UMONS
POLYTECH MONS

UNIVERSITÉ DE MONS
FACULTÉ POLYTECHNIQUE
Service de Mécanique Rationnelle,
Dynamique et Vibrations
Boulevard Dolez, 31. B-7000 Mons (Belgium)

Thèse de doctorat déposée en vue de l'obtention du grade de
Docteur en Sciences de l'Ingénieur
par
Juliette FLORENTIN

Reconnaissance automatisée de sons naturels - Application aux pics (*Aves*)

Automated Recognition of Natural Sounds - Application to Woodpeckers (*Aves*)

Membres du Jury :

Prof. Francis Moiny – UMONS (Président)
Prof. Georges Kouroussis – UMONS (Secrétaire)
Prof. Olivier Verlinden – UMONS (Promoteur)
Prof. Thierry Dutoit – UMONS (Co-Promoteur)
Prof. Pierre Rasmont – UMONS
Dr. Stéphane Dupont – UMONS
Prof. Dick Botteldooren – Universiteit Gent
Prof. Jean-Jacques Embrechts – Université de Liège
Prof. Thierry Tison – Université Polytechnique Hauts-de-France
Dr. Dan Stowell – Queen Mary University of London

May 2019

À mes deux grand-mères, Lise et Iolanda.
L'une était docteur en médecine,
l'autre n'a jamais eu la possibilité d'étudier.

Le travail fut long, car l'ingénieur, voulant produire un effet formidable, ne comptait pas consacrer moins de dix litres de nitro-glycérine à l'opération.

Jules Verne, *L'Île mystérieuse*.

La technique moderne est basée fondamentalement sur une science préindustrielle : la mécanique. Or celle-ci élimine le vivant. L'exclut. On a fabriqué des écoles d'ingénieurs - selon une tradition militaire - où on enseignait les maths, la physique, un peu de chimie, ensuite l'électricité, l'électronique, etc. La modernité occidentale a fabriqué un univers mécanique oubliant seulement que nous étions, nous, des êtres vivants dans une nature vivante. La société industrielle a des fondements scientifiques incomplets et anachroniques.

Claude Lorius & Laurent Carpentier, *Voyage dans l'Anthropocène*.

Et maintenant, il venait d'achever un nouveau poème. Dédié au pic mar, le bel oiseau qu'on ne voyait plus en Suède. Le poète des oiseaux, pensa-t-il. Presque tout ce que j'ai écrit les concerne.

Henning Mankell, *La cinquième femme*.

Acknowledgments

First and foremost, I would like to thank Olivier Verlinden for welcoming me to the Theoretical Mechanics, Dynamics and Vibrations department, and for giving me the opportunity to pursue my interest in Bioacoustics. Under his guidance I have been spoiled with unprecedented scientific freedom. There were also times in which I was busier with personal difficulties than research, and in these days his patience and humanity helped me see it through.

I would also like to thank the engineers on my thesis committee, Georges Kouroussis, Francis Moiny, Thierry Dutoit and again Olivier Verlinden, for the open-mindedness they demonstrated while accompanying this unusual work. To answer that quote I copied from French glaciologist and CNRS gold medal Claude Lorius, I attempted to put a few living organisms into my engineering mind. I deeply appreciate that my committee endured the experiment. It was probably not that hard for Thierry Dutoit, who is an amateur birder himself, but I nonetheless thank him for his unwavering support for my work. He facilitated it any way he could.

From the other side of campus, I am most thankful for the guidance Pierre Rasmont offered me throughout this thesis and most crucially in its early stages, when my knowledge about birds was non-existent. He welcomed me into his ornithology class, took me on field trips, explained species, gave me access to the researchers in the Zoology department, and then time for meetings, appreciation and encouragements. All of it was deeply appreciated.

This work would not have gotten far without the researchers in Thierry

Dutoit's group. My deepest thanks to Christian Frisson, Thierry Ravet, Sohaib Laraba, Stéphane Dupont, Omar Seddati, Jérôme Urbain, Gueorgui Pironkov and others. Thierry Ravet and Sohaib Laraba most helped me through the final lap, Sohaib by sharing his Pytorch codes, Thierry by installing some proper CUDA for me, both for giving me access to the computer that made this work possible.

For their help and advice along the way, I would like to thank a number of my colleagues at the University of Mons: Maxence Gérard in Zoology, Benoît Fauville, Pierre Lecomte and Frederic Coquelet in Physics, Katshidikaya Tshibangu and Damien Bury in Mining Engineering, Aurélien Van Laere in Telecommunications, and Oleksandr Skrylnyk and Claudine Judex in Thermodynamics.

Jean-Yves Paquet, Alain de Broyer, Franck Hidvegi, Didier Vieuxtemps and Anne Weiserbs from Aves took plenty of time to discuss woodpeckers with me. Philippe Moës of the DNF gave me access to the forest road in Tenneville and often transported me and my batteries to the station site. Patric Lorgé of the Biodiversum Centrum in Luxemburg spent his entire spring of 2017 changing the station's batteries. My stepfather Alain Remy proposed himself for the same tedious task for the 2018 campaign. I know how much hassle it is and they both have my deepest gratitude.

I would also like to thank the bioacousticians and ecoacousticians who took time to exchange with me and significantly influenced my work: Michael Towsey (Queensland University of Technology, Australia), Nicolas Mathévon and Maxime Garcia (University of St-Etienne, France) and Karl-Heinz Frommolt (Humboldt University Berlin). Through these encounters and many others, the IBAC and Ecoacoustics conferences have greatly enriched my knowledge and my work.

Many researchers in the machine learning community publish their scripts and their neural networks. Without this, the present thesis would have a far more limited scope. Hence I owe much gratitude to Thomas Grill at the Austrian Research Institute for Artifical Intelligence in Vienna, and to the authors of AlexNet, Inception, ResNet, DenseNet and VGG. The same goes for all the enthusiasts who upload their recordings of bird calls to Xeno-Canto and thus contributed to a ground-breaking major resource for bioacousticians. To that list I add Kyle Turner, who contacted me in 2017 after having read my article on woodpecker drumming and offered his own expertise and his vast recordings collection. This proved invaluable for my work.

Closer to us, what work could be achieved in Theoretical Mechanics without the help of Kevin Nis and Régis Berton? Simply put, they built the recording station. Kevin's knowledge of electroacoustics, informatics and Arduino board programming was instrumental in that regard. I give them both my warmest thanks.

In the name of camaraderie, I would also like to thank all my colleagues past and present in Theoretical Mechanics: Georges Kouroussis, Pierre Tihon, Lassaad Ben Fekih, Quentin Oggero, Bryan Olivier, Joseph Tsongo Vughuma, Hoai Nam Huynh, Loïc Ducarne, Python Kabeya Tshibamba, Georgios Alexandrou, Philippe Brux, David Wattiaux, Gentiane Dirksen, Marjorie Dequenne, Monsieur l'Ancien Recteur Calogero Conti, Monsieur l'Ancien Recteur Serge Boucher, once more Kevin Nis, Régis Berton and our unit head Olivier Verlinden. The best was saved for the end: our secrétaire extraordinaire, Marjorie Godart!

My friend Barbara kindly proofread this manuscript.

Abstract

There are eleven species of woodpeckers on the European continent. Ten of them drum on trees and seven have long-distance advertising calls. Every year from March to May, these signals contribute to forest soundscapes while woodpeckers draw territories, find mates and dig tree cavities. Each drum and each call is species-specific and easily picked up by a trained ear. In this thesis, we have worked toward automating this process and thus toward making the continuous acoustic monitoring of woodpeckers practical. There were two main steps to implement: first the detection of woodpecker signals against the backdrop of diverse acoustic communities and secondly the identification of the different species. Because continuous monitoring generates hundreds of gigabytes of data, detection had to be progressive; first we coarsely trimmed the datasets using a simple indicator, the Acoustic Complexity Index (ACI), then we analyzed more elaborate sound features. Species identification required mostly a description of duration and rhythm for the drums and an analysis of the spectrograms for the calls. For both detection and species identification, for both the drums and the calls, deep neural networks provided the most efficient, if not the only solution. Two favorable circumstances made this possible: 1) legacy very deep image nets (up to 169 layers) were made public and could be re-trained to address specific image problems and 2) the sound problem could be transformed into an image problem via the spectrogram. When tested on development datasets obtained from online archives such as Xeno-Canto, very deep nets easily recognized 95% of submitted drums and calls, also alongside other noises. For real-life datasets, the false positives came in larger numbers. The nets

get confused by the countless birds that could not be taken into account during training. Another point that calls for caution is the fact that the image invariants that sustained the original training of the deep nets (e.g. the enlarged image of a car still represents a car) do not necessarily apply to spectrograms. Overall, the woodpecker signals were recognized with a high accuracy in March and early April, when the forest is relatively quiet. Later in the season, the false positives crept in, but the nets still allowed discarding more than 95% of the recordings. This number further increased when the nets were trained with known confusing signals. In the end, a reasonable number of audio files was left that could be reviewed manually. This dataset reduction is a consequent improvement compared to other techniques and allowed very deep nets to make the acoustic monitoring of woodpeckers a reality.

Contents

Chapter 1

Introduction

The Nature Conservancy has worked for a long time with communities in Papua New Guinea on planning their land use. In 2015, the organization's scientists started to research how to demonstrate that the empirical solutions they were putting forward were truly beneficial to biodiversity. They installed an array of sound recorders in locations with different types of land occupation such as gardens, hunting grounds, conservation land, etc. In the recordings they found the correlation they were looking for between land use and the number and variety of observed species (Burivalova et al. [9]). Impressed with the technology, they started to promote acoustic monitoring in Indonesia to investigate which logging practices best preserved the forest biodiversity[1].

In Belgium, when the Leffe quarry applied for an extension of operating hours into the night, the regulator complained that it might impact the bats in the surrounding Natura 2000 zone. How could it be assessed? Plecotus[2] stores survey data from known hibernation sites, but that does not tell if and how bats use the quarry. Four nights of recording ultrasounds later, a picture emerged of *Pipistrellus pipistrellus* hunting insects in the quarry lights and *Myotis myotis* crossing over to return to the woods just before dawn.

These stories unfolded in the wake of a rising new scientific discipline, *ecoacoustics*, powered up by advances in machine learning and the wide availability of cheap electronics. The International Society of Ecoacoustics (ISE) was founded in 2014 after a successful symposium in Paris, entitled

[1]This story is from the keynote speech of Eddie Game, scientist at the Nature Conservancy, delivered at the Ecoacoustics 2018 conference in Brisbane, Australia. More at https://blog.nature.org/.

[2]The NGO that surveys bats in Brussels and Wallonia.

"Ecology and Acoustics: Emergent Properties from Community to Landscape". To understand how ecoacoustics came to existence, let us first backtrack to the common difficulties associated with biodiversity monitoring.

Surveying species is at the heart of every environmental assessment. From plants to frogs, from birds to fishes, from terrestrial to marine mammals, everything gets counted. From these numbers, indices can be calculated that quantify the health of a habitat, for example Shannon's diversity index or the species richness. However, inevitably, surveys are restricted in scope to amount to a manageable effort; the time spent, the number of families surveyed or the geographical area are reduced. The European Union, for example, has decided to focus its conservation efforts on birds, which materialized in the Birds Directive[3]. Birds are a common survey target for a number of reasons. First, they are iconic animals and as such have long been the focus of conservation efforts. Secondly, they are excellent bio-indicators, i.e. studies show that if the avian population of an ecosystem is healthy, then other animal populations are likely to be in a good shape as well (Voříšek et al. [87]). Finally and most crucially, birds are relatively easy to detect because they reveal their presence through their singing.

As a result of the Birds Directive, in member countries of the European Union, bird surveys are carried out yearly by a multitude of volunteers, 170 for Wallonia alone[4]. Each volunteer either visits a handful of preset counting stations, spending 3 to 5 minutes per location, 2 or 3 times per year, or walks along a line transect and documents every bird seen or heard on the way. The locations are chosen to satisfy a statistical scheme that extrapolates the sparse counts to country estimates[5].

The methodology for bird counts is not without a few loose threads. First, identifying birds is a skill that is neither easily acquired nor widely available. Maintaining volunteer pools with sufficient expertise is a struggle. Then, even though the statistical foundation of the estimations has gained

[3]Directive 2009/147/EC of the European Parliament and of the Council of 30 November 2009 on the conservation of wild birds. The European Bird Census Council (EBCC) coordinates survey efforts from all member states and compiles the European Breeding Bird Atlas (EBBA). In Belgium, bird monitoring is delegated to the two ornithological societies, Natuurpunt in the North and Aves-Natagora in the South.

[4]Surveillance des Oiseaux Communs en Wallonie (SOCWAL) at:
http://biodiversite.wallonie.be/fr/socwal.html?IDC=3730.

[5]It has also become commonplace to supplement the traditional bird counts with citizen science portals where anyone can log in their sightings. In Belgium, this is the popular observations.be / waarnemingen.be.

widespread acceptance, it cannot shake off questions about the extremely sparse sampling involved, elusive species (e.g. nocturnal species) or hard-to-reach areas. In 2014, Aves launched a project to estimate the remaining population of grey-headed woodpeckers (*Picus canus*) in Belgium. Despite a good effort, the study produced only meager results and was abandoned in 2015, leaving questions unresolved. Had *P. canus* left Belgium for good, or had it become nearly impossible to detect? In a country like Indonesia, surveys are an almost impossible feat. The country is home to 17% of the world's bird species. Tropical forests are full of little-known species, whose names change from village to village. This is the place where taxonomists err[1].

Then came the age of artificial intelligence and big data. There were breakthroughs in human speech recognition and this was expected to translate to bird calls. A technology roadmap started to form (Blumstein et al. [7]). Microphones would be deployed in the wild and would continuously scrutinize their environment. The audio stream would be mined for bird calls, and the vocalizing species would be automatically identified. There would be a smartphone application for amateur birders to use on their walks[6]. This scientific promise has been actively pursued by the ecoacoustics community. Every year since 2014, the BirdCLEF challenge submits new datasets for audio-based bird species identification[7]. Since 2016, the Bird Audio Detection (BAD) challenge aims to develop a robust bird detection algorithm[8]. Microphone arrays have already been deployed. A conservation project on the mountain Hymettus near Athens, Greece, had 17 recording stations in place, transmitting data through the cell phone network (Jahn et al. [37]). The Queensland University of Technology recorded 28 TB of audio over several years throughout Australia (Towsey et al. [81]). Puerto Rico and Costa Rica installed semi-permanent stations with an online data access (Aide et al. [3]). Towsey et al. [82] talked about a "data deluge". Only a fraction of the recordings could be reviewed, and with significant manual labor.

For this reason, the ability to automatically detect and identify species is a pressing need. Here, the difficulty is the adaptation of technologies intended for human signals (speech, music, photographs), with their specific

[6]There are several. http://www.ornithomedia.com/pratique/equipement/applications-pour-smartphones-pour-identification-oiseaux-00678.html.

[7]https://www.imageclef.org/.

[8]http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/.

context of form and meaning, to another set of signals (bird sounds) that operate under a different referential and where the important attributes are not the same. The most creative singers, birds in the order of the passerines, are the most challenging in that respect. On the other side of the spectrum, bat algorithms were faster to reach maturity, because bat signals are simpler, with a fixed shape and at frequencies disputed by no other animals. This is how the quarry data mentioned at the beginning of this introduction could be analyzed.

Nevertheless, the lack of species detection and identification algorithms has not deterred ecoacousticians from analyzing audio. The early attempts included tentative metrics, like the "vocal activity" (number of calls) in Frommolt & Tauchert [26]. Then more relevant indicators were designed: the Acoustic Complexity Index (Pieretti et al. [62]), the Normalized Difference Soundscape Index (NDSI) (Kasten et al. [39]), β-indices (Sueur et al. [77]) and false-color spectrograms (Towsey et al. [81]). Their purpose was to a) estimate species richness and diversity without the actual counts and b) visualize acoustic data in a compact way and thus provide an overview of what the audio contains without actually listening to it. These indicators have opened new doors in conservation. They are the basis of the Nature Conservancy's assessment of habitats in Asia.

With such accomplishments, ecoacoustics is shaping up as a bypass to traditional species counts in a number of applications. A new goal surfaced, besides helping large-scale bird monitoring surveys: addressing conservation questions from another angle, using the emerging techniques of ecoacoustics. As in the Nature Conservancy case, they have proven able to provide answers in a time- and cost-efficient manner. This is a decisive advantage for ecoacoustics. Another one is that the competing technologies (radar, flying camera) require a perfect visibility, a circumstance that might not exist outside of marine applications. Only acoustic monitoring may work when the view is obstructed by clouds or vegetation, or when the setting is as in Frommolt & Tauchert [26], where the authors searched for a rare nocturnal bird (*Botaurus stellaris*) in remote wetlands. Finally, the beauty of the acoustic solution is that it hacks into the actual communication channel of birds. As long as birds talk to each other, acoustic monitoring will work.

The present thesis offers a contribution to ecoacoustics research in the form of a complete treatment of the European woodpeckers case. The purpose is not quite to perform surveys for this family of birds, but rather to

develop the tools to detect them in audio streams and identify the species. With this focused application case, we hope to confront the hurdles that prevent current algorithms from delivering a reliable mass processing of large audio datasets. Here, by processing, we mean the isolation and analysis of multiple target signals. Woodpeckers were chosen because they form a coherent group whose acoustic signals are innate, and therefore have little plasticity; these were necessary qualities to design a problem of reasonable complexity. Chap. 2, which reviews birdsong in general, sheds further light on these aspects. Chap. 2 also presents the technologies currently involved in bird species detection and identification, and points out the assumptions and limitations we hope to test on woodpeckers. The acoustic signals used by woodpeckers are described in Chap. 3, including the most iconic of all: drumming. For our work, woodpecker calls and drums are collected from public archives and through measurement campaigns in Belgium, Luxembourg and France (Chap. 3 & Appendix B). From the measurement campaigns, we expect to gather realistic datasets, but also hands-on knowledge about such datasets and practical woodpecker information that could guide algorithm design. The processing of the data is addressed in Chap. 4 through 6. Chap. 4 discusses the detection of woodpeckers in long audio recordings, i.e. the production of a clean set of acoustic signals out of the data deluge. Then, the way drums can be used to identify the species is presented in Chap. 5. Chap. 6 is about detecting and identifying calls using Deep Neural Networks (DNN). Eventually, by conducting our application case to its conclusion, we aim to outline a basic framework for the acoustic monitoring of woodpeckers.

Chapter 2

The Many Folds of Automated Bird Song Identification

2.1 Species Traits in Birdsong

Birdsong[1,2,3] is primarily a manifestation of reproductive behaviors. The vocalizations are broadcast to attract a mate and/or to signpost a territory. In most species, these are adult male functions[4]. Hence not all birds sing, and not all the time. In temperate climates, the bulk of birdsong occurs in the spring. When the weather is bad, it gets dropped to prioritize food search.

Because of their function, bird vocalizations must convey information about whether a sexual partner is fit and belongs to the proper species. In other words, they carry individual and species markers. They are also loud and clear, because this information must go through to conspecifics (sexual partners, intruders) and heterospecifics (neighbors, intruders). Aside from the territorial calls, birds have a host of other calls, e.g. alarm calls, in-flight

[1]The term "song" is reserved for the birds from the order of the Passerines, also aptly called songbirds; for other species, the correct term is "calls". "Vocalizations" is the generic term, although the term "song" is commonly abused. The motivation for maintaining such a difference is that calls are innate, whereas songs are learned. As passerines mature from juveniles to adults, songs go through a plastic (intermediate) form before reaching a stereotyped (final) form. The young do not strictly copy other males; instead, they draw inspiration from their songs to create a new and personal version that retains the key elements indicative of species.

[2]The animal kingdom has the following subdivisions: phylum, class, order, family, genus (pl. genera), and species (pl. species). Birds form the *Aves* class. In our context, a taxon (pl. taxa) is a species.

[3]Unless indicated otherwise, the description of birdsong in this section stems from Catchpole & Slater [11].

[4]Female song actually exists but is understudied (Odom et al. [56]). Juveniles also have their own vocalizations.

calls, contact calls, mob calls, dispute shrieks, juvenile calls. These are often only captured at short distances and have foreseeably weak, if any, species markers.

Hence, the realistic scope of purely acoustic methods is territorial vocalizations, with the known limitation that they can only survey a fraction of the bird population, i.e. adult males engaged in reproductive behavior. Silent males, females and juveniles will not be accounted for.

Birds structure their vocalizations in a variety of ways, using both spectral parameters (e.g. pitch, frequency sweep) and temporal parameters (e.g. duration of the song, interval between syllables). Different songs from the same bird form a repertoire. A song can be broken down into phrases, syllables and elements. Most songs are found in the 1 kHz – 8 kHz frequency range; some reach up to 15 kHz. Because of the reduced length of their vocal tract, smaller birds will vocalize at higher frequencies. In all parameters, the intraspecific variations (between individuals of the same species) might be significant and overlap with other species. For example, the present thesis documents multiple confusions between *P. canus*, the grey-headed woodpecker, and *Picus viridis*, the green woodpecker, two closely related species.

Nevertheless, the vocal species present at a given location[5] are able to share the acoustic space, both in time and frequency. This is the acoustic niche hypothesis (Krause [43]): bird communities compose a symphony, in which every species occupies its own place on the music sheet. Incidentally, the niches depend on location. The characteristic traits of the calls of a given bird species have some flexibility in order to adapt to different contexts. Some species will make their songs really distinctive only when faced with competition. Kirschel [40] studied the case of two closely related species of African tinkerbirds[6] with overlapping habitats; he showed that these species altered their songs to reinforce their differences where they were sympatric (i.e. when they lived in the same geographic area). The birds that had fine-tuned their songs did not recognize the more generic songs of isolated populations of their own species. In the case of *P. canus* and *P. viridis* cited above, hybridation occurs when better partners cannot be found; the calls do not establish the two species as irreconcilable. The hybrid offspring is sterile and its calls are a mix of the two parent species (Schmitz [69]; Ławicki et al. [48]).

[5] Including frogs, bats, insects...

[6] From the order of the Piciformes, like woodpeckers.

The above has direct consequences on automated bird song classification. Classification algorithms need to be trained on a reference database and the scope of the training database conditions the scope of the algorithm. If the aim is to recognize all bird species, then should repertoires not be catalogued in extenso, including regional dialects? There are 10,711 species of birds on Earth[7] (a more modest 884 in Continental Europe[8]). Some repertoires have more than 200 songs. The repertoires are also renewed in time, when outsiders join the community or when the young propose new versions of the songs. After a span of twenty years, less than 10% of the original songs may be left. In addition, some species improvise their songs or imitate the songs of others. Drafting a full catalogue seems both out of reach and a questionable approach. However birds have somewhere in their brain a simple template that they use to read vocalizations and assess who is in their species and who is not, at least in a local context. This is the relevant information that the acoustic analysis should chase. The training database does not need to collect all potential variations of a call, but rather enough variations of a call so that its revealing feature(s) are brought out.

In practice, species markers can be assessed through audio replay experiments. For example, Garcia et al. [28] replayed drums to *Dendrocopos major* individuals, both original recordings and modified versions in which some of the parameters were modified[9]. He observed that *D. major* still responded when its characteristic acceleration pattern was removed from the drums. The birds recognized a conspecific and reacted by drumming back to defend their territory. Only when both the acceleration and the amplitude decay were replaced did the birds stop responding. This would indicate that *D. major* uses the combination of accelerated strikes and amplitude decay to identify its conspecifics. Evidently though, such results are available for only a few acoustic signals and a few species.

[7] http://www.worldbirdnames.org/

[8] https://avibase.bsc-eoc.org/

[9] See Chap. 3: woodpecker drums are territorial signals and carry species markers. The work of Garcia et al. is reproduced in Appendix C.

2.2 Detection of Bird Sounds in Audio Streams

There are two critical functions required of algorithms: 1) detecting bird sounds in audio streams and 2) identifying the species emitting these sounds. The two tasks are not necessarily separate: the most common and long-standing birdsong detection technique, *spectrogram cross-correlation*[10], inherently includes the species identification step because it searches audio recordings for one given sound. Its principle is to have a template of the target spectrogram[11], i.e. an image, sliding over a continuous audio stream until a maximum in cross-correlation is reached. Performance-wise, it is well-suited to sounds with a rigid and simple form, and little intraspecific variation. For example, Ulloa et al. [85] were able to detect 34.9% of *Lipaugus vociferans* calls, with zero false positives[12]. These calls are advantageously loud and stereotyped. Swiston & Mennill [78] studied the detection of double-knocks from two species of woodpeckers (*Campephilus guatemalensis* and *Campephilus principalis*). There was no expectation that potential species traits in double-knocks would play a role and the same template was used for both species. Respectively 24% and 8% of double-knocks were detected, with respectively 97% 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. In summary, spectrogram cross-correlation is suitable for signals that can hardly be confused with other sounds. The proportion of false positives is otherwise a strong inconvenience: the results must be reviewed by a human in a time-consuming process. False nega-

[10] Sometimes also called *template matching*.

[11] Generic processing for audio files starts with the Discrete Fourier Transform (DFT). As birdsong is a varying signal, the DFT is repeated at short intervals. The audio segment is cut into *frames* of 10 to 50 ms, with overlap between the successive frames, and the DFT is performed on the individual frames. This is called the Short Time Fourier Transform (STFT). The STFT spectra are stacked along the horizontal axis to form *spectrograms*: contour plots with time on the x-axis, frequency on the y-axis, and spectral amplitude on the color scale. This is a comprehensive representation of how the spectral content evolves throughout a sound. In human speech processing, frames of 10 ms are typical because this is the smallest time window the human ear will capture. There is less consensus on what is appropriate for birds. Most analyses in the present thesis use windows of 23.2 ms with 50% overlap. A sampling frequency of 12 kHz was deemed appropriate for woodpeckers. A Hamming window is used in the FT, and light spectrum smoothing over 5 bins. In addition, noise filters are a requirement for most datasets.

[12] False detections.

tives[13] can be traced back to faint signals captured at a distance or to signals overlapped with other noises. They are foremost problematic for elusive species, but fortunately most birds tend to repeat their calls ad nauseam to get the message through.

Dynamic time warping (DTW), where the signals can be warped in time to account for variations in their duration, may be used to improve detection rates. However, the increase in computational complexity yields limited gains (Stowell et al. [75]). Other methods are designed to fit a few species of interest. For example, Bardeli et al. [4] were able to isolate the calls of two species (*Botaurus stellaris* and *Locustella luscinioides*) by tracking either the energy in a given frequency band or the repetition rate of song elements. The detection rates were respectively 96% and 93% (33% and 1% of false positives). Dong et al. [16] looked at increasing the number of taxa that could be detected. They did so by searching for ridges in the spectrogram images of whistles, clicks, chirps and harmonic tones. For 16 out of 20 species, there was a 100% chance of finding a true positive in every set of 10 retrievals. Recently, Deep Neural Networks (DNN)(see Section 2.5) pushed the subject forward. Most participants in the Bird Audio Detection challenge (BAD) (Stowell et al. [76]) used DNNs to separate bird calls from other noises. Adavanne et al. [1] estimated 16% of false positives, 8% of false negatives, and 42% of incorrect labels in all false identifications. Pellegrini [57] estimated 13–22% of false positives and 4–8% of false negatives (accuracy 88.3–90.7%).

For the BAD challenge above, approximately 100 hours of recordings were made available, i.e. 5 days. This is a scale that still permits detailed analysis, even if tedious. It is common for ecoacoustics projects to involve months of recordings. Towsey et al. [81] proposed *false-color spectrograms* to address the gap. False-color spectrograms are both a data reduction tool and a descriptive tool: 1) they display 24 hours of data on a single image by compressing standard spectrograms by a factor 2000 and 2) they show the different components of the local biophony[14]. The images can be used to track certain species or disruptions of the soundscape as a whole. Degraded soundscapes are indicators of degraded environments, as acoustical niches reflect ecological niches (Farina et al. [20]).

[13]Missed detections.

[14]A soundscape is the sum of its anthropophony (sounds made by humans and human machines), its geophony (geological sounds like rain, rivers, wind, etc.) and its biophony (animal sounds).

False-color spectrograms are obtained by superposing three different index spectrograms[15] that share common axes and whose color scales are respectively from black to red, black to green and black to blue. Towsey et al. [81] calculate indices in 1-minute blocks[16].

The first index, the Acoustic Complexity Index (ACI, red scale), was proposed by Pieretti et al. [62]. The ACI spectrum is calculated as:

$$\text{ACI(f)} = \frac{\sum_{k=1}^{n-1} |I_k(f) - I_{k+1}(f)|}{\sum_{k=1}^{n-1} I_k(f)} \tag{2.1}$$

$I_k(f)$ is the acoustic intensity of the k^{th} frame (frequency-dependent) and n the number of frames in the 1-minute time interval under consideration. High ACI values reflect a strong vocal activity of birds, because bird songs are a rapidly varying signal. Strong acoustic intensity variations from one frame to the next increase the numerator of the ACI. Monotonous sounds such as an engine running at constant rpm would have low ACI values.

An entropy spectrum is used as the second index (green scale). For each frequency bin, the temporal entropy proposed by Sueur et al. [77] is computed over the 1-minute interval. With the n values of the intensity $I_k(f)$ normalized so that they sum up to 1, the temporal entropy formulation is:

$$H_t(f) = -\sum_{k=1}^{n} \frac{I_k(f) * \log_2(I_k(f))}{\log_2(n)} \tag{2.2}$$

The entropy of a flat function is 1, whereas a lonely peak concentrating the energy yields zero. The latter is the expectation for rapidly varying birdsong, and thus Towsey et al. [81] use $1 - H_t$ so that high entropy values indicate the presence of wildlife.

The third index (blue scale) used by Towsey et al. [81] is the Acoustic Cover (CVR). It is computed for each frequency bin as the fraction of cells in the 1-minute interval where the intensity exceeds a threshold.

There are three dimensions in sound: the temporal variations, the spectral content and the sound intensity. The ACI and the entropy spectrum both

[15] An acoustic index is a statistic that describes an aspect of a sound. If the index is a spectrum, it can be used to build a spectrogram. The vertical lines are the successive index spectra, evolving through time following the x-axis.

[16] At 17,640 Hz with frames of 512 bins, there are 2067 frames per minute, without overlap. This is where compression is achieved.

characterize the first one; the CVR the third. The spectral content is visualized through the spectrogram form, but not commented upon by the chosen indices. We note that Sueur et al. [77] also proposed a spectral entropy index, similar to the temporal entropy but calculated along the frequency dimension. The reasoning is that for a spectral entropy of 1, all frequency bins have content, and supposedly all the acoustic niches are populated: the diversity is high in the recording. However, spectral entropy is not a spectrum as it sums ups the frequency bins. Consequently, it cannot be visualized in a spectrogram form.

Recently, Phillips et al. [61] achieved a 10^6 data reduction and produced Diel plots showing important ecological information over the span of a year. They first replaced each minute of recording by a vector of 12 acoustics indices; then they grouped 1.5 million of such vectors into 60 clusters with ecological meaning (e.g. predominantly birds, predominantly cicadas...). The Diel plots have time of day on the x-axis, day of the year on the y-axis and the cluster's assigned color on the color scale. It is a potent overview of seasonal variations in acoustic communities.

2.3 Audio Feature Extraction for Classification

The datasets used in classification tasks are assemblages of sound files that contain one song each (single-label), or songs from several species (multi-label). The files are preferably short, i.e. a few seconds long. Prior to classification, each one has to be reduced to a manageable and meaningful vector of audio features. The feature vectors are the basis on which sound files will be deemed similar or not. This similarity is most often measured by the Euclidian distance between feature vectors. Alternatively, the angle between two feature vectors is a suitable metric for large vectors with comparable norms.

Audio features[17] comprise all numbers calculated on sound files that provide a description of these sounds: duration, peak frequency, frequency range, full spectrum, octave bands, sound pressure level, energy, loudness, sharpness, zero crossing rate, linear predictor coefficients (LPC), Mel-frequency cepstral coefficients (MFCC), etc. Using more parameters at once improves the description. Early bioacoustics applications considered hand-crafted audio features (Schrader & Hammerschmidt [70]; Brandes [8]; Acevedo et al. [2]);

[17] Also acoustic features, markers, parameters, indicators.

then the LPC (Kogan & Margoliash [41]; Lee et al. [51]) were introduced, and finally the MFCC and their time derivatives (Kogan & Margoliash [41]; Somervuo et al. [71]; Ranjard & Ross [66]; Fox et al. [25]; Lee et al. [51]), which have known great success in human speech recognition and subsequently became a default solution for bird voices as well.

Mel-frequencies are frequencies that were rescaled to better take into account the actual perception of the human ear. To calculate the MFCC, the STFT is first converted to the Mel scale. Then the logarithm of its amplitude goes through another Fourier transform. The obtained function is called a cepstrum and the MFCC are its coefficients. Authors typically retain the first 15–30 at most, doubled or tripled by appending the first and second time derivatives.

The second Fourier transform in the MFCC gives weight to the repetition of frequency peaks in the spectrum, i.e. the harmonic content, which is a key aspect of the human voice. By design, the MFCC are tailored for human speech applications and thus their suitability to parameterize bird song remains questionable. In terms of quantity of information, the Mel scale is neutral at low frequencies and integrates larger frequency bands as frequency increases. The cepstrum calculation reduces the Mel spectrum to a limited set of essential MFCC coefficients. This being said, very large feature vectors tend not to be an issue anymore. Stowell & Plumbley [74] used the full Mel spectrum as a feature vector and further increased the dimensionality by projecting it onto an overcomplete basis. Here one should mind that classifiers do not see that adjacent bins of a spectrum have a meaningful connection and that energy can easily slip from one bin to the next while the overall sound remains similar. The MFCC on the other hand are approximately decorrelated (Stowell & Plumbley [74]).

Feature sets must be normalized so that different features have comparable ranges and therefore comparable weights in the classification. The usual practice is to scale all numbers back to the [0,1] or to the [-1,1] interval. The features can also be further transformed to improve their decorrelation (van der Maaten & Hinton [86]; Stowell & Plumbley [74]); Principal Component Analysis (PCA) is commonly used to that effect.

The features are calculated for all time frames of the sound files, including frames of silence. The values that are retained for further analysis are statistics such as the mean and standard deviation over all frames. This is a gross simplification of the temporal patterns in birdsong. Stowell & Plumb-

ley [74] experimented with alternatives; they stacked the features of several successive frames together to capture a temporal evolution, or used the first ten coefficients of a Fourier transform. Nevertheless, these evolutions did not perform better than the simple mean and standard deviation.

As a matter of fact, in spectrograms, bird songs produce image patterns from which species are readily identifiable (Fig. 2.1). Following this thought, Lee et al. [51] implemented a direct parameterization of MPEG images of the spectrograms. In the same vein, Potamitis [65] and Lasseck [46] derived their features from cross-correlation scores with a set of template images. Lasseck [46] showed that such features outperform spectral features[18] in the classification of syllables and elements of songs. Ultimately, the image analysis approach leads to Deep Neural Networks (DNN) (see Section 2.5).

By design, neural networks can work from a raw signal and build up their own features in their lower layers, prior to classification. Before their coming of age, Giret et al. [29] explored the possibility of automatically identifying the features that would work best with a given set of audio files. From a set of basic functions (root-mean-squared, zero crossing rate...), the authors produced new arbitrary mathematical functions, which evolved until their classification performance was maximized. Note that the new features remained subordinated to the original signal description; there was no new characterization.

Metadata (recording location, time of day and habitat characteristics) was included in the parameter set by Lasseck [46]. Such information may guide the classifier by factoring in a probability for a given species to be present in the recording.

[18] Lasseck used the OpenSMILE library (Eyben et al. [18]) of 57 low-level descriptors, including 35 spectral features and 13 MFCC. For all, the first and second derivatives were added. The time dimension was reduced using 39 functionals (mean, moments, percentiles...).

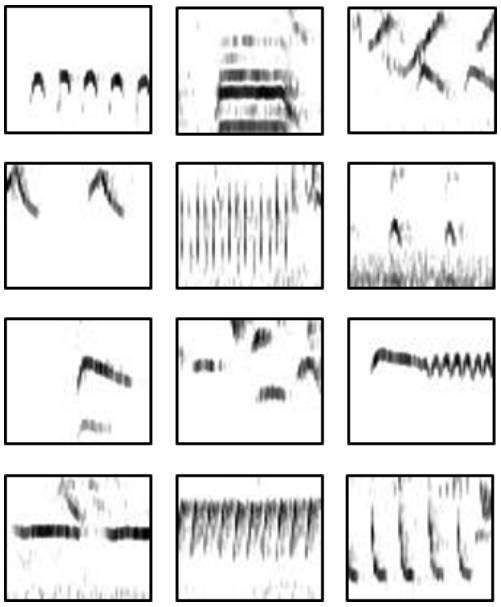


Figure 2.1: Images from Audio Recordings
in the Nature Reserve of Remerschen, Luxemburg

The spectrograms have a duration of 1 sec and span
the frequencies 1–3.5 kHz. All represent bird calls.

2.4 Classification up to 2015

The *classes* represent the different vocalizations up for identification (e.g. the song of Species A, the contact call of Species B, etc.). *Classifiers* first build an understanding of similarity/dissimilarity by processing a *training set* of known sounds. Then, unlabeled sounds from a *test set* are identified on the basis of a comparison to the training set. The performance of the classifier is established from the results of this second step. This process is called supervised machine learning, because there are labels in the training set to tell the algorithm what to learn. Unsupervised methods (k-means, t-SNE) look for patterns in a dataset based on the sole merit of the acoustic features. The purpose is to make data clusters appear[19].

Many supervised classifiers, of which Figure 2.2 depicts a few, have been employed to recognize bird vocalizations : k-NN (Connor et al. [12]), probabilistic models such as Hidden Markov Models (HMM) (Kogan & Margoliash [41]; Somervuo et al. [71]; Brandes [8]; Aide et al. [3]) and Gaussian Mixture Models (GMM) (Somervuo et al. [71]; Lee et al. [51]), Support Vector Machines (SVM) (Fagerlund [19]; Acevedo et al. [2]), decision trees (Ranjard & Ross [66]; Acevedo et al. [2]), random forests (Stowell & Plumbley [74]; Potamitis [65]) and early Artificial Neural Networks (ANN) (Ranjard & Ross [66]; Fox et al. [25]).

The performance of the above algorithms fluctuates with the following: 1) how comprehensive the training set is and how different from the test set; 2) how well the acoustic features describe the sounds; 3) how well the algorithm renders the topology of the data, while avoiding overfitting; 4) which metric is used to qualify performance. Classification problems are constrained by these four elements operating in conjunction with each other: the data, the features, the classifier and the uncertainty about the "true" potential of the methodology. Regarding 1), the recommended size for the training set would depend on the resilience of the algorithm with respect to slight variations in songs. Unwanted sounds such as noise or silences may need to be included as a separate class. The test set should be as dissimilar as possible from the training set. At minimum, all calls from the same bird or same recording should fall on the same side. Regarding 4), until 2015 the preferred measures of performance were accuracy and Area Under the Curve (AUC). Accuracy is the percentage of samples in the test set that

[19]Details in Appendix A, *More Machine Learning Concepts*.

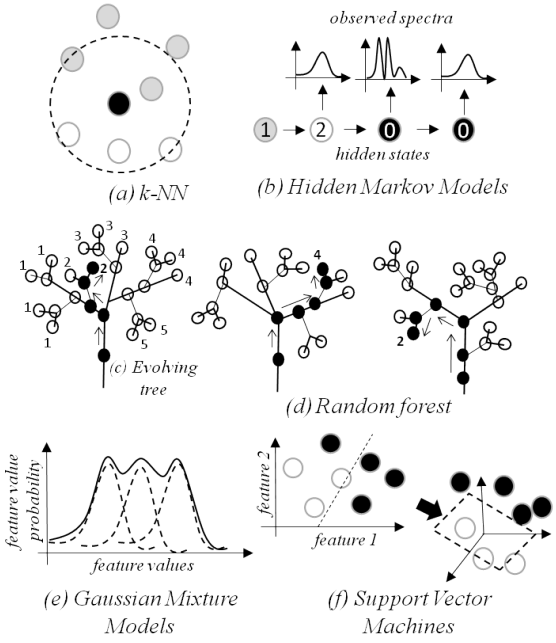


Figure 2.2: Classifiers

(a) With $k = 1$, the nearest neighbor to the black test sample is grey. With $k = 5$, the black sample joins the white group (white= 3, grey= 2). (b) The HMM structure shows a sequence of 4 hidden states, which have a fixed number of possible values. Three states produce observations, chosen from a limited set of spectra. For each class i, the probabilities to transition from one state value to the next and to produce a certain observation given the hidden state value are optimized. The optimized HMM instance is a generator of class i vocalizations. Then, given a test observation, the HMM that has the highest probability to generate this observation indicates the class. (c) Each new sample travels up the trunk of the evolving tree. It follows branches whose children are the most similar to itself and keeps climbing until it reaches a final leaf where it is positioned. Leaves crowded beyond a threshold are split into new leaves and thus the tree grows. Each end leaf is associated with a class. Once the tree is final, new samples are classified by sending them up the tree. (d) Trees of the random forest are built from randomly sub-sampling the training set and the feature vectors. Thus the same dataset is interpreted in a number of different ways. The forest returns a probabilistic class assignment, based on the results of the various trees. (e) The GMM computes the optimal Gaussian distributions, one per class, that best explain the features' multimodal distributions. Then the probability for a test sample to belong to class i is known. (f) The SVM artificially increases the dimension of the feature space until it becomes possible to define hyperplanes that separate the classes.

Table 2.1: Bird Classification 2006–2015

Authors	**Training Set**	**Features & Classifier**	**Results**
Somervuo et al. (2006)	14 sp. birds 792 calls Avg. 57/class	MFCC GMM or HMM	Accuracy 71.3%
Brandes (2008)	27 birds/crickets/ frogs 908 calls Avg. 34/class	Handcrafted features HMM	Accuracy 75.4–95.8% (birds)
Acevedo et al. (2009)	12 birds/frogs 10061 calls Avg. 838/class	Handcrafted features LDA[a] Decision tree SVM	Accuracy: 71,45% 89,2% 94,95% (birds 96.5%)
Connor et al. (2012)	67 birds 568 calls Avg. 8/class	Gabor transform[b] k-NN	Accuracy 59–98%/sp. Overall 90%
Lee et al. (2013)	28 birds 2627 calls 39–170/class	Image parameters GMM	Accuracy 0%[c]–100%/sp. Overall 94.6%
Potamitis (2014)	87 (78 birds) 687 calls Est. 24/class	Cross-correlation scores Random forest	ACU 91.7%
Stowell & Plumbley (2014)	501 birds 9688 calls Avg. 19/class	Coordinates in a basis of Mel spectrum clusters Random forest	ACU 85.4%[d] mAP 42.9%
Lasseck (2015)	501 birds 9688 calls Avg. 19/class	Cross-correlation scores reselected per species and metadata Random forest	ACU 91.5% mAP 51.1%

[a] Linear Discriminant Analysis, see description in Appendix A.
[b] Time-averaged, compact alternative to the STFT.
[c] Insufficient training data or high background noise.
[d] ACU 89.8% on the same dataset as Potamitis (2014).

are correctly predicted[20]. The AUC is presented as a more robust metric, insensitive to the composition of the dataset. Values above 90% are excellent and values below 60% poor[21]. Recently, the mean Average Precision (mAP) and the Mean Reciprocal Rank (MRR) were also used[22].

Table 2.1 presents a list of studies from 2006–2015 that are representative in terms of training set size, employed techniques and performance. Some results were inflated by advantageous study designs. In Acevedo et al. [2], the bandwidths of the three bird calls are distinct. The confusion of one of the calls with a frog call confirms that the description power of the handcrafted acoustic features was limited. There is no question however that the superiority of SVMs over LDA and decision trees was demonstrated. SVMs are skilled at revealing manifolds in a dataset. In Connor et al. [12], a study with a sizeable training set, every vocalization in the database had a twin to be matched to by k-NN. Lee et al. [51] were dissatisfied with the vulnerability of their method to noise. Because of its originality and isolation, it is hard to judge the viability of the approach.

In general, authors give little detail about the failures of their algorithms. The "variability" of songs is invoked as a difficulty (Somervuo et al. [71]; Potamitis [65]) in relation to passerines, i.e. the order with the most elaborate songs. A typical percentage of correct identifications for passerines is 70–80% (Somervuo et al. [71]; Brandes [8]; Fox et al. [25]). In recent efforts, datasets and evaluation metrics were mutualized (Stowell & Plumbley [74]; Potamitis [65]; Lasseck [46]) to help compare techniques, but it remains difficult to grasp the reach of these works, i.e. the level of complexity in bird vocalizations that they can reliably support. Do the low mAP values in Stowell & Plumbley [74] and in Lasseck [46] suggest a large number of false positives?

After 2015, deep convolutional neural networks took over (Joly et al. [38]; Lasseck [47]).

[20] Along this line, one can also describe the true positives (TP), correctly assigned to class i, the false positives (FP), wrongly assigned to class i, the true negatives (TN), correctly not assigned to class i, and the false negatives (FN), wrongly not assigned to class i.

[21] The AUC is further described in Appendix A.

[22] Details also in Appendix A.

2.5 Deep Neural Networks

Figure 2.3 depicts the architecture of a simple neural network, similar to the one used by Fox et al. [25] to identify the songs of seven individual birds from three species. This network had four layers: the inputs, two hidden layers and the outputs. The inputs were 12–15 MFCC, the hidden layers had 5–60 neurons each and the output layer 7 neurons, corresponding to the 7 birds to identify. A neuron produces an *activity* y_j, calculated from the activities x_i of the neurons in the previous layer, in two steps:

$$\begin{aligned} &(1)\ \ z_j = \sum_i w_{ij} x_i + b_j \\ &(2)\ \ y_j = f(z_j) \end{aligned} \tag{2.3}$$

In step (1), the w_{ij} coefficients are weights and b_j is a bias. It is common but not mandatory for a neuron to be connected to all the neurons in the previous layer. The particularity of neural networks is the f function in step (2). By requirement, f is non-linear, continuous and has a simple derivative. Common options for f are the hyperbolic tangent or the logistic function which is defined as:

$$y_j = \frac{1}{1 + e^{-z_j}} \tag{2.4}$$

Both project the output z_j to the $[0, 1]$ interval. When y_j is 1, the neuron is activated. When it is 0, the neuron is dormant. These two functions are also sigmoids (S-shaped), which presents two disadvantages. First, negative inputs produce a non-zero output and thus connections that should be abandoned for good might be revived at a later stage during training. Secondly, the slope is steep for small positive inputs and the sigmoid rapidly converges toward 1. This means that connections with a modest positive contribution are promoted almost as much as the best connections. For these reasons, the *rectified linear unit* (ReLU) and the *leaky*-rectified linear unit gained traction in recent works, e.g. in Grill & Schlüter [33]. Following Maas et al. [52], the mathematical expression for these nonlinearities is:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{otherwise.} \end{cases} \tag{2.5}$$

The rectified linear unit uses $\alpha = 0$ and the leaky-rectified unit $\alpha = 0.01$.

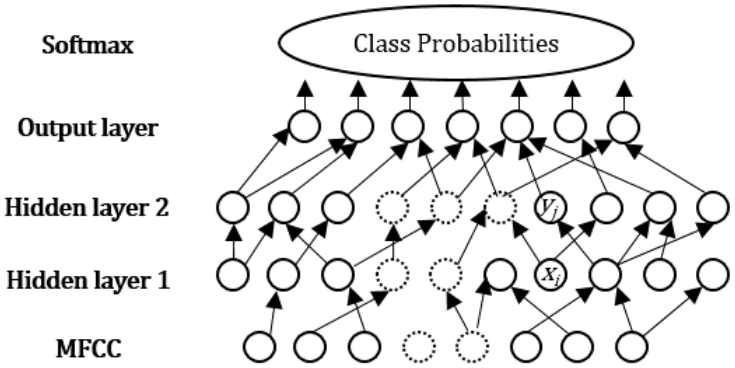


Figure 2.3: Neural Network Similar to Fox et al. [25]

In the output layer, the neuron that corresponds to the most probable class will compute the highest activity. The class probabilities are calculated in a "softmax" layer that comes on top of the output layer. It has the same number of neurons as the output layer and neuron j produces the value y_j:

$$y_j = \frac{e^{x_j}}{\sum_i e^{x_i}} \tag{2.6}$$

The training of a neural net consists in optimizing the values of the weights and biases. This is achieved by *backpropagation*: given a cost function C that measures the gap between the actual and desired activities of the end neurons, the calculation of $\frac{\partial C}{\partial w_{ij}}$ and $\frac{\partial C}{\partial b_j}$ is straightforward because of the simple derivative of f. Requests for change to the value of the cost function can be backpropagated to the weights and biases using partial derivatives:

$$\frac{\partial C}{\partial w_{ij}} = \frac{\partial C}{\partial y_j}\frac{\partial y_j}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} = \frac{\partial C}{\partial y_j} f'(z_j) x_i \tag{2.7}$$

A proper cost function for a "softmax" type of end layer with target values t_j is the categorical cross-entropy:

$$C = -\sum_j t_j \log y_j \tag{2.8}$$

For efficiency, large training sets are handled by mini-batches. The cost function and the network weights are updated after each batch. This is a

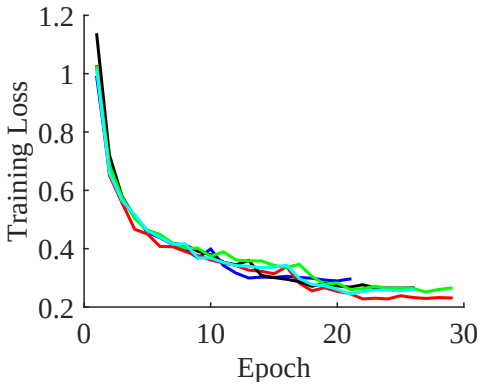


Figure 2.4: Training Loss

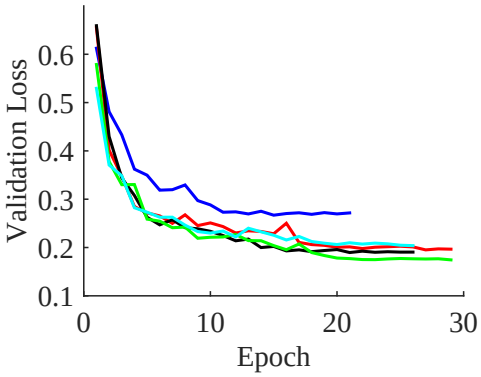


Figure 2.5: Validation Loss

trade-off between convergence speed and proper direction[23]. One iteration over the full training set is called an epoch. The amount of change imposed on a weight after an iteration is $\Delta w_{ij} = -\epsilon \frac{\partial C}{\partial w_{ij}}$, where ϵ is the learning rate (common value: 0.01 at most). In some implementations, training starts with a high learning rate to move quickly toward the region of the expected minimum. Then the learning rate is decreased to explore the minimum region in greater detail. In practice, it gets divided by ten after the training loss has stalled for a number of epochs. Fig. 2.4 and 2.5 show an example of training loss (cost function aggregated on the training set) and validation loss (id. on a validation set) progressing through training for five different training sets. Four of the five runs achieve comparable training and validation loss; only the dark blue one has a larger than average validation loss, indicative of over-fitting. The dark blue network improved by fitting the peculiarities in the training set but is unable to generalize its analysis to previously unseen data. Over-fitting, along with the mathematical dead end of neurons compensating each other but producing no valuable information altogether, have long caused neural networks to underperform. A successful technique that has then been employed to refocus learning is dropout. With every new batch of training samples that are evaluated, the net randomly drops a given percentage of its connections; as many as 50%. The remaining neurons must still predict the correct class. Hence the technique forces every neuron to contribute meaningful information on its own.

Fig. 2.4 and 2.5 also illustrate the practice of training a network on several

[23]See Geoffrey Hinton's 2012 online course "Neural Networks for Machine Learning", at https://www.cs.toronto.edu/ hinton/nntut.html.

folds of the training set. Here the network is trained five times, but every time a different fifth of the data is held back for validation. Then the five resulting models render different trends in the data. The final class probabilities are averaged over the five models and may be squared before average to enhance the confident predictions (Lasseck [47]). It is also common practice to pool together models that were trained using different parameters. Prediction scores always seem to benefit from the alliance of various approaches. Statistically, more predictions mean more chances of a good prediction.

Again, neural networks are intended to design their own features. The MFCC on the contrary represent an advanced stage of acoustic analysis. Using them as inputs, as Fox et al. [25] did, necessarily limits the scope of the neural net. In effect, Fox et al. [25] supplied the features and used the net as a mere classifier. The performance did not exceed other contemporary attempts: the accuracy was 70% for canaries, whose songs are highly variable. Proper feature extraction would require a rawer input and a deeper net. The depth of a net, i.e. the number of hidden layers, embodies its analytical power.

Research on the recognition of handwritten digits in images precipitated the rise of Deep Neural Networks (DNN), more specifically of a type of DNN that uses convolutional layers to analyze images (LeCun et al. [50]). Because sounds are easily replaced by (Mel) spectrogram images, the adoption of DNNs for sound problems was straightforward. Spectrograms are a good trade-off between working from a raw input and not having to reinvent the primary analytical tool in acoustics, i.e. the Fourier Transform. However, as always, the choices for frame duration and overlap are hardly adapted for all signals.

Salamon & Bello [68] used a DNN to identify noise sources (dogs, honks, music, etc.) in the New York City soundscape. The network (3 convolutional layers, 2 dense layers) was trained on a database of 8732 labeled samples (10 classes) and achieved a mean accuracy of 79% across the classes. More to the point, Adavanne et al. [1], Çakir et al. [10], Grill & Schlüter [33], Kong et al. [42] and Pellegrini [57] all used a DNN to build a bird call detection system in the BAD challenge. The performance of these nets was discussed in Section 2.2: less than 20% of false positives. The top-scoring network of Grill & Schlüter [33] (AUC 88.7%) had 4 convolutional layers, 3 dense layers and 373169 trainable parameters. A similar architecture is depicted in Fig. 2.6.

Convolutional layers are image-analysis tools. In Fig. 2.6, the bottom con-

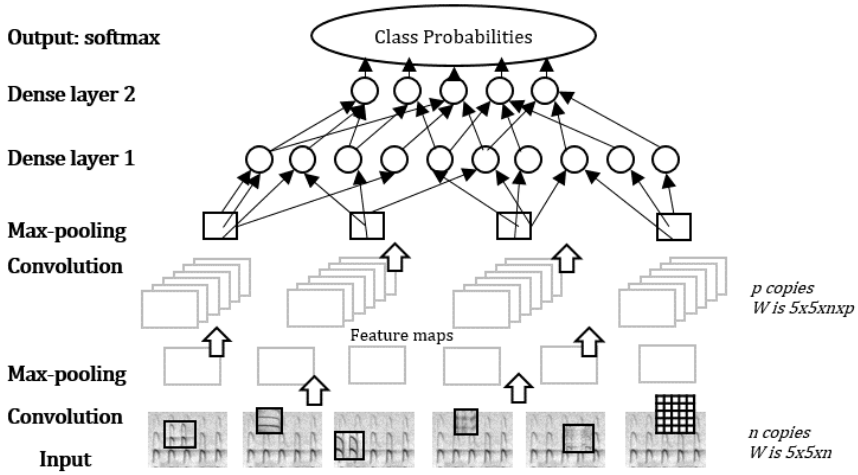


Figure 2.6: Neural Network Similar to Grill & Schlüter [33]

volutional layer has a width n: the input image is replicated n times to be analyzed by n different filters. The filters are small patterns, here squares of 5×5 pixels, that slide over the input image. At each position, a 5×5 portion of the input image is multiplied by the 5×5 weights of the filter. This is similar to spectrogram cross-correlation. The result is a new image of the same size as the input, each pixel storing the result of a local multiplication (and transformation through a non-linear function). The new image is called a feature map: an interpretation of the information contained in the original image. As there are n filters, the matrix of weights for the first convolutional layer is $5 \times 5 \times n$. The layer output is a stack of n images. Convolutional layers are often immediately followed by max-pooling layers. These have no weights; they simply downsize the images by retaining the pixel with the highest value in every 3×3 patch (for example) of the image. The result is a stack of n smaller images. In the next convolutional layer, of width p, the stack of n images is copied p times. The filters are three-dimensional, $5 \times 5 \times n$, as they multiply aligned sub-regions of the n images simultaneously, then sum and nonlinearly transform the outcome into one number. The full weight matrix for this layer is $5 \times 5 \times n \times p$. Again, one filter produces one feature map and the output of the layer is p feature maps, further reduced in size by the next max-pooling layer. Eventually the feature maps become small enough that the pixels can be unfurled into a line of classic neurons. From then on, subsequent layers are ordinary dense layers and are

made progressively smaller until the last layer, which contains as many neurons as there are classes. This is the classifier part of the net; the final feature maps have been computed and now the features are examined and the class of the sample is diagnosed. Dense layers require significantly more weights and computations than the convolutional layers, hence their name. A neuron in a dense layer proceeds from every bit of available information, whereas the filters in convolutional layers are selective and focus on small areas of the feature maps. Image analysis is a computationally light process.

The above structure may be tripled to accomodate the three RGB components of color pictures. As spectrograms are black and white, authors that reused networks designed for RGB simply triplicated the spectrograms and sent the copies to the three channels, sometimes with augmentation tactics such as variations in brightness or contrast (Lasseck [47]).

Because the main limitation of DNNs is the scarcity of tagged data available for training, data augmentation is a key companion of DNNs. Great representational power demands a large number of layers and therefore weights; but if there is insufficient data to learn these weights, the network overfits. Data augmentation consists in artificially increasing the size of the training set by modifying the original images in a way that does not compromise their meaning. This could be controlled shifts in time and frequency, slight distortions of the time and frequency scales, removal of random pixels, addition of noise, etc. Other transformations common in image analysis, such as rotations, are not permissible for sound. The purpose is to teach the network basic invariants, for example the fact that a call may occur at any moment in time and it will not change the species. To an extent invariance along the time and frequency direction is already hard-coded into the network, as the same filters slide over the full image and the max-pooling layer retains only the best results locally.

Work on the ImageNet database of 1.28 million images has pushed DNNs further ahead. “The most straightforward way of improving the performance of deep neural networks is by increasing their size” (Szegedy et al. [79]), and consequently their representational capabilities. The challenge is to be able to train such networks, i.e. 1) to propagate decaying information through a large number of layers, 2) to maintain a reasonable number of weights and 3) to avoid overfitting as the representational capability of the network starts to exceed the quantity of data available for consideration. Table 2.2 describes a few of the landmark models.

Table 2.2: Deep Neural Networks for Image Recognition

Authors	Name	Layers	Weights	Performance
Krizhevsky et al. (2012)	AlexNet	8	60 mil.	Top-5 error rate 15.3%[a]
Simonyan & Zisserman (2015)	VGG	19	144 mil.	Top-5 error rate 6.8%
Szegedy et al. (2014)[b]	Inception v3	22	7 mil.	Top-5 error rate 6.7%
He et al. (2015)	ResNet	152	1.7 mil.	Top-5 error rate 3.6%
Huang et al. (2018)	DenseNet	264[c]	25 mil.	Top-5 error rate 5.3%

[a]Percentage of samples with no good identification among the top 5.
[b]The authors have further developed Inception, also known as GoogLeNet, in other publications.
[c]In this thesis we used the 169-layer version.

AlexNet (Krizhevsky et al. [44]) was the first implementation of GPU computation and batch normalization[24]. Then the authors of Inception (Szegedy et al. [79]) focused on the reduction of the number of weights. On the premise that an ideal architecture would be sparsely connected (not all features are useful to recognize all classes) but that large sparse matrices are not efficient from a computational standpoint, they worked to replace classic convolutional layers by successions of lighter convolutions. For example, a convolution with a $3 \times 3 \times n$ filter ($9n$ weights) can be replaced by a convolution by a $1 \times 1 \times n$ filter followed by a convolution by a 3×3 filter ($9 + n$ weights). The partial representation of Inception in Fig. 2.7 also shows that there are several convolutional layers running in parallel (blue and red blocks) on the same feature maps. Some are with 1×1 filters, some with 3×3 filters, some with 5×5 filters and some are max-pooling. The network simultaneously looks for patterns of different sizes in the feature maps.

ResNet (He et al. [34]) and DenseNet (Huang et al. [36]) considered the difficulty of propagating information all the way through a deep network. Note that with only 22 layers, Inception already required side injections of information during training (circled in Fig. 2.7; the class targets were sup-

[24]Details further ahead.

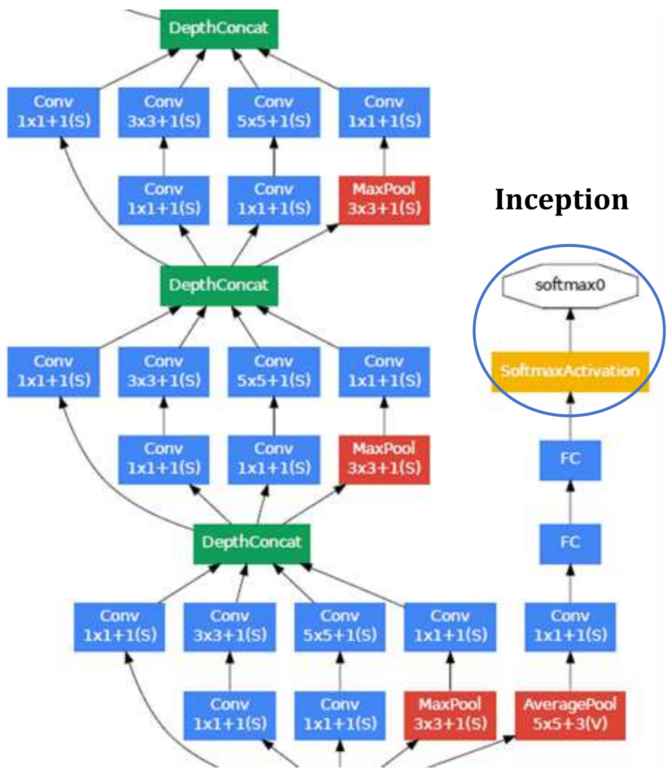


Figure 2.7: A Section of the Inception Deep Net from Szegedy et al. [79]

The network runs from the bottom toward the top. Boxes on the same horizontal line correspond to operations running in parallel, e.g. convolutions with different filter sizes and max-pooling. The specificity of Inception is the decomposition of complete convolutions into two successive simpler convolutions. Hence the parallel operations run on two successive horizontal lines. Then the "depth concatenation" box agglomerates the feature maps from parallel operations back into a single set. At two intermediate positions in the network, a side access to a softmax output, i.e. to the class information, was implemented to facilitate training.

plied to lower layers). The so-called *vanishing gradients* are a well-known manifestation of these depth issues. Through backpropagation, the gradients most removed from the cost function tend to become very small, if not zero, because they are the multiplication of successive partial derivatives, all small and getting smaller with every layer. Training the weights in the remote layers is a challenge when the prescribed weight changes are invariably minimal. Researchers have solved this problem by normalizing the neuron activities at intermediate layers. This operation, *batch normalization*, brings the neuron values up.

However, for a while, adding layers to a network still meant degrading the accuracy. Eventually He et al. [34] cleared the obstacle by complementing the output of convolutional layers with the original feature maps (see Fig. 2.8). This allowed a reformulation of the optimization problem into an optimization of residuals, i.e. a difference between feature maps and their predecessors, hence the name ResNet. The residuals problem is an easier mathematical problem and allowed the training of a 152-layer network which outperformed its predecessors[25]. Huang et al. [36] interpreted the same issue as loss of information. Every layer generates new feature maps that serve the next layer, but the information that was in previous feature maps, once interpreted, is lost. The later layers do not have access to all the existing information to make their decisions. The authors then passed all available feature maps, old and new, to the next layer within a block of layers

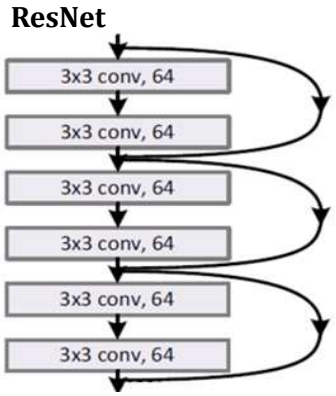


Figure 2.8: A Detail in ResNet from He et al. [34]

The network runs from top to bottom. Here is a succession of convolutional layers with 3×3 filters. Every two layers, the feature maps from two layers up are called back and subtracted from the current feature maps. The net further conveys information in the form of residuals.

[25] A fuller explanation: there is a trivial solution to the weight optimization problem; starting from a network with a given number of layers, one can add layers that operate like the identity function. In this way, the deep net cannot perform less well than the shallower net. The difficulty is to train networks to operate like the identity function. The solution is to train not for the identity $f(x) = x$ but for a residual $h(x) = f(x) - x$; it is indeed simpler for most mathematical systems to converge toward zero than toward another objective. Zero is the default solution pre-training. In other words, the default solution for the deep net is to function as well as its shallower predecessor.

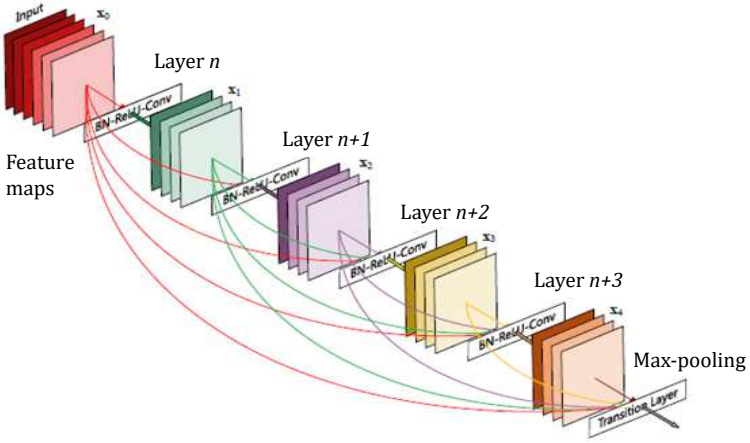


Figure 2.9: A Block of Densely Connected Layers in DenseNet from Huang et al. [36]

The network runs from left to right. The image represents a "Dense Block", i.e. a succession of several convolutional layers and a transition layer. The convolutions are preceded by batch normalization (BN) and followed by the rectified linear unit function (ReLU). The input of a convolutional layer is every feature map entering the layer block and generated in previous layers. The transition layer typically involves max-pooling.

where the size of feature maps remained the same (Fig. 2.9). This is how DenseNet could be trained with 264 layers.

The legacy networks from Table 2.2 are powerful image analysis tools. They know what to look for in an image submitted to them. They have learnt form, color, patterns. Also, all are publicly available as Pytorch[26] implementations. The consequence for audio analysis is the following: instead of training from scratch a model to analyze spectrograms, it is more efficient to restart from the legacy models, who know how to handle images, and to further their training to have them learn the spectrogram problem specifically. This is particularly convenient knowing that bioacousticians do not have training sets available that could compare in size to ImageNet. As Joly et al. [38] reports, in 2018 this approach has become the norm.

Lasseck [47] recently applied it on a training set of 36496 single-label and 381 multi-label recordings from 1500 South America bird species (4–

[26]https://pytorch.org/

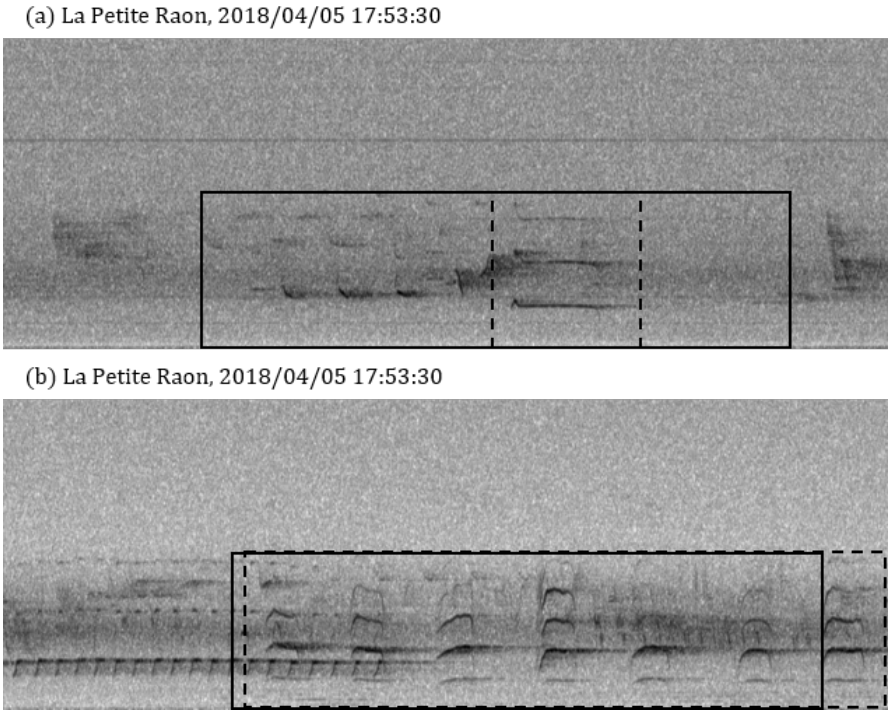


Figure 2.10: Trade-Offs in Image Size Choices

Solid rectangle: 5 seconds, 0–6000 Hz.
(a) Dashed rectangle: *D. martius* contact call.
(b) Dashed rectangle: *D. medius* call, overlapped with *D. martius* territorial call.

160 calls/species). The purpose was to identify the foreground species in a subset of the data (Task 1: MRR 82.7%) and to annotate long-duration soundscapes (Task 2: mAP 19.3%). The base net was Inception v3, although others were tried out, including DenseNet and ResNet. Regardless of the duration of recordings, Inception accepts 299×299 pixel input images; the inputs were resized accordingly. The recordings were processed in random chunks of 5 seconds (*crops*), with consistent bandwidth, frame and FFT parameters in order to maintain the relative proportions of patterns when resizing the images. The process of taking crops from a recording is illustrated in Fig. 2.10; the examples depict calls from woodpeckers *Dryocopus martius* and *Dendrocoptes medius*.

Mel spectrograms were used and performed better than linear frequency scale spectrograms (Fig. 2.11). A range of augmentation techniques were used: adding background noise to the images, adding chunks from other recordings of the same class, skewing or stretching intervals of the data in time and in frequency, applying a cyclic shift to the data, randomly dropping time intervals, randomly omitting elements of the calls, shifting the spectrogram in the frequency direction, etc. Although mostly unphysical (Fig. 2.12;

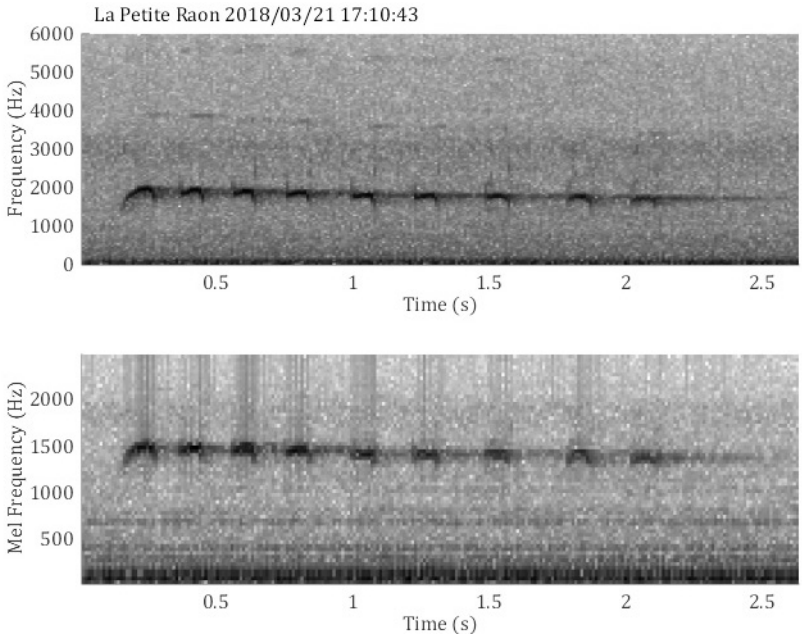


Figure 2.11: Linear Frequency Scale and Mel Scale Spectrograms

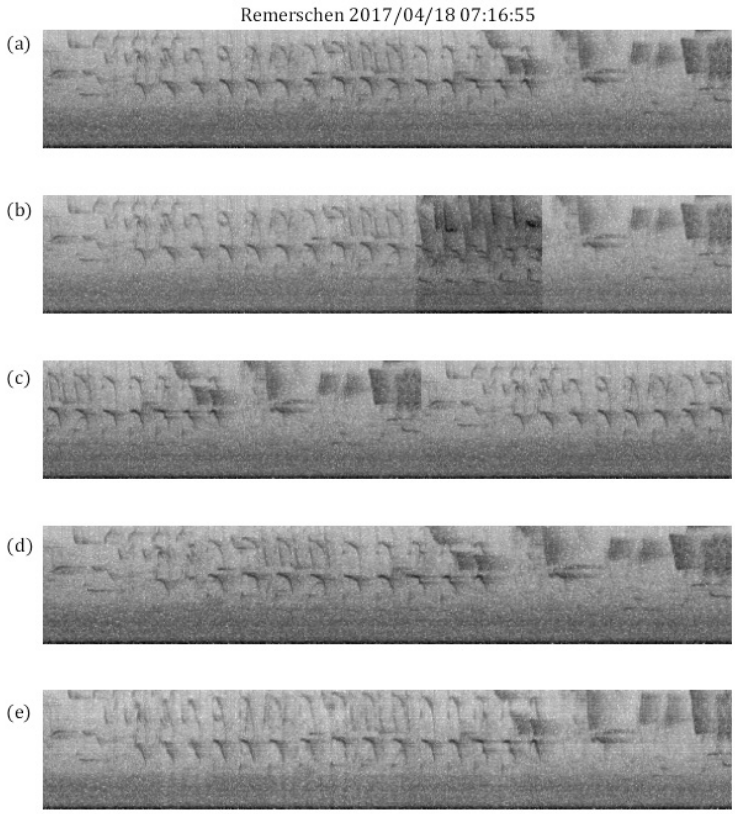


Figure 2.12: Augmentation Strategies

(a) Original *J. torquilla* call (b) Addition of a section from another *J. torquilla* call (c) Cyclic shifting (d) Omission of a time section (e) Time and frequency sections are skewed or stretched by no more than 10%.

with a call from woodpecker *Jynx torquilla*), these measures were responsible for a 10% score increase. The addition of background noise was the most successful step, followed by deformed spectrograms and incomplete spectrograms with time intervals missing. Sections of the spectrograms were sketched or skewed by no more than 10%.

On the other hand, the scores on soundscape analysis have stalled at around 15% (Joly et al. [38]). The 19.3% score in Lasseck [47] is a jump ahead, yet still promises 80% of false identifications in average. Unfortunately, this is the task that most resembles the desired end application: self-annotating audio streams. A gap seems to remain between the algorithm operating in a development setting (single-label short files) and in an application setting (multi-label long audio files). According to Stowell & Plumbley [74], the reason is that there are no true single-label recordings to train the nets on. The unlabeled background calls confuse the algorithm.

2.6 Contributions from the Present Thesis

The present work started as an examination of the two founding hypotheses of earlier classification studies: 1) it is possible to define a set of acoustic features that will grasp the species traits in the vocalizations of any species; 2) these features are not necessarily related to the biological features birds use in their daily lives to know their kins. Consequently, the species traits in birdsong do not need to be studied in depth to make the algorithms work.

In 2015, the first hypothesis was suffering a few cracks. Lasseck [46] reselected subgroups of features for each species to significantly improve his classification score. He also indicated that his feature extraction process did not render temporal structures and repetition rates, which was a problematic drawback. In 2018, on the contrary, DNNs have come close to validating 1). There is no information reported on species resisting identification by DNN. Naturally though, the literature tends to focus on successes rather than failures and thus in certain aspects remains uninformative. Let us consider for example that datasets are commonly augmented by varying the temporal structures and repetition rates. Would that not be problematic for some species? And how do DNNs resolve species when the borders between them are not clear-cut? We have seen that some species traits are quite relative, depending on the region and on other species in contact. Even with the right quantity of data at disposal, it might not be possible to segregate

the species further than at a limited regional scale. In 2015, Lasseck [46] integrated the Xeno-Canto metadata, the most complete to date (date, location, author), to his analysis for modest gains (MAP +3.2%, AUC +0.1%).

The second hypothesis seems irrelevant in 2018. Authors have taken a step back from using the speech analysis features that were likely unfit for birds. The spectrogram is one of the lowest-level processing of sound. The bird auditory system, like the human one, intakes sound in a similar form. In that sense, the DNNs operate with the same signal as birds. There is no over-interpretation of the audio signal and no bypass of biological meaning[27].

As it appears, the remaining *terra incognita* relates to the first hypothesis. A cloud of uncertainty remains on the performances of DNNs in real-life applications. For example, unphysical data augmentation raises questions. At what point does it become problematic that a species is identified from an image where it is represented singing at the wrong frequency and with the wrong silence intervals between syllables? Would the nets recognize a spectrogram that has been flipped left-right or up-down? Image nets have learnt to consider this modification as an invariant (looking left or right, a cat is still a cat) but sounds would definitely be unrecognizable with the same treatment. Does this kind of flexibility (Fig. 2.12) cope well with the small margins that exist between species in the real world (Fig. 2.13)?

The present thesis proposes to have an in-depth look at species detection and recognition in an attempt to shed some light on the above questions. The work will focus on a manageable subset of species, the European woodpeckers, for which it will be possible to examine the specifics in detail. The woodpeckers, not being passerines, have simple, innate calls, although their most famous acoustic signal is their drumming on trees. Drumming is nothing but a time signal and as it will turn out, some of its characteristics are not well rendered on spectrograms. In truth, the design space for drumming is so archaic that its analysis does not warrant using complex classifiers. It is also archaic enough that without context, it barely allows identifying the species. The woodpecker calls are better candidates for DNNs, and will offer us the chance of a look under the hood. Then we might come to form an opinion on whether it is technically feasible to monitor woodpeckers at a large scale in Europe using live audio recordings.

[27] Which brings up an interesting question: do the features designed by the DNNs in their lower layers correspond to higher-level biological features used by birds?

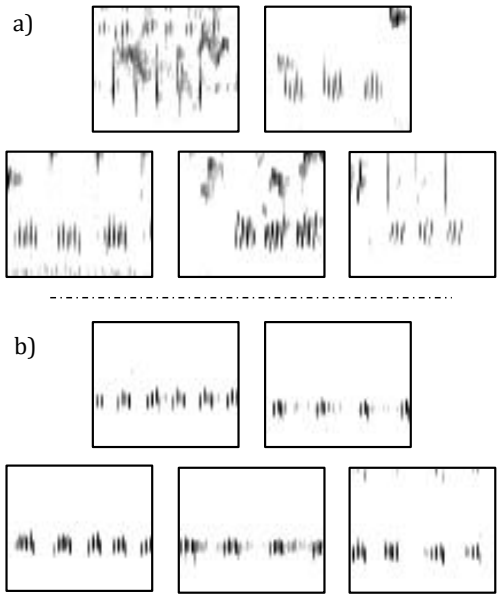


Figure 2.13: Images from the *D. martius* Flight Call

(a) Not the *D. martius* flight call; (b) the actual *D. martius* flight call.

Chapter 3

Woodpecker Sounds

Winkler & Short [90] wrote a 109-page document on the acoustical signals of a subset of 33 woodpecker species, which is to say that the subject is vast. In this chapter, we focus on a description of the sounds and behaviors that will allow or facilitate the detection and identification of woodpeckers in Europe. We also introduce the recordings that we used throughout our analyses.

3.1 European Woodpeckers

The Picidae family, in the order of the Piciformes, counts three hundred species of woodpeckers. The European continent has 11 species, under 7 genera, presented in Table 3.1, Fig. 3.1 and Fig. 3.2. *Dryobates minor* and *J. torquilla* are the smallest, respectively 14–16 cm and 16–19 cm in length. *D. martius* is the largest with a length of 45–50 cm[1]. Hybrids exist between *D. major* and *Dendrocopos syriacus* (Michalczuk et al. [53]), between *P. canus* and *P. viridis* (Schmitz [69]; Ławicki et al. [48]) and between *P. viridis* and *Picus sharpei* (Pons et al. [64]). Until 2011, *P. sharpei* was considered a subspecies of *P. viridis* (Perktas et al. [58]).

The reasons that make woodpeckers an interesting target for acoustic monitoring are plenty. They are valued as ecosystem keystones (Gorman [31]) and indicators of forest health (Mikusinski & Angelstam [54]). Some species are targeted by regional conservation programs, e.g. *D. medius* in Sweden (Pettersson [60]) or *P. canus* in Belgium. Most importantly for our purpose,

[1]The length estimations are from Gorman [31].

Figure 3.1: European Woodpeckers (Adapted from Peterson [59])

Table 3.1: European Woodpecker Species

Genus	**Species**[a,b]	**Vernacular**	**French**	**Area**
Jynx	*torquilla*	Eurasian wryneck	Torcol	Migrant, open habitat
Picoides	***tridactylus***	Eurasian three-toed woodpecker	Pic tridactyle	Not in BE, FR except Alps
Dendrocoptes[c]	***medius***	Middle spotted woodpecker	Pic mar	Mature oak woodlands
Dryobates[c]	*minor*	Lesser spotted woodpecker	Pic épeichette	Mature woodlands
Dendrocopos	***syriacus***	Syrian woodpecker	Pic syriaque	Eastern Europe
	major	Great spotted woodpecker	Pic épeiche	Common
	leucotos	White-backed woodpecker	Pic à dos blanc	Not in BE, FR except Alps/Pyr.
Dryocopus	***martius***	Black woodpecker	Pic noir	Common
Picus	*viridis*	Eurasian green woodpecker	Pic vert (pivert)	Common
	sharpei[d]	Iberian woodpecker	Pic de Sharpe	Iberian peninsula
	canus	Grey-headed woodpecker	Pic cendré	Mainly old forest, rare in BE

[a] Following the order in the list of the International Ornithologist's Union, posted at http://www.worldbirdnames.org/.

[b] In bold, species listed under the Annex I of the EU Birds Directive (threatened species). Two subspecies of *D. major* found only in the Canary Islands are also listed.

[c] Formerly *Dendrocopos*, reclassified after Fuchs & Pons [27].

[d] Formerly a subspecies of *P. viridis*, reclassified after Perktas et al. [58].

Dendrocopos major *Picus canus* *Dryocopus martius*

Figure 3.2: European Woodpeckers in the Flesh

woodpeckers have relatively simple vocalizations. Some of their calls are iconic landmarks in forest soundscapes, but they most distinctively reveal their presence by drumming on trees. Drumming is a form of long-distance communication. It is used for territory marking and mate attraction (Zabka [91]; Tremain et al. [83]), which implies that drums carry the species and individual information to some extent (Zabka [91]; Stark et al. [73]; Dodenhoff et al. [15])[2]. Thus, there is a functional overlap with advertising calls. Some species drum abundantly and forego advertising calls, others do not drum at all. Monitoring woodpeckers is a two-sided problem: on the one hand the calls, on the other hand the drums.

J. torquilla, *D. minor*, *D. medius*, *D. major*, *D. martius*, *P. viridis* and *P. canus* are present in Belgium. *P. canus* was never abundant, as Belgium sits on the edge of its distribution area (Schmitz [69]). Across the European continent, the species is declining : -21% between 1990 and 2009 (Sordello [72]). The decline is most pronounced to the North of the distribution area, i.e. in Belgium. Testaert [80] mentions 10-15 pairs in 1998 but according to Aves[3], only three or four males remain today. Most Belgian observations occurred in the Hertogenwald to the East and in the Lorraine region to the South-East of Wallonia. The preferred habitat is in the heart of old forests, where the trees are at least 50 years old and some are standing dead (Sordello [72]). In 2014, Aves launched a special monitoring program for *P. canus* in Wallonia[4].

[2]Stark et al. [73] and Dodenhoff et al. [15] actually debate whether the drums convey the species information.

[3]Aves is the organization that monitors bird populations in Wallonia.

[4]"Rechercher le Pic cendré en Wallonie, programme d'inventaire et surveillance des

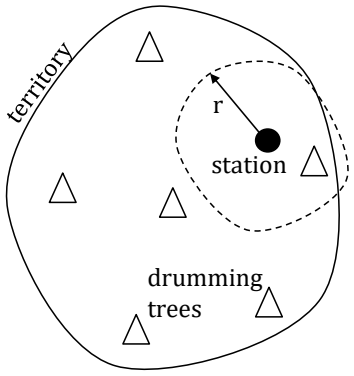


Figure 3.3
Station Range *r* (unknown)
and Woodpecker Territory

Multiple field visits starting in mid-March turned out fruitless. There was no response to playback either. This prompted new questions about the specific difficulties in monitoring on the distribution edge. A. de Broyer (Aves) hypothesized that the *P. canus* pairs reunite with their partner of the year before in February and stop vocalizing beyond that date. For that reason, the detection of pairs and nesting proves particularly problematic.

Because of this local context, *P. canus* receives special attention in the present thesis. *P. canus* is a conspicuous bird (Sordello [72]) and hence a fitting subject for continuous audio monitoring. Its range in the spring is 1–2 km^2 (Sordello [72]). During that time, *P. canus* flies around its territory from one drumming spot to the next, i.e. trees it previously identified as good substrates, often dead or hollow. The reach of recording stations was never substantiated but is thought to be comparable to human ears. For reference, in bird surveys it is commonly accepted that ornithologists detect birds up to 100 m away, which is less than the *P. canus* territory size by an order of magnitude. The success of audio monitoring using a fixed station is contingent on *P. canus* patrolling its territory and eventually drumming on a tree within the range of the station, as schematized in Figure 3.3.

3.2 Drumming in European Woodpeckers

Drumming is easily spotted in spectrograms. Fig. 3.4 shows three *P. canus* samples: the first drum of the season, tentative and incomplete (a), then a

oiseaux nicheurs de Wallonie", at:
http://www.aves.be/fileadmin/Aves/Bulletins/Articles/48_1/Rechercher_le_Pic-cendre_en_Wallonie_fev2014.pdf.

full drum recorded at close range, with content up to 6000 Hz and beyond (b), and finally a distant drum of which only the main frequency is left after propagation and absorption through the forest (c).

The European woodpeckers can be split into three groups depending on their use of drumming (see Table 3.2). The drummers (*Dendrocopos leucotos*, *D. major*, *D. syriacus* and *Picoides tridactylus*) drum abundantly and seem to not possess an advertising call. The versatile species (*D. minor*, *D. martius* and *P. canus*) both drum and call. The recordings in Section 3.4.3 contain several examples of this versatility. In Tenneville, on the day when a *D. martius* invaded the *P. canus* territory, they fought by both drumming and calling at each other. The *D. martius* individual used three different calls on top of his drums. A fourth one, his long-distance contact call, was heard on a different day. Roughly 2000 *P. canus* drum rolls were eventually retrieved from the Tenneville recordings, but *P. canus* was never heard drumming in Remerschen[5].

The final group of woodpeckers, *D. medius*, *J. torquilla*, *P. sharpei* and *P. viridis*, are the species that do not truly drum. Drumming in *D. medius* was in turn defended (Wallschläger [88]) and debunked (Turner [84]). Drumming was recorded in *J. torquilla*, *P. sharpei* and *P. viridis*, but is rare. In reality, the drumming in these species is a sub-type called *soft drumming*, which is quieter than *territorial drumming* and has a different function. Soft drumming achieves intimate communication within the pair.

As implied in Table 3.2, woodpeckers produce an array of sounds with their bills. The shocks from foraging and hole digging are not intended as communication, but there are two categories of signals which exhibit structure and meaning: drumming and tapping[6].

[5]See Section 3.4.3 for a description of these recording campaigns.

[6]The bill signals of woodpeckers were birthed through *exaptation*: a shift in the function of a trait during evolution (Gould & Vrba [32]). For woodpeckers, the original function of their peculiar bill was to dig for food in dead trees, or to dig cavities. Sound was simply a byproduct of these acts. But then, out of opportunism, woodpeckers have turned it into a tool for communication, first creating loose signals (tapping), then eventually faster, louder and more structured sequences (drumming). The most ancestral woodpecker species use the rawest signals: uneven taps that come in series of small bouts (N. Mathevon, personal communication). Exaptation shows in Table 3.2, which summarizes which signals the different European species produce. All have come up with tapping; *D. medius* went as far as the tap/drum mixed signals but never developed drumming; *J. torquilla*, *P. sharpei* and *P. viridis* stopped at soft drumming. One could actually question whether the latter species never developed territorial drumming or gave up on it, choosing to rely on their calls instead. A

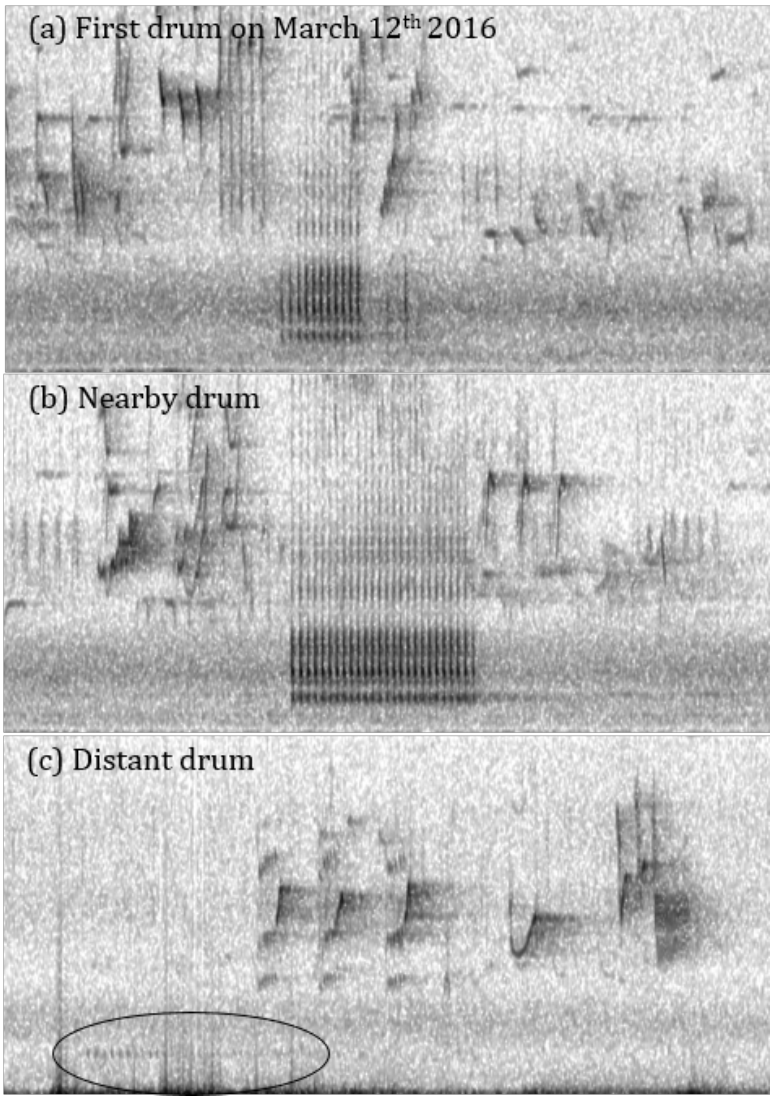


Figure 3.4: Drumming Recorded in Tenneville, BE (Section 3.4.3)

Bandwidth 0–6 kHz. Spectrogram durations ∼5.5 s.

Table 3.2: Summary of Woodpecker Signals (Florentin et al. [23])

Species	**Nest Showing**	**Nest Relief**	**Soft Drums**	**Territorial Drums**	**Advertising Calls**
Drummers					
P. tridactylus	v	v	v	v	-
D. syriacus	v	v	v	v	-
D. major	v	-	v	v	-
D. leucotos	v	-	v	v	-
Versatile					
D. minor	v	-	v	v	v
D. martius	v	v	v	v	v
P. canus	v	-	v	v	v
Vocal					
J. torquilla	v	v	v	-	v
D. medius	v	v	-	-	v
P. viridis	v	nr	v	-	v
P. sharpei	nr	nr	v	-	v

v: exists; -: does not exist; nr: not recorded/unknown

In territorial drumming, both sexes drum. Male-female pairs have synchronized drumming duets during the mating season. Drumming contests also occur between males of different species. The drums are slightly different between male and female, and drumming is potentially affected by geographical and seasonal variations (Zabka [91]). Drumming produces a distinctive pattern on spectrograms: a succession of vertical lines, produced as each strike of the bill excites a range of frequencies (Fig. 3.4). The succession of strikes is fast and for most European species, accelerated.

Tapping is slower than drumming, quieter and less frequent. Its rhythm is mostly even-speed but often irregular. There is barely any characterization of it in the literature. The observations in Winkler & Short [90] suggest that there are at least two functions for tapping: *nest showing* (suggesting a location for a hole) and *nest relief* (to request a changeover in excavating or watching the eggs).

K. Turner's recordings (Section 3.4.2) also contain a small number of intermediate signals between tapping and drumming, such as accelerated

symmetrical question could be asked of the *Dendrocopos* and of *P. tridactylus*: did they ever possess an advertising call that they let go of?

tapping or tap/drum mixes. Eventually, only territorial drumming (and the rare soft drum) is addressed in the present thesis. Territorial drumming is the one woodpecker bill signal that is far-carrying and potentially marked with the species, because these qualities are essential to perform the functions of territorial defense and mate attraction. It is also abundant and heard throughout the spring (see Fig 3.7 in Section 3.4.2). In other words, this is the only signal that an autonomous recording station could reliably detect and recognize.

3.3 Calls of European Woodpeckers

Woodpeckers have a multitude of calls (call notes, rattle, kweek, wicka, chirp, etc.), some rare, some frequent, used singly or in combinations. Winkler & Short [90] drew a tentative list for the pied woodpeckers, which included *D. major*, *D. leucotos*, *D. medius*, *D. minor*, *D. syriacus* and *P. tridactylus*.

The three *Dendrocopos* and *P. tridactylus* abundantly use a *call note*, which serves as a contact call or as a distancing signal (Fig. 3.5); however this does not have the long-distance reach of their drums. The *rattle call* seems more involved in setting up territories, and in many ways it resembles the songs of passerine birds. Other than drumming, the rattle is the major long-distance signal of woodpeckers. It is the most frequently heard call of *D. minor*, often combined with drumming (Winkler & Short [90]). The non-pied woodpeckers (the three *Picus*, *J. torquilla* and *D. martius*) also have a prominent rattle call. Such a call has around 10 notes per call, and the notes last for about 80–150 ms[7]. For *D. medius*, which does not drum, the territorial-sexual role is taken up by the *kweek* call, heard at its peak in the spring. This call has 6 long notes (400-500 ms) in average, is loud, frequency-modulated and rich in overtones. Finally, the vocal repertoire of *D. martius* includes the *kru-kru-kru* flight call, similar in structure to a rattle call, and the wailing *kleee* contact call, which consists of one long note of approximately 0.6–1 seconds (Gorman [31]).

Our study focuses on the long-distance, emblematic and sufficiently available calls: the rattles, the *D. medius* kweek and the flight and contact calls of *D. martius*. Fig. 3.6 shows spectrograms of these nine calls.

[7]Syllable durations are estimated on calls retrieved from Xeno-Canto, discussed in Section 3.4.1.

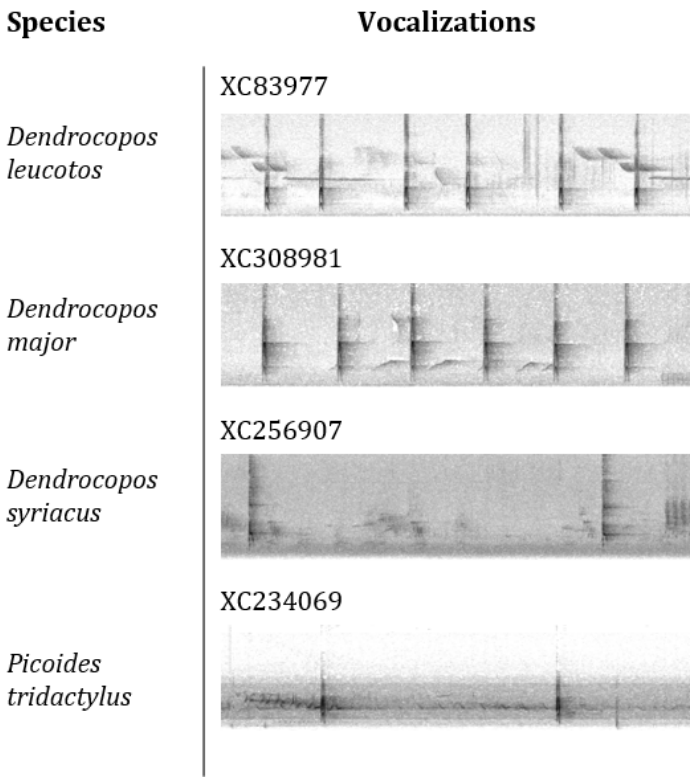


Figure 3.5: Call Notes of European Woodpeckers

Bandwidth 0–12 kHz. Spectrogram durations 5 s (all). The typical duration of a call note is 50 ms[7]. The files contain the following number of call notes: 15 notes (*D. leucotos*), 21 notes (*D. major*), 3 notes (*D. syriacus*) and 3 notes followed by tapping ((*P. tridactylus*). [Note: the spectrograms only show a part of the recordings.] The XC code is the file identification from Xeno-Canto – see Section 3.4.1.

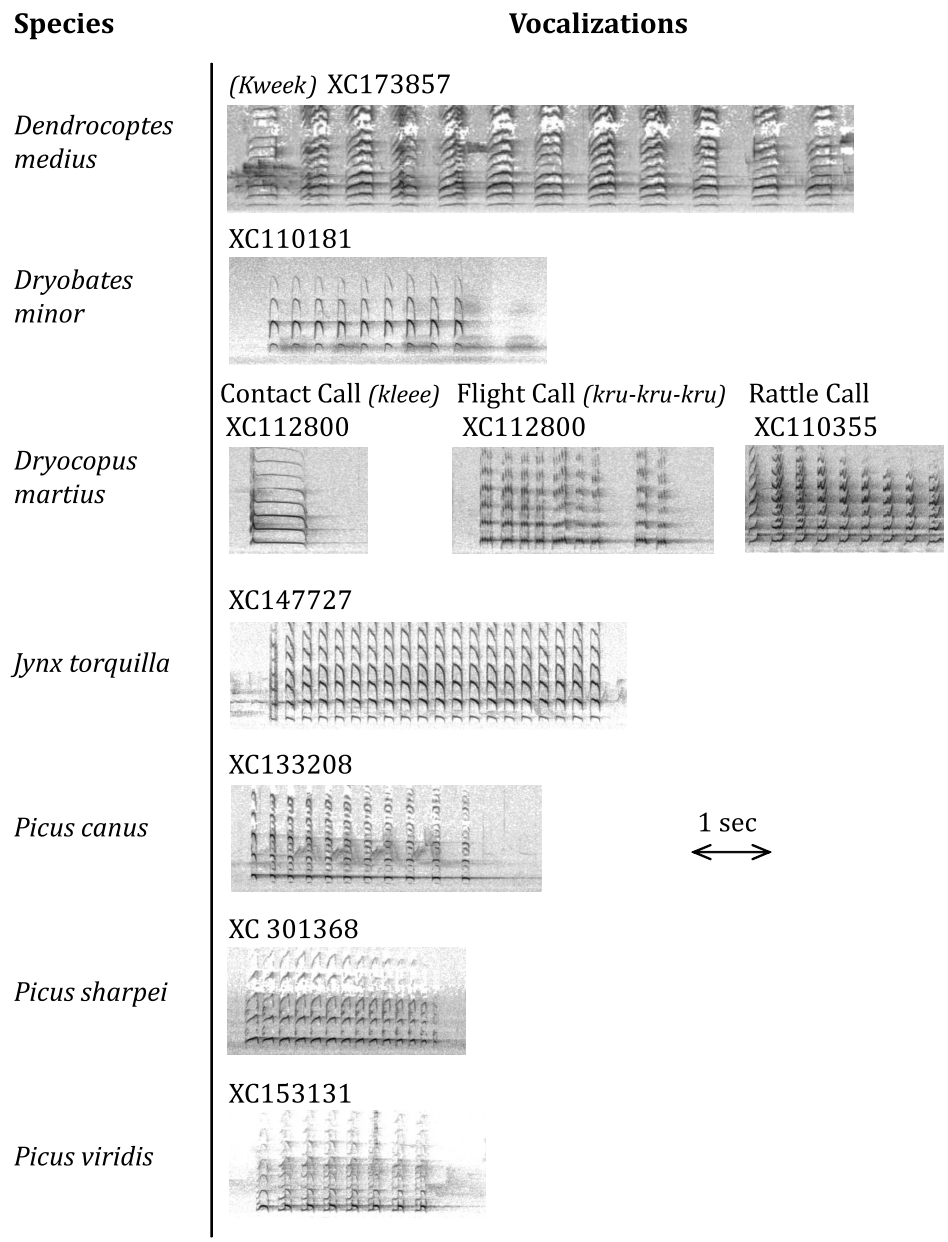


Figure 3.6: Target Vocalizations for the Present Thesis

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 [31]).

3.4 Available Recordings

The recordings in the present thesis were obtained from the sources listed in Table 3.3. Datasets sampling all species were gathered from the online archives Xeno-Canto[8], a major resource for bioacousticians, and Tierstimmen[9]. British birder Kyle Turner kindly shared his private collection. Last, continuous field recordings were acquired in Tenneville, Belgium, Remerschen, Luxembourg and La Petite Raon, France (map in Fig. 3.8). The datasets are subsequently referred to by their code, e.g. XC for Xeno-Canto. The identification number for XC files is the one used by the website. A file with an additional trailing index is a segment from the original recording (e.g. XC171084_21).

The original files from Xeno-Canto, Tierstimmen and Kyle Turner last up to a few minutes. The sampling frequency is either 44.1 kHz or 48 kHz. The online archives use mp3 compression. Some files were low-pass filtered or edited by the recordists. The sound quality of files in Xeno-Canto is rated A to E. We favored the A-quality files, where the background noise is low.

Table 3.3: Recordings Available for the Present Thesis

Collection Name	**Code**	**Type of Collection**	**Content**
Xeno-Canto	XC	Online archive, crowd-sourced	400,000 recordings, 10,000 bird species
Tierstimmen	TS	Online archive	120,000 recordings, incl. 1,800 bird sp.
Kyle Turner	KT	Personal collection	11 Europ. Woodpeckers, 27 GB
Tenneville (BE)	TN	Continuous recording	8 GB
Remerschen (Lux.)	RM	Continuous recording	128 GB
La Petite Raon (FR)	LPR	Continuous recording	118 GB

[8]The Xeno-Canto Foundation, https://www.xeno-canto.org/
[9]Museum fur Naturkunde Berlin, http://www.animalsoundarchive.org/

3.4.1 Recordings from Xeno-Canto and Tierstimmen

Drums

The composition of our drums collection is shown in Table 3.4. We retrieved 267 files from Xeno-Canto and 94 from Tierstimmen for nine species. In total, they amount to two hours and 20 minutes of recordings (4 GB). We then extracted 324 drum rolls from the TS data and 2342 from the XC data[10]. The few drums collected for *P. viridis* are evidently soft drums. The drums labeled *D. medius* in the archives are dubious (Turner [84]) and were eventually discarded. The dataset is skewed toward the most common species (*D. major*, *D. minor*) and species without an advertising call (*D. leucotos*, *P. tridactylus*). *D. syriacus* is only present in Eastern Europe where fewer recordings are available.

Table 3.4: Xeno-Canto / Tierstimmen Drumming Database

Species	**Number of Original Files**	**Number of Drumming Rolls**
D. leucotos	43	248
D. major	115	818
D. martius	27	84
D. medius[a]	3	8
D. minor	67	832
D. syriacus	3	8
P. canus	29	104
P. tridactylus	68	547
P. viridis	6	16
TOTAL	**361**	**2665**

[a]Dubious *D. medius* labels. These drums were eventually discarded.

[10]See Chap. 4 for methodology.

Calls

The composition of our calls database is shown in Table 3.5. No calls were sourced from Tierstimmen. Instead, the database was completed with recordings from K. Turner's collection. The recordings of *P. sharpei* are scarce. The high number of *J. torquilla* calls is due to XC177894 yielding 276 calls from a single pair. The Xeno-Canto files are more diverse in that they were gathered from multiple recordists. Kyle Turner, on the other hand, might have had access to fewer birds.

Table 3.5: Corpus of Woodpecker Vocalizations

Species		**Number of XC Calls**	**Number of KT Calls**	**Total**
D. martius	Ad.	90	30	120
	Flight	73	16	89
	Contact	103	51	154
D. medius		120	48	168
D. minor		66	105	171
J. torquilla		572	56	628
P. canus		175	33	208
P. sharpei		29	6	35
P. viridis		213	50	263
TOTAL		**1441**	**395**	**1836**

3.4.2 Recordings by Kyle Turner

Kyle Turner has been recording woodpeckers since 2002. His collection covers all European species, with recordings from France, Hungary, Slovakia, Spain and England. His purpose is to document the various behaviors associated with drumming and tapping. The recorded material is the foundation for Table 3.2. For each species, the signals for the subcategories territorial drumming, soft drumming, nest relief and nest showing were shown to be significantly different (Florentin et al. [23])[11].

Fig. 3.7 shows the spread of signal categories versus recording dates. The

[11] Detailed results unpublished to date.

recording season starts at the end of February, hence the absence of data in January and February. The limited data posterior to June is not shown. This plot was also redone with a larger body of territorial drums, which brought no difference.

Outside the breeding season, woodpeckers do not remain in pairs. They compete with each other for most of the year. In winter they begin to drum loudly in order to claim a territory for breeding and to attract a mate. Blume & Tiefenbach [6] show low to moderate drumming by *D. major* in February through to mid-March, peaking in early April before dropping again through May and stopping by mid-June (in Germany). Fig. 3.7 (all species, all locations) is consistent with Blume & Tiefenbach: the territorial drums peak at the end of March and in April.

Nest showing peaks in March, before territorial drumming, and is fol-

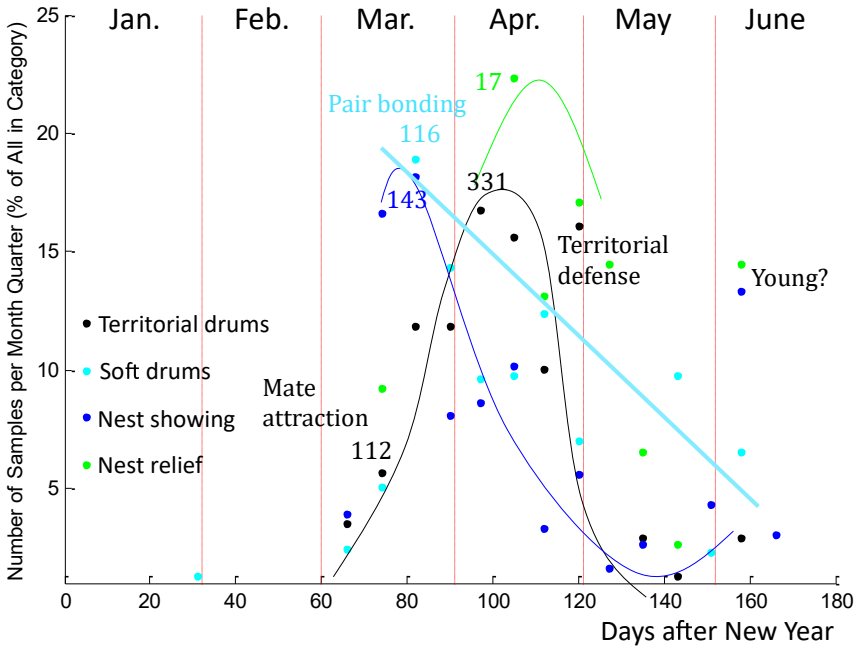


Figure 3.7: Recording Dates of Drumming and Tapping Samples

The number of samples is aggregated for every quarter of the month and normalized by dividing by the total number of samples in each category. The trend lines are drawn by hand to guide interpretation. Numbers annotated on the plot are the unscaled figures.

lowed by nest relief in April. The pair first selects a location, then claims the territory and takes turns digging the nest and brooding the eggs. This calls into question the true contribution of territorial drumming to pair formation: the territorial drumming peak comes too late, after the selection of a nest location. At the beginning of March, the pairs are already formed. Hence territorial drumming appears foremost as a distancing signal.

Soft drumming is present throughout the spring, yet diminishes. The assumed function is pair-bonding, or seduction. As the breeding season ends, the need for close communication within the pair vanishes. We note that pair-bonding is also achieved through territorial drumming duets.

The anomalous-looking signals from June might be from the young. Late bursts are another possibility; Blume [6] mentions such late territorial drumming in *D. martius* in August.

3.4.3 Tenneville, Remerschen, La Petite Raon

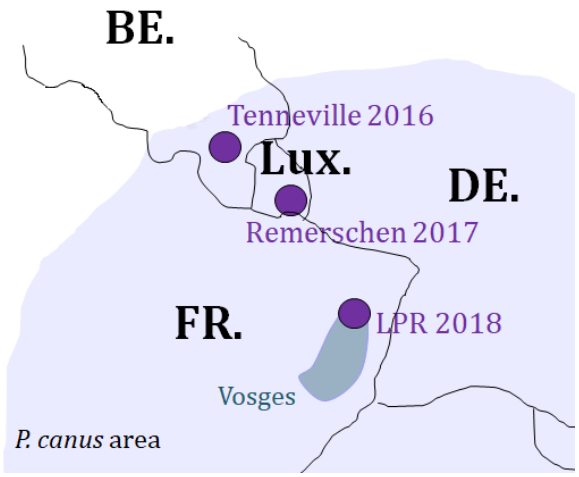


Figure 3.8: Area Map with Station Positions

In 2016, we built an autonomous station to acquire recordings of woodpeckers in the wild[12]. The purpose was to test the performance of species detection and identification programs in a realistic setting, and to shed some light on the drumming and calling habits of *P. canus* that could facilitate its

[12]Design details in Appendix B.

Table 3.6: Data Collected with the Recording Station

	Tenneville	**Remerschen**	**La Petite Raon**
Code	TN	RM	LPR
Dates in operation	25/02/2016 - 22/04/2016	01/03/2017 - 20/05/2017[a]	17/02/2018 - 21/05/2018[b]
Hours in operation	435	850	1291
Scanned bandwidth	300–1500 Hz	300–3000 Hz[c]	300–2100 Hz
Stored data			
– in nb. of WAV files	11527	52224	47695
– in hours	96	435	397
– in % of listening time	22%	51%	31%
– in GB	8 GB	128 GB	118 GB

[a]Fully operational from 15/04/2017.
[b]Interrupted 22/02/2018–04/03/2018.
[c]Eventually brought back to 2100 Hz.

monitoring in the future. The first campaign took place in Tenneville, on the plateau north of Saint-Hubert, Belgium, where Aves had spotted a single individual: the one pictured on Fig. 3.2. The station was deployed to the field from February 25th to April 22nd 2016. Having observed a single individual for the better part of a spring, we then turned our focus on pairs.

In 2017, the station was deployed in the nature reserve at Remerschen, Luxembourg, which is known to host 3 to 4 *P. canus* territories, including breeding pairs. The reserve is partly a sand quarry; a succession of ponds has formed where the ground has been excavated. These wetlands are an important stop for migratory birds and for ornithologists. Interestingly, *P. canus* does not drum in Remerschen[13]. The local trees are willows, soft and impractical for drumming. In Tenneville, where we collected more than 2000 drums in a month and a half, the *P. canus* had a drumming spot on an old

[13]Per Patric Lorgé, resident ornithologist at the Centre Nature et Forêt Biodiversum, Bréicherwee 5, L–5441 Remerschen.

beech, which is hard wood. Thus, in Remerschen, *P. canus* can only be detected through its call. In addition, in that campaign, the exact position of the birds was unknown. The station was installed by a group of older trees that *P. canus* was likely to visit. It operated intermittently from March 1st to April 15th 2017, impaired by software glitches, then continuously from April 15th to May 20th 2017, when the woodpecker season was winding down. 2017 turned out to be a rough year for woodpeckers; P. Lorgé was unable to spot *P. canus* on the reserve.

In 2018, we installed the station in the northernmost stronghold of *P. canus* in France, the Vosges mountains (Sordello [72]). The coniferous forest in La Petite Raon is old, with dead trees left standing[14]. Prior visits to the site had revealed the presence of *D. martius*, *D. major* and *P. canus*. The station stayed from February 17th 2018 to May 21th 2018, with an 11-day interruption starting February 21st. It was positioned next to a stag bearing marks of woodpecker activity.

Photographs of the three sites are in Fig. 3.9. A schematic map showing the locations of the stations and the *P. canus* distribution area is in Fig. 3.8. Additional numbers on the data collected are in Table 3.6. The station stored 30-sec WAV files, but not indiscriminately. It examined ACI values in a chosen bandwidth to select the ones where woodpecker sounds were most probable[15]. The bandwidth for drumming is up to 1500 Hz and for songs, conservatively, up to 3000 Hz.

[14]The dominant species is the Norway spruce, *Picea abies*. *L'épicéa*, in French.
[15]This aspect is further developed in the next chapter.

Figure 3.9: Tenneville (top) - Remerschen (middle) - La Petite Raon (bottom)

3.5 Conclusions

Of the multitude of woodpecker acoustical signals, the ones that are most easily detected are the ones intended to travel long distances, i.e. the advertising signals used to claim a territory or to attract a mate. Depending on the species, these functions are filled by drums, calls or both. Moving forward, our study will consider these two broad categories, which entail ten drumming species and nine calls from seven species. In the context of either territorial conflicts or reproduction, the species information has to be conveyed to the other party; this gives us hope that we will be able to find it in the target acoustic signals and use it to identify the birds. The Xeno-Canto and Tierstimmen online archives, as well as British ornithologist Kyle Turner, supplied us with ample datasets to develop proof-of-concept algorithms to that end. The real-life test will be the analysis of the data collected through three field campaigns in Belgium, Luxemburg and France.

Chapter 4

The Detection of European Woodpeckers in Audio Recordings

We start this chapter with 281 GB of field recordings in our hands and hundreds of files downloaded from on-line archives in which woodpecker signals are scattered. We need to locate the woodpecker drums and calls in this haystack. The desired outcome is a collection of files limited to one drum or one call each. We will reach this point progressively; first we will decide which half-minutes of the recordings are worth processing, then we will segment the audio into short files containing candidate signals and eventually, we will assess whether the sounds belong to woodpeckers or not.

4.1 The ACI, a First-Level Woodpecker Detector

Detection Above a Threshold

The Acoustic Complexity Index (ACI) is a simple indicator that has shown great sensitivity to passerine songs. Its formulation emphasizes the intensity variation from one frame of signal to the next (Eq. 4.1) and thus brings forward all fast-varying sounds. This includes drumming and to an extent woodpecker calls.

$$\text{ACI(f)} = \frac{\sum_{k=1}^{n-1} |I_k(f) - I_{k+1}(f)|}{\sum_{k=1}^{n-1} I_k(f)} \tag{4.1}$$

f is the frequency bin, $I_k(f)$ the acoustic intensity spectrum of the k^{th} frame and n the number of frames in the time interval under consideration.

It follows that the ACI can be used to preselect segments of long audio recordings in which woodpeckers might be present. In this approach, the audio stream is processed in successive 30 s segments[1] and in a target bandwidth. Segments for which the maximum value in the ACI spectrum is greater than a given threshold are selected for further analysis. The target bandwidth is inferred from the target sounds. The threshold as well but not only; ACI values are impacted by the frame duration and the segment duration. Our results stand for 30 s segments and frames of 46 ms[(2)].

We selected a threshold of 1.2 and a bandwidth of 500–1500 Hz for drums and 500–2100 Hz for calls, based on an examination of the samples at our disposal, i.e. 2665 drums and 1836 calls (see Chap. 3, Table 3.4 for drums and Table 3.5 for calls). Most of these samples are less than 30 seconds long and were de facto padded with silence. In this manner, the ACI is not overestimated: padding with background noise or other signals would yield higher ACI values. Regarding the bandwidth, common background noise is circumvented with a low bound in the range 300–500 Hz. The higher bound has different values for drums and for calls. We will see in Chap. 5 that drumming has characteristic low frequency content below 1500 Hz. For the calls, the mean ACI spectra per call class are shown in Fig. 4.3 (XC/KT dataset). All have significant content below 2000 Hz but *D. minor*, which peaks at 2400 Hz. Unfortunately, above 2000 Hz

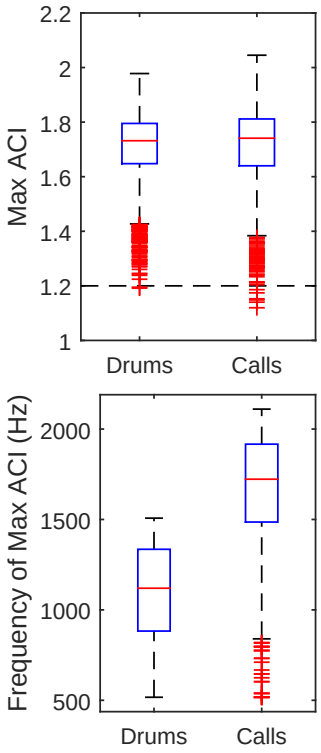


Figure 4.1: Maximum ACI in XC/TS/KT Drums and Calls

The ACI is calculated in the interval 0.5–1.5 kHz (drums) and 0.5–2.1 kHz (calls). The dashed line indicates the 1.2 ACI threshold.
Box plots: the box is bounded by the first and third quartiles of the data. The solid line within the box is the median. The whiskers extend to the values whose distance to the box is at most 1.5 times the height of the box. The crosses are outliers.

[1]Some authors use 60 s or more; see Chap. 2.

[2]Sampling frequency 12 kHz, Fourier transform size 2048 bins; or any numbers that maintain the ratio between these two.

the overlap with passerines is important and the selection capacity of the ACI is diminished. This is the reason why we limited the ACI bandwidth to 500–2100 Hz for calls, in a configuration that is not ideal for *D. minor*.

Fig. 4.1 shows the distribution of maximum ACI values for drums and calls. With a threshold of 1.2, only 2 drums and 6 calls miss the cut (lowest ACI 1.12). The median frequency at which the maximum ACI occurs is 1120 Hz for the drums and 1723 Hz for the calls. The drums with ACI values below 1.2 are from *D. minor*. For this species, it is interesting to look at the impact of the frame duration on the ACI (Fig. 4.2). With shorter frames, the ACI values are all above 1.2 and the distribution is more contained, with fewer outliers. As we will see in Chap. 5, *D. minor* is the fastest drummer; the interval between two successive strikes might be as low as 40 ms. Here the 46 ms frame duration is too high for the ACI to capture the variations in the signal.

Among the six failing calls, one has significant background noise and an overlapping call[3], two were recorded at a distance and in a highly reverberant environment that smooths out the time variations and the last three are *D. minor* and occur at frequencies higher than 2100 Hz.

We preselected recordings using an ACI threshold in all our field experiments (TN, RM, LPR; see Chap. 3 and App. B), primarily to reduce data storage. The ACI allowed us to discard 50–80% of the audio (see Table 3.6 in Chap. 3). In the TN dataset, in which the ACI scanned the 300–1500 Hz interval, 18% of the retained audio files contained drumming rolls. Rain drops were a large subset of the false positives. Starting in April, when the rain receded and bird activity picked up, the percentage climbed to 34% (see Table 4.1).

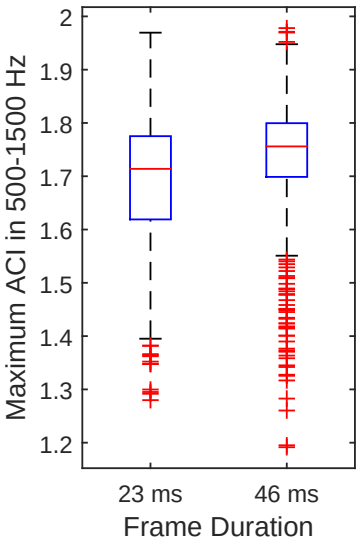


Figure 4.2: Maximum ACI in *D. minor* Calls vs. Frame Duration

[3]In return, for faint drums and calls, it is not impossible that the ACI reaches values above 1.2 *because* of other sounds present in the recordings. However, the XC/TS/KT recordings are more often than not captured at close range.

Table 4.1: Drumming Rolls Detected in Tenneville after April 6th

Date	**Rain**	**Recorded Sound Files**	**Detected Drums**[a]	**Useful Sound Files**
06/04/2016		183	28	15%
07/04/2016	1 h	303	103	34%
08/04/2016	30 min	140	51	36%
11/04/2016		83	3	4%
12/04/2016	3 h	648	285	44%
13/04/2016	1 h	482	175	36%
14/04/2016		544	416	76%
15/04/2016	30 min	241	27	11%
16/04/2016	2 h 30	580	79	14%
17/04/2016	2 h	406	57	14%
18/04/2016		383	112	29%
19/04/2016		384	158	41%
TOTAL		4377	1494	34%

[a]Through repetitions analysis (Section 4.3), which detects fewer drums than other methods described in the present chapter.

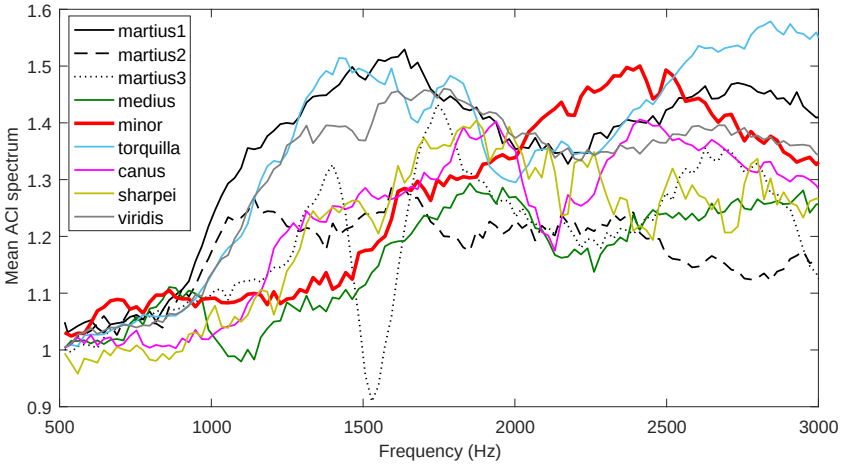


Figure 4.3: Mean ACI Spectrum over XC/KT Calls, by Call Class

Detection from Image Patterns

ACI spectrograms are remarkable tools to review large amounts of audio data at once. For example, the spectrogram in Fig. 4.4 shows a full day of recordings in a compact display. The black vertical patterns starting at 7:00, 7:45, 9:00, 10:00, etc., are occurrences of drumming. In the morning, a woodpecker repeatedly drummed in sessions of roughly half an hour. The stability of the spectral profile indicates that the drumming spot was always the same. The thinner lines at around 13:00 are isolated drums. These patterns cannot be mistaken: few sounds below 2000 Hz have the inherent variability that generates high ACI values on wide bandwidths. Hence the ACI spectrograms enable us to discard hours of recordings from further analysis in one look. Notably periods of rain, a recurring source of difficulty in drumming detection, are easily set aside. They form the opaque black regions starting at 11:30, 15:15, 16:30 and 18:30 in Fig. 4.4[4]. A third of the LPR1 dataset could be set aside in this manner (Table 4.3 in Section 4.2).

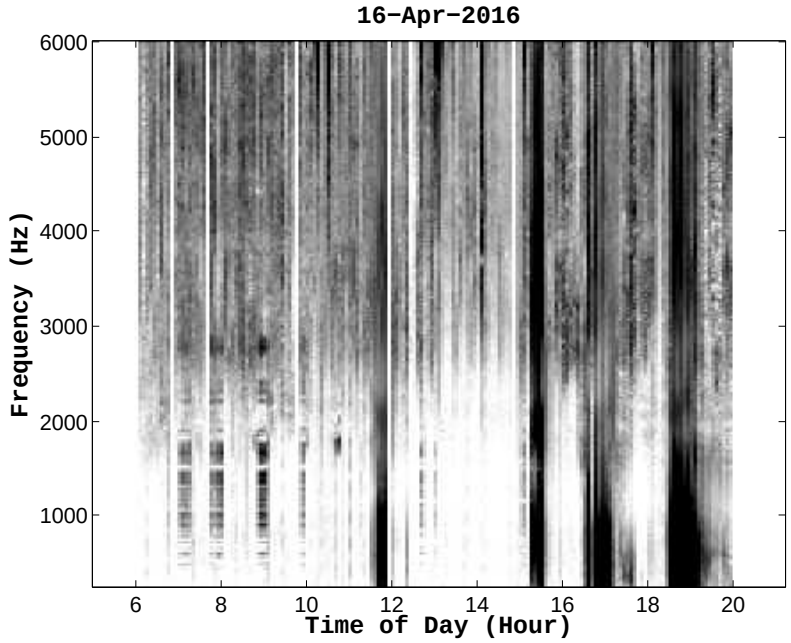


Figure 4.4: ACI Spectrogram Recorded in Tenneville, April 16th 2016

[4]Woodpeckers still drum in light rain. The drumming that started around 15:00 might have continued during the shower. Some caution must be used when discarding data.

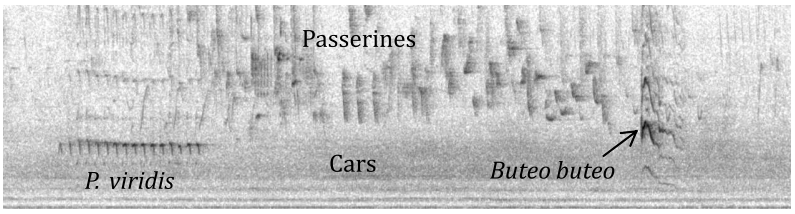


Figure 4.5: Remerschen, March 29th 2017, 9:35

Bandwidth 0-6 kHz. Spectrogram duration 15 s. *P. viridis*: green woodpecker. *Buteo buteo*: common buzzard.

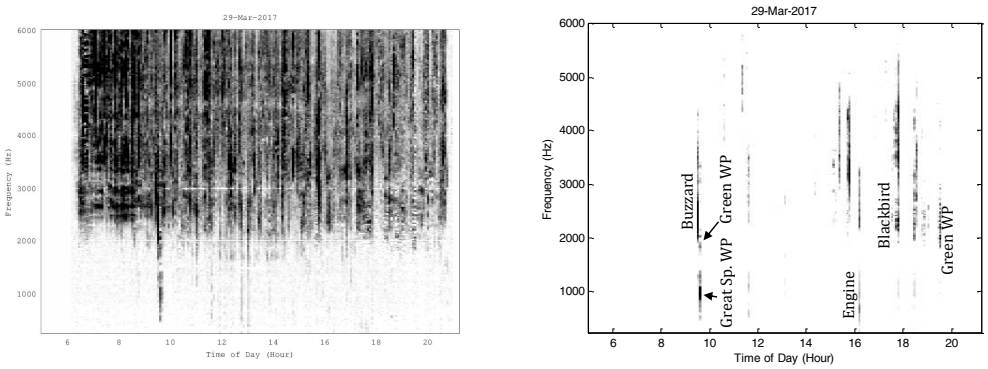


Figure 4.6: Original and Modified ACI Spectrogram for March 29th, 2017 at Remerschen

On the other hand, the calls yield only small and rather unremarkable stains on the ACI images. Compare the ACI spectrogram on the left of Fig. 4.6 and the interpretation on the right; the woodpecker calls are quasi invisible in the ACI spectrogram and cannot be distinguished from the blackbirds, buzzards and ducks that also use the woodpecker bandwidth. Only the *D. major* drums shortly before 10:00 stand out.

Limitations of the ACI

Fig. 4.6 is a telling example of the limitations of ACI spectrograms. The context is peculiarly tough because birdlife is abundant and diverse in the Remerschen wetlands and because the recording station was not placed next to a known woodpecker hangout. As it turned out, the woodpecker calls were often partially masked in the RM recordings; 9% of all identified calls in the

dataset generate ACI values less than 1.2. Again, the 1.2 threshold was inferred from recordings in which the target calls were generally center stage. It is not well suited to more distant calls. The ACI is more reliable with drumming because its formulation matches the structure of drums rather well (one frame of high amplitude followed by one frame of silence, repeated 13–32 times[5]) and because drums reside in a lower and less populated[6] frequency range.

Looking back at the ACI formulation in Eq. 4.1, two parameters affect the ACI negatively: the frequency range and the background noise. The ACI is the ratio between the average acoustic *intensity variation* over 30 seconds and the average acoustic *intensity* in the same time period. As a rule of thumb, the intensity is high at low frequencies because of cars, planes, human activities and other background noise. It goes down by an order of magnitude around 2000 Hz as the anthropophony dies out, and the values in the passerine range are residual[7]. This is why the ACI is biased toward passerine songs: it works very well with sharp temporal variations in intensity, even of a low amplitude, over low intensity values. On the contrary, the ACI can only detect the strongest intensity variations at low frequencies, because large background intensity values cancel them out. As a matter of fact, the ACI tends to increase with frequency; see for example Fig. 4.3. In Remerschen, the bird chatter is permanent and the background noise is further elevated by the nearby roads and planes (Fig. 4.5). The passerine songs black out the upper bandwidth (above 2500 Hz) and the background noise makes it difficult for signals to emerge out of the lower bandwidth (Fig. 4.6, left).

We attempted to reformulate the ACI in order to counterbalance the effect of the denominator. Using $I(f)$ as the average intensity spectrum over 30 seconds (minus the last frame) and assuming the 30-second signal contains n time steps, the ACI formula is first rewritten:

$$ACI(f) = \frac{\frac{1}{n-1}\sum_{k=1}^{n-1}|I_{k+1}(f) - I_k(f)|}{I(f)} \tag{4.2}$$

[5]In average, depending on the species, see Chap. 5.

[6]Less populated *by birds*, but the most common anthropogenic sounds do not trigger the ACI.

[7]The acoustic intensity in the ACI formula is on a linear scale. Decibels are not used. They would equalize the intensity across the frequency range to an extent, but also diminish the small variations in acoustic intensity that occur in birdsong and contribute to the success of the ACI.

Now, instead of normalizing by the average intensity $I(f)$, we can instead use the average of $I(f)$ over a range a frequencies of interest to us:

$$ACI(f)_{mod1} = \frac{\frac{1}{n-1}\sum_{k=1}^{n-1}|I_{k+1}(f) - I_k(f)|}{\frac{1}{N}\sum_{F=F_1}^{F_2} I(F)} \tag{4.3}$$

Here we assume that there are N frequency bins in the $[F_1, F_2]$ interval, which in practice is the 500–2700 Hz interval. The denominator is now independent of frequency: all intensity variations are scaled by the same reference. The expected effect is that at frequencies greater than 2700 Hz, temporal variations in intensity will be negligible compared to the denominator, and only the loudest calls will come up in the ACI spectrograms. At low frequencies, calls that were previously canceled out could appear, although low frequency calls tend to not have strong temporal variation. Our purpose is ultimately to favor the woodpecker calls in the 500-2700 Hz frequency range.

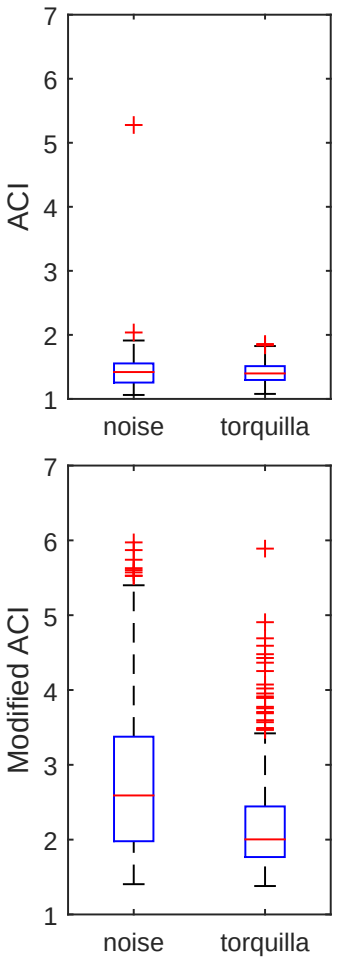


Figure 4.7: ACI and Modified ACI in Signals Recorded at Remerschen

The proposed reformulation brings ACI values down significantly. More nuanced results are obtained when the original ACI is scaled using the base 10 logarithm of the intensity spectrum[8]:

$$ACI_{mod2}(f) = ACI(f) \times \frac{\log_{10} I(f)}{\frac{1}{N}\sum_{F=F_1}^{F_2} \log_{10} I(F)} \tag{4.4}$$

In Eq.4.4, the scaling factor will be high at low frequency (the average intensity is greater than in the 500–2700 Hz range), close to 1 in the

[8]The intensity spectrum is divided by its minimum value (over all frequencies and all frames) prior to the ACI computation to avoid negative logarithm values.

woodpecker bandwidth and less than 1 in the passerine range. The logarithmic scale prevents it from reaching excessive values. Figure 4.6 shows ACI spectrograms in the original and final formulation. The picture is clarified through prioritizing vocalizations near the microphone, but the modified formula still does not provide a recognizable pattern for woodpecker calls. The choice of a threshold for the ACI remains a trade-off between false positives and false negatives, as both cannot be minimized at once. Fig. 4.7 illustrates how the modified formulation brings the ACI values up, but does not provide additional discrimination between woodpecker calls (here, *J. torquilla*) and other signals. In the end, the original ACI remained our pragmatic choice to implement a preselection of recordings.

4.2 Segmenting the Recordings

Preselection by the ACI leaves us with a collection of medium-size recordings[9] from which it is desirable to extract candidate bird sounds of a few seconds at most. The Region of Interest procedure (Potamitis [65]; Lasseck [46]) is likely the most advanced scheme to perform this task. It boxes every pattern in the spectrogram and isolates syllables of songs, sometimes even harmonics. This is perhaps too radical in the woodpecker case. Woodpecker sounds have a characteristic structure, hence it makes sense to extract them in one piece. The ideal outcome is audio segments centered around one drum or one call that can subsequently be identified. To that purpose, we opted to look for moments of high energy in recordings. We evaluated the acoustic intensity (or a similar metric) for each frame of sound in a bandwidth of interest and built segments from the frames whose intensity exceeded a threshold. This is a common approach known under many variants and names, e.g. median clipping (Potamitis [65]; Lasseck [46]), median-based background subtraction (Glotin et al. [30]), median-based thresholding (Stowell & Plumbley [74]), etc. Our implementations are described in Table 4.2. Our core routine was built with the following rules: a) the segments had a required minimal length; b) the intensity remained above the threshold for the duration of the segment, but silence gaps were allowed

[9]The TN/RM/LPR field recordings are 30 s long files. For the XC/TS/KT recordings, the ACI is not calculated because the presence of woodpeckers is not in doubt. The files are up to a few minutes long (XC/TS: mean duration 72 s, max 47 min, min 1 s; KT: mean 159 s, max 27 min, min 7 s).

Table 4.2: Parameters Used in Audio Segmentation

Dataset	**Sampling & Frames**	**Bandwidth**	**Time Knobs**	**Selection & Threshold**
XC/TS Drums	11025 Hz 23 ms frames 50% overlap	300–2500 Hz	gap 0.15 s min length 0.25 s lead 0.15 s	1.5×median [a]
KT Drums	12000 Hz 21 ms frames 50% overlap	300–1500 Hz	length 3.5 s (default) lead 0.15 s	manual[b]
TN Drums	12000 Hz 21 ms frames 50% overlap	300–1500 Hz	gap 0.15 s min length 0.4 s lead 0.15 s	max–30dB [c,d]
RM/LPR Drums	12000 Hz 21 ms frames 50% overlap	300–1500 Hz	gap 0.225 s min length 0.4 s lead 0.15 s	max–25dB
XC/KT Calls + TN (try 1)	12000 Hz 21 ms frames 25% overlap	500–3000 Hz	gap 0.3 s min length 1 s lead 0.15 s	2×median[e] OR max–30dB if more demanding
RM/LPR Calls + TN (try 2)	12000 Hz 21 ms frames 25% overlap	1000–2700 Hz	gap 0.3 s min length 1 s lead 0.15 s	1.5×median OR max–30dB if more demanding

[a] We choose $g_i = median(|p_i(t)|)$ as the evaluation function, where $p_i(t)$ is the audio wave (time series) for all t in frame i. Frames for which $g_i > 1.5 \times median(|p(t)|)$, where $p(t)$ is the audio wave of the whole file, are selected. The sound is bandpass-filtered before evaluation.

[b] The drums starting times were known.

[c] Here and onward, the frame evaluation function is $g_i = sum_{f=f_1}^{f_2} I(f)$, where $I(f)$ is the intensity spectrum for frame i and f_1 and f_2 are defined in the "Bandwidth" column.

[d] Frames for which $g_i > max(g_i, i = 1..N) - 30dB$ are selected, where $max(g_i, i = 1..N)$ is the maximum value of g_i over all frames, numbered 1 to N.

[e] Frames for which $g_i > 2 \times median(g_i, i = 1..N)$ are selected, where $median(g_i, i = 1..N)$ is the median value of g_i over all frames, numbered 1 to N.

between syllables or to account for the occasional interruption in the middle of a drum (e.g. XC150529); c) leading and trailing silences were included for margin, knowing that the first or last strikes in drums are sometimes quieter. Table 4.2 also documents the values we used for the temporal parameters involved in these rules.

We used frames of 21–23 ms for the segmentation, i.e. a greater level of detail than for the ACI preselection. The recordings were resampled to 12 kHz at most to downscale the Fourier transform computation efforts. The overlap between frames was decreased for calls, also in hope of downsizing the calculations through the vast RM/LPR datasets. The bandwidth choice followed the target signals: 1500 Hz for drums, 2700–3000 Hz for calls. Only for the XC/TS drums was the upper bound extended to 2500 Hz, in an attempt to use a greater part of the drumming signal. the XC/TS drums are often recorded at close range and exhibit a broad frequency content[10]. Naturally, this choice increased false detections triggered by passerine songs.

The different functions and thresholds used for the evaluation of frames bear no conceptual advantage over each other. Higher thresholds ensure that fewer sounds are selected for further processing, but diminish the probability of detecting faint signals. Table 4.3 describes some of the resulting datasets for drums. The segmentation reduces the duration of the audio to analyze by a factor 10. In the TN dataset, the mean duration of a segment is approximately 2.5 s; this seems appropriate considering mean drum durations are 0.65–1.86 s (Chap. 5). For the LPR1 dataset the extracts are shorter, which suggests a lot of non-woodpecker content. Files longer than 5 s usually correspond to planes or wind gusts. The selection process described in the next section (repetitions analysis) left out more than half of the remaining TN data. Between the ACI preselection, the segmentation and the repetitions analysis, the LPR1 dataset was reduced from roughly 136 hours to 14 minutes. We will soon comment on the efficiency of these processes.

[10]Fig. B.7 in App. B shows how drums loose most of their upper frequency content with increasing distance. At 100 m, all that remains is the main frequency peak.

Table 4.3: Output of Audio Segmentation for Selected Datasets, Targetting Drums

Dataset	Original Duration	After Extraction	Nb. of Extracts	Extracts Durations Mean	Min	Max
TN						
All dates	47:10:29	04:21:49	6760	2.32 s	0.26 s[a]	80.69 s
6-19/04	23:36:00	02:52:21	4145	2.49 s	0.26 s[a]	72.85 s
6-19/04	23:36:00	01:04:59[b]	1534	2.54 s	0.82 s	4.99 s[c]
LPR1						
Rain[d]	42:57:56					
No rain	92:41:14	00:14:14[b]	672	1.27 s	0.70 s	31.72 s

[a] A lesser issue with the code allowed 3 files to escape the set minimum duration of 0.4 s.
[b] Segmentation *and* selection using the repetitions analysis (see Section 4.3).
[c] In this variant files longer than 5 s were recut using a 10% threshold increase.
[d] Discarded using the ACI images.

4.3 Detecting Repeated Patterns in Sounds

The next step is to assess whether the extracted segments contain signals of interest or not. To this end, we first designed a methodology that exploited the inherent characteristics of the target signals, starting with drumming. Drumming is a succession of repeated strikes with almost identical spectra. We set out to search for these cues using tools from the music analysis field, notably the similarity matrix and the beat curve that were initially designed for the study of rhythm. The routes we examined in this direction involved a number of empirical parameters and still retained a high number of false positives. The next level in woodpecker detection was eventually reached using neural networks, as will be discussed in Section 4.4 and Chap. 6.

Similarity Matrix and Beat Curve

The strikes in a drum have a near-identical spectrum, with variations mainly due to the strength of the strike. A drum strike spans 1–2 frames with our signal processing parameters. The similarity matrix measures the resemblance between the spectra of different frames and thus appears like a suitable tool to detect the repeated strikes. Foote et al. [24] suggested the

following formulation for the similarity matrix S[11]:

$$S(i,j) = 1 - cos(\mathbf{u^{(i)}}, \mathbf{u^{(j)}}) \tag{4.5}$$

In the above, vectors $\mathbf{u^{(i)}}$ and $\mathbf{u^{(j)}}$ are the spectra[12] for respectively frame i and frame j. The frequency bandwidth is indicated in Table 4.4. The output is close to 1 for similar frames and close to 0 for dissimilar frames. The cosine function used as a distance metric discards the amplitude variations; it only matters that the spectra of the different frames have similar profiles.

Unfortunately, drums recordings contain only a few frames of signal and many frames of silence. Silence has a random direction and the similarity matrix calculated as in Eq. 4.5 yields little contrast. An alternative is to use the Euclidian distance instead of a cosine. With $u_k^{(i)}$ as the k^{th} coordinate of vector $\mathbf{u^{(i)}}$, S is then:

$$S(i,j) = 1 - \sqrt{\sum_k (u_k^{(i)} - u_k^{(j)})^2} \tag{4.6}$$

In this version, the output is close to 1 for similar frames and negative, possibly with great values, for dissimilar frames. As the frames containing the strikes are rather dissimilar from silence, this formulation generates the type of contrast visible in Fig. 4.8. Most of the matrix expresses similarity (silence with silence, strikes with other strikes), but the strikes stand apart from silence.

The beat curve measures the time interval between repeated frames. It is derived from the similarity matrix by summing along diagonals (Foote et al. [24]). With R the range of frames allowable for j and p the size of R, the beat B is:

$$B_i = \frac{1}{p} \sum_{j \in R} S(j, i+j) \tag{4.7}$$

The frame index i can then be replaced by the starting time t of frame i and B_i becomes $B(t)$. Following Table 4.2, in our case the time step is

[11]The similarity matrix, beat curve, filterbanks, envelope, STFT and peak picking were implemented using functions of the MIR Toolbox v.1.6.1 Lartillot & Toiviainen [45].

[12]Linear rather than logarithmic amplitude, in some cases with smoothing over 5 bins or 20 bins. A questionable choice in many instances, but not for drumming. Drumming excites a wide frequency band but concentrates most of the energy in one main peak, as we will see in Chap. 5. With the linear scale, the similarity matrix focuses the evaluation on the dominant peak.

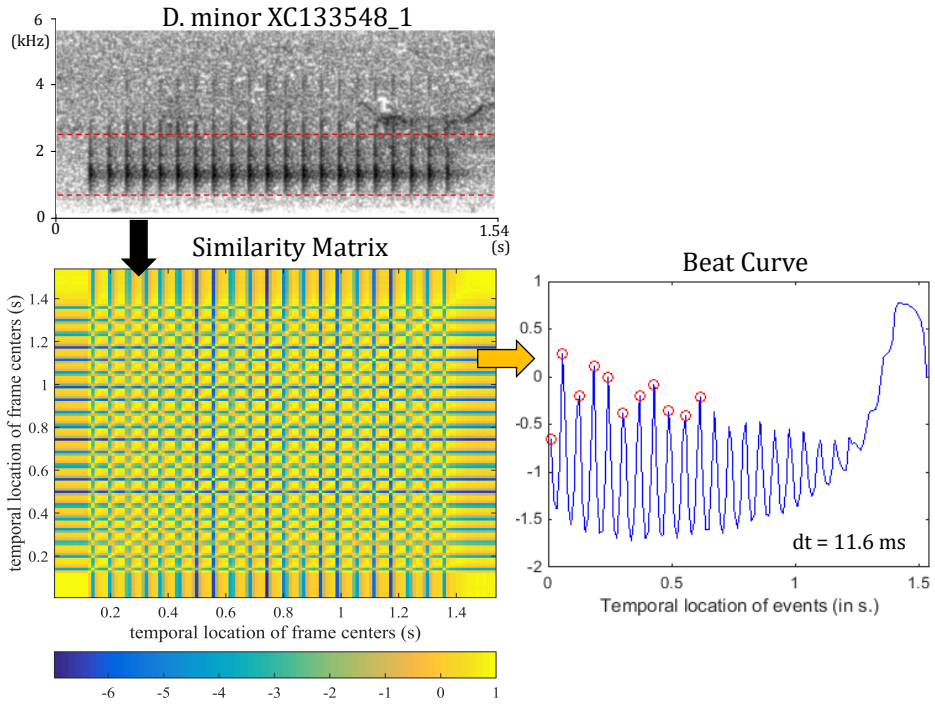


Figure 4.8: Similarity Matrix and Beat Curve

either 10.7 ms or 11.6 ms for drums and 16.0 ms for calls. The beat picks up on the periodicities in the similarity matrix and produces an oscillating curve (Fig. 4.8). The peaks mark the height of silence and the valleys the positions of the strikes. The time intervals between the peaks (or between the valleys) are of interest; their mean value can be compared to known woodpecker values. From Zabka [91], the strikes are repeated at intervals in the range 40–90 ms[13]. The time intervals are calculated with a limited precision, constrained by the time step resolution (11.6 ms in Fig. 4.8), but the estimates are sufficient for an evaluation against the 40–90 ms range.

The mean number of strikes per drum is 13–32 depending on the species

[13]Beyond the 11th strike, the intervals of *D. major* and *D. syriacus* might reach 35 ms. However as mentioned further down in the text, we focused on the first 11 strikes. Regarding the higher bound, the graphs in Zabka [91] show the greatest intervals for *P. tridactylus* at 80 ms (in average). From our observations in the XC dataset, this is often exceeded. Occurrences of time intervals reaching 96 ms were found (e.g. for XC153234). Because of this species, the upper limit was raised to 90 ms.

(see Chap. 5). Often though, the beat curve contains just a few clear oscillations before degradation. Another common difficulty is missed peaks (faint strikes) or extra peaks (overlapped signals, reverberation) that produce wrong time intervals. Thus, only the first 11 peaks were analyzed (at most), i.e. 10 time intervals, and the criteria for a positive output were: a) a minimum of 5 strikes were detected, b) less than 10% of the time intervals were longer than 90 ms and c) less than 10% of the intervals were shorter than 40 ms.

Peaks in the Signal Envelope

When the drums are faint or overlapped with loud calls, then the beat curve might not have enough of an oscillatory pattern in order to proceed with peak picking. The procedure described above altogether fails. For that case, a recovery path was implemented using peaks of the signal envelope. In this variant, we assume that the envelope peaks simply correspond to the strikes. Hence the calculations are more robust but there is little certainty on the content of the signal.

The envelope is derived as follows: first, the audio waveform is bandpass-filtered using the bandwidth specified in Table 4.4; then its absolute value is low-pass filtered twice[14] with an infinite impulse response (IIR) filter to remove all rapid oscillations. The smoothing of the signal by the IIR filter is controlled by a time constant τ. We used $\tau = 0.01$ s, which, applied twice, smoothed over 20 ms in total, about the duration of a frame, or the duration of a drum strike.

As before, we focused on the first 11 peaks. A minimum interval of 40 ms was imposed between the peaks. Those that fell into frames of silence (as determined during segmentation) were rejected. The criteria for positives were: a) a minimum of 5 peaks were detected, b) less than 10% of time intervals were longer than 90 ms.

Similarity Matrix and Beat Curve: Variant For Calls

With the exception of the *D. martius* contact call, the woodpecker calls in our study are series of repeated syllables. This means that the similarity matrix can also be considered in this case. However, the repetitions are not as perfect. The syllables morph, the pitch changes and the sounds extend across

[14] A common practice with IIR filters to preserve the phase of the original signal.

more frames. The resulting beat curve does not have the clean oscillations of the one in Fig. 4.8. Hence we had to modify the previous procedure. First, we smoothed out the beat curve to address the particular case of the *D. martius* flight call that has fine patterns below the syllable level – which we were not interested in at this point. Then, we adapted the peak picking to feeble oscillations. Instead of selecting peaks above a threshold, we selected peaks whose height was above a threshold. Here, peak height is understood as the peak emergence compared to the surrounding valleys. The minimum allowable height was either 20% of the maximum height or 15% of the height difference between the maximum peak and the minimum valley, depending on which one was the most demanding. Only the first 5 syllables were analyzed (there might not be more). The minimum time interval between peaks was set to 100 ms (for rattle calls, the typical duration between the start of two successive syllables is 150–300 ms). As before, the mean time interval between peaks was evaluated on the final selection.

The criteria for positives were : a) a minimum of 3 syllables were identified, b) the mean time interval was longer than 125 ms and c) the mean time interval was less than 350 ms.

Commentary

Table 4.4 lists the datasets on which the procedures were tested. In all cases, the bandwidth was identical to the one used during segmentation. The recovery path using the signal envelope was enabled for the XC/TS drums dataset only. As already expressed, the XC/TS files are rich in drums. They can be processed using methods that will bring forward dubious signals because the probability that these are truly drums is high. In real-life datasets such as TN/LPR, the envelope method would disproportionately increase the false positives.

In all cases, the segments that passed the acceptance criteria were reviewed to exclude false positives. For the XC/TS drums dataset, the outcome was controlled for all segments and false negatives were rescued. The performance numbers in Table 4.4 indicate that the assignments were correct in approximately 80% of the cases (XC/TS dataset). This setup using the envelope method as recovery path and a wider bandwidth is conducive to more false positives. In a more conservative setup, the TN dataset produced only 3% of false positives. The algorithm efficiently detected drums with

Table 4.4: Variants Used in Repetitions Analysis

Dataset	Bandwidth	Target Time Interval	Methods	Performance
XC/TS Drums	500–2500 Hz	40–90 ms	Similarity matrix & Beat curve Recovery: envelope	80% correct[a]
TN/LPR Drums	300–1500 Hz	40–90 ms	Similarity matrix & Beat curve	3% FP[b] (TN) 83–89% FP (LPR)
TN Calls	500–3000 Hz	125–350 ms	Similarity matrix & Beat curve	92.6% FP

[a]Results monitored approximately on a subpart of the dataset: 80% of the assignments, either positive or negative, were correct. This still conceals the drums that were not picked up by the segmentation step.

[b]False Positives (FP).

a high signal-to-noise ratio, i.e. when woodpeckers were drumming on the trees close to the microphone. On the contrary, the analysis was hampered for the feeble signals of distant drums. Masking by noise or other birds is another difficulty, but not as prevalent. The ACI images collected in Tenneville support the fact that there is little competition in the frequency range of drumming.

The results for the LPR dataset paint a less optimistic picture, as shown in Table 4.5. The first issue is that for LPR, the default preselection using the ACI worked on the 500–2100 Hz bandwidth in order to include the calls as well. In this configuration, the number of positives was too high to review. Even random background noise sometimes exhibits a structure that fits the requirements of the repetitions analysis. Hence the ACI preselection first had to be redone in a bandwidth limited to 1500 Hz. This ensured that the repetitions analysis would be performed on actual content from the drumming bandwidth. Still, the number of false positives remained as high as 83% in March (LPR1) and 89% in April (LPR2). Focusing on the LPR1 cohort, we noticed a drop in true positives (TP) when restricting the bandwidth; this corresponds to faint drums that had previously been ACI-preselected by chance because of superimposed passerine songs. For the FP,

the drop amounted to roughly 100 drums; mostly background noise that had been validated by the repetitions analysis. The other false positives included rustling sounds, planes, man-made noise, a tawny owl[15], etc. In the LPR2 cohort, the number of positives was too large to sort out the true ones by hand (10177). Thus, our methodology failed. In a practical dataset, too many signals occur in the target bandwidth that can fool the algorithm. For example, the tawny owl call is nothing like a drum and yet passed all the gates of our detection scheme (Fig. 4.9). The TN results might have been better simply because there was less happening in the woodpecker bandwidth at that location.

Finally, the poor performance makes the methodology terminally impractical for calls (Table 4.4). For the XC/KT datasets, the segments were assessed manually.

Table 4.5: Drums Detected in La Petite Raon and in Tenneville for Comparison

Cohort	**ACI Bandwidth**	**TP**	**FP**	**Not Analyzed**
TN[a]	300–1500 Hz	1494	40 (3%[b])	0
LPR1[c]	500–2100 Hz[d]	105	567 (84%)	0
LPR1	300–1500 Hz	98	468 (83%)	0
LPR2[e]	500–2100 Hz	0	0	13893
LPR2	300–1500 Hz	426	3620 (89%)	6131

[a] 06/04/2016–19/04/2016, see Table 4.1.
[b] A spike at 13% on 15/04/2016, caused by strong wind and microphone static.
[c] Recordings from March 2018.
[d] These settings of the recording station are suited for both drums and calls.
[e] Recordings from April 2018.
With difficulties arising, the LPR3 and RM dataset were not processed.

[15] *Strix aluco*. La chouette hulotte en français.

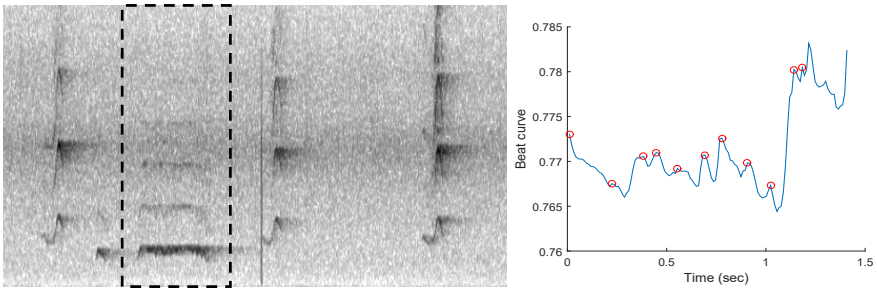


Figure 4.9: Tawny Owl Call and Beat Curve

Spectrogram bandwidth 0–6 kHz and duration 7.5 s. The beat curve on the right corresponds to the segment in the dashed rectangle. The curve does not possess the oscillatory shape as in Fig.4.8, but peak selection proceeds nonetheless and returns a mean time interval between peaks in the acceptable range for woodpeckers.

A Comparison with Cross-Correlation

We benchmarked the above detection procedure against spectrogram cross-correlation[16,17] (see Chap. 2, Section. 2.2) on the TN dataset. As template, we used a *P. canus* extract from XC98174, clear and with low background noise (Fig. 4.10). The template slides horizontally and vertically over the larger image being searched and the cross-correlation is calculated at each position. For each time position, we retained the maximum cross-correlation over all vertical positions (drumming can shift in frequency). We considered a match if this value was greater than 0.4. Both template and test images were limited to the 300–1500 Hz frequency range. The test image does not need to be segmented prior to analysis.

Table 4.6: Performance Data (I) for Repetitions Analysis Vs Cross-Correlation

	Repetition Analysis	**Cross-Correlation**
Run Time	4 h 48 min	21 min
True Positives	1494	1204
False Positives	3%	77%

[16] At the time, the most widespread detection method.

[17] Cross-correlation was implemented using Matlab's "normxcorr2" function. As "normxcorr2" responds best to matching white areas, we reversed the usual color scheme in order to match signal rather than silence (Fig. 4.10).

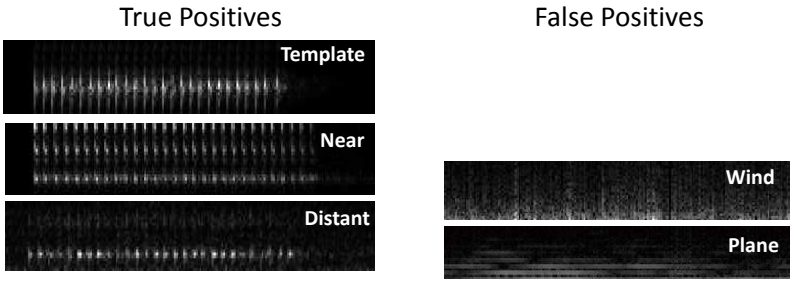


Figure 4.10: Spectrogram Images in Cross-Correlation Detection

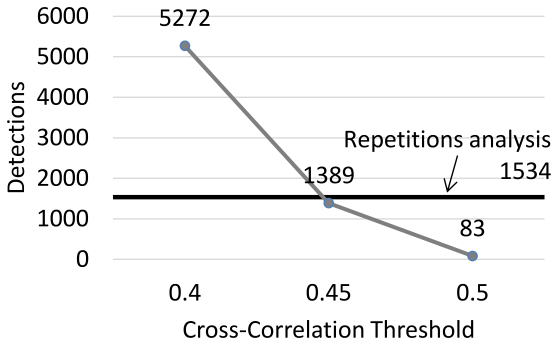


Figure 4.11: Impact of Cross-Correlation Threshold on Total Positives (True or False)

Table 4.6 and Fig. 4.11 document the relative performances of the repetitions analysis versus cross-correlation. Starting with Table 4.6, the difference in calculation speed is expected; cross-correlation uses a built-in Matlab function and has fewer tasks, e.g. segmentation is not required. However the repetition analysis is far more accurate (3% of FP versus 77% for cross-correlation). Fig. 4.11 shows the stark decrease in the number of detections as the selection threshold is increased. The original 0.4 threshold is the one that ensures a number of true positives on par with the repetition analysis. Still, the repetition analysis method extracts almost 300 additional drums (Table 4.6). It can be argued that *D. martius* also drums in Tenneville and that a second template would be required, but this accounts for only 120 drums[18]. The main difficulty is illustrated in Fig. 4.12, which displays the maximum cross-correlation curve for a recording with four drums. The margin between the mean cross-correlation (approximately 0.35) and the desired threshold of

[18]Details in Chap. 5. The identification yielded 2447 *P. canus* and 120 *D. martius*.

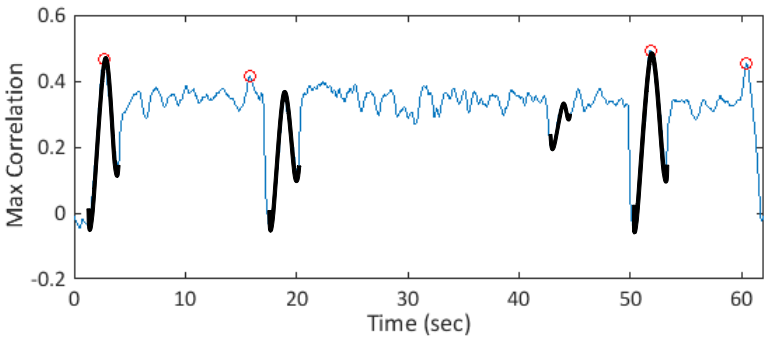


Figure 4.12: Selection of Peaks in Cross-Correlation - TN April 6th, 06:58:08

The circles indicate the instants where the maximum cross-correlation in the range 300–1500 Hz exceeds 0.4. Four peaks are selected. The recording contains four drums, indicated with the black superimposed curves. The third one is faint. Drums produce first a dip in correlation, then a peak that does not necessarily exceeds the preset threshold of 0.4, then a second dip. The dips occur at the points of first and last contact between the template and the candidate drum, and the peak when the drums overlay in the two images. Selection above a threshold yields 2 TP and 2 FP. Two drums are missed.

0.4 is small. In the end, the method detects a lot of erroneous content in the 300–1500 Hz frequency range, including wind and planes (Fig. 4.10), but not all target signals.

Fig. 4.12 also shows that drums induce first a dip in correlation, then a peak, then a second dip. The dips seem like better cues for the presence of a drum than the peak in correlation. We thus designed an improved drum detection procedure as follows: from the maximum correlation curve, we removed the mean value and selected the dips below a negative threshold. We imposed a minimum time interval between the selected dips, so as to retain only one per drum. We then extracted the sounds at the corresponding time stamps. Table 4.7 shows a comparison with repetitions analysis for two different days from the TN dataset. The new approach improves the performance of cross-correlation significantly. More concerning, it reveals that the combination of segmentation and repetitions analysis dissimulates a large number of false negatives. However, the number of false positives for cross-correlation remains important, and thus the method impractical.

Table 4.7: Performance Data (II) for Repetitions Analysis Vs Modified Cross-Correlation

Day	**Repetition Analysis**		**Mod. Cross-Correlation**	
	TP	FP	TP	FP
06/04/2016	28	1 (3%)	103	124 (55%)
13/04/2016	175	7 (4%)	218	402 (65%)

4.4 Drums Detection through Neural Networks

Net Retraining

Drums produce a distinctive pattern in spectrograms: a succession of vertical lines, reduced to small batons when the drum has traveled a certain distance[19]. To some extent it is even possible to identify the species on sight. The drums of *D. martius* look rugged, the ones of *P. canus* are clean, with strikes of constant strength and the ones of *D. major* are brief and with a sharp decay in strength. These species are easily discriminated from each other. However in a broader context, the differences between the taxa can be subtle. We leave the discussion of identifying the drums to the next chapter and focus here on detecting the drums.

To this end, we used what is currently the top-performing approach: very deep neural networks, pretrained on a large image bank and retrained to identify more specific images. This approach was described in Chap.2, Section 2.5, and will be discussed again in Chap. 5 and 6.

We used four pre-trained networks: Inception v3, ResNet with 34 layers, ResNet with 152 layers and DenseNet with 169 layers (see Table 2.2 in Chap. 2). The original nets all have a Softmax top-layer with 1000 entries (elephant, car,...) and a dense layer underneath that brings all previous results to converge toward the final 1000 neurons. These last two layers were removed and replaced by a dense layer converging toward 2 neurons and then a 2–neuron Softmax. The final outcome became binary: a drum (category 1) or not a drum (category 0). We then retrained all layers using woodpecker signals[20].

[19] See Chap. 3, Fig. 3.4, and App. B, Fig. B.7.

[20] Stochastic gradient descent, adaptive learning rate, 60 epochs at most. Further details in

Table 4.8: Drums Database for the Fine-Tuning of Deep Neural Networks

Dataset	**Number of Images**	
	Not a Drum	A Drum
XC	0	2343
TS	0	324
KT	55	2002
TN	1177	0
LPR1	571	103
LPR2	3632	426
TOTAL	**5435**	**5198**

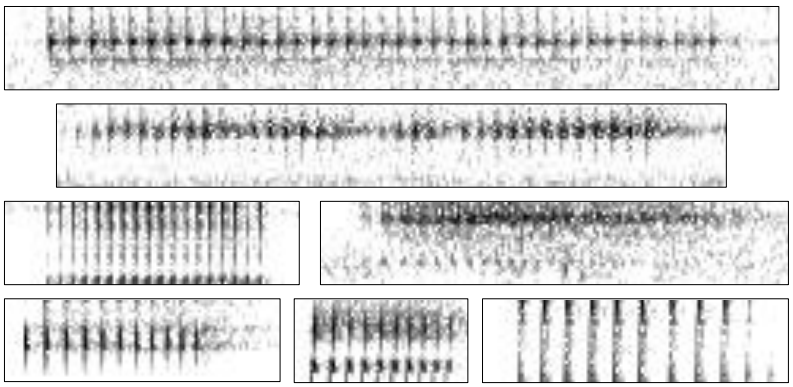
A database of 10633 spectrograms was assembled to retrain the nets. Table 4.8 details its composition. Drums stem from the XC/TS dataset (see Table 3.4 in Chap. 3), from Kyle Turner (Section 3.4.2 in Chap. 3) or were extracted through repetition analysis from the LPR datasets. All drumming species were sampled. The "not a drum" category was populated with false positives from the repetition analysis[21]. It seemed on point to include the signals that had previously been mistaken for drums. The database was then randomly split into a training set (90%) and a test set (10%). When a single root recording had yielded several sounds (drums or not drums), all were included on the same side, training or test.

We formatted all spectrogram images to a 300–1500 Hz bandwidth, a 12 kHz sampling frequency and 21.3 ms frames. The color scale was set to represent the top 30 dB in the sound fragment. When the original sound was longer than 5 s, several images were issued with a 25% overlap. The shorter images were not zero-padded; hence the images have a varying number of pixels in the horizontal direction. Fig. 4.13 shows a few samples. The mean image width is 143 pixels (1.5 s) and the maximum size for a 5-second sound is 26×469 pixels. Incidentally, all nets cited above accept 224×224 images as inputs, except Inception which takes 299×299 images. Our images were thus rescaled to fit this frame[22]. Approximately 10% of the images underwent

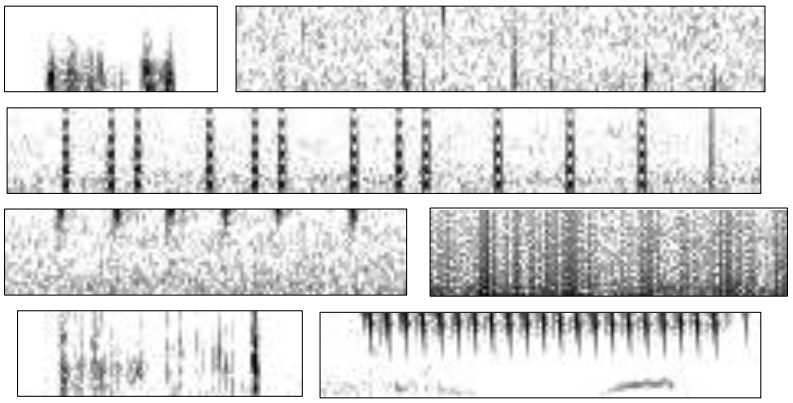
Chap. 6.

[21]Either in its final version or in intermediate trials. The KT dataset for example was eventually annotated by hand. The false positives come from an incomplete repetitions analysis.

[22]The ratio between width and height was preserved and the width was adjusted to the maximum allowable width, i.e. 224 or 299 pixels.



(a) Drums



(b) Not Drums

Figure 4.13: Images Used in Drums Detection Deep Net Training

a compression in width. For the 5-second sounds, the compression factor was 2.1. The ideal duration to preserve the original image details would be 2.4 seconds (Inception 3.2 s). A notable consequence of this rescaling is that the resulting images loose their shared time and frequency scales. The nets cannot sense whether a drum has a main frequency at 800 Hz or 1200 Hz, or whether it is fast or slow. A common alternative is to take multiple 224×224 crops from the original image; then the image fragments enter the nets at their original scale. Splitting the sounds longer than 5 seconds into multiple images was already a form of cropping. We will systematize this approach in Chap. 6.

Table 4.9 documents the performance of the nets on the test dataset.

Inception is the top performer, although all four nets exhibit outstanding accuracy. The nets are slightly more efficient in predicting the drums than the non-drums. Fig. 4.14 shows examples of mispredicted images. All four nets are fooled by the single knock in noise (a) and by rapid series of call notes (c,d). There are 18 instances of (c) in the test set (3.4%), which explains the lower accuracy for non-drums. Inception and ResNet 34 are fooled by a sample that contains nothing specific (b). Regarding the drums, Inception has a 100% accuracy but it should have rejected demonstrative tapping (e) and drums too short to be of interest (f), as the other nets rightly did. Drums that seem neither too faint nor masked were missed by ResNet 34 (h,i) and ResNet 152 (g). The results for (e) and (f) might reflect the capacity of the nets to recognize the same pattern at different scales. The short drum (f) is likely strongly distorted when the small image is scaled up to 224×224 pixels. In comparison, most of our drumming samples are in a 1–3 s duration range and are likely scaled by comparable factors. Hence only Inception seems able to recognize a (partial) drum at a widely different scale[23]. The same quality prevents it from differentiating between drumming and tapping. The other nets reject demonstrative tapping (e) and include fast-paced series of call notes (c,d); they seem to understand that there is an acceptable time interval between drum strikes, 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.

Table 4.9: Accuracy of Predictions on Test Dataset After Retraining

Deep Net	**Accuracy for Negatives**	**Accuracy for Positives**	**Overall Accuracy**
DenseNet 169	97.37%	99.44%	98.40%
Inception	97.19%	100.00%	98.59%
ResNet 34	97.37%	99.06%	98.21%
ResNet 152	97.75%	99.25%	98.50%

[23]When Inception was trained by its authors, 144 crops were generated for each image of the training set. The images were first rescaled to different sizes, mirrored, then cropped at a number of positions. In comparison, DenseNet used 10 crops and ResNet only one, taken from a randomly resized image.

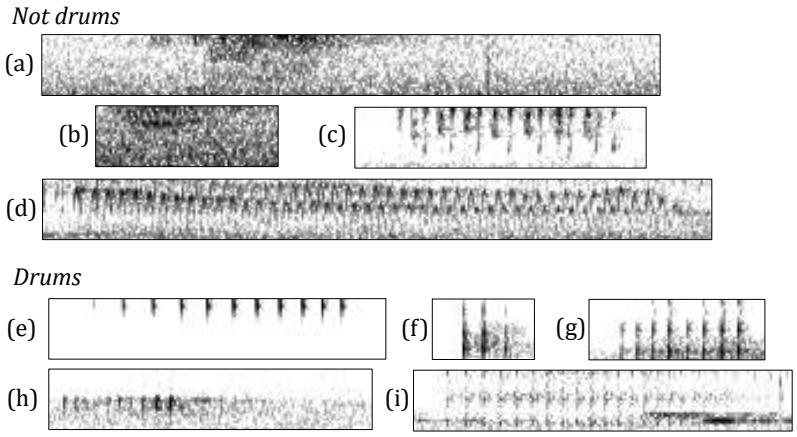


Figure 4.14: Images Mispredicted by Deep Nets

Not drums: (a) a lot of noise and one single knock at 3/4 of the image width; (b) noise; (c),(d) rapid series of call notes, with 60 ms intervals. *Drums*: (e) demonstrative tapping; (f) very short drum; (g),(h),(i) well-developed drums that are not particularly faint nor masked by other signals.

Drums Detection in Field Datasets

The retrained nets were then used to detect drums in the three field datasets TN, RM and LPR. The most frequent prediction among the four nets was adopted. In case of a tie, the prediction of ResNet with 152 layers was retained[24]. The results are shown in Tables 4.10 and 4.11.

The number of extracted drums for TN is on par with what was obtained with the modified spectrogram cross-correlation (Table 4.7). This confirms that the repetitions analysis missed a large number of drums. That the two image-based techniques, cross-correlation and deep nets, generate similar results was to be expected. We explained in Chap. 2 that deep nets are fundamentally based on image cross-correlation. The strong benefit of deep nets is the low production of false positives.

The number of false positives dropped to marginal numbers for TN and LPR2, helped of course by the inclusion of samples from these datasets in the retraining set from Table 4.8. The effect is not as strong for LPR3 (still 10% of FP), recorded later in the season when new birds have started to sing

[24]ResNet 152 came in second place for drums, but had a good accuracy with calls. Considering Table 4.11, it was not a bad choice after all.

that the nets do not know. In the RM dataset, the FP stand at 74.7%, yet the analysis still allowed discarding 94.7% of the initial dataset. The deep nets successfully detected drums although they are only marginally present in the datasets (in 4.0% of the analyzed sounds for LPR, 1.3% for RM). Examples of false negatives and false positives are shown respectively in Fig. 4.15 and 4.16. The false negatives are blurry images of quieter or more distant drums. The false positives comprise a variety of bird songs.

Having a closer look at the LPR results, for 98.7% of the images, the four nets were in perfect agreement. Table 4.11 illustrates the trade-off between precision and recall[25]; Inception and ResNet 152 missed the fewest drums, but produced twice as many false positives. The net with the simplest architecture, ResNet 34, was the one that yielded the fewest false positives. A possible interpretation is that a net with a (relatively) low analysis power, such as ResNet 34, can reliably identify the simple patterns of drums; nets with a greater analysis power can catch less obvious drums, but make more mistakes as they rely on tenuous details. Then the deeper nets would need more data or more data augmentation to learn and avoid the pitfalls. In the RM dataset, ResNet 152 does not quite fit this narrative; the FP increase by 14% compared to ResNet 34, whereas the increase is 71% for Inception. It is possible that the lesser flexibility of ResNet 152 when it comes to the size of patterns eventually turns up as an advantage. In any case, pooling together the four models improved detection for all datasets.

Demonstrative tapping was found in the RM dataset (52 occurrences). In agreement with previous observations, Table 4.11 shows that all nets mainly classified tapping as “not a drum” (as they should) except Inception. Pooling reinforced the negative assessment.

The RM results suggest that the deep nets can be further improved. In the present analysis, we did not employ all available techniques, e.g. data augmentation. Yet the good results for LPR indicate that drums were already sufficiently characterized in our training set from Table 4.8. Distorting the images might improve the results only marginally. What the RM analysis needs is the inclusion of more examples in the non-drums category. The images of false positives in Fig. 4.16 are proof that there is a multitude of natural sounds that can be confused with drumming by a deep net.

Finally, we remark that some datasets appear to be more difficult than

[25]Definition in App. A, Section A.2. Precision increases when there are fewer false positives, recall increases when there are fewer false negatives.

Table 4.10: Drumming Rolls Detected in TN/RM/LPR Through Deep Nets

Cohort	**Sounds**	**Images**[a]	**TP**	**FP**[b]
TN (All)	6760	7875	2570	4 (0.2%)
TN (6–19/04)	4145	4946	1548[c]	1 (0.1%)
LPR2[d]	5619	5619	347	4 (1.1%)
LPR3[e]	8933	10862	237	26 (9.9%)
RM	20866	28601	278	822 (74.7%)

[a]The sounds that lasted more than 5 s were split into several images. A sound is deemed positive if at least one of the derived images is positive.

[b]Negative predictions were not reviewed.

[c]115 TP on 6/04 and 186 TP on 13/04, compared to respectively 103 and 218 with cross-correlation (see Table 4.7). Note that in the present table, long sounds might contain more than one drum, e.g. when a plane increases the background noise and prevents a finer segmentation.

[d]Recordings from April 2018. This set comprises recordings shorter than 5 s and "not analyzed" in Table 4.5.

[e]Recordings from May 2018.

Table 4.11: Drums and Deep Nets: False Positives and False Negatives

Deep Net	False Positives[a]				False Negatives[b]				Missed Taps[c]
	TN	**LPR2**	**LPR3**	**RM**	**TN**	**LPR2**	**LPR3**	**RM**	**RM**
DenseNet	24	12	36	1263	9	3	7	15	50
Inception	34	16	70	1446	14	6	1	6	18
ResNet 34	3	8	25	846	26	8	4	15	40
ResNet 152	13	12	70	967	14	6	1	1	43
Pool	4	4	26	822	12	5	1	2	52

[a]False positives include demonstrative tapping.

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

[c]52 occurrences of demonstrative tapping were identified in the reviewed recordings.

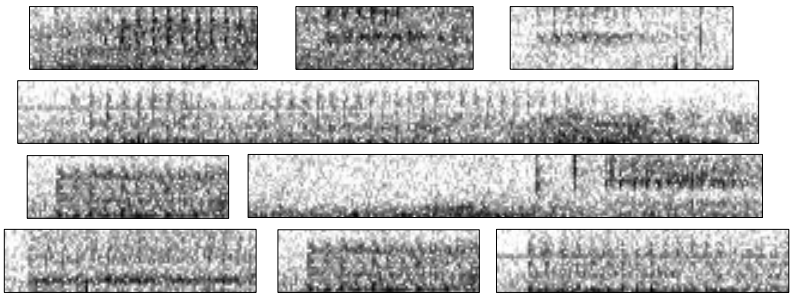


Figure 4.15: Examples at LPR of False Negatives
from Deep Nets Drums Detection

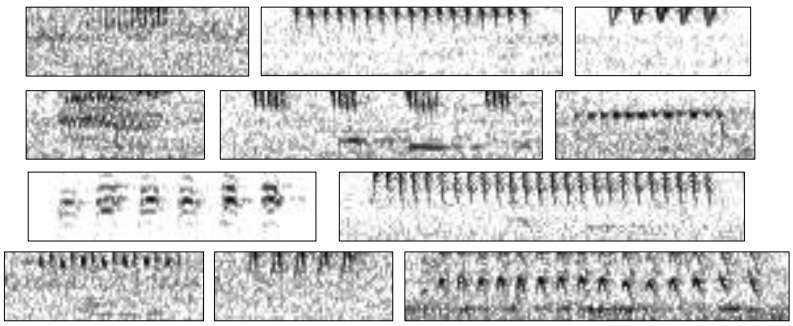


Figure 4.16: Examples at RM of False Positives
from Deep Nets Drums Detection

others. We achieved 3% of false positives in TN (Table 4.5) and 1.1% for LPR2 (Table 4.10), but 9.9% for LPR3 (Table 4.10) and 74.7% for RM (Table 4.10). LPR3, and particularly RM, are more complex cases. This can be traced to location and month of recording. TN is recorded in March and in the first half of April, LPR2 in April, LPR3 in May, and RM in the second half of April and in May. In March, the woodpeckers have the forest more or less to themselves. In May, the passerines have completely taken over and provide a variety of new noises that confuse the algorithms. This is exacerbated in Remerschen where the avian community is remarkable. On the other hand, as the woodpeckers are less active in May, the false negatives are less of an issue for LPR3 and RM (Table 4.11).

4.5 Conclusions

We structured drums detection in field recordings as a three-step operation: preselection, segmentation and detailed analysis. In the first step, the ACI was used in (almost) real time to prune 50–80% of the recordings. The remaining data was then segmented into sounds of a moderate duration, preferably 5 sec at most, by selecting high-amplitude signals in a target bandwidth. The third step entailed the most elaborate analysis and decided whether the sound was a drum or not. For this analysis, deep image nets outperformed other experiments, with less than 1% of false positives on the TN dataset, less than 10% of false positives on the LPR dataset and the elimination of 94.7% of the vast RM dataset.

Our focus was first and foremost on reducing the number of false positives. In the LPR2 case, analyzing the repetitions in the data left us with 89% of FP (3620 sounds) a third of the way through the dataset. Deep nets provided a more practical solution that reduced the FP to 1.1% for the remainder of the dataset. This is a ground on which further analysis can proceed. Only the RM dataset suggested that the deep nets could be further improved upon. In this dataset, 75% of the positives were false. The recording conditions were also the harshest we met: late in the season, amidst an extremely abundant and diverse avian community.

The number of drums extracted from the TN dataset with the two image-based techniques (spectrogram cross-correlation and deep nets) contrasts with what the repetitions analysis produced. The latter had a large number of false negatives. Considering that the deep nets analyzed audio fragments

that had been generated by our segmentation step and that cross-correlation worked on the full 30-second recordings, the comparable results indicate that the segmentation does not produce false negatives in significant numbers. In any case, our field recordings are proof that woodpeckers drum abundantly (Fig. 4.17). The consequences of missing a few drums are limited.

Note that the deep nets do not require smart segmentation per se; if vast amounts of data are not an issue, they can be split into regular 5-second crops and processed in this form. However, smart segmentation reduces the number of FP in that it limits the analysis to segments where the target signals are most probably present. Less noise passes the barriers. Without segmentation (and the increased ability to identify non-drums), cross-correlation still had 55–65% of FP (Table 4.7).

Overall, image-based methods (deep nets, spectrogram cross-correlation) retain a strong advance on signal-descriptive methods (repetitions analysis). Here we find back the idea expressed in Chap. 2 that the spectrogram is a complete and compact description of sounds and that any attempt to synthesize its content results in a loss of information. It follows that the parameters used in the STFT have to be carefully chosen. We saw for example that 46 ms frames in the ACI calculation were problematic for some *D. minor* drums. However, in practice, we were able to preselect *P. canus* drums, which are equally fast.

The issue of detecting woodpecker calls is left out for Chap. 6, in which deep nets will be used to detect *and* identify calls in a single step. However, note that for the field datasets TN/RM/LPR, the ACI preselection was operational during the measurement campaigns. In Tenneville, the upper frequency band was 1500 Hz, meaning that mainly the calls issued in succession with drums were recorded. In Remerschen and La Petite Raon, we used an upper bound of 2100 Hz that remains detrimental to *D. minor*.

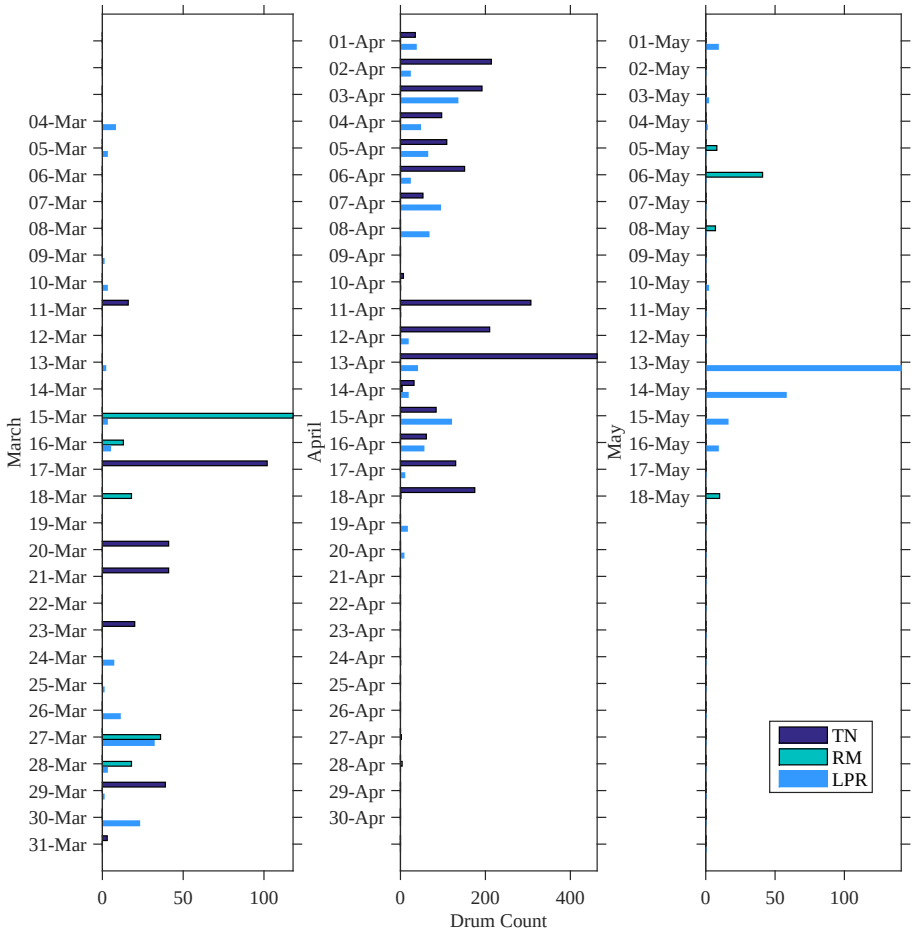


Figure 4.17: Drums Detected at TN, RM and LPR

Chapter 5

The Identification of European Woodpeckers from their Drums

After drums have been detected in recordings, comes the task of identifying the species that produced them. The drums of different species have different characteristics (Zabka [91]; Garcia [28]). In the present chapter, we address the calculation of acoustic features that capture this species character, then the classification of drums, both in demonstration datasets and in our field recordings.

5.1 Acoustic Features for Drumming

Past Solutions and Present Approach

Drumming is foremost a time signal. Its main parameters are depicted in Fig. 5.1. They are the time between rolls, the time between strikes and the drum duration. These traits are natural features to classify Drumming Rolls (DR)[1].

[1]Drumming roll, drum roll, roll and drum are equivalent terms in the present text.

Figure 5.1: Temporal Parameters of a Drumming Roll (DR)

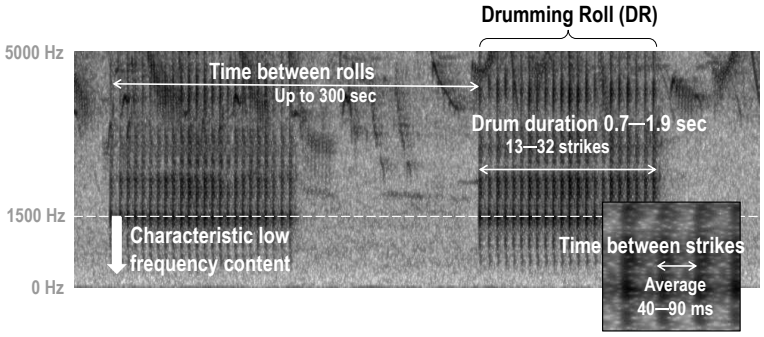


Table 5.1: Parameter Mean Values and Variation Published in Zabka [91]

Species	**Number of Files**	**Delta Interval**	**Initial Interval**		**DR Duration**		**Nb. of Strikes**	
		Mean (ms)	Mean (ms)	Variation (%)[a,b]	Mean (s)	Variation (%)	Mean	Variation (%)
D. leucotos	17	-1.2[d]	80.2[d]	8	1.64	11.3	34.4	13.8
D. major	104	-2.0[d]	61.5[d]	8	0.56	28.7	13.1	27.2
D. martius	21	-0.6[d]	72.0[d]	12	1.61	31.3	29.2	27.2
D. medius[c]	17	0.2	57.1	4	1.29	10.9	23.0	26.4
D. minor	40	0.0	48.9	12	1.19	24.1	24.6	22.9
D. syriacus	11	-2.0[d]	72.9[d]	8	0.89	16.8	21.6	20.6
P. canus	16	0.1	52.3	13	1.37	21.0	26.4	21.2
P. trid.	9	-0.5	76.1	3	1.34	25.4	20.8	19.1
P. viridis	5	0.6	40.4	-	1.15	24.3	25.8	27.5
Total	240							

[a]Standard deviation over mean. See Eq. 5.2.

[b]The numbers in this column were inferred from illustrations.

[c]Questionable data. *D. medius* does not drum.

[d]For a straightforward comparison with our results, the exponential forms used by Zabka [91] were converted to linear curves. E.g. for *D. major*, a set of 12 (n, t) samples were generated from the equation $t = 63.56 \times n^{-0.1827}$, where t is the time interval between the n^{th} strike and the next one; 12 intervals were considered because the mean number of strikes is 13 for this species; then a new polynomial was fitted through the 12 samples. The slope gave the delta interval and the y-intercept led to the initial interval through the same formula as in Table 5.2.

Zabka [91] ran a previous study on woodpecker drumming using 240 DRs from nine European species, i.e. all but *J. torquilla* and *P. sharpei*[2]. *D. major* made up almost half of the collection (Table 5.1). The time intervals between strikes and the duration of rolls were measured manually on spectrograms and signal envelopes. This author rejected the time between rolls as a viable acoustic feature because of excessive variability. His findings were that *D. major* had the shortest DR, and *D. leucotos* and *D. martius* the longest. The time structure (evolution of the time between strikes) followed either a linear law or a decreasing exponential law and was a critical species trait. Some woodpeckers accelerated (e.g. *D. major*); others maintained a quasi-constant speed or decelerated slightly (e.g. *D. minor*).

Stark et al. [73] performed another statistical analysis on drumming parameters for 11 woodpecker species occurring in California (3347 DRs). DR duration, number of strikes per DR, average time interval between strikes and cadence (strikes per second) were considered. Cadence was found to be the best indicator for species differentiation. Overall, 78% of all samples were correctly reclassified using their parameter set; 91% when considering only sympatric species (i.e. with overlapping geographical ranges). The authors argued that only species susceptible to live in the same areas needed to differentiate their acoustic signals. Admittedly, the design space for drumming does not have enough parameters to fully differentiate 300 woodpecker species, unless the breeding range is factored in. Our approach, considering only the European species, is consistent with this observation.

Zabka [91] and Stark et al. [73] both determined that the temporal parameters of drumming were the diagnostic ones. In the two studies the time between strikes was the primary variable for the separation of species (the cadence in Stark et al. [73] is the inverse of the mean time between strikes). Here we see how the drumming case deviates from the general problem of bird song classification described in Chap. 2. In Chap. 2, we presented the evolution of the technique from using acoustic feature sets designed for human voice recognition (primarily the MFCC) toward considering spectro-

[2]Until recently, the remaining nine species were thought to drum. Then Turner [84] showed that the *D. medius* samples in circulation were misidentified *D. minor*. Kyle Turner's recordings helped with clarifying the use of drumming in European species. As seen in Chap. 3, it was assessed that *D. medius* does not drum and that *J. torquilla*, *P. viridis* and *P. sharpei* produce only soft drums, rather than the far-carrying territorial drums. Our own work on drumming features (Florentin et al. [21]) considered the same species list as Zabka. In the present text, we removed mentions to *D. medius*.

grams as the one compact presentation of sounds that can retain both the time and the frequency dimensions. The drumming case has specific demands that none of the two approaches fully meets. The MFCC and their derivatives for example can be excluded on two grounds, first because they are designed to capture a piece of information that is scarce in drums (meaningful frequency content, rich in harmonics) and secondly because they are classically averaged over time and thus retain no description of rhythm. Spectrograms might not be suitable for a precise description of drums either, because the time step is imposed and limited by the STFT processing parameters. A precision to the millisecond is required (e.g. consider the initial intervals in Table 5.1), whereas the spectrogram time step we used up until now was approximately 11 ms, which, admittedly, could be refined at a computational cost. However the detailed calculations might be excessive for a signal as simple as drumming. Lasseck [46] also made a comment that spectrogram images do not render temporal structures and repetition rates in bird songs well. For context, his study focused on disassembled song elements. He thus suggested that different time scales could be used in order to eventually consider entire songs, and then their repetition at intervals. He later used random 6-second segments with deep nets (Lasseck [47]); the method scored well but the results were not autopsied any further. The capacity of spectrogram-based techniques to treat the layout of elements in function of time is essentially unproven in the ecoacoustics literature. On the other hand, the images have one potential upside; deep neural networks might be able to capture more subjective qualities of sound, in particular whether a drum sound is rough or clear, which might be species-specific. Describing sound quality, e.g. timbre, has always been a grail of audio feature extraction.

In the present work, we will discuss two approaches: simple acoustic features in conjunction with the k-NN classifier, and fine spectrogram images with deep neural nets. Regarding the acoustic features, we relied first on the temporal parameters indicated in Fig. 5.1. This being said, the spectral information is not entirely without merit. Woodpeckers choose the tree on which they drum, potentially because they like its sound. This preferred sound might be a species trait. Stark et al. [73] supported this hypothesis. Some birds might also have the strength or skill to excite higher trunk harmonics. Therefore, we included spectral parameters in our analysis.

The final eight features we selected are summarized in Table 5.2, which

includes mathematical formulations for the temporal features. For each drum, the series of inter-strike intervals was line-fitted, resulting in an inter-strike interval versus strike number function that captured speed and acceleration or deceleration through the drum. The upcoming sections further discuss this design and other parameters in Table 5.2. We computed parameter ranges on the XC/TS dataset (Table 3.4 in Chap. 3), which comprises the same nine species as the Zabka dataset in Table 5.1. The most represented species are *D. major*, *D. minor* and *P. tridactylus* (>500 samples). Less than 30 samples are available for *D. syriacus*, which is rarely recorded, and *P. viridis*, which does not use drumming as a territorial signal; this is insufficient for statistical significance. Of the remaining species, the smallest class is *D. martius* (84 samples). Again, we excluded the *D. medius* samples.

Table 5.2: Overview of Drums Acoustic Features

Numbering	Features	Mathematical Formulations	
		For n strikes with (t_i,y_i) coordinates in the envelope curve:	
f_1 f_2	First Interval Delta Interval	$t_{i+1} - t_i = f_2 \times (i-1) + f_1$	for $1 \leq i \leq n-1$
f_3	DR Duration	$f_3 = t_n - t_1 + \epsilon$	ϵ takes into account the width of the first and last peaks
f_4	Number of Strikes	$f_4 = n$	
f_5	Amplitude Slope	$y_i / y_{max} = f_5 \times t_i + \nu$	for $5 \leq i \leq n$ ν not used
f_6 f_7	Spectrum Centroid Spectrum Peak		
f_8	Time between Rolls	for the p^{th} DR in a series of P DRs:	
		$f_8 = (t_1)_2 - (t_1)_1$	for $p = 1$
		$f_8 = \frac{1}{2}((t_1)_{p+1} - (t_1)_{p-1})$	for $2 \leq p \leq (P-1)$
		$f_8 = (t_1)_P - (t_1)_{P-1}$	for $p = P$

The full process flow, from the extraction of drumming rolls from audio records to the identification of woodpecker species, is sketched in Fig. 5.2. In Chap. 4, the XC/TS data yielded 2665 DRs, which were saved to individual files. The upcoming section addresses the identification of different birds in the case when more than one is heard drumming in a recording. This was used to calculate the time interval between rolls. Subsequent sections discuss the parametrization of individual DRs and their classification.

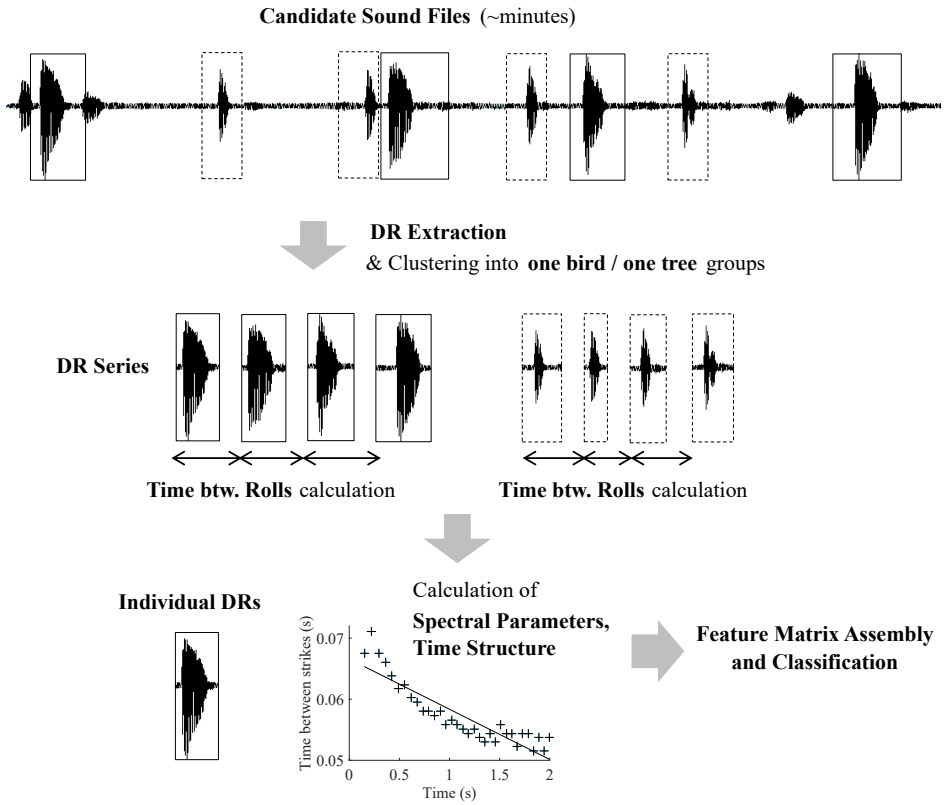


Figure 5.2: Drumming Species Identification Process Flow (Example: XC89410)

Time Interval Between Rolls

The time between the start of two successive drumming rolls was estimated from the DR file start times[3]. For isolated DRs (only one DR in the original audio record), it could not be computed. The formulation in Table 5.2 (feature f_8) describes series of P drums. We identified such series by including DRs that were separated by at most 300 s (5 minutes). We assumed that pauses longer than 5 minutes meant that the bird had moved on to another activity. This time limit is arbitrary but large. As the formula shows, the time interval between rolls for one given DR was calculated by averaging the time intervals to the previous and to the following DR in the series.

Multiple woodpeckers drumming in succession were a complication for the analysis (e.g. XC169038 contains DRs from six confirmed individuals, XC89410 from four). The DRs in such a file have to be assigned to distinct individuals. We did it by using k-means unsupervised clustering and spectral features. Indeed, drumming on different trees produces different nuances of sound, which is directly reflected in the spectral content[4]. We thus based our feature vector on the strike spectrum. This is appropriate because all DRs from one bird hitting on the same tree spot have a stable spectral content; the sorting algorithm does not need to cope with shifts in frequency[5]. The strike spectrum is the average of the spectra of all the frames that overlap a strike. The energy (integral of the squared spectrum) in these frames is greater than the median frame energy for the DR, which allows automating the selection. Eventually, the feature vector we used was the derivative with respect to frequency of the normalized strike spectrum. This proved more resistant to variations in signal amplitude. The feature matrix was then complemented with a training set of 76 manually preselected vectors. These were triplets of DRs issued by the same bird hitting the same tree. Within the triplets the DRs were thus similar in spectrum, and different triplets had distinct timbres. We selected three triplets per species if available (9 files). This added training set was meant to help k-means gauge the level of discrepancy that

[3]Each detected drum was saved to a separate audio file that started 0.15 s before the drum.

[4]In practice, the situation where several birds are drumming cannot be told apart from the situation where one bird produces a series of drums and then moves to another drumming spot with a different sound (e.g. XC171084). The clusters regroup distinct birds at distinct locations.

[5]Considering two spectra with one peak each, the two peaks being in neighboring frequency bins; then the sounds are similar but the feature vectors are unrelated. The sorting algorithm does not grasp the closeness of frequency bins.

makes two spectra unalike. For the initial cluster centers, we picked a set of vectors which were different from each other, yet similar to other vectors, i.e. not isolated cases. As the number of bird/tree combinations (clusters) was a priori unknown, a loop was implemented where k-means was run for an increasing number of clusters. As soon as the cluster centers became similar (the dot product of the corresponding vectors yielded an angle which was less than 10°), the loop was stopped. We enforced that the final number of clusters had to remain below 6 to keep the worst mathematical artifacts at bay.

The detection of individual bird/tree combinations was tested on 16 *D. major* files for which the XC or TS annotations reported several birds. These files typically contained 20 to 40 DRs and two to four birds. The solutions were perfect for nine files; three more had two wrongly classified DRs; only one had poor results (5 misclassified DRs out of 15). Further checks on other species confirmed that the bird/tree assignments were realistic. For example, for *D. minor* XC171084 where 109 DRs were identified in a 35-minute file, long series of identical DRs (same bird/tree) were detected, followed by transitional phases. This complies with the ornithologist's notes, which indicate that the bird kept going back and forth between two trees. The bird/tree assignments were also in agreement with the audio. In the full drumming

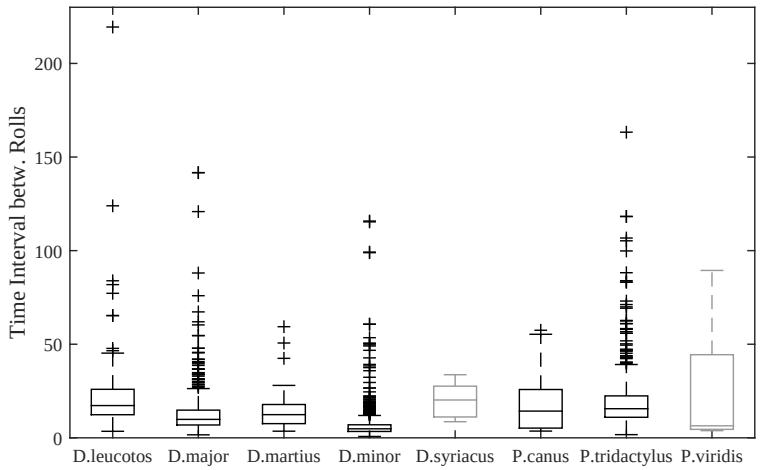


Figure 5.3: Time Interval Between Rolls, per Species

Light gray boxes are used for species with less than 30 samples.

database, 42% of the recordings (153 out of 361) were found to have more than one bird/tree combination. Eventually, it was possible to assign a value of time interval between rolls to 92.7% of the 2665 DRs. Among the remaining 7.3%, 1.6% correspond to files where only one DR was recorded and 5.7% (153 DRs) are DRs that were isolated by the clustering (i.e., associated with no other DR).

Fig. 5.3 shows the spread of calculated values per species. *D. minor* appears to drum in series with short intervals between the drums. For the species with a significant number of samples, the third quartile of the distribution (box top) remains below 30 s and the whiskers below 60 s. The most populated classes (*D. major*, *D. minor* and *P. tridactylus*) exhibit a large number of outliers above the box. This could mean that a) it is not important for the birds to constrain this parameter, b) the bird/tree identification algorithm artificially produces these larger values or c) the duration of the silence after which a series is broken is overestimated (here, 5 min for all species). Note that a few estimates were terminally unreliable as some recordists posting on XC edit silence out of their files (e.g. XC171084). No evidence was found in the data that the time interval between rolls depends on the bird drumming alone or with other birds.

Frequency Parameters

The typical strike spectrum, produced by averaging the spectra of all the frames that overlap a strike, has one main peak. Fig. 5.4 through 5.7 primarily describe the calculation of temporal parameters, but let us see a few examples of strike spectra, shown in the top right plots, with the corresponding spectrograms shown on the top left. Fig. 5.7 is a common example, with one main peak near 1000 Hz. The spectrogram, with amplitudes (colors) in a decibel scale, shows significant spectral content roughly from 500 Hz to 2000 Hz. This is deceiving; the logarithmic scale creates the expectation that the strike spectrum should show a plateau over a large bandwidth. In reality, the linear scale strike spectrum, on the right, is dominated by one peak, i.e. the first mode of the substrate. The rest of the spectral content is barely visible.

Harmonics are scarce and generally have a low amplitude. The *D. minor* sample in Fig. 5.4 is a counterexample with a remarkably rich spectral content, likely recorded at close range. The presence of harmonics depends on

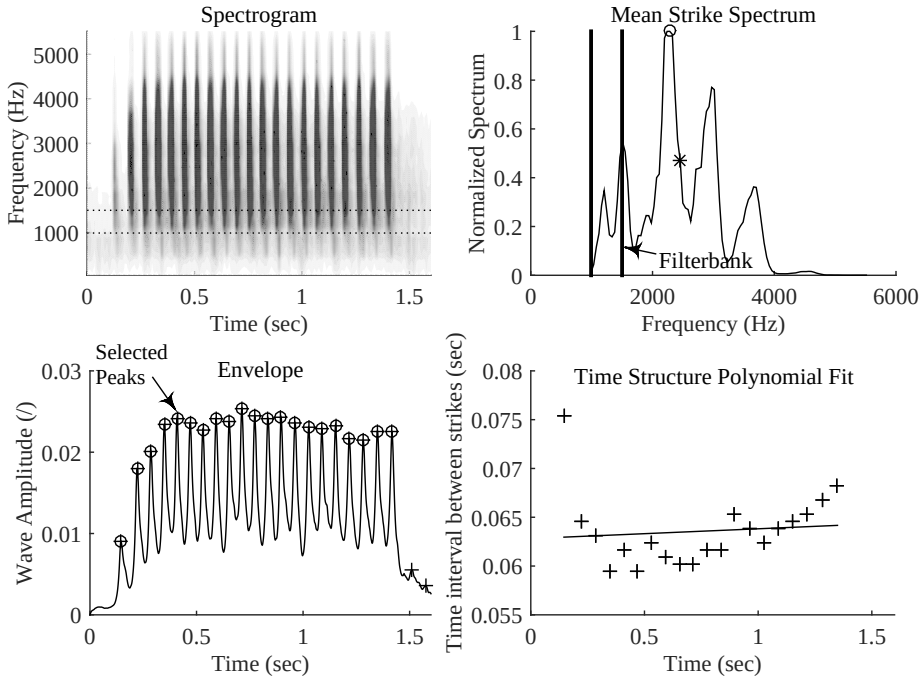


Figure 5.4: Acoustic Parameter Calculation from Spectrum and Envelope (*D. minor*, XC129193)

the excitation (the strength of the bird and the location of the hits), the resonator (the tree modes, easily excited or not) and the medium in which the sound propagates (absorption is frequency-dependent). The environment of the forest trims drums until only a thin dominant peak remains. The higher frequencies are washed away first. In essence, the argument for tracking harmonics is weak.

In the end, the strike spectrum is simple enough that we described it using two numbers: the maximum peak (shown with a circle in the figures) and the spectrum centroid (shown with a star). These are features f_6 and f_7 in Table 5.2. In most cases, the two frequencies coincide or nearly coincide. The correlation between the spectrum centroid and the maximum spectral peak is strong (0.83). Deviations occur in the presence of harmonics, as the centroid shifts to higher frequencies (Fig. 5.6). Fig. 5.4 is an example in which a second moment (standard deviation) could have advantageously complemented the spectrum description. The numbers are also polluted by

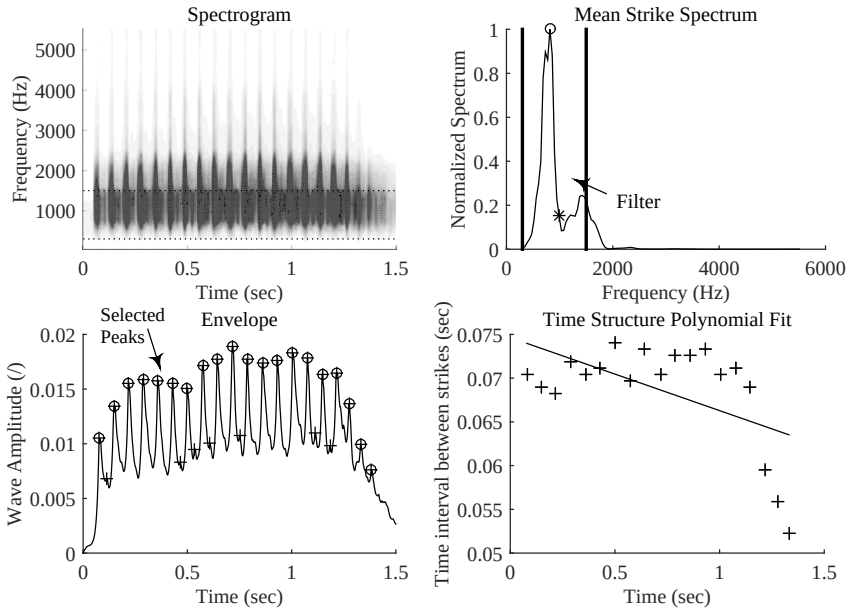


Figure 5.5: Acoustic Parameter Calculation from Spectrum and Envelope (*P. tridactylus*, XC153234)

the occasional co-occurring passerine song.

Fig. 5.9(b) in the next section shows the frequency of the main peak for all samples in the XC/TS dataset. *D. minor* uses the highest pitch (and is the smallest bird) and *D. martius* the lowest (and is the largest bird)[6]. For all species but *D. minor*, the third quartile of the distribution of main frequency peak values (i.e. the box top) lies below 1500 Hz. Three species produce frequencies above 2000 Hz; *D. major* and *P. tridactylus* exceptionally and *D. minor* for a quarter of the distribution.

Another frequency-related parameter of importance is the band-pass filter used in the calculation of the signal envelope, further discussed in the next section on temporal parameters[7]. This envelope is necessary for their derivation; the strikes will be identified as its peaks. As a consequence, the

[6]For these two species, size or strength necessarily affects the drums. *D. martius* nests on large trees and uses trunks as substrates, while *D. minor* resides in the upper, thinner branches. Evidently, *D. minor* would not be able to efficiently excite a large trunk. Overall, considering all species and using the mean bird length in Gorman [31], the correlation between length and frequency of the main peak is only -0.36.

[7]Calculation details for the signal envelope are in Chap. 4.

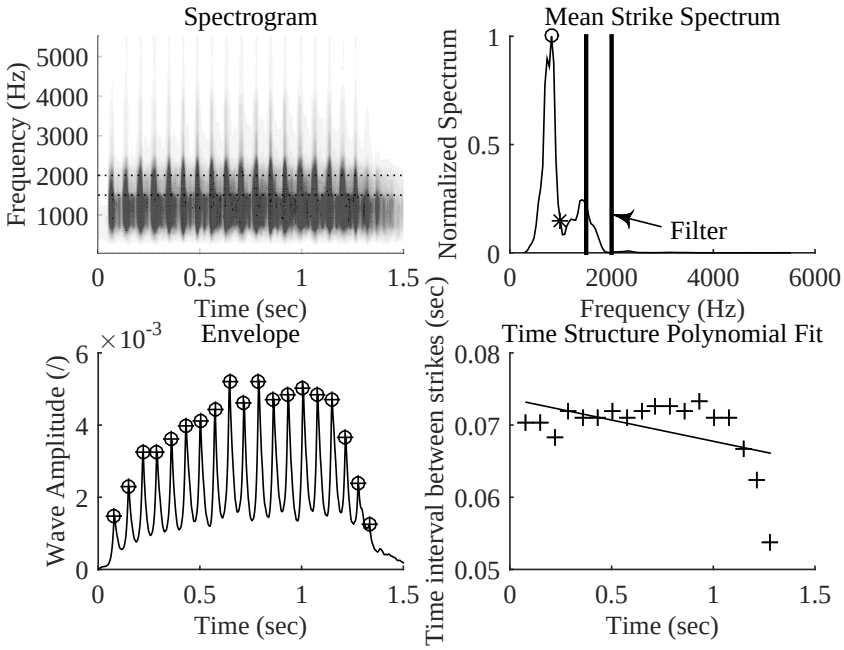


Figure 5.6: Acoustic Parameter Calculation from Spectrum and Envelope (*P. tridactylus*, XC153234) - Using a Different Filter

In this example the filter bounds were specified by user input, through the manual mode of the program for the calculation of drumming parameters.

filter must satisfy two requirements: 1) the bandwidth must contain a relevant part of the drum signal and as little as possible of other signals and 2) the signal in this bandwidth must yield a clean envelope curve, with clear peaks.

Regarding 1), as in Chap. 4, the default band-pass filter in our analyses is 300-1500 Hz. The low cut at 300 Hz allows setting aside common background noise and the high bound helps with minimizing the interference with other birds. Not all of the spectral content that is readily visible in the spectrogram must be included to produce a clean envelope; a subpart suffices, as long as the different strikes are clearly detached. The low frequencies are the most probable in drums, as they propagate further in the forest. However the data in Fig. 5.9(b) suggest that this might be an issue for *D. minor*.

We then implemented a procedure to adapt the upper bound, using the

variations in the spectra of the different frames containing the strikes. While drumming produces a stable strike spectrum, the passerine songs which populate the higher frequencies are inherently frivolous. The distinction can be made visible using the standard deviation of the strike spectrum (dash-dot line in the top right plot in Fig. 5.7). Above a threshold in the standard deviation, the signal corresponds either to passerines (5000 Hz peak on Fig. 5.7) or to the upper tail of the strikes (2200 Hz peak). In the example considered in Fig. 5.7 the 1500 Hz upper bound was maintained.

Regarding the second requirement on the filter, i.e. a clean envelope with well-detached peaks, the main issue is reverberation. In a reverberant environment the sound of the strikes trails off, particularly at the main frequency peak. Then the strikes do not emerge as clearly above the background noise, and the envelope has either weak peaks or intermediate peaks caused by the echoes in between the strikes. Fig. 5.5 is an example with intermediate peaks in the envelope. Here the peak selection routine successfully eliminated the lower intermediate peaks. Note the darker background behind the strikes in the spectrogram; it is a sign of high reverberation. This example illustrates the fact that better envelopes can be obtained by keeping the main and most reverberating peak out of the filter. In Fig. 5.6, the same sample was treated using a 1500-2000 Hz filter that encompasses a part of the secondary spectrum peak. At these frequencies the strikes are not as strong but they are clearer. The envelope generated with this filter allowed a straightforward peak selection. Here the filter was modified through the manual mode in the drums features calculation program[8].

Time Structure and DR Duration

The determination of the time structure begins with picking the time positions of the strikes in the DR. We used the peaks of the waveform envelope, because it allowed a high precision[7]. The original signal was downsampled by a factor 8; this yielded a time step of 0.7 ms. We calculated two envelopes. The first one had a high time constant for the low-pass filter (30 ms) and captured only rough energy bursts in the signal. It was used to select a time interval where the curve was above a threshold and then deduce the DR duration (feature f_3, Table 5.2; Fig. 5.7, lower left plot). The second enve-

[8]In that case, the envelope is recalculated using the new filter but the spectrum centroid and maximum peak are unchanged.

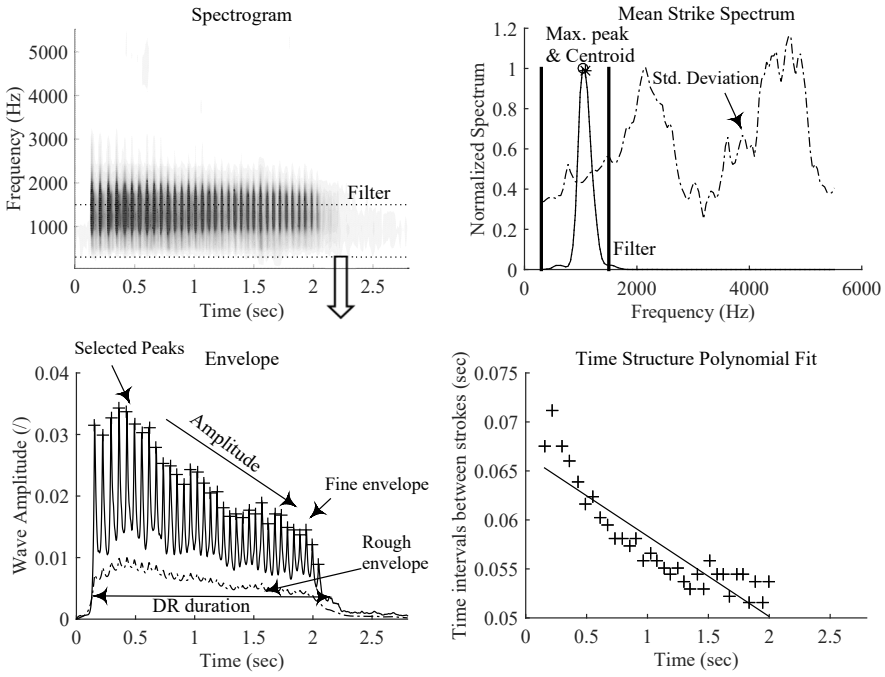


Figure 5.7: Acoustic Parameter Calculation from Spectrum and Envelope (*D. martius*, XC83280)

lope had a short time constant (10 ms) and retained more details from the original waveform. The individual strikes were visible and well detached. This envelope was used for peak picking and the determination of the time structure. Peaks located within the DR duration interval were selected. The fringes were checked for additional peaks that approximately matched the time structure. The number of strikes in the DR was set as the number of selected peaks (feature f_4, Table 5.2; Fig. 5.7, lower left plot).

The time structure is a first-degree polynomial fit through the (t,δt) data where t is the time position of a strike and δt the time interval between this strike and the next one (Fig. 5.7, lower right image). The final polynomial was modified to match the form in Zabka [91], i.e. time interval versus strike number. The two polynomial coefficients were saved as acoustic parameters for the classification. They are the initial time interval (feature f_1, Table 5.2) and the delta interval (feature f_2, Table 5.2). The delta interval is the slope of the polynomial or the difference in the duration of two successive intervals. In mechanical terms, it relates to acceleration, while the initial time interval

is the inverse of speed.

Fig. 5.7 summarizes the calculation of features f_1 through f_7 for a typical DR. The top left image is the spectrogram of the segment; the boundaries of the band-pass filter are showed with dashed lines. The bottom left image contains the envelopes and the peak selection. The DR duration and amplitude trend are annotated. The bottom right image shows the time intervals involved in the determination of the time structure polynomial. We saw other examples in Fig. 5.4 and 5.6. The cases in which the amplitude of the strikes is not much greater than the background noise are the most challenging. This occurs with faint drums or in reverberant environments. The odd superposed signals from other birds can be set aside by modifying the filter.

Parameter mean values are documented in Table 5.3 in the summary section. Fig. 5.8 and 5.9 show box plots for the time structure (initial interval and delta interval), the number of strikes and the maximum spectral peak. The light grey boxes are used to differentiate the small classes (*D. syriacus* and *P. viridis*), for which the results should be treated with caution. As expected, there is a strong correlation between the number of peaks and the DR duration (0.91); see Table 5.4 in the summary section as well. All outliers were checked manually for the time structure and the number of strikes. Because of the short DR duration for *D. major*, there were fewer peaks available for a robust polynomial fit and ultimately more algorithm failures for this taxon.

Most species accelerate while drumming (negative polynomial slope in Table 5.3 and Fig. 5.8(b); the time interval between strikes diminishes), with *D. major* exhibiting the largest gradients. *D. minor* and *P. canus* are the two significant exceptions amongst the species with large datasets; here the values indicate slight deceleration or constant speed drumming. This is coherent with the fact that *D. minor* and *P. canus* are the fastest drummers, as reflected by their short initial intervals (Table 5.3 and Fig. 5.8(a)). Starting at a high speed, they tend to decelerate. *P. tridactylus* is the overall slowest bird, with multiple initial interval values far above the likely range of 40–90 ms inferred from Zabka [91] and marked on Fig. 5.8(a). We note that pairs of twin species (*D. major* and *D. syriacus*; *P. canus* and *P. viridis*) have similar time structures but differ in the duration of the DR or in the number of strikes. This resonates with Miles et al. [55]; these authors observed that the drumming speed was constrained by body size, which twin species have in

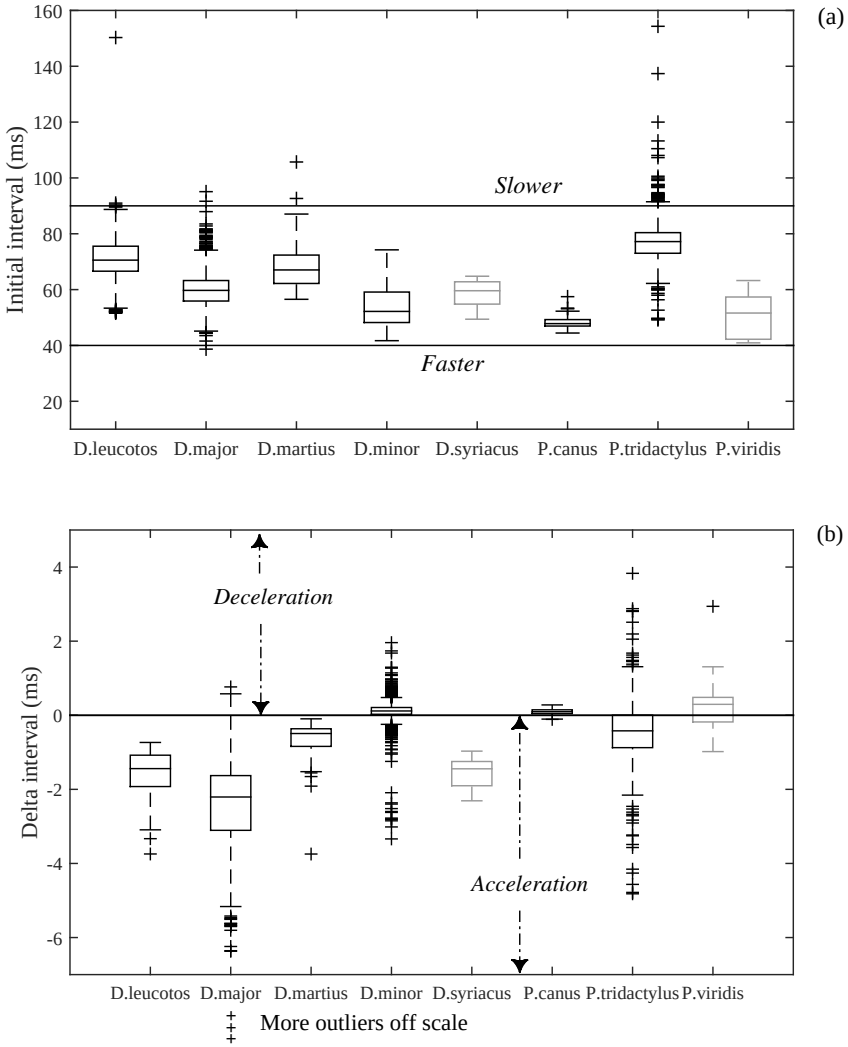


Figure 5.8: Parameter Distributions per Species for (a) Initial Interval (b) Delta Interval

Light gray boxes are used for species with less than 30 samples.

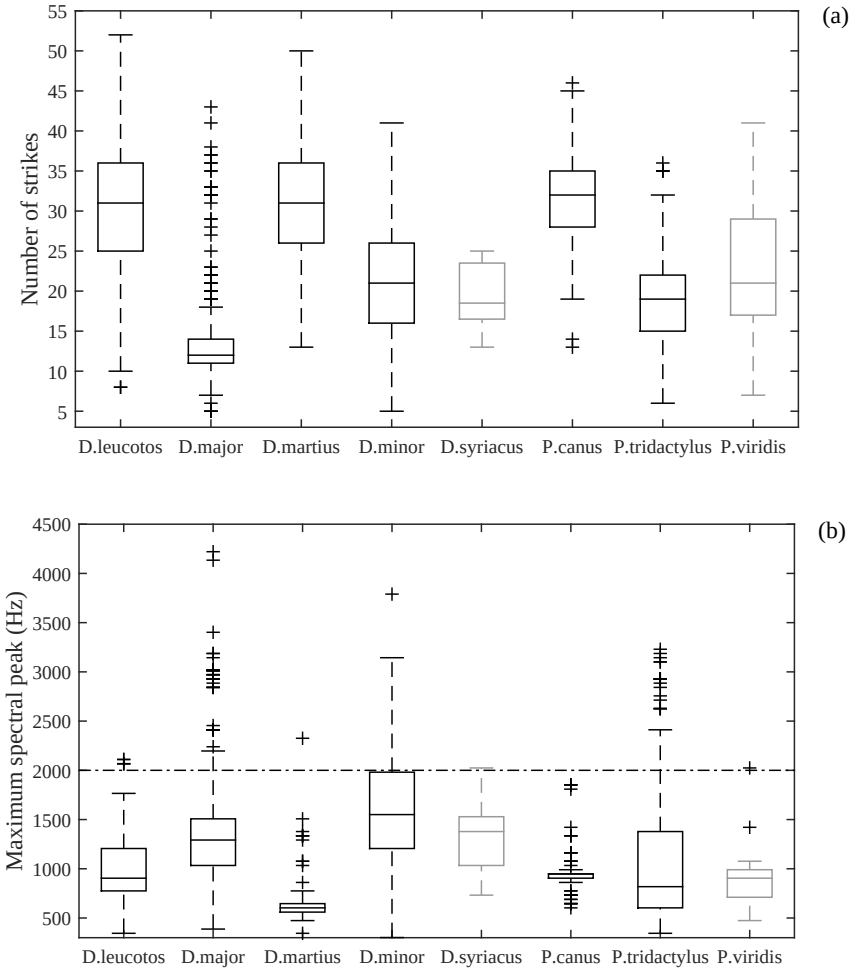


Figure 5.9: Parameter Distributions per Species for (a) Number of Strikes (b) Maximum Spectral Peak

Light gray boxes are used for species with less than 30 samples.

common, whereas the DR duration was a free parameter.

Table 5.3 and Fig. 5.9(a) show that *D. major* consistently has the lowest number of strikes per DR, while *D. leucotos*, *D. martius* and *P. canus* have the greatest numbers of strikes. There is some correlation between the delta interval and the DR duration (0.46) as it is difficult to produce long drums while accelerating. For example, *D. major* has the steepest accelerations but the shortest drums. *D. martius* has the longest drums but moderate accelerations (and it decelerates at the end of long drums, as we see in Fig. 5.10). The decelerating *D. minor* and *P. canus* are both able to produce drums with more than 40 strikes.

Fig. 5.10 illustrates the limitations of the polynomial fit in a number of examples. The accelerating species accelerate strongly at first, then when the

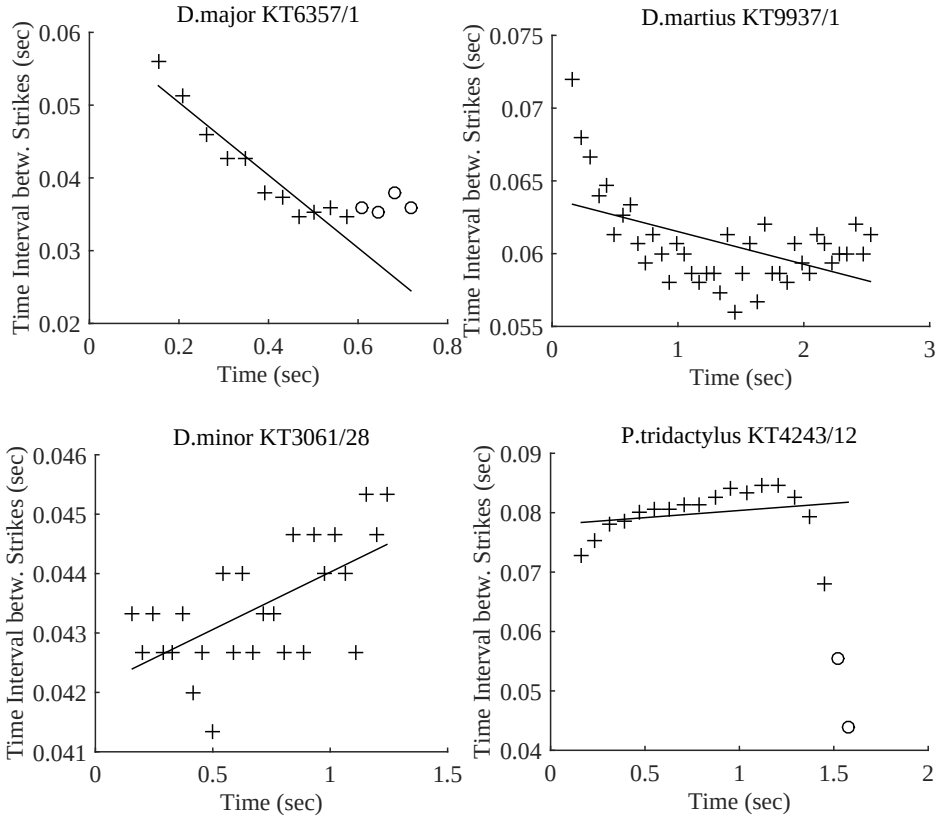


Figure 5.10: Failures in Line Fits through Time Intervals between Strikes

Circles indicate rejected data.

drum is long enough, such as is often the case for *D. leucotos* or *D. martius*, the bird hits a speed ceiling and needs to decelerate toward the end of the drum. In the *D. major* example (top left), the last four points were discarded (manually) and thus the steep slope of the initial descent was well captured by the line fit. In the *D. martius* example (top right), all data points are considered, including in the final plateau and deceleration; the slope of the initial descent is strongly diminished. The accelerating species are the species for which Zabka used an exponential form $\delta t = b \cdot exp^{-a \cdot t}$. This provides a better fit to the data in cases similar to the *D. major* example above, but not in the *D. martius* example. For our matter, the critical point is that the different species produce a characteristic range of line slopes, distinct from the other species, off-trend points included or not. In Fig. 5.8(b), *D. leucotos*, *D. major* and *D. martius* appear to successfully separate on their delta interval values[9]. There might be limited value in using a more complex fitting function. For *D. major*, the drums are short and the end part where the speed stabilizes is rarely present. Perhaps a focus on the initial descent only would lead to more consistent results, i.e. more compact boxes in Fig. 5.8(b). Note how the decelerating species, *D. minor* and *P. canus*, have thin boxes.

The shape of the fitting function is also an issue for some *P. tridactylus* samples. When the drum is long enough, the time intervals align along a bell curve, either partial of full (Fig. 5.10, bottom right; see also Fig. 5.6). In Fig. 5.10, the last two data points were discarded manually; the fitting line would otherwise have had a negative slope. Considering the full XC/TS dataset, the delta intervals for *P. tridactylus* span a wide range of values, both positive and negative (Fig. 5.8(b)). Zabka used a line fit as well. The delta interval is obviously not an optimal feature to characterize *P. tridactylus*. The most singular trait of this taxon appears to be a low drumming speed, evidenced in the initial interval numbers.

For the above three examples (*D. major*, *D. martius* and *P. tridactylus*), a possible improvement would be to consider either a higher-degree polynomial, or multiple measurements: the line fit through the first 8 points (a characterization of the steep initial slope), the one through the first 80% of the data (a characterization of the global slope through the data, minus the end point effects), and the average of the final 5–6 points (a characterization of the drum end). This could be completed by the maximum and the

[9] *D. major* and *D. syriacus* do not separate, but these are closely related species; see discussion two paragraphs up.

minimum values in the intervals between strikes.

In the last example in Fig. 5.10 (*D. minor*, bottom left), the first degree polynomial is accurate enough, but the bird is too fast for the resolution of the calculated signal envelope: the time intervals can only take discrete values which are multiples of the envelope time step. In any case, *D. minor* produces a slight deceleration; in the example all the time intervals lie in the range 41.0-45.5 ms. The graph illustrates the limitations in the design of our approach, but does not jeopardize the differentiation of *D. minor* from the accelerating species. Alas, the speed and acceleration of *D. minor* are terminally similar to those of *P. canus*. A finer envelope would not solve this issue.

Amplitude Slope

A second polynomial fit was produced on the amplitude of the peaks (normalized so that the maximum amplitude is one). All trends were observed in the data: increasing, decreasing and constant intensity, as well as more complex variations. Only the slope coefficient was saved (feature f_5, Table 5.2; annotation in Fig. 5.7). From the summary in Table 5.3, *D. major*, *D. syriacus*, *D. leucotos* and *D. martius* exhibit overall decreasing peak amplitudes, while other species appear to drum with a constant intensity. Amplitude is correlated with the delta interval (0.63). This indicates that accelerating the drum comes at the expense of maintaining a stable strike strength. *D. major*, *D. syriacus*, *D. leucotos* and *D. martius* are the accelerating species.

In replay experiments, Garcia et al. [28] observed that *D. major* still responded to a *D. major* drum in which the acceleration had been artificially removed (the initial speed was maintained). Only when both the acceleration and the amplitude decay were replaced did the birds stop responding[10]. This is actually compatible with our results. The *D. major* data in Fig. 5.8(b) shows that a range of accelerations are possible for this taxon. Steep accelerations are specific to *D. major* but not always present. Most importantly, *D. major* has a characteristically low number of strikes, and this was left untouched by Garcia et al. When the acceleration is lessened or removed, and the speed, number of strikes and amplitude decay are conserved, the drum remains in the design space of *D. major*. When the amplitude decay

[10]The results of Garcia et al. provide an interesting counterview to our work but to date are only available in a poster form. With the authorization of the authors, we reproduced the text and illustrations in Appendix C.

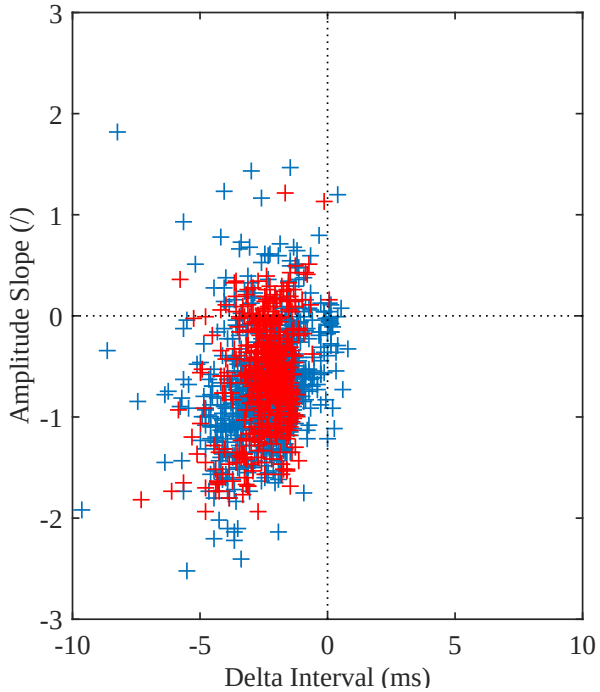


Figure 5.11: Amplitude Slope versus Delta Interval in *D. major*

Blue: XC/TS dataset;
Red: territorial drums from the KT dataset.

is removed, then the drum exits the design space. In Europe, the species that drum at more or less constant speed and with a constant intensity are *D. minor* and *P. canus*. *D. minor* sometimes uses a low number of strikes and speeds comparable to *D. major* (Fig. 5.8(a) and 5.9(a)). Garcia et al. showed that *D. major*, expectedly, responds to neither *D. minor* nor *P. canus*.

Fisher's discriminating power (Table 5.5) shows that the amplitude slope is a secondary indicator of species, unlike the delta interval to which it is correlated. Even though the mean amplitude slope values documented in Table 5.3 indicate a clear difference between the species with strikes of decreasing strength and the species with strikes of constant strength, the distributions are wide, span positive and negative values and have numerous outliers. The scatter plot in Fig. 5.11 shows that a tolerance exists in *D. major* for both positive amplitude slopes and null accelerations, but not simultaneously. The requirement for a marked acceleration is drastic in territorial drums, with no comparable restriction on the amplitude slope. In the end, the width of the distribution neutralizes the discriminating power of the amplitude slope.

Statistics on the Final Parameter Set

Table 5.3 summarizes the mean parameter values for all species. Table 5.4 documents the correlations between the different parameters. The parameter values for the time structure and DR duration are remarkably in line with Zabka's numbers for the classes that are well represented in both studies (this study: Table 5.3; Zabka: Table 5.1). The delta interval value is particularly stable. This speaks to the independence of these results from the content and the size of the database. Zabka's sound files were not recovered for the present study and the ratio in database size is 11 to 1.

Table 5.3: Mean Parameters Values

Species	Number of Files	First Interval (ms)	Delta Interval (ms)	DR Duration (s)	Nb. of Strikes	Amplitude Slope (/)	Spectrum Centroid (Hz)	Spectrum Peak (Hz)	Time betw. Rolls[a] (s)
D.leucotos	248	71.3	-1.6	1.60	30.1	-0.45	1206	1041	21.9
D.major	818	60.1	-2.4	0.65	13.0	-0.68	1481	1323	12.7
D.martius	84	68.7	-0.7	1.86	31.0	-0.22	801	683	14.4
D.minor	832	53.6	0.1	1.16	21.1	0.07	1708	1581	6.8
D.syriacus	8	58.6	-1.6	0.97	19.4	-0.58	1486	1330	20.1
P.canus	104	48.3	0.1	1.57	31.5	-0.03	1116	965	19.1
P.trid.	547	77.6	-0.5	1.36	18.9	0.07	1231	1041	20.2
P.viridis	16	50.6	0.3	1.27	23.0	-0.02	1150	918	23.7

[a]The maximum allowable time between drum rolls is 5 min compared to 60 s in Florentin et al. [21].

Table 5.4: Parameter Correlations

	Delta Interval	First Interval	DR Duration	Nb. of Strikes	Spectrum Centroid	Spectrum Peak	Amplitude Slope	Time betw. Rolls
Delta Interval		-0.26	0.46	0.39	0.11	0.11	0.63	-0.03
First Interval			0.25	-0.05	-0.26	-0.24	0.02	0.23
DR duration				**0.91**	-0.15	-0.15	0.35	0.12
Nb. of Strikes					-0.12	-0.11	0.23	0.09
Spect. Centroid						0.**83**	0.07	-0.17
Spectrum Peak							0.06	-0.15
Amplitude Slope								-0.02

Table 5.5 shows Fisher's discriminating power for all parameters and the variation for the ones with positive values. Fisher's discriminating power D_k is calculated to calibrate expectations regarding the acoustic indicator k (Eq. 5.1). Note that the formula gives a greater weight to large classes:

$$D_k = \frac{\sum_{i=1}^{C} n_i \cdot (\mu_{ik} - \mu_k)^2}{\sum_{i=1}^{C} n_i \cdot \sigma_{ik}^2} \quad (5.1)$$

C is the number of classes, n_i the number of samples of class i, μ_{ik} the mean of the k^{th} indicator over class i samples, μ_k the mean of the k^{th} indicator over all samples and σ_{ik}^2 the variance of the k^{th} indicator over class i samples.

Following Zabka [91], the variation V_{ik} for the k^{th} indicator over class i samples is computed as the standard variation over the mean (Eq. 5.2). Table 5.5 documents its average over all classes.

$$V_{ik} = \frac{\sigma_{ik}}{\mu_{ik}} \quad (5.2)$$

We obtained 11% for the first interval and around 30% for DR duration, number of strikes and spectrum centroid. This compares reasonably well to Zabka (in Table 5.1: 10% for the initial interval, 10–30% for the DR duration, 15–30% for the number of strikes). With a variation of 91%, we confirm that the time interval between rolls is a volatile parameter. The discriminating powers are in agreement with these numbers: the time structure parameters are the critical ones, with DR duration, number of strikes and perhaps amplitude in second line. Other parameters (spectral parameters, time between rolls) are less promising.

The numbers in Table 5.5 prompted us to eventually drop the time interval between rolls ; the computational effort was not worth it. We were left with seven parameters per drum roll, of which three were discriminating and independent.

All features were normalized prior to classification, using max-min normalization: for the k^{th} sample in the dataset, the value of the i^{th} parameter was projected to the $[0, 1]$ interval following Eq. 5.3. The maximum and minimum are over the full dataset.

$$p_k^{(i)} = \frac{(p_k^{(i)} - p_{min}^{(i)})}{(p_{max}^{(i)} - p_{min}^{(i)})} \quad (5.3)$$

Table 5.5: Parameter Prospective Quality

	Fisher Discriminating Power[a]	Variation[a] (%)
First Interval	1.57	11%
Delta Interval	1.24	
DR Duration	1.01	29%
Number of Strikes	0.85	29%
Amplitude Slope	0.76	
Spectrum Centroid	0.37	27%
Spectrum Peak	0.27	37%
Time between Rolls[b]	0.17	91%

[a]Figures from Florentin et al. [21] were re-evaluated after removal of *D. medius*.

[b]The maximum allowable time between drum rolls is 5 min compared to 60 s in Florentin et al. [21]. The standard deviations increase accordingly.

5.2 Mapping and Classification of Drums

In this section we discuss two mapping techniques, Linear Discriminant Analysis (LDA) and t-Distributed Stochastic Neighbor Embedding (t-SNE), and one classification scheme, k-Nearest Neighbor (k-NN). Mapping is a transformation of the parameter set into a new, low-dimensional space, where coordinates are decorrelated. This allows a straightforward visualization of the dataset using a 2D or 3D scatter plot. Ideally, when the original parameters capture the critical differences that exist between the species, then the different species occupy different regions of the final map. LDA is an extension of Fisher's Discriminant and t-SNE a nonlinear optimization scheme[11]. Classification uses either the original or the transformed coordinates to identify species in test samples.

Some of the visualizations in the present section incorporate, in addition to the XC/TS dataset, 2215 drums recorded in Tenneville (2133 *P. canus* and 82 *D. martius*[12]) and 2769 drums (all species but *D. medius)*, both soft and territorial, recorded by Kyle Turner.

[11]Details in App. A.

[12]These identifications were intermediate results, detected through analysis of repetitions and identified using the XC/TS dataset as reference. Actualized numbers are provided further down in the text.

5.2.1 Linear Discriminant Analysis (LDA)

The results of the LDA canonical analysis for the combined XC/TS and TN dataset are in Table 5.6. Following the χ^2_{obs} calculation suggested by Dagnelie [13], it appears that the first five eigenvalues have significance. There is a gap between the importance of the first three and the next two. This is consistent with the prior assessment of our parameter set: two parameters in the set are highly correlated to others (five significant parameters remain) and three parameters (delta interval, initial interval and DR duration) have the bulk of the discriminating power[13]. Looking at the eigenvectors, there is a visible alignment of the first three with delta interval, initial interval and number of strikes. The number of strikes being highly correlated with the DR duration, this is similar to the previous result. The fourth eigenvector is again the initial interval, along with a combination of DR duration and number of strikes that seem to cancel each other. The fifth one introduces the spectral parameters. The amplitude slope makes no significant contribution to any of the eigenvectors.

Fig. 5.12 shows the dataset projected onto the first two eigenvectors and with boundaries between species. The segregation of species is uneven. *D. leucotos* for example spills into the neighboring regions; its own region has little homogeneity. The cloud of *P. canus* points is quite compact[14], but on top of the *D. minor* region. Fig. 5.13 addresses the question about whether other dimensions would offer a better separation of the classes. It displays the LDA 2D maps for all possible pairs of eigenvectors. These maps also show the 90% confidence ellipses for *P. canus* and *D. minor*[15]. We see in Fig. 5.13 that *P. canus* and *D. minor* almost always completely overlap. The greatest separation occurs when the third vector is involved, i.e. the number of strikes. As a second dimension, marginal features like the spectral parameters (5^{th} eigenvector) give better results than the initial interval (1^{st} eigenvector). For other species, eigenvector two efficiently differentiates *D. minor* and *D. major*; the combination of one and three works well for *P. tridactylus*; two and three for *D. leucotos* and to a lesser extent for *D. martius*. Like *P. canus*, *D. martius* seems difficult to separate from the rest.

[13] A logical convergence, as LDA is a generalization of Fisher's discriminant.

[14] Expectedly: the 2133 *P. canus* drums from TN all belong to the same individual.

[15] It is more common to use 95%, but there is not enough class separation with the higher value.

Table 5.6: LDA Eigenvalues and Eigenvectors

	Vect. 1	**Vect. 2**	**Vect. 3**	**Vect. 4**	**Vect. 5**	**Vect. 6**	**Vect. 7**
Eigenvalues	2.611	2.358	0.915	0.170	0.031	-0.001	-0.002
χ^2_{obs}	16275	9999	4078	903	135	-14	-8
Physical Parameters							
Delta interval	-5.9	**14.5**	9.2	-9.8	8.6		
Initial interval	**-26.5**	8.4	-5.9	**-22.1**	7.3		
DR duration	0.0	4.0	8.7	**31.2**	**-12.8**		
Number of strikes	2.8	3.2	**-15.4**	**-28.2**	6.2		
Centroid	-0.9	-5.8	6.1	-7.8	**-15.1**		
Maximum peak	4.0	2.8	-2.2	2.0	**12.5**		
Amplitude slope	1.2	2.5	3.7	-1.2	0.0		

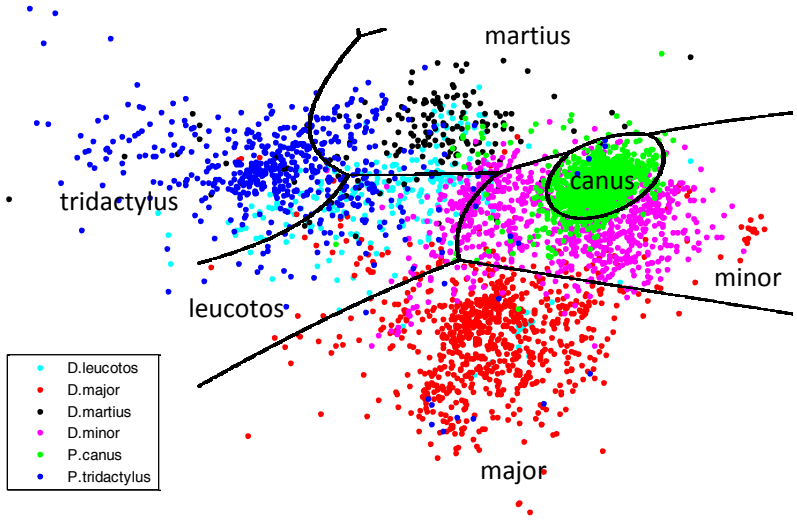


Figure 5.12: LDA Projection Using Eigenvectors 1 & 2

Note to the reader: not all printers render this plot faithfully. If the darker colors cannot be distinguished, please consult a pdf version of this text. Here the *P. tridactylus* samples spill into the *D. martius* and *D. leucotos* regions.

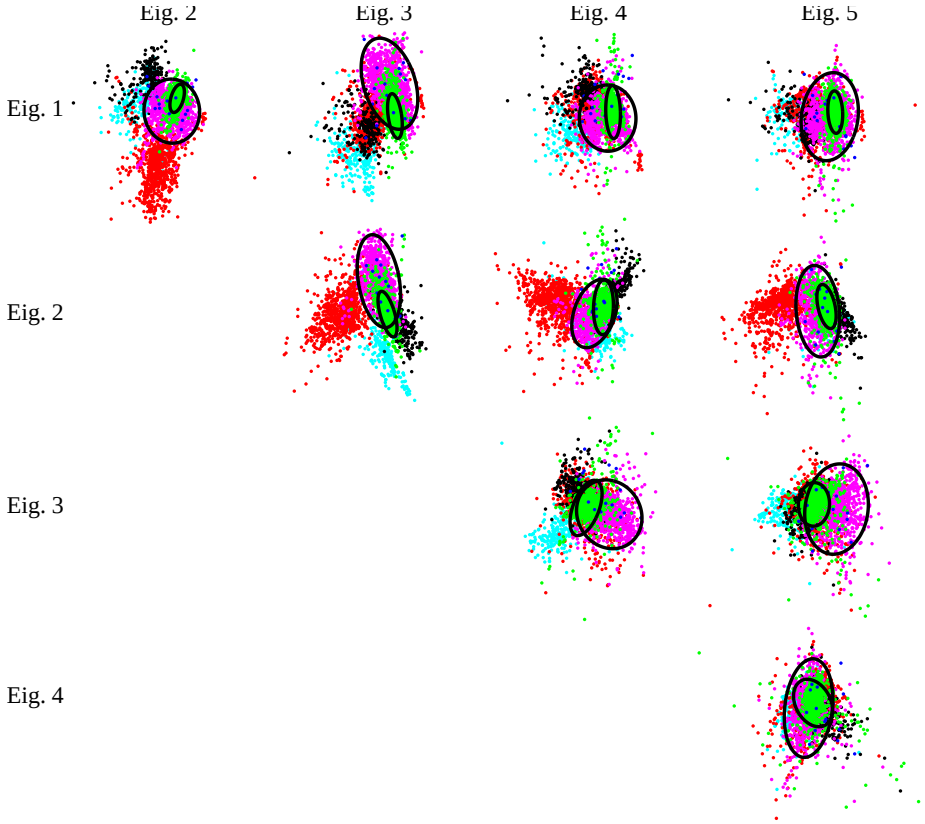


Figure 5.13: All 2D LDA Projections with Ellipses for *D. minor* & *P. canus*

5.2.2 t-Distributed Stochastic Neighbor Embedding (t-SNE)

A t-SNE map for the XC/XS dataset is presented in Fig. 5.14. The colors are added a posteriori to visualize the different classes. Here, they are clustered into separate regions of the map, which indicates that the parameter set allows discriminating the different species. The figure also showcases the ability of t-SNE to exploit small variations in parameters to spread out classes. Although the point coordinates in t-SNE maps have no physical meaning, the optimization seemingly exploited DR duration in the horizontal direction and drumming speed in the vertical direction. *D. major* with the shortest DRs is to the left, whereas *D. leucotos*, *D. martius* and *P. canus* with DRs longer than 1.5 s are to the right. The fast *D. minor* and *P. canus* are at the top and the slow *P. tridactylus* at the bottom. All the accelerating species are in the bottom part of the figure. As it seems, the parameters with the most discriminating power were prioritized.

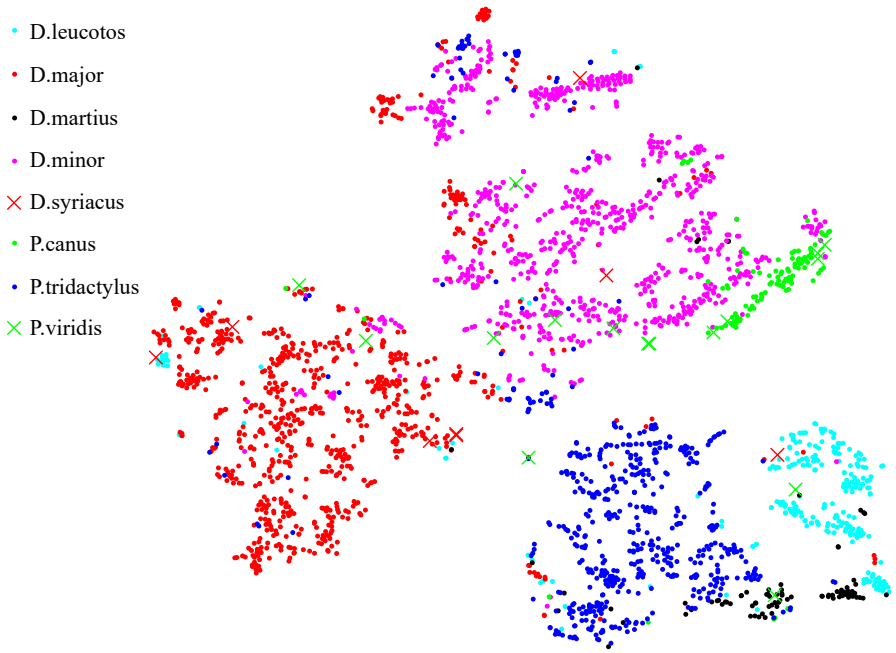


Figure 5.14: t-SNE Map with 2657 XC/TS Drums

Note to the reader: not all printers render this plot faithfully. If the darker colors cannot be distinguished, please consult a pdf version of this text. Here the *D. martius* samples are clustered at the bottom right of the *P. tridactylus* region and create a bridge to the *D. leucotos* region. The overlap between the three groups is minimal.

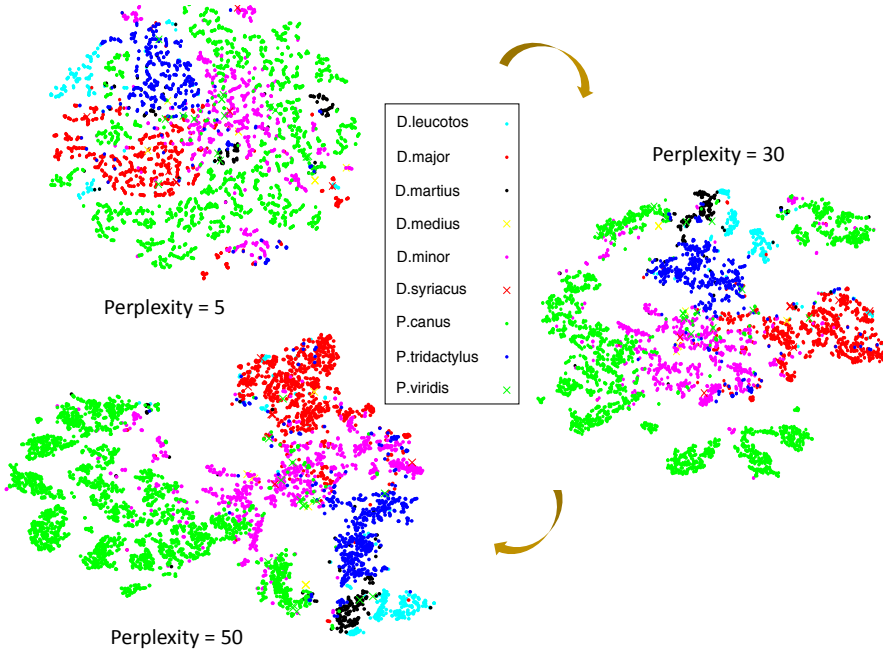


Figure 5.15: t-SNE Maps as a Function of Perplexity

We used a perplexity of 30 in Fig. 5.14, as it seemed to aggregate the largest classes well. Perplexity relates to the size of the typical core group of samples that can be considered as close neighbors. The optimization behind the t-SNE map is primarily influenced by two factors: the perplexity and the order in which the data points are supplied.

Fig. 5.15 shows variations in the t-SNE map in function of perplexity for the XC/TS and TN datasets. In that case, using 5 is too small and 30 not as efficient as 50. The optimal perplexity depends on the size of the classes. It cannot be efficient to describe a class of 2000 samples by making groups of 5. Here we have classes with 800 samples (e.g. *D. minor*, *D. major*), 2000 samples for *P. canus* and numbers in the hundreds for the medium size classes.

On the maps, it might seem as though *P. canus* is capable of using an extraordinary range of variation in drumming parameters compared to other species but this is not the case. t-SNE makes space for points: if a group is large, it will occupy a large region. The purpose is to see all points rather than to crowd certain hot spots. The fact that there seems to be sub-clusters[16]

[16]Further investigation would be required to determine why t-SNE forms these sub-clusters.

in the *P. canus* group is an artifact. We actually see the groups of 50 points.

Fig. 5.16 shows additional maps, this time with 7641 samples from XC/TS, TN and KT. In each map, the samples in the dataset were randomly permuted, and thus the optimization proceeded from a different starting point and toward a different minimum. A number of variations were tried, yet we failed to keep all species in compact groups. In the three examples shown, first *P. tridactylus*, then *P. canus*, then *D. minor* appears in split regions. With few samples available, the positioning of the soft drummers remains ambiguous. *P. viridis* can be spotted on the margins of *P. canus* and *D. minor*. The various t-SNE maps presented in this section capture the overlap between these three species well. The proximity of *P. tridactylus*, *D. martius* and *D. leucotos* is recurrent as well. Thanks to Kyle Turner's *D. syriacus* samples, the closeness of *D. syriacus* and *D. major* can be visualized. These pictures are quite successful at representing the inner structures of drumming datasets and thus offer a preview of what will be achieved through classification.

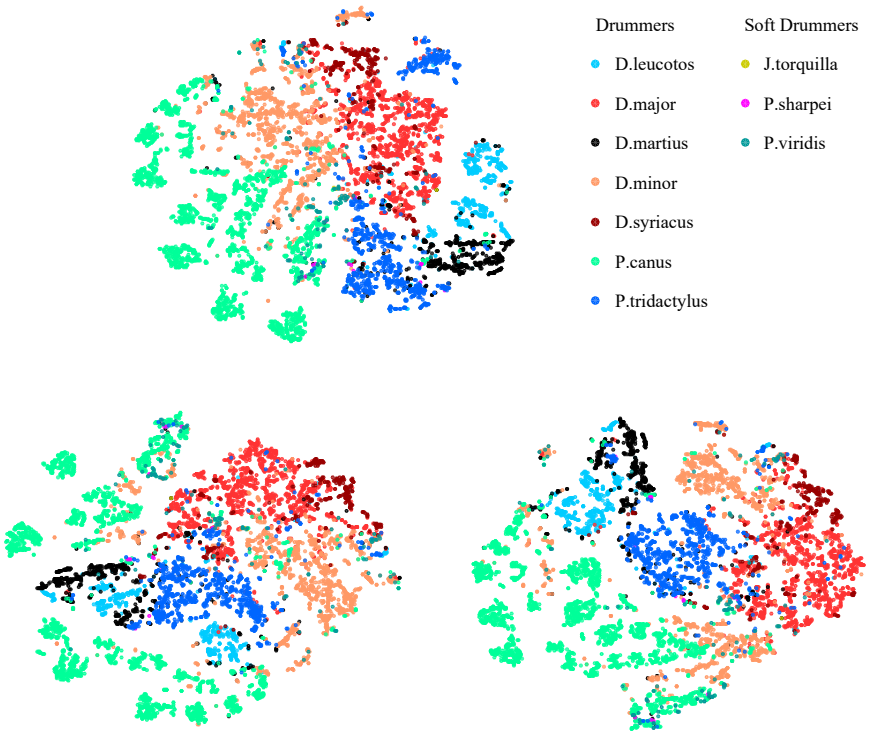


Figure 5.16: Different Drumming Galaxies Generated from Permutations of the Same 7641 Samples (Perplexity=50)

5.2.3 Classification Using k-NN

With seven parameters per drum, eight if the time between rolls is included, we constructed a highly compressed representation of drums. Through the t-SNE maps, we controlled that this representation was faithful to the inner structures of the datasets and allowed a segregation of the classes. At this point, the interpretative work on the data is done. There is little value in using a classifier with sophisticated analysis capabilities, such as a deep net. There is also no point in using a classifier intended to tackle large parameter sets, such as a random forest. Our remaining task consists in connecting seven (eight) normalized drum parameters to class assignments. This can be done in a straightforward manner by classifier k-NN.

We used k-NN with $k = 5$, i.e. test samples were matched to their five nearest neighbors in the training set. The point coordinates were the drum parameters and the distance metric the Euclidean distance. The most frequent class among the five neighbors was assigned to the test sample under consideration[17]. With $k = 1$, there is a risk that the test samples are matched to anomalies in the training base; with values of k greater than 5, the samples are abusively matched to dominant groups. Trial runs showed that a training set comprising 20% of the total database was adequate. The accuracy of the classification plateaued for greater percentages[18]. A minimal representation of the smaller classes had to be forced; see the final training set composition in Table 5.7. For *D. martius*, *P. canus* and *P. viridis*, approximately 50 training samples were selected, and for the larger classes, approximately 100. The sample selection for the training set was restricted to samples having a value defined for the time interval between rolls (feature f_8).

As seen in Table 5.7, we experimented with two datasets, the aforementioned XC/TS dataset and the KT dataset (Chap. 3, Section 3.4.2). The KT dataset has a greater supply of samples from *D. martius*, *D. syriacus*, *P. canus* and *P. viridis*. The number of samples for *D. syriacus* and *P. viridis* in the XC/TS dataset (resp. 8 and 16) was insufficient to include these classes in the study. Similarly, the KT dataset possesses 4 *J. torquilla* samples that cannot be considered here. In the end we used 2633 drums of the XC/TS dataset

[17] In case of ties, the Matlab *knnclassify* function selects the closest sample among the leading classes.

[18] Other classification problems have easily used 80%. Note that such a problem design is suitable at the development stage, but for proper validation the training set and the test set should stem from entirely different data pools.

and 2749 drums of the KT dataset. Another specificity of the KT dataset is that the soft drums are labeled as such and can thus be either included or rejected. In the XC/TS dataset, it is unknown whether the drums are territorial or soft; the presence of soft drums is however likely. The only drawback of the KT dataset is that it was collected by a single man; many signals come from the same birds or same places, whereas the crowdsourcing nature of Xeno-Canto guarantees a greater diversity.

The k-NN classification was run two hundred times, each time with a different randomly selected training set. Fewer runs would have worked as well, as eventually the standard deviation in the accuracy results was only 0.6–1.2% (Table 5.8).

Table 5.7: Training Set Composition

	XC/TS Dataset			**KT Dataset**		
	Total Files	Training Set	(in %)	Total Files	Training Set	(in %)
D. leucotos	248	83	33.5	288	74	25.7
D. major	818	113	13.8	589	70	11.9
D. martius	84	43	51.2	388	78	20.1
D. minor	832	113	13.6	528	74	14.0
D. syriacus				308	76	24.7
P. canus	104	50	48.1	130	54	41.5
P. tridactylus	547	112	20.5	418	78	18.7
P. viridis				100	46	46.0
Total			**19.7**		**20.0**	

Table 5.8: Overall Accuracy of k-NN Classification over 200 Trials

Dataset	**Features**	**Mean**	**Maximum**	**Standard Deviation**
XC/TS	8 parameters	87.8%	90.0%	0.8%
XC/TS	7 parameters	87.4%	89.4%	0.8%
XC/TS	4 parameters[a]	84.4%	86.2%	0.8%
KT (terr. drums)	7 parameters	93.6%	94.8%	0.6%
KT (all drums)	7 parameters	86.7%	88.8%	0.8%
XC/TS	t-SNE coord. Fig. 5.14	86.0%	89.0%	1.2%
KT (terr. drums)	9 t-SNE maps as in Fig. 5.16	89.1%	91.5%	1.0%
KT (all drums)	9 t-SNE maps as in Fig. 5.16	81.4%	84.3%	1.2%

[a]Delta interval, initial interval, DR duration and number of strikes.

In the XC/TS case, the accuracy[19] for the full test dataset, i.e. excluding the training set, is 87.8% (Table 5.8). The differences in the results when the time between rolls is included in the parameter set or not is marginal. The classification is based almost entirely on the main four temporal parameters (delta interval, initial interval, DR duration and number of strikes), which achieve a 84.4% accuracy on their own.

For the configuration in which all eight parameters are used, the accuracy per species and the confusion matrix are documented in Table 5.9. The scores follow the class size: the classes with more than 200 samples are identified with an accuracy greater than 85%; the smaller classes have lower accuracies (*P. canus* 78.9%, *D. martius* 67.3%). Observations made on the t-SNE map are recalled. For example, some *D. martius* samples are misclassified as *D. leucotos* and *P. tridactylus*. These three species were neighbors on the t-SNE maps. The expected confusions are observed between *D. minor* and *P. canus*. A communication channel also exists between *D. major* and *D. minor*, caused by the overlaps in parameters in these two large distributions. Some *D. major* drums are barely accelerated, some *D. minor* drums have moderate speeds and less than 20 strikes. The t-SNE maps showed a zone of interpenetration between the two groups.

Following Tables 5.8 and 5.10, the performance on the KT dataset, including only the territorial drums, is significantly higher: the overall accuracy is 93.6%. The lowest score is for *D. syriacus* (87.3%). As expected, the taxon is confused with *D. major*. No class suffers from a deficit of samples in the dataset. Compared to XC/TS, the confusions are much decreased, e.g. between *D. minor* and *P. canus* or between *D. minor* and *D. major*. Once the soft drums are included back into the dataset, the overall accuracy goes down to 86.7% (Table 5.8) and the confusions return (Table 5.11). *P. viridis* whose drums are all soft is particularly poorly predicted. This taxon notably overlaps with *P. canus*, whose accuracy drops from 95.1% when only territorial drums are considered, to 80.9% when the soft drums are included. The deterioration is the combined result of the presence of a new species in the set (*P. viridis*) and of the loss of clarity in the boundary with *D. minor* (from 1.9 to 6.8 confusions). The species character is less pronounced in soft drums. Typically, the soft drums are slightly slower and contain a few strikes less, which fuzzies the boundaries between species (Florentin et al. [23]).

Following Dupont et al. [17], we also attempted the k-NN classification

[19]Percentage of test samples that were correctly identified.

Table 5.9: XC/TS Dataset, 8 Parameters:
k-NN Confusion Matrix and Accuracy

Actual	**Predicted Classes: Mean Nb. of Samples over 200 Trials**						**Accuracy**
Classes ↓	D.leuc.	D.maj.	D.mart.	D.min.	P.can.	P.tri.	
D.leucotos	139.8	11.0	2.3	1.2	0.1	10.7	84.7%
D.major	16.8	624.6	0.0	45.8	0.6	17.2	88.6%
D.martius	5.2	0.7	27.7	2.6	0.4	4.3	67.6%
D.minor	0.9	34.6	1.8	646.6	27.0	8.0	89.9%
P.canus	0.1	1.5	0.8	8.0	42.6	1.0	78.8%
P.tridactylus	3.3	13.9	11.0	26.3	1.9	378.7	87.1%

Table 5.10: KT Dataset (Territorial Drums), 7 Parameters:
k-NN Confusion Matrix and Accuracy

Actual	**Predicted Classes: Mean Nb. of Samples over 200 Trials**							**Accuracy**
Classes ↓	D.leuc.	D.maj.	D.mart.	D.min.	D.syr.	P.can.	P.tri.	
D.leucotos	181.6	2.0	8.5	0.1	3.0	0.2	3.5	91.3%
D.major	0.0	352.0	0.0	5.2	31.2	0.5	1.1	90.2%
D.martius	7.8	0.8	154.8	1.1	0.3	1.1	2.1	92.1%
D.minor	0.0	1.4	0.0	367.9	1.8	2.9	0.0	98.4%
D.syriacus	1.8	21.1	1.5	0.6	172.9	0.1	0.0	87.3%
P.canus	0.6	0.4	0.2	1.9	0.1	60.9	0.1	95.1%
P.tridactylus	1.0	2.2	2.0	0.9	0.6	0.1	311.2	97.9%

Table 5.11: KT Dataset (All Drums), 7 Parameters:
k-NN Confusion Matrix and Accuracy

Actual	**Predicted Classes: Mean Nb. of Samples over 200 Trials**								**Accuracy**
Classes ↓	D.leuc.	D.maj.	D.mart.	D.min.	D.syr.	P.can.	P.tri.	P.vir.	
D.leucotos	192.0	1.3	12.1	0.1	3.6	0.6	4.3	0.0	89.7%
D.major	0.2	446.2	1.5	7.6	58.6	1.1	1.0	2.8	86.0%
D.martius	21.9	9.9	249.8	1.7	3.6	1.9	21.0	0.2	80.6%
D.minor	0.0	10.8	0.1	412.6	3.3	3.7	0.0	23.3	90.9%
D.syriacus	1.4	28.5	2.0	1.4	198.3	0.4	0.1	0.0	85.5%
P.canus	0.4	0.1	0.8	6.8	0.1	61.5	0.2	6.1	80.9%
P.tridactylus	2.2	1.9	13.9	0.9	0.3	0.1	320.5	0.2	94.3%
P.viridis	0.0	3.2	0.1	13.7	0.4	10.4	0.6	25.5	47.1%

Table 5.12: k-NN Classification: Between the Two Datasets

The KT dataset includes the soft drums, as the XC/TS dataset includes them too. The analysis was run using only the main four temporal parameters (delta interval, initial interval, DR duration and number of strikes).

	Accuracy	
Training	KT	XC/TS
Test	XC/TS	KT
D. leucotos	77.4%	94.4%
D. major	83.7%	93.0%
D. martius	83.3%	67.8%
D. minor	42.0%	85.2%
D. syriacus	75.0%	4.5%
P. canus	50.0%	25.4%
P. tridactylus	82.6%	94.7%
P. viridis	18.8%	2.0%
Overall	68.1%	71.4%

on the low-dimensional space, i.e. on the coordinates of the t-SNE maps. For the XC/TS dataset, we used the map in Fig. 5.14 and for the KT dataset, a combination of nine maps including the three in Fig. 5.16. The overall accuracy is documented in Table 5.8. For both datasets, with or without the soft drums, the results deteriorated. Therefore t-SNE does not interpret the data in a way that facilitates classification. The projection to a low-dimensional space, even to several low-dimensional spaces, induces a loss of information.

Note that the split between test and training data in all of the above results does not take into account the source recording. The same bird might have produced several drums that were indifferently distributed between the training and the test sets. This inflates the results. The effect is even more pronounced for the KT dataset in which the same birds were recorded at length.

Hence the results in Table 5.12 should offer a more balanced view of the methodology, and temper expectations. Here the data from one dataset was used as a training base to identify the data from the other one, and the accuracy fell to around 70%. The XC/TS dataset appears as a better training set; it still produces an accuracy over 85% for *D. leucotos*, *D. major*, *D. minor* and *P. tridactylus*. At the same time, *D. syriacus* and *P. viridis* are poorly recognized because these classes are under-represented in the XC/TS set. When the KT dataset becomes the training set, then the classes that were previously successful lose at least 10%. *D. minor* falls to 41.9%. This

is imputable to the lack of diversity in the KT dataset. On the contrary, the *D. martius* and *P. canus* classes are better populated, and their results improve. Still, for *P. canus*, a species of primary interest in our work, the results never exceed 50.0%. The reference datasets remain largely incomplete for this taxon.

5.2.4 Identifying Drums in Field Datasets

We now contemplate the identification of the drums that were detected in the three field datasets TN, RM and LPR. The first dataset, TN, contains primarily *P. canus* drums, as the station was installed next to a known hole. The corresponding k-NN results are presented in Table 5.13.

We started with the KT set as training set and used all seven parameters. In this configuration, the *P. canus* drums were massively identified as *D. minor*. Using only the four temporal parameters (delta interval, initial interval, DR duration and number of strikes) dramatically shifted this trend. In a context where the reference database does not contain enough *P. canus* samples, the known range of frequency values for *P. canus* is too restrictive

Table 5.13: Species Identified in Tenneville by k-NN

	Training							
Set	KT (TD)	KT (TD)	KT (All)	XC/TS	KT(All)& XC/TS	KT (TD)	KT (TD)	**Truth**
Parameters	7	4	4	4	4	4	4	
Taxa	All	All	All	All	All	No minor	No leu./trid.	
	Number of Detections							
D. leucotos	6	2	0	3	0	3	0	
D. major	16	17	13	14	14	24	25	14
D. martius	127	113	118	78	96	113	122	120
D. minor	2099	163	152	1923	671	0	163	
D. syriacus	0	0	0	0	0	0	0	
J. torquilla	0	0	1	0	1	0	0	
P. canus	312	2271	2278	525	1766	2426	2272	2447
P. sharpei	0	0	2	0	0	0	0	
P. tridactyl.	22	16	13	39	34	16	0	
P. viridis	0	0	5	0	0	0	0	

Table 5.14: Training: Full KT; Test: TN ;
k-NN Confusion Matrix and Accuracy

Actual Classes ↓	**Predicted Classes** D.maj.	D.mart.	D.min.	J. torq.	P.can.	P. sha.	P.tri.	P.vir.	**Accuracy**
D.major	12	0	1	1	0	0	0	0	85.7%
D.martius	0	108	0	0	0	0	12[d]	0	90.0%
P.canus	1[a]	10[b]	151	0	2278[c]	2[a]	0	5[a]	93.1%

[a]Distant, shortened or interrupted *P. canus* drums.
[b]Long *P. canus* drums.
[c]Not reviewed.
[d]The 13th drum is... wing flaps.

and the frequency parameters are misleading[20]. Their discriminating power being low, they can be discarded without much regret. We discarded the amplitude slope as well, perhaps lightly.

There was not much difference in the predictions depending on whether the soft drums were included in the training set or not. With the soft drums, a few original proposals appeared (*J. torquilla*[21], *P. sharpei*, *P. viridis*). The results were again unbalanced toward *D. minor* when the XC/TS dataset was used for training. This is consistent with the numbers in Table 5.12; KT is a better predictor of *P. canus* than XC/TS. The bias against *P. canus* in the XC/TS set is rather strong since the combination of the two sets offered worse results than KT alone. When *D. minor* was suppressed from the training set, the predictions shifted toward *P. canus* and a few *D. major*. The data was eventually reviewed (Table 5.14). The TN dataset captures a full season of drums from one *P. canus* individual, including many failed executions of the proper drumming signal. Naturally, this disturbed our attempts at identification.

The DR duration should be the parameter to tell *P. canus* and *P. minor* apart, but there is an overlap. There is little margin to distinguish a long *D. minor* roll from a short *P. canus* one. In the field, the sound intensity helps

[20]For the *P. canus* group, the maximum peak distribution extends as follows: 560–1000 Hz for the KT dataset, 860–1000 Hz for the XC/TS dataset, 1260–1600 Hz in Tenneville.

[21]We subsequently found one *J. torquilla* call in the TN recordings. However this species has only soft drums, which occur in pairs and therefore in the context of a settled territory, not for a passing bird.

ornithologists make the difference, as *D. minor* has quieter territorial drums. Alas this property cannot be exploited in recordings because a) by default the microphones are not calibrated, b) the distance to the bird is unknown and c) the soft drums of *P. canus* are quiet too. The problem with *P. viridis* and *P. sharpei* is the same; the parameters are close, even more so because the three species are genetically close. *D. minor*, *P. canus*, *P. viridis* and *P. sharpei* form a continuous space, visible in Fig. 5.18, which shows t-SNE maps of the Tenneville drums with and without the training set.

The next group of associated species is *D. martius*, *D. leucotos* and *P. tridactylus*, of which only *D. martius* is present in Belgium. The t-SNE maps in Fig. 5.18 show this proximity, and the experiment in which *D. leucotos* and *P. tridactylus* were removed from the training set (Table 5.13) confirmed that they were confused with *D. martius*. Using the full KT set for training, no *D. leucotos* was found in the TN dataset, but 13 *P. tridactylus* were identified, which we all determined to be *D. martius*. Among the *D. martius* identifications, only the ones of April 14th were correct (Fig. 5.17). The others were long *P. canus* drums.

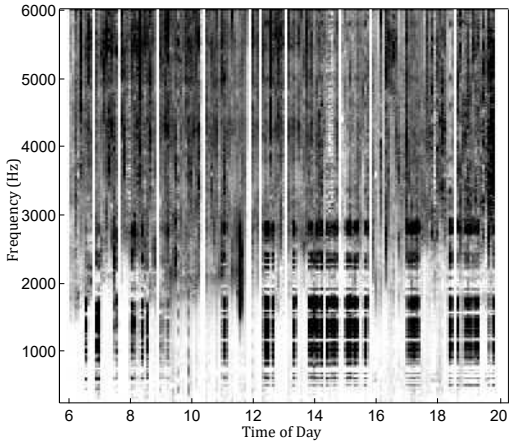


Figure 5.17: ACI Spectrogram - Tenneville 14/04/2016

The identifications from April 14th are supported by the presence of the characteristic vocalizations of *D. martius* in the recordings. On that day, *D. martius* attempted to invade the territory of *P. canus*. The ACI image sheds further light onto what went on. It is clear in the spectrogram that the morning drums have a different spectrum from the afternoon ones. The *D. martius* attack happened in the morning. The birds exchanged drums and calls; the contributions from two different birds and the rough sound of *D. martius* drums produce a somewhat fuzzy pattern in the ACI spectrogram. Apparently the *P. canus* won the fight. The *D. martius* gone, the *P. canus* spent its entire afternoon drumming to re-establish its territory, not moving from its favorite drumming spot and producing a remarkably stable drum spectrum from 11:00 AM on. This was likely a significant effort as the number of drums was divided by 10 on the next day (it went from 343 to 32).

The presence of *D. major* is highly probable nearly everywhere in Europe. It was detected sporadically on a number of days, almost always correctly. These were distant drums; in the midst of its territory, *D. major* is a much more prolific drummer. This will show in the RM and LPR datasets. The specificity of this taxon is that its drums are the shortest. Often, incomplete drums or drums for which the parameters were not properly calculated are identified as *D. major*; in the present case this happened with a shortened *P. canus* drum.

Table 5.15 shows the number of detections for the RM and LPR datasets. In Remerschen, we found a large *D. major* contingent. The *D. syriacus* identifications were mistaken *D. major*, *D. syriacus* being a bird of Eastern Europe. The *D. minor* identifications were nearly all *D. major* as well. The two species have a common ground where the *D. minor* drums are only moderately fast and shortened, and where the *D. major* drums are not strongly accelerated. Then, a few more drums were scattered in the *D. martius* group and another few in the *D. minor* group (Fig. 5.18). The three *P. canus* drums were seemingly correct. Among the calls, there was one occurrence of *P. canus*, but at another date. The drums were on a day where *P. viridis* was caught calling. *P. viridis* rarely drums, but this option cannot be ruled out. *P. canus* is not known to drum in Remerschen, the wood being too soft.

Most interestingly, the calls analysis revealed a strong territorial activity by *J. torquilla* near the microphone. Most of the *P. tridactylus* are demonstra-

Table 5.15: Species Identified in Field Datasets by k-NN

	TN			**RM**			**LPR**			
Training Set	KT (All)	XC /TS	Truth	KT (All)	XC /TS	Truth	KT (All)	KT (slope)	XC /TS	Truth
D. leucotos	0	3		0	2		5	4	35	
D. major	13	14	14	204	226	257	808	816	910	1001
D. martius	118	78	120	2	0		88	87	66	108
D. minor	152	1923		6	36		21	9	82	
D. syriac.	0	0		48	0		158	165	1	
J. torquilla	1	0		0	0	6	0	0	0	
P. canus	2278	525	2447	3	0	3	16	15	6	10
P. sharpei	2	0		0	0		0	0	0	
P. tridact.	13	39		14	14		5	5	19	
P. viridis	5	0		1	0		18	18	0	

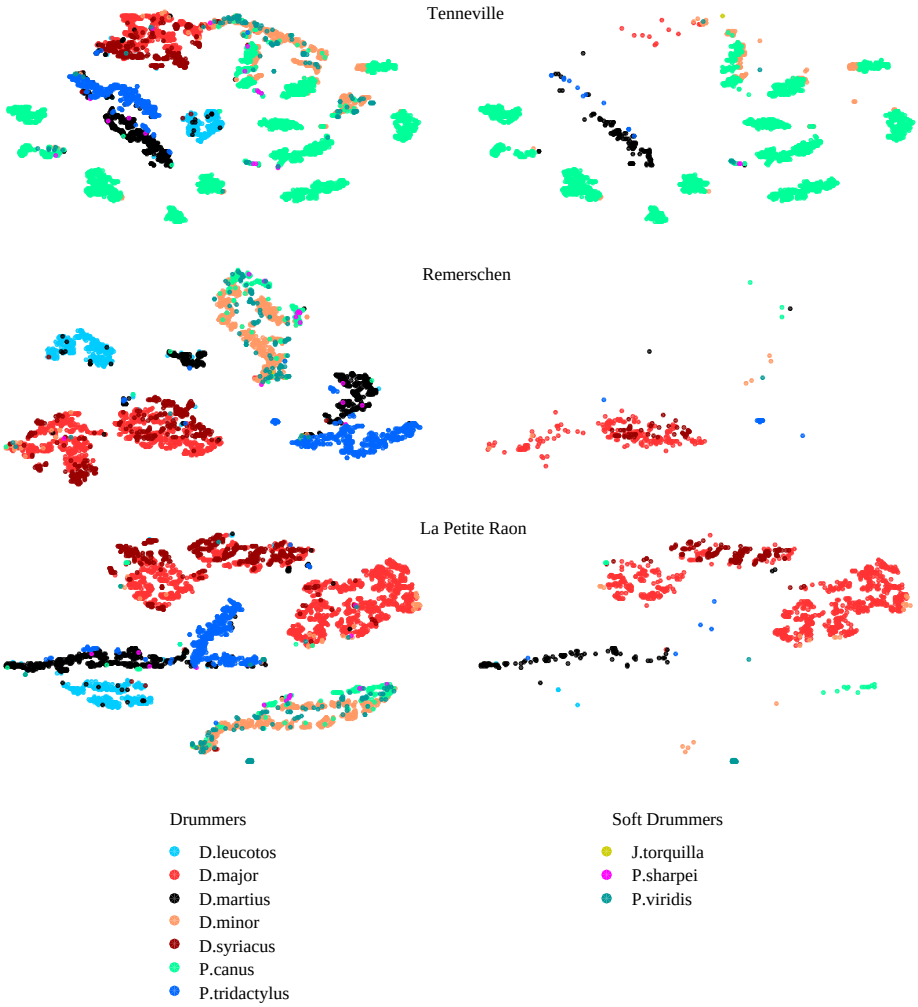


Figure 5.18: t-SNE Maps of Field Samples Identified by k-NN

Training: Full KT Set, 4 Parameters.
Left: With Training Set; Right: Field Samples Only.

tive tapping, the confusion being due to the fact that *P. tridactylus* is a slow drummer. Demonstrative tapping is reproductive behavior, which must belong to the species claiming the territory, i.e. *J. torquilla*. This leaves a list of misplaced soft drums, identified as *D. martius* (2), *D. minor* (1), *P. tridactylus* (2) and *P. viridis* (1). Many were issued in association with *J. torquilla* calls, and we concluded that they were rare recordings of *J. torquilla* soft drumming. Recordings of *J. torquilla* tapping are equally rare.

In La Petite Raon, we found again a strong *D. major* showing. There are multiple *D. major* territories at this location. We see in the t-SNE maps the detachment of two other groups (Fig. 5.18). First, the place is also a *D. martius* territory: all the *D. martius* calls are in the recordings. The *D. leucotos* and *P. tridactylus* detections were misidentified *D. martius*. Then, a second group is formed by a number of *D. minor*, *P. viridis* and *P. canus* drums. Many of these were faint drums, whose identification is delicate and often wrong, such as the distant *D. martius* drums tagged as *D. minor*. Then the undetermined nature of the *P. viridis* drums is another source of error. When the drums are short, *P. viridis* clusters along with *D. minor* at the border with *D. major*; when they are long, *P. viridis* stays with *P. canus* at the border with *D. martius* (see Fig. 5.14 and 5.16). In La Petite Raon, the drums identified as *P. viridis* were either *D. major* or *D. martius*. Only 10 of the *P. canus* were correct; the species calls a lot in La Petite Raon but barely drums. The contrast with Tenneville is striking: 10 drums versus 2447 drums. And we recorded in La Petite Raon for a longer period.

We see again that the KT dataset is better at detecting *P. canus*. It also produces less bogus *P. tridactylus* identifications. On the other hand, *D. syriacus* being well represented in the KT set, the number of confusions between *D. major* and *D. syriacus* is important. The same occurs between *P. viridis* and *P. canus*, but in lesser numbers. In the end, the accumulation of results, the context, the comparison with the detected calls enables us to recognize the meaningful identifications.

Because the LPR dataset provided an interesting number of confusions between *D. major* and *D. minor*, we reassessed the amplitude slope parameter. *D. minor* supposedly drums with a constant amplitude and *D. major* with a decaying amplitude. The updated predictions are shown in Table 5.15. The discrimination between *D. major* and *D. minor* was improved. Twelve *D. minor* drums were reclassified as either *D. major* or *D. syriacus*. This indicates that the amplitude slope has more merit than we previously thought.

5.3 Could We Do the Same with DNNs?

We previously argued that DNNs would be impaired by the low time resolution of spectrograms if tasked with identifying drumming species. The pre-trained nets introduced in Chap. 2, Table 2.2, use either 224×224 or 299×299 pixel images. When larger images are submitted, they are downsampled. Now, 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, the time step in the envelope in Fig. 5.7 and similar is 0.7 ms. Such a precision is out of reach for spectrograms. We opted to create images using 224 pixels per 1 sec of data (a time step of 4.5 ms). Images have three channels (RGB), hence we were able to store 3 s of spectrogram at best. For further gains, we removed any lead time in the sound files and started the spectrogram at the instant when the drum started. Only in a few cases did a parasitic noise cause the wrong part of a long sound to be considered. Beyond 3 s, the exact duration of the drum does not matter anymore. The few species that can produce such long drums are already singled out.

Fig. 5.19 shows a few examples; 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). The purpose was to have the nets unlearn that the same objects can come in different sizes in images; here the size, notably of the time intervals between strikes, is fixed and a criterion for differentiation.

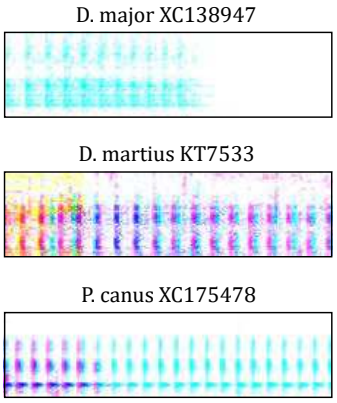


Figure 5.19: Images of Drums Submitted to DNN

Note that the capacity of the nets to measure a time interval between strikes with a 4.5 ms precision is jeopardized by the various max-pooling layers, in which the results are summarized over several adjacent pixels. Fortunately, in the very deep nets, the max pooling layers are few and far between and do not occur after the first convolution layer, where the time step is processed in its raw form.

Table 5.16: DNN Re-Training: XC/TS/KT Database Composition

Species	Training	Validation
D. leucotos	480	56
D. major	1238	169
D. martius	406	66
D. minor	1189	171
D. syriacus	278	38
P. canus	207	27
P. tridactylus	857	108
P. viridis	100	16

We re-trained DenseNet, Inception v3, ResNet 34 and ResNet 152 using drums from the XC/TS and KT datasets (5406 drums). We included both territorial and soft drums, but *J. torquilla* and *P .sharpei* were not considered as the number of available samples was too low. We saved approximately 12 % of the database for validation (Table 5.16). This time, all the drums stemming from the same recording were assigned to the same side, either training or test. This is a more demanding configuration performance-wise than what we had for k-NN. Still, particularly in the KT dataset, the same bird might be present in multiple recordings.

The results are excellent on the validation set: the accuracy is 94.8% for Inception (Table 5.17). Pooling the models, using DenseNet as solution in case of ties[22], does not improve on the DenseNet results. *P. canus* is perfectly detected. The soft drums of *P. viridis* are still identified with a 68.8% accuracy in the worst case. The confusion matrix for the case when the four models are pooled is in Table 5.18. With the exception of the *D. martius*/*P. tridactylus* pair, the confusions are not what was observed in previous experiments. The *P. canus*/*D. minor* confusion is absent, whereas a *P. tridactylus*/*D. minor* confusion, which was already somewhat present for the XC/TS dataset with k-NN (Table 5.9), becomes prominent. This might indicate that the nets did not identify the same discriminant features as we did, or found more information in the images. Was this relevant information or a case of overfitting? The fact that the nets produced 22 false *P. tridactylus* is a concern. At the very least, it indicates that the nets did not find enough information to recognize *P. tridactylus* decisively.

[22]Justified by the TN results.

Table 5.17: Drums Identification by DNN: Accuracy on Validation Set

Species	**DenseNet**	**Inception**	**ResNet34**	**ResNet152**
D. leucotos	94.6	94.6	96.4	91.1
D. major	90.5	89.9	85.2	87.6
D. martius	95.5	93.9	95.5	93.9
D. minor	92.4	98.2	95.3	90.1
D. syriacus	94.7	92.1	94.7	84.2
P. canus	100.0	100.0	100.0	100.0
P. tridactylus	99.1	99.1	98.1	98.1
P. viridis	81.3	81.3	75.0	68.8
All	93.7	94.8	92.9	90.8

Table 5.18: XC/TS/KT Dataset: DNN Confusion Matrix and Accuracy

Actual	**Predicted Classes: Mean Nb. of Samples over 200 Trials**								**Accuracy**
Classes ↓	D.leuc.	D.maj.	D.mart.	D.min.	D.syr.	P.can.	P.tri.	P.vir.	
D.leucotos	53	2	0	0	0	0	0	1	94,6%
D.major	1	153	3	2	1	0	8	1	90,5%
D.martius	0	1	63	0	0	0	1	1	95,5%
D.minor	0	1	0	158	0	0	12	0	92,4%
D.syriacus	2	0	0	0	36	0	0	0	94,7%
P.canus	0	0	0	0	0	27	0	0	100,0%
P.tridactylus	0	0	0	1	0	0	107	0	99,1%
P.viridis	0	1	0	1	0	0	1	13	81,3%

As usual, the predictions for the field datasets offer a more complex picture (Table 5.20). Here, we see that the confusion between *P. canus* and *D. minor* was not overcome. The pooled nets still predicted *P. canus* more often than *D. minor* because DenseNet was selected as the net to decide ties. DenseNet is the best *P. canus* predictor (Table 5.19). This is the net that best understood the critical differences between the two species. Only Inception predicted more *D. minor* than *P. canus*. Otherwise, the nets predicted the classes that they saw the most during training, e.g. *D. major* was predicted rather than *D. syriacus*. The numbers for *D. minor* are also systematically higher than what k-NN predicted, because a neural net that saw a lot of *D. minor* during training thinks that *D. minor* is highly probable. Even with an overwhelming number of *D. minor* in the training set, k-NN will still assign a class based on the few samples that are close to the test sample. If we consider the wrong *P. viridis* identifications in LPR (18 for k-NN and 16 for the nets), 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 nets 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 accuracy for *D. martius* and *P. canus* is very poor (Table 5.21).

Overall, the nets do not appear to perform better than k-NN. The accuracy numbers in Table 5.20 simply reflect the fact that the nets differentiate *P. canus* less well than k-NN, and on the other hand produce fewer false *D. syriacus*. The confusion matrix in Table 5.21 shows that random (wrong) identifications are more common for the DNN than for k-NN. It is hard to know whether the predictions would improve with more data (although 5000 drums is already a significant dataset) or whether the DNN hit a wall in terms of the amount of info that they can extract from the images. As mentioned, imbalance between the classes in the training set causes the nets to wrongly learn that some classes are more probable than others. Odd samples are also washed away by the repetition of more typical examples. k-NN preserves the ability to match test samples to oddities. However the nets retain a strong advantage in terms of the simplicity of the process; the image generation is neither subtle nor long; training the nets requires two hours at most and running the test samples a few minutes.

Table 5.19: Drums Identification by DNN: Details for TN Dataset

Species	**DenseNet**	**Inception**	**ResNet34**	**ResNet152**
D. leucotos	10	5	31	20
D. major	52	25	50	48
D. martius	135	117	162	140
D. minor	153	1434	589	446
D. syriacus	0	1	2	2
P. canus	2197	962	1717	1888
P. tridactylus	14	25	19	25
P. viridis	21	13	12	13

Table 5.20: Drums Identification by DNN: In Field Datasets

Species	**TN**			**RM**			**LPR**		
	k-NN	DNN	Truth	k-NN	DNN	Truth	k-NN	DNN	Truth
D. leucotos	0	7		0	0		5	19	
D. major	13	44	14	204	233	257	808	945	1001
D. martius	118	139	120	2	9		88	50	108
D. minor	152	329		6	18		21	80	
D. syriacus	0	0		48	5		158	0	
J. torquilla	1			0		6	0		
P. canus	2278	2037	2447	3	2	3	16	5	10
P. sharpei	2			0			0		
P. tridactylus	13	15		14	11		5	4	
P. viridis	5	11		1	0		18	16	
Accuracy[a]	92.9%	83.0%		74.5%	83.5%		81.0%	87.0%	

[a]For the reference labels, the large *P. canus* and *D. major* sets were assumed to be correct. Other drums were listened to and confronted with the presence of calls in the data.

Table 5.21: LPR: DNN Confusion Matrix and Accuracy

Actual	**Predicted Classes**								**Accuracy**
Classes ↓	D.leu.	D.maj.	D.mart.	D.min.	D.syr.	P.can.	P.tri.	P.vir.	
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%

5.4 Conclusions

As it appears, the identification of woodpecker species from their drums is not always clear cut. 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*. This is further impaired by the realities of field recordings: distant drums, poorly executed signals, significant variations in the productions of a single individual[23]. Improvements of the parameter set could only address these issues to a marginal extent.

We note on this subject that the parameters that proved the most reliable were the simplest temporal parameters: the delta interval, the initial interval, the DR duration and the number of strikes. Other parameters either obscured the analysis or proved ineffective. Zabka's assessment that the time interval between rolls had no value was confirmed.

It is particularly frustrating that we were not able to fully exploit the eccentricities of *D. minor* (quiet drums, quickly issued in succession, rich in harmonics) to efficiently separate this species from *P. canus* and *D. major*. The sound intensity of drums cannot be measured, the time between rolls disappointed and considering the lack of diversity in our training sets, the spectral parameters degraded rather than helped the results.

We had discarded the amplitude slope early. Fisher's discriminating power had placed it in second line, the LDA had deemed it without value. Then Garcia et al. [28] presented it as a critical parameter for *D. major*. In the end we found it some merit in helping to differentiate *D. minor* from *D. major*. It would not help for the differentiation of *P. canus*.

The most confident predictions, either with k-NN or with a deep net, were the ones for which there was a volume of observations. The presence of *P. canus* in Tenneville, and *D. major* in Remerschen and La Petite Raon was immediately confirmed by the analysis. Thankfully, birds naturally repeat their signals in order to increase their reach. For the birds that occupy the territory, great numbers are easily reached. This fact works to our advantage because it means that detection and identification algorithms do not need to be perfect. As long as there is a sufficient number of drums that are detected and correctly identified, the information has substance. On the contrary, the absence of repetition is suspicious.

[23] Are drums modified to suit the circumstances? It seemed that the *D. martius* that attacked the *P. canus* territory in Tenneville had exceptionally long drums, maybe as a show of force?

The less voluminous results must be confronted with contextual data. Background information about the habitat and the range of species is critical. For example, it is impossible to separate *D. major* and *D. syriacus* otherwise[24]. The other decisive point is the co-occurrence of calls. We front-loaded results from the next chapter on calls to decipher the drums identification. Without this secondary information, the scope of drums analysis is somewhat limited.

With either recognition method, the results are vulnerable to the quality of the training set. Large classes and diversity within the classes are essential. For the deep nets, balance between the classes is another requirement. The training sets we used had multiple shortcomings, starting with some classes being insufficiently populated. That the XC/TS set was unable to detect *P. canus* was a fatal hurdle. Eventually, we chose to include the soft drums in the training sets. They create their lot of confusion, but to exclude them is to shut the door to the detection of rarer signals. If the microphone is close enough to a nest, soft drums are very likely to be recorded. Our Remerschen data was found to contain *J. torquilla* soft drums, although none were correctly detected.

The procedures we developed in the present chapter are most efficient for drums that are loud and clear. Otherwise the calculation of temporal parameters becomes challenging. This is an additional restriction on the reach of recording stations. Parasitic signals superposed to the drums are another source of difficulty, but one that can more easily be worked around. It is also not that common in the data because the frequency range of drumming is depleted in European forests. In any case, failure to process sporadic drumming rolls is not a road block because birds produce them in large quantities. There will be enough clean drumming rolls to make an identification.

In an unusual turn of events, the deep nets did not perform better than the combination of handcrafted acoustic features and k-NN. We explained the difficulty in generating images with a sufficient time resolution. In addition, the nets have to capture indirect information from the images; it is not the patterns of drums that are of interest (straight vertical lines), it is the spacing of strikes, the number of strikes, the length of drums, indirectly inferred from the use of the different colors. It is quite possible that the nets need a much bigger dataset to develop this sophisticated knowledge. We managed 5000 samples, which is not negligible. Expanding all categories

[24]Winkler & Short [90] thought that they were better separated by their call note.

equally and with a diversity of sources is a significant challenge.

We finish with a few words on whether the drums truly carry the species information; this is a necessary assumption in the present work but one that has been doubted (Dodenhoff et al. [15]; Garcia et al. [28]). As we observed, they do not in an absolute sense because at best, only the sympatric species can be discriminated. Can the birds themselves tell the difference based solely on the drums? Garcia et al. [28] showed that woodpecker species that are close genetically often have similar drums. In his replay experiments, *D. major* responded to drums of *D. syriacus* and *Dendrocopos hyperythrus*, both genetically close[25] as if they were the drums of a conspecific. On the contrary, it responded to neither *D. minor* nor *P. canus*. The causality might however not be straightforward. The genetically close species might be correctly identified but perceived as competition. And there are multiple occurrences of genetically distant species answering each other in our own field recordings, e.g. *D. martius* and *P. canus* in Tenneville, *D. major* and *P. canus* in Tenneville and La Petite Raon, or *D. major* and *D. martius* in La Petite Raon. There is a social context (good neighboring, aggression); the reasons why birds answer each other are not fully determined by their capacity to identify the other species. Perhaps closing the debate, Winkler & Short [90] claim that drumming is a territorial signal and not a mate attraction one[26]. This would be supported by the recording dates in Fig. 3.7 (Chap. 3): drumming comes in too late to be meant to attract a partner. A consequence is that drumming does not need to be the best signal to discriminate species, because the purpose is not reproduction but keeping intruders at bay.

If it was critical to encode the species information in drums, then we would see a reinforcement of character in the hybrid zone[27], similar to what Kirschel [40] describes for tinkerbirds. For example, where *D. major* and *D. syriacus* coexist, their drums would become more distinct, with perhaps *D. major* shortening the rolls and *D. syriacus* lengthening them. However in their overlapping range in Eastern Europe, *D. major* tends to restrict itself to mountain forests, leaving the lower altitudes to *D. syriacus* (Michalczuk et al. [53]); the sympatry is limited. In addition, hybrids between *D. major* and *D. syriacus* are not rare (Michalczuk et al. [53]), which indicates that rein-

[25] *D. major* covers most of Eurasia; it is sympatric with *D. syriacus* in Eastern Europe. *D. hyperythrus* resides in South-East Asia and meets *D. major* in China.

[26] The contrary is more often expressed in the literature.

[27] Thought offered by M. Garcia in personal communication; untested to date.

forcement of character might not be a priority. The only certainty is that the technology we developed could help with such investigations.

We close this chapter with a summary illustration of the drums detected in the three field datasets: Figs. 5.20, 5.21 and 5.22. The Tenneville *P. canus* appears as an outstanding drummer, perhaps due to its peculiar situation of living on the edge (of the distribution zone). The Vosges mountains are the definite place to visit in April to listen to a drumming concert.

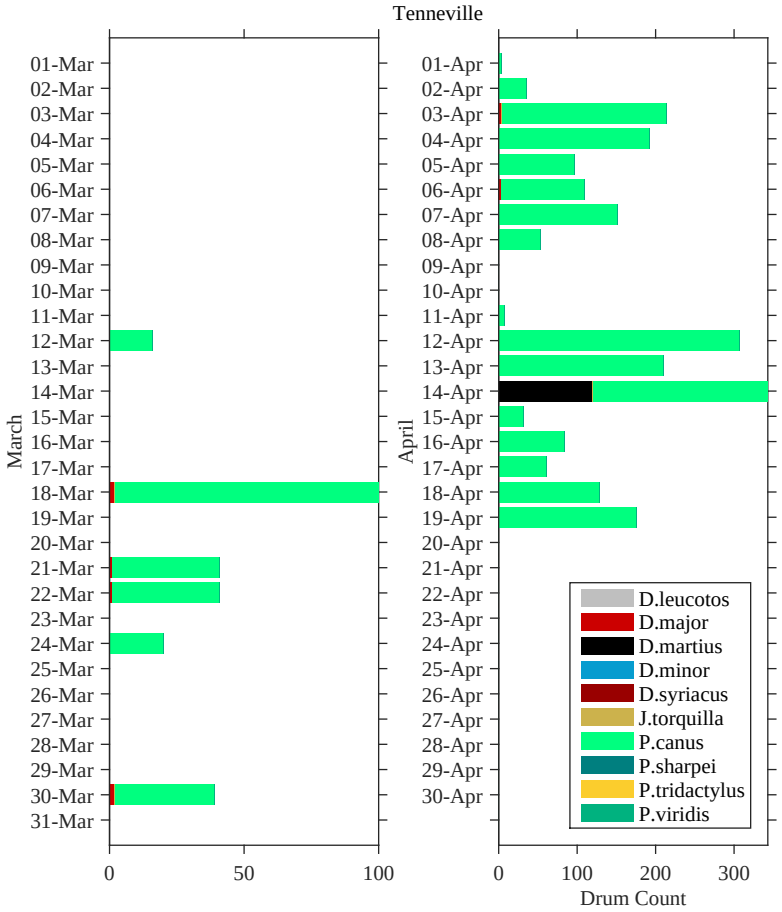


Figure 5.20: Tenneville: Drums Identifications by Date

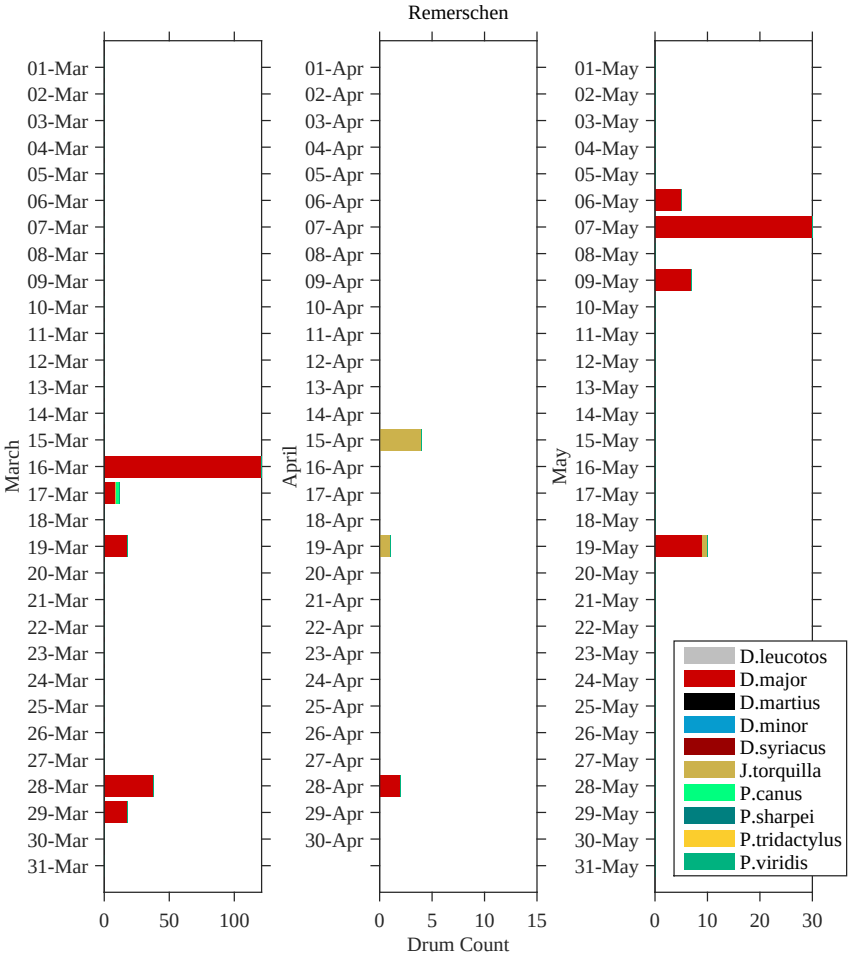


Figure 5.21: Remerschen: Drums Identifications by Date

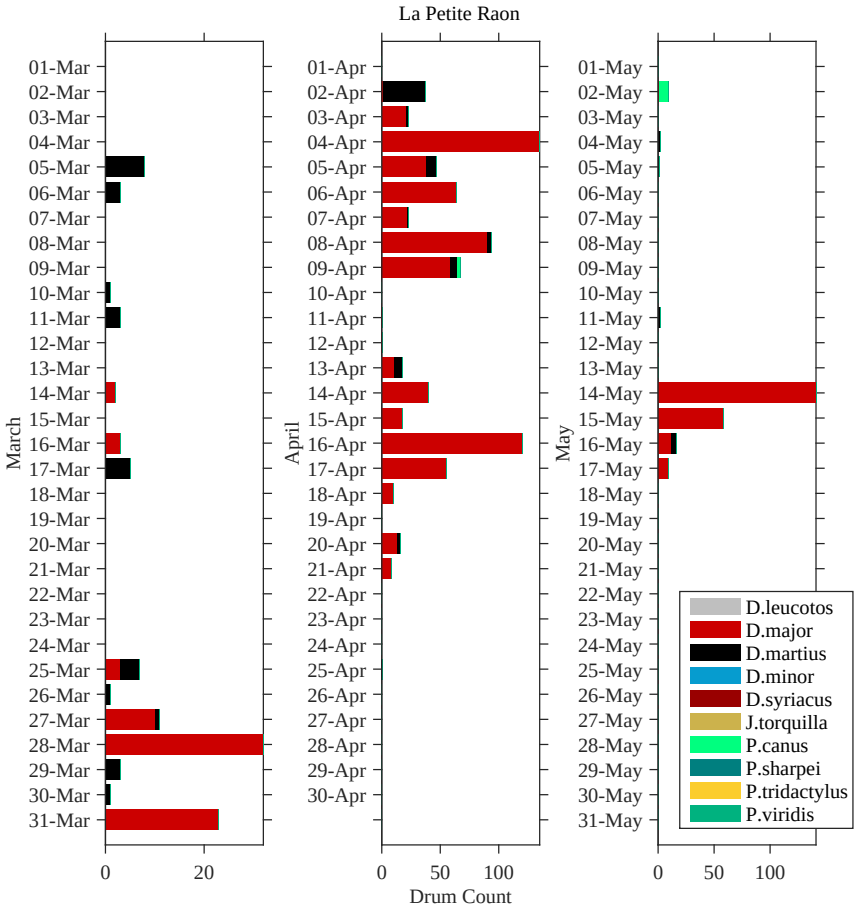


Figure 5.22: La Petite Raon: Drums Identifications by Date

Chapter 6

The Identification of European Woodpeckers from their Calls

Our last task consist in detecting and identifying woodpecker calls in recordings. We saw with drumming that the two problems have a different scope. In the detection step, the woodpecker calls must be set apart from every other sound, biotic or abiotic, that might possibly occupy the same bandwidth. In the identification step, the woodpecker calls must be told apart from each other.

The woodpecker calls were introduced in Chap. 3, Section 3.3. We opted to focus our study on the nine calls represented in Fig. 3.6. Among these, the rattle calls have many similarities in terms of structure and frequency range. There is a certain number of syllables, typical interval durations during the syllables, a characteristic evolution of the syllable main frequency. But what is most distinctive is the sound produced by each syllable, i.e. the kleees and kru-kru-krus that produce different shapes in the spectrograms. Parameterization of the different types of syllables is an arduous endeavor, and one that seems rather vain when convolutional neural nets have developed such image analysis capabilities. We also saw in Chap. 5 that neural nets have some difficulties with time structures but are not completely incompetent in that matter either. In any case, convolutional Deep Neural Nets (DNN) are the most promising way forward to analyze woodpecker calls.

In the present chapter we explore two paths: first a simple 4-layer network[1] that we designed and trained from scratch, then the legacy very deep image nets that we retrained for woodpecker calls. In the chronology of our research, the work on the simple network predates all uses of the very deep image nets, including the analyses in Chap. 4 and 5. The results are modest

[1] Counting only the layers that have weights.

but help outlining the challenges specific to our data, namely their paucity and their lack of variety. As in Chap. 5, we seek to control the time and frequency scales in the images.

Our ultimate goal is to analyze the field recordings of Tenneville, Remerschen and La Petite Raon in full. For the development of the neural nets, we use the mixed collection of XC and KT recordings detailed in Table 3.5 (Chap. 3). It contains 1836 calls from *D. martius* (3 different calls), *D. medius*, *D. minor*, *J. torquilla* and the three *Picus*. Two known shortcomings are that the number of samples for *P. sharpei* is too low (35) and that the number of *J. torquilla* calls is artificially high (628) due to one single recording yielding 276 calls. For all other classes, the mean number of samples is 168 (minimum 89, maximum 263).

The first section in the present chapter discusses the formatting of the images of woodpecker calls that are fed to the neural nets. Then we proceed with the simple net, and finally with the very deep image nets.

6.1 Constraints on Images and Methodology

Following the indications in Table 4.2, the XC/KT recordings containing calls were segmented by searching for emerging signal in the 500–3000 Hz bandwidth. The segments were then manually reviewed; the 1836 ones containing calls were saved to individual files.

Compared to other datasets (Chap. 4, Chap. 5, Table 6.2), this one is small and likely insufficient to train a deep neural net. Producing several partial images of the calls (*crops*), focused on a few syllables, is an option to augment the dataset. A second benefit is that the smaller images will require less analytical power for their examination, and thus a shallower, less data-hungry and easier to train net. In essence, we substitute the recognition of the syllables for the recognition of the calls. This has been done by other authors as well. Potamitis [65] classified syllables or elements of songs extracted from spectrograms. Brandes [8] 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. In our case, there is a chance that the syllable identification will be enough.

We settled for 54×63-pixel images (1000–3500 Hz × 1 sec, using 21 ms frames with 25% overlap). The calls were segmented with a 15% overlap be-

Table 6.1: Database of Spectrogram Images: Woodpecker Calls

Species	**Call**	**Images**
Noise		6081
D. martius	Ad	543
	Flight	307
	Contact	207
D. medius		625
D. minor		451
J. torquilla		2429
P. canus		595
P. sharpei		117
P. viridis		799
Total		**712154**

tween consecutive images[2]. 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. Fig. 6.1 shows the images extracted from a *D. minor* and from a *J. torquilla* call. In the *D. minor* example, the bird followed its call by drumming; the last image does not contain its voice at all. We foresaw an additional "noise" class to label such images. Contributors to this class include passerine calls, other woodpecker calls not in our study, anthropogenic sounds or various instances of background noise. For the full database, we obtained 12154 images (Table 6.1), half of which were noise. We see in the *J. torquilla* example in Fig. 6.1 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.

The frequency bandwidth for the images was selected after reviewing the mean spectra of the various calls in Fig. 6.2. For all species, the fundamental frequency is most often found in the 1500–2500 Hz range. As for drumming, the limiting taxon is *D. minor*, which uses higher frequencies.

In the image generation, amplitudes less than 20 dB below the maximum amplitude in the image were discarded. For the XC/KT recordings, a version where the spectrogram amplitudes were kept linear was also used[3]. The calls in these recordings are usually taken at close range; they are clearly

[2]Lasseck [47] also tested extracting random audio chunks with some success.

[3]In the small net simulations only.

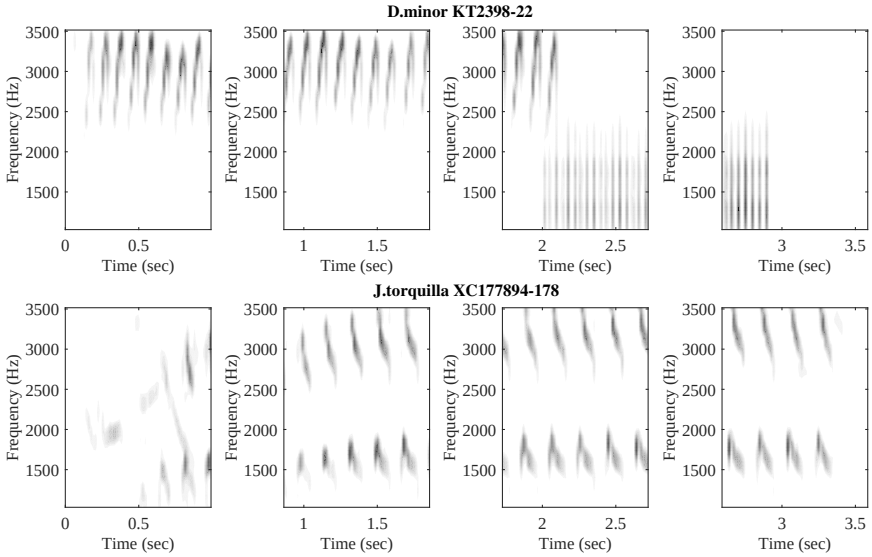


Figure 6.1: Images Extracted from a *D. minor* and from a *J. torquilla* Call

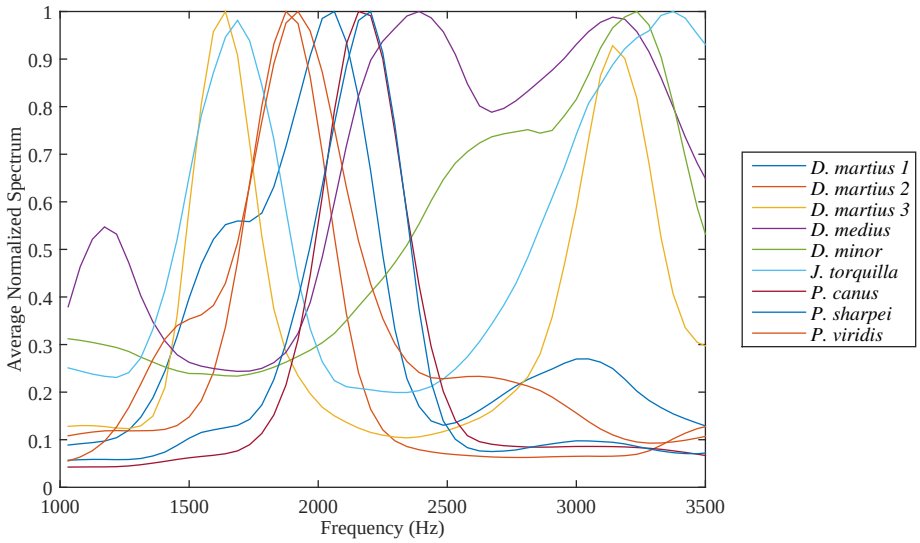


Figure 6.2: Spectral Profiles from Woodpecker Calls,
Averaged over the Database

Table 6.2: Image Sizes Used by Various Authors

Source	**Nb. of Images** (/)	**Image Size** (pixels)	**Width** (sec)	**Time Step** (ms)	**Height** (Hz)	**Freq. Step**
Salamon [68] (simple conv. net)	8732	128×128	3[b]	23	0–22050	128 mel bands[c]
Grill [33] (simple conv. net)	16000	80×1000	10	10	50–11000	80 mel bands[c]
Lasseck [47] (very deep nets)	36496	299×299[a]	5[b]	17	160–10300	256 mel bands[c,d]
This text (both)	12154	54×63	1	16	1000–3500	46 Hz

[a]Input size for Inception v3.

[b]Every time an audio file is considered during training, a 3 or 5 sec-long crop is taken at a random position in the spectrogram. The frequency range is always considered in its entirety.

[c]In the mel-scale, the frequency bins are not of equal width.

[d]The width is larger than the height. The images are white-padded in the vertical direction.

visible in linear scale. This is of course atypical; in field recordings, the decibel scale helps with bringing forward secondary calls. For the field data, we also increased the cut-off to 30 dB. We note that Grill & Schlüter [33] observed that background noise removal was essential, but that other authors did not consider it. Grill subtracted the mean per frequency band from the spectrograms, the intent being of suppressing colored noise[4]. In our work the -20 dB or -30 dB cut-off removed the lowest noises from the images and allowed using the available color range to describe the main sounds in the segment.

Table 6.2 shows how our solution compares with other works, two using simple convolutional nets (Salamon & Bello [68]; Grill & Schlüter [33]) and one using a very deep net (Lasseck [47]). Our images are small by design

[4]But does this approach compromise the images? Adjacent frequency bands are adjusted differently.

but still ensure a good time and frequency resolution, comparable to the numbers in Lasseck [47]. Computing mel-spectrograms was not considered[5].

LeCun et al. [49] issued the following guidelines for the normalization of features entering a neural net. The features should have zero mean and unit standard deviation over the training set and be decorrelated from each other. The latter is traditionally achieved through PCA, but cannot apply to images; neighboring pixels in an image often maintain a strong relation by design. LeCun's rules were written in 2012, before convolutional networks became the norm (AlexNet was published in 2015). The other rules, on mean and standard deviation, were meant to ensure that there were negative and positive multipliers to the weights[6], allowing them to increase and decrease, and that all weights evolved independently from each other and brought a unique contribution to the network.

For the small net, the spectrograms were passed on using the HDF format. We used two normalization schemes for our images. In option (1), the spectrogram amplitudes (in dB) were normalized to the [0,1] interval. This option considered that there was no physical basis to divide the pixel amplitudes into positive and negative values. Since the noise had been cut off, the data in the images was almost binary: either there was content, or not. This normalization was physically sound but did not abide by LeCun's rules. In option (2), the null values in the image were replaced by negative values, all equal to $-c$, where c was a positive constant. The value of c was chosen so as to approach zero mean and unit standard deviation for the set of all pixels from all images in the database.

The legacy deep image nets accept images in the JPEG format. JPEG images have three channels (RGB) and pixel amplitudes in the interval [0,1][7].

[5]Mel-spectrograms have been successful in human speech recognition and have since been used by default by many authors. The mel-frequencies are obtained through a logarithmic transformation of the linear frequencies. They aim to capture the response of the human auditory system better. To compute a mel-spectrogram, a filter-bank is determined with central frequencies spread linearly along the mel-frequency range. Each mel-band integrates one or several linear frequency bands. In Grill & Schlüter [33] the compression factor from the linear spectrogram to the mel-spectrogram was 6; in Salamon & Bello [68] 4, in Lasseck [47] 2. Lasseck obtained better results with mel-spectrograms than spectrograms with a linear frequency scale.

[6]See Equation 2.7 in Chap. 2. The partial derivative $\frac{\partial C}{\partial w_{ij}}$, which is a factor in the prescribed change to weight w_{ij}, depends on the neuron activity x_i. If x_i is zero, the weight never changes.

[7]Actually, [0,255]. The conversions from [0,1] to [0,255] and vice-versa are implied.

The triple one corresponds to white and the triple zero to black. In our implementation, we took the images from option (1) above, triplicated them to fill the three RGB channels and flipped the pixel values to have the calls appear in black over a white background, as is usual. The Pytorch scripts that command the nets then read in the images and converted the pixel values to the [-1,1] interval using:

$$x_{ij} = \frac{x_{ij} - \bar{x}}{\sigma} \tag{6.1}$$

Here $\bar{x}$ is supposedly the mean pixel value and σ the standard deviation, but both were simply taken to be 0.5. After this, the pixels capturing the calls were the ones with high negative values. Consequently, the weights in the first convolutional layers had to be negative as well to flip the problem back to a natural order in which positive content corresponds to positive values.

Grill & Schlüter [33] used batch normalization, i.e. they normalized not the whole training set at once, but every mini-batch of data entering the input layer in a new training sequence. The images were normalized per frequency bin, i.e. all the pixels from the image batch that belonged to the same frequency bin were corrected to collectively achieve zero mean and unit variance[8]. In the deeper networks, the pixels in the input layer had values in the range [-1,1] as in option (3) and batch normalization was used in intermediate layers for avoid vanishing gradients (Szegedy et al. [79]; He et al. [34]; Huang et al. [36]), which occur when neuronal activities become too small. Batch normalization uses Eq. 6.1: the neuron activities are collectively scaled back up to achieve zero mean and unit variance.

Normalizing pixels of the input layer using Eq. 6.1 has become common, but it does not strongly connect with the physics of sound. In our case, out of the usable 30 dB, 15 dB are deemed "positive" content and 15 dB "negative" content. This considers that the two halves describe opposed phenomena. However, calls might appear on either side, depending on their intensity. Some calls might also be split between the two sides, with perhaps some important components ending up near zero. How will the nets handle these situations? The [-1,1] normalization was conceived for color ranges.

As in previous analyses, we assembled a test set by setting aside either 10% of each class or 50 samples at most. If one image extracted from a

[8]Here as well, the continuity between vertically adjacent pixels might be lost. The patterns that span multiple frequency bands are not scaled at once.

given call was in the test set, then all images extracted from this call were. However the calls from a single individual could be split between the test and training sets. Admittedly, 276 *J. torquilla* calls were extracted from the same XC recording (987 images).

Besides size, the other issue with our training set is its unevenness. The discrepancy in size between for example *J. torquilla* and *P. sharpei* may cause the network to learn that the first one is highly frequent and the second one rare. Our wish is for a network that can detect both without bias. For this reason we also prepared a modified dataset in which the classes are more balanced in size (Table 6.3). The number of *J. torquilla* and *P. viridis* samples was limited to 600, while the *D. martius* contact call samples were copied twice and the *P. sharpei* samples four times to artificially increase their numbers. This is not true diversity, but it makes the frequency of occurrence of these classes in the training process fair compared to other calls.

Table 6.3: Full and Balanced Datasets: Composition

	Full Dataset		**Balanced Dataset**	
	train	test	train	test
D. martius (1)	495	48	494	49
D. martius (2)	269	38	280	27
D. martius (3)	192	15	380	17
D. medius	577	48	576	49
D. minor	405	46	415	36
J. torquilla	2376	53	600	52
P. canus	556	39	554	41
P. sharpei	103	14	416	13
P. viridis	756	43	600	45

6.2 A Simple Convolutional Neural Network

6.2.1 Initial Model Setup

Starting from Grill's architecture, we designed a simple network that aimed to differentiate the nine woodpecker calls from each other. This task is potentially simpler than the one that detects woodpecker calls among a variety of noises, because most woodpecker calls, notably the rattles, are variations on a similar structure.

The architecture of this simple convolutional deep net is depicted in Fig. 6.3. The details are in Table 6.4. Compared to the nets in Salamon & Bello [68] (3 convolutional layers with either 24 or 48 feature maps, 2 dense layers) and Grill & Schlüter [33] (4 convolutional layers with 16 feature maps, 3 dense layers), this one is smaller. Given the difference in training set sizes (Table 6.2), we could afford about a third of Grill's network. We implemented two convolutional layers, each followed by a max-pooling layer; then two dense layers and the final softmax layer in which the results of the previous layer are converted into class probabilities.

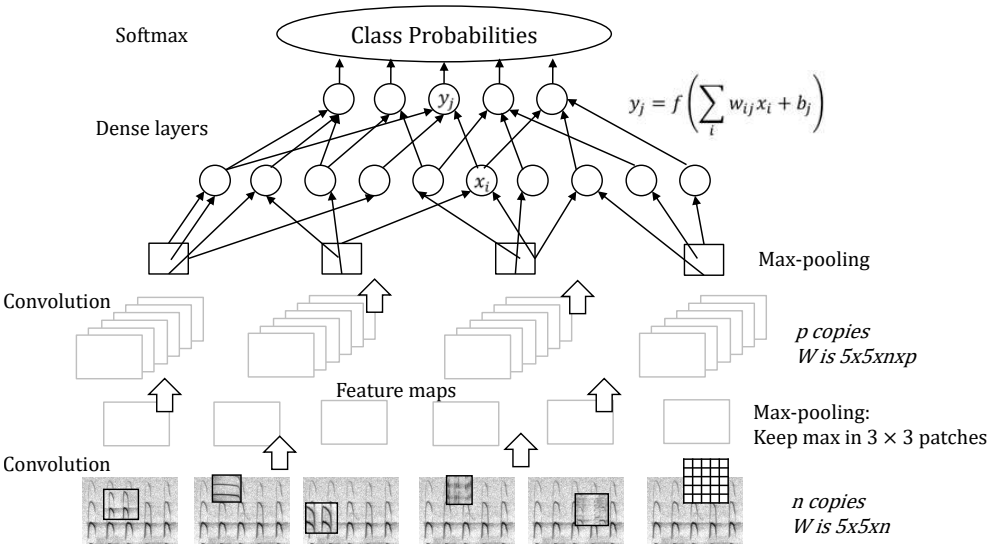


Figure 6.3: Deep Neural Network for the Identification of Woodpecker Calls

In the convolutional layers, fourteen 5x5 filters were used. This filter size amounts to 250 Hz by 80 ms in the first layer, on par with syllable dimensions. In all layers, the non-linear function used behind the weights multiplications is the leaky-rectify function. Given that there are 9 classes to identify, it seemed that the minimal necessary number of filters was 9; using 14 gave us a margin. A typical number of filters in the lower layers of very deep nets (1000 classes) is 64; it then increases in deeper layers to compensate for the decrease in the size of the feature maps (Szegedy et al. [79]; He et al. [34]).

The total number of trainable weights and biases in the network is 42961. Table 6.4 puts in evidence the fact that dense layers are heavier than convolutional layers. Whereas convolutional layers involve very selective multiplications, dense layers involve every available bit of information, hence the prohibitive number of weights that results from their use.

6.2.2 Model Training

The network was implemented from Grill's scripts[9], using the Theano/Lasagne libraries (Dieleman et al. [14]), and essentially retained the training scheme of Grill's network. Training consisted in the minimization of the categorical cross-entropy cost function through stochastic gradient descent. The initial weights were sampled from the uniform distribution (Glorot initialization). Then the training data was processed in mini-batches of 64 samples. The cost function sums the errors for the full mini-batch; this is the training loss. From the training loss, the necessary modifications to the weights of the network are determined. At the end of an epoch, the training loss is evaluated on the full training set, along with a validation loss computed on the validation set, i.e. a random 20% of the training data that was set aside. An epoch is complete when a given number of mini-batches have been processed; typically, when all the samples in the training set have been reviewed. The purpose of the validation loss is to control that the optimization truly progresses, and this requires a different set of samples than the ones used for training. If the validation loss does not decrease, then the net is overfitting the training samples. It will eventually be able to recognize test samples very similar to the training samples, but will lack the capacity to generalize.

A relatively large learning rate of 0.005 was used at the beginning of training. In this configuration, each iteration brings large changes to the

[9] https://jobim.ofai.at/gitlab/gr/bird_audio_detection_challenge_2017.

Table 6.4: Simple Convolutional Net: Layers Description

Layer	Parameters	Output	Weights & Biases
Inputs		54×63 image	0
Convolution	5×5 filters 14 filters	14 feature maps 54×63 f.maps[a]	364[b]
Max-pooling	3×3 filters	14 18×21 f.maps	0
Convolution	5×5×14 filters[c] 14 filters	14 feature maps 18×21 f.maps	4914[d]
Max-pooling	3×3 filters	14 6×7 f.maps i.e. 588 neurons	0
Dense	Fully connected	63 neurons	37107[e]
Dense	Fully connected	9 neurons	576
Softmax		9 class probabilities[f]	0
		Total	**42961**

[a] The input images are zero-padded so that the 5×5 filters can be superimposed at every location, including on the image edges.

[b] Per filter: 5×5 weights and 1 bias.

[c] The filters act on all the input feature maps at once and sum up the results.

[d] Per filter: 5×5×14 weights and 1 bias.

[e] 588×63 weights and 63 biases.

[f] Whose sum is one.

weights. The learning rate was then re-evaluated at the end of epochs. If the training loss on the full training set had not decreased, then the learning rate was divided by 10. Below a given learning rate threshold (0.00005) or after 300 epochs, the training was stopped.

Dropout was used in the upper layers. Half of the feature maps were randomly ignored at the entrance of the first dense layer, then a fourth of the pixels of the remaining feature maps. At the entrance of the second dense layer, half the inputs were dropped. This setup forces every weight in the dense layers to contribute beneficial information or vanish; they cannot rely on other weights to compensate their own shortcomings, as said other weights might be missing on a random basis. Dropout does not apply to the validation or the test set; in predictions, every connection in the network must inform the final decision.

The net was trained five times with a different random selection for the training set and the validation set, always with a 80%–20% split. Two aspects, on the one hand the varying composition of these folds of the training set and on the other hand the presentation of training samples to the network in a random order, ended up producing different models with different prediction capabilities. They were eventually pooled together to increase the probability of a correct answer. The final performance of the network was evaluated on the test set. The class probabilities from the different models were averaged[10] and the top prediction, or the top three predictions, were retrieved.

6.2.3 Results and Variants

Baseline & Training Parameters

Table 6.5 shows the top-3 accuracy for the baseline simulation and for a few variations on the training parameters. For the baseline, the accuracy when only the top prediction is considered is 24% (average of the accuracies for individual classes). This means that the network is only twice as good as a random prediction (1/9). The network is good at predicting *D. minor*

[10] In some simulations we selected the maximum probability achieved for each class over the different models, and then the three classes which achieved the best probabilities. This was meant to put forward the models that had produced confident predictions. However averaging the class probabilities over the different models consistently returned a higher accuracy and is more common in the literature. Some authors also square the probabilities before adding them up (Lasseck [47] for example).

Table 6.5: Baseline Simulation and Impact of Training Parameters

			Accuracy(%)[a]									
Epoch Size[b]	**Nb. Epochs**[c]	**LR D.**[d]	mart.1	mart.2	mart.3	med.	min.	torq.	can.	sh.	vir.	Avg.
Baseline[e]												
1	16–33	0.1	72.9	78.9	60.0	45.8	93.5	100.0	71.8	0.0	86.0	67.7
2	25–31	0.1	60.4	13.2	93.3	4.2	91.3	100.0	38.5	0.0	90.7	54.6
20	33–50	0.1	2.1	0.0	100.0	0.0	39.1	100.0	94.9	0.0	74.4	45.6
1	251–264	0.9	8.3	63.2	20.0	100.0	0.0	0.0	94.9	0.0	0.0	31.8

[a]Percentage of samples for which the correct answer is in the top three predictions. Top-1 prediction accuracy in the range 20%–24%. Class probabilities averaged over the five models.

[b]Expressed as the number of times the training set is visited in an epoch.

[c]After which training stopped.

[d]Learning rate decay: multiplying factor for the learning rate when the training loss fails to decrease at the end of an epoch.

[e]Full dataset, pixel normalization [0,1]. Fourteen (14) feature maps in the convolutional layers.

(distinctively high frequency), *D. martius* (3)[11] (peculiar syllable shape), and *J. torquilla* and *P. viridis* (abundance of data). It is average at predicting the other two *D. martius* calls and *P. canus*, and rather bad for *D. medius* (distinctive but highly variable syllable shape) and *P. sharpei* (too little data). The correct answers for the simulations in Table 6.5 are produced with probabilities in the range 0.21–0.27 in average: the confidence in the answers is not overwhelming. The probability climbs to about 0.4 for *J. torquilla*, the best predicted class.

Training was extended by either increasing the size of the epochs or slowing the decay of the learning rate. Both approaches led to a fall in accuracy (Table 6.5). Over all simulations on the small net, the number of epochs after which training stopped lied in the range 20–35 in most cases. This is short but dictated by the size of the training set. Longer training can only lead to overfitting[12]. We observed throughout simulations that the effective number of training epochs did not have an impact on the validation loss. The net-

[11]Notation in column headers: mart.1, mart.2, mart.3. If space allows: *D. martius* (1), *D. martius* (2), *D. martius* (3). Respectively: the advertising call (rattle), the flight call (kru-kru-kru) and the contact call (kleee).

[12]Keeping the training short is a known regularization technique.

works that underwent a longer training did not achieve a better accuracy. They were simply slower learners.

Dropout

Grill had omitted to disable dropout for the validation and test sets in his code, which produced an interesting experiment. Table 6.6 shows the top-3 accuracy calculated for the baseline model; the calculation is repeated several times with and without dropout enabled. Thus in runs (a) through (e), more than half of the net was discarded to make the predictions. Yet the accuracy exceeds the baseline accuracy by up to 10%. In a single class (*D. martius* (1)), the accuracy goes up by almost 40% in run (d). If better predictions are obtained with a decimated net, then the net must be too big[13]. These results motivated a reduction of the number of feature maps from 14 to 8.

Table 6.6: Dropout in Evaluation of Test Samples

Training	**Testing**	**Accuracy(%)** Average over Species	mart.1
Dropout	No Dropout	56.84	45.83
Dropout	Dropout(a)	64.05	81.25
Dropout	Dropout(b)	66.63	70.83
Dropout	Dropout(c)	61.86	72.92
Dropout	Dropout(d)	56.88	83.33
Dropout	Dropout(e)	48.45	81.25
Less Dropout	No Dropout	46.92	43.75

Percentage of samples for which the correct answer is in the top three predictions. Top-1 prediction accuracy in the range 18%–34%. Maximum class probabilities over 5 models. Full dataset, pixel normalization [0,1]. Dropout(a) through (e): all predictions were based on the same trained network as in the first row, but sections of it were randomly dropped. Less Dropout: no dropout was used in the evaluation of the test samples; during training, 20% of the feature maps were dropped after the second convolution, versus 50% in the other simulations. Dropout in the dense layers was unchanged.

[13] Alternate explanation: once dropout is removed for the validation, the weights in the net should also be scaled down to account for the increased number of connections. Otherwise the calculations are off. These could explain the poor performance in the first table row.

In the last simulation in Table 6.6, the amount of dropout in the model was reduced. The accuracy dropped, which indicates that the dropout implemented in the baseline model is not excessive.

Normalization and Balanced Dataset

Table 6.7 shows the benefits of a balanced training set (where the cardinality of the classes is homogeneous) and of input features that overall satisfy LeCun's condition of zero mean and unit variance. Without this, the null values in the feature maps perpetually prescribe a null change to the corresponding weights, rather than a reduction. The average accuracy over the classes increases with the balanced dataset (+2.3%) and with the normalization (+6.2%). *J. torquilla* is strongly affected by the decrease in samples, whereas *P. viridis* is not. This might mean that the *J. torquilla* call was previously well identified solely based on the perception by the network that it was highly probable. Both *D. martius* (3) and *P. sharpei* benefit from the artificial increase of their numbers. *D. minor* is negatively affected by the normalization.

Note that to accompany the balanced training set, a revised test set was produced (Table 6.3). This implies that the networks in Table 6.7 were not all evaluated on the same data. We still observe, in Table 6.7 as well as in Table 6.5, that classes are either correctly predicted or moderately well predicted or poorly predicted, and that the last two categories encompass a wide accuracy range. In other words, the results are volatile.

Table 6.7: Balanced Dataset and Normalization

	Accuracy(%)									
Dataset	mart.1	mart.2	mart.3	med.	min.	torq.	can.	sh.	vir.	Average
Original	81.3	94.7	60.0	93.8	100.0	100.0	43.6	0.0	88.4	73.5
Balanced	93.9	88.9	82.4	100.0	100.0	63.5	22.0	38.5	93.3	75.8
Bal.& Norm.	91.8	92.6	100.0	98.0	88.9	86.5	63.4	23.1	93.3	82.0

Percentage of samples for which the correct answer is in the top three predictions. Top-1 prediction accuracy in the range 26%–38%. Class probabilities averaged over 10 models (10 random splits of the training set into 80% training samples and 20% validation samples). Eight (8) feature maps in the convolutional layers. The augmented set is 13 times the size of the balanced set.

Number of Models

Table 6.8 documents simulations in which the number of training set/validation set splits, i.e. the number of pooled models, was increased. This has a direct positive effect on the accuracy, which reaches 84.8% with 100 models. Within a model pool, some are excellent and other awful, which prompted us to wonder whether there was a way of discarding the ones that deteriorate the accuracy. The scatter plot in Fig. 6.4 shows the accuracy reached by single models versus the final validation loss achieved when training these models. All started from the original, full training set but used different training parameters. We see that the relationship between accuracy (on the test set) and validation loss is rather random. This is a dispiriting result, as it appears that a model that underwent a successful training is not necessarily competent.

We find again important variations between similar runs (the two simulations with 10 models in Table 6.8). This is connected to the small size of our dataset. Removing 20% of the samples from the training set generates radically different solutions, spread over a myriad of local minima on the optimization surface. Reducing the number of potential local minima to obtain more stable results with a limited number of models might require further reducing the net size.

Table 6.8: Number of Models and Impact on Accuracy

	Accuracy(%)[a]									
Nb. Models	mart.1	mart.2	mart.3	med.	min.	torq.	can.	sh.	vir.	Average
5[b]	60.4	13.2	93.3	4.2	91.3	100.0	38.5	0.0	90.7	54.6
50[b]	62.5	39.5	100.0	31.3	97.8	100.0	69.2	0.0	74.4	63.9
10[c]	91.8						63.4			82.0
10[c]	93.9						36.6			79.0
100[c]	89.8						48.8			84.8

[a] Percentage of samples for which the correct answer is in the top three predictions. Top-1 prediction accuracy in the range 21%–42%. Class probabilities averaged over the models.

[b] An epoch goes twice over the training set. Full dataset, pixel normalization [0,1]. Fourteen (14) feature maps in the convolutional layers.

[c] An epoch goes once over the training set. Balanced dataset, pixel normalization using negative values in the voids. Eight (8) feature maps in the convolutional layers.

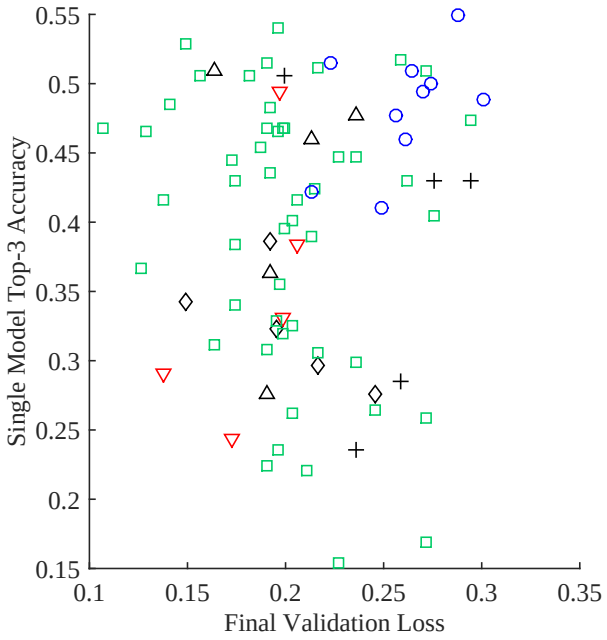


Figure 6.4: Accuracy of Single Models versus Validation Loss

From Table 6.5: first row (+), second row (△), third row (◊). From Table 6.6: last row (▽). From Table 6.7: first row (◯). From Table 6.8: second row (□). Full dataset, pixel normalization [0,1].

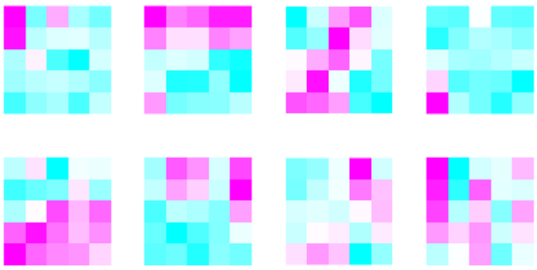


Figure 6.5: Optimized Filters in the First Convolutional Layer

Cyan: negative filter values. Magenta: positive filter values.

The optimized filters from the first convolutional layer are shown in Fig. 6.5 for one of the most accurate models. The patterns are simple enough: horizontal, vertical and diagonal lines. The last filter might be the kru-kru call (see Fig. 3.6). The top left and top right filters might not be able to detect anything significant.

6.2.4 The Failure of Data Augmentation

Having looped back to the insufficient size of our dataset, we considered the following strategies for data augmentation:

- The images were shifted left and right by 1, 2, 4 and 8 pixels. Multiple of 3 pixels were avoided, as the 3×3 max-pooling layer would have ended up considering the same patches as in the no-shift case. Instead of zero-padding the shifted images, the correct second of signal was re-extracted from the original audio files.

- The time axis was scaled by +/-10% to make syllables shorter or longer.

- The frequency axis was dilated and contracted by 2.5%. Here, a scaling rather than a shift was essential to preserve a physical relationship between the different harmonics of a call. The scaling factor was assessed from the data in Fig. 6.2.

These transformations are meant to help the network learn that small variations in the time or frequency directions are of no importance. They are not a substitute to the additional recordings from other birds that would represent diversity within a call class. But we still aimed to produce physically possible outputs. One does not know what the consequences of teaching a network wrong information would be in the long run. On the other hand, Lasseck [47] introduced unphysical transformations that successfully contributed to learning.

In our case, with a training set whose size is multiplied by 13, the accuracy falls from 82.0% to 66.5% (Table 6.9). Two *D. martius* calls and *D. medius* experience a sharp decrease. *P. canus* is the lone species for which data augmentation brings an improvement.

The augmentation did not induce a significantly longer training. Over the 10 models, the average number of epochs is 29.8. A validation loss of 0.21 is achieved on average, which compares to the numbers in Fig. 6.4. In

essence, training proceeded successfully, but the resulting models have a poor ability to generalize. Is the data augmentation inappropriate for some of the calls? For example, for the *D. martius* flight call, the duration of the syllables barely varies. Stretching the time axis forces the network to design a sophistication that does not exist in real life. With a drop in accuracy from 92.6% to 18.5%, certainly the proper way to augment this call has to be reconsidered. On the other hand, the syllables of *D. medius* do vary in pitch and duration. The augmentation had a much smaller chance of misrepresenting them, and it did. An explanation could be that the network is now too shallow to represent such a range of variation. A minimum number of neurons and connections is needed to assimilate all the information from the expanded dataset. In essence, there has to be a convergence between the complexity of the analysis and the size of the (augmented) dataset, which is not achieved here.

There are other approaches to augmentation that could be more effective. Pironkov [63] showed that the coupling of two neural nets treating somewhat related but different problems is actually beneficial. When attempting to recognize words in recordings of human voice, he improved his results by first training to identify whether the speaker was male or female. In our case, we could retrain the network of Grill & Schlüter [33] that differentiates bird calls from other noises and ask it to identify woodpeckers. In effect, this would be a split of the detection problem and of the identification problem but most importantly, it would capitalize on knowledge built on an additional 16000 spectrograms and using augmentation techniques. This is a net with some experience of spectrogram analysis. Our contribution would be

Table 6.9: Dataset Augmentation

	Accuracy(%)									
Dataset	mart.1	mart.2	mart.3	med.	min.	torq.	can.	sh.	vir.	Average
Bal.& Norm.	91.8	92.6	100.0	98.0	88.9	86.5	63.4	23.1	93.3	82.0
↑ Augmented	77.6	18.5	100.0	38.8	86.1	88.5	97.6	0.0	91.1	66.5

Percentage of samples for which the correct answer is in the top three predictions. Top-1 prediction accuracy respectively 38% and 30%. The augmented set is 13 times the size of the balanced set. Class probabilities averaged over 10 models. Eight (8) feature maps in the convolutional layers.

to reorient it toward woodpecker spectrograms.

These thoughts bring us to the approach developed in our last section: the retraining of existing deep image nets. The philosophy behind it is the same, except that these nets developed an expertise not on spectrograms but on pictures of random objects, and that this expertise was built from one million images.

6.3 Very Deep Nets

Retraining the legacy very deep image nets is expected to solve most of the issues identified so far. Owing to their depth[14], these nets have a great command of image analysis; they can identify many patterns and have learnt through their training a number of invariants, e.g. the fact that objects in images can change position, scale and orientation. Data augmentation similar to what we implemented above should not present any new information to such nets. For example, they already know that shifting an image horizontally does not change its meaning.

It thus appears that the only things that the very deep nets have left to learn are the specific patterns of woodpecker calls. For this, our small dataset could possibly suffice. If one difficulty is expected, it is the fact that image nets have learnt many invariants that do not exist in spectrograms, e.g. one might not flip a spectrogram left-right without significantly affecting its meaning. One might not displace syllables to higher frequencies. These invariants need to be unlearnt, because nature is full of funny replicas.

6.3.1 Setup and Retraining of the Nets

The very deep nets we consider in the present section are the ones listed in Table 2.2 in Chap. 2: AlexNet, VGG, Inception v3, ResNet 34 and 152, and DenseNet 169. All were originally trained on the one-million image database ImageNet and are available from the Pytorch libraries. Model downloads from the Pytorch libraries, retraining and making new predictions was managed through a set of Python/Pytorch scripts[15]. This process was also described in Chap. 4.

[14]Note that Grill & Schlüter's net, with 7 layers [33], is almost as deep as AlexNet, which has 8 layers (Szegedy et al. [79]).

[15]First obtained through Sohaib Laraba at the UMONS Numediart Institute. A tutorial on how to set-up a similar process is available at

Calculations were performed on a GPU[16]; the GPU memory constrains the batch size[17], which was set conservatively at 20 samples. An epoch was systematically considered as one pass over the training set. Some simulations had a fixed learning rate (0.001) and a limiting number of epochs (15 or 30), others an adaptive learning rate. In the adaptive version, the learning rate was divided by 10 at the end of an epoch if the training loss had not decreased. The corresponding Pytorch function also allowed the specification of a *patience*, i.e. a waiting period before actually proceeding with the learning rate decay. With a patience of 2, the training loss had to increase for another two epochs before considering an alteration of the learning rate. Training stopped when the learning rate reached 0.00001 or when the number of epochs reached 60.

The training procedure also used *momentum*, a second order parameter that affects the way the weights are changed based on the training loss. The use of momentum diminishes the oscillations in the training loss as it progresses toward a minimum[18]. The changes to the weights of the network are computed as follows:

$$\Delta \boldsymbol{w}(t) = \alpha \cdot \Delta \boldsymbol{w}(t-1) - \epsilon \cdot \frac{\partial E}{\partial \boldsymbol{w}} \quad (6.2)$$

$\boldsymbol{w}$ is the vector of weights, α the momentum (in our case, $\alpha = 0.9$), ϵ the learning rate, E the cost function, i.e. the training loss aggregated on the current mini-batch. Instant t corresponds to the current mini-batch, instant $t-1$ to the previous minibatch.

The images were resized to fit the input channels of the different nets (299×299 pixels for Inception, 224×224 otherwise). Both dimensions of the image were scaled to match the target dimensions; there is no requirement to conserve the proportions of the image. Hence our original 54×63 images were enlarged by a factor 3 to 5, all identically and using bilinear interpo-

https://pytorch.org/tutorials/beginner/finetuning_torchvision_models_tutorial.html. Retraining existing very deep nets has become a common approach in the machine learning community. In ecoacoustics, Lasseck [47] also deployed it, in parallel to the present work.

[16]Make & model: NVIDIA GEFORCE GTX 1080 Ti, GP102 processor, 11 GB in RAM.

[17]The net itself must also be loaded on the GPU. For the heaviest net, VGG, this blocks 574 MB (144 million weights). Most of the capacity is used for the calculations.

[18]If the optimization surface is a bowl, instead of jumping back and forth between opposite slopes, the optimization follows the steepest path to descend straight to the minimum. See Geoffrey Hinton's 2012 online course "Neural Networks for Machine Learning"(Lecture 6), at https://www.cs.toronto.edu/~hinton/nntut.html.

lation. This is to say that we could have afforded a greater frequency range (which might have introduced new confusions with more passerines) and longer audio clips. The pixel amplitudes were also stretched from the [0,1] interval to the [-1,1] interval, as discussed before.

By default the training set was considered in full, with its discrepancies between the class cardinals. The nets were trained for 10 classes, which included the noise class. The test set was again the one described in Table 6.3 (full dataset).

Table 6.10 shows the overall accuracy on the test set. This considers only the top prediction from each net, not the top three. All nets excel, the top performance being DenseNet with an adaptive learning rate and a patience of 2 (Run #5); the accuracy is 95.4%. Run #3, in which the nets were trained with a fixed and small learning rate, produces the lowest results. The optimization of the weights was too slow and was not allowed enough time to reach an optimal state. The accuracy per class is documented in Table 6.11. The noise class is well differentiated from the woodpecker calls; the top performance for this class is 99.3%, achieved by ResNet 152 (Run #4), which augurs a certain ability to shed false positives. For the call classes, accuracies greater than 90% are routine. Even for the underrepresented *P. sharpei* class, the accuracy rarely drops below 70%. *D. martius* (1) & (3), *D. medius* and *J. torquilla* reach 100%. Only the *D. minor* result might be seen as a shortcoming (top performance 83.3%). This is a call that should be easily identified because of the specific frequency range, but on the other hand it has a greater plasticity than other calls. The syllable production rate varies significantly from one sample to the next[19]. Finally, the *P. canus* result is en-

[19]Should this have been tagged as distinct calls? The profane does not know.

Table 6.10: Very Deep Net Retraining: Training Parameters

Run	**Nb. Epochs**	**LR**	**Patience**	**Overall Accuracy(%)**					
				Alex	Dense	Incep.	Res152	Res34	VGG
(1)	15	0.001		93.6	94.0	92.7	94.4	93.3	93.6
(2)	30	0.001		92.3	94.0	94.6	93.6	93.8	94.8
(3)	30	0.0001		92.9	92.9	93.6	91.9	90.4	92.3
(4)	60(max)	adaptive	0	92.5	94.6	94.2	94.2	94.4	94.4
(5)	60(max)	adaptive	2	94.4	**95.4**	94.8	94.2	94.0	92.9

Table 6.11: Very Deep Net Retraining: Results Per Class and Overall

Net	**Run**	noise	mart1	mart2	mart3	med	min	torq	can	shp	vir	avg	all
AlexNet	(1)	98.7	93.9	92.6	82.4	98.0	77.8	100.0	97.6	69.2	84.4	89.4	93.6
DenseNet	(1)	97.4	100.0	96.3	88.2	95.9	77.8	100.0	90.2	76.9	88.9	91.2	94.0
Inception	(1)	98.0	98.0	96.3	82.4	98.0	69.4	100.0	97.6	38.5	86.7	86.5	92.7
ResNet152	(1)	98.7	100.0	96.3	88.2	98.0	77.8	96.2	95.1	76.9	86.7	**91.4**	**94.4**
ResNet34	(1)	98.0	95.9	96.3	100.0	95.9	69.4	98.1	97.6	69.2	84.4	90.5	93.3
VGG	(1)	97.4	93.9	96.3	94.1	98.0	80.6	100.0	97.6	84.6	75.6	91.8	93.6
AlexNet	(2)	98.7	83.7	88.9	82.4	100.0	72.2	100.0	87.8	76.9	93.3	88.4	92.3
DenseNet	(2)	96.7	98.0	96.3	76.5	98.0	77.8	100.0	92.7	92.3	88.9	91.7	94.0
Inception	(2)	97.4	93.9	96.3	100.0	98.0	**83.3**	96.2	92.7	84.6	91.1	**93.3**	**94.6**
ResNet152	(2)	98.0	95.9	92.6	88.2	95.9	72.2	96.2	95.1	92.3	88.9	91.5	93.6
ResNet34	(2)	98.7	98.0	96.3	82.4	95.9	72.2	98.1	95.1	76.9	88.9	90.2	93.8
VGG	(2)	97.4	98.0	96.3	88.2	100.0	77.8	100.0	95.1	76.9	91.1	**92.1**	**94.8**
AlexNet	(3)	98.0	93.9	92.6	88.2	98.0	69.4	98.1	92.7	84.6	86.7	90.2	92.9
DenseNet	(3)	96.1	87.8	96.3	82.4	100.0	77.8	96.2	97.6	84.6	88.9	90.7	92.9
Inception	(3)	98.0	95.9	96.3	82.4	98.0	69.4	100.0	95.1	76.9	88.9	90.1	93.6
ResNet152	(3)	98.7	93.9	92.6	82.4	100.0	61.1	100.0	95.1	53.8	84.4	86.2	91.9
ResNet34	(3)	97.4	91.8	92.6	58.8	93.9	66.7	100.0	92.7	92.3	77.8	86.4	90.4
VGG	(3)	98.0	91.8	88.9	76.5	98.0	69.4	100.0	95.1	61.5	91.1	87.0	92.3
AlexNet	(4)	98.7	93.9	92.6	76.5	93.9	72.2	98.1	95.1	76.9	86.7	88.5	92.5
DenseNet	(4)	97.4	100.0	96.3	88.2	98.0	77.8	100.0	95.1	76.9	88.9	**91.9**	**94.6**
Inception	(4)	98.0	95.9	96.3	82.4	100.0	72.2	100.0	95.1	76.9	91.1	90.8	94.2
ResNet152	(4)	99.3	98.0	92.6	88.2	95.9	75.0	98.1	95.1	84.6	86.7	**91.4**	**94.2**
ResNet34	(4)	99.3	100.0	96.3	88.2	98.0	77.8	100.0	92.7	69.2	84.4	90.6	94.4
VGG	(4)	98.0	100.0	96.3	94.1	95.9	75.0	100.0	95.1	76.9	86.7	**91.8**	**94.4**
AlexNet	(5)	99.3	95.9	92.6	88.2	98.0	77.8	96.2	92.7	76.9	93.3	**91.1**	**94.4**
DenseNet	(5)	98.7	100.0	96.3	88.2	98.0	**83.3**	100.0	**97.6**	76.9	86.7	**92.6**	**95.4**
Inception	(5)	98.7	95.9	96.3	88.2	98.0	80.6	100.0	92.7	69.2	93.3	**91.3**	**94.8**
ResNet152	(5)	98.7	95.9	96.3	82.4	100.0	77.8	100.0	92.7	76.9	86.7	90.7	94.2
ResNet34	(5)	98.0	98.0	96.3	82.4	98.0	75.0	100.0	95.1	76.9	86.7	90.6	94.0
VGG	(5)	96.1	93.9	96.3	88.2	95.9	80.6	100.0	92.7	69.2	86.7	90.0	92.9

Class accuracy: percentage of samples in this class that were correctly predicted. Average: average accuracy over classes. All: percentage of samples in the test set that were correctly predicted. In bold, the models that exhibit the best performances and were selected for further predictions. The numbers that particularly motivated this choice are also in bold.

couraging. The maximum accuracy is 97.6% and is reached by six models, four of which in Run #1 and the other two being DenseNet architectures. Early stopping, i.e. keeping the training short as in Run #1 (15 epochs), is a common tactic to avoid overfitting. Obviously, it worked well for *P. canus*, which could mean that we still lack data for this taxon. The top performing DenseNet (Run #5) posts a maximum score for five classes, including *P. canus*.

6.3.2 Analysis of the Field Datasets

The field datasets were segmented into short audio files (see Chap. 4), from which series of 1000–3500 Hz $\times$ 1 s images were generated with a 15% overlap. The spectrogram amplitudes, in dB, were cut-off at 30 dB below the maximum in the picture, and normalized to [-1,1]. The resulting image sets range in size from 13051 images in Tenneville, where the recording station only scanned below 1500 Hz, to 643901 images in Remerschen, where the biodiversity is extreme; see Table 6.12.

Table 6.12: Audio Segments and Images in Field Datasets

Dataset	**Audio Segments**	**Images**
TN	3732	13051
LPR1	21831	73883
LPR2	30072	98450
LPR3	52061	172992
RM	150894	643901

We used an ensemble of nine models to label these images. The selection is highlighted in bold in Table 6.11. These are the models for which the average class accuracy exceeds 91% and the overall accuracy exceeds 94%. For each class except *P. sharpei*, at least one of the selected models produces the maximum score for that class. The predictions of the different models were pooled together by counting votes and retaining the most-voted class[20]. In case of ties, we went with the choice of top-performing model DenseNet(5).

[20]This is majority-voting, also used in Random Forests. We otherwise experimented with averaging the class probabilities produced by the various nets, but the accuracy was significantly and systematically lower. We concluded that class probabilities are not an absolute measure, but should be read in the context of a given net. The numerical outputs of different nets cannot be compared.

Table 6.13: Accuracy (%) on Images of Woodpecker Calls in Field Datasets

Net	TN	LPR1	LPR2	LPR3	RM
AlexNet(5)	91.4	74.4	50.2	57.7	56.0
Inception(2)	93.4	78.9	40.6	43.4	51.4
Inception(5)	91.4	82.1	46.2	56.6	54.8
ResNet152(1)	89.8	75.8	49.6	50.2	50.9
ResNet152(4)	96.4	76.7	40.8	40.8	53.0
DenseNet(4)	95.4	79.2	76.0	60.3	63.1
DenseNet(5)	96.4	83.4	72.2	37.8	57.3
VGG(2)	96.4	83.1	58.0	41.9	55.4
VGG(4)	94.4	79.6	53.4	52.1	57.2
Ensemble	97.5	83.4	64.5	55.1	55.3

Only the predictions other than noise were reviewed to assess the ground truth. Calls that were not detected by any model (false negatives) are a blind spot.

Table 6.13 documents the accuracy on the subset of woodpecker calls; the overall accuracy would merely reflect the performance on noise, as the noise segments outnumber the woodpecker calls. The numbers in Table 6.13 reflect two trends we have already seen in previous analyses. First, the performance on the field datasets is much decreased compared to what was obtained on the XC/KT data; the underlying reason is the limited ability of the nets to generalize. The XC/KT training dataset remains small; it does not allow the deep nets to learn to identify woodpecker calls in all their variations. Secondly, the performances decrease as we move toward the right of the table. The datasets were intentionally ordered by month of recording: TN was March–early April (and with a restricted bandwidth), LPR1 was March, LPR2 April, LPR3 May and RM end of April–May. As in Chap. 4, we observe that woodpeckers become harder to identify as the passerines take over the acoustic space. This being said, the accuracy for TN is outstanding and for LPR1 rather good. DenseNet(4) offers the best sustained performance, with accuracy over 60% for all datasets. Moving on from the deepest net to the most shallow, AlexNet(5) also posts decent results. On the XC/KT test set, AlexNet(5) accurately detected noise (99.3%, top accuracy) and DenseNet(4) reached the maximum accuracy for *D. martius* (1) and *J. torquilla*, calls that are massively present in LPR and RM. These links have a limited justification

value though; ResNet(4) also achieved 99.3% on the XC/KT noise class and many models had an identical 100% on *J. torquilla* which dominates the RM dataset. The XC/KT test set only has 300 samples; percentage gains on this set might be affected by oddities rather than by a true ability to identify a class. In effect, we ranked the different models on marginal results. We notice in particular that DenseNet(5) is not after all the best model and thus not the optimal choice to arbitrate ties in majority-voting. Thus the accuracy for the pooled models does not necessarily exceed the accuracy of individual models. For the three toughest datasets (LPR2, LPR3 and RM), using the DenseNet(4) model alone is a better choice than the ensemble.

For the pooled predictions, the accuracy per class is documented in Table 6.14. The results are somewhat erratic, owing to the fact that the different classes are not necessarily well-represented in all datasets. There is only one *P. canus* call in RM and it was correctly identified. The corresponding 100% accuracy sheds limited insight.

The decay in accuracy for *D. martius* (1) has more significance, and above all, so does the decay in accuracy for the noise class. 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 641561 noise samples, and thus the ability to isolate these samples is critical.

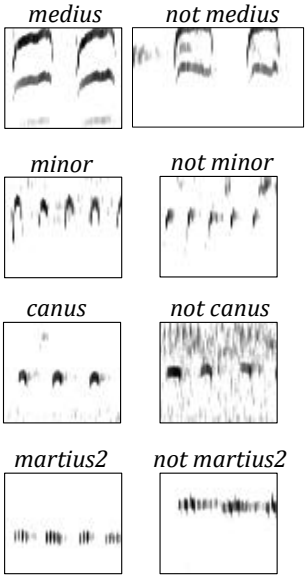


Figure 6.6: Confusions in Remerschen and La Petite Raon

Tables 6.15 and 6.16 show this same issue from another perspective. Table 6.15 lists the number of false positives per woodpecker call class, i.e. the misclassified images that are either noise or another call. The false positives are manageable in TN or LPR1, then escalate to exceed 10,000 images in LPR3 and RM. This is put in perspective 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 nets: *D. martius* (1), *D. medius*, *P. canus* and *P. viridis*. We seem to have favored models that predicted these classes a lot, rather

Table 6.14: Accuracy (%) Per Class in Field Datasets, Using Pooled Models

Class	**TN**	**LPR1**	**LPR2**	**LPR3**	**RM**
D. martius1	96.2	84.7	71.7	30.8	0.0
D. martius2	100.0	75.5	54.2	75.0	
D. martius3	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

D. minor and *P. sharpei* were not found in the recordings.

Table 6.15: False Positives and Actual Calls in Field Datasets

Class	**TN**		**LPR1**		**LPR2**		**LPR3**		**RM**	
	FP	Truth	FP	Truth	FP	Truth	FP	Truth	FP	Truth
D. martius1	1	26	64	2819	1754	996	2285	13	572	6
D. martius2	1	14	5	294	5	48	5	24	563	0
D. martius3	4	5	92	71	70	5	72	8	356	0
D. medius	21	0	132	111	2673	13	3970	14	4594	0
D. minor	21	0	191	0	46	0	60	0	2684	0
J. torquilla	5	4	15	0	231	2	554	0	2476	2263
P. canus	3	148	57	42	1134	425	2368	208	1187	4
P. sharpei	0	0	10	0	15	0	34	0	147	0
P. viridis	6	0	71	51	1411	0	2189	0	805	67
Total	62	197	637	3388	7339	1489	11537	267	13384	2340
FP(%)	24.4		18.4		88.4		98.7		91.2	

False positives generated by the pooled models. They can be either noise or another woodpecker call class. The "truth" is the actual number of images of the corresponding class in the dataset, given for perspective. As the numbers show, a *D. medius* was on site in LPR for about a week at the end of March 2018, and then sporadically. Its calls were answered by the *D. martius* whose territory overlapped. *D. medius* is about 10 times more abundant than *P. canus*, but this was our first recording of the species. *D. minor* and *P. canus* were not recorded.

Table 6.16: Confusion Matrix Accumulated over all Field Datasets

Actual	**Predicted Classes: Nb. of Samples**									
Classes ↓	mart1	mart2	mart3	med	min	torq	can	shp	vir	noise
D. martius1	3131	3	1	4	7	4	39	15	251	405
D. martius2	3	280	0	1	0	1	1	0	3	91
D. martius3	0	0	79	0	0	0	0	0	0	10
D. medius	0	0	2	113	0	0	1	0	0	22
J. torquilla	0	0	0	5	2	1273	0	0	0	989
P. canus	101	5	17	10	0	6	482	0	133	73
P. viridis	2	0	0	0	13	2	2	0	60	39

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. *D. minor* is abundantly 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. 6.6 for a few examples.

Table 6.16 shows the confusion between the different classes, consolidated over the five datasets. *D. minor* and *P. sharpei* were detected at none of the locations and thus do not appear in the table rows. There are relatively few confusions between the classes of woodpecker calls, aside from a triangle between *D. martius* (1), *P. canus* and *P. viridis*. All are rattle calls and the last two are indeed sometimes hard to discriminate. Yet 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. The weakness of the woodpecker/noise discrimination is exemplified with *J. torquilla*; most of the samples are from RM, and 44% of them (989 images) are missed by the model pool. Half of these (23% of all *J. torquilla* images) were actually correctly identified by one or more models and not properly promoted by the vote because the correct models were in the minority.

Figure 6.7 shows the distribution of votes for all classes and all datasets. We see that the models are unanimous (and correct) for 41% of the images and near unanimous for 51% of the images. For 17% of the images, 1–4 models had the correct answer but were overruled. 12% of the images are not recognized at all. The models that are right but overruled are most often DenseNet ($>$500 images) and least often ResNet ($\tilde{2}$00 images). Here again, we find the idea that the initial models were either not correctly evaluated

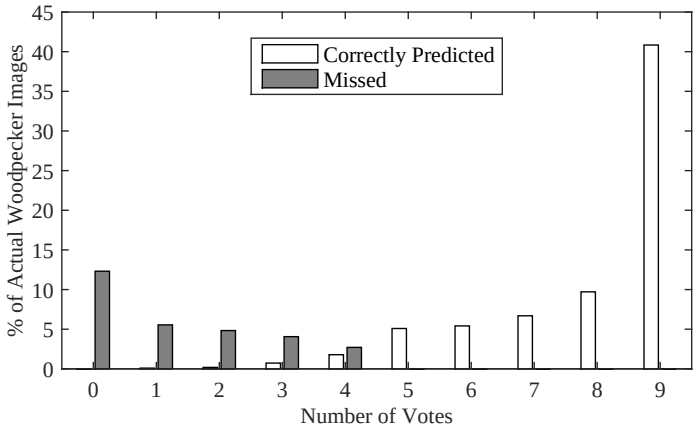


Figure 6.7: Number of Votes When Identifying Images of Woodpecker Calls in Field Datasets

and/or not correctly chosen. We also find a critique of the majority-voting procedure; if too many models in the ensemble are incompetent, the good models will not save the day.

In the end, 21% of the woodpecker images are misdiagnosed as noise. Many of these misses were to be expected, because these are poorer images at the beginning or at the end of calls. We also note that we were not able to review all the images predicted as noise. To construct a ground truth, we examined all the calls for which at least one image was identified as something other than noise by at least one of the models. The analysis has a blind spot for the woodpecker calls that are entirely predicted as noise by all the models. And this begs the question: how many more images did the ensemble unanimously fail to detect?

There is also ample ground to defend these results. The numbers discussed above deal with the identification of partial images of woodpecker calls. Some caught only a part of a syllable or the fuzzy tail of a call. In truth, not all images that contain part of a woodpecker call enable the identification of said call. Hence we also computed the accuracy at the call level, i.e. for a given sound segment containing a woodpecker call. The identification was deemed correct if at least one of the images generated from this segment was correctly identified as this call. The numbers are in Table 6.17 and are significantly more pleasing. In many cases, 100% of the calls are identified. Even late in the season, 72.5% and 81.7% of the calls from the birds that owned the

Table 6.17: Accuracy (%) Per Class in Field Datasets, at the Call Level

Class	**TN**	**LPR1**	**LPR2**	**LPR3**	**RM**
D. martius1	100.0	98.7	92.3	50.0	0.0
D. martius2	100.0	94.1	66.7	100.0	
D. martius3	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

territory (*P. canus* in LPR; *J. torquilla* in RM) are detected. The segmenting of calls into several images seemingly drives a spotty recognition; some images are unfocused, some syllables obscured, the beginning and end of calls are not necessarily telling. But the overall picture is that a significant amount of calls are correctly picked up. We might conclude that the images we used are too small and that larger images would be more reliably identified; we also saw above that the short image duration did not help the discrimination of some of the rattle calls. However we recall that our choice of image size was driven by a need to compensate for the small size of the training database.

From Table 6.17, the false positives now amount to 1.4%–18.3% of the noise audio segments. In a dataset such as RM, having to review only 5.9% of the irrelevant data is what makes any analysis possible to start with. The ability to discriminate noise and woodpeckers is a tributary of the training set completeness; ideally all the variations on woodpecker calls should be represented, and all the possible imitators should be accounted for on the noise side. For the detection of drums in Chap. 4, we assembled a noise class with false positives from detection by other means; it was therefore quite relevant. Here, our noise class was built out of the background of XC/KT files. We did not assemble it with a consideration for the most likely sources of confusion. In particular, the buzzards and owls that are easily mistaken for *D. medius* were omitted. Hence, after the number and diversity of singers picked up in the late spring, false positives increased dramatically.

The analyses in the present section showcase DenseNet (169 layers) as the most potent net. There is however a random nature to training; the samples are shuffled and presented to the net in a random order. In other simula-

tions (not shown), ResNet 152 was the most able net. Above, we saw an example of AlexNet performing surprisingly well for a small net; elsewhere (not shown), it was ResNet 34. The VGG net (19 layers) is impractical because of its excessive size, but its large number of weights gives it the analysis breadth of a deeper net. It performed reasonably well in a number of simulations. Almost all training runs, using different configurations for the learning rate, generated models that exceeded 94% of accuracy. However, the models trained with an adaptive learning rate fared better with the field datasets (Table 6.13). Overall, our view of the capabilities of the different net architectures evolved throughout our work and depending on the latest results. This is how the analyses in Chap. 4 and 5 used only the deeper nets, that showed more promise.

6.3.3 Variants

Models Trained on Additional Data

We retrained the six nets using the XC/KT data and data from LPR1: 3388 images of woodpecker calls, from six classes, and 2252 images of noise, either from false positives or from segments preceding or following calls. The additional woodpecker calls bring in limited diversity as they were uttered by only a handful of birds, but the noise group is of particular interest. It contains calls likely to be confused with woodpeckers and the ordinary avian community of LPR.

ResNet 152 took the lead in this Run #6, closely followed by DenseNet and VGG (Table 6.18). We added these three models to the previous ensemble and computed predictions for the other datasets. The results are in Table 6.19. The accuracy for the woodpecker calls is shown first for the individual models, each time with a comparison with an equivalent model trained only on XC/KT, and then for the ensembles, old and augmented.

The additional training data leads to an improved performance across the board. The TN accuracy with the old ensemble of models was already very good and does not change; for the other datasets, LPR2 gains 7 points and LPR3 and RM gain 2. The improvements for LPR2 and LPR3 were expected because of the relation with LPR1, but the gain for RM is fully independent. The DenseNet(6) model reaches 67.3% for RM. The previous best was 63.1% with DenseNet(4). The RM recordings contain mostly *J. torquilla* calls that

Table 6.18: Very Deep Net Retraining on XC/KT and LPR1: Results Per Class and Overall

Net	**Run**	noise	mart1	mart2	mart3	med	min	torq	can	shp	vir	avg	all
AlexNet	(6)	98.0	100.0	96.3	94.1	93.9	66.7	98.1	92.7	69.2	91.1	90.0	93.3
DenseNet	(6)	98.0	95.9	96.3	100.0	98.0	75.0	100.0	95.1	92.3	88.9	**94.0**	**95.0**
Inception	(6)	98.7	95.9	96.3	76.5	95.9	72.2	100.0	95.1	84.6	82.2	89.7	93.1
ResNet152	(6)	98.7	98.0	96.3	94.1	98.0	77.8	100.0	95.1	84.6	91.1	**93.4**	**95.4**
ResNet34	(6)	98.0	95.9	96.3	88.2	98.0	75.0	100.0	95.1	84.6	84.4	91.6	94.0
VGG	(6)	99.3	98.0	96.3	88.2	98.0	80.6	100.0	92.7	69.2	88.9	**91.1**	**94.8**

Table 6.19: Accuracy (%) on Woodpecker Calls in Field Datasets, Using XC/KT/LPR1 Training Data

	TN	**LPR2**	**LPR3**	**RM**
ResNet152(4)	96.4	40.8	40.8	53.0
ResNet152(6)	93.9	76.4	61.0	56.9
DenseNet(5)	96.4	72.2	37.8	57.3
DenseNet(6)	94.9	63.0	56.6	67.3
VGG(4)	94.4	53.4	52.1	57.2
VGG(6)	94.4	74.5	52.4	63.5
Old models only	97.5	64.5	55.1	55.3
All models	97.5	71.9	57.3	57.4

Table 6.20: False Positives in Field Datasets, Using XC/KT/LPR1 Training Data

	TN	**LPR2**	**LPR3**	**RM**
D. martius1	1	1285	1552	589
D. martius2	1	2	1	436
D. martius3	2	39	37	324
D. medius	13	1629	2249	3616
D. minor	18	29	32	2317
J. torquilla	5	164	414	2440
P. canus	3	785	1612	1111
P. sharpei	0	5	19	146
P. viridis	6	912	1379	811
Total	49	4850	7295	11790

were not found in LPR1; the gain is without doubt the work of the more relevant noise class, including birdlife present at both locations, which are roughly 200 km apart.

Table 6.20 presents the number of false positives with the augmented ensemble, to be compared with the numbers in Table 6.15. A third of the false positives are removed at LPR (-34% for LPR2, -37% at LPR3); for RM the decrease is a more modest -12%, but this translates into approximately 2000 fewer false positives. The greatest success is for the reduction of false *D. medius* detections (LPR2 -39%, LPR3 -43%, RM -21%). This time, the training set included owls and buzzards.

Forcing the Nets to Unlearn Invariance

In this variant, an arrow was added at the top left of all pictures. The intention was to imprint a scale and a left-right direction so that the nets could unlearn invariance to size and to left-right flips of the images. The results on the XC/KT test set are encouraging (Table 6.21), with a top performance for almost all nets. Inception, the net with the most to unlearn, posts the best score. However we have learnt that a good performance on the XC/KT test set does not necessarily translate into a great ability to generalize. The accuracy was also computed for the field datasets (Table 6.22). The results are mixed. Inception(!) and VGG perform less well than in the baseline simulation (Table 6.13). *P. canus* loses accuracy at all locations. Noise is better predicted at LPR2 and LPR3, which translates into fewer false positives, but the performance falls for the other three datasets. Overall, the benefits from the arrows are not convincing.

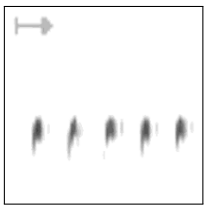


Figure 6.8: Call Image with an Arrow

There are two issues with our implementation: 1) the arrows occupy the same pixels in all images, and likely cause the nets to ignore these pixels as devoid of information; 2) the arrows are in the upper left corner, most often removed from the important patterns, and are thus not considered together with the calls in the lower convolutional layers. Regarding 1), the arrows would be better suited to sounds of different durations; scalable arrows would help with comparing images in which an identical pattern appears with different sizes. Regarding 2), a grid overlaid on the full image might be a better option.

Table 6.21: Accuracy (%) on XC/KT Test Set Using Image Modifications

	Arrows	**Arrows & Pixel Dropout**	**Arrows & Added Noise**
AlexNet	93.1	86.1	91.3
DenseNet	95.0	92.1	94.2
Inception	95.4	92.7	92.9
ResNet152	95.0	72.8	93.1
ResNet34	94.6	87.7	94.0
VGG	92.7	72.6	93.1

Table 6.22: Accuracy (%) in Field Datasets with Arrows Added

Class/Model	**TN**	**LPR1**	**LPR2**	**LPR3**	**RM**
			Woodpecker Call Images		
AlexNet	92.9	77.5	55.9	59.9	58.7
DenseNet	95.9	85.1	63.9	33.3	55.7
Inception	94.9	78.8	45.7	50.2	56.4
ResNet152	94.9	71.8	67.4	47.6	58.6
ResNet34	94.9	81.9	64.0	49.1	60.0
VGG	94.4	76.6	49.7	38.6	56.8
			Woodpecker Calls, Pooled Models		
D. martius1	100.0	98.7	90.9	66.7	0.0
D. martius2	100.0	94.1	75.0	100.0	
D. martius3	100.0	100.0	100.0	80.0	
D. medius		91.2	100.0	100.0	
J. torquilla	100.0		100.0		83.5
P. canus	98.2	84.2	59.4	67.5	100.0
P. viridis		94.1			40.9
All woodpeckers	98.6	97.4	77.4	70.4	81.7
Noise	98.3	97.6	87.7	87.3	93.4

Pixel Dropout

In this variant, 25% of the pixels were set to white (i.e. 1) randomly. The selected pixels differ in the three RGB components of the image. This yields poor results on the XC/KT test set (Table 6.21). For reference, DenseNet had already an accuracy of 88% after one epoch of training in the next variant. Here, four nets do not even reach that number.

All the legacy nets were trained using dropout but ResNet (He et al. [34]). Consequently, they have already learnt to maximize the potential of all their connections. ResNet 152, which has not, has a particularly poor 72.8% result. Incidentally, there might not be much sense in supplying partial images to the nets when the initial material is already quite concise. With 54×63 images, mostly white, the pixels that hold the key to species identification are few.

Noise Addition

In Lasseck's study [47], the most effective augmentation technique consisted in adding random segments of noise to the audio of calls (in succession, not superimposed). The calls were processed using random crops of 5 s, and thus during the training phase, images were not necessarily centered on calls and included undesirable signals or relative silence. Lasseck's nets targeted 1500 classes of bird calls but had no noise category. The separation between noise and calls had been done upfront. This is an important difference with our nets because in Lasseck's case, adding noise chunks to the signals was less likely to confuse the classifier. In addition, the noise chunks contained little biophony; the purpose of the nets was to identify all bird species, hence no bird species were rejected to the unwanted class. Finally, Lasseck's noise chunks came from a separate dataset that the nets had not previously seen. They were new information.

In our version, we concatenated images of the training set with an image of the noise class, along the time dimension. This produced spectrograms with one second of interesting signal, followed by one second of unwanted calls or relative silence. For images of the noise class, this simply resulted in 2 seconds of noise. To maintain a consistent image size throughout the analysis, we white-padded the images of the test set. The arrow in the top left corner was also added.

Here the results are fair (Table 6.21) but down from the variant with only the arrows.

6.4 Conclusions

As in Chap. 5, we present in Figs. 6.9, 6.10 and 6.11 summary illustrations of the signals that were detected in the three field datasets. These final pictures are the desired outcome of acoustic monitoring. Whereas the drums analysis was dominated by *D. major*, we found a greater variety of species in the calls analysis. The LPR recordings were particularly fruitful. On Fig. 6.11, we see that the territory was initially occupied by *D. martius*, which used its advertising 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. 5.22). 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. We note that calls identifications are more confident than drums identifications; we did not need to seek context or other signals in the recordings to confirm the species. To the contrary of drums, advertising calls are without question species-specific. A slight disappointment in our work was to not record *D. minor*. Of all the species that we had a chance to capture, this is the only one missing. We have seen that 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. Did we miss these signals? Since we recorded in *P. canus* territory and had the calls on tape, we never envisioned that some of the drums might have been *D. minor*. As for the calls, without a positive identification in our data, we cannot be assured that we are able to detect them.

Considering the developments that led to these results, there is no doubt that without deep neural networks, we would have been unable to analyze the RM and LPR datasets. This is the primary gain from the technology: deep nets reduce vast datasets into tentatively annotated datasets of a manageable size. A large part of these tentative annotations are actually correct: between 73.5% (LPR3) and 100% (TN) of the calls were successfully detected and identified. The manual review is still necessary because of a non-negligible amount of false positives. They totaled around 20% of the detections early in the season and around 90% late in the season (LPR3, RM). In practice, in the worst case (RM), the false positives amounted to 13384 images out of a dataset of 643901 images. The reality behind the seemingly dispiriting late season numbers is that the deep nets transformed an impossible review into

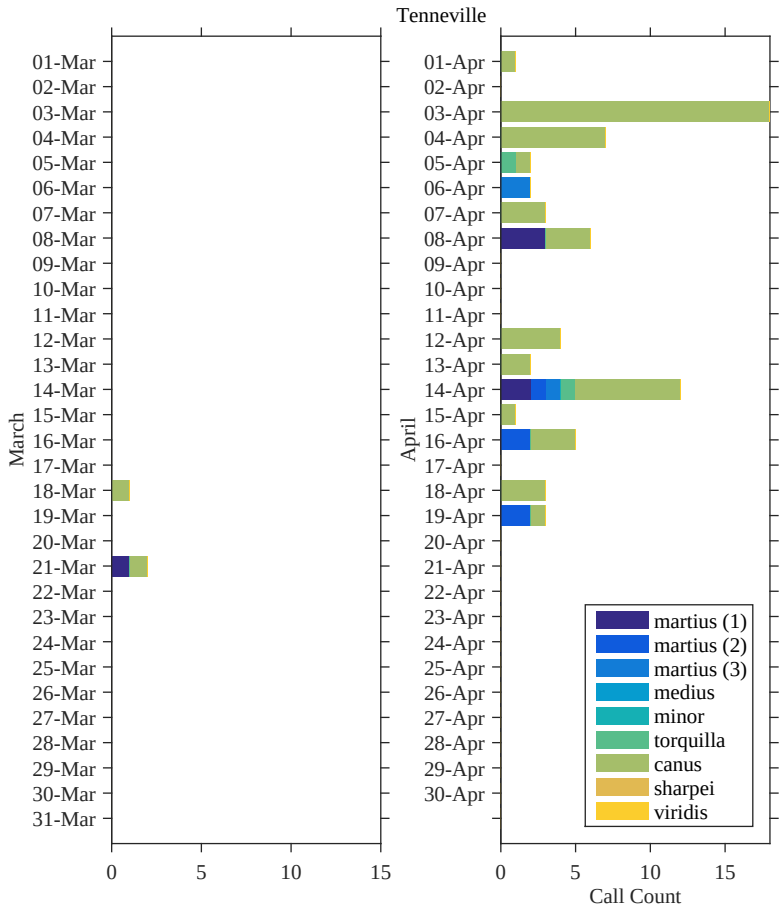


Figure 6.9: Tenneville: Calls Identifications by Date

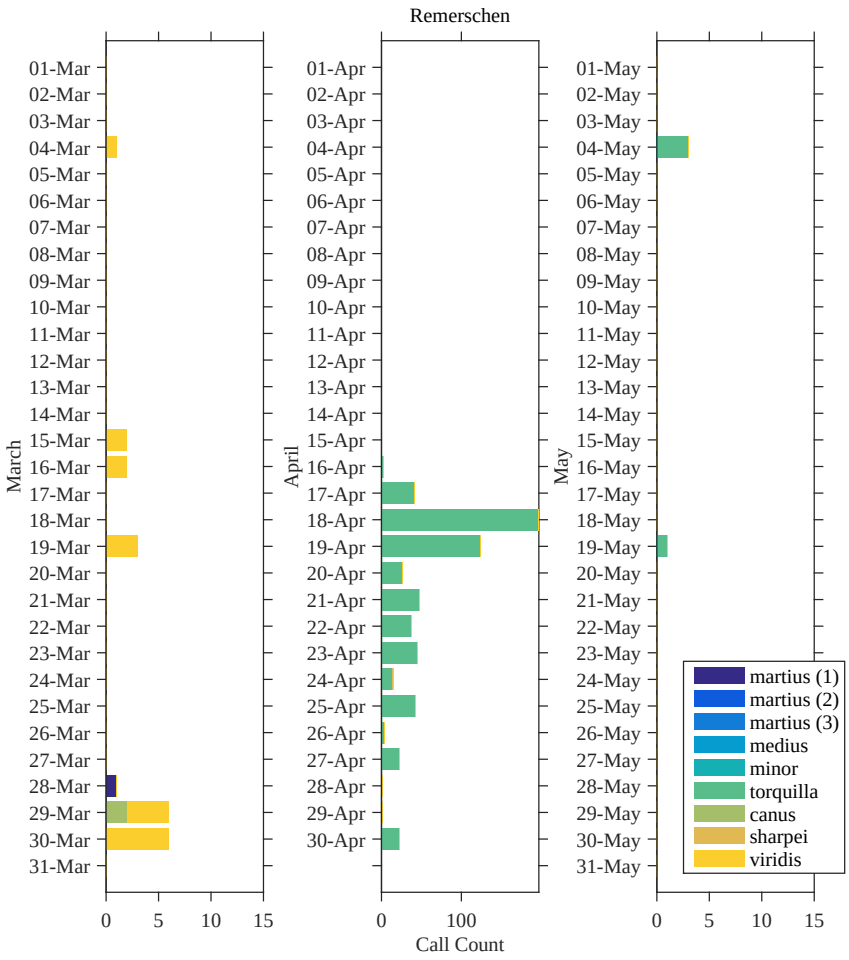


Figure 6.10: Remerschen: Calls Identifications by Date

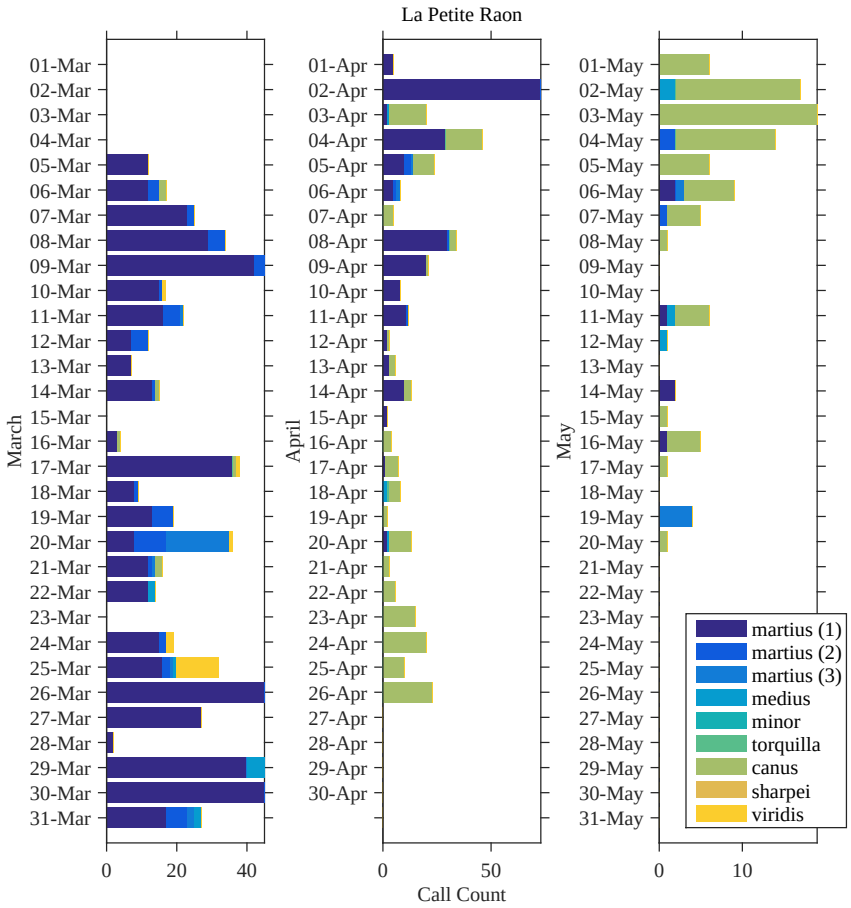


Figure 6.11: La Petite Raon: Calls Identifications by Date

a tedious review.

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). Improvement in the predictions is achieved when the training set is complemented with images of the noise class that had previously been mistaken for woodpecker calls. There is definite value in building a noise class that represents the full scope of biodiversity that the nets might subsequently have to deal with. This had been done more diligently for the detection of drums. Here we started out with 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. We recorded unsuspected copycats in Remerschen and La Petite Raon.

This remark brings the following issue: to be complete, the noise class has to be large. In nature, the woodpecker/not a woodpecker problem is not symmetrical, except maybe in March. Now, 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, that 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. Our study is designed to minimize the false negatives, and thus has to cope with the false positives. We would rather have audio to review in excess than risk missing woodpecker calls. This being said, the vastness of the datasets limited us to reviewing only the calls that the nets had qualified as woodpecker calls. Our study is blind to false negatives. Table 6.16 indicated that most of the misidentifications (among the calls we knew of) had been confused with noise. How many calls are there that we did not know of, i.e. that none of the nets detected? The predictions eventually improved when we aggregated the results from

the several crops taken out of the same call. Still, if we compare to Chap. 4, it seems like detecting drums is a simpler problem than detecting calls, and therefore one that could be taught with a more modest training set.

In Lasseck [47], Inception was the best performing network. In our case, we obtained the best results with different architectures in different runs and sometimes in repeated identical runs. The deeper nets often outperform the shallower ones, but it remains hazardous to pick just one. Using a ensemble of models either improves or deteriorates performance compared to a single good model.

The different strategies we implemented to augment the training set had limited success. The original training of deep image nets was extensive and already deployed most of the basic tricks of data augmentation. Modifying the position of objects, degrading the image quality, changing their scale, flipping or rotating them were things that could not be taught again with the same benefits. Dropout was also already used. Whereas it had been a success in other studies, we were not able to gain anything from adding noise to our data. This was likely because we did not have noise data at our disposal that the nets had not already seen before. The only augmentation technique that improved our results was to actually add data from one of the field datasets.

In the list above, not all transformations are desirable for spectrograms. They should not be flipped or rotated and should maintain fixed axes. This is where image analysis and sound analysis differ. We saw an undesired impact in Chap. 4 when Inception, which is quite apt at handling changes of scale, picked up the much slower demonstrative tapping when searching for drums. Here we tried to counter the transforms by adding a fixed arrow to the images, but the results were modest. Forcing the nets to unlearn their unnecessary invariants might require just as much data as it took to learn them.

Taking multiple 1-second crops from the calls is also an augmentation technique, and one that allowed us to proceed with training deep nets. The images were short; some represented just one syllable or one part of a syllable and were difficult to distinguish from non-woodpecker calls. Longer crops would allow incorporating some of the calls structures into the analysis, which the deeper nets certainly have the analytical power to study, but again, the limitation is on the dataset size. The multiple 1-s crops proved adequate enough to detect and identify woodpeckers. Notably, the identifi-

cation of calls occurred with a great accuracy and yielded few confusions between the different species. Only *D. martius* (the rattle call), *P. canus* and *P. viridis* were found to exchange samples. This was understandable for the last two, whose calls are sometimes rather similar, but less for *D. martius*. We anticipate that this point could also be improved with longer images.

In the end the proper synthesis of Pironkov [63], Grill & Schlüter [33] and Lasseck [47] could have been to retrain the legacy image nets (one million images) with the data from the BAD challenge used by Grill (16000 images) for the same two-class objective of bird/not bird, and from that base to retrain for woodpeckers or noise. Anything to incorporate more data.

Conclusions

At the conclusion of this work, we have contemplated many facets of the acoustic monitoring of woodpeckers. We considered the practical aspects of recording audio in the wild. An autonomous recording station was built and operated for three successive springs in different places and environments. It had the capability to calculate the Acoustic Complexity Index (ACI) on board, which proved beneficial beyond the original scope of this indicator. The ACI detects temporal variations in acoustic intensity such as the ones caused by birdsong and woodpecker drumming. In consequence, the vast portions of datasets that do not contain these signals can be rapidly discarded. We turned an indicator that was intended to measure species richness into a mean to scale down the field datasets.

To support the development of woodpecker detection and identification algorithms, we gathered training data from crowd-sourced public archives such as Xeno-Canto and Tierstimmen. We also gained access to the private recordings collection of British birder Kyle Turner. We mapped out the signals we needed to target: on the one hand the drums, both territorial (seven species) and soft (ten species), and on the other hand a set of calls that were either particularly revealing (two calls of *D. martius*) or used in an advertising capacity (seven species). Drums and calls were subsequently treated separately.

For both, we implemented detection by using legacy deep image nets (AlexNet, VGG, Inception v3, ResNet and DenseNet) that we retrained on spectrogram images of woodpecker signals. The approach was a success; at a minimum, the outcome was a significant further reduction of the datasets. The other methods we explored were unable to detect as many signals while keeping the number of false positives manageable. We experienced the most difficulty with data that had been recorded late in the season when the avian

population increases. The variety of bird songs caused the deep nets to commit more errors.

We were able to identify drumming species in two different manners, the first one by handcrafting acoustic parameters and combining them with the simple k-NN classifier and the second one by using deep image nets again. Because of the specificity of drumming, a precise time signal, the handcrafted parameters and k-NN fared better. Temporal constructions are not the strong point of deep image nets. This being said, deep image nets produced results that were not far behind k-NN and required no manual labor to implement. Image generation is straightforward and the analysis of the largest dataset (RM) took only a few minutes. On the contrary, calculating handcrafted features demands a time-consuming manual oversight. Depending on the number of drums, it can take days. The k-NN analysis on the other hand is instantaneous.

We implemented the identification of calls using deep image nets as well. In this case, with syllables readily recognizable on spectrograms, the method was a perfect fit. The success of deep image nets is also the consecration of spectrograms as the best audio features, regardless of species. The shortcomings we suffered were essentially linked to the insatiable hunger of deep nets for more data.

In the end we were able to successfully monitor woodpeckers for three reproductive seasons. We detected drums and calls and we identified the species issuing these signals. We found an abundance of drums and calls, and a fair amount of communication between individuals, intraspecies and interspecies. Distant signals are harder to analyze, but as woodpeckers circle around over their territory, a time comes when they make noise in the vicinity of the recording station. In other words, woodpeckers are good subjects for acoustic monitoring. We were able to capture rare *J. torquilla* drums and taps in Remerschen, iconic *D. medius* calls in La Petite Raon and many occurrences of *P. canus*, both as a rarity in Belgium and in the heart of its territory in the Vosges mountains. Our final results depict the complete acoustic activity of woodpeckers in three different environments. Even before going through the entire analysis, the ACI spectrograms let us watch an entire spring in the life of woodpeckers unfold in less than 100 images. These outcomes were previously inaccessible and open up new possibilities for ornithological research. The technology we developed can support a number of behavior, evolution and conservation studies for this group of birds. The

question of species character in drums comes to mind.

Deep image nets played a decisive role in our achievements. Although they yielded outstanding results overall, we singled out two limitations that could translate into future research: the need for very large datasets and the fact that the image invariants that the deep nets learned in their original training are improper for spectrograms. For the first point, we concluded that adding data by any means was the right direction for future work; this could entail for example adding unseen noise data like Lasseck [47] or using a secondary objective like Pironkov [63], as long as there is more data available to train for this objective. For the second point, we suspected that having the image nets unlearn their superfluous invariants might require the same amount of data as in their original training. In that case, why not build nets directly for sound? The different architectures could be retrained from scratch using a database of sounds and possibly different spectrogram resolutions instead of RGB channels. The problem is of course to find a general-purpose database of one million sounds. Salamon & Bello [68] gathered about 9000 images of sounds from the New York city soundscape, and many more went untagged. An open platform like freesound.org[21] has a few thousands. The yearly DCASE challenges[22] provide datasets for the classification of audio scenes and to develop automatic audio tagging. Besides birds, the featured applications are home surveillance, human-machine interfaces and urban planning. The common theme is the lack of reliably annotated audio. The construction of sounds databases is thus a legitimate research concern.

In a different direction, as big data and big computations take a hold on ecoacoustics, another important axis of development could be the search for low-tech approaches. Indeed, it might seem paradoxical to design conservation tools that inherently have a high environmental impact. How can we conceive the evolution of the discipline by taking into account future needs, future hardware and future energy availability? What are the indices with the most value for computational cost in ecoacoustics? And what would it take to get a low-consumption recording station like AudioMoth (Appendix 2) to operate solely from energy harvested from its environment? The urgency of developing such thoughts can only increase in the decade to come.

[21] Sound effects, stored at the Universitat Pompeu Fabro, Barcelona, Spain.
[22] http://dcase.community/challenge2019/.

Appendix A

More Machine Learning Concepts

This appendix first describes machine learning algorithms that are used in Chap 5 as subsidiaries or additional investigative techniques. All but one are unsupervised methods, i.e. methods that do not need to know which sample belongs to which class. Then, in a second section, we detail how some of the performance metrics used in machine learning competitions are computed.

A.1 Unsupervised Learning

Clustering

The k-means algorithm is a simple and popular clustering procedure. Its purpose is to split the samples in a dataset into a chosen number (k) of clusters. For the n^{th} cluster, the sum C_n of the distances between the samples in the cluster and the cluster's center is minimized (Eq. A.1). The k clusters are jointly optimized by moving samples from one cluster to another. In the final configuration, a sample belongs to the cluster whose center is the nearest. As for k-NN (Chap. 2, Fig. 2.2), the distance metric is central in k-means. The spherical k-means variant examines the angle rather than the Euclidian distance between feature vectors (Fig. A.1). The sum of the angles between the test samples and the cluster's center is then the cost function to minimize within a cluster (Eq. A.2).

With the Euclidian distance:

$$C_n = \sum_i ||\vec{u_i} - \vec{v_n}|| \qquad \text{(A.1)}$$

Or using angles between vectors:

$$C_n = \sum_i \angle(\vec{u_i}, \vec{v_n}) \qquad \text{(A.2)}$$

C_n is the cost function for cluster n (out of k clusters). u_i and v_n are the feature vectors respectively associated with the i^{th} sample in the cluster and the cluster's center.

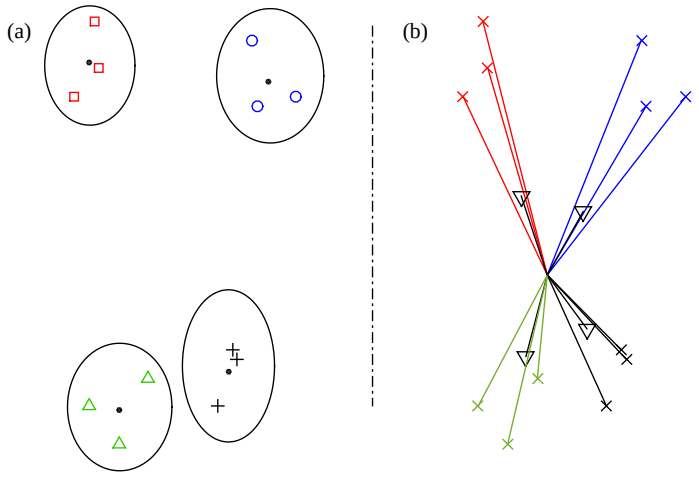


Figure A.1: A Visualization of k-Means

(a) k-means using the Euclidian distance. Four clusters (red squares, blue circles, green triangles, black crosses) with the cluster centers as black dots. (b) Spherical k-means, four clusters. The vectors for the cluster centers have a triangle tip. The angle between vectors is the distance metric used.

Dimensionality Reduction

Data represented by high-dimensional feature vectors is not easily comprehended unless some form of projection brings it back to a meaningful 2D or 3D visualization. The most desirable solution entails coordinates that push the classes apart from each other. In the new domain, clustering and classification would be greatly facilitated. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and t-distributed Stochastic Neighbor Embedding (t-SNE) work to that effect. PCA is a linear projection that maximizes the variance in the dataset; the directions that explain most of this variance are retained. LDA on the other hand maximizes the separation between the classes[1].

LDA is an established technique in biology (Rao [67]). In essence, it is a projection onto a modal base with a ranking of the most discriminant eigenvectors. The modal coordinates for the top two or three eigenvectors

[1] And is therefore a supervised method.

can be used to build a 2D or 3D scatter plot. Then, the plots can be enriched either with boundaries between the classes or confidence ellipses (see the illustrations in Section 5.2.1 in Chap. 5). The boundaries are equations of conic sections and are defined between pairs of species[2]. After regions are assigned to the different species, new samples can be classified according to the partitions. The confidence ellipses on the other hand materialize how well the species cluster. For a 95% confidence ellipse, 95% of the samples for a given species are inside the ellipse.

The recent t-SNE is a truly unsupervised scheme (van der Maaten & Hinton [86]). The 2D coordinates it produces are based solely on the features, not on the class information. The classes form separate clusters in the resulting scatter plot if and only if the class information is correctly reflected in the acoustic features (see the illustrations in Section 5.2.2 in Chap. 5). t-SNE is non-linear and the final 2D coordinates have no physical interpretation. Instead of working with distances between feature vectors, t-SNE replaces them with conditional probabilities between pairs of points both in the initial high dimensional space and in the destination low dimensional space. The probability that point A_j is similar to point A_i belongs to a distribution centered on A_i. It is high for points close to A_i and low at greater distances. A Gaussian distribution is used in the high-dimensional space and a Student-t distribution is used in the low-dimensional space. The variance of the distributions is chosen so as to keep 5–50 points in the bulk of the distribution close to A_i. This parameter is called the perplexity and is left to the preference of the user.

t-SNE optimizes the coordinates in the low-dimensional space so as to match the pairwise probabilities of both spaces. As the Student-t distribution has heavier tails, there is more margin to position points away from the distribution center. A modest distance in the high dimension can produce a greater distance in the low dimension. The dataset spreads out on the map instead of crowding central regions. This is how the t-SNE visualization typically offers more separation of the classes than LDA. Another advantage is that two dimensions are sufficient by design, whereas with LDA the number of directions that must be retained to satisfactorily separate the classes depends on the dataset. However, contrary to LDA, the coordinates produced by t-SNE cannot be re-used if the dataset is extended. For LDA, once the

[2]A conic section is either an ellipse, a hyperbola or a parabola. The cartesian form is $ax^2 + bxy + cy^2 + dx + ey + f = 0$.

projection is determined, it can be applied to any new data point. For t-SNE, new data demands a new optimization.

In his Matlab implementation of t-SNE[3], van der Maaten performs a PCA on the features before launching the optimization of the low-dimensional coordinates.

A.2 Performance Metrics

Area Under the Curve

The Area Under the Curve (AUC) quantifies the ability of a binary classifier, with values above 90% being excellent and values below 60% poor. In multi-class problems, the AUC is computed for all pairs of classes.

The curve in question is the Receiver Operating Characteristic (ROC). When presented with a test sample, a classifier returns a score between 0 and 1; above a threshold, the sample is class i and under the threshold, the sample is *not* class i. The false positives (FP) and true positives (TP) vary with the threshold: a parametric ROC curve can be drawn in the (FPR, TPR) map. Here the TP rate (TPR) is TP over the number of positives in the dataset and the FP rate (FPR) is FP over the number of negatives in the dataset. With a perfect classifier, $TPR = 1$ regardless of the FPR[4]. The AUC is the integral of the ROC with FPR varying from 0 to 1, i.e. the area under the AOC curve. The AUC of the perfect classifier is 1 or 100%. When the algorithm gives random results, the ROC curve is the diagonal and the AUC is 50%. Figure A.2 shows a few configurations for predictions and the resulting ROC.

Figure A.3 illustrates why the AUC gained traction in the machine learning community. It is insensitive to the composition of the dataset, whereas the accuracy is directly impacted when one of the classes is poorly predicted. However the accuracy retains the advantage of a more immediate physical meaning. It is the number of samples, either positive or negative, that were correctly predicted. The AUC on the other hand can be interpreted as the expected proportion of positives with scores higher than a random negative; a less suggestive picture.

[3] https://lvdmaaten.github.io/tsne/

[4] The perfect classifier gets all the actual positives right but might still generate false positives. The decision threshold only affects the negatives. Naturally, it should be set at a value that minimizes the FP. In this case, minimizing the FP does not lead to increasing the false negatives: the positive samples are always correctly detected.

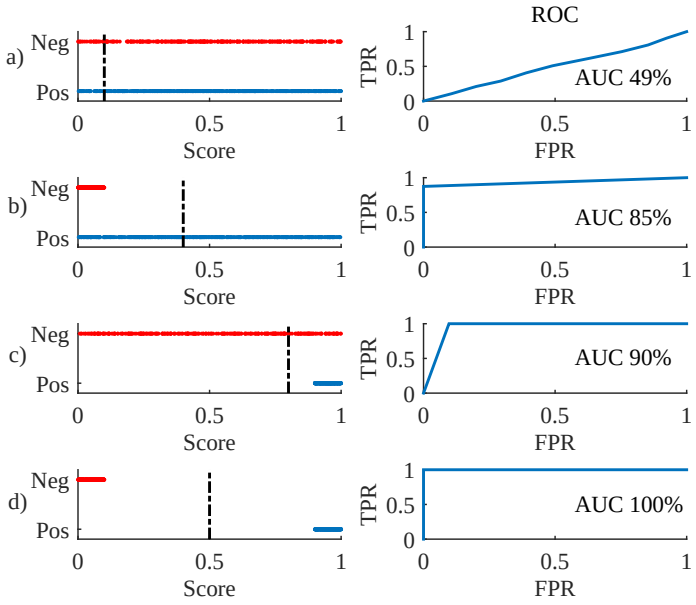


Figure A.2: Examples of Receiver Operating Characteristic Curves

The figure presents the ROC (plots on the right) in four cases a) through d). In the left-side plots, the red dots ("Neg" ordinate) show the distribution of negative samples and the blue dots ("Pos" ordinate) the distribution of positive samples, in function of the score returned by the classifier for these samples. a) The predictions are random. Both distributions spread over the full range of scores. If the threshold to consider a prediction positive is set at 0.1, then 90% of negatives are FP and 90% of positives are TP. The ROC, built from varying the threshold, follows the diagonal. b) The negatives are correctly predicted with low scores, the positives are random. With a threshold at 0.4, the FPR is 0% (there are no FP) and the TPR is 60%. The ROC sticks to the y-axis. c) The positives are correctly predicted with high scores, the negatives are random. The ROC sticks to the $y = 1$ line. d) The predictions are correct for both positives and negatives. The ROC follows the top-left path.

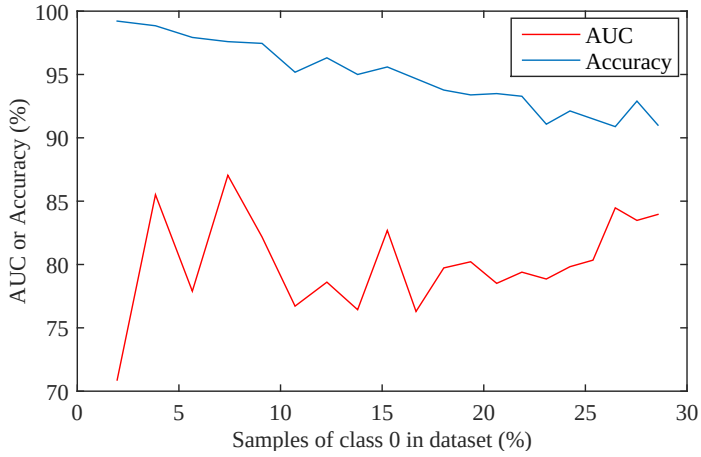


Figure A.3: AUC versus Accuracy

In a binary problem where the positives (class 1) are well predicted (mean score 0.9, standard deviation 0.05) and the negatives (class 0) poorly predicted (mean 0.1, standard deviation 0.9), the accuracy decreases with an increasing proportion of negatives in the dataset. When the dataset contains 100% of positives, the accuracy is 100%; when the dataset is split 70% positives / 30% negatives, the accuracy falls to 90%. Meanwhile, the AUC remains "stable" around 80%.

In the present thesis, by default, we used the accuracy. We complemented it with the accuracy per species, in order to not let a well-performing class hide more problematic results.

Mean Average Precision (mAP)

The *Precision* is defined as $\frac{TP}{TP+FP}$. This is the number of correct predictions amongst the positives. The *Recall* is $\frac{TP}{TP+FN}$. This is identical to the TPR, i.e. the number of correct positives over the number of predictions that should have been positive. A high precision is typically associated with a low recall: few errors are made on the positive predictions because only the most confident predictions were made. On the contrary, when the recall is high, many samples are proposed as positive, and many are wrong, hence a low precision.

If the predictions made by a model are ranked from the most confident (highest probability) to the least, precision and recall can be evaluated for

the first prediction, then the first two, then the first three, etc. The precision will gradually decrease and the recall increase. One can tabulate, for a range of recall values (e.g. from 0 to 1 in steps of 0.1), the highest precision reached under each recall value. The mAP averages these precision values.

Mean Reciprocal Rank (MRR)

The MRR is used with algorithms that return a list of possible identifications, ranked from the most probable to the least. For a given test sample, if the correct answer comes in j^{th} position, the reciprocal rank is $\frac{1}{j}$. The MRR is the mean of the reciprocal ranks for all test samples. At 100%, the correct answer is always in first position, at 50% it is often in second position, etc.

Appendix B

Autonomous Recording Station

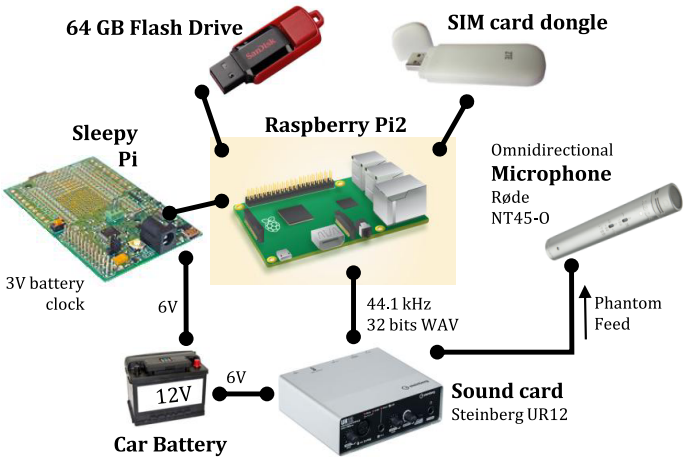


Figure B.1: Station Parts and Connectivity

B.1 Components and General Operation

The architecture of our autonomous recording station is shown in Fig. B.1 (Florentin et al. [22]). The station is built around a Raspberry Pi 2 mini-computer[1]. The Pi runs a Linux OS, has a quad-core processor, 40 GPIO pins and four USB ports, but no sound card. Auxiliaries connected through USB comprise an external sound card (and microphone), a dongle holding a SIM card and a flash disk for data storage. The GPIO pins are used to connect

[1]The Raspberry Pi Foundation, Cambridge, United Kingdom.

a Sleepy Pi 2[2], a power board fitted with an Arduino microcontroller and a real-time clock (RTC). It allows scheduling the Pi's activity and controlling the power supply. The battery connects to the Sleepy Pi which redistributes power as suitable. Only the sound card is powered independently.

The batteries are set inside a plastic container and elevated on wood planks to limit exposure to potential water infiltrations. Small holes were drilled under the handles of the container for the ventilation of hydrogen emissions. All electronics are kept dry in a sealed plastic box that is meant to be attached to a tree (Fig. B.2). In winter, their own heat maintains the box above freezing temperatures.

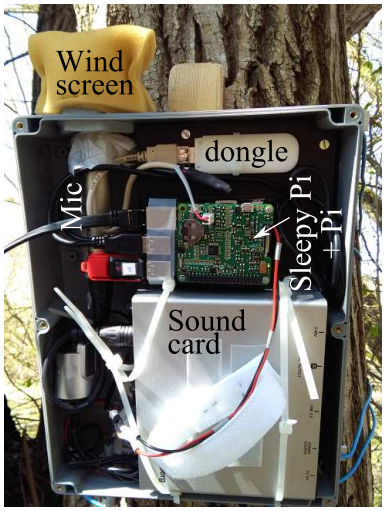


Figure B.2: The Electronics Box in Remerschen

The station tasks are automated through three levels of scripting. There is first an Arduino script for the Sleepy Pi which defines wake-up and sleep times for the Pi, reads battery voltage and powers the SIM-card dongle when needed. The Pi's activity is directed by a shell script that launches at power-up. This script organizes the station's work day and operates the auxiliaries, i.e. it starts or stops audio recording, moves files around and sends SMS. Finally, GNU Octave[3] scripts perform the signal processing tasks, which consists primarily in calculating the Acoustic Complexity Index (ACI)[4] and deciding which recordings to keep. The shell script has self-checks to recover from a number of failures and the Sleepy Pi controls at intervals that the Pi is still running when it should.

The Pi is turned on shortly before dawn (5:00) and turned off at dusk (21:00)[5,6]. It records continuously throughout the day. SMS are sent twice daily, at 9:00 and at 21:00, with information about the battery and the recorded

[2]Spell Foundry, Wokingham, United Kingdom.

[3]GNU Octave, version 3.8.2 (2014), https://www.gnu.org/software/octave/.

[4]See Chap. 2 for a definition, Chap. 4 for usage in the present context.

[5]The RTC clock is programmed to follow the daylight saving time. Thus recording starts earlier, in absolute time, from the end of March on.

[6]At 21:00, the Pi finishes its current tasks and shuts down. The Sleepy Pi discontinues the current supply to the Pi at 21:15.

data (see Section B.4). Sound is recorded in mono, at a sampling frequency of 44.1 kHz, which is the lowest the Steinberg sound card allows. For reference, all analyses in the present thesis use 12 kHz. The main frequency content of woodpecker drumming rolls generally lies below 1500 Hz (Chap. 5) and the first harmonics of woodpecker calls below 3000 Hz (Chap. 6). Recording proceeds without interruption on one of the cores; every ten minutes, the other three post-process the accumulated data in Octave. Data is stored in chunks of 30 seconds using 32-bit WAV encoding. At 21:00, the station stops recording and sends another status SMS. The Sleepy Pi shuts the Pi down for the night at 21:15.

B.2 Microphone

An omnidirectional microphone (Røde NT55-O) is used to record ambient sound around the station. This is the choice of most authors recording birdsong (see Table B.2), although Bardeli et al. [4] suggest that four cardioid microphones in a cross are a better option for detection. The head of the microphone is tilted so that it is ideally positioned to capture sounds at 7-15 meters of height, in a group of trees 3-5 meters apart. These considerations are superfluous in theory for omnidirectional microphones, but we anticipated a deterioration of performance at higher frequencies and because of the harsh conditions the microphone would be subjected to. The manufacturer indicates perfect omnidirectionality up to 4000 Hz (Fig. B.3).

The frequency response in Fig. B.3 is on par with the Primo EM172 capsule favored by many designs (see the section on other recording stations, in particular Table B.2). The sensitivity of the Røde microphone is 10 dB lower. Recently, the cheap MEMS microphones have been popular in autonomous recording stations (Bartalucci et al. [5]). These are less power-hungry, but their durability is questioned. The Røde microphone still captures excellent-quality sound after three springs outdoors. The Steinberg sound card seems to have suffered in its last deployment.

The microphone is mounted with its body inside the electronics box (Fig. B.2) to keep it warm. Inevitably, there is a blind spot for direct acoustic waves coming from the other side of the supporting tree. The microphone head is covered with a latex film cut from household gloves, then clamped into a cable gland to seal the set-up. In the lab, we measured the frequency

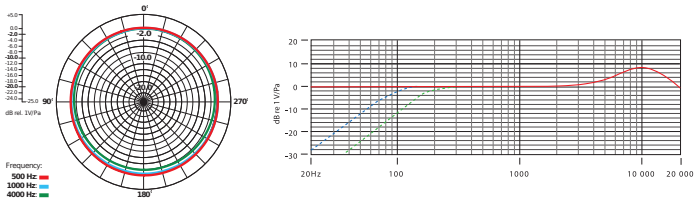


Figure B.3: Røde NT55-O Polar Pattern and Frequency Response

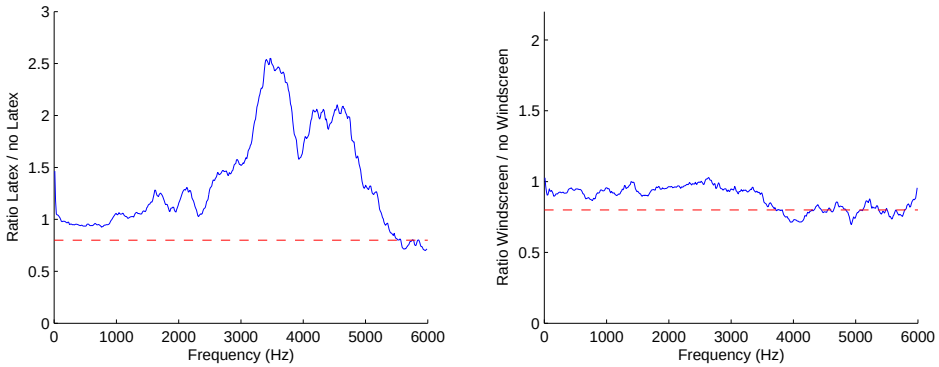


Figure B.4: Lab Tests on Microphone Response

response to white noise for two varieties of latex[7]. Figure Fig. B.4 (left) shows the ratio between the spectra with and without latex, for the type of latex we ultimately selected. We deem acceptable that the latex increases the signal amplitude, i.e. that the ratio is greater than 1, and that the latex does not decrease the signal amplitude excessively, i.e. that the ratio remains greater than 0.8 (dashed red line). From 1000 Hz on, the signal is often enhanced by the latex, likely because of the modes of the latex membrane or of the cavity between the latex membrane and the microphone head. Drums below 1500 Hz should be fairly rendered, while we can expect passerine songs above 3000 Hz to be enhanced. In any case, the temporal and frequency parameters of songs should not be affected; only their amplitude. Finally, a windscreen was shaped up from surplus foam. Again, the impact on the microphone response was tested and found satisfactory (Fig. B.4, right). We did not encounter issues of wildlife eating up the foam or perching on the microphone.

[7]The room was not anechoic. The source and microphone were in fixed positions.

B.3 Power board

The Arduino code in the Sleepy Pi 2 loops continuously through 3-min cycles during the day and 10-min cycles during the night. At each iteration it checks whether the Pi is running and takes action if it needs to be restarted or shut down. Then, in the daytime, it monitors battery voltage and passes the information to the Pi using the serial GPIO port. In the nighttime, it enters a deep sleep, low-consumption mode. Right before 9:00 and 21:00, it starts supplying the SIM card dongle with a separate 5 V. It turns it off half an hour later.

The Sleepy Pi 2 is a fragile piece of equipment; no current above 2 A can be drawn through the board. This is the primary reason why the sound card has a parallel current supply. Luckily, the sound card turns off automatically when the Pi is down. Both the Sleepy Pi 2 and the RTC endured long deployments irregularly. The RTC, powered by a 3 V lithium button cell, drifted through the recording season by no more than a couple of minutes.

B.4 Communication in the Field

In the field, the Pi can be connected to a laptop using a network cable. Then the Pi's console is accessed using a terminal emulator such as PuTTY[8]. To that end, the IP address of the Pi must first have been imposed so that it fits the IP address of the laptop.

The station sends SMS using a program called gammu[9]. The texts contain information about disk space usage (Disk: 61%), percentage of audio stored to disk on the previous day (Wav: 36%), number of audio files saved since campaign start (FileID: 14205) and battery voltage (Fig. B.5). These SMS, through their non-delivery or through erratic metrics, are essential indicators of station malfunction. However, they bring in ad-

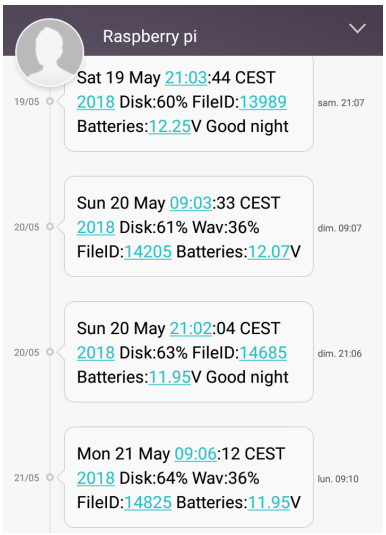


Figure B.5: SMS Sent by the Station

[8]https://www.putty.org/.
[9]https://wammu.eu/gammu/

ditional difficulties. Communication through Wi-Fi or the cellular phone network is known to be power hungry; this is the reason why such a functionality is not implemented in commercial boxes. We address this by letting the Sleepy Pi 2 disconnect the dongle when not in use. Then, the cellular network is not entirely reliable, especially in remote areas or across operators. Finally, the cellular waves interfere with the sound card electronics. We placed the dongle as far apart from the sound card as possible.

At increased costs in terms of complexity and energy, full remote control of the Pi can be set-up through the cell phone network. The benefit would lie in the debugging phase. Log files could be retrieved and updated scripts remotely uploaded.

B.5 On-the-Spot Processing

The station calculates and saves the ACI spectrum for each successive 30-second segment of audio. Segments for which the maximum ACI in the target bandwidth exceeds a threshold are stored to disk. The thresholds, and the value of using the ACI to detect woodpeckers in general, are discussed in detail in Chap. 4; Table 3.6 in Chap. 3 demonstrates the benefits of this approach. Between 50–80% of the data does not need to be saved. This diminishes the disk space requirements and future time spent on data analysis. To date, the onboard ACI calculation of our recording station is unique. The makers of AudioMoth (Hill et al. [35]) have contemplated programming the set of indices from Towsey et al. [81] in future releases.

In the morning the station reviews the ACI spectra from the previous day and generates an ACI spectrogram image (Fig. B.6). This gives a daily overview of recorded events, as a host of bird songs and most notably woodpecker drums produce remarkable patterns in ACI images. Blanks appear when the station was not operating properly or shut down, e.g. at around 13:00 in Fig. B.6. Horizontal lines indicate interference with the cell waves (2016) or continuous electronics noise (2018).

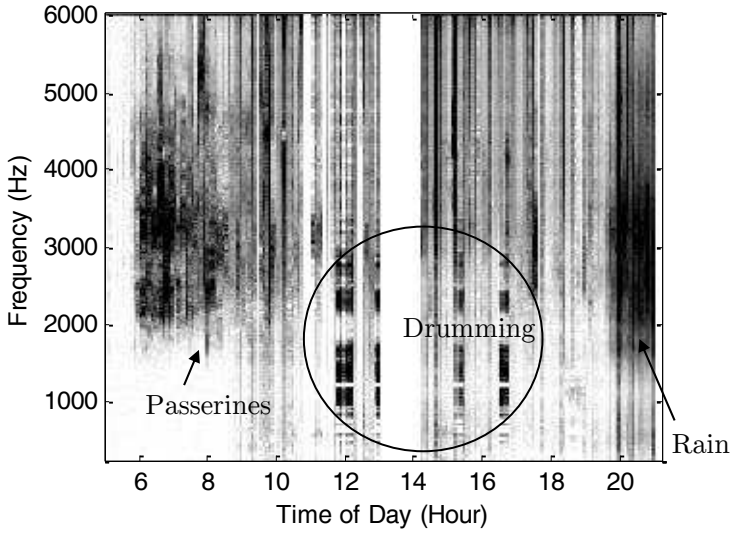


Figure B.6: ACI Spectrogram for March 18th, 2016

B.6 Batteries

The station needs a minimum of one 12 V battery to operate. We used either one 60 Ah car battery (i.e. starting lead-acid) or two in series. Aide et al. [3] and Whytock & Christie [89] (battery information from Hill et al. [35]) with Solo, another Raspberry Pi-based design, used similar equipment. The Sleepy Pi, the Pi and the sound card all expect 5 V. The Sleepy Pi accepts 5.5–30 V and can be wired directly to the battery. The sound card is connected through a smartphone car charger socket and plug, which ensures the conversion to 5 V.

Car batteries are not intended for full discharges. The voltage should not drop below 11.9 V to avoid irreversible damage. A slow charge also helps with subsequently sustaining a suitable voltage for a longer time period. Deep cycle batteries would allow lowering the damage threshold by another volt or two, but the additional station time is not worth the increase in price. Solar panels could extend the battery cycle but not make the station fully self-sufficient.

The great benefit of the car battery is its capacity. For comparison, a high quality, not rechargeable AA alkaline battery offers 2.6 Ah. The capacity of lithium-ion batteries, widespread in consumer electronics, has recently improved but they remain expensive and perform poorly in cold weather.

The 12 V 60 Ah battery weights roughly 15 kg and lasts about a week in the field (100 hours). This is far less efficient than the commercial solutions in Table B.2. At present time, the battery replacement cycles are not well-suited to remote locations. The station is also not easily displaced because of the battery weight.

B.7 Power Consumption

The first steps toward reducing power consumption were letting the station sleep through the night and powering the SIM-card dongle off when not in use. The next one was the reduction of intensive computations, namely audio post-processing. The Octave scripts were rewritten to execute faster. The downsampling of audio recordings to 12 kHz was abandoned, as data storage was not an issue. On the contrary, the live calculation of the ACI has key benefits and was maintained.

Table B.1 shows the proportions of power used by different features and parts in the final version of the station. Energy consumption is driven primarily by the Pi/sound card/microphone block. The Raspberry Pi is confirmed as power-hungry, even when idle; this is the cost of running the Linux OS. It gets worse in newer versions: the power draw with four cores running is 4.6 W for the Pi 2 and 8.1 W for the Pi 3. Solo (Whytock & Christie [89]) uses the old Raspberry A+ for this reason. Finally, we have no reading on how energy-efficient our sound card/microphone combination is. In any case, low consumption is not a design target for music studio equipment.

Table B.1: Power Consumption of Main Devices

Equipment	**Operating Time**	**Power Consumption**
SIM card dongle	2*30 min	0.2%
Raspberry Pi 2 + sound card + microphone	continuous	94.0%
Sound processing using Octave on 3 cores	60 sec every 10 min	5.8%

B.8 Distance Tests

Few attempts have been made to date[10] to estimate the reach of recording stations, i.e. the *r* parameter in Fig. 3.3 (Chap. 3). Yet this parameter is essential knowledge for the planning of large scale audio surveys. The question is complex; in addition to the distance factor, the proper detection of a species depends on the strength of the original signal, on the amount of absorption through the environment and on the capacity of the post-processing scripts to deal with a diminished signal. Nevertheless, orders of magnitude would advance the subject.

We thus staged an outdoor experiment in which we replayed calls and drums recorded at Tenneville (visible in Fig. B.7) using an i-Pod and a Bluedio BS-3 loudspeaker (frequency response 20-20000 Hz, output power 2*5 W). The sounds were replayed at varying distances from the station, in steps of 10 m up to 100 m. This was done twice, in Remerschen and in La Petite Raon, on the access roads that led to the station. In Remerschen, this was a private dirt road and in La Petite Raon, a forest road. Sadly, in Remerschen, the summer vegetation at ground level blocked almost all signals from reaching the station. In La Petite Raon, the vegetation is persistent and the undergrowth limited. The reverberation time remains cathedral-like throughout the year. We thus could test at the end of May in acoustic conditions not too dissimilar from March-April. Still, the setup of our experiment ensured that the direct acoustic waves followed the road and did not have to travel through dense vegetation.

All signals up to 100 m were correctly re-recorded by the station. On the spectrograms one can see the calls lose their harmonics as they move further away (Fig. B.7). Fig. B.8 shows the calls that passed the ACI selection threshold. This data put the reach of the station in the range 50–70 m. Note that the recordings were also tainted by shock noises, most significantly for the higher distances. We removed most of them from the data.

Then, from the recordings of the call on the left of Fig. B.7, we extracted 90 ms windows encompassing the 13 syllables and calculated the acoustic pressure spectrum for each. We then measured the level of the maximum peak[11]. Fig. B.9 shows the resulting average level over all syllables against

[10] None...?

[11] Probably not the most sensible choice due to the risk of leakage, yet the results seem reasonable.

distance from the station. All dB levels are uncalibrated. The theory assumes that acoustic pressure is proportional to the inverse of the distance (spherical spread). This model seems satisfying for most points. The decay at 10 m is surprisingly sharp; possibly a near-field effect. Further on, the point at 50 m is erratic and the station might not recapture the sounds over 70 m well enough.

We also compared the re-records to the original from Tenneville. If the Tenneville *P. canus* was 1.45 m away from the microphone, then the levels from the LPR recordings would match well (22 dB in average over the syllables at 1 m). But the branches of the tree where the station was attached in Tenneville were much higher up. A more accurate estimate would be 3–7 m. In truth, the Bluedio BS-3 does not generate as much acoustic power as the actual bird. The levels, also for drums, are 10–20 dB too low. However we can use the comparison with the original recording to recalibrate the distances measured in LPR. Assuming the Tenneville bird was e.g. 5 m from the station when singing, we can calculate its acoustic power using the inverse distance law. Then we can estimate at what various distances the acoustic pressure levels measured in LPR correspond. This is the principle behind Fig. B.10. The data at 50 m and for distances greater than 70 m is not shown, as the results were unrealistically high and we already have evidence that these distances might be beyond the station's reach. The curve corresponding to a bird at 1.45 m roughly follows the bisectrix (the $y = x$ line), i.e. this is the case for which the distances measured at LPR are valid. For this case, recaptures with the ACI and deviations from the acoustic propagation law in Fig. B.9 indicate that the station's reach is roughly 60 m. However, because the Tenneville bird was at greater distances from the microphone, Fig. B.10 tells us that this boundary should be amended to at least 135 m (bird at 3 m from the microphone). It would be excessive at this time to contemplate the higher distances, as our experimental conditions were favorable in terms of acoustic propagation. 135 m still encompasses most datapoints in Fig. B.10.

There are two conclusions to the experiment. The first is that it seems justified to consider that the recording station has a reach comparable to the human ear. The second is an evidence: the environment strongly impacts this reach.

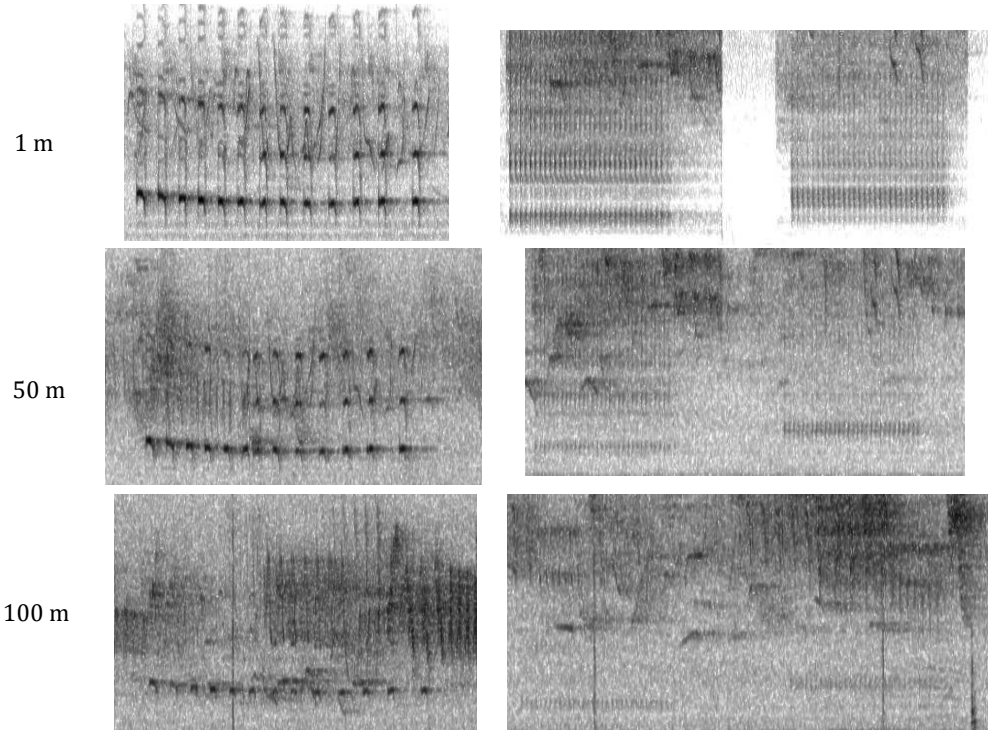


Figure B.7: Spectrograms of Recorded Signals at Various Distances

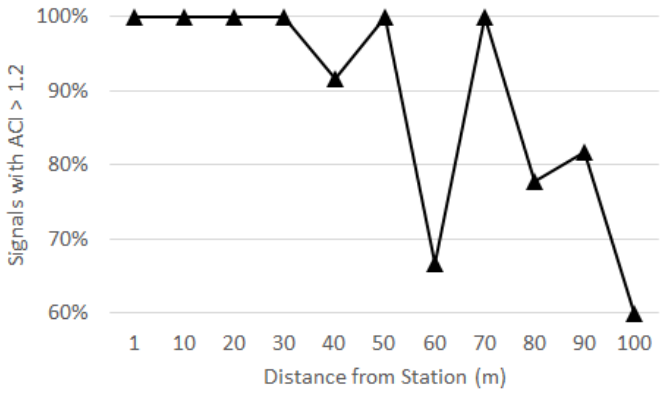


Figure B.8: Signals Detected by the ACI

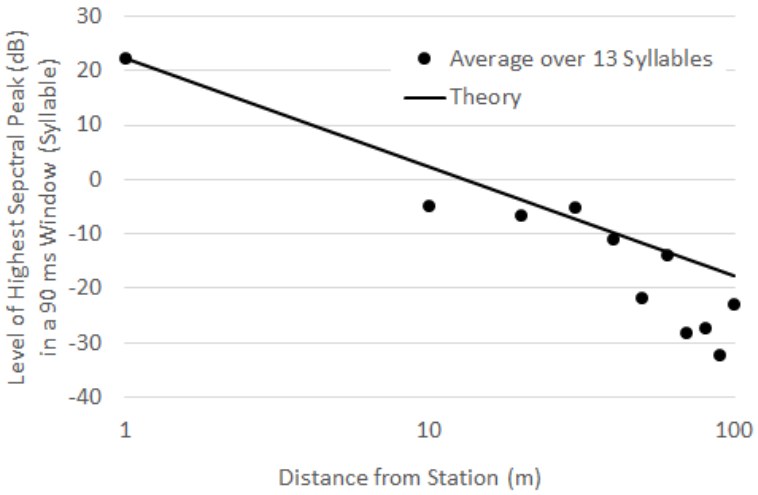


Figure B.9: Acoustic Pressure Amplitude Decay with Distance

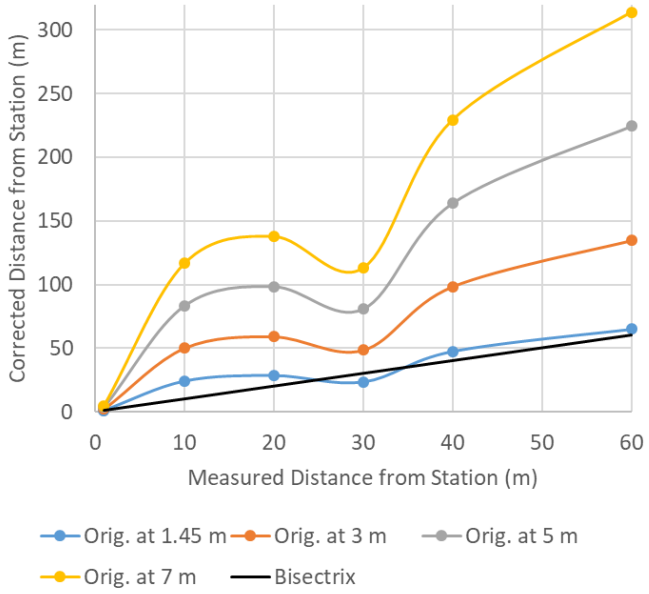


Figure B.10: Distances Corrected for Actual Bird Position

B.9 Other Recording Stations

Table B.2 lists a number of other recording stations. Their photographs are shown in Fig. B.11. Some are custom-made, including other designs revolving around the Raspberry Pi, and others are commercial products. Most have options to program a recording schedule. These designs primarily target bird vocalizations. A few others exist for bats[12].

For comparison, running our station's schedule (16 hour per day) and sampling at 48 kHz, AudioMoth would last approximately 66 days. However, using AudioMoth at the best of its ability requires buffering data and saving sporadically to disk. Indeed, AudioMoth is conceived on the premise that not all data will be saved. The target sounds are detected through an onboard real-time acoustic analysis performed by the microcontroller. The purpose is to avoid large datasets and energy-intensive writes to the SD card. Thus, Hill et al. [35] programmed a frequency-band selection for cicadas and a Hidden Markov Model for gunshots. As our review shows, the bird detection problem is not that clear cut. For woodpecker calls, there are three levels of audio processing: preselection through the ACI, production of images and identification through a neural network (see Chap. 6). AudioMoth cannot run a deep neural network such as the ones in Chap. 2, Section 2.5. The ACI is our realistic real-time pre-processing scheme.

This being said, with the RAM of AudioMoth amounting to 256 kB, we would not be able to buffer 30 second segments. The files we recorded at 12 kHz in Tenneville weighted 700 kB. In addition, it would be necessary to split the RAM between the ACI calculation and the ongoing recording. The duration of sound segments could not exceed 5 seconds and this is not adapted to woodpecker vocalizations. Therefore all segments would have to be saved to disk. In that case, the energy consumption of AudioMoth more than triples and the projected autonomy falls to 20 days.

Another difficulty of AudioMoth is the limited data storage. The WAV files are saved onto a 32 GB SD card. At 5 GB a day on our schedule without the ACI preselection, the card must be replaced every sixth day. By sampling at 12 kHz, this could be brought down to 1.4 GB per day. Then the

[12]The SM4BAT (Wildlife Acoustics), the Anabat Swift (Titley Scientific) or Aurita, an extention of the Solo design. AudioMoth has also been used for bats (Hill et al. [35]), although the MEMS microphone (100 Hz – 10 kHz) is ill-suited for this application. The ultrasonic microphone in the SM4BAT retains a high performance until at least 120 kHz. For reference, *Rhinolophus hipposideros* echolocates at 112 kHz.

autonomy of AudioMoth would be 24 days. This is not yet in the range of the Wildlife Acoustics SM. Recording 16 hours a day, it lasts approximately 25 days, supposedly at a higher sampling frequency. There is no onboard pre-processing, but also no limitation on data storage.

Figure B.11: Autonomous Recording Stations

Table B.2: Specifications of Various Autonomous Recording Systems

Design	**Core Equipment**	**Microphone**	**Batteries**
ARUPi[a] (A. Turner, 2015) 150£	Raspberry Pi A+ Sleepy Pi 1 USB sound card (48 kHz 32-bit mono)	Primo EM172[g]	8x AA 20 Ah[i] 11h recording
Solo[b] (J. Christie, 2018) 120£	Raspberry Pi A+ Witty Pi 2 USB sound card (48 kHz 16-bit mono)	Primo EM172	USB battery bank 24 Ah 5days 24h/24
soundCamp[c] (G. Smith, 2016) Price unknown	Raspberry Pi B+ Custom power board Cirrus Logic Audio Card (44.1 kHz 16-bit stereo)	Primo EM172	Unspecified
AudioMoth[d] (Open Acoustic Devices, ongoing) $49–$79	EFM32 Gecko processor Custom parts 8–384 kHz On-board processing	MEMS[h]	3x Li AA-cell[i] 10 mAh/hour, 44 days at 48 kHz
SM[e] (Wildlife Acoustics, ongoing) $1231	Commercial box Sampling 8–96 kHz	Primo EM172	4x D-cell Alkaline 400 hours
Bioacoustic Audio Recorder[f] (FrontierLabs, ongoing) €560	Commercial box Sampling 8–96 kHz	Primo EM172	4x Li AA-cell 14 Ah[i] 320 hours

[a] https://www.instructables.com/id/ARUPi-A-Low-Cost-Automated-Recording-Unit-for-Soun/.

[b] https://solo-system.github.io/home.html, [89].

[c] http://soundtent.org/. This station is designed to stream live from the wild at http://locusonus.org/locustream/.

[d] https://www.openacousticdevices.info/, [35].

[e] https://www.wildlifeacoustics.com/.

[f] http://www.frontierlabs.com.au/.

[g] 60 Hz - 10 kHz. Omnidirectional.

[h] 100 Hz - 10 kHz. Omnidirectional.

[i] We consider, in average, 2.5 Ah for an AA alkaline battery and 3.5 Ah for a lithium AA-cell battery.

B.10 Conclusions

The SM boxes of Wildlife Acoustics have become a popular solution to record the wild. Compact, easy to use, best-in-class for autonomy, they fit the bill for a wide array of projects, especially if the scope is restricted to a few minutes of recordings every hour or every day. Our ambition was to record continuously 16 hours a day for an entire spring. No station could have handled the battery issue gracefully. In addition, this setup implied an excessively large amount of recordings, which would have been another potential dead end save for onboard processing, i.e. the capacity to shed large portions of the recordings using the ACI. At the time of our field campaigns (and to our knowledge, still today), this was only possible through a custom design. Seeing how remote sensors and machine learning are moving forward, there is little doubt that Wildlife Acoustics and Frontier Labs will consider the issue at some point. Meanwhile, we found the open design AudioMoth worth keeping an eye on.

Appendix C

Encoding Species Identity in a Phylogenetically Constrained Signal: Woodpeckers' Drumming

This appendix reproduces text and illustrations from Garcia et al. [28] with authorization from the author. The footnotes in the figures are our addition.

Introduction

Drumming (defined as a repetitive striking of the bill on a substrate) is a rhythmic signal only found in the woodpecker family, typically occurring during the breeding season. Despite overall diversity, woodpecker's drumming appears to be constrained mechanically and well conserved within genera. To which extent can thus drumming encode species-specific information in reproductive contexts?

Methods

- Phylogenetic mapping of acoustic characters
- Investigation of drumming acoustic spaces
- Field playbacks on *D. major*:
  - Conspecific VS. hetero-specific signals (natural)
  - Acoustically modified signals (resynthesis)

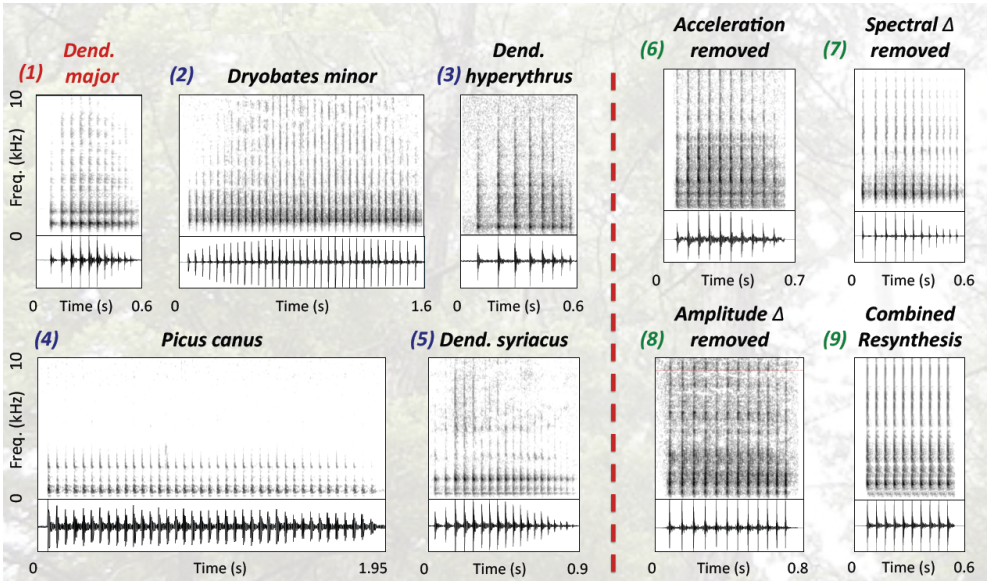


Figure C.1: Signals Used in the Playback Experiments

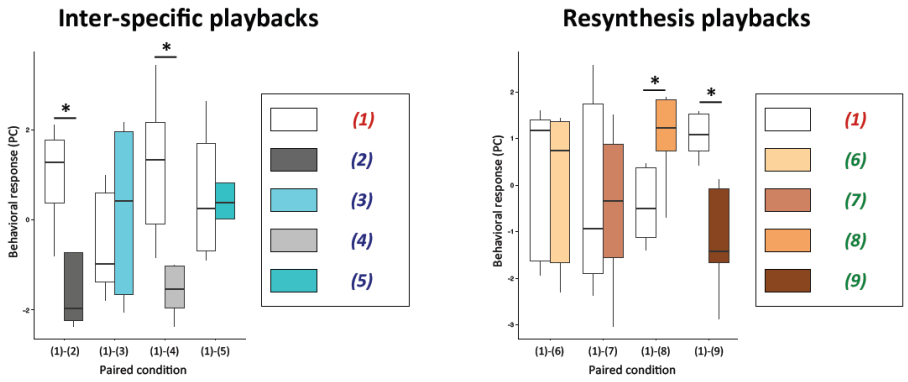


Figure C.2: Behavioral Response in Playback Experiments

Stars mark a significant difference in the reaction.

Results

Tested individuals of *D. major*:

- Respond to similar drumming patterns and NOT to different drum-

ming patterns.

- Require multiple acoustic degradations to stop responding to their species-specific signal.

Discussion

Woodpecker's drumming bears a strong phylogenetic signal. Limited species-specific recognition when drums occupy a similar acoustic space (determined phylogenetically?). Integration of species-specific information relies simultaneously on multiple parameters. Decoupled mechanisms: drum production - drum perception.

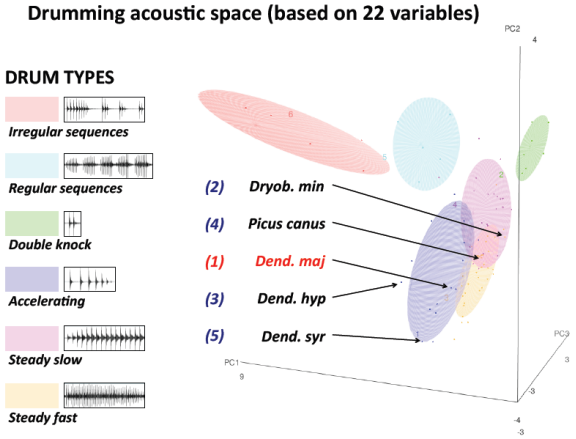


Figure C.3: Drumming Acoustic Space

The authors collected three drums per species and calculated 22 parameters on each drum, including the ones reported in Florentin et al. [21]. The three main directions of a PCA were used for the plot. Six broad categories of drums were deduced from the results. With the six groups marked on the phylogenetic tree (next figure), one can contemplate the link between drumming style and genetics.

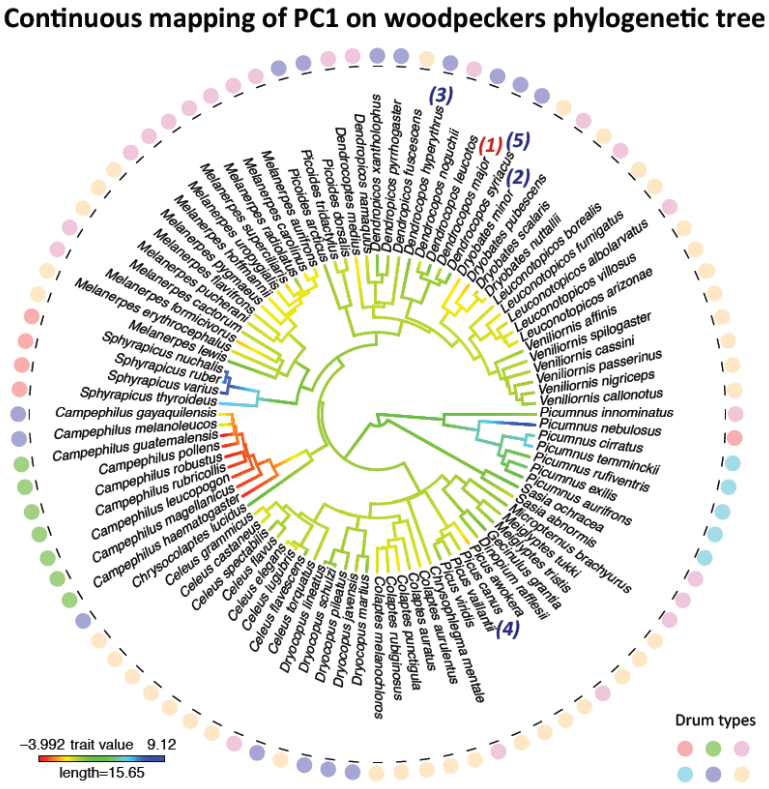


Figure C.4: Phylogenetic Tree

Bibliography

[1] 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, 1729–1733. IEEE.

[2] Acevedo M.A., Corrada-Bravo C.J., Corrada-Bravo H., Villanueva-Rivera L.J., Aide T.M. (2009) Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics, 4(4):206–214.

[3] 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.

[4] Bardeli R., Wolff D., Kurth F., Koch M., Tauchert K.H., Frommolt K.H. (2010) Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognition Letters, 31:1524-1534.

[5] Bartalucci C., Borchi F., Carfagni M., Furferi R., Governi L. (2017) Design of a prototype of a smart noise monitoring system. XXIV International Congress on Sound and Vibration (ICSV). London, UK.

[6] Blume D., Tiefenbach J. (1997) Die Buntspechte (Gattung Picoides). Die Neuhe Brehme-Bücherei. Westarp Wissenschaften. Magdeburg.

[7] Blumstein D.T., Mennill D.J., Clemins P., Girod L., Yao K., Particelli G., Deppe J.L., Krakauer A.H., Clark C., Cortopassi K.A., Hanser S.F., McCowan B., Ali A.M., Kirschel A.N.G. (2011) Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. Journal of Applied Ecology, 48:758–767.

[8] Brandes T.S. (2008) Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Transactions on Audio, Speech, and Language Processing.

[9] Burivalova Z., Towsey M., Boucher T., Truskinger A., Apelis C., Roe P., Game E. T. (2018) Using soundscapes to detect variable degrees of human influence on tropical forests in Papua New Guinea. Conservation biology, 32(1):205–215.

[10] Çakir E., Adavanne S., Parascandolo G., Drossos K., Virtanen T. (2017) Convolutional recurrent neural networks for bird audio detection. In Signal Processing Conference (EUSIPCO), 2017 25th European, 1744–1748. IEEE.

[11] Catchpole C.K., Slater P.J.B. (2008) Bird song: Biological themes and variations, 2nd edition. Cambridge University Press.

[12] Connor E.F., Li Shidong and Li Steven (2012) Automating identification of avian vocalizations using time-frequency information extracted from the Gabor transform. Journal of the Acoustical Society of America, 132(1):507-517.

[13] Dagnelie P. (1975) Analyse statistique à plusieurs variables. Presses Agronomiques de Gembloux.

[14] Dieleman S. et al (2015) Lasagne: First release.

[15] Dodenhoff D.J., Stark R.S., Johnson E.V. (2001) Do woodpecker drums encode information for species recognition? The Condor, 103(1):143-150.

[16] Dong X., Towsey M., Truskinger A., Cottman-Fields M., Zhang J., Roe P. (2015) Similarity-based birdcall retrieval from environmental audio. Ecological Informatics, 29:66–76.

[17] Dupont S., Ravet T., Picard-Limpens C., Frisson C. (2013) Nonlinear dimensionality reduction approaches applied to music and textural sounds. Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME). San Jose, USA.

[18] Eyben F., Wllmer M., Schuller B. (2010) Opensmile: The Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia, 1459–1462. ACM.

[19] Fagerlund S. (2007). Bird species recognition using support vector machines. EURASIP Journal on Applied Signal Processing, 2007(1):64–64.

[20] Farina A., Lattanzi E., Malavasi R., Pieretti N., Piccioli L. (2011) Avian soundscapes and cognitive landscapes: Theory, application and ecological perspectives. Landscape Ecology, 26:1257–1267.

[21] Florentin J., Dutoit T., Verlinden O. (2016) Identification of European Woodpecker Species in Audio Recordings from their Drumming Rolls. Ecological Informatics, 35:61-70.

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

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

[24] Foote J., Cooper M., Nam U. (2002) Audio retrieval by rhythmic similarity. Proceedings of the 3rd International Conference on Music Information Retrieval. Paris, France.

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

[26] Frommolt K.H., Tauchert K.H. (2014) Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. Ecological Informatics, 21:4–12.

[27] Fuchs J., Pons J.M. (2015) A new classification of the Pied Woodpeckers assemblage (Dendropicini, Picidae) based on a comprehensive multi-locus phylogeny. Molecular Phylogenetics and Evolution, 88:28–37.

[28] Garcia M., Sèbe F., Marin-Cudraz T., Mathevon N. (2017) Species recognition in Great Spotted Woodpeckers (Dendrocopos major): A multilevel perception system. XXVI International Bioacoustics Congress. Haridwar, India. [Poster]

[29] Giret N., Roy P., Albert A., Pachet F., Kreutzer M., Bovet D. (2011) Finding good acoustic features for parrot vocalizations: The feature generation approach. The Journal of the Acoustical Society of America, 129(2):1089–1099.

[30] Glotin H., Clark C., LeCun Y., Dugan P., Halkias X., Sueur J. (2013) Proceedings of the first international workshop on Machine Learning for Bioacoustics, joint to the 30^{th} International Conference on Machine Learning (ICML 2013). Atlanta, USA.

[31] Gorman G. (2014) Woodpeckers of the world; the complete guide, Bloomsbury Publishing.

[32] Gould S.J., Vrba E.S. (1982) Exaptation : A missing term in the science of form. Paleobiology, 8(1):4–15.

[33] Grill T., Schlüter J. (2017) Two convolutional neural networks for bird detection in audio signals. 25th European Signal Processing Conference (EUSIPCO). Kos, Greece.

[34] He K., Zhang X., Ren S., Sun J. (2016) Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778.

[35] Hill A.P., Prince P., Pia Covarrubias E., Doncaster C.P., Snaddon J.L., Rogers A. (2018) AudioMoth: Evaluation of a smart open acoustic device for monitoring biodiversity and the environment. Methods in Ecology and Evolution, 9(5):1199–1211.

[36] Huang G., Liu Z., Van Der Maaten L., Weinberger K.Q. (2017) Densely connected convolutional networks. In CVPR, 1(2):3.

[37] Jahn O., Mporas I., Potamitis I., Kotinas I., Tsimpouris C., Dimitrou V., Kocsis O., Riede K., Fakotakis N. (2013) The AmiBio Project – automating the acoustic monitoring of biodiversity. 24th International Bioacoustics Congress (IBAC). Pirenopolis, Brazil [Poster].

[38] 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, 247–266. Springer, Cham.

[39] Kasten E.P., Gage S.H., Fox J., Joo W. (2012) The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. Ecological Informatics, 12:50–67.

[40] Kirschel A. (2017) Differential patterns of song similarity in Tinkerbirds leads to contrasting interactions among contact zones. XXVI International Bioacoustics Congress. Haridwar, India. [Abstract]

[41] Kogan J.A., Margoliash D. (1998) Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. Journal of the Acoustical Society of America, 103(4):2185–2196.

[42] Kong Q., Xu Y., Plumbley M.D. (2017) Joint detection and classification convolutional neural network on weakly labelled bird audio detection. In Signal Processing Conference (EUSIPCO), 2017 25th European, 1749–1753. IEEE.

[43] Krause B.L. (1993) The niche hypothesis: A virtual symphony of animal sounds, the origins of musical expression and the health of habitats. The Soundscape Newsletter, 6:6–10.

[44] Krizhevsky A., Sutskever I., Hinton G.E. (2012) Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097–1105.

[45] Lartillot O., Toiviainen P. (2007) A Matlab toolbox for musical feature extraction from audio, International Conference on Digital Audio Effects. Bordeaux, France.

[46] Lasseck M. (2015) Towards automatic large-scale identification of birds in audio recordings. Lecture Notes in Computer Science, 9283:364-75.

[47] Lasseck M. (2018) Audio-based bird species identification with deep convolutional neural networks. Working Notes of CLEF 2018.

[48] Ławicki L., Cofta T., Beuch S., Dmoch A., Sikora A., Aftyka S., Czechowski P., Bocheński M., Sieczak K., Mazgaj S. (2015) Identification and occurrence of hybrids Grey-headed × European Green Woodpecker in Poland. Dutch Birding 37:215–228.

[49] LeCun Y.A., Bottou L., Orr G.B., Müller K.R. (2012) Efficient backprop. In Neural networks: Tricks of the trade, 9–48. Springer, Berlin, Heidelberg.

[50] LeCun Y., Bengio Y., Hinton G. (2015) Deep Learning. Nature. 521.

[51] Lee C.H., Hsu S.B., Shih J.L., Chou C.H. (2013) Continuous birdsong recognition using Gaussian Mixture modeling of image shape features. IEEE Transactions on Multimedia, 15(2):454–464.

[52] Maas A.L., Hannun A.Y., Ng A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In Proc. ICML, 30(1):3–8.

[53] Michalczuk J., McDevitt A.D., Mazgajski T.D., Figarski T., Ilieva M., Bujoczek M., Malczyk P., Kajtoch, Ł. (2014) Tests of multiple molecular markers for the identification of Great Spotted and Syrian Woodpeckers and their hybrids. Journal of ornithology, 155(3):591–600.

[54] Mikusinski G., Angelstam P. (1998) Economic geography, forest distribution, and woodpecker diversity in Central Europe. Conservation Biology, 12:200–208.

[55] 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. Proceedings of the Royal Society B: Biological Sciences, 285(1873):20172628.

[56] Odom K.J., Hall M.L., Riebel K., Omland K.E., Langmore N.E. (2014) Female song is widespread and ancestral in songbirds. Nature Communications, 5:3379.

[57] Pellegrini T. (2017) Densely connected CNNs for bird audio detection. In Signal Processing Conference (EUSIPCO), 2017 25th European, 1734–1738. IEEE.

[58] 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. Biological Journal of the Linnean Society, 104(3):710–723.

[59] Peterson R.T. (2008) A field guide to the birds, sixth edition. Houghton Mifflin Company.

[60] Pettersson B. (1985) Extinction of an isolated population of the middle spotted woodpecker *Dendrocopos medius* (L.) in Sweden and its relation to general theories on extinction. Biological Conservation, 32(4):335–353.

[61] Phillips Y.F., Towsey M., Roe P. (2018) Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation. PloS one, 13(3):e0193345.

[62] Pieretti N., Farina A., Morri D. (2011) A new methodology to infer the singing activity of an avian community: The Acoustic Complexity Index (ACI). Ecological Indicators, 11(3):868–873.

[63] Pironkov G. (2018) Acoustic modelling using deep neural networks for automatic speech recognition. PhD Thesis. Université de Mons, Belgium.

[64] Pons J.M., Olioso G., Cruaud C., Fuchs J. (2011) Phylogeography of the Eurasian green woodpecker (Picus viridis). Journal of Biogeography, 38(2):311–325.

[65] Potamitis I. (2014) Automatic classification of a taxon-rich community recorded in the wild. PLoS ONE 9(5):e96936.

[66] Ranjard L., Ross H.A. (2008) Unsupervised bird song syllable classification using evolving neural networks, Journal of the Acoustical Society of America, 123(6).

[67] Rao C.R. (1948) The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B (Methodological), 10(2):159–203.

[68] Salamon J., Bello J.P. (2017) Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3).

[69] Schmitz L. (2004) Hybridation des Pics vert et cendré (*Picus viridis, P. canus*) en Belgique. Aves, 41(1-2).

[70] Schrader L., Hammerschmidt K. (1997) Computer-aided analysis of acoustic parameters in animal vocalisations: A multi-parametric approach. Bioacoustics, 7(4):247–265.

[71] Somervuo P., Harma A., Fagerlund S. (2006) Parametric representations of bird sounds for automatic species recognition. IEEE Transactions on Audio, Speech, and Language Processing.

[72] 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. Service du patrimoine naturel du Muséum national d'Histoire naturelle. Paris.

[73] Stark R.D., Dodenhoff D.J., Johnson E.V. (1998) A quantitative analysis of woodpecker drumming. The Condor, 100(2):350–356.

[74] Stowell D., Plumbley M. D. (2014) Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ, 2:e488.

[75] Stowell D., Wood M., Stylianou Y., Glotin H. (2017) Bird detection in audio: A survey and a challenge. 25th European Signal Processing Conference (EUSIPCO). Kos, Greece.

[76] 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 in Ecology and Evolution, 10(3):368–380.

[77] Sueur J., Farina A., Gasc A., Pieretti N., Pavoine S. (2014) Acoustic indices for biodiversity assessment and landscape investigation. Acta Acustica United with Acustica, 100:772–781.

[78] 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. Journal of Field Ornithology, 80(1):42–50.

[79] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich, A. (2015) Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9.

[80] Testaert D. (1998) Découverte de la présence du Pic cendré (*Picus canus*) dans le sud de la Province de Namur. Aves, 35/1.

[81] Towsey M., Zhang L., Cottman-Fields M., Wimmer J., Zhang J., Roe P. (2014) Visualization of long-duration acoustic recordings of the environment. 2014 International Conference on Computer Science, Procedia Computer Science, 29:703–712.

[82] Towsey M., Wimmer J., Williamson I., Roe P. (2014) The use of acoustic indices to determine avian species richness in audio-recordings of the environment. Ecological Informatics, 21:110–119.

[83] Tremain S.B., Swiston K.A., Mennill D.J. (2008) Seasonal variation in acoustic signals of pileated woodpeckers. Wilson Journal of Ornithology, 120(3):499–504.

[84] Turner K. (2011) The case against drumming in Middle Spotted Woodpecker (Dendrocopos medius). Limicola, 1:37–53.

[85] Ulloa J.S., Gasc A., Gaucher P., Aubin T., Réjou-Méchain M., Sueur J. (2016) Screening large audio datasets to determine the time and space distribution of Screaming Piha birds in a tropical forest. Ecological Informatics, 31:91–99.

[86] van der Maaten L.J.P., Hinton G.E. (2008) Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research, 9:2579–2605.

[87] Voříšek P., Klvaňová A., Wotton S., Gregory R.D. (2008) A best practice guide for wild bird monitoring schemes, first edition. CSO/RSPB.

[88] Wallschläger D. (1980) Über das Trommeln des Mittelspechtes. Falke, 27:310–312.

[89] Whytock R.C., Christie J. (2017). Solo: An open source, customizable and inexpensive audio recorder for bioacoustic research. Methods in Ecology and Evolution, 8(3):308–312.

[90] Winkler H., Short L.L. (1978) A comparative analysis in acoustical signals in Pied Woodpeckers (Aves, Picoides). Bulletin of the American Museum of Natural History, 160.

[91] Zabka H. (1980) Zur funktionellen Bedeutung der Instrumentallaute Europäischer Spechte under besonderer Berücksichtigung von Dendrocopos major und D. minor. Mitteilungen aus dem Zoologischen Museum in Berlin 56, Suppl.: Ann. Orn. 4:51–76.