classes, test set.

Problem	Networks	Classes	Samples	Test set
Drums detection	Inception v3	Not a Drum	5435	10%
	ResNet 34	A Drum	5198	
	ResNet 152	(Total)	(10633)	
	DenseNet 169			
Drums identification	Inception v3	*D. leucotos*	536	12%
	ResNet 34	*D. major*	1407	
	ResNet 152	*D. martius*	472	
	DenseNet 169	*D. minor*	1360	
		D. syriacus	316	
		P. canus	234	
		P. tridactylus	965	
		P. viridis	116	
		(Total)	(5406)	
Calls detection and identification	AlexNet	*D. martius*		6%
	VGG	Ad	543	
	Inception v3	Flight	307	
	ResNet 34	Contact	207	
	ResNet 152	*D. medius*	625	
	DenseNet 169	*D. minor*	451	
		J. torquilla	2429	
		P. canus	595	
		P. sharpei	117	
		P. viridis	799	
		Noise/other	6081	
		(Total)	(12154)	

drums faithfully. The time resolution in spectrograms is constrained by the frame size and overlap (i.e. 10.5 ms for 21 ms frames with 50% overlap). As the differentiation between some species hangs on a few milliseconds (Florentin et al., 2016; Zabka, 1980), we expect image-based methods to be challenged by the task of identifying drums. On the contrary, the detection of drums is more naturally suited to an image treatment, because drums produce a characteristic pattern in spectrograms (a series of vertical lines) that few other sounds mimic.

3.3. Deep convolutional neural networks

3.3.1. Deep networks retraining and use

We considered the following six network architectures for our analyses: AlexNet (8 layers), VGG (19 layers), Inception v3 (22 layers), ResNet (34 and 152 layers) and DenseNet (169 layers). All are milestone architectures that won the ImageNet competition in the last decade and are available from the Pytorch libraries. Having observed that the deeper networks identified the calls more accurately, we did not use AlexNet and VGG again for the drums. As all these networks already have a great command of image analysis, they only have the specific patterns of woodpecker sounds left to learn. For this, our modest datasets in Table 4 could possibly suffice.

The retraining targeted all layers, but mostly affected the last, diagnostic layer. We performed stochastic gradient descent with an adaptive learning rate, i.e. we started with 0.001 and divided it by 10 when the training loss had stalled for two epochs (i.e. we used a patience[5] of 2). We also used a momentum of 0.9 (Hinton et al., 2012) and stopped training if 60 epochs were reached. A percentage of the training images was set aside for validation (Table 4). When a single root recording had yielded several sounds through the segmentation step, all were included on the same side, training or test.

To classify the sounds from the recording station, we pooled the predictions of the different network architectures. The class with the most votes was retained (majority-voting). For the drums, the ensemble comprised all available models. For the calls, we tried a number of variations on the training parameters: learning rate, number of epochs, fixed versus adaptive learning rate and patience. We then picked the nine models for which the average class accuracy exceeded 91% and the overall accuracy exceeded 94% on the test set to populate the ensemble. In case of ties in the votes, we went with the diagnosis of ResNet 152 (drums) or DenseNet (calls), which had performed well in pre-trials. We observed on the drums detection case that the different networks were in perfect agreement for 98.7% of the images. For long sound files from which several images were derived (see below), a positive identification was granted if one of the images was positive.

The deep networks cited above all accept 224×224 images as inputs, except for Inception v3 which takes 299×299 images. Thus both dimensions of our images were rescaled, independently from each other, to fit this frame. For the color scale, we retained the top 30 dB in the images. This range was then normalized to [0,1] (max-min normalization) to issue JPEG images. The same spectrogram was triplicated to fill the red, green and blue (RGB) channels, except for the identification of drums, for reasons explained below. Computing mel-spectrograms was not considered.

3.3.2. Images for the detection of drums

For the drums detection, the "a drum" class in Table 4 comprises 4669 XC/TS/KT drums, and 529 drums extracted from the LPR dataset (LPR1/LPR2) using the repetitions analysis method, which we experimented with first. The "not a drum" class was populated with the false positives from repetitions analysis trials (KT/TN/LPR1/LPR2). It seemed on point to include the signals that had previously been mistaken for drums (rain, wind, cell-phone interference, fast series of chirps, wing flaps).

Spectrograms were generated in the 300–1500 Hz bandwidth and for the duration of the sounds, unless they exceeded 5 s. Then the sounds were split into successive images, with an overlap of 25%. The mean image width was 143 pixels (1.5 s) and the maximum width 469 pixels (5 s). Along the frequency dimension, there were 26 pixels. Approximately 10% of the images underwent a compression in width at the entrance of the networks. For the 5-s sounds, the compression factor was 2.1. Hence the resulting images do not have a common time scale, and the networks cannot sense whether a drum is fast or slow. Another issue with the time scale compression is the potential loss of fine details: drum strikes span 1–3 spectrogram frames, hence 1–3 pixels, and inter-strike intervals span 2–5 frames in the fastest drums.

3.3.3. Images for the identification of drums

For the identification of drums, the precision of the images along the time dimension is key. The longest drum in the XC/TS dataset has a duration of 3.3 s; using 224 pixels, a time step of 15 ms is possible. With such an error on the intervals between strikes, the different species cannot be distinguished. For comparison, in the calculation of the simple temporal parameters described earlier (acceleration, speed, etc.), we used a time step of 0.7 ms. Eventually, we opted to create images using 224 pixels per 1 s of data (a time step of 4.5 ms). With the three RGB channels, we were thus able to store 3 s of spectrogram at best. Beyond 3 s, the exact duration of the drum does not matter anymore: only *D. martius* produces such long drums. Fig. 5 shows a few examples (before rescaling); the short *D. major* drum uses only the blue channel, the long *D. martius* drum uses the three colors. With this approach, all drums were represented with the same time and frequency scales (300–3000 Hz). For this task the networks must unlearn that the same objects can come in different sizes in images; the size, notably of the time intervals between strikes, is a criterion for differentiation.

We included soft drums in the training set as they remain probable in the recordings, but did not consider *J. torquilla*, *P. sharpei* and *D. medius*, as we did not possess enough data to properly train the networks to recognize these species.

3.3.4. Images of calls

For the calls, we were only able to collect 1836 samples. For

[5] See the Pytorch documentation at https://pytorch.org/docs/stable/optim.html.

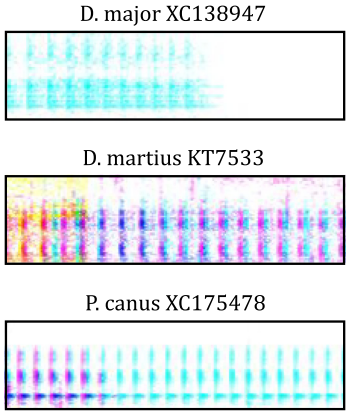

Fig. 5. Images of drums submitted to DCNN.

Table 5
Audio segments and images in field datasets.

Dataset	Audio segments	Images
TN	3732	13,051
LPR1	21,831	73,883
LPR2	30,072	98,450
LPR3	52,061	172,992
RM	150,894	643,901

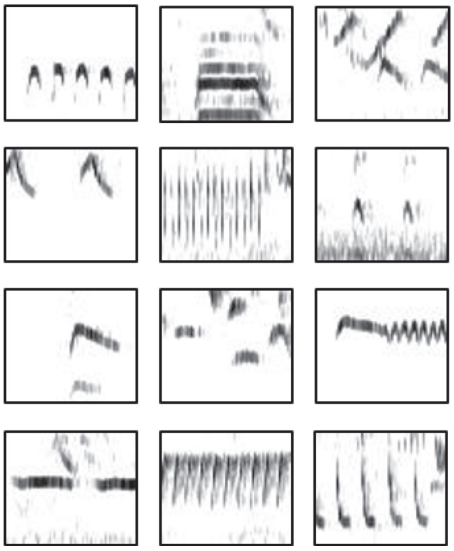

Fig. 6. Images of bird calls from audio recordings in the nature reserve of Remerschen, Luxemburg.

shallower networks, Salamon et al. (2017) had 5428 audio clips and Grill and Schlüter (2017) 16,000. Lasseck (2018) trained Inception v3 with 36,496 samples. We thus augmented our dataset by producing, for each call, several partial images focused on a few syllables. We settled for 54 × 63-pixel images (1000–3500 Hz × 1 s, using 21.3 ms frames with 25% overlap). This time and frequency resolution compares to the setup in Lasseck (2018), although he used larger images. In our solution, at the entrance of the deep networks, the images were enlarged by a factor 3.6–5.5. The calls were segmented with a 15% overlap between consecutive sub-images. Up to 10 images were retained per call, if needed selected among the ones where the signal was the loudest. More often, the calls were spread over 2–5 images. For the largest dataset, Remerschen, we generated 643,901 images (Table 5). Fig. 6 shows a few examples. The first one on the top left is *P. canus*.

The full training set comprised 12,154 images, half of which were "noise" (Table 4). Fig. 7 shows the images extracted from a *D. minor* and from a *J. torquilla* call. We see in the *J. torquilla* example that some of the call structure is captured in the images. The first and second images show the ascent in frequency before the syllable stabilizes in the last two images. Similarly, the decaying notes of *P. canus* were at times captured. In the *D. minor* example, the bird followed its call by drumming; the last image does not contain its voice at all and was labeled as noise. The noise class was entirely populated with similar rejects: passerine calls, other woodpecker calls not in our study, anthropogenic sounds or various instances of background noise. We did not have information to help us include more relevant examples. Gorman (2014) and Del Hoyo et al. (2002) mention a few similarities of woodpecker calls with raptors, but nothing exhaustive.

In essence, with the sub-images, we substituted the recognition of the syllables for the recognition of the calls. This was done by other authors as well. Potamitis (2014) classified syllables or elements of songs extracted from spectrograms. Lasseck (2018) tested extracting random audio chunks with some success. Brandes (2008) is an example of a carefully constructed classification in successive steps: a first Hidden Markov Model identified the syllables, then the song structure was modeled using a second one.

4. Results

4.1. Drums detection

The four networks that were trained to detect drums exhibit outstanding accuracy on the validation set (Table 6). Inception accepted bouts of demonstrative tapping and very short drums, when other networks did not. Short drums are strongly distorted when the small image is scaled up to 299 × 299 pixels. In comparison, half of our drumming samples are in a 0.7–1.5 s duration range and are scaled by comparable factors. Tapping is slower than drumming; the vertical lines are more spaced out. Hence Inception seems to have the greatest capacity to recognize a drum at a different scale. As we recall, during its original training by Szegedy et al. (2015), 144 crops were generated for each image. The images were first rescaled to different sizes, then transformed and cropped at a number of positions. In comparison, DenseNet used 10 crops and ResNet only one, taken from a randomly resized image. These other networks rejected demonstrative tapping and included fast-paced series of chirps: they seem to understand that there is an acceptable time interval between the vertical lines, and this despite the loss of a shared time scale between the images. Again, this might be favored by the fact that the durations of most drumming samples form a narrow distribution. The time interval between drum strikes is rescaled similarly for all.

The retrained networks were then used to detect drums in the three field datasets TN, RM and LPR. The comparison with the repetitions analysis and spectrogram cross-correlation is presented in Table 7 for 2 days of the TN dataset. The repetitions analysis missed a large number of drums. The deep networks extracted a similar number of drums as spectrogram cross-correlation, but the deep networks produced far fewer false positives (FP). Considering that the deep networks analyzed audio fragments that had been generated by our segmentation step and that cross-correlation worked on the full 30-s recordings, the comparable results between the two methods indicate that the segmentation does not produce false negatives in significant numbers.

In Table 8 (all datasets), the number of false positives for TN and LPR2 is marginal, helped of course by the inclusion of samples from these datasets in the retraining set. The effect is not as strong for LPR3 (still 10% of FP), recorded later in the season when new birds, that the networks did not know, had started to sing. In the RM dataset, the FP stand at 74.7%, yet the analysis still allowed discarding 94.7% of the initial dataset. The deep networks successfully detected drums although they are only marginally present in the datasets (in 4.0% of the

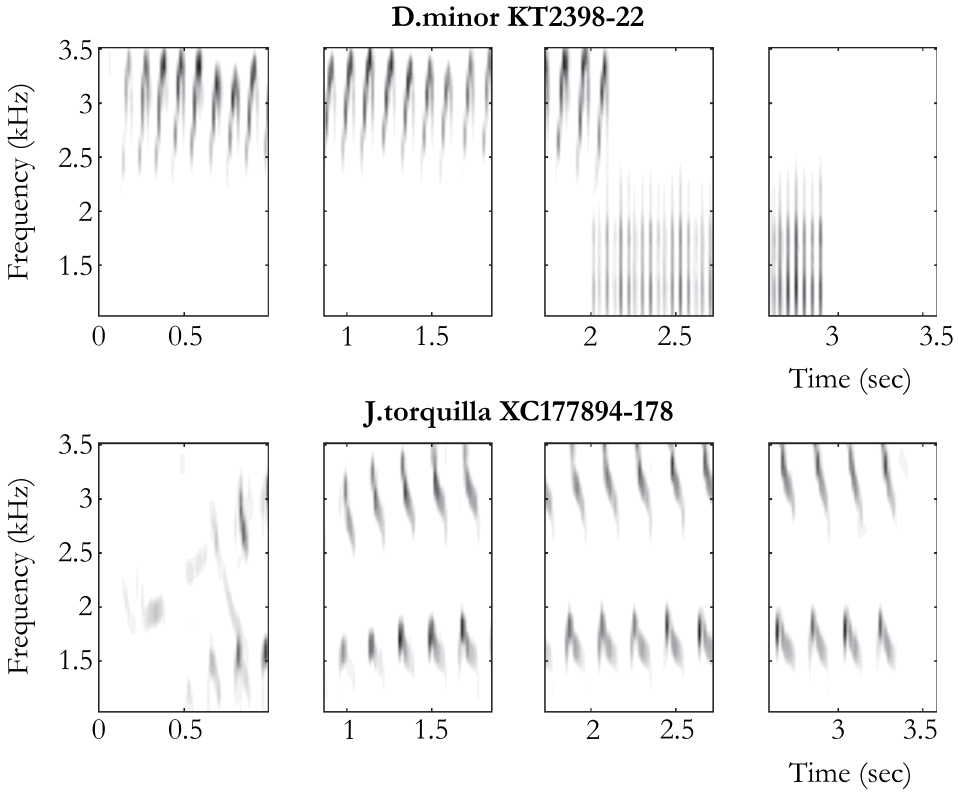

Fig. 7. Images extracted from a *D. minor* and from a *J. torquilla* call.

Table 6
Accuracy (%) of predictions on test datasets after retraining.

Deep network	Drums detection	Drums Id	Calls Det./Id.
AlexNet			94.4
VGG			92.9
Inception	98.59	94.8	94.8
ResNet 34	98.21	92.9	94.0
ResNet 152	98.50	90.8	94.2
DenseNet 169	98.40	93.7	95.4

Table 7
Comparison between drums detection methods.

Cohort	Repetitions analysis		Cross-correlation		Deep networks	
	TP[a]	FP[b]	TP	FP	TP	FP
TN 06/04	28	3.0%	103	55.0%	107[c]	0.0%
TN 13/04	175	4.0%	218	65.0%	210[d]	0.0%

[a] True positives.
[b] False positives.
[c] Add 4 drums in sound files with multiple drums.
[d] Add 6 drums in sound files with multiple drums.

Table 8
Drumming rolls detected in TN/RM/LPR.

Cohort	Sounds	Images	TP	FP
TN	6760	7875	2570	4 (0.2%)
LPR2 (Part 2)[a]	5619	5619	347	4 (1.1%)
LPR3	8933	10,862	237	26 (9.9%)
RM	20,866	28,601	278	822 (74.7%)

[a] LPR1 and LPR2 (Part 1) were processed through repetitions analysis.

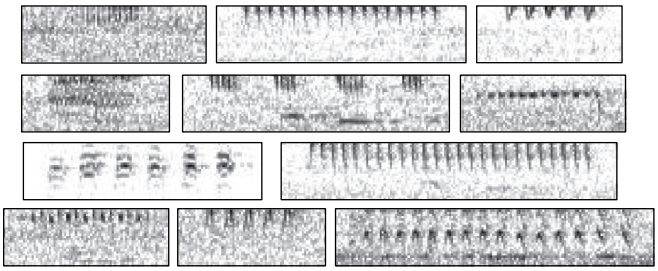

Fig. 8. Examples at RM of false positives from DCNN drums detection.

Table 9
Drums detection: false positives and false negatives.

Deep network	False positives				False negatives[a]			
	TN	LPR2	LPR3	RM	TN	LPR2	LPR3	RM
DenseNet	24	12	36	1263	9	3	7	15
Inception	34	16	70	1446	14	6	1	6
ResNet 34	3	8	25	846	26	8	4	15
ResNet 152	13	12	70	967	14	6	1	1
Pool	4	4	26	822	12	5	1	2

[a] Not all negative predictions were reviewed. This is an evaluation of false negatives that were positively identified by at least one network.

analyzed sounds for LPR, 1.3% for RM). Examples of false positives in Remerschen are shown in Fig. 8. They comprise a variety of bird calls. Among the negatives, only the sounds that were predicted positive by at least one of the networks but turned down by the model ensemble were reviewed. We found a few faint or distant drums that had been missed.

In Table 9, Inception and ResNet 152 missed the fewest drums, but Inception produced 50% more false positives. ResNet 34, the network with the simplest architecture, was the one that yielded the fewest false

Table 10
Drums identification: confusion matrices and accuracy.

Tenneville

Actual classes ↓	Predicted classes								Accuracy
	D.leu.	D.maj.	D.mart.	D.min.	D.syr.	P.can.	P.tri.	P.vir.	
k-NN with XC/TS training set									
D. major	0	11	0	3	0	0	0	0	78.6%
D. martius	2	0	76	18	0	0	24	0	63.3%
P. canus	1	3	2	1902	0	525	14	0	21.5%
								Average	54.5%
k-NN with KT training set									
D. major	0	12	0	1	0	0	0	0	85.7%
D. martius	0	0	108	0	0	0	12	0	90.0%
P. canus	0	1	10	151	0	2278	0	5	93.1%
								Average	89.6%
Ensemble of deep networks									
D. major	0	8	0	0	0	4	0	2	57.1%
D. martius	0	2	104	0	0	1	13	0	86.7%
P. canus	7	34	35	329	0	2032	1	9	83.0%
								Average	75.6%

Remerschen

Actual classes↓	Predicted classes								Accuracy
	D.maj.	D.mart.	D.min.	D.syr.	J.tor.	P.can.	P.tri.	P.vir.	
k-NN with KT training set									
D. major	204	0	5	48	0	0	0	0	79.4%
J. torquilla	0	2	1	0	0	0	2	1	0.0%
P. canus	0	0	0	0	0	3	0	0	100.0%
								Average	59.8%
Ensemble of deep networks									
D. major	232	4	14	5	0	1	1	0	90.3%
J. torquilla	1	3	0	0	0	1	1	0	0.0%
P. canus	0	0	3	0	0	0	0	0	0.0%
								Average	30.1%

La Petite Raon

Actual classes ↓	Predicted classes								Accuracy
	D.leu.	D.maj.	D.mart.	D.min.	D.syr.	P.can.	P.tri.	P.vir.	
k-NN with KT training set									
D. major	0	808	0	19	158	5	0	11	80.7%
D. martius	5	0	88	2	0	1	5	7	81.5%
P. canus	0	0	0	0	0	10	0	0	100.0%
								Average	87.4%
Ensemble of deep networks									
D. major	3	934	8	42	0	3	2	9	93.3%
D. martius	16	11	39	32	0	2	2	6	36.1%
P. canus	0	0	3	6	0	0	0	1	0.0%
								Average	43.1%

positives; seemingly, low analysis power is sufficient to reliably identify the simple patterns of drums. Networks with a greater depth can catch less obvious drums, but make more mistakes as they rely on tenuous details. The great flexibility of Inception toward the size of patterns made it a false prediction machine. It was not supposed to, but Inception also detected 34 out of the 52 occurrences of tapping in the RM dataset; the next "best" was ResNet 34 with 12. In any case, pooling together the four models improved the precision for all datasets.

LPR3, and particularly RM, are more complex cases because of location and month of recording. In March (TN, LPR1), the woodpeckers have the forest more or less to themselves. In May (LPR3, RM), the passerines have taken over and provide a variety of new sounds that confuse the algorithms. This is exacerbated in Remerschen where the avian community is remarkable.

4.2. Drums identification

The predictions on the test set are again very accurate (Table 6), 5–8% above the accuracy Florentin et al. (2016) obtained with k-NN and simple parameters. DenseNet is the best performing network, with accuracy per species ranging from 81.3% (the soft drums of *P. viridis*) to 100% (*P. canus*). Soft drums have a weaker species character (Florentin et al., 2017).

The results on the field datasets paint a more contrasted picture; see the confusion matrices and accuracies per class in Table 10 (only three species were heard drumming at each location). The performance of k-NN with two different training sets is shown for the TN dataset. The three species present at that location were *D. major*, *D. martius* and most of all *P. canus*. The accuracy per species is higher with the KT set as

reference. The XC/TS set confuses *P. canus* with *D. minor* and *D. martius* with other classes, because it does not contain enough samples of *P. canus* and *D. martius*. Also, because of the greater variety of sources, XC/TS exhibits larger parameter distributions, and supposedly a few labeling mistakes.

The deep networks perform less well than k-NN. Despite the 100% accuracy for *P. canus* on the test set, the *P. canus/D. minor* confusion is not sorted out. The best performance for this aspect is from DenseNet (predicted 2197 *P. canus* and 153 *D. minor* in TN; all should be *P. canus*) and the worst, Inception (962 *P. canus* and 1434 *D. minor*). None of the *P. canus* drums in RM and LPR are correctly identified. The accuracy for *D. major* improves in RM and LPR as the number of confusions with *D. syriacus* diminishes; however this is because the networks predicted the classes that they saw the most during training (1407 *D. major*, 316 *D. syriacus*). The poor results for *D. martius* comprise a number of unusual confusions. It is common for *D. martius* to be confused with *Dendrocopos leucotos* and *Picoides tridactylus* (Florentin et al., 2016) but not with *D. major* and *D. minor* which have much shorter drums. Given our training set, *D. major* and *D. minor* are three times more probable classes than *D. martius* to the networks. Finally, if we consider the wrong *P. viridis* identifications in LPR (18 for k-NN and 16 for the networks), only three drums are in both groups. The difficult drums are not the same for the two methods, although overall, the mispredicted drums were distant drums in both cases. Like k-NN, the networks struggle to differentiate a distant *D. martius* drum from a soft *P. viridis* drum. They also make blatant and inexplicable mistakes, e.g. a *D. major* drum at close range mistaken for *P. viridis*. The excellent performance on the test set (Table 6) does not mean that the networks identified discriminant features from the images, but that the test set is not different enough from the training set. Indeed, for example, XC98152 is in the test set, and XC98153 and XC98154 (different recordings, same bird) are in the training set.

As mentioned earlier, imbalance between the classes in the training set causes the networks to wrongly learn that some classes are more probable than others. On the contrary, k-NN preserves the ability to match test samples to smaller classes. Even with an overwhelming number of *D. minor* in the training set, it successfully assigned the *P. canus* class based on the few samples that were close to the candidate drums.

Overall, deep convolutional networks are not incompetent with time structures (the TN predictions are fair, for example) but k-NN produces more reliable, more physical results. The networks retain a strong advantage in terms of the simplicity of the process; the image generation is neither subtle nor long; training the networks requires 2 h at most and running the test samples a few minutes. In the k-NN method, the calculation of the drum parameters is tedious.

4.3. Detection and identification of calls

All networks excel on the validation set (Table 6). DenseNet delivers the top performance. The accuracy for the noise class is greater than 96% for all networks and promises few false positives. For the call classes, accuracies greater than 90% are routine. Only the *D. minor* result can be viewed as a shortcoming: the top accuracy for this class is 83.3% with DenseNet. This call should be easily identified because of the specific frequency range, but the networks have learned vertical translational invariance. It also has a greater plasticity than other calls. The syllable production rate varies significantly from one sample to the next.

Table 11 documents the accuracy for the field datasets. The most abundant calls are in bold, to separate them from circumstantial data (e.g. the 100% *P. canus* accuracy in RM corresponds to one call). The accuracy for TN is outstanding and for LPR1 rather good. Then the performances decrease as we move toward the right of the table. The datasets were intentionally ordered by month of recording, and we observe that woodpeckers become harder to identify as the passerines take over the acoustic space. The decay in accuracy for the *D. martius* rattle and above all for the noise class is significant. A drop from 99% to 93% in noise identifications means an increase in false positives in the woodpecker identifications. The drop to 97.9% in RM might seem limited, but the RM dataset includes 641,561 images in the noise class, and thus the ability to isolate these samples is critical.

Table 11
Calls identification: performance on images from the field datasets.

Net	TN	LPR1	LPR2	LPR3	RM
			Accuracy (%)		
D. martius1 (Rattle)	96.2	**84.7**	**71.7**	30.8	0.0
D. martius2 (Flight)	100.0	**75.5**	54.2	75.0	
D. martius3 (Contact)	100.0	**88.7**	80.0	87.5	
D. medius		79.3	84.6	100.0	
J. torquilla	75.0		100.0		**56.0**
P. canus	**98.0**	59.5	**48.0**	**50.0**	100.0
P. viridis		**76.5**			31.3
Noise	99.5	99.2	92.8	93.4	97.9
			Number of images		
Actual calls	197	3388	1489	267	2340
False positives	62	529	6938	11,440	13,356
False negatives	5	455	127	23	1019

Only the audio segments for which one of the images was predicted as a call were reviewed to assess the ground truth. Calls that were not detected by any model are a blind spot. In bold, the most abundant calls. False positives and false negatives evaluate any woodpecker call versus noise.

The number of false positives for the woodpecker call classes are manageable in TN or LPR1, then escalate to exceed 10,000 images in LPR3 and RM. This is put in perspective in Table 11 with the true number of images from woodpecker calls in the dataset; in LPR3 particularly, the number of files that have to be manually set aside becomes out of proportions with the number of interesting images in the set. Four classes are systematically over-predicted by the networks: the *D. martius* rattle, *D. medius*, *P. canus* and *P. viridis*. These are the most abundant in the training set. We seem to have built models that predicted these classes a lot, rather than well. On the other hand, considering the amount of *J. torquilla* samples in the training set, the numbers of false positives for this class is not excessive: for 2263 true positives in Remerschen, there are 2476 false positives. These samples were quite redundant (from the same two birds). *D. minor* is abundantly

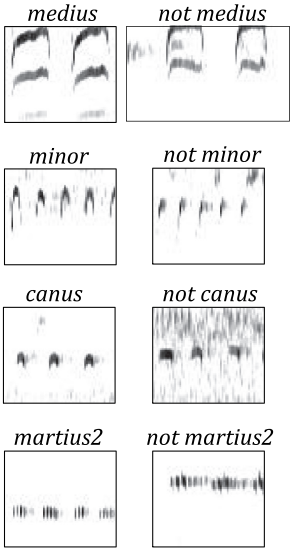

Fig. 9. Confusions in Remerschen and La Petite Raon.

Table 12
Calls identification: performance at the call level.

Class	TN	LPR1	LPR2	LPR3	RM
		Accuracy[a] (%)			
D. martius1 (Rattle)	100.0	**98.7**	**92.3**	50.0	0.0
D. martius2 (Flight)	100.0	**94.1**	66.7	100.0	
D. martius3 (Contact)	100.0	**97.3**	100.0	80.0	
D. medius		97.1	100.0	100.0	
J. torquilla	100.0		100.0		**81.7**
P. canus	**100.0**	94.7	**65.5**	**72.5**	100.0
P. viridis		**94.1**			40.9
All woodpeckers	100.0	97.9	80.5	73.5	80.0
Noise	98.6	98.0	81.7	81.8	94.1
		Number of calls			
Actual calls	72	699	394	98	656
False positives[b]	51	429	5437	9449	8905
False negatives	0	3	7	2	112

Bold indicates the most abundant calls.

[a] As we used majority-voting to pool the different models, the Mean Reciprocal Rank (MRR) is at least equal to the accuracy. Both would decrease if additional false negatives were uncovered. For comparison, Lasseck (2018) obtains an MRR of 82.7% on foreground species, 74.0% when including the background species as well.

[b] False positives and false negatives evaluate any woodpecker call versus noise.

predicted foremost in RM; there are indeed species in the dataset that can provoke this confusion. In the end, every class found its imitator; see Fig. 9 for a few examples. In the other direction, 21% of the woodpecker images were misdiagnosed as noise. Many of these misses were to be expected, e.g. when the images caught only a part of a syllable or the fuzzy tail of a call.

The numbers discussed above relate to partial images of woodpecker calls. At the call level, the resulting accuracy is documented in Table 12. Here, the identification was deemed correct if one of the sub-images was correct. The accuracy increased by approximately 20% for the abundant calls compared to Table 11. Even late in the season, 72.5% and 81.7% of the calls from the birds that owned the territory (*P. canus* in LPR; *J. torquilla* in RM) were detected. We conclude that some groups of syllables enable the identification of calls better than others. This suggests it would be beneficial to group them into larger images, assuming the better syllables would prevail in the analysis.

The false positives now amount to 1%–18% of the noise audio segments. In a dataset such as RM, having to review only 6% of the irrelevant data is what makes any analysis possible to start with. After the number and diversity of singers pick up in the late spring, false positives increase dramatically.

There are relatively few confusions between the classes of woodpecker calls, aside from a triangle between the rattles of *D. martius*, *P. canus* and *P. viridis* (Table 13). The last two are indeed sometimes hard to discriminate, but the first one has an early ascent in pitch whereas the two *Picus* tend to decrease in pitch through the call. Naturally, it is arduous to capitalize on these characteristics when considering only a fraction of the call. Besides this, the bulk of the confusions are with noise.

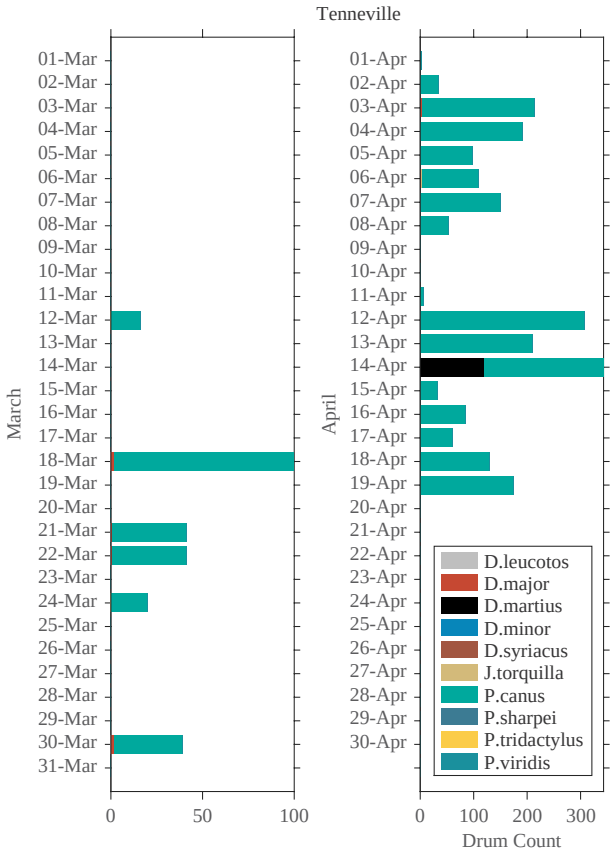

Fig. 10. Tenneville: Drums Identifications by Date.

4.4. Overall woodpecker monitoring

Figs. 10 through 15 show a time line of drums and calls detected in the three field datasets. The Tenneville *P. canus* appears as an outstanding drummer, perhaps due to its isolated position on the edge of the distribution zone. The same species did not drum as much at the other locations. Tenneville was the site of a territorial dispute between *P. canus* and *D. martius* on April 14th 2016, rather visible in the drums; the calls actually show that the *D. martius* had been in the vicinity for a while. A *J. torquilla* also called on 2 days, including on the day of the dispute.

There is not much drumming at Remerschen due to the lack of

Table 13
Calls identification (at the call level): confusion matrix accumulated over all field datasets.

Actual classes↓	Predicted classes: number of samples									
	mart1	mart2	mart3	med	min	torq	can	shp	vir	noise
D. martius1	718	3	2	10	9	9	27	13	152	6
D. martius2	14	80	1	2	1	2	3	0	4	0
D. martius3	0	0	46	2	0	0	0	0	0	2
D. medius	3	0	2	42	0	0	1	0	0	0
J. torquilla	2	0	1	10	5	517	1	0	0	108
P. canus	95	4	14	17	0	4	242	0	97	4
P. viridis	1	0	0	1	11	3	1	0	25	4
noise	4094	487	510	9063	2326	2266	4046	178	3610	232,395

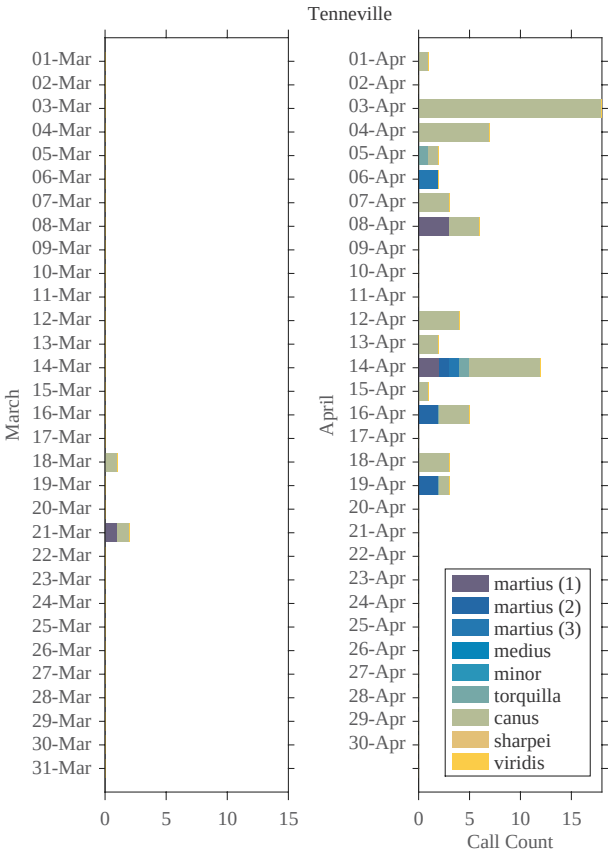

Fig. 11. Tenneville: calls identifications by date.

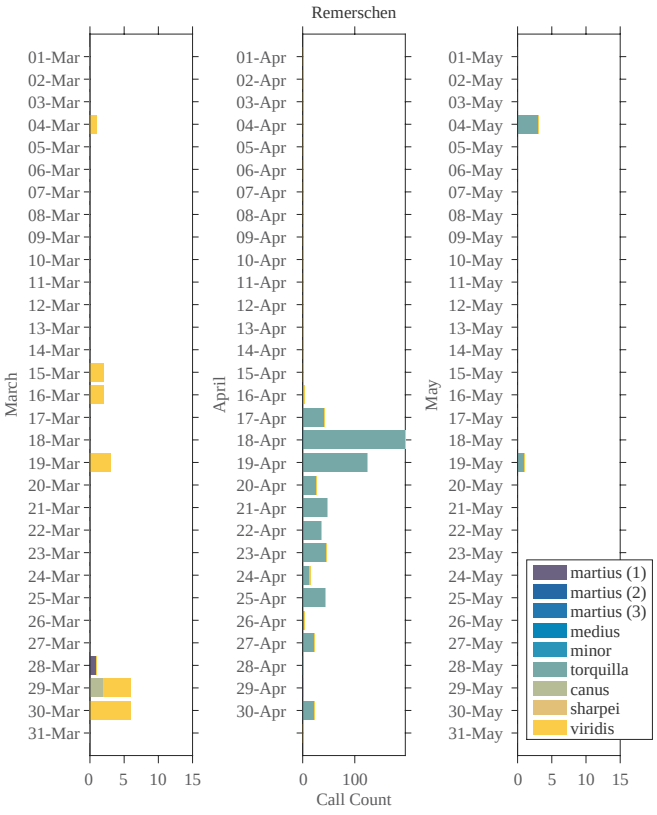

Fig. 13. Remerschen: calls identifications by date.

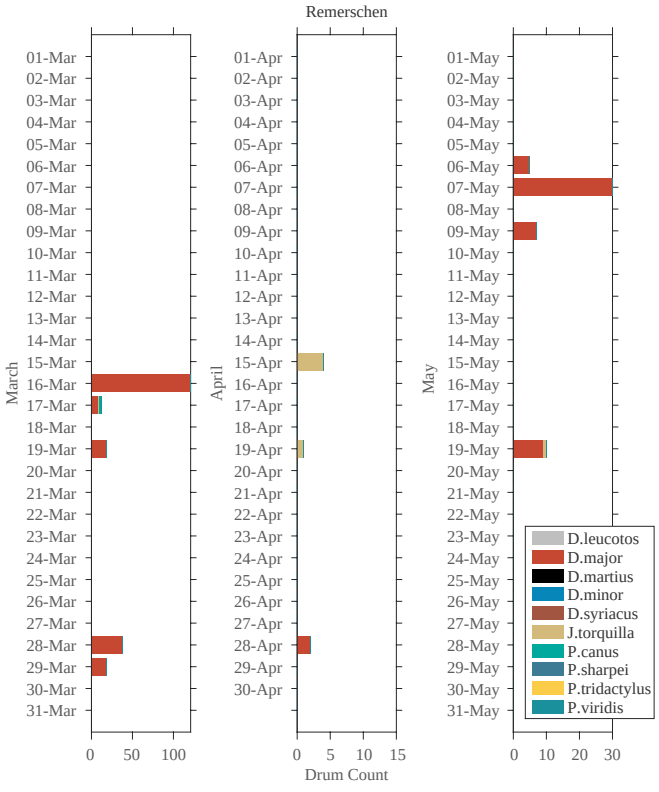

Fig. 12. Remerschen: drums identifications by date.

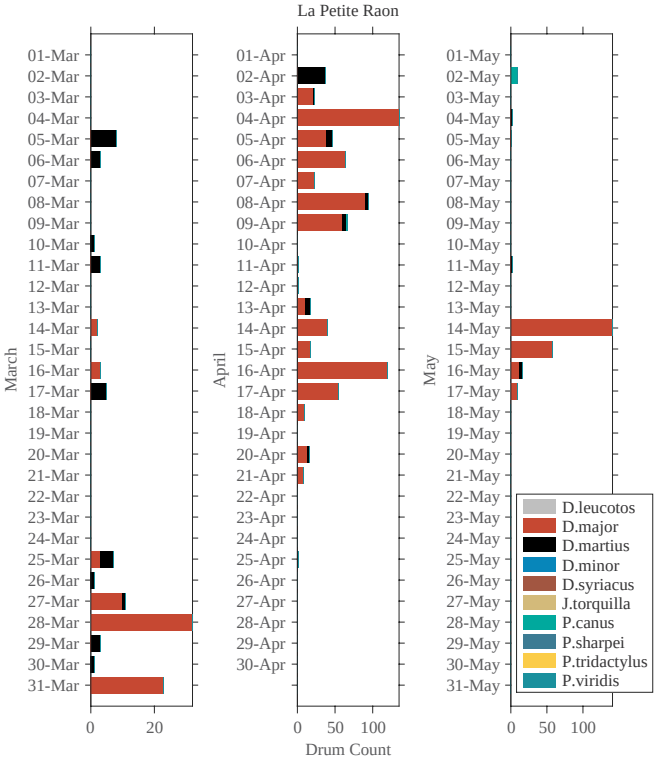

Fig. 14. La Petite Raon: drums identifications by date.

proper substrates. The willow trees that surround the ponds are too soft. As in La Petite Raon, the drums analysis is dominated by *D. major*. The station was installed in the midst of a *J. torquilla* territory, and an

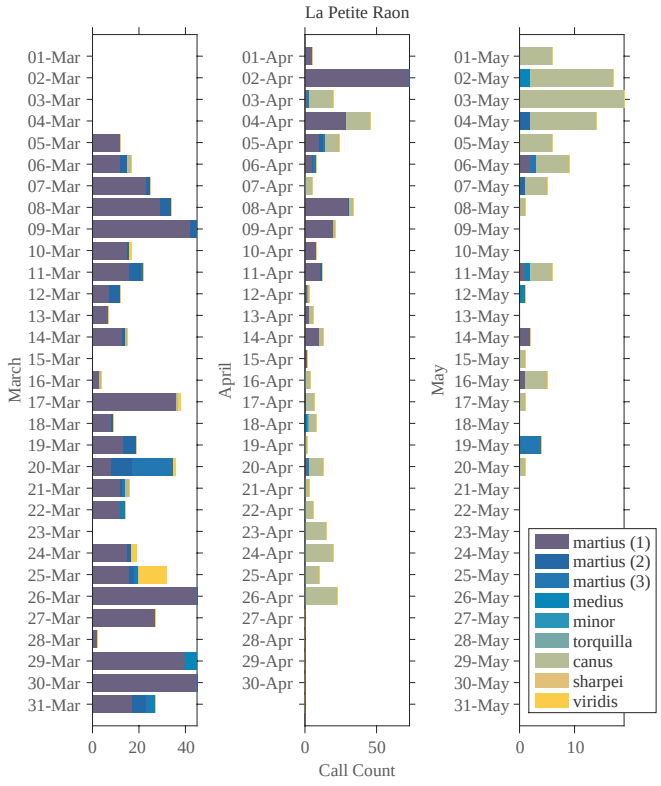

Fig. 15. La Petite Raon: calls identifications by date.

abundance of calls were recorded. In the mistakes of the different networks, particularly Inception, we discovered rare *J. torquilla* drums and taps (only the drums are showed in Fig. 12).

The LPR recordings show different woodpecker species sharing the same territory. On Fig. 15, we see that it was initially occupied by *D. martius*, which used its rattle call above all and the other two sporadically. The site was visited on occasion by *P. viridis* and *D. medius*. At the end of March, *D. major* started drumming and continued throughout most of April (Fig. 14). In early April, *P. canus* also claimed the territory and *D. martius* gave up ground. *P. canus* was still calling at the beginning of May. Then woodpecker activity receded.

Of all the species that we could have recorded, *D. minor* is the only one we missed. The signals from this species were always a bit fringe; fast drums, high-pitched calls. The drums are almost impossible to tell apart from *P. canus*, and the calls are confused with passerines. Without a positive identification in our data, we cannot be assured that we are able to detect these signals.

5. Discussion

5.1. On the success of deep convolutional neural networks

There is no doubt that without DCNNs, we would have been unable to analyze the RM and LPR datasets. The networks reduced these vast datasets into tentatively annotated datasets of a manageable size. A large part of these tentative annotations were actually correct (Tables 8, 10, 12). A manual review was still necessary because of a non-negligible amount of false positives, which increased as the reproduction season progressed. In the worst case (RM calls), the false positives amounted to 13,384 images (total positive predictions: 14677 images) out of a dataset of 643,901 images. In terms of recording time, the RM dataset amounts to 435 h of audio; the segmentation into individual acoustic events reduced it to 164.7 h, and the DCNNs made positive identifications in audio files totaling 13.9 h. Thus the deep networks transformed an impossible review into a tedious review. With the help of a good setup to look at images in batches, the task could be completed. The drums datasets are an order of magnitude smaller because of the restricted frequency range; RM, again the most difficult case, generated 17.3 h of audio to process (28,601 images), which was reduced to 1.6 h of positive predictions by the DCNNs (1000 images).

This success of deep networks in sound problems is also the success of spectrograms as acoustic features. We saw with the drums detection exercise that image-based methods (deep networks, spectrogram cross-correlation) performed much better than signal-descriptive methods (repetitions analysis). This reflects the efficiency of spectrograms in describing sounds. The spectrogram is not exactly the raw sound, but it spares the networks from having to reinvent the primary analytical tool in acoustics, i.e. the Fourier Transform. It follows that the time parameters used in its calculation have to be carefully considered, namely the frame duration and the overlap between the successive frames. We saw for example that 46 ms frames were problematic for some drums (Fig. 1), notably the *D. minor* ones. Then the time resolution became critical in the drums identification.

The drums identification could probably have been done together with the detection, as we did for the calls. For this task, the signal-descriptive method (handcrafted features and k-NN) worked better than the deep networks. This is not only connected to the limited precision of the images, but also to a conceptual difficulty. The temporal features that discriminate the different species (acceleration, speed, duration, number of strikes) are not immediate information in the images. Image analysis first picks up on shapes; it works well on spotting the peculiar patterns of woodpecker call syllables or the series of vertical lines that drumming produces. But the time intervals between the lines and the number of lines is another level of information processing, just like the agency of syllables in a call. We supply the spectrograms on the left of Fig. 1, and the deep networks must recover the graphs on the right. Performing this sophisticated task requires a deep network, which is not an issue here, but learning it demands a large quantity of data. Five thousand drums is a small dataset to retrain the networks through their entire depth. The consequences were felt in our study as some classes were assigned seemingly based on their perceived probability. This is a sign that the DCNNs were not able to construct reliable discriminant features, and therefore that their analysis of time structures was not thorough enough. As often with DCNNs, improvement would be best achieved using a larger training set. Here, the additional data would help the networks unlearn that objects have varying sizes (drums have fixed dimensions) and develop their notion of rhythm (not only the presence of objects is discriminant, but also their respective positions). In a second step, using larger images could be considered to accommodate finer spectrograms.

The task of distinguishing the woodpecker calls from each other is simpler than distinguishing them from non-target signals. Real-life datasets are full of bird calls that can legitimately be confused with woodpecker calls on sight, either in the same bandwidth and with somewhat similar syllables (owls) or in a different bandwidth and with almost identical syllables (unidentified passerines). There is definite value in building a noise class that represents the full scope of biodiversity that the networks might subsequently have to deal with. We had done it more diligently for the drums detection. For the calls we assembled a noise class that comprised a lot of non-target woodpecker signals (e.g. drums, call notes), which were likely to be encountered in woodpecker territory, but we had not sufficiently represented other species. In effect, the construction of the positive class is obvious (woodpecker calls) but the construction of the noise class is a notch more difficult. The sounds that are confused with woodpeckers are mostly known from accumulating bad experiments.

This remark brings the following issue: to be complete, the noise class has to be large. Yet, to teach neural networks that the different classes all have, a priori, the same probability, one needs to populate the different classes in the training set equally. If the noise class has to be large, then we need more data for the woodpecker classes, otherwise

their detection probability decreases. On the other hand, such an effect could be desirable, because woodpecker calls are a less probable occurrence than a lot of other noises in forests. Our considerations loop back to a choice between false positives and false negatives. In the context of a woodpecker monitoring scheme, it is preferable to minimize the false negatives, and thus to cope with the excess audio to review. This being said, the magnitude of the datasets limited us to reviewing only the calls that at least one of the networks had qualified as woodpecker signals. Our study is blind to pure false negatives. Fortunately, we observed that the birds that own the territory call and drum abundantly, thereby increasing their detection probability.

Our end-to-end methodology was designed for woodpeckers and as such will not readily address other species, but most of the aspects that do not transfer pertain to the steps ahead of the DCNNs rather than to the DCNNs themselves. The frequency ranges used for the ACI calculations and for the segmentation into acoustic events are too restrictive for other species, yet were instrumental in downsizing our datasets. The spectrogram computation parameters we used might elsewhere be ill-suited. The use of one-second audio excerpts to identify the calls, i.e. the focus on syllables rather than full calls, might also be problematic for other species. Finally, the woodpecker signals, both drums and calls, are heavily stereotyped and thus generate similar images as they are repeated. The capacity of the DCNNs is untested when it comes to handling variable structures. For example the blackbird (*Turdus merula*), whose phrases are series of random elements with infinite variations, could be a difficult case. Still, we recall that Sevilla and Glotin (2017) and Lasseck (2018) used the same computer vision DCNNs on large groups of species with some success. Additional studies focused on difficult species would be informative.

5.2. On the identification of species

The positive results in Table 12 lead us to believe that most woodpecker calls can be identified at the syllable level. They were very few confusions between the different species; some were observed between the rattles of *D. martius*, *P. canus* and *P. viridis*. Larger images would allow incorporating some of the call structures into the analysis to improve this result. The deeper networks certainly have the analytical power to study call structure, but again, the limitation is on the available training data.

Calls identifications are more confident than drums identifications, with either method; we did not need to seek context or other signals in the recordings to confirm the species. Advertising calls are without question species-specific. For drums, the debate is ongoing (Dodenhoff et al., 2001; Florentin et al., 2016). There is enough overlap between the various parameter ranges that some of the species cannot be differentiated in practice, e.g. *P. canus* and *D. minor*. Sex, region and function all affect the structure of drums, notably by a modification of the number of strikes (Blume, 1996; Blume and Tiefenbach, 1997). The shorter soft drums also increase the confusions in drums identifications (Florentin et al., 2017). This is further impaired by the realities of field recordings: distant drums, poorly executed signals, significant variations in the production of a single individual. The most confident predictions are the ones for which there is a volume of observations. Otherwise the predictions must be confronted with contextual data: location, habitat, co-occurrence of calls. Without this secondary information, the scope of drums analysis is somewhat limited.

5.3. On the respective merits of the different networks

In Lasseck (2018), Inception was the best performing network. In our case, we obtained the best results with different architectures depending on the problem and the training parameters. We also obtained different standings from repeated identical simulations, and there was no correlation between the networks that performed well on the validation set and the networks that performed well on the field data. As it stands, our validation sets were poor windows into the actual performance of the trained networks, which explains part of the randomness in the results, but not all. During training, the samples are shuffled and presented to the network in a random order, which can lead to different results in repeated simulations. In the calls analysis, almost all training runs, using different configurations for the learning rate, generated models that exceeded 94% of accuracy on the validation set. However, the models trained with an adaptive learning rate fared better with the field datasets. Most often, the deeper networks outperformed the shallower ones on the field datasets. ResNet 152 was the most accurate to detect drums, DenseNet to identify drums and calls. Inception produced poor results, likely because both its architecture (with filters of different sizes) and its original training (with 144 crops) made it skilled at recognizing objects at different scales. For the same reason, this was the best network to detect secondary signals like demonstrative tapping. It remains hazardous to designate one architecture as superior.

Using a ensemble of models either improves or deteriorates the accuracy compared to a single good model. For the three toughest calls sets (LPR2, LPR3 and RM), using one of our instances of DenseNet alone was a better choice than the ensemble. If we consider the *J. torquilla* calls that dominate the RM dataset, 44% of the images (989 images) were missed by the model pool. Half of these were actually correctly identified by one or more models and not properly promoted by the vote because the correct models were in the minority. Considering all species in all datasets, 12% of the images were not recognized at all (but adjacent images were, and allowed detecting these calls). Again, with our poor validation set, we did not pick models wisely for the ensemble. The majority-voting procedure still deserves some criticism; if too many models in the ensemble are incompetent, the good models will not save the day. We otherwise experimented with averaging the class probabilities produced by the different architectures, but this lowered the accuracy further. Note that the above comments pertain to the number of images that were correctly detected; for the number of false positives, the ensemble performed better than any individual network, by far. DenseNet was one of the poorest networks for false positives.

5.4. On data augmentation

Data augmentation would be a logical option to try and compensate for the modesty of our datasets. However, as mentioned, the original training of the legacy image networks already deployed most of the basic tricks of data augmentation. Applying a time shift to the images, stretching them or degrading their quality were things that could not be taught again with the same benefits. Dropout was also already used, in all networks but ResNet.

We still ran limited experiments with the calls. We were not able to gain anything from adding noise segments to our data, whereas it had been a success in other studies (Lasseck, 2018). This was likely because the noise we added was sampled from the existing noise class. Unlike Lasseck, we did not have noise data at our disposal that the networks had not already seen before. What actually improved our results was to add new data: we retrained the networks using the results from the LPR1 dataset, and this led to a decrease in false positives in LPR2, LPR3 and most significantly RM. This is consistent with our previous assessment that there is value in populating the noise class using past mistakes.

But the Lasseck (2018) result is interesting in that the added noise is not focused on audio segments that bring in differential information. It is just more sound data. Another interesting perspective is from Pironkov et al. (2018). These authors trained a speech recognition network simultaneously on two different objectives (recognizing speech and denoising sound clips), with different data feeding into the two tasks. The performance of the final network on speech recognition was improved. Lu et al. (2004) also improved speech recognition by first training on speaker gender recognition. Hence an interesting direction for future work on woodpecker sound identification would be to

mobilize auxiliary sound data by training the networks on different objectives. For example, the 16,000 sound clips from the BAD challenge (Stowell et al., 2019) could be used to train on the detection of bird calls prior to training for woodpeckers, or simultaneously. We could also envision that the secondary objective might not even be bird-related. The networks could first be trained to identify a number of familiar sounds, in the same way they were trained to detect common objects (cars, animals…) in images.

This brings a last remark on the image invariants that are inappropriate for sound. An analysis based on spectral content would not have confused the *D. martius* flight call in Fig. 9 with the unknown call on the right; the vertical translational invariance was not properly unlearned by the deep networks. However, unlearning invariants might require just as much data as it took to learn them in the first place: 1 million images. In that case it could be more profitable to retrain the different architectures from scratch with sounds and toward an acoustic objective. Instead of color channels, different spectrogram scales could be used (different time steps, frame durations, etc.). However, this is contingent on the publication of large and correctly annotated audio collections.

6. Conclusion

In this work we presented models to detect and identify the drums and the characteristic calls of European woodpeckers in audio streams. Two technologies proved decisive in our endeavor: the acoustic complexity index and deep convolutional neural networks. Calculating the ACI aboard our recording station allowed an early assessment of potential woodpecker content. We turned an indicator that was originally intended to measure species richness into a mean to scale down the field datasets. They were still too large for most detection and classification techniques, but deep image networks enabled their detailed analysis.

We singled out two limitations that could translate into future research: the need for very large training sets and the fact that the image invariants that the deep networks learned in their original training are improper for spectrograms. For the first point, we concluded that adding data by any means, including data serving peripheral objectives, was the direction to explore. The second point is an argument in favor of building deep networks directly for sound, and therefore supports the establishment of a public million sound collection.

We tested our algorithms on continuous field recordings and this brought an uncommon insight into the practicality of the developed methods. For three successive years, we were able to record woodpeckers in the wild and to analyze the audio in full. Such an outcome was previously inaccessible and opens up new possibilities for ornithological research. Deep networks can support a number of behavior, evolution and conservation studies for European woodpeckers.

Our recordings show that they drum and call abundantly, with intra-species and inter-species exchanges. We lament the absence of *D. minor*, but on the positive side, the capture of *J. torquilla* drums and taps in Remerschen, the *D. medius* calls in La Petite Raon and the many recordings of *P. canus*, both as a rarity in Belgium and in the heart of its territory in the Vosges mountains, are all satisfying outcomes.

Acknowledgments

The authors wish to thank Sohaib Laraba and Thierry Ravet at the UMONS Numediart Institute for their advice on deep networks; Kyle Turner, the Xeno-Canto foundation, the Xeno-Canto recordists and the Museum für Naturkunde Berlin for sharing their data; Jean-Yves Paquet and Alain de Broyer from Aves-Natagora, the Département de la Nature et des Forêts in Wallonia, Patric Lorgé from the Biodiversum Centrum in Remerschen and Alain Remy in La Petite Raon for their assistance with the field measurements; and Kevin Nis and Régis Berton from the UMONS Theoretical Mechanics Dynamics and Vibrations Unit for their assistance with equipment design and construction. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2019.101023.

References

Adavanne, S., Drossos, K., Çakir, E., Virtanen, T., 2017. Stacked convolutional and recurrent neural networks for bird audio detection. In: Signal Processing Conference (EUSIPCO), 2017 25th European. IEEE, pp. 1729–1733.

Blume, D., 1996. Schwarzspecht, Grünspecht, Grauspecht: *Dryocopus martius, Picus viridis, Picus canus*, 5th ed. Die Neuhe Brehme-Bücherei. Westarp Wissenschaften, Magdeburg.

Blume, D., Tiefenbach, J., 1997. Die Buntspechte (*Gattung picoides*). Die Neuhe Brehme-Bücherei. Westarp Wissenschaften, Magdeburg.

Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J.L., Krakauer, A.H., Clark, C., Cortopassi, K.A., et al., 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J. Appl. Ecol. 48 (3), 758–767.

Brandes, T.S., Aug 2008. Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Trans. Audio Speech Lang. Process. 16 (6), 1173–1180.

Del Hoyo, J., Elliott, A., Sargatal, J., 2002. Handbook of the Birds of the World. Vol. 7. Jackamars to Woodpeckers. Lynx, Barcelona.

Dodenhoff, D.J., Stark, R.D., Johnson, E.V., 2001. Do woodpecker drums encode information for species recognition? Condor 103 (1), 143–150.

Florentin, J., Verlinden, O., 2017. Autonomous wildlife soundscape recording station using Raspberry Pi. In: XXIV International Congress on Sound and Vibration (ICSV), London, UK.

Florentin, J., Dutoit, T., Verlinden, O., 2016. Identification of european woodpecker species in audio recordings from their drumming rolls. Ecol. Inform. 35, 61–70.

Florentin, J., Gérard, M., Turner, K., Rasmont, P., Verlinden, O., 2017. Towards a full map of drumming signals in European woodpeckers [abstract]. In: XXVI International Bioacoustics Congress. Haridwar, India.

Foote, J., Cooper, M.L., Nam, U., 2002. Audio retrieval by rhythmic similarity. In: Proceedings of the 3rd International Conference on Music Information Retrieval. Paris, France.

Fox, E.J., Roberts, J.D., Bennamoun, M., 2008. Call-independent individual identification in birds. Bioacoustics 18 (1), 51–67.

Fuchs, J., Pons, J.-M., 2015. A new classification of the pied woodpeckers assemblage (Dendropicini, Picidae) based on a comprehensive multi-locus phylogeny. Mol. Phylogenet. Evol. 88, 28–37.

Gorman, G., 2014. Woodpeckers of the World; the Complete Guide. Bloomsbury Publishing.

Grill, T., Schlüter, J., 2017. Two convolutional neural networks for bird detection in audio signals. In: 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, pp. 1764–1768.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 770–778.

Hinton, G., Srivastava, N., Swersky, K., 2012. Lecture 6a: overview of mini-batch gradient descent. In: Neural Networks for Machine Learning. URL https://www.cs.toronto.edu/\textasciitilde{}hinton/nntut.html

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 4700–4708.

Joly, A., Goëau, H., Botella, C., Glotin, H., Bonnet, P., Vellinga, W.-P., Planqué, R., Müller, H., 2018. Overview of lifeclef 2018: A large-scale evaluation of species identification and recommendation algorithms in the era of ai. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, pp. 247–266.

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

Laraba, S., Brahimi, M., Tilmanne, J., Dutoit, T., 2017. 3d skeleton-based action recognition by representing motion capture sequences as 2d-rgb images. Comp. Animat. Virtual Worlds 28 (3–4), e1782.

Lartillot, O., Toiviainen, P., 2007. A matlab toolbox for musical feature extraction from audio. In: International Conference on Digital Audio Effects. Bordeaux, France.

Lasseck, M., 2015. Towards automatic large-scale identification of birds in audio recordings. In: International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, pp. 364–375.

Lasseck, M., 2018. Audio-based bird species identification with deep convolutional neural networks. In: Working Notes of CLEF, pp. 2018.

Lu, Y., Lu, F., Sehgal, S., Gupta, S., Du, J., Tham, C.H., Green, P., Wan, V., 2004. Multitask learning in connectionist speech recognition. In: Proceedings of the Australian International Conference on Speech Science and Technology.

Mikusiński, G., Angelstam, P., 1998. Economic geography, forest distribution, and woodpecker diversity in Central Europe. Conserv. Biol. 12 (1), 200–208.

Miles, M.C., Schuppe, E.R., Ligon IV, R.M., Fuxjager, M.J., 2018. Macroevolutionary patterning of woodpecker drums reveals how sexual selection elaborates signals under constraint. Proc. R. Soc. B Biol. Sci. 285 (1873), 20172628.

Pellegrini, T., 2017. Densely connected cnns for bird audio detection. In: Signal Processing Conference (EUSIPCO), 2017 25th European. IEEE, pp. 1734–1738.
Perktas, U., Barrowclough, G.F., Groth, J.G., 2011. Phylogeography and species limits in the green woodpecker complex (Aves: Picidae): multiple *Pleistocene refugia* and range expansion across europe and the near East. Biol. J. Linn. Soc. 104 (3), 710–723.
Pieretti, N., Farina, A., Morri, D., 2011. A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). Ecol. Indic. 11 (3), 868–873.
Pironkov, G., Wood, S.U., Dupont, S., Dutoit, T., 2018. Investigating a hybrid learning approach for robust automatic speech recognition. In: International Conference on Statistical Language and Speech Processing. Springer, pp. 67–78.
Potamitis, I., 2014. Automatic classification of a taxon-rich community recorded in the wild. PLoS One 9 (5), e96936.
Salamon, J., Bello, J.P., Farnsworth, A., Kelling, S., 2017. Fusing shallow and deep learning for bioacoustic bird species classification. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 141–145.
Schmitz, L., Dumoulin, R., 2004. Hybridation des pics vert et cendré (*Picus viridis*, *P. canus*) en belgique. Aves 41 (1–2), 91–106.
Sevilla, A., Glotin, H., 2017. Audio bird classification with inception-v4 extended with time and time-frequency attention mechanisms. In: CLEF (Working Notes).
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
Sordello, R., 2012. Synthèse bibliographique sur les traits de vie du pic cendré (*Picus canus*, gmelin, 1788) relatifs à ses déplacements et à ses besoins de continuités écologiques. In: Tech. rep., Service du patrimoine naturel du Muséum national d'Histoire naturelle.
Stark, R.D., Dodenhoff, D.J., Johnson, E.V., 1998. A quantitative analysis of woodpecker drumming. Condor 100 (2), 350–356.
Stowell, D., Plumbley, M.D., 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ 2, e488.
Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y., Glotin, H., 2019. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods Ecol. Evol. 10 (3), 368–380.
Sueur, J., Farina, A., 2015. Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics 8 (3), 493–502.
Swiston, K.A., Mennill, D.J., 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. J. Field Ornithol. 80 (1), 42–50.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1–9.
Turner, K., 2011. The case against drumming in middle spotted woodpecker (*Dendrocopos medius*). Limicola 1, 37–53.
Winkler, H., Short, L.L., 1978. A comparative analysis of acoustical signals in pied woodpeckers (Aves, Picoides). Bull. Am. Mus. Nat. Hist. 160.
Zabka, H., 1980. Zur funktionellen bedeutung der instrumentallaute europäischer spechte unter besonderer berücksichtigung von *Dendrocopos major* (l.) und d. minor. Mitt. Zool. Mus. Berl. 56 (Suppl.: Ann. Orn. 4), 51–76.