Development of artificial intelligence methods for marine mammal detection and classification of underwater sounds in a weak supervision (but) Big Data-Expert context

Paul Nguyen Hong Duc

École Doctorale de Science Mécanique, Acoustique, Electronique et Robotique

Doctoral Thesis

Discipline : Acoustics and Signal Processing

presented by

Paul NGUYEN HONG DUC

Development of artificial intelligence methods for marine mammal detection and classification of underwater sounds in a weak supervision (but) Big Data-Expert context.

---

supervised by Olivier ADAM, Paul R. WHITE, Odile GERARD and Dorian CAZAU

Defended on 17 December 2020 in front of the jury :

Pr. Holger Klinck	Cornell Lab of Ornithology	Rapporteur
Dr. Julie N. Oswald	University of St Andrews	Rapporteur
Dr. Jérôme Habonneau	Thalès	Examiner
Pr. Jean-Dominique Polack	Sorbonne Université	Examiner
Pr. Olivier Adam	Sorbonne Université	Supervisor
Pr. Paul R. White	University of Southampton	Supervisor
Dr. Odile Gerard	DGA-TN	Co-Supervisor
Dr. Dorian Cazau	ENSTA Bretagne	Co-Supervisor

Institut Jean-le-Rond-d'Alembert
Equipe Lutherie-Acoustique-Musique
11, rue de Lourmel
75015 PARIS

Sorbonne Université
Ecole Doctorale de Sciences Mécanique,
Acoustique, Electronique et Robotique
4 place Jussieu
75252 Paris Cedex 05
Boite courrier 290

Remerciements

Je tiens à remercier mon encadrement de thèse pour m'avoir accordé leur confiance et m'avoir aidé tout au long de ce travail : Olivier, Paul, Odile et Dorian.

Ce travail n'aurait pas été possible non plus sans le soutien de la DGA qui m'a permis de travailler sereinement grâce à une allocation de recherches.

Je tiens à remercier toutes les personnes qui ont oeuvré pour la prolongation de mon contrat que ce soit côté DGA, toute l'équipe Formation par la Recherche avec en tête Mme Anglade, ou côté Sorbonne Université, Sandrine, Simona, Mme Lapert. Je tiens également à les remercier ainsi que Mme Vallin pour leur assistance dans les dossiers administratifs.

Je remercie également le jury pour avoir accepté de lire et d'évaluer mon manuscrit.

J'ai eu la chance d'avoir pu enrichir ce travail à travers de belles collaborations : Joseph, Alexandre, Erwan, Romain, Anouck, Sean, Jérôme, Yann, Ronan, Julien, Florent, Oliver, Fabio, Bruno, Jean-Yves. Merci pour votre temps et vos conseils, en espérant que ces collaborations durent.

Un grand merci à l'équipe de l'ENSTA Bretagne (ordre aléatoire, j'en ai sûrement oublié) : Flore, Angélique, Charles, Samuel, Julie, Gilles, Maëlle, Morgane, Ewen, Rémi, Gaëtan, Adrien, Fabio, Guillaume, Thomas, Milan, Clément, Clémentin, Jean-Yves.

Merci aux amis qui m'ont soutenu moralement ou autre pendant ces trois années, ils se reconnaîtront.

Enfin, je tiens à remercier mes proches pour leur soutien journalier (toute la famille). Mon grand petit cousin pour m'avoir prêté sa chambre pour mes incursions à Paris ainsi que ses parents. Evidemment, mes co-locataires parents, pour une cohabitation plus ou moins réussie et mon frère.

Abstract

Abstract

The subject of our doctoral research project, initiated by Pr. Olivier Adam, Dr. Dorian Cazau, Dr. Odile Gérard and Pr. Paul R. White, is entitled: "Development of artificial intelligence methods for marine mammal detection and classification of underwater sounds in a weak supervision (but) Big Data-Expert context."

Whales have become an important topic in marine science. In addition to founding highly developed societies that are constantly evolving and worthy of individual study, whales also play a crucial role in the overall health of the ocean ecosystem. However, their role is not recognized as such and they are endangered worldwide. One of the best solutions for their protection and conservation is to better understand their social structure and the way they communicate with each other. To this end, passive acoustic monitoring (PAM) offers a unique solution for simultaneously capturing source-specific and more contextual information, thus providing a better understanding of both the behavior of whales and their relationship with the ecosystem around them.

Due to the development of underwater cabled observatories, which now have virtually unlimited power for high bandwidth, as well as continuous data acquisition, increased storage capacity, and longer life of temporary recorder batteries, the total volume of PAM data to be processed has become increasingly important. As a result, the development of automated AI methods to analyze these data is now of prime importance in marine bioacoustics. In the age of deep learning, the use of supervised approaches is known to achieve the best recognition performance in the majority of AI tasks. As they also rely heavily on the quantity and quality of the annotated data, this process of collecting annotations is the main bottleneck in the development of such methods. And because such a process is tedious, resource-intensive and dependent on human error, the annotated PAM data needed for AI methods are scarce. Subsequently, we will call "weakly supervised", a context in which a supervision signal for the annotation of large amounts of data will be hampered by noisy, limited or imprecise sources.

Overall, this paper explores two approaches in a weakly supervised context, by acting or:

— on the context itself, through the question of where and how one can better extract useful information to supervise AI methods, a study carried out in the first two chapters of the thesis ;

— or by acting on AI methods in this context, through the question of how to develop AI methods to better respond to weak supervision, which is addressed in the third chapter.

Before detailing these chapters, we would like to point out that, since the beginning of this doctoral research project, we have chosen to work within the OSmOSE project, a

collaborative and interdisciplinary environment built on the principles of Findable, Accessible, Interoperable and Reusable (FAIR) data. This explains the use of various methodological approaches such as the development of big data oriented engineering systems and AI models, and the use of diverse scientific disciplines including ecoacoustics, detection and classification of acoustic scenes and events, present in this doctoral thesis.

In Chapter 1, to answer the first question of where and how to extract useful information, we exploit a greater amount of data covering longer time periods and multiple marine sites among available big data. Faced with the limitations of tools for processing such traditional datasets, we first propose a distributed cloud computing based system to perform scale-intensive, context-sensitive and fast computation of soundscapes descriptors. This system has been benchmarked to evaluate its computational performance and has been validated on different typical uses of contextual bioacoustic analysis, such as the generation of automatic reports on the soundscape under consideration and the visualization of long-term data, for example through an exhaustive study of the soundscape of Saint-Pierre-et-Miquelon. It has also been used to conduct in-depth research on key methodological issues such as the variability of biological presence indices.

In chapter 2, again on the origin and the way in which useful information is extracted, we rely on the knowledge provided by a bioacoustic expert, in order to efficiently collect manual annotations through collaborative annotation campaigns. Following this line of work, we will insist on two aspects:

— 1) Development of new technological tools and transfer of state-of-the-art AI methods to accelerate the annotation process: we first developed a new web-based but scalable annotation platform, and then we studied two AI methods, namely active few-shot learning methods ;

— 2) a better understanding of inter-native variability and the implementation of recommendations for future effective collaborative annotation campaigns: we conducted a comprehensive study of inter-native variability based on an original collaborative annotation campaign on a subset of low-frequency public data from the challenge proposed at the Detection, Classification, Localization and Estimation conference.

In the last and third chapters, we addressed our second research question, how to develop AI methods to better respond to weak supervision, as defined above, i.e., we can only provide a small amount of annotated data to train an AI model. First, we explored different data augmentation strategies to generate new samples to significantly increase the AI training database. In particular, the results were obtained by the contribution of physics through the integration of underwater propagation models parameterized by site characteristics to describe the physical deformations of a source signal through the propagation medium. We have also performed a large-scale comparative analysis of different transfer learning approaches, using either pre-trained neural networks from other audio related tasks, or synthetic bioacoustic datasets.

Our contribution has allowed us to develop new tools to help the recognition of whale sounds through a more open science.

Keywords

Passive acoustic monitoring, Bioacoustics, Deep learning, Big data, Weak supervision, Open science

Développement de méthodes d'intelligence artificielle pour la détection des mammifères marins et la classification de sons sous-marins dans un contexte de faible supervision (mais) de Big Data-Expert

Résumé

Le sujet de notre projet doctoral, initié par le Pr. Olivier Adam, le Dr Dorian Cazau, le Dr Odile Gérard et le Pr. Paul R. White, est intitulé : ≪ Développement de méthodes d'intelligence artificielle pour la détection de mammifères marins et pour la classification de sons sous-marins dans un contexte de faible supervision (mais) Big Data-Expert ≫.

Les baleines sont devenues un sujet important dans les sciences marines. Outre le fait qu'elles fondent des sociétés très développées qui évoluent constamment et qui méritent d'être étudiées à titre individuel, les cétacés jouent également un rôle crucial dans la santé générale de l'écosystème des océans. Cependant, le rôle les baleines n'est pas reconnu comme tel et elles sont en danger dans le monde entier. L'une des meilleures solutions pour leur protection et leur conservation consiste à mieux appréhender leur structure sociale et la façon dont elles communiquent les unes avec les autres. À cette fin, l'acoustique passive sous-marine -Underwater Passive Acoustics en anglais, U.P.A.- offre une solution unique pour saisir simultanément des informations spécifiques à la source et des informations plus contextuelles, permettant ainsi de mieux comprendre à la fois le comportement des baleines et leurs relations avec l'écosystème qui les entoure.

En raison du développement des observatoires câblés sous-marins qui sont désormais dotés d'une puissance pratiquement illimitée pour une bande passante élevée, ainsi que de l'acquisition continue de données, de l'augmentation de la capacité de stockage et de la durée de vie des batteries d'enregistreurs temporaires, l'ensemble des volumes de données U.P.A. à traiter est devenu de plus en plus important. En conséquence, le développement de méthodes automatisées d'intelligence artificielle (I.A.) pour analyser ces données est dorénavant de première importance en bioacoustique marine. À l'ère de l'apprentissage profond -deep learning, en anglais, l'utilisation des approches supervisées est réputée pour obtenir les meilleures performances de reconnaissance dans la majorité des tâches d'I.A.. Comme également, elles reposent fortement sur la quantité et la qualité des données annotées, ce processus de collecte des annotations constitue la principale limite dans la mise au point de telles méthodes. Et comme un tel processus est fastidieux, gourmand en ressources et dépendant de l'erreur humaine, les données U.P.A. annotées nécessaires aux méthodes d'I.A. sont rares. Par la suite, nous appellerons ≪ faiblement supervisé ≫, un contexte dans lequel un signal de supervision pour l'annotation de grandes quantités de données sera gêné par des sources bruyantes, limitées ou imprécises.

Globalement, cette contribution explore deux approches dans un contexte faiblement supervisé, en agissant :

— sur le contexte lui-même, à travers la question de savoir où et comment on peut mieux extraire de l'information utile pour superviser les méthodes d'I.A., étude menée dans les deux premiers chapitres de la thèse ;

— ou en agissant sur les méthodes d'I.A. dans ce contexte, à travers la question de savoir comment développer des méthodes d'I.A. pour mieux répondre à une supervision faible, point traité dans le troisième chapitre.

Avant de détailler davantage ces chapitres, nous soulignons que, depuis le début de ce doctorat, nous avons choisi de travailler au sein du projet OSmOSE, un environnement collaboratif et interdisciplinaire construit sur les principes de données Faciles à trouver, Accessibles, Interopérables et Réutilisables (F.A.I.R.). Cela explique l'utilisation d'approches méthodologiques variées comme le développement de systèmes d'ingénierie pour traiter les données massives et de modèles d'I.A., tout autant que le recours à des disciplines scientifiques diverses notamment l'écologie acoustique, la détection et la classification des scènes et des événements acoustiques, présents dans ce doctorat.

Dans le chapitre 1, pour répondre à la première question d'où et comment est extraite l'information utile, nous avons cherché à exploiter une plus grande quantité de données couvrant des périodes plus longues et des sites marins multiples parmi les données massives disponibles. Face aux limites des outils de traitement de tels jeux de données traditionnels, nous proposons d'abord un système distribué basé sur l'informatique en nuage pour effectuer à l'échelle, avec rapidité et en contexte, un calcul intensif des descripteurs de paysages sonores. Ce système a fait l'objet d'une analyse comparative pour évaluer ses performances de calcul et a été validé sur différentes utilisations typiques de l'analyse bioacoustique contextuelle, comme la génération de rapports automatiques sur le paysage sonore considéré et la visualisation de données long terme, par exemple par une étude exhaustive du paysage sonore de Saint-Pierre-et-Miquelon. Il a également été utilisé pour effectuer des recherches approfondies sur des questions méthodologiques clés telles que la variabilité des indices de présence biologique.

Dans le chapitre 2, toujours sur cette même première question, nous nous sommes appuyés sur la connaissance apportée par un expert en bioacoustique, de façon à collecter efficacement des annotations manuelles à travers de campagnes d'annotation collaborative. En suivant cet axe de travail, nous insisterons sur deux aspects :

— 1) le développement de nouveaux outils technologiques et le transfert de méthodes de l'état de l'art en I.A. pour accélérer le processus d'annotation : nous avons d'abord développé une nouvelle plate-forme d'annotation basée sur le web mais évolutive, puis nous avons étudié deux méthodes d'I.A., l'apprentissage actif et les méthodes d'apprentissage par quelques clichés -few shot learning, en anglais ;

— 2) une meilleure compréhension de la variabilité inter-annotateurs et de la mise en place de recommandations pour de futures campagnes d'annotation collaborative efficaces : nous avons réalisé une étude exhaustive de la variabilité inter-annotateurs à partir d'une campagne d'annotation collaborative originale sur un sous-ensemble de données publiques basse fréquence issues du défi proposé lors de la conférences Détection, Classification, Localisation et Estimation de la Densité 2015 ,et l'impact de cette variabilité sur les méthodes d'I.A..

Dans le dernier et troisième chapitre, nous avons abordé notre deuxième question de recherche, afin de savoir comment développer des méthodes d'I.A. pour mieux répondre à une supervision faible, définie plus haut, c'est-à-dire que nous ne pouvons fournir qu'un petit nombre de données annotées pour entraîner un modèle I.A.. D'abord, nous avons exploré différentes stratégies d'augmentation des données pour générer de nouveaux échantillons afin d'accroître de façon conséquente la base de données d'entraînement I.A.. En particulier, les résultats ont été obtenus par l'apport de la physique grâce à l'intégration des modèles de propagation sous-marine paramétrés grâce aux caractéristiques les caractéristiques d'un site pour décrire les déformations physiques d'un signal source à travers le

milieu de propagation. Puis, nous avons effectué une analyse comparative à grande échelle de différentes approches d'apprentissage par transfert, en utilisant ou des réseaux neuronaux pré-entraînés provenant d'autres tâches liées à l'audio, ou des ensembles de données bioacoustiques synthétiques.

Notre contribution à permis de développer de nouveaux outils pour l'aide à la reconnaissance de sons de baleines à travers une science plus ouverte.

Mots-clés

Surveillance acoustique sous-marine passive, Bioacoustique, Apprentissage profond, Big data, Faible supervision, Science ouverte

Contents

List of Figures

List of Tables

Notation

All notations of terms are listed.

List of acronyms

Disciplines

AI Artificial Intelligence
ASC Acoustic Scene Classification
CDM Cetacean Distribution Modelling
DCL Detection, Classification and Localization
DL Deep Learning
ML Machine Learning
PAM Passive Acoustic Monitoring

Algorithms

CNN Convolutional Neural Network
CRTAL Certainty-based active learning
FDS Fast Dawid-Skene
GAM General Additive Model
GBDT Gradient Boosting Decision Tree
GEE Generelized Estimating Equations
GMM Gaussian Mixture Model
HAC Hierarchical Agglomerative Clustering
HMM Hidden Markov Model
MV Majority Voting
NN Neural Network
SVM Support Vector Machine

Acronyms of notions

ACI Acoustic Complexity Index
AL Active Learning
DFT Discrete Fourier Transform
FFT Fast Fourier Transform
FSL Few-Shot Learning

HDFS Hadoop Distributed File System

lgbm or LightGBM Light Gradient Boosting Model classifier (Ke et al., 2017)

LTALR Long Term Average Learning Representation

MFCC Mel Frequency Cepstrum Coefficient

PSD Power Spectral Density

ResNet Residual Network

RMS level Root-Mean Square level

TOL Third Octave Level

Sound Database

DCASE Detection and Classification of Acoustic Scenes and Events http://dcase.community/

DCLDE Detection, Classification, Localization and Density Estimation http://cetus.ucsd.edu/dclde/

Macaulay Library Environmental sound database https://www.macaulaylibrary.org/

MobySound Marine mammal sound database (Mellinger and Clark, 2006)

PNMI Parc Natural Marin d'Iroise

OHASISBIO Observatoire Hydro-Acoustique de la SISmicité et de la BIOdiversité (doi: 10.18142/229)

Watkins Marine mammal sound database (Sayigh et al., 2016)

Other notation

AIS Automatic Identification System for vessel identification

APLOSE Annotation PLatform for Ocean Sound Explorers

ARGO International program that uses floats to collect oceanographic data http://www.argo.ucsd.edu/

CPOD Underwater recorder with a odontocete acoustic detector https://www.cpodclickdetector.com/

DC Detection and Classification Essential Ocean Variables

ELT Extract, Load and Transform process

FAIR Findability, Accessibility, Interoperability, and Reusability

MEOPS http://www.meop.net/

MFSD EU Marine Strategy Framework Directive of the European Union

MM Marine Mammal

MPI Message Passing Interface

OSmOSE Open Science meets Ocean Sound Explorers https://osmose.xyz/

PALAOA The Perennial Acoustic Observatory in the Antarctic Ocean

Introduction

General

This introduction will cover the general concepts of this doctoral thesis mentioned in the title (namely, "passive acoustic monitoring", "big data", "deep learning" and "weak supervision").

In particular, concepts in AI are presented to a broad audience, assuming that this is not (yet) common knowledge in PAM communities, who generally comes from signal processing or physics disciplines. Key concepts were drawn from modern well-known references for AI, such as Chapelle et al. (2010); Goodfellow et al. (2016); Bianco et al. (2019).

Why studying whales?

Whales play a very crucial role in the health of both ocean and terrestrial ecosystems. Indeed, a stable food chain is achieved thanks to whales as they act by limiting the expansion of certain marine species. A blue whale for example can consume as much as 40 million krill per day, making it a crucial element of stabilization of the entire food chain (Hildyard, 2001). They also act as a reservoir of carbon, with about 400,000 tonnes of carbon that are extracted from the air due to these whales per year (Roman and McCarthy, 2010). Even whale fecal matter stimulates the growth of phytoplankton supplying other species that need it to survive (Smith et al., 2013). Whales are also an attraction for tourists and they help growing economies that rely on whale watching and other activities bring in capital through tourism.

Despite these important roles, whales are endangered worldwide. They were targeted by the whaling and even after international conservation laws against whaling, this activity still goes on in several countries for commercial and cultural purposes. Nowadays, whales are also threatened by environmental changes that impact krill stocks. Anthropogenic activities such as fishing, military operations, tourism are a threat to whales and their habitats. These cetaceans can be struck by vessels, caught in fishing nets. Another threat is ocean sound pollution. Anthropogenic sound can impact whales in different ways such as behavioral responses, communication masking and hearing losses (Mann et al., 2010).

One of the best solutions for their protection and conservation involve a better understanding about their social structure and how they communicate with one another, as whales form highly developed relationships and societies with one another, whales also display a high level of intelligence and self-recognition (Delfour and Marten, 2001) (a characteristic that is found in only a few animal species).

Passive Acoustic Monitoring (PAM)

A bit of history

Passive Acoustic Monitoring is a scientific discipline starting during World War I. In that time, a first passive sonar was set up to detect German U-boat submarines which aimed at sinking merchant shipping. Major breakthroughs were essentially achieved between the two World Wars. In the late 1930s, an anchor buoy system named "Herald" was developed by the Naval Research Laboratory to detect submarines entering port areas (Klein, 1968). Then, the US pursued their efforts in the underwater warfare by creating the National Defense Research Committee to improve their underwater acoustic systems (Lasky, 1977). Military applications greatly helped in the development and improvement of the underwater acoustic systems.

Due to its history, underwater passive acoustics is strongly related to military applications. However, this eclipsed many other applications among which oceanography, tomography and bioacoustics (Bjørnø, 2003; Muir and Bradley, 2016). The impact of anthropogenic noise on marine ecosystem was not considered until the 1970s (Bradley, 2008) when researchers demonstrated that some cetaceans could be potentially affected by oil and gas exploration that used drilling and active acoustics. Thus, the protection and conservation of cetaceans is becoming a hot topic with an increase in the number of studies on this topic over the years (cf Fig. 1).

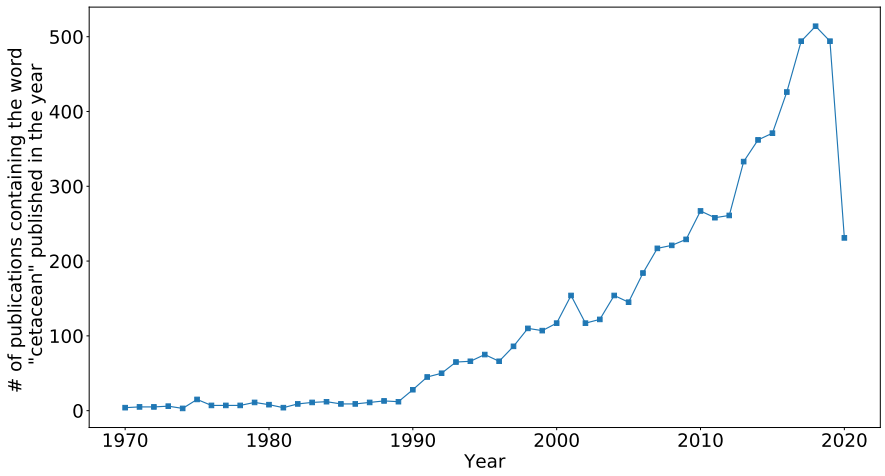


Figure 1 – Number of publications per year containing the word "cetacean" (source: Web of Science, July 26th, 2020).

Instrumentation side

PAM is used as one of many monitoring systems to gain knowledge about cetaceans. It is often assimilated to a sensor-based observation method of ocean processes and human activities based on a hydrophone, which is the underwater equivalent to the microphone. As a monitoring method, it exhibits several advantages over surveying traditional methods which require manual and field workers to visit the study site: it enables to record underwater audio data continuously and autonomously for weeks or months (depending on their power source, the acoustic sources of interest), it costs relatively little compared

to field work, it is minimally invasive, and it creates a permanent, objective record of a site. Sometimes because of the lack of access to the ocean (due to bad weather conditions, lack of observers, or the expense of field work), PAM becomes an attractive option to analyze underwater soundscapes.

Data analysis side

From a signal processing point of view, the general character of the ocean ambient noise (NRC, 2003) is a slowly varying background interrupted by shorter time scale sound events. Thus, PAM encompasses different analysis methods that depend on both the spatio-temporal scale of the analysis and on the end-user application. Analysis of the ambient noise is a major application of PAM and provides information on a coarse temporal scale. Several applications of PAM analysis from Robinson et al. (2014) are listed and explained. Each of them can be useful in the management of marine and coastal environments.

— "To provide a descriptor of the overall noise level due to all noise sources in the location, perhaps to evaluate the effect on marine species (e.g., from masking)": when conducting oil and gas or seismic explorations, whales avoid the area due to noise (Gordon et al., 2003; Williams et al., 2015). It can help set up regulations for such operations to limit the noise exposure experienced by marine mammals (Farcas et al., 2016).

— "To characterise the nature of the noise and the soundscape": the main aims are to determine the acoustic sources which generates the underwater noise (is it from anthropogenic activities (anthrophony)? From marine wildlife (biophony)? Or from geophysical processes (geophony)?) and to understand the interactions between anthrophony, biophony and geophony (NRC, 2003; Pieretti et al., 2011; Pijanowski et al., 2011; Miksis-Olds et al., 2018).

— "To provide a descriptor of the typical background noise in the location in the absence of temporary or transient events": this could help better understand marine processes and propagation in a specific area which is difficult to access (Kinda et al., 2013). It could also be used as a monitoring system to determine rainfall or wind speed (Nystuen and Selsor, 1997; Pensieri et al., 2015).

— "To determine the performance of a system in the presence of known background noise": ambient noise is an input for the sonar equation and it is the basis of every underwater acoustic system (Etter, 2012; Siddagangaiah et al., 2015).

— "To compare the noise levels with those in other locations": retrieving several PAM datasets from different sites allows scientific to compare them and define which are the main contributors of the soundscape at a specific location (Haver et al., 2019);

— "To determine trends in noise levels": for this purpose, several trends can be determined such as seasonality. For example, regarding biophony, it can match migration patterns from breeding to mating grounds for cetacean or diurnal variations for coral reefs (Bertucci et al., 2017; Haver et al., 2019). In anthrophony, summer and weekends can experience an increase in underwater noise levels due to recreational boating (Samuel-Rhoads et al., 2005).

— "To provide a measure of background noise to estimate the spatial range of audibility of a specific sound source": sound reception heavily depends on the characterization of propagation environment and thus on the ambient noise. It helps regulating

anthropogenic activities to lower their effect on marine mammal hearing system (Bailey et al., 2010).

Furthermore, by understanding and reducing the sources of variability in an analysis, the framework proposed could be more robust and reused for different locations and time periods. Another major type of PAM analysis focuses on shorter acoustic events (during from less than one to several seconds), which aims at classification (Bergler et al., 2019; Kirsebom et al., 2020) and when possible source localization (Niu et al., 2017; Gemba et al., 2019).

General applications

PAM data are exploited across a wide variety of applications such as acoustical oceanography, sonar, geophysical exploration, underwater communications, offshore engineering. PAM is also one of the primary methods used in ecology, where long-duration acoustic recordings of the environment are increasingly used to monitor species diversity in both terrestrial and marine ecosystems. More precisely, in the marine realm, PAM can be used to measure the presence, abundance, and migratory patterns of marine mammals (Zimmer, 2011), and for the assessment of risk to marine life (Barlow and Gisiner, 2006).

Promise for the future

PAM already shows great promise regarding a great diversity of applications in a near-future. One of these research avenues is the on-board integration of PAM in large-scale networks of mobile multi-physics platform (Howe et al., 2019a) (e.g., gliders (Hildebrand et al., 2010; Cauchy et al., 2018), ARGO profilers (Yang et al., 2015), bio-logged animals (Cazau et al., 2017)), with the long-term goal of routinely sampling new PAM-based Essential Ocean Variables (EOV) within existing monitoring programs such as ARGO [1] or MEOPS [2]. These EOV concern both geophysical variables like wind speed as well as biological ones like presence indexes of marine mammal populations.

General challenges

This general introduction to PAM is concluded with current open challenges on the data analysis side. A great number of these challenges result from the complex multi-stage processing chain of a source information in PAM, as illustrated in figure 2. For example, a whale produces a vocalization. This sound is propagated through the underwater environment which is characterized by several parameters such as bathymetry, temperature and salinity. Consequently, the vocalization is first altered by the local propagation conditions. Other acoustic sources contribute to the overall ambient noise in the ocean such as geophony (wind, rain, underwater earthquakes) and anthrophony (vessel noise, drilling, sonar operations) and may mask the sound emitted by the whale. The vocalization is recorded by an hydrophone. Once the hydrophone is retrieved, underwater recorded audio data are analyzed. Finally, useful information is extracted. Through this whole and long process to extract information from underwater recordings, the signal is altered by different sources such as the acquisition site, the material to record the signal and the softwares used to process it. The main challenge is to reduce as much as possible this variability to build robust detectors and classifiers of cetacean vocalizations.

1. http://www.argo.ucsd.edu/ .
2. http://www.meop.net/.

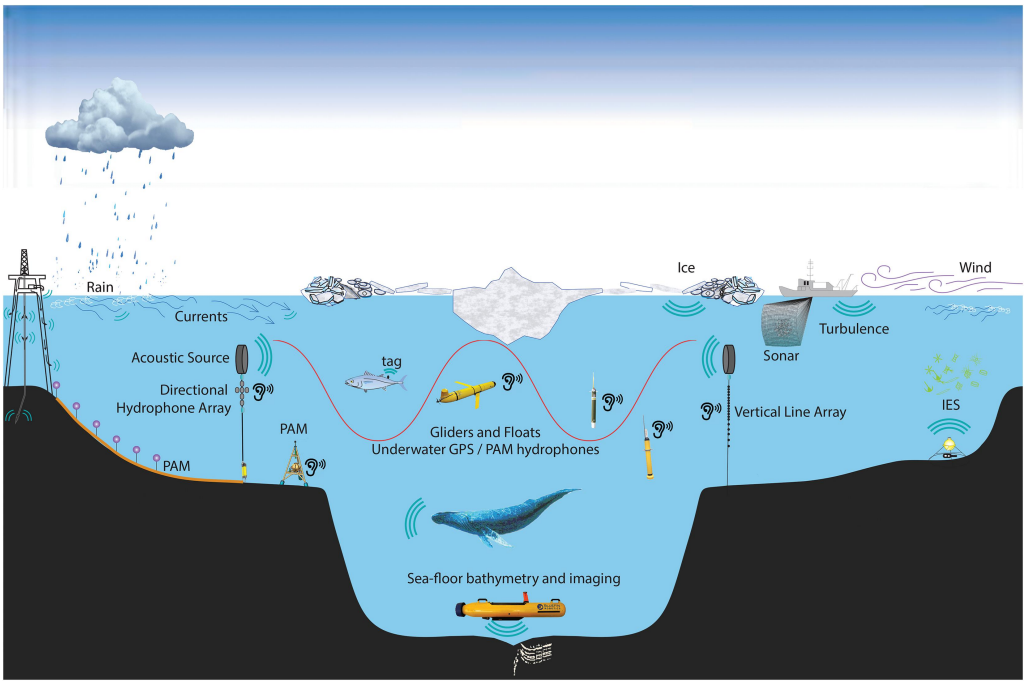


Figure 2 – Figure showing how the different PAM generation and processing steps may induce variability in the process of source information extraction from (Howe et al., 2019b).

PAM-based whale presence index

Before defining the special meaning to this concept in this work, the reader is first introduced the two main methodological approaches classically used to inform such an index, namely through the two disciplines of Ecoacoustics and Detection, Classification, Localisation and Density Estimation (DCLDE).

Ecoacoustics

Ecoacoustics is a relatively new discipline that studies soundscapes for ecological purposes (Farina and Gage, 2017). Ecoacoustics investigates sound, not only as a mechanism of animal communication, but also as an indirect indicator of animal diversity, abundance, dynamics, distribution, and of ecosystem function. An important analytical tool is the "acoustic index, a mathematical function that summarizes some aspect of the distribution of acoustic energy in a recording" (Phillips et al., 2018) "to evaluate some aspects of biodiversity" (Sueur et al., 2014). Some indices are traditional, such as signal-to-noise ratio and some describe the spectral distribution of energy or its segmentation into acoustic events but more have been developed to highlight specific trends in a soundscape (Sueur et al., 2014; Harris et al., 2016).

Characterizing an underwater soundscape can provide an insight into the status of the ocean ecosystem at a particular location. Underwater soundscapes (Miksis-Olds et al., 2018; Pijanowski et al., 2011) are a valuable source information to describe marine ecosystems as they are the result of complex interactions between biophony, geophony and anthrophony at various spatio-temporal scales (Lillis et al., 2018). Describing the soundscape and then understanding relationships between these different types of sounds provides information which can be used in developing conservation strategies for marine ecosystems to mitigate the constantly increasing disturbance from anthropogenic activities (Frisk, 2012;

McDonald et al., 2006; McKenna et al., 2012). This marine anthropogenic sound pollution is a threat to the fauna, especially marine mammals (Southall et al., 2008). Indeed, sound is one of the key elements of vital activities in marine life. For example, marine mammals perceive and generate sounds for foraging, social interactions, mating, escaping from predators and also navigating (Clark et al., 2009). Moreover, it has been shown that marine mammals are affected by the deterioration of the sound underwater environment, mainly resulting from increasing maritime traffic (Putland et al., 2017; Tasker et al., 2010).

Acoustic indices as features are also used to classify soundscapes of ecological interest. This task is a particular application of computational auditory scene analysis (CASA) (Barchiesi et al., 2015).

Detection and Classification (DC) of cetacean

It is noteworthy that PAM for marine mammal detection and classification (MMDC) is part of a wider scientific discipline called automatic environmental sound classification, whose terrestrial applications range from context aware computing (Chu et al., 2009) and surveillance (Radhakrishnan et al., 2005) to noise mitigation enabled by smart acoustic sensor networks (Mydlarz et al., 2017). In this discipline, the main goal is to find the best system able to automatically detect and classify underwater sounds. Detection is defined as identifying if a specific signal is present in an audio recording and classification refers to determining the type of the sound (Van Trees, 1968).

The two previous disciplines share a common goal of detecting and classifying marine mammal presence within complex multi-source PAM recordings. However, when considering their application to marine mammal presence indexes, they may be differentiated at least regarding the following aspects:

— temporal scales: Ecoacoustics tends to process soundscapes as a "whole", where biological sources are only one component. DC usually aims at identifying specific biological sources in a soundscape

— interpretability: low-level physical descriptors used in Ecoacoustics offers direct interpretation of the contents. DC is often seen as a "black box" when using artificial intelligence systems

— computational efficiency: current methods in Ecoacoustics are better suited to process high-volume data. DC can be integrated into real time systems.

These two disciplines give different insights in the audio dataset. Ecoacoustics provides long-term information whereas DC can refine this analysis. However, DC systems can be difficult to design and they are mostly used for specific tasks and datasets whereas ecoacoustics methods can be applied irrespective of that.

Big data and Machine learning

Big data

A first definition of Big Data could be[3]: "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze". Digital data volume is increasing year on year in every research area. Even if

3. From https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_exec_summary.ashx.

existing technologies are not able to keep up the pace imposed by the growth of data, they help to process efficiently and at scale the huge amount of data.

Computer clusters are sets of distributed computer resources that aim to achieve a common task. They can be made up of storage and computational resources that work together. They are divided into nodes and they can be networked to storage devices.

Machine and Deep learning

This general introduction is drawn from classical references in the machine / deep learning community Goodfellow et al. (2016).

Classification is the main task covered in our work. In this type of task, some input (image, audio, text, ...) is fed into a computer program that is asked to predict from which k categories it belongs to. A simple example is classifying images as "cat" or "dog". Classification algorithms are usually asked to fit a mapping (or predictor) function $f : \mathbb{R}^n \rightarrow \{1,\cdots,k\}$ from input variables x to y. The output of the classification model is not always a class or label, but it can also be probabilities such as a probability distribution over the classes.

Unsupervised Raw dataset or pre-computed features from a dataset are fed into an unsupervised learning algorithm that learn valuable characteristics of the dataset structure. It determines the inferences of the input data without any expert supervision and it tries to derive input data hidden patterns on the fly. In machine and deep learning, it is assumed that a dataset was generated by a probability distribution and that the main goal is to find this distribution "whether explicitly as in density estimation or implicitly for tasks like synthesis or denoising" (Goodfellow et al., 2016). However, there are also weaker forms of unsupervised learning, such as quantile estimation, clustering, outlier detection, and dimensionality reduction. Unsupervised learning attempts to explicitly or implicitly find either structure or the probability distribution (here the mapping function f is a probability distribution p) based on observations represented by a vector $\mathbf{x}$.

Supervised Supervised learning algorithms are trained based on a training dataset containing features, and each example of this dataset is associated with a class or label which is seen as the truth. Supervised learning learns to predict the label, y, of an example of the dataset, representing by a vector $\mathbf{x}$, with y and $\mathbf{x}$ given as inputs to the supervised system by approximating $p(\mathrm{y}|\mathbf{x})$. As its name implies, supervised learning systems can be seen as a student system who has been provided an example input, $\mathbf{x}$ and their associated desired outputs y, and the student has to learn how to map the inputs to the outputs.

Discriminative algorithms, on the contrary to generative algorithms, do not try to estimate how the $\mathbf{x}^i$ have been generated, but instead concentrate on estimating $p(\mathrm{y}|\mathbf{x})$. Some discriminative methods even limit themselves to modeling whether $p(\mathrm{y}|\mathbf{x})$ is greater than or less than 0.5; an example of this is the support vector machine (SVM). It has been argued that discriminative models are more directly aligned with the goal of supervised learning and therefore tend to be more efficient in practice.

But note that the division between unsupervised and supervised learning is often blurred as supervised learning can be resolved by using unsupervised approaches to learn a representation of the data (Goodfellow et al., 2016). These two concepts are not absolutely formal, but they enable to distinguish two types of problems: whether an historical labeled dataset is available or not to help the learning system.

Weak supervision Weak supervision can be defined as [4] "a branch of machine learning where noisy, limited, or imprecise sources are used to provide supervision signal for labeling large amounts of training data in a supervised learning setting. This approach alleviates the burden of obtaining hand-labeled data sets, which can be costly or impractical."

Evaluation To evaluate if a machine learning model performs well at predicting new inputs, the system predictions are assessed on a test set which is a collection of examples generated from the same probability distribution as the training set. This probability distribution is called the data generating process (Goodfellow et al., 2016). The independent and identically distributed assumptions are made: examples in the datasets are independent from each other, the train and test set are identically distributed and drawn from the same probability distribution, the data generating distribution denoted p_{data} (Goodfellow et al., 2016). Thanks to this probabilistic framework and the independent and identically distributed (i.i.d.) assumptions the training error and test error relationship can be mathematically analyzed (Goodfellow et al., 2016).

In practice, the system is first trained with the training set, then its parameters are adjusted to reduce the training error and the system is checked on the test set. In this case, "the expected test error is greater than or equal to the expected value of training error" (Goodfellow et al., 2016). Two main issues arise from these errors. When the gap between the training and test error is too large, it is called overfitting. In practice, it means that the system is able to almost perfectly predict each label of the training set, but it poorly does for new unseen data. The system does not generalize well. On the contrary, when the system is trained and it does not achieve a low enough training error, this is called underfitting. In practice, it means that the system is not complex enough to capture the structure of the examples of the training set. One concept behind these two issues is the capacity of a system which is the ability to be able to capture a broad spectrum of functions. In an underfitting scenario, the system has a low capacity whereas in an overfitting one, its capacity is too high. The desired system with best performances lies in the middle of overfitting and underfitting.

Generalization "The central challenge in machine learning is that we must perform well on new, previously unseen inputs, i.e. not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called generalization" (Goodfellow et al., 2016).

Real-life AI study cases are presented to make it more concrete. Figure 4 represents three pictures of two different species of cats in different environmental contexts and recording setup, from left to right: in A), a clean close-up photo of an Asian golden cat ; in B), the photo of the same cat species but now taken in the wild at night and further away from it ; in C), still the photo of a brown cat but that is now in the presence of another animal, both looking more like play animals rather than wild ones.

The AI task of "recognizing the cat species in B) among a large corpus of cat photos using some learning knowledge extracted from A)" is now considered. This is a case of **domain adaptation**, i.e. a same task of recognizing a unique cat species but showing highly different input distributions. Another possible AI task would be to "give a yes/no answer to the question: is there a cat in C) ?, still using some learning knowledge extracted from A)". This is a case of **transfer learning**, i.e. a different recognition task because a different cat species but with potentially related information.

4. From Wikipedia page Weak Supervision https://en.wikipedia.org/wiki/Weak_supervision.

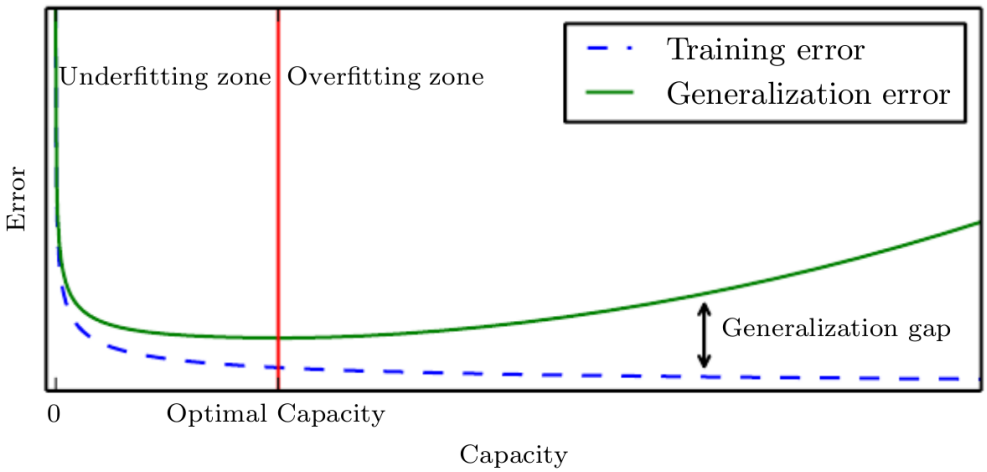


Figure 3 – Typical relationship between capacity and error. Training and test error behave differently. At the left end of the graph, training error and generalization error are both high. This is the underfitting regime. As capacity increases, training error decreases, but the gap between training and generalization error increases. Eventually, the size of this gap outweighs the decrease in training error, and the overfitting regime is reached, where capacity is too large, above the optimal capacity. Figure taken from Goodfellow et al. (2016).

Domain adaptation and transfer learning are two well-known situations where the issue of generalization needs to be addressed, which can be done through the common fundamental question of "Is my learning knowledge extracted from A) general enough to make subsequent learning task easier, or in other words to be able to recognize cat species in a robust way that make it independent from the picture characteristics". As an AI system is never and must not be learned to retrieve information that are already well known by the system.

Digging into the "learning knowledge" expression, the previous question can be now reformulated into "which learned features from the cat photo A) would allow for an efficient generalization on a sub-task?". Naturally, optimal features here will depend on the subtask objective. For the cat species identification task, i.e. learning from A) and testing on B), generalizable features would likely be high-level characteristics of the cat species like the color of its hairs and the relative size of its ears. For the cat presence recognition task with transfer learning, i.e. learning from A) and testing on C), the learned features that will be useful will likely more general ones for cats, such as their relatively small size and long hair.

Generalization in marine bioacoustics Mirroring figure 4 used as illustrative case studies of generalization, figure 5 represents in A) a knowledge source corpus on humpback whale vocal sounds, consisting in a small catalogue of isolated template-like sound units over different call types (as defined in literature, e.g. Au et al. (2006)) ; in B) the spectrogram of a sound unit recorded in the wild at a location not included in the knowledge corpus ; in C) a long-term averaged soundscape excerpt where multiple PAM sources are present, including whale vocal sounds from species other than humpback whales. As before, similar tasks can be formulated, a domain adaptation one from A) to B) when "trying to recognize cetacean species in a sound-long recording", and a transfer learning one from A) to C) when "trying to recognize that whether there is a cetacean in a long-

Figure 4 – Three pictures of two different species of cats in different environmental contexts and collection setup, from left to right: A) an Asian golden cat zoomed and centred in the picture, in very clean recording conditions; B) the same cat species than A) but in the wild at night; C) another cat species with a dog.

term recording or not". The same fundamental question as before needs to be addressed: "which learned features from the sound unit catalogue A) would allow for an efficient generalization on our two sub-tasks ?".

The vocalisations of species are classically divided into catalogue of sounds following the work by Payne and McVay (1971) on humpback whale songs. But to the best of our knowledge, no clear experimental evidence has ever been shown on how well such "template" features generalize to different sites and contexts. On the contrary, some studies claim that such features may suffer from high relative variability, with dependencies on the analyzer (setting for example the number of different types of vocal sounds and the features judged the most salient in a spectrogram for annotation), from the recording site, from the surrounding soundscape. Erbe et al. (2017) says "it might be less arbitrary to group them according to how they are produced", suggesting anatomy-based features would show more invariance. Such an anatomy-based approach has been explored (Mercado et al., 2010; Adam et al., 2013; Cazau et al., 2013; Cazau and Adam, 2016) but has not so far provided so far objective cues of vocal features that would be discriminative from one species to another (or other sub-category) based on its specific sound producing anatomy. Although Cazau and Adam (2016) suggested that vocal non-linearities like frequency jumps might be robust cues for individual or species discrimination as they are highly robust to acoustic propagation and might also be customized in relation to fine anatomical specificities. While other authors support the idea that humpback whales adapt their vocal features to their surrounding environment. To date, there has been no definitive evidence to support this idea.

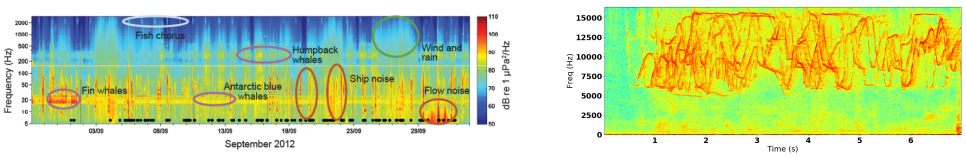


Figure 5 – Two spectrograms of cetaceans in different environmental contexts and recording setups from Erbe et al. (2015) and Saint-Pierre-et-Miquelon dataset.

Doctoral research project overview

The key ingredients at a glance

Context The main issue addressed in this work is the weak supervision for assessing cetacean acoustic presence, i.e. how to deal with the lack of labels for underwater acoustic

scenes or events either for AI applications or for ecoacoustics?

Methodological approach Two complementary approaches are proposed: i) by tackling the weak supervision issue at the root with the help of technologies able to handle the huge volume of acoustic data or by using human experts and analyzing how they bias the performances of the learning systems; ii) by fine-tuning the learning models;

AI (methodological) objective The most common use case when developing an AI system is the capability of re-using a trained system to a new dataset: this is called generalization;

Ecological (thematic) objective The higher level of application of this work is to be able to assess marine mammal acoustic presence through ecoacoustics and DC approaches based on PAM;

AI tasks In ecoacoustics, the major concern is to compute a continuous presence index of cetaceans to target cetacean vocally active time periods whereas in DC underwater bioacoustic task, it is to be able to discriminate which scene or event contains a cetacean vocalization;

Datasets 5 datasets, labeled either for scenes or events, as detailed in tables 1 and 2;

Target species Vocalizing marine mammals, including blue whales (southern and central coast of California), fin whales (southern and central coast of California);

Disciplines AI, big data management and processing, marine bioacoustics;

name	start_date	end_date	duty cycle [%]	lat [°N]	lon [°E]
SPMAuralB2010	08/19/2010	11/02/2010	75.0	47.09	-56.69
synthetic	08/01/2017	10/06/2017	100.0	45.00	-10.00
chagos	04/14/2017	09/06/2017	44.6	-6.66	71.40
argoLOV	05/05/2017	03/19/2018	0.3	43.26	8.00
DCLDE2015HF-CINMS_17_B	12/26/2011	01/01/2012	100.0	34.00	-120.00
ohasisbio_2015_wker	01/16/2015	08/01/2016	99.9	-46.57	60.52

Table 1 – Table of PAM datasets used in the doctoral research project, on the bioacoustics side.

Context & challenges

All our doctoral research project motivations come from the weak supervision context and associated challenges in marine bioacoustics, which has been the main hurdle to the development of machine learning and now to deep learning in our research community described in Chapter 2. In marine bioacoustics, a noisy supervisory signal might be an unexperienced annotator in a crowd-sourcing context who has a low level of confidence in recognizing different acoustic sources. An imprecise supervisory signal might be a presence

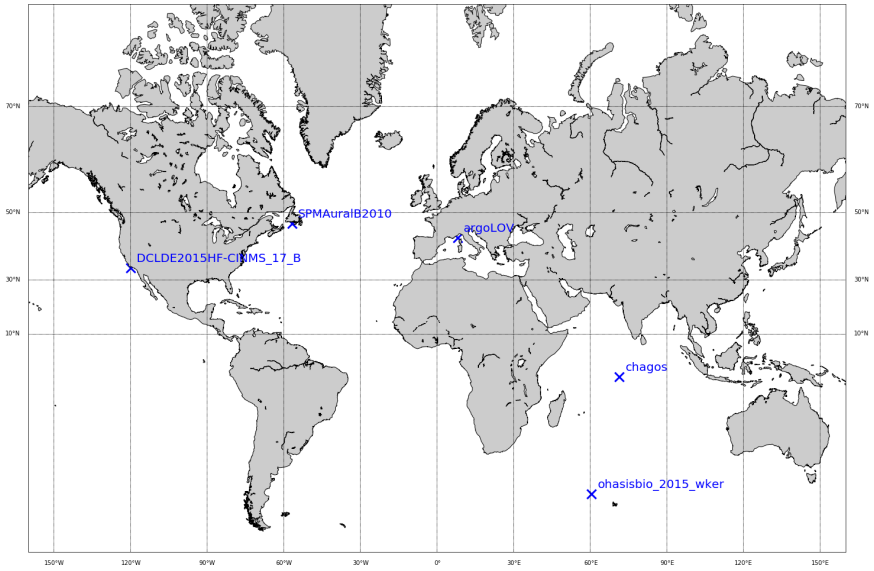


Figure 6 – OSmOSE datasets.

Name	Total volume [GB]	Number of audio files	File size [MB]	Number of hours	File duration [s]
SPMAuralB2010	312.6	1807	173.0	1354	2700.05
synthetic	67.2	1600	42.0	1599	3600.00
chagos	259.5	8370	31.0	1555	669.00
argoLOV	11.7	278	42.0	22	290.10
DCLDE2015HF-CINMS_17_B	235.6	268	879.0	166	2250.00
ohasisbio_2015_wker	33.4	2085	16.0	13489	23301.69

Table 2 – Table of PAM datasets used in the doctoral research project, on the data processing side.

label of a whale species over a long time period when you are interested in detecting time occurrences of its vocal sounds.

Objectives & methodological approaches

This doctoral research project globally aims to propose innovative methods to tackle at its source the challenges of weak supervision. Put it in other words, it aims to revisit the questions of how obtaining the "knowledge basis" from which to train AI models, and how developing models that are better suited to the small amount of training data and/or partially labelled data. The first question will be addressed through two "big", consisting basically in increasing the amount and diversity of knowledge sources. Two sources have been included: raw data and expert annotators. Methodological-wise, two main approaches are used more and less conjointly, which can be roughly categorized into i) big-data oriented technological, and ii) AI model development.

The long term motivation underlying this work is to be able to bridge more efficiently the two counterparts of any AI system, i.e. development and validation, towards improved "knowledge"-in-the-loop AI systems, while making them more generalizable.

Big Solution 1: learning from data

Unless specified otherwise, most of analyzed data are high volume, mostly unlabeled, widely available and can be obtained at no cost. For example, audio data retrieved from an underwater recording campaign that lasts several months can reach several terabytes of data depending on the parameters of the recording system. These data cannot be analyzed by a human expert, they should be directly processed by a system to extract useful information. The proposed solution unlocks the potential of open source scalable big data frameworks to efficiently process long-term underwater audio datasets.

Standard operations such as intensive computation of low-level sound descriptors may seem quite "trivial", consisting in aggregating different multimodal observations into a same unity of space in time, but it becomes complex when achieved it at scale with speed and with valuable information that makes sense for expert users.

In this axis, one major concern is the overall quantity (in Bytes) of data involved in each task T. This could help experts design their ecoacoustics systems to extract useful information from raw audio data.

Big Solution 2: learning from experts

To face the weak supervision challenge, experts are involved. To facilitate and make the annotation process faster, an open source collaborative platform is proposed. Experts can share their knowledge by aurally and visually identify acoustic marine mammal presence in an underwater audio dataset.

Moreover, like the marine environment is highly variable depending on the audio sampling site, expertise in physical modeling, especially underwater sound propagation, could help addressing the weak supervision issue by increasing the training data but also to help in migrating deep learning models from one site to another without having to annotate new datasets.

Finally, effort is made to leverage knowledge derived from other research areas to assist in the development of the PAM MMDC.

In this solution, describing to which extent a human expert has been involved in the training and validation of the model used in a given task T is the main aim.

Method approach 1: Big-data oriented technology

As already mentioned above, before big data being an opportunity, it first raises sever technological challenges. Overall, the proposed solutions here have been to import state-of-the-art "big data" oriented frameworks to be able to perform standard management and processing data operations at scale with speed and in a global multimodal context.

Method approach 2: AI model

The predominant modern ML paradigm is deep learning, a representation learning method in which the machine automatically discovers the representations that are required for carrying out a feature detection or classification task using raw input data or spectrogram-based images. The proposed models are designed for mitigating the data-dependence of deep learning for PAM processing.

Doctoral thesis plan

The different configurations of AI development in the different doctoral thesis chapters, from big data unsupervised approaches to collaborative supervision ones, are described as follows:

— Chapter I has been entitled "The cetacean ecoacoustic approach". It is in a high volume data context using the unsupervised analysis from standard soundscape studies. The tasks tackled here are T_1^1: developing operational solutions in a big data context and T_2^1: assessing variability between individual practices. Keywords in this chapter are automatic report content in high-volume PAM datasets, distributed systems on cloud-based multi-user infrastructure, open data science tools, robust Long Term Average Learning Representation.

— Chapter II has been entitled "Putting cetacean expert in the learning loop". In this chapter, current practices for underwater DC (regarding annotation protocol, problem definitions, evaluation frameworks, ...) are reviewed and a Findability, Accessibility, Interoperability, and Reusability (FAIR) dataset for the international community is built to allow AI benchmarks. The tasks addressed here are T_1^4: developing operational solutions to allow for reproducible and collaborative annotations

on PAM datasets and T_2^4: understanding individual labelling variability with inter-agreement metrics between raters. Keywords in this chapter are annotation platform, collaborative annotations, inter-annotator agreement, inter-annotator variability on AI systems.

— Chapter III has been entitled "The marine mammal detection and classification (MMDC) approach". It is a weakly supervised PAM data context. The tasks addressed here are T_1^2: developing operational solutions in a weakly-supervised context and T_2^2: assessing the generalisation and robustness of the solutions on new datasets acquired on a different site. Keywords in this chapter are weak supervision context for MMDC, data augmentation, transfer learning.

Chapter 1

Big Data oriented cetacean ecoacoustics approach

1.1 Introduction

1.1.1 Background on soundscape analysis

Underwater soundscapes can provide information about marine fauna which complements that available from visual surveys (Staaterman et al., 2017). Moreover, marine spatio-temporal trends can be identified at different scales and acoustic variations in this environment can be recognized using PAM (Lillis et al., 2018).

Soundscape ecology is an emerging field of research, but it needs to be based on relevant, accurate soundscape assessments. Several methods have been employed to characterize the underwater soundscapes. Acoustic indices were first used in the assessment of terrestrial soundscapes (Sueur et al., 2009; Bradfer-Lawrence et al., 2019) and then for marine ones (Bohnenstiehl et al., 2018; McPherson et al., 2016; Harris et al., 2016; Lindseth and Lobel, 2018). However, whilst these methods have demonstrated utility in some marine environments, e.g. coral reefs (Harris et al., 2016), this is not universally true (Eldridge et al., 2016). Whilst there is currently no consensus on the relevance of acoustic indices for marine soundscapes, it is generally accepted they should not be used in isolation but in conjunction with other descriptive metrics such as long-term spectrograms (Blondel and Hatta, 2017). Probabilistic methods with different factorization methods, e.g. Principal Latent Component Analysis (Eldridge et al., 2016) or Non-negative Matrix Factorization (Lin et al., 2017b), have been proposed to take into account both time and frequency variations, these approaches can be regarded as extending existing acoustic indices that only consider one of these.

Furthermore, to enhance the analysis of long-term soundscape variations, environmental variables such as wind speed and rainfall, are often used in the analysis. Trends in sound level across years and seasonality patterns can be revealed by correlating descriptive metrics such as sound pressure levels (SPL) and power spectral density (PSD) with wind speed (Ahonen et al., 2017; Erbe et al., 2015; Mathias et al., 2016; Putland et al., 2017; Romagosa et al., 2017), moon phase (Staaterman et al., 2014), temperature (Romagosa et al., 2017; Bohnenstiehl et al., 2016), day / night variations (Romagosa et al., 2017; Freeman and Freeman, 2016; Kaplan et al., 2015), ship traffic (Putland et al., 2017; Romagosa et al., 2017; Gendriz and Padovese, 2016; Viola et al., 2017). In addition to these auxiliary variables, inter-site studies were carried out to observe seasonality or sound patterns (Romagosa et al., 2017; Bertucci et al., 2017; Haver et al., 2017; Marley et al., 2017; Pine et al., 2015; Staaterman et al., 2013).

All the presented methods aim at reducing the time required to analyze large underwater acoustic datasets because usually these are visually and aurally analyzed to quantify sound abundance and diversity which is resource-consuming. For several underwater soundscape studies, visual and aural analysis is still largely used (Ahonen et al., 2017; Pierretti et al., 2017; Putland et al., 2017; Harris et al., 2016; Freeman and Freeman, 2016; Staaterman et al., 2013; Lillis et al., 2018; Bolgan et al., 2016; Picciulin et al., 2016) However, only small effort has been made to propose best practices or guidelines to analyze audio recordings (Bradfer-Lawrence et al., 2019) which makes reproducible studies between sites difficult.

Following our general introduction to the discipline of ecoacoustics in the Introduction, a review of traditional methods in this discipline is presented in Sec. 1.1.2. Then, the main challenge (the “big data” nature of PAM data used in ecoacoustics) of this discipline addressed in this work is introduced in Sec. 1.1.3.

1.1.2 Background on ecoacoustic indices

To characterize the underwater soundscapes, long-term descriptive representations such as root-mean square level (RMS level), octave frequency bands and the acoustic complexity index (ACI) are generally used.

1.1.2.1 RMS level

The RMS pressure level is prone to be impacted by outliers in the ambient noise spectrum (Merchant et al., 2016), but it is useful to describe ambient noise and continuous sounds (Erbe, 2011). It is computed as follows (Erbe, 2011):

$$SPL_{rms} = 20\log_{10}\left(\sqrt{\frac{1}{T}\int_T P(t)^2 dt}\right) \tag{1.1}$$

where $P(t)$ is the amplitude of a pressure time series, T is the time period over which the indicator is computed.

1.1.2.2 1/3-octave frequency bands and third octave level (TOL)

While spectrograms have fine-scale frequency resolution suitable for source identification, 1/3-octave frequency bands are more appropriate to describe distributions and trends in noise levels as they combine noise levels over a standardized frequency range into a single metric (Merchant et al., 2015). Moreover, the MSFD EU Indicator 11.2.1 requests the monitoring of the RMS levels in the 63- and 125-Hz centered 1/3-octave frequency bands to describe low frequency anthropogenic noise (Buscaino et al., 2016).

Center frequencies can be computed in base-two and base-ten. In our computations, only base-ten exact center frequencies were used. It has to be noted that the nominal frequency is not the exact value of the corresponding center frequency. Readers are referred to Wikipedia and ISO (1975) to have the first center frequencies of the TOLs. Center frequencies of the TOLs can be calculated as follow:

$$toCenter = 10^{0.1\times i} \tag{1.2}$$

with i the number of the TOL. In order to determine the band edge frequencies of each TOL, ANSI and ISO standards give the following equations:

$$lowerBoundFrequency = \frac{toCenter}{tocScalingFactor} \tag{1.3}$$

$$upperBoundFrequency = toCenter \times tocScalingFactor \tag{1.4}$$

with $toCenter$ the center frequency of the TOL and $tocScalingFactor = 10^{0.05}$. From (Merchant et al., 2015, Appendix 1) and Richardson et al. (1995), a TOL is defined as the sum of the sound powers within all 1-Hz bands included in the third octave band (third octave band). Mathematically, according to Merchant et al. (2015), Supplementary Material, it can be expressed as:

$$TOL(\text{toCenter}) = 10\log_{10}\left(\frac{1}{p_{\text{ref}}^2}\sum_{f=\text{lowerBoundFrequency}}^{f=\text{upperBoundFrequency}}\frac{P(f)}{B}\right) - S(\text{toCenter}) \tag{1.5}$$

with "B the noise power bandwidth" and "S the correction factor" (Merchant et al., 2015). For computational efficiency, TOLs are computed by summing the frequency bins of the power spectrum that are included in a TOL. Filters with specific characteristics should be designed to compute TOLs with the time-domain signal (ISO, 1975). Note that for accurate representation of third-octave band levels at low frequencies, a long snapshot time is required (sufficient accuracy at 10 Hz requires a snapshot time of at least 30 seconds).

1.1.2.3 Acoustic Complexity Index (ACI)

Acoustic indices are another way to describe underwater soundscapes. More than 60 acoustic indices exist (Bradfer-Lawrence et al., 2019). The ACI is one of the most common and used used as a metric for detecting the possible presence of biological sounds, with the aim of isolating the biophony from continuous anthropophonic and geophonic components of the soundscape (Pieretti et al., 2011; Buscaino et al., 2016). It was computed to calculate the difference in amplitude (I) between adjacent temporal steps (k) using the following formula (Pieretti et al., 2011; Buscaino et al., 2016):

$$ACI_{ij} = \frac{\sum_{k=1}^{n} |I_k - I_{k+1}|}{\sum_{k=1}^{n} I_k} \tag{1.6}$$

"where n is the number of temporal steps (k), i is a frequency bin and j is the interval of time considered. The sum of all the frequency bins (i) and temporal intervals (j) was calculated for every recorded file" (Buscaino et al., 2016).

To allow comparison of the ACI across different sites and audio durations, the sum of the ACI is divided by the number of minutes in the audio file.

The computation of this ecoacoustic index is known to be resource-intensive (Lindseth and Lobel, 2018) and it has been chosen to check the efficiency of our big data framework.

1.1.3 Context and challenges

1.1.3.1 The Big Data challenge

The volume of underwater audio data to process should progress from individual and independent (hand-curated data) to Findability, Accessibility, Interoperability, and Reuse (FAIR) practices. PAM data is opaque and by definition always represents data over time. This makes it difficult to summarize, visualize, or even manually preview individual files (Foote, 1999). Effectively and efficiently describing specific underwater areas as well as large number of audio recordings, collected across different spatio-temporal periods, is challenging.

With the increase of underwater sensors and the large capacity of high-volume data storage, the collection of underwater audio dataset is no longer critical. This presents new challenges. Ecoacoustic research is now focused on how to process such datasets equally managing and analyzing audio data (Truskinger et al., 2014; Frasier et al., 2018). One of the main processing bottlenecks in such high volume of data is the transfer of data that generates a large number of Input/Output (I/O) operations that slow down the computation, and the allocated resources (number of processors, RAM) used for computations (Truskinger et al., 2014).

Put it another way, analysis and management of ecoacoustics has now become a big data problem, as already formalized by other projects (e.g. Ecosounds Acoustic Workbench (Truskinger et al., 2014)) through the 5Vs of big data (Dumbill, 2012; Demchenko et al.,

2013; Hey, 2014) as metrics. Big data challenges in the PAM community are similar to those in the terrestrial ecoacoustic community (Truskinger et al., 2014) (all italicised text are direct quotes):

— Volume: PAM devices now routinely generate datasets that are larger than off-the-shelf software tools or spreadsheet applications can handle. Due to the development of cabled observatories that now provide virtually unlimited power for high bandwidth, continuous data acquisition, and the increase of storage capacity and life battery of temporary recorders, the volume of datasets to process has become larger and larger. For instance, the PerenniAL Acoustic Observatory in the Antarctic Ocean (PALAOA) observatory has been recording quasi-continuously the underwater soundscape of the Southern Ocean since 2005 (Boebel et al., 2006), generating about 140 GB per day (Kindermann et al., 2008), and the Ocean Network Canada has collected more than 300 TB of PAM data in their database (Biffard et al., 2018). In France, governmental agencies like Service Hydrographique et Océanographique de la Marine and Agence Française de la Biodiversité are also experiencing similar challenges of processing large volume of data in the MFSD context, where anthropogenic ambient noise analysis and marine mammal census have to be performed on a long-term continuous effort. The main factors that impact the size of a dataset are (Dugan et al., 2011):

  — *The number of sensors; when trying to characterize a sound over a large area, a distributed network of sensors is required.* The use of hydrophones array enables to record with more than one hydrophone (Reeves et al., 2017).

  — *The recording duration; the total time for which the sensors were recording sound.* With cabled observatories, recording time can be continuous over several years (Biffard et al., 2018).

  — *Sample rate; the number of samples per second at which the sensors are acquiring data.* Depending on the study object, the sample rate will be set differently according to the Nyquist criterion which allows to study sounds up to half the sample rate. For example, for low frequency sounds such as blue whale calls, the OHASISBIO recorder network collects data with a sample rate of 200 Hz (Leroy, 2018) whereas for studying whistles or clicks from delphinids or beaked whales sample rate must be much more higher (Frasier et al., 2018).

  — *Bit depth; the resolution at which the sensor determines the loudness of the sound. The bit depth also determines the dynamic range over which the sound can be measured. The dynamic range is the difference between the maximum value that can be recorded and the minimum value.* The bit depth depends on the recorders used for collecting data. It can reach 32 bits for audio WAV files. One of the most common bit-depth for WAV files is a 16-bit resolution leading to a dynamic range of about 96 dB.

  For example, given a sample rate of 96 kHz, a recording campaign duration of 61 days, more than 505,958,400,000 data points will be generated which makes it hard for usual laptops to process.

— Velocity: *It is defined as the pace at which data is generated or created.* The use of several underwater recorders set with different recording parameters and underwater cabled observatory lead to an exponential growth of data generation.

— Variety: It is defined as the data produced by different sensors used to help analyze audio datasets. For example, for underwater soundscapes, audio data recorded in a

WAV file can be used with wind speed timeseries retrieved in CSV files but also with annotations of the audio also in a CSV formats. In that case, two different formats of data (text and audio) are used. The variety of processing tools such as automatic methods, visualizations can also provide different types of outputs.

— Veracity: The objectivity of the raw data collected from different sensors is an advantage over traditional manual analyse (Truskinger et al., 2014) that involves human perception. When human judgement is involved, a bias and as a result an uncertainty are introduced. Hence, in order to ensure long-term accessibility and re-use of PAM data, there is a need for a consensus with regard to best practice and data format standards between researchers, statisticians, tag manufacturers and database developers.

— Value: In this doctoral research project, the useful information to extract from the high volume of audio datasets is for ecoacoustic research to help protect marine mammal species from anthropogenic activities. Furthermore, in many cases, the data collected is not only valuable for only one objective but also valuable for broader, longitudinal comparisons across studies, species, geographical range and years.

1.1.3.2 The uncertainty/variability challenge

Most of the techniques used to describe a soundscape are prone to uncertainty and variability depending on the settings for the collection of the data, the site characteristics, and the used processing tools. No standardized frequency bands or parameters are used in the underwater soundscape studies. Of more than 50 research papers investigated, no one used the same parameters or methods to analyze the underwater soundscape. One example of such variability between underwater soundscape studies is the ACI where no standardized parameters are proposed and each study computed it with different parameters (Bohnenstiehl et al., 2018; Bolgan et al., 2018).

Eldridge et al. (2016) suggested that existing indices operating in time or frequency domain may be insensitive to the dynamic patterns of interaction in the soundscape which characterize specific acoustic communities and proposed a sparse and shift-invariant Probabilistic Latent Component Analysis as a promising new tool for research. Within the conceptual framework of ecoacoustics, such techniques provide a means to investigate the composition of the acoustic community as a whole in terms of dynamic interactions between spectro-temporal patterns of vocalizing component species, as acoustic communities may be structured according to competition across acoustic niches through spectro-temporal partitioning. In other words, if acoustic niches exist that they do not lie neatly along 1D vectors in the frequency or time domain but dance dynamically across pitch timbre-time space. Frequency-based indices fail to track variation in species richness in the wild is because the defining feature of acoustic communities are global patterns of interaction across a more complex spectro-temporal domain, rather than frequency band occupacy or amplitude variation alone.

Similarly, Lin et al. (2017a) proposed a periodicity coded non-negative matrix factorization for separating different sound sources from a spectrogram of long-term recordings, as an alternative of indices that may be biased when environmental and anthropogenic noises are involved. The PC-NMF first decomposes a spectrogram into two matrices: spectral basis matrix and encoding matrix. Next, based on the periodicity of the encoding information, the spectral bases belonging to the same source are grouped together.

This question is mainly addressed in regard to environmental management of underwater noise pollution (Merchant et al., 2016). Indeed, the current recommendation for the

MSFD is to use the RMS level (Dekeling et al., 2014), but this metric is strongly influenced by outliers in the distribution (Merchant et al., 2012, 2016), and so can be skewed away from the general trend in noise levels by a few high amplitude but unrepresentative events in the time series. Furthermore, the current choice of MSFD frequency bands at 63 Hz and 125 Hz may inadequately reflect the risk of acoustic masking (Hermannsen et al., 2014), and can be contaminated by flow noise (Merchant et al., 2014), and higher frequency bands (e.g. at 250 or 500 Hz) appear to better correlate with broadband levels of shipping noise (Merchant et al., 2014). Merchant et al. (2016) rather recommend the use of percentile-based metrics that also directly related to the temporal distribution of noise levels, making them more appropriate for assessing the risk of acoustic masking (Hatch et al., 2012), as well as being more straightforward to interpret and communicate to policymakers.

Consequently, like for marine mammal conservation, environmental management of underwater noise pollution is greatly constrained by a lack of baseline data on noise levels (Merchant et al., 2016). This limits the ability of managers to make informed decisions at a range of scales, from the regulation of individual developments through to large-scale ecosystem-based management via legislative instruments such as the MSFD.

Finally, the uncertainty/variability challenge is closely related to big data processing. To partially address this challenge, long-term high-frequency recording campaigns, multimodal data sources and the methods validity checking are more suitable (Mooney et al., 2020). This can be done with specific frameworks to handle the huge volume of heterogenous data and to fit the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles.

1.1.4 Contributions

Original contributions in this chapter are organized around the following three actions:

— **Developing & Transferring**: in Sec. 1.2, facing the big data challenge, a first technological chapter aims at i) propose an open source big data oriented system for PAM analytics to handle high volume of audio data with a description of the methods and a tutorial to allow new users to use it, ii) validate the system through extensive computational performance experimentations, in particular on the scalability of the proposed distributed system in the case of feature intensive computation for PAM datasets. Our scientific production is composed of:

  — **OSmOSE report 1** : **Nguyen Hong Duc, P.**, Cazau, D. et al. (2019) "Pushing the standards forward in Underwater Passive Acoustics processing for both theory and code" OSmOSE report 1, arXiv:1902.06659 (`https://arxiv.org/abs/1902.06659`)

  — **OSmOSE report 2** : **Nguyen Hong Duc, P.**, Cazau, D. et al. (2019) "Achieving basic processing of PAM data at scale - with speed" OSmOSE report 2, arXiv:1903.06695 (`https://arxiv.org/abs/1903.06695`)

  — **OSmOSE report 3** : **Nguyen Hong Duc, P.**, Cazau, D. et al. (2019) "Inform, Compute, Visualize, Estimate: a notebook-based processing chain for Underwater Passive Acoustics", OSmOSE report 3

— **Using**: in Sec. 1.4, the main contribution has been to perform a large-scale empirical analysis of the robustness of whale presence indices, and more generally of Long Term Average Learning Representation (LTALR). This is a first important step towards empirical uncertainty modeling of such metrics. Our main contribution in assessing

robustness (or invariance) of LTALR was to define variability metrics w.r.t processing parameters. Our scientific production is composed of:

— **Conference: IEEE OCEANS 2019 Nguyen Hong Duc, P.**, Degurse, A., Adam, O., White, P. R., Gérard, O., Fablet, R. and Cazau, C., A scalable Hadoop/Spark framework for general-purpose analysis of high volume passive acoustic data, IEEE OCEANS 2019, Marseille, France

— **Journal article** : **Nguyen Hong Duc, P.**, Cazau, D., Gérard, O., White, P. R., Detcheverry, J., Urtizberea, F. and Adam, O., (2020) "The marine soundscape off the North Atlantic French Saint-Pierre-et-Miquelon Archipelago", Applied Acoustics, *in revision.*

1.2 Developing: big data system

Below we provide a summary of our published OSmOSE report 2.

1.2.1 Key methods

Currently, computing resources are not great enough to process huge amounts of data in one process. The process needs to be addressed with the Extrac-Load-Transform (ELT) process to divide it into smaller chunks and to be able to handle the amount of data. The proposed distributed computing system relying on ELT and based on the Apache Hadoop and Spark frameworks, as shown in OSmOSE report 2. Details can be found in the paper.

The proposed system is efficient and scalable thanks to either the computing system or the implementation of the computational problems. The use of Hadoop Distributed File System (HDFS) enables parallel access to data by splitting each file into smaller file blocks. This allows the distribution files to different machines and reduce the execution time of the whole processing. This is particularly optimal in the case of files bigger than the size of a block (chosen by the user). The time saving also depends on other parameters such as the number of workers and the number of blocks to process. For example, an audio file is divided in blocks of several seconds that a worker can access

The second element making the proposed system efficient and scalable is specific to the implementation of the FFT-based features. The processing chain was implemented to be "trivially parallelizable". This means that each worker does not need to wait for the result of another worker to achieve its computational task before sending its result to the main program. In that way, network I/O operations, which can be the main bottleneck across distributed computers, are reduced.

Other frameworks could have been used for these tasks such as Message Passing Interface (MPI), MATLAB or the Dask library in Python. However, for MPI and Dask, the entry-level skill requirements for them is much higher than for Apache Spark (Dunner et al., 2017). The latter can be used in Python with PySpark but it was implemented first in Scala. Scala runs on the Java Virtual Machine like JAVA and both can be mixed. This choice was also motivated by the wish to allow other researchers to contribute to the framework. All the codes are available on GitHub.

1.2.2 Key results

To determine the gain in performances with the Hadoop-Spark system, a benchmark was set up between the proposed system, MATLAB and Python frameworks. The experimental set-up is described in OSmOSE report 2. Key results are presented below.

1.2.2.1 Benchmark on a single node system

It was shown that the Hadoop-Spark system performs reasonably well in its standalone mode and reaches the MATLAB system performances in terms of execution time. MATLAB on local computer is one of the most common used frameworks. Although the expected advantage of Apache Spark technology is to scale out processing over several nodes, showing that the proposed big data system is able to be almost as fast as the main framework in the PAM community demonstrates that it is also a valuable tool on a local computer.

1.2.2.2 Benchmark on a multi node system

The speed up of the propose big data system relatively to the MATLAB implementation as a function of the number of nodes for different workloads was evaluated. This is basically the factor of how fast the proposed system with n nodes runs compared to the MATLAB one for a same workload. It was demonstrated that the proposed system scalability was close to the ideal case of scalability for datasets bigger than 200 GB. The ideal case is reached if when you double the number of nodes, your system is twice as fast as the initial one. However, when processing small datasets smaller than 200 GB, the scalability of the proposed system was not optimal. One reason highlighted in the working report is the number of idling executors for small datasets resulting in unused resources.

1.2.3 Contributions of the doctoral research project

The proposed big data system was validated through intensive performance evaluations with common frameworks such as MATLAB. Results made clear the execution time gain of the proposed work compared to other frameworks showing the relevance of such big data system for computing FFT-based features. This will considerably decrease the processing time, allowing to free time for the in-depth analysis and enabling the user to replay and experiment more parameters.

1.3 Transferring: open sciences and standardization

The promotion of the development and use of open data science tools towards a more efficient collaborative research is now commonly pushed forward in ecology (Lowndes et al., 2017; Bradfer-Lawrence et al., 2019; Mooney et al., 2020).

In this doctoral research project, efforts were made towards this goal of open sciences and standardization to promote best practices for processing the data, and for reporting the measurements using appropriate metrics. For this purpose, several working reports were made available online describing the big data framework with complementary results and the details of the implementation of the processing chain. Moreover, another report with how to exploit results is also available through Jupyter notebooks. More details can be found in OSmOSE report 3.

1.3.1 Findability, Accessibility, Interoperability, and Reusability (FAIR) principles

Besides complying to computational performance, our OSmOSE system also aims to comply at best with the FAIR principles. Two principles have been specially addressed in this doctoral research project, namely Accessibility and Reuse, detailed in this section.

1.3.2 Accessibility

One of the main bottlenecks in underwater acoustic studies is the access of the data used. Most of the data are proprietary and not open source. More and more efforts are made to make data accessible possibly by authentication and authorization such as Zenodo, the Cornell Lab audio, DCLDE challenge datasets and Watkins database. Moreover, the International Quiet Ocean Experiment linked underwater audio projects with the associated data.

1.3.3 Reuse

Reuse of the metadata, data and processing methods on audio datasets is one of the goals of the FAIR principles. This principle is composed of two principles: standardization and reproducibility. The former is a way to allow comparisons of results between underwater studies with the use of guidelines, best practices, specifications or requirements for material and methods to fit specific tasks. The latter allows the possibility of results being reproduced, reused, and checked by other researchers (Ivie and Thain, 2018). Enabling reproducibility is not simple. Describing the processing workflow, allowing other researchers to access data are key steps towards reproducibility.

OSmOSE report 1 addresses the need for a common approach, and the desire to promote best practices for processing the data, and for reporting the measurements using appropriate metrics. Having details of the workflow and the implementations in a same place was set up to facilitate the reuse of the proposed system.

1.3.4 Contribution to open sciences

By following these two FAIR Principles, the added value of this effort enables the user to be certain of what, and how, the features are implemented and to be able to perform variability analysis of these features for their specific applications.

Several working reports are available for the big data system proposed in this doctoral thesis. OSmOSE report 1 is the description of the workflow used to compute the different FFT-based features. It compares the different codes in different programming languages giving the benchmark results. OSmOSE report 2 is an extension of the conference paper with more results. OSmOSE report 3 is a user guide that details how to connect to the cluster, to launch a job to compute FFT-based features and notebooks to utilize them.

1.4 Using: big data analysis study cases

This theme provides concrete applications of our big data framework previously described on specific study cases on marine mammals with PAM. Three applications were investigated in this section to show the relevance of developing such a big data system.

1.4.1 Underwater soundscape study case

In this work, an underwater soundscape study near Saint-Pierre-et-Miquelon (SPM) using the developed big data system (more details in soundscape study). The main objective was to determine the contributions of biophony, geophony and anthrophony at this site. It has to be noted that the audio files were contamined by a self-generated noise.

1.4.1.1 Material & Methods

For this study, three years of audio recording was analyzed from two locations off SPM. One was near a corridor vessel lane and the other was near fishing grounds and deeper waters suitable for deep diving cetaceans. The audio data were aurally and visually assessed to determine if a file contained mysticete calls, odontocete calls, ships, rain, self-generated noise, or none of these. Moreover, wind speed and daily precipitation data were also added to the study to understand the contribution of the geophysical processes for SPM.

Different metrics were used for each acoustic source and all metrics were computed using the big data system. For evaluating anthrophony, the 63-Hz and 125-Hz centered TOL were correlated to the expert annotations for each file following the requirements of the MSFD EU Indicator 11.2.1. Expert manual annotations were used for both assessing the presence of marine mammals but also to understand their interactions with other contributors such as anthrophony and geophony (more details can be found in the soundscape study). The ACI was also computed to study if it could be used for shallow water environments and if it is a good proxy for assessing marine mammal vocal presence following the four criteria in (Harris et al., 2016). Wind speed data and daily precipitation were correlated to SPL in 8 octave bands for geophony.

1.4.1.2 Key results

The anthrophony in our recordings consisted of vessels travelling to and from SPM along with distant ship traffic and fishing vessels. Only a small effect size of the two concerned TOL was found with our manual annotations raising the question about the use of these two TOL to monitor for shallow waters (Kinda et al., 2017).

A non-exhaustive sound catalogue was established thanks to the expert annotations. Seasonality patterns were also found: odontocetes were identified, mostly in summer and autumn, on both sites whereas it seems that mysticete acoustic season is winter for SPM. The ACI was shown to be robust to anthrophony and geophony interferences but in our case study it did not describe well acoustic presence of cetaceans. This could be due to the self-generated noise that masked all frequencies lower than 3kHz.

Wind has been shown to be the most dominant abiotic source in the underwater soundscape of SPM. The effects of the Hurricane Igor (2010/09/08 - 2010/09/21) are thought to have been recorded in our underwater campaigns.

1.4.1.3 Conclusions

This first research case study of the proposed big data system demonstrated one possible uses of the big data system by analyzing an underwater soundscape. Underwater long-term audio datasets of a same location allow to monitor changes in the soundscape over the years, which might contribute to develop marine life conservation programs. It was shown that PAM is a good proxy to assess the vocal presence of marine mammal, but it is also a good proxy to determine the effect of anthrophony and geophony on underwater sound levels.

This work is also part of a complete tool chain for assessing the acoustic cetacean presence in audio recordings. Indeed, by monitoring over years a same location, cetacean acoustic seasonal patterns could reveal the presence of different species on specific sites. This helps to build habitat maps and, if needed, marine mammal sanctuaries.

1.4.1.4 Contributions to the doctoral research project

The big data system is used to generate a soundscape. Underwater sound levels are assessed for a long-term PAM dataset. It allows the identification of to identify the main contributors to the SPM underwater soundscape. Anthrophony, geophony and biophony are assessed with specific descriptors that can be computed with the proposed system. The acoustic presence of cetacean is assessed both by expert annotation and the ACI. It was shown that, due to a transient self-generated noise from the recorders, only a small effect size is observed between the annotations and the ACI.

1.4.2 Metric variability study cases

Measured noise levels in PAM are sometimes difficult to compare because different measurement methodologies or acoustic metrics are used, and results can take on different meanings for each different application, leading to a risk of misunderstandings between scientists from different PAM disciplines.

To comply with reproducibility, the understanding of how parameters can change an underwater soundscape analysis is needed. Even if standards were set up to harmonize studies, few studies addressed the effect of parameter choices such as window length analysis and duty cycle. The influence of these settings was analyzed for the detection of killer whale vocalizations (Riera et al., 2013). The authors showed that reducing the duty cycle of 1/3 could lead to missing about a quarter of the detections of killer whale sounds. Thomisch et al. (2015) also studied the influence of duty cycle on the probability of detecting whale vocalizations. They showed that setting the duty cycle lower than 1/10th yielded a probability of detection ranging from 0% to 0.05% for whale sounds in the worst scenario. Stanistreet et al. (2016) agreed that reducing duty cycle could lead to an underestimation of the acoustic presence of cetaceans. Bohnenstiehl et al. (2016) also demonstrated the need to get long-term recording campaigns for analyzing diurnal patterns in marine soundscapes. They used a diurnal sampling strategy over several months instead of using only dusk or nighttime recordings over a few days as usual for snapping shrimps environment. Shabangu and Charif (2020) also highlighted the difficulty to compare studies due to the different duty cycles used to compute call rate of crabeater seals. Hawkins et al. (2014) provided a comprehensive study of the impact of Discrete Fourier Transfrom (DFT) size on the mean spectral levels and they examined the influence of duty cycle in specific frequency bands for the analysis of transient signals. They showed that these two parameters in the computation of spectral level could lead to a difference as high as 4 dB and 6 dB respectively. Harris et al. (2019) also analyzed in the uncertainty on sound exposure level and they proposed a regression model to estimate the uncertainties of the sound pressure levels.

In this section, the big data system was used in order to reproduce the frameworks proposed in Hawkins et al. (2014) to study the influence of the window length of analysis and the duty cycle effect on the FFT-based features.

1.4.2.1 Material & Methods

The OHASISBIO (Observatoire Hydro-Acoustique de la SISmicité et de la BIOdiversité) hydrophone network was deployed in the Southern Indian ocean since 2010 (Royer, 2009). Continuous recordings are sampled at 240 Hz. Only 25 weeks of the dataset were used for this work from January 01, 2015, to June 25, 2015.

Both influences of the window length and the subsampling interval parameters were

assessed in terms of difference, in dB, that could be found in a frequency band for a low frequency dataset. The same protocol as in Hawkins et al. (2014) is followed to compute the mean difference in sound levels. No statistical analysis is performed.

Window length analysis Raw audio data were converted to spectral density using the following FFT sizes: 2400 and 14400 which correspond to 10s and 60s-long window lengths. Mean spectral band levels were computed using an Hamming window with no overlap (Hawkins et al., 2014) to generate sequential power spectrum estimates over the selected week in the following frequency bands: 10 Hz-30 Hz, 40 Hz-60 Hz and 5 Hz-120 Hz. In these bands, one expects to find blue whale calls and seismic activity (Leroy, 2017).

Subsampling analysis The subsampling interval analyses tested the null hypothesis that mean ranks and distributions of data were equal over three subsampling intervals: 1, 30, and 60 min. The analysis was conducted twice using only a 60s-long window length.

1.4.2.2 Key results

Window length analysis For all frequency bands, the mean differences in dB between the two window length was lower than 0.03 dB with a maximum difference of 0.2 dB found in the 10 Hz-30 Hz frequency band. Hawkins et al. (2014) found a maximum difference of 1 dB in mean spectral levels due to the window length parameter. This result demonstrated that the mean spectral levels are not highly impacted by the window length analysis.

Subsampling analysis Differences in dB varied substantially across all frequency bands. The biggest differences were observed between the continuous and the 1 min per hour subsampling strategies in the 5 Hz-120 Hz frequency band. The maximum difference was 15 dB for several weeks for the 15 Hz-30 Hz band and the differences ranged from 2 dB to 15 dB depending on the week and the frequency band. By analyzing the specific weeks that experienced such difference, it was found that blue whale D-calls were observed for a long period but also airguns and more rarely underwater earthquakes.

1.4.2.3 Discussion

Results showed changing either the window length or the subsampling led to a change in the mean spectral levels meaning that data are nonstationary (Hawkins et al., 2014). They also showed that the biggest differences in the mean spectral levels were found when changing the subsampling strategy (Hawkins et al., 2014). The frequency band most affected by the choice of subsampling is the band containing anthropogenic activities (airguns) but also blue and fin whale vocalizations (Royer, 2009; Hawkins et al., 2014; Leroy, 2018; Torterotot et al., 2019). Subsampling only 1 min per hour would have lead to underestimation of the acoustic presence of blue and fin whales which have been studied on these sites for several years now (Royer, 2009; Leroy, 2018; Torterotot et al., 2019) but also to an underestimation of the anthropogenic activities. The sampling site is also prone to underwater seismic activity (Royer, 2009; Leroy, 2017) and subsampling strategies could also lead to underestimation of such underwater events. The 10 Hz-30 Hz and the 40 Hz-60 Hz frequency bands also experienced great differences with a maximum of 14 dB and 10 dB respectively. These bands are also concerned by different blue whale calls.

This work describes the uncertainty that can be estimated in mean spectral level over a 25-week period by changing two parameters of analysis: the window length and the subsampling strategy. The uncertainty evaluated in this study is directly comparable

with the work of Hawkins et al. (2014) as the sampling site, sampling rate and frequency bands are similar. However, greater difference in the frequency bands was found for the subsampling analysis in the present results with a dB difference ranging from 2 dB to 15 dB compared to 6 dB in (Hawkins et al., 2014).

This work highlights another application of the proposed big data system. It allows one to play different uncertainty modeling scenarios by computing FFT-based metrics efficiently, and at scale, on long term datasets. Modelling systematically uncertainty could improve underwater soundscape studies like the one presented in the soundscape study which did not include such analysis. It is believed that more studies could benefit from this system by integrating uncertainties in their sound pressure levels (Harris et al., 2019) to have a better insight according to the variability patterns results. This could result in better suited management regulations for specific underwater sites.

1.4.2.4 Conclusion

This work reproduced the Hawkins et al. (2014) study on a similar dataset. Uncertainty was measured for two different processing parameters (window length and subsampling strategy). Similar uncertainties were obtained compared to the study of reference for the window length, but higher uncertainty was found for the subsampling strategy analysis. All the low-level features of this study were generated using the big data system proposed in this chapter. It is believed that systematically modelling uncertainty in underwater soundscape could lead to better interpretations and solutions for managing marine areas (Harris et al., 2019).

1.4.2.5 Contributions to the doctoral research project

Leveraging the computational performance of the proposed system allows one to study the influence of processing parameters for underwater soundscapes. Having this information is a step towards comparative studies across different underwater sampling sites. This also informs the choice of recording parameters, such a duty cycle in a recording campaign.

1.5 Highlights & Summary

In this chapter, we present and describe one solution to our initial issue on the assessment of marine mammal presence from underwater passive acoustics with deep learning-based methods in a weakly-supervised (but) big data context:

— A specific big data system was proposed to deal with the huge volume of underwater audio data that is still increasing. Its performances when computing the FFT-based features was assessed against common frameworks of the PAM community. It was shown that the execution time was significantly reduced (T_1^1).

— The proposed system should be a first step towards multiple applications such as underwater soundscape analysis, metric variability, and feature extraction for deep learning methods for assessing cetacean vocal activity (T_1^2).

— The proposed system was improved by efforts made to follow FAIR principles of open science with the working reports made available to the community to help them use this system.

Chapter 2

Putting cetacean experts in the AI loop

2.1 Introduction

Before developing the current consequences brought by weak supervision in the PAM community, we first assess in more detail the root of the problem.

2.1.0.1 Causes of weak supervision

The traditional approach to collect annotations in PAM most often involve bioacousticians (with different levels of expertise) who manually annotate the data. Such an approach is currently thought to be the most accurate one (e.g. in comparison to automatic labeling), and always serves as a reference (often referred to as ground truth) for further analysis (Ahonen et al., 2017; Heenehan et al., 2019; Bergler et al., 2019). The counterpart of manual annotation is to be resource intensive, laborious, time consuming, as well as subjective. The study of this latter aspect has only received a regain of interests recently (Leroy, 2018), with the recognition e.g. that annotation in marine bioacoustics may be highly compounded by the intrinsic difficulty in discriminating underwater acoustic sources. Even experts recognize some inextricable ambiguities.

Advances in recording hardware speeds, battery life and data storage capacity have increased the rate of acoustic data accumulation to a point where relying on manual analysis has become unmanageable. Due to the large amount of data to annotate, automatic labelling is sometimes carried out to quickly gather new collections of annotated datasets at the risk that the machine learning model learn to reproduce the behavior of the algorithm generating training labels more than true processes.

The first and most likely cause explaining weak supervision in marine bioacoustics is the difficulty to collect underwater data compared to other datasets such as images of cats and dogs. Collecting underwater audio data is a long process. Recording system are expensive, missions need to be carefully planned and even if an underwater recording is retrieved, it is possible that no cetacean sounds are recorded. Moreover, the second most important cause is the overall difficulty of manual annotation. Evidence of this comes from how often the question around the quality and quantity of training data used in marine bioacoustics has been addressed in the different DCLDE workshop editions (e.g. the DCLDE 2013 discussion panel emphasized the need for more exhaustive and reliable annotation campaigns based on consistent annotation protocols [1].)

Other causes might be related to typical profiles of researchers from the PAM community, coming majorly from signal processing and physical acoustics disciplines more than AI and computer science fields, besides being highly interdisciplinary. Logistical issues around managing (e.g. how to make available huge amount of data to a community) and processing big data are also likely responsible (Chapter 1).

2.1.0.2 Consequences of weak supervision

Direct consequences of the weak supervision context in marine bioacoustics are presented in this section. Before discussing machine learning related issues, labelled data are critical in marine bioacoustics as they serve as a reference (often referred to as ground truth) for many different types of analysis, and are consequently at the basis of complex processing chains where initial errors can be propagated. They are for example directly used as an analytical support of humpback whale song analysis (Au et al., 2006). Given the dominance of large-scale machine learning in marine bioacoustics today, issues related

1. See Summary / Concluding remarks in http://cetus.ucsd.edu/dclde/docs/pdfs/Wednesday/14-Gillespie.pdf and https://www.onr.navy.mil/reports/FY13/mbgilles.pdf

to this field are presented. A first direct consequence of previous effects is the absence of widely used referenced publicly datasets and a reproducible state-of-the-art.

It is interesting to put this into perspective with concrete numbers in comparison to neighbored communities, note that Mobysound, one of the most well-known published dataset of annotated vocalizations of marine mammals, have only a few thousand of annotated sound samples for approximately 20 classification classes corresponding to different species, whereas in the vision computer community a database like ImageNet (a database of labeled images, with labels, such as "cat" or "dog" (Deng et al., 2009)) gathered more than 3.2 million images in 2014. It is noteworthy that in urban acoustics, manually annotated datasets for sound event detection with strong labels (i.e. timestamped labels) are also very limited in size (e.g. the TUT Sound Events 2016 development set is 78 minutes long), although considerable effort is ongoing to improve the situation (e.g., Piczak (2015); Mesaros et al. (2016) , and the Google Audio Set (Gemmeke et al., 2017)). A small number of publicly available datasets for acoustic scene classification exist. For example DCASE 2013 (Stowell et al., 2015) acoustic scene development dataset contains 10 classes, 10 examples of 30 seconds length per class, with an evaluation set of the same size. Another example is the LITIS Rouen Audio scene dataset (Rakotomamonjy and Gasso, 2014) containing 3026 examples for 19 classes, audio of length 30s. But general sound event detection (SED) is only recently picking up pace, as reflected in the availability of general sound event databases, which are more and more numerous Kinoshita et al. (2017); Harper (2015); Eaton et al. (2015); Mesaros et al. (2016); Fonseca et al. (2017); Trowitzsch et al. (2019), and the creation of DCASE.

One may argue that the sound scene and event recognition community has a relatively short history in marine bioacoustics (first DCLDE edition in 2003), however much younger communities than the PAM one have quickly adopted best practices (e.g. Detection and Classification of Acoustic Scenes and Events, DCASE).

Hence, ML development in marine bioacoustics has fallen behind. The paradox is that, in most MMDC systems, to be fully helpful e.g. in aid annotation, they first need a certain quantity of training labels. In most data science domains, deep learning methods have outperformed state-of-the-art methods. However, the main drawback of deep learning methods is the huge amount of labelled data they need to supervise the learning of their networks, and such requirement often limits their applications to many Earth sciences (Reichstein et al., 2019), including marine bioacoustics. Most of them rely on the amount and quality of annotated training data. The process of collecting annotations is thus the main bottleneck in building such methods. Moreover, in music research, it was found that evaluating automatic method performances strongly rely on which expert annotator is taken as reference (Balke et al., 2016). No such analysis was performed for underwater DC tasks.

Having said that, in spite of these constraints and uncertainties, some freely available annotated datasets do exist, such as the dataset from DCLDE workshops which allow participants to directly compare algorithms and methodologies datasets [2] and more recently the challenge Dyni Odontocete Click Classification, 10 species (DOCC10) (Ferrari et al., 2020). But it is now urgent that the PAM community acquires well-established sources for these kinds of "big data" training sets, as it has already been observed that large, publicly available data sets (e.g. Eaton et al. (2015); Harper (2015); Eaton et al. (2015); Mesaros et al. (2016); Kinoshita et al. (2017)) have stimulated a variety of innovative research across forms of acoustics.

A direct consequence of the small amount of freely available annotated datasets is that

2. http://cetus.ucsd.edu/dclde/datasetDocumentation.html.

to compare algorithms is a difficult task, for example they are often only evaluated on one dataset and they are only used for some specific cetacean sound types. No consistent benchmarks have been proposed yet, but some efforts are being made towards this objective (the DCLDE challenges, (Bouffaut, 2019; Ferrari et al., 2020)). In the following, a brief review of DC methods for cetacean acoustic presence is compiled based on the works (Roch et al., 2008; Bittle and Duncan, 2013; Usman et al., 2020). Due to the time and frequency heterogeneity of cetacean vocalizations, many methods have been employed to detect and classify them.

2.1.1 Context and challenge

Automated DC algorithms have become necessary to perform accurate acoustic surveys and improve knowledge of marine ecosystems. These algorithms provide more consistent and comparable estimates throughout a study period and across studies when processing long-term time series. They are less prone to bias than human analysts and can be quantified more objectively. However, they cannot be used without supervision, and typically require performance evaluation or correction at some point in the processing pipeline. For instance, labelled datasets are used for training and evaluation of machine learning models, and misclassifications may need to be quantified or corrected (Marques, 2013). Manual review remains also an important part of the process for additional scientific insights, since analysts are best able to judge the context-dependent nature of biological data. Such a supervision often goes through a manual annotation process by one or several analysts. Obtaining such expert annotations is resource intensive and laborious, especially for long recording campaigns as a reliable annotation often needs a careful listening of sounds. Overall, either the lack or quality level of annotated datasets are now frequently criticized (e.g. Leroy (2018), see last slide from `http://cetus.ucsd.edu/dclde/docs/pdfs/Wednesday/14-Gillespie.pdf`), preventing our community to comply with the best practices in machine learning development, with for example the construction of sustainable reference benchmarking datasets. In the PAM area, the amount of annotations is ridiculously small compared to the overall amount of collected data.

2.1.2 Contributions

Original contributions in this chapter are two-fold:

— **Developing & Transferring**: in Sec. 2.2, addresses the challenge of multi-annotator analysis of PAM datasets, a technological description of the collaborative annotation platform is proposed. Like in the previous chapter, several working reports about this technology are presented. Our scientific production is composed of:

  — **OSmOSE report 4** : **Nguyen Hong Duc, P.**, Cazau, D. et al. (2019) 'APLOSE: a scalable web-based annotation tool for marine bioacoustics" OSmOSE report 4, OSmOSE report 4.

— **Using**: in Sec. 1.4, the main contribution has been to perform a large-scale empirical analysis of the robustness of whale presence indices, and more generally of Long Term Average Learning Representation (LTALR). This is a first important step towards empirical uncertainty modeling of such metrics. Our main contribution in assessing robustness (or invariance) of LTALR was to define variability metrics w.r.t processing parameters. Our scientific production is composed of:

— **Inter-annotator variability Nguyen Hong Duc, P.**, Torterotot, M., Samaran, F., White, P. R., Odile Gerard, O., Olivier Adam, O. and Cazau, D., Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics, *Accepted in Ecological Informatics*

2.2 Developing and transferring: Annotation PLatform for Ocean Sound Explorers (APLOSE), an open source collaborative annotation platform

2.2.1 Related works: Tools for annotation in bioacoustics

Raven In the bioacoustics community, similar development trends can be observed. Highly specialized tools have been developed such as the famous Raven Pro software [3]. It is a prevalent software program for the acquisition, visualization, measurement, and analysis of sounds, used by many bioacousticians to annotate their datasets (e.g. Leroy (2018)). The Pro version (Raven Pro), costs from $50 up to $800 dollars depending on the license type and terms. It allows the user to visualize the sound as a waveform and/or a spectrogram. Multiple parameters may be modified by the user, such as the spectrogram window size-, the contrast, the colorbar. Presets can be created, saved, and downloaded with custom parameters. The user can also zoom in and out and play the sound at different rates. The annotations are made by drawing a time/frequency box (or time only if the annotations are done on the waveform) around the acoustic event of interest. Annotation are then stored in a table with the start and end time and upper and higher frequency values of each "box". Multiple annotation tables can be filled at the same time when annotation is performed on numerous call types simultaneously. The Lite version (Raven Lite) is free and provides the basic functions of Raven Pro, but does not allow for advanced and customizable control of spectrogram parameters and advanced sound measurements and annotation.

Audacity This is an open source software [4], also offering many features for sound annotation. The user needs the dataset in local to be able to perform audio transformation and annotation. Both annotation in time and frequency are available in Audacity.

Koe This is a web-based annotation tool and application for classifying and analyzing animal vocalizations (Fukuzawa et al., 2020). Koe offers bulk-labelling of units via interactive ordination plots and unit tables, as well as visualization and playback, segmentation, measurement, data filtering/exporting and new tools for analyzing repertoire and sequence structure in an integrated environment.

ARBIMON It uses OpenLaszlo for visualizing, listening and annotating audio recorders which enables compatibility across several web browsers. All users can help in annotating and the platform has a dedicated space for citizen scientists. All raw data are first sent to a program to compute spectrograms Aide et al. (2013). Annotation help can also be performed with a program that spots high energy regions in the spectrograms. Once the dataset is annotated, the user can perform vocalization identification with Hidden Markov

3. http://ravensoundsoftware.com/software/raven-pro/
4. https://www.audacityteam.org/, see manual at https://manual.audacityteam.org/man/spectrogram_view.html

Models. Moreover, deep learning methods to help the annotators are under development and notebooks are available. Meanwhile, a pattern matching algorithm enables the user to extract and classify several audio segments using Random Forest models.

Manual analysis A traditional approach in underwater bioacoustics to annotate an audio file is to use a software to listen and visualize spectrogram such as Audacity, Adobe Audition and annotating using a spreadsheet.

In the following section, a brief description of APLOSE can be found. A full description of the system is presented in OSmOSE report 4.

2.2.2 Key components and features of APLOSE

APLOSE is an open-source, web-based tool programmed in JavaScript with React and Node.js libraries. The front-end is heavily inspired by the extended version of wavesurfer.js used in the CrowdCurio project. Mozilla Firefox is full-featured whereas Google Chrome does not support the sound playback function yet.

2.2.2.1 Spectrogram generation

A custom Python code is used in the current version of APLOSE to generate spectrograms. This code (available on GitHub https://github.com/ixio/ODE-Scripts) performs classical audio pre-processing operations such as filtering, amplification of raw audio signals. This enables the user to specify the frequency band of interest depending on the vocalization to identify. To efficiently compute spectrograms, raw audio data are chunked into smaller segments with a duration defined by the user.

In this doctoral research project, each spectrogram was stored as an image file called a tile. All tiles can be pre-generated at the beginning of an annotation campaign or they can be generated on the fly and cached in memory. This method of tile-based rasterization was motivated by its successful usage for geospatial map applications where panning and zooming was achieved without a loss of image resolution. This method is also thought to be at the core of the Pattern Radio, Google [5]. For each zoom level, different tiles are computed and concatenated to provide the whole spectrogram.

Once all the tiles are computed, the lag time to display the spectrogram depends on the internet connection speed of the user.

2.2.2.2 APLOSE Features

Several features are available on APLOSE. It is a web-based platform enabling several annotators to annotate the same files. Zooming in and out is allowed and either the spectrogram or waveform representation can be annotated. When annotating a spectrogram, a 2D box in time and frequenye can be drawn. Moreover, a speed up parameter can be adjusted to listen to low or high frequency vocalizations because marine mammal sounds can be infrasonic or ultrasonic. Campaign instructions are available at any time during a campaign. In this way, the campaign administrator can share audio examples of the sounds to identify. Annotating can be tedious and time-consuming, annotations made for a task are stored and the user can exit an annotation task without losing all the annotations he/she performed. At the end of the annotation campaign, two CSV files are available for the campaign administrator: one containing all the annotations performed by the annotators and another with the time spent on each annotation task by the annotators.

5. https://medium.com/@alexanderchen/pattern-radio-whale-songs-242c692fff60

2.2.3 Transferring: open sciences and standardization

Like chapter 1, efforts were made towards FAIR principles by making available detailed technical reports on APLOSE and sample codes to analyze results from an annotation campaign. OSmOSE report 4 presents APLOSE in more details. A link to a demo version is available in this report enabling anyone to try it. Some datasets are already available for annotation. A Jupyter notebook for simple metrics is available on the OSmOSE Github at `Project-ODE/ODE-Scripts/blob/master/APLOSE-Simple%20Metrics.ipynb`.

2.3 Using: an annotation platform to address scientific questions

2.3.1 Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics

This work aims at pursuing recent efforts (Sirovic, 2016; Leroy, 2018; Torterotot et al., 2019) in better understanding inter-annotator agreement within collaborative annotation campaigns in marine bioacoustics. For this purpose, a collaborative annotation campaign was performed on the DCLDE 2015 low frequency dataset, involving 6 annotators with different profiles, in addition to the two experts who originally annotated this dataset for the DCLDE 2015 challenge. The main objective of this work was to show how strong inter-annotator variability was regarding different potential sources of variability. The full version of the article can be found Inter-annotator variability.

2.3.1.1 Material & Methods

Material APLOSE was used as the annotation software. The DCLDE 2015 low frequency dataset was used in this study. Two whale calls had to be identified by the annotators: blue whale (*Balaenoptera musculus*) D-calls (Thompson, 1965) and fin whale (*Balaenoptera physalus*) 40-Hz calls (Watkins, 1981) identification. This dataset has already been annotated by two annotators on the occasion of the 2015 DCLDE challenge. The annotation file proposed by the challenge committee consists in the fusion by majority voting of these two individual annotations. For this work, only 50 consecutive hours of the CINMS18B file, recorded within the Channel Islands National Marine Sanctuary and starting on the 23rd June 2012 was used by the 6 annotators of the collaborative annotation campaign.

All the annotators could refer to a campaign guide with audio and visual examples of spectrograms of the two calls to identify. They could choose among 3 labels: D-call (for blue whale D-calls), 40-Hz call (for fin whale 40-Hz calls) and Unknown call. Annotators were instructed to use this latter label in case of doubt between the two call types, but not to annotate a call type from an unknown source. They were grouped into three categories depending on their annotation experience: neophyte (no annotation experience at all), biacoustician (one that has already annotated underwater audio recordings but not the two concerned calls) and expert (one that has already annotated blue whale D-calls and fin whale 40-Hz call).

Methods The inter-annotator agreement was assessed using the Fleiss κ score(Fleiss, 1975; Zapf et al., 2016) as all annotators were considered as equally important. It is evaluated using only pseudo-presence observations resulting from the annotation process

meaning that only events annotated at least by one annotator are considered in the agreement evaluation. The agreement on pseudo-absence was not evaluated.

To better understand the inter-annotator agreement variability, this work focused on three potential sources: the Signal-to-Noise Ratio (SNR) of the identified calls, the time that each annotator spent to annotate each audio file and the annotation behaviors (how many annotations differs between the annotators).

2.3.1.2 Key results and discussion

The inter-annotator variability heavily depends on the call type. The fin whale 40-Hz call showed the worst inter-annotator agreement between the two calls to identify with a Fleiss κ score near 0. The similarity between the blue whale D-call and fin whale 40-Hz call signatures could explain this result.

This annotation campaign also revealed an annotation behavior for low frequency calls. The time spent on each annotation task was shown to be smaller than the duration of the audio files revealing that the annotation of low frequency calls was visually performed instead of both aurally and visually.

This annotation behavior could explain that the salience of the calls, partially measured by the SNR of the annotated calls, could be a potential source of inter-annotator agreement variability. Most of the annotators visually identify calls, the higher the SNR was, the higher the Fleiss κ score. This was only verified with the blue whale D-calls in this study, but other work has observed the same result for different type of calls (Urazghildiiev and Clark, 2007; Leroy, 2018). For the fin whale 40-Hz calls, the Fleiss κ score was already low and the influence of the SNR on this call was not considered.

It was also shown that the annotation experience was not necessarily a source of variability as the neophyte in underwater audio annotation had a similar annotation pattern to 2 of the experts. More surprisingly, one expert had a divergent annotation behavior compared to the other 2 experts. This confirms that even neophytes in annotating underwater audio recordings can provide high-quality annotation labels, which has already been observed in other research areas (Rogers, 2003; Snow et al., 2008; Snel et al., 2012; Hantke et al., 2016). More investigations on other cetacean sounds should be carried out because this result might not be valid for odontocete clicks or unknown calls.

Finally, this work highlighted that underwater audio annotation is highly dependent on the human perception and judgment.

2.3.1.3 Guidelines for future annotation campaigns

These results enabled us to propose a few guidelines on how to set up an annotation campaign in marine bioacoustics. First, the need to involve more than one annotator (at least two annotators) is suggested to allow for a minimal sanity check to inform the user of the future created dataset on the difficulty level of the annotation task with respect to the different call types to be annotated. Annotating is a resource-consuming activity, experts or neophytes who wish to take part in an annotation campaign without being financially compensated are rare but even neophytes (e.g. citizen scientists) should not be rejected for annotating underwater audio recordings.

Another guideline would be to assess the inter-annotator agreement to be able to evaluate the reliability of an annotation task and to find the minimal number of annotators required for that annotation task. The use of the Fleiss κ score is one possible solution to achieve the first objective while the Δ Fleiss κ can be used to determine the minimal number of annotators.

Annotation is tedious and can be rushed, to keep the annotator focused on the task, short duration annotation files having an open-ended question (e.g. identification of time and frequency of whale calls instead of a yes/no question about the presence of a specific sound in an audio clip) is recommended. To make the annotation easier, a closed task such as annotating a small audio sample (max 10s long like in the DCASE challenge) with predefined labels could be a good alternative.

At the end of each annotation campaign, as suggested in Lee et al. (2018), a quality control could be set up with experts to review labels that divide annotators. Both annotator quality and ground truth inference strategies for label aggregation could be determined during this process.

It also seems that high SNR calls are easier to annotate and present higher inter-annotator agreement. This raises the question about the need for integrating low SNR calls in studies.

Finally, the need to systematically perform inter-annotation variability studies prior to development of machine learning methods and validation is necessary, especially due to the renewed interest for the DCLDE challenge datasets (Socheleau and Samaran, 2018; Guilment et al., 2018; Shiu et al., 2020) that should now act as reference datasets for the PAM community. It is believed that such needs will require collaborative open tools and datasets and APLOSE and the DCLDE challenge datasets are a first step towards this goal.

2.3.1.4 Conclusions

This work presents a new annotator subjectivity dataset of two cetacean call types based on the DCLDE 2015 low frequency dataset. Each annotator has a distinct annotation behavior (annotation experience, annotation time, number of identified acoustic events). Three potential sources of inter-annotator agreement variability were investigated. It was shown that the annotation experience was not a prerequisite to providing high quality labels for this specific annotation campaign. The SNR and the type of call to identify played a key role in the inter-annotator agreement score. The results also highlighted that subjectivity was the decisive factor in annotating underwater sounds.

Finally, this annotation campaign can still be joined by anybody, and the annotation results will be automatically updated. New annotation contributions from the community would provide even stronger experimental evidence of our findings.

2.3.2 Assessing the variability of the annotations on DC systems and their evaluations

Automated methods of underwater sound recognition are a necessary step to develop our knowledge on the underwater environment that can improve ecosystem management. Supervised artificial intelligence techniques rely on the amount and quality of annotated training data. To evaluate the performance of such methods, they must be developed and tested on reference annotated datasets. Consequently, the process of collecting human annotations is the main bottleneck in building such databases. In PAM, the number of annotations is very small in relation to the overall amount of collected data.

Annotation task is resource intensive and laborious. It is also due to the intrinsic difficulty in discriminating underwater acoustic sources in comparison for example to daily visual objects (cats, dogs, chair, ...). The traditional approach to collect annotations in the PAM scientific community is through manual annotations by bioacoustician experts.

Such an approach is likely to be the most accurate solution (e.g. in comparison to automatic labelling) and has allowed publicly available labelled datasets to be built (e.g. DCLDE datasets, Orca sounds, Watkins database and MobySound). However, one of its most reported drawbacks in the literature is being highly resource intensive and time consuming. This major drawback justifies the current expansion of ML approaches as already mentioned.

Another drawback of manual annotation, much less analyzed in our community, is that it can be error prone, with potential annotator-specific biases for multiple-annotation campaigns. Although almost all PAM studies mention the use of several annotators, only a very few recent works either addressed this question as a product of their study, Leroy (2018) and the work presented above being the only exceptions to the best of our knowledge. As a result, annotated data is often unreliable as indicated by the poor agreement between the annotator and the same annotator analyzing the same soundsome time later (Leroy, 2018).

This work presents the consequences of annotator subjectivity for underwater classification system performances. This issue was investigated for music research in Balke et al. (2016). The authors showed that the performances of automated methods were highly dependent on the expert that annotated the dataset taken as the ground truth.

2.3.2.1 Material & Methods

Dataset The collected annotations from inter-annotator variability was used in this analysis.

Clustering aggregation methods Methods from inter-annotator variability were used for clustering the aggregations methods.

Neural network for deep learning experiments In these experiments, "Unknown call" and "None" labels are not considered because DCLDE experts did not use these. Only 3 classes were considered in this set up: D-call, 40 Hz-pulse and Noise.

Training and test sets After labels from all annotators are collected, audio events for each are extracted. Annotators do not annotate the same events, the size of the training and testing sets differ across annotators. A stratified 5-fold cross-validation is used to keep the proportions of the imbalance dataset. The training size is set to 80% of the annotator dataset. For each annotator, a ResNet 18-layer model (ResNet18) (He et al., 2015; Schaetti, 2018) and an extreme gradient boosting model (LightGBM) are trained from scratch. The focus of the paper is not to find the optimal parameters for a specific neural network architecture, the aim is to show difference in performance for a given architecture. The number of epochs and the batch size are set to 20 and 128 respectively when using ResNet18. Default parameters are kept for training the LightGBM (lgbm) classifier. Balanced accuracy score is used to evaluate the performances of the classifiers to take into account the fact that fin whale 40-Hz calls are underrepresented in each dataset.

Pseudo absence (noise samples) generation Noise samples are extracted in files where all annotators have not identified any calls. Consequently, noisy samples come from only one site for a specific period. For each file identified as having "no activity", 3s long random noise samples were extracted. They could overlap.

Features Annotated acoustic events are bandpass filtered between 15 Hz and 150 Hz using a third-order Butterworth filter. Magnitude spectrograms are extracted with temporal and frequency resolutions of 30 ms and 0.48 Hz, respectively. Inputs to the neural network are images of size 290×96. Audio samples are zero-padded to be 3s long when necessary. If the annotated acoustic event is longer than 3s, the first 3 seconds of the signal are extracted. This choice is motivated by the mean durations of fin whale 40-Hz and blue whale D-calls which are about 1s and 3s, respectively.

Label aggregation methods For aggregating labels, the "Unknown call" label is taken into account in analyses. Another label is also added when at least one annotator identifies an acoustic event whereas others do not. The label for the others is "None". However, if the resulting aggregation gives an "Unknown call" or "None", the annotation is removed.

— Majority Voting (MV): This is the simplest label aggregation method and yet one of the most popular. It considers all worker annotations as equal. The true label is estimated by the label with more votes. It is usually used as a baseline (Kairam and Heer, 2016).

— Fast Dawid-Skene (FDS): Another popular approach in label aggregation is the Dawid-Skene algorithm relying on the Expectation-Maximization principle. It chooses a label based on an estimated quality of an annotator. In this study, an efficient version of the Dawid-Skene method with similar performances is used: the Fast Dawid-Skene (FDS) algorithm (Sinha et al., 2018).

Time differences between overlapping annotations The time differences between the overlapping annotated events were computed. Only annotations when at least two annotators identified a call are considered. Time differences between all annotators are then computed as well as associated descriptive statistics.

2.3.2.2 Key results

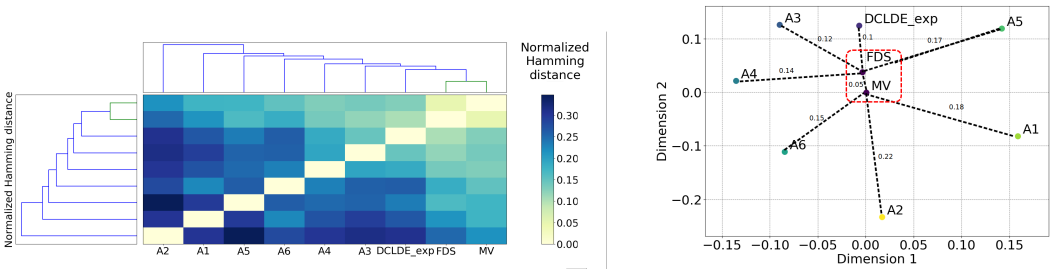


Figure 2.1 – Left: HAC with heatmap representing Hamming distance metric between each pair of annotators. Hierarchical trees are shown on the upper left of the heatmap. Right: Divergence between annotators plotted using multidimensional scaling. The Hamming distance between some annotators was displayed on the dotted lines. "Unknown call" label was not considered in both figures.

Aggregation methods in the annotation behavior analysis With the new cluster analysis in Fig. 2.1, it is shown that the FDS and MV aggregations methods are grouped in the same cluster. The Hamming distance between these two methods is 0.05 meaning that they disagree on about 70 annotations. The annotator which seems closest to the cluster is

the DCLDE experts with a Hamming distance of 0.1 and 0.13 to the FDS and MV methods respectively. Both methods follow more closely the DCLDE expert annotations instead of other annotators. It has to be noted that the MV method is closer to all annotators, except the DCLDE_exp and A3 than the FDS by almost 0.04. For DCLDE_exp and A3, the Hamming distance is greater than the FDS one by 0.02 and 0.01, respectively.

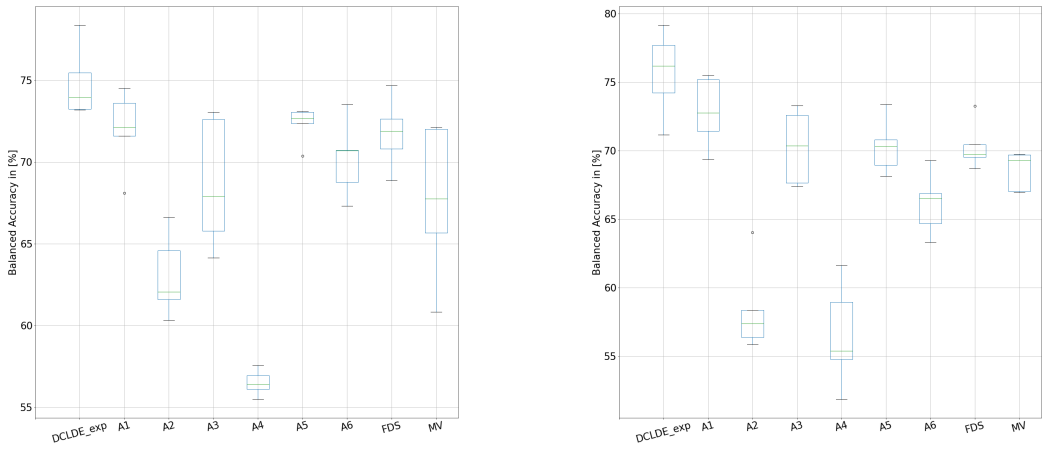


Figure 2.2 – Balanced accuracy for the lgbm (left) and the Resnet18 model (right) for each annotator and aggregation method.

Accuracy of the models for each annotator Fig. 2.2 represents the balanced accuracy for each machine learning method. Results are highly variable depending on the annotator annotations. A gap of almost 20% is observed for both machine learning methods. Depending on the training dataset, the lgbm model has a higher median balanced accuracy score than the Resnet18 one (A2, A4, A5, A6 and FDS) or lower (DCLDE_exp, A1, A3, MV). Interquartile ranges are also highly dependent on the dataset used as reference.

Outliers are found for A1, A2, A5 and FDS showing the dependency of the supervised methods on the training set like balanced accuracy can be either reduced or improved. This also shows that the identified acoustic events are not easy samples. Identified acoustic events could be noisy or ambiguous which can make them hard for the model to learn.

Great interquartile ranges of accuracies reflect the high heterogeneity in the annotated acoustic events for each annotator. These range from less than 1 to almost 10 and they are different for each model like different characteristics of the input data are captured by each which can lead to better or worse discrimination of the call type.

Note that FDS and MV show similar balanced accuracy scores for the Resnet18 model with a median near 70% while for the lgbm model, the gap between the two aggregation methods is about 5%.

Time differences between annotations Table 2.1 shows descriptive statistics for the time differences in the start and end times of the annotations. It shows that all 2D annotation boxes in time are offset of about 0.6s and 0.9 for the start and end times, respectively. In the campaign guidelines, it was required to make the 2D box annotation

as close as possible to the identified acoustic event. This result shows that in average the time differences are less than 1s. However, maximum value of the time differences are 3s and about 6s. These highest values may reflect a lack of concentration from the annotator as the task is tedious.

Time differences for start / end times [s]	start	end
mean	0.61	0.86
std	0.47	0.67
max	3.28	5.97

Table 2.1 – Descriptive statistics for time differences for start and end times.

2.3.2.3 Discussion

This work presents an analysis of the direct consequences of the subjectivity of annotating underwater audio on label aggregation and machine learning methods.

Regarding the label aggregation methods, it is shown that the DCLDE_exp annotator strongly influences both methods as suggested by the small Hamming distance to them. DCLDE experts are already a majority voting between 2 annotators, the resulting annotations could be easy signals with high SNR or not ambiguous. By computing the label aggregation following the collaborative annotation campaign, the annotations found by the two DCLDE_exp and A3 experts could be the ones with high inter-annotator agreement. This could explain why the FDS method gives a higher supposed reliance and is closer to them. For the MV method, if only acoustic events with high inter-annotator agreement were the ones found by the DCLDE experts, this could also explain the closeness of them to the MV method. However, as MV gives equal importance to all annotators, if a majority disagree with the DCLDE experts, the majority label will be given to the identified acoustic event which could explain a greater distance between the DCDLE_exp annotator. Moreover, by giving the same weight in the final decision, all the Hamming distances to the MV method are smaller than for FDS except for the annotators that most influenced the FDS method (DCLDE_exp and A3).

Moreover, the results show also high variability depending on the supervised methods and the reference datasets used for training. Differences in the balanced accuracies between two different training datasets could be up to 20% for the two models tested. This result is observed in music research where the performance evaluation of automated methods is also highly dependent on the dataset used for training and testing (Balke et al., 2016). This could also reveal that some annotators identified faint acoustic events which was not requested in the guidelines. Furthermore, depending on the used model for classifying the identified acoustic events, up to 5% of differences can be found. The highest performance of the DCLDE_exp annotator could be because easier samples to classify are fed as inputs into the supervised methods. However, A3 has a similar annotation behavior to the DCLDE_exp, intuitively it could have been believed that similar results could have been found. This is not the case with a difference of more than 5% in the balanced accuracies. Substantial variations between the different training datasets and methods used are found for this annotation campaign.

The use of a classifier to help discriminate sound type and label aggregation method analysis could also be a way to assess the reliability of annotations and it can be complementary to an inter-annotator agreement score. Using label aggregation, annotators that are too far from the label aggregation method could be considered as noisy annotators.

This also raises the question of the existence of a unique "ground truth" (Balke et al., 2016). However, maximizing annotator inter-agreement during the dataset creation may also result in poor generalization capabilities of models and they could only be able to recognize stereotyped calls after such an operation.

This work also investigated the start / end time differences between annotators. It has to be noted that the maximum time resolution of that could be displayed by the annotation platform was 40s which could explain the small start / end time differences. Knowing such a parameter could help in designing performance evaluation metrics. For example, Serizel et al. (2020) used a tolerance of 0.2s for start times and a varying tolerance for end time with a minimum tolerance of 0.2s. In this work, annotators show greater differences in start and end times with more than 0.6s. The tolerance for start and end times also depends on the predefined length of the supervised method inputs. All these parameters should be taken into account before setting hard tolerances for performances evaluation.

2.3.2.4 Conclusion

The present investigation shows that automated methods are directly impacted by the subjectivity of the annotators either in their design or in their evaluation. Moreover, label aggregation methods are also biased by the annotator perception and judgment. It is believed that unlocking the use of collaborative annotation campaigns for underwater audio recordings will help in building reference dataset for the PAM community with not only a single "ground truth" but multiple. It could help build more robust automated methods. All this work is made possible thanks to APLOSE.

2.3.3 Reducing the annotation effort: few-shot embedding and active learning

2.3.3.1 Few-shot learning: Siamese networks

The main bottleneck in supervised automated DC methods is the need of annotated data. Few-shot learning (FSL) aims at generalizing to new data using fewer training samples. In this doctoral research project, the focus is on Siamese networks to learn useful embeddings with a smaller number of training samples.

FSL refers to the methods that train a neural network model with small amounts of training data. The idea is to try to obtain an accurate model and able to generalize to unseen classes from small to very small training datasets (Wang et al., 2020). This is of particular interest in PAM to reduce the annotation effort and to generalize to datasets from different sampling sites.

FSL is usually found to train on a very small amount of annotated data. One-shot learning refers to the case when only one annotated sample for a category is available for training. This number is defined by the user and zero-shot learning also exists (Wang et al., 2020). The term N-way K-shot learning is also found which means that for N different classes, K annotated samples are available for the training of the model.

In this work, FSL is used as a feature extractor to be able to differentiate between two imbalanced classes: blue whale D-calls and fin whale 40-Hz.

For the feature extraction model, a Siamese network is used. This kind of network was successfully used in different domains such as character recognition (Koch et al., 2015), audio embedding extraction (Manocha et al., 2018) and object tracking (Rahul et al., 2017).

A Siamese network is made up of two models sharing weights. Two inputs are fed into this architecture. One of these is a reference image and it is compared to another one in the training dataset. The model embeds both inputs into a lower-dimension space and computes a similarity metric (e.g. Euclidian distance) to discriminate if yes or no the two inputs are similar. This can be used for classification tasks but also as a feature extractor.

In sound classification, Siamese networks can be used as a feature extractor. In Zhang et al. (2018), they used a Siamese style convolutional neural network to be able to search sound by vocal imitation in a a database. They outperformed the state-of-the-art method by 2% to 20% depending on the class. In Nanni et al. (2020), bird and cat sounds are classified. The Siamese network is used to learn a dissimilarity space from prototypes obtained by clustering methods. A support vector machine (SVM) then classifies the sounds. Even if it did not outperform best models proposed on the two datasets, their performances were 1% and 4% less than the best models for the bird and cat datasets respectively. Their model does not need the optimization of hyperparameters whereas one of the best classifiers for the bird dataset does require such an optimisation.

2.3.3.2 Active learning (AL)

The promise of AL is that by iteratively increasing the size of our carefully selected labeled data, it is possible to achieve similar (or greater (Ilhan and Amasyali, 2014)) performance to using a fully supervised data-set with a fraction of the cost or time that it takes to label all the data. AL is considered to be a semi-supervised method, between unsupervised and fully supervised in terms of the amount of labeled data, i.e., for unsupervised data we use 0% labelled samples and for fully supervised we use 100% labeled samples. Therefore, the decision of how much data to use or alternatively how much performance is required from the model relies on a resource management decision.

The maximum number of labels that can be assigned is called a labeling budget, which is used to quantify a limited annotation effort. When labelling budget is small, there are two established techniques to utilize the abundant amount of unlabeled data: active learning and semi-supervised learning.

The applicability of AL to real-world problems remains an open question. While in supervised learning, practitioners can try many different methods, evaluating each against a validation set before selecting a model, AL affords no such luxury. Over the course of one AL run, an agent annotates its dataset exhausting its labeling budget. Thus, given a new task, an active learner has no opportunity to compare models and acquisition functions.

In our work, we evaluate the applicability of AL for PAM annotation. AL aims to maximize the performance of a classifier while minimizing the number of instances requiring manual annotation (Settles, 2012). AL is a human-in-the-loop process in which informative instances are automatically selected at each iteration for manual annotation by a human expert. Informative instances are those that carry additional useful information: when annotated and added to the training set, they lead to a classifier with a higher performance compared to adding other instances. Therefore, AL builds a training set in an iterative process with less manual annotation effort through the smart selection of instances. AL has been previously applied to audio tasks such as automatic speech recognition (Riccardi and Hakkani-Tur, 2005), speech emotion classification (Zhang and Schuller, 2012), audio retrieval (Roma et al., 2012), and sound classification (Shuyang et al., 2017). Only one study has explored active learning for classifying bird sounds (Qian et al., 2017). Their experimental results showed that active learning, with kernel-based extreme learning machine as the base learner, reduces up to 47 % of the number of instances requiring manual annotation, to reach 80 % unweighted average recall. It was also recently explored for

classifying sounds in long-duration recordings of the environment (Kholghi et al., 2018). These authors developed an active learning framework for classifying major categories of sound sources (i.e., bird, wind, rain, ...) contributing to the soundscape derived from 13-month continuous recordings of the environment. The main purpose of the framework is to study the potential of AL in reducing the manual annotation effort when describing the content of the recordings by their dominant sound source.

The key element of a pool-based AL process is the query strategy (also known as sampling criterion) which typically selects instances for annotation based on the useful information each instance contains for the classifier.

Though other types of active learning methods exist, only certainty-based active learning (CRTAL) (Cohn et al., 1994) has been studied in the field of acoustic pattern recognition. It has been proposed to speech recognition in Hakkani-Tür et al. (2002). As is shown in (Riccardi and Hakkani-Tur, 2005), as long as less than 10 % (about 3000) utterances are labeled, performance of CRTAL is behind random sampling. An ideal way to deal with a small labeling budget is to utilize the internal structure of the dataset so that the method starts to outperform random sampling from the very beginning of a labeling process. Shuyang et al. (2017) propose a method to optimize the sound event classification performance when labeling budget is limited and only a small portion of data can be annotated. The proposed method is called medoid-based active learning. K-medoids clustering is performed on sound segments, and the centroids of clusters (medoids) are selected for labeling. The medoid label is then propagated to the entire cluster to enable a fast labelling of the data.

In this context, we wish to combine recent advances in few-shot learning and the active learning in a practical way for PAM. Through our experiments, we wish to show that our framework can accurately learn annotator expertise, infer true labels, and effectively reduce the amount of annotations in model training as compared to state-of-the-art approaches.

2.3.3.3 Material & Methods

Dataset The 2015 low-frequency DCLDE dataset was recorded with High-frequency Acoustic Recordings Packages deployed off the southern and central coast of California at different locations, spanning all four seasons, over 2009-2013 period (see the full dataset documentation at `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`). The sampling rate is 2000 Hz for recorders Channel Islands National Marine Sanctuary and 3200 Hz for Diablo Canyon Power Plant recorders, and depths of 600, 65 and 1000 m, respectively. As a consequence, iablo Canyon Power Plant recordings were resampled at 2000 Hz. As this dataset was used in the DCLDE international challenge on detection and classification of marine mammals in 2015, it has already been annotated by two independent experts, with a total of 5211 strong labels (i.e. with start and end times of events) over 2 whale species classes that are highly unbalanced: blue whale D-calls (4796 samples) and fin whale 40-Hz calls (415 samples).

Inputs The same inputs as in Sec. 2.3.2.1 were used.

Neural network architecture: FSL and AL A ResNet is a kind of deep neural network using skip connections or short-cuts which jumps over some layers (He et al., 2015; Schaetti, 2018). ResNet models were introduced in 2015 and won several competitions in computer vision. The motivation behind skipping layers in artificial neural network is to avoid the well-known problem of vanishing gradients using activation from a previous layer

until the next one has learned its weights. Usually, ResNet architectures are designed with a hundreds of layers because its structure enables to successfully deal with the vanishing gradient issue when training a neural network with a lot of layers. In our study, only 18 layers were stacked in the ResNet to avoid overfitting as the training set is not very large. It was trained from scratch to handle the size of the spectrogram images (290×96) instead of the initial shape of 224×224. This is the whole model which is retrained in the active learning process. Our implemented version is based on existing open source codes [6]. A baseline performance with the whole dataset was computed. For the baseline, the number of epochs and the batch size were set to 20 and 64 respectively and an Adam optimizer with a learning rate of $1e^{-3}$.

Moreover, for the FSL neural network architecture, one dense layer is added to the ResNet model that shares weights between the two inputs to make the final decision. For the classification decision, a simple dense layer is added.

FSL process Six training sets are built with 100, 200, 300, 400, 1000, 2000 samples. For each training set, the Siamese network is trained to learn the embeddings. A new model is trained when the Siamese network ended its training. This procedure is longer than the training process where only one network is trained. This method is compared to the training of the ResNet alone. For each training set, a new model is trained from scratch unlike the AL process. The batch size of the Siamese network is 64, an Adam optimizer with a learning rate of $5e^{-5}$ and the number of epochs is set to 100. For classification, these parameters are a batch size of 64, an Adam optimizer with a learning rate of $1e^{-3}$ and the number of epochs of 100. This model is compared to a baseline of the ResNet model with no optimization but the same number of training samples, number of epochs, batch size and optimizer. The focus was not to find the best baseline architecture but to show the efficiency of the Siamese network approach to better generalize to new unseen data in a specific configuration.

AL process The modAL Python framework was used in this work to perform the active learning process (Danka and Horvath, 2018). The uncertainty sampling method was used as the query strategy (Lewis and Catlett, 1994). The number of queries was set to 20 and for each query the 100 samples the classifier was the most uncertain was given to the model to learn. The initial dataset size was fixed according to the previous experiment results to 200 which was not too bad classification scores for beginning with. This initial dataset was randomly chosen. The ResNet neural network was used as the classifier. When new instances (samples of a class) were fed into it, it was retrained from scratch with the new bigger dataset with batch size and number of epochs set to 32 and 20 respectively. To assess the performances of the active learning process, the balanced accuracy score was computed at the end of each query. The test set contained the same proportion of blue whale D-calls and fin whale 40-Hz calls as the original one. The experiment was launched 5 times but only the best model accuracy was plotted. A baseline with a random sampling is also evaluated and compared to the uncertainty sampling.

Imbalance dataset PAM datasets are mostly imbalanced (either for presence vs absence or for different whale calls). Consequently, two different experiments were carried out: with or without weight balancing class for the computation of the loss function.

6. https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/applications/resnet.py

Weight balancing consists in setting different class weights in the loss function of the trained model to over-penalize the miss-classification of the under-represented class compared to other classes. Using cross entropy as loss function, it can be modified for this purpose as follows. Let y_{ik} be 1 if k is the true class of data point i, and be 0 otherwise, and $y'_{ik} \in (0, 1]$ be the corresponding model estimation. The original cross-entropy can be written as:

$$H_y(y') = -\sum_i \sum_{k=1}^{K} w_k y_{ik} log(y'_{ik}) \tag{2.1}$$

This is called class weight balancing. When used, the class weights were set to the original proportions of calls (blue whale D-call: 1 and fin whale 40-Hz: 11.57) during the training phase of the model (class_weight parameter in Keras). This means that 1 occurrence of fin whale 40-Hz weighs 11 blue whale D-call instances in the loss function.

2.3.3.4 Key results

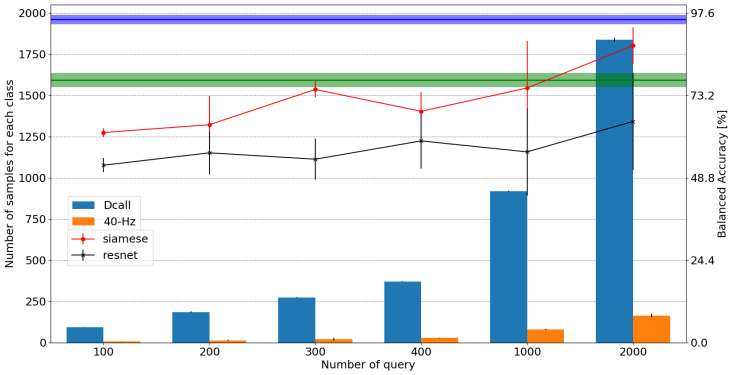


Figure 2.3 – Balanced accuracy for the without balancing classes with weights (left) and with balancing classes with weights (right). Bar plots are the number of samples for each class after a query. Blue and green horizontal lines with shadow area represent the mean balanced accuracy for the baseline with the standard deviation for the ResNet and LightGBM models with all training samples, respectively.

FSL Fig. 2.3 shows the balanced accuracy score for each training dataset size. Bar plots represent the number of total samples for each class for the training dataset. Blue horizontal line with shadow area represents the mean balanced accuracy for the ResNet baseline with the whole training dataset with the standard deviation and the green one is for the LightGBM model baseline with the whole training dataset. The ResNet baseline with smaller sizes of training dataset is shown in black. The same scale was used for both figures. The Siamese network feature extractor performs well compared to the ResNet model for the same training dataset size. It also outperforms the LightGBM accuracy when using one fifth of the original training dataset.

AL Fig. 2.4 shows the balanced accuracy score after each query for the best model in the experiments. Bar plots represent the number of total samples for each class after a query. Blue horizontal line with shadow area represents the mean balanced accuracy

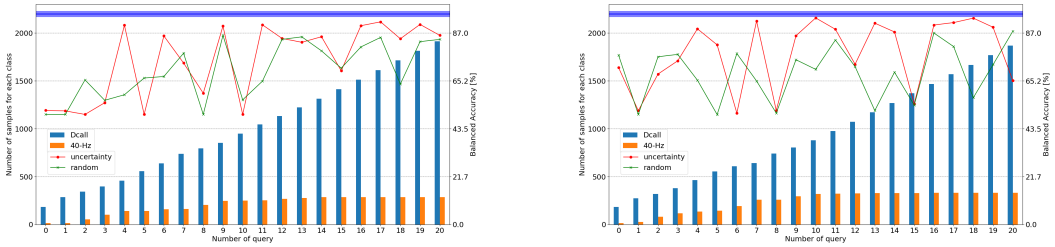


Figure 2.4 – Balanced accuracy for the system with (right) and without (left) class weight balancing. Bar plots are the number of samples for each class after a query. Blue horizontal line with shadow area represents the mean balanced accuracy for the baseline system with the standard deviation.

for the ResNet baseline only with the standard deviation. The same scale was used for both figures. Both active learning processes reached balanced accuracy scores close to the original baseline system at the fourth query (about one fifth of the original dataset). The maximum balanced accuracies are 91% and 93% without and with the use of weight balancing, respectively. The number of samples for each class also depends on the weight balancing. When no weight balancing is used, less fin whale 40-Hz instances are queried. For the 10th query, the number of fin whale 40-Hz is about 320 and 250 with and without the weight balancing, respectively. The random sampling strategy has a mean balanced accuracy over all queries of 69% with and without weight balancing and the uncertainty sampling achieves 77% and 75%, respectively. The uncertainty sampling reaches highest values compared to the random sampling for weight and not weight balanced systems.

2.3.3.5 Discussion

Deep learning has been shown to be a powerful machine learning method in several research areas such as image classification (Krizhevsky et al., 2012) and whale call recognition (Bermant et al., 2019; Ferrari et al., 2020; Shiu et al., 2020), its main drawback is its inefficient training process which requires large (Sun et al., 2017), well-balanced (Johnson and Khoshgoftaar, 2019) and sufficiently high quality (Hendrycks et al., 2019) annotated training sets. This step is resource intensive. In this work, two different deep learning approaches are investigated to reduce the annotation effort.

FSL The performances are better than using the ResNet model. It is believed that using Siamese network for feature extraction can help in reducing the annotation effort (Wang et al., 2020). It also outperforms a state-of-the-art machine learning method with less annotated data making it a good candidate for further investigations on how to use it with 1- or 5- shot learning for example. Moreover, using Siamese network in an imbalance classification task (blue whale D-call class is overrepresented compared to fin whale 40-Hz one) is suitable for this kind of dataset (Wu et al., 2019; Wang et al., 2020). It also seems that Siamese network method can generalize to unseen data better than the ResNet model in our case.

In this work, the Siamese network learns to discriminate similar or dissimilar embeddings using dense layers instead of a distance metric (Figueroa-Mata and Mata-Montero, 2020) or specific loss (Melekhov et al., 2016). This is another way for the Siamese network to learn inter- and intra- class relationships. By gaining knowledge about the relationships

between the pairs of inputs, the Siamese network acts like a feature extractor (Melekhov et al., 2016).

Next steps will be to use it to detect on sliding windows the presence of these two calls in the original dataset. Another imbalance issue will appear as the two considered calls are underrepresented compared to noise in the original dataset. Another issue that may arise by trying sliding window detection in audio data is the disagreement between annotators for the fin whale 40-Hz class. The neural network detector will be more to prone to be uncertain between noise or fin whale 40-Hz.

AL This investigation of active learning with a simple uncertainty sampling strategy shows that selecting appropriate samples leads to similar results to a model trained on the whole dataset with a decrease of the manual annotation effort. The uncertainty sampling strategy is shown to be better than the random one by selecting samples that the model is the most uncertain. It reaches highest prediction scores and faster. Using AL process could enable to save about 50% of the annotation effort to reach similar performances to a model trained on the whole dataset. Similar trend is described in other research areas such as medical imaging (Yang et al., 2017; Shi et al., 2019) and sound classification (Shuyang et al., 2017).

The initial dataset of 200 samples was chosen randomly. However, this initial dataset plays a key role as shown Fig. 2.4. The first balanced accuracy score gap between the two different initial random dataset is more than 20%. This shows that some samples are more discriminative than others and the main aim of active learning is to find them to better and faster train a model (Shuyang et al., 2017; Kholghi et al., 2018).

The imbalance issue of the PAM dataset is also partly addressed in this work by balancing the weight in the loss function of the neural network. A small improvement of 2% in the maximum balanced accuracy is found for the weight balanced model. AL can partly remove the class imbalance Attenberg and Ertekin (2013). When sampling, it can be seen on Fig. 2.4 that when the underrepresented class increase, the balanced accuracy score can decrease. This raises the question of the quality of the labels that add confusion to the model for its prediction. This class imbalance issue could lead to regions of the problem space that are totally missed or misclassified (Attenberg and Ertekin, 2013). This could explain the instability of the balanced accuracy that has a increasing trend but for some queries a huge decrease in the score is observed.

Another explanation of such instability is the non-deterministic nature of the training of deep learning models. They did not always learn the same embeddings. In the proposed AL process, the model is retrained from scratch at each query which could explain such difference from one query to another.

Finally, Kholghi et al. (2018) showed that AL is scalable and can be used for large unlabeled dataset like ones that can be found in PAM.

2.3.3.6 Conclusion

In this work, two different state-of-the-art methods are investigated to reduce the annotation effort. It is shown that using Siamese network as feature extractor or active learning could enable to reach similar performances that those found when training a model with the whole dataset.

2.4 Highlights & Summary

In this chapter, we present and describe another solution to our initial issue on the assessment of marine mammal presence from underwater passive acoustics with deep learning based methods in a weakly-supervised (but) big data context by involving cetacean experts:

— An opens source collaborative annotation platform was proposed to deal with the need of having several ground truths for a same PAM dataset for better DC systems (T_2^1).

— The proposed platform should be a first step towards multiple applications such as understanding and quantifying the inter-annotator variability, aggregating and evaluating the impact of the different annotations of DC systems to better understand their weaknesses and strengths and finally to accelerate the annotation process with annotation help by using deep learning systems (T_2^2).

— The proposed work was improved by efforts made to follow FAIR principles of open science with the working reports proposed to the community to help them using APLOSE.

Chapter 3

The marine mammal detection and classification (MMDC) approach

3.1 Introduction

MMDC consists in developing automated DC systems for marine mammal vocalizations. It is expected to accurately detect and assign a label to an instance of an acoustic recording containing one or more vocalizations produced by a specie of interest.

Only binary classification is considered in this work: binary classification of single label acoustic event using isolated sounds (referred to as acoustic event classification) and single acoustic event detection but also binary acoustic scene classification (presence or absence of biophony in an audio recording).

3.1.1 Landscape of DC methods

Different methods are used for MMDC task. Here, traditional approaches are presented. They can be split into two categories.

3.1.1.1 Parametric methods

Audio descriptor-based approaches (not image-like representations) The use of audio descriptors to describe the marine mammal signals of interest is a first approach to compare and classify signals. They are often used as inputs to detectors or classifiers.

Pace et al. (2010) used linear prediction coding, Mel Frequency Cepstrum Coefficients (MFCCs) and Cepstrum coefficients to classify humpback whale call. They compared these 3 features set to classify humpback whale subunits and they showed that MFCCs gave better classification results.

Shamir et al. (2014) proposed the use of more than 2800 image features such as textures, statistical distribution of the pixel intensities among others from different 2D representations of the audio signal. It has to be noted that the set of features used in their work is a type of transfer learning because the feature set was originally designed to classify biomedical images. To gather whale calls into sound groups, the features were weighted by a Fischer discriminant score and a similarity distance between the computed set of features from different whale calls was computed. They quantified the similarity of different whale calls in an unsupervised algorithm.

Bouffaut (2019) used spectral features as inputs to a classifier to identify different low frequency whale calls. They reached more than 99% of precision and recall for several call types. Patris et al. (2019) also used spectral features for classifying low-frequency blue whale tonals. To describe a tonal, the first four peak frequencies should be integer multiples of the average band interval.

Spectrogram correlation detector Another studied approach for detecting whale calls is the spectrogram cross-correlation. This method enables to measure a similarity distance between a reference (also called kernel or template) and a test spectrogram. A threshold is is then applied on the similarity score to perform the detection. Mellinger and Clark (2000); Munger et al. (2005); Baumgartner and Fratantoni (2008); Shabangu et al. (2017) used the spectrogram correlation to build a template function for the acoustic signal of interest. Then, they cross-correlated it with a target spectrogram to calculate their similarity and perform detection.

Nanaware et al. (2014) compared energy summation and spectrogram cross-correlation of the softwares ISHMAEL (Mellinger, 2002) and PAMGUARD (Gillespie et al., 2009), respectively. They found that spectrogram cross-correlation outperformed the energy summation method for call detection.

3.1.1.2 Parametric modeling approaches

Associated stochastic frameworks of parametric model makes these approaches not suffering from some drawbacks of correlation-based detectors. In particular, it does not require the choice of an a priori fixed and subjective template. Socheleau et al. (2015) developed a method for blue whale call detection based on a subspace detector of sigmoidal-frequency signals with unknown time-varying amplitude. Many marine mammal vocalizations are frequency modulated and can be modelled as polynomial-phase signals Urazghildiiev and Clark (2006). Chen et al. (2019) proposed to extract whistle component with sparse decomposition combined with spectrogram ridge feature extraction. The classification step is performed with an SVM or a Random Forest (collection of decision trees whose outputs are aggregated for the final decision) fed with ridge features. They used only 5 samples maximum to train and test their algorithm. They reached an accuracy of more than 98% when using a Random Forest.

3.1.1.3 Physical model-based approaches

Propagation modeling of the underwater acoustic environment is also considered to improve the acoustic detection of marine mammal calls. Küsel et al. (2011) modelled the transmission loss of their study site to estimate the SNR of beaked whale click detection for density estimation of the beaked whale population. They used simulated data to estimate the probability of detection at their hydrophone of beaked whale clicks. They had to estimate the sound propagation model because no in situ measurements of the site characteristics were available. They concluded that best probability of detection could be found with in situ measurements. Marques (2013) used empirical measurement to model the sound propagation environment to estimate the probability of detection. They proposed to validate theoretical sound propagation modeling with empirical data when possible.

3.1.1.4 Machine learning

Usual classifier is the SVM with an acoustic feature set (Jarvis et al., 2008). They used the times between consecutive zero crossings and the envelope amplitude were selected as inputs to classify Mesoplodon clicks. They reached an accuracy of more than 90% accuracy to classify foraging clicks and about 80% to classify buzz. Mazhar et al. (2007) also used SVM for classifying humpback whale calls. They used cepstral coefficients as inputs and they reached 99% of accuracy. They outperformed another machine learning approach, Gaussian Mixture Models (GMMs), on the same dataset of more than 10%. Ness (2013) also investigated the variability in the parameters for SVM. He found that the best kernel was the radial basis function. His proposed system achieved about 92% of accuracy on a subset of the Orchive database.

GMMs and Hidden Markov Models (HMMs) are also popular methods to automatic identification of whale calls from different species such as humpback whales, killer whale calls, beaked whales, pilot whales and Risso's dolphins (Mazhar et al., 2007; Roch et al., 2008; Brown et al., 2010). One of the main advantage of GMMs is that any data distribution can be modelled by adjusting the parameters and number of Gaussian probability density functions used in the mixture model. Compared to the HMMs, GMMs cannot capture the temporal structure of the considered signals (Usman et al., 2020).

Another traditional approach is decision trees. Ones of their main advantages is their output prediction can be easily interpreted and they can perform well with small anno-

tated datasets. Chen et al. (2019) compared the performances of Random Forest and SVM to classify odontocetes. They found that Random Forest outperformed SVM when smaller amount of annotated data is available, but they performed equally when increasing the dataset size. Garland et al. (2015) used classification and regression tree and Random Forest to classify the beluga calls into different categories to build a beluga sound repertoire. They proposed this method to ensure better classification compared to parametric cluster analysis such as discriminant function analysis and principal component analysis which have too strong assumptions to be checked.

3.1.1.5 The DL era

Motivations The traditional approaches exploit features from PAM time series with which information-extraction models can be constructed. Handcrafted features have proved effective and can represent a variety of spectral, textural, and geometrical attributes of the sounds. However, since these features cannot easily consider the details of real data, it is impossible for them to achieve an optimal balance between discriminability and robustness. We utilize a large amount of labeled data and state-of-the-art deep learning techniques effectively trained to tackle one main challenge in bioacoustics research: develop an automatic, robust, and reliable detection and classification of useful and interesting marine mammal signals from large bioacoustic datasets.

The different approaches presented above are resource intensive in terms of expert inputs. For example, for the spectrogram correlation detector, this is specific to one call of one species and the threshold needs to be adjusted for one dataset. All the expert knowledge is another bias that can contamine the detection results. Most of the approaches described above are not robust to new unseen dataset (poor generalization) or to new calls of different species. Consequently, thanks to the successful use of deep learning in other research communities (computer-vision (Krizhevsky et al., 2012), biomedical (Cao et al., 2018), finance (Heaton et al., 2017)), deep learning is more and more used for bioacoustics detection and classification tasks (Bergler et al., 2019; Bermant et al., 2019; Poupard et al., 2019; Thomas et al., 2019; Ferrari et al., 2020; Kirsebom et al., 2020; Shiu et al., 2020).

In marine bioacoustics, NN-based methods date back to the end of 1990's, for example the use of unsupervised self-organizing map networks to categorize the bioacoustic repertoire of false killer whale (Pseudorca-crassidens) vocalizations (Murray et al., 1998). It has also been demonstrated that a radial basis function network could effectively distinguish between six individual sperm whales (Schaar et al., 2007). NN have been constructed to classify the bioacoustic signals of killer whales based on call type, individual whale identity, and community dialect (Gaetz et al., 1993; Deecke et al., 1999).

Nowadays, deep learning experiences a renewed interest due to the recent technological advances and available annotated datasets. In PAM community, several studies of marine mammal acoustic presence using deep learning methods have recently been published. (Harvey, 2018) used a computer-vision deep learning architecture (ResNet (He et al., 2016)) to detect humpback whale vocalizations. However, DL methods ask for more data, and it is with the advent of DL that the PAM community started evaluating generalization by testing their DL methods on different spatio-temporal datasets. (Kirsebom et al., 2020) also used the ResNet architecture to detect and classify at once North Atlantic right whale calls. They achieved a precision of 90% for detection and classification on unseen data. They proved the ability of deep learning approaches to better capture intrinsic features of data and as a result to better generalize to unseen data over sites and periods compared to traditional methods described above. Similarly, Shiu et al. (2020) showed this powerful capacity of deep learning approach to generalize to unseen data. They also detected North

Atlantic right whale calls but near California, USA. To achieve better performances, they used data augmentation techniques to feed deep learning methods with more training samples to help them generalize. They achieved almost about 90% of true detections on unseen data.

Convolutional neural network (CNN) based image-like spectrograms Spectrogram images and their derivatives are the most common inputs fed into deep learning approaches for PAM datasets (Bermant et al., 2019; Bergler et al., 2019; Poupard et al., 2019; Thomas et al., 2019; Kirsebom et al., 2020; Li et al., 2020; Merchan et al., 2020; Shiu et al., 2020). The reason for using images instead of raw audio waveforms as inputs to neural networks is computer-vision architectures have been successfully tested on image datasets and transforming raw audio to image such as spectrograms can be done at no cost. Moreover, feeding into deep learning systems image-like spectrograms can enable the PAM researcher to explain / interpret the output of the neural network. Indeed, as seen in Chapter 2 and Thomas et al. (2019), a great part of PAM dataset analysis is performed visually by an expert instead of aurally. For example, it is easier to identify a pattern in an image than processing low frequency audio data (infrasonic and not all the whale calls are audible for humans) to recognize a sound.

Thomas et al. (2019) proposed a new image-like spectrogram input to be fed into deep learning systems to address the variability issue of the spectrogram representation and capture features for both low and high frequency whale calls. For that purpose, they interpolated and then stacked spectrograms of different time and frequency resolutions. Their best model combined with their novel representations achieved an accuracy of about 96%.

3.2 Transfer learning

3.2.1 Background

Transfer learning refers to "the situation where what has been learned in one setting (i.e., distribution P_1) is exploited to improve generalization in another setting (say distribution P_2)" (Goodfellow et al., 2016). In transfer learning, the learner must perform two or more different tasks, but we assume that many of the factors that explain the variations in P_1 are relevant to the variations that need to be captured for learning P_2.

In general, transfer learning can be achieved via representation learning when there exist features that are useful for the different settings or tasks, corresponding to underlying factors that appear in more than one setting. Many visual categories share low-level notions of edges and visual shapes, the effects of geometric changes, changes in lighting, etc. Here, we have shared lower layers and task-dependent upper layers.

However, sometimes, what is shared among the different tasks is not the semantics of the input but the semantics of the output. For example, a speech recognition system needs to produce valid sentences at the output layer, but the earlier layers near the input may need to recognize very different versions of the same phonemes or sub-phonemic vocalizations depending on which person is speaking. In such cases, it makes more sense to share the upper layers (near the output) of the neural network and have a task-specific preprocessing. The lower levels (up to the selection switch) are task-specific, while the upper levels are shared. The lower levels learn to translate their task-specific input into a generic set of features.

As detailed in our introduction, for many domains, sufficiently large datasets used to

train deep neural networks from scratch are scarce, so transfer learning is often used to solve the problem. If there is significantly more data in the first setting (sampled from P_1), then that may help to learn representations that are useful to quickly generalize from only very few examples drawn from P_2.

In the related case of domain adaptation, the task (and the optimal input-to-output mapping) remains the same between each setting, but the input distribution is slightly different. For example, consider the task of sentiment analysis, which consists of determining whether a comment expresses positive or negative sentiment. Comments posted on the web come from many categories. A domain adaptation scenario can arise when a sentiment predictor trained on customer reviews of media content such as books, videos and music is later used to analyze comments about consumer electronics such as televisions or smartphones. One can imagine that there is an underlying function that tells whether any statement is positive, neutral or negative, but of course the vocabulary and style may vary from one domain to another, making it more difficult to generalize across domains. Simple unsupervised pretraining (with denoising autoencoders) has been found to be successful for sentiment analysis with domain adaptation (Glorot et al., 2011). A related problem is that of concept drift, which we can view as a form of transfer learning due to gradual changes in the data distribution over time.

In the following, we shortly review two special techniques of transfer learning.

Pre-trained models as feature extractors Deep learning models are made up of layers that learn different features at different stages of the training on a sample. These layers are used as feature extractors. The last layer of the feature extractor part is connected to a final layer to perform the classification of the input sample.

Thanks to this layered architecture, deep learning models are modular meaning that several layers can be removed or replaced after a particular layer. Moreover, this enables to use pre-trained layers in an architecture and to use the final layers for a specific task.

In this specific transfer learning task, the main idea is to use pre-trained weighted layers (usually on a bigger database) to extract features instead of training the deep learning model from scratch. Only the last layers will be updated to learn how to classify the new data. The knowledge transfer is the feature extraction process.

Fine tuning pre-trained models Fine-tuning a model is defined by the fact that the user choose which layer of the network to retrain on the new task. Lower layers capture generic features while the higher ones can learn task-specific ones, another use of pre-trained weighted layers is layer weight initialization. For that purpose, some layers of the NN are retrained and the other layers are frozen (their weights are fixed). The knowledge transfer is the layer weight initialization.

3.2.1.1 Challenges

In this chapter, our investigations on transfer learning mainly addressed the questions of **"what to transfer" and "how to transfer" in MMDC**, referring to which part of knowledge can be transferred across domains or tasks and how it can be efficiently done in terms of MMDC performance, as presented in Pan and Yang (2010). Some knowledge is specific for individual domains or tasks, and some knowledge may be common between different domains such that they may help improve performance for the target domain or task. After discovering which knowledge can be transferred, learning algorithms need to be developed to transfer the knowledge, which corresponds to the "how to transfer"

issue. "When to transfer" asks in which situations, transferring skills should be done. We are interested in knowing in which situations, knowledge should not be transferred. In some situations, when the source domain and target domain are not related to each other, brute-force transfer may be unsuccessful. In the worst case, it may even hurt the performance of learning in the target domain, a situation which is often referred to as negative transfer. Most current work on transfer learning focuses on "what to transfer" and "how to transfer", by implicitly assuming that the source and target domains be related to each other. However, how to avoid negative transfer is an important open issue that will be attracting more and more attention in the future becauset two different domains can have different feature spaces and can have different label spaces.

3.2.1.2 Our contributions

Following similar works from ocean sciences (e.g. Racah et al. (2017) for climate simulations), bioacoustics datasets may not share the same statistics as natural images or sounds, which are mainly covered by many popular deep learning techniques, raising the question whether we can build off successes from the computer vision community such as using pretrained weights from CNNs Simonyan and Zisserman (2015); Krizhevsky et al. (2012) pretrained on ImageNet (Russakovsky et al., 2015) or not. In this chapter, we then propose a series of numerical experimentations to better understand the questions of **"what to transfer" and "how to transfer" in MMDC** using both pre-trained NN from other audio tasks, as well as NN fully trained on MM datasets.

Original contributions in this chapter are two-fold:

— **Working report Nguyen Hong Duc, P.**, Adam, O., Gérard, O., White, P. R., and Cazau, D. "Data augmentation for marine mammal vocal sound classification: from naive to physically-based schemes"

— **Working report** Cazau, D., **Nguyen Hong Duc, P.**, Druon, J.-N., Matwins, S., Fablet, R., "Multimodal deep learning forcetacean distribution modeling. Case study of fin whales (*Balaenoptera physalus*) in the western Mediterranean Sea."

3.2.1.3 Related works

The questions of "what to transfer" and "how to transfer" have been widely addressed in the literature. Yosinski et al. (2014) presented their findings on how the lower layers act as conventional computer-vision feature extractors, such as edge detectors, while the final layers work toward task-specific features.

One of the first transfer learning from computer-vision research to the PAM one was for the Kaggle competition initiated by the Marinexplore and the Cornell University with the use of a convolutional neural networks used modified from AlexeNet Krizhevsky et al. (2012). The winner of the challenge was able to reach an accuracy of almost 97%.

Another transfer learning case is described in Thomas et al. (2019). Besides using computer-vision deep learning architectures (ResNet50 and VGG19 (Simonyan and Zisserman, 2015)) to classify whale calls, they proposed to transfer the knowledge of the system to a new class of whale call.

Nanni et al. (2020) built a large-scale benchmark on using computer-vision architectures, SVM with handcrafted acoustic features and ensemble models on the DCLDE 2013 dataset. They showed that the best model on this dataset was the ensemble one using 6 deep learning models with either spectrograms or scattergrams as inputs. They reached about 97% on the dataset.

Zhang et al. (2019) used ResNet101 and Xception (Chollet, 2017) as feature extractors to classify odontocete calls. The inputs of their architectures were either raw waveforms or melspectrogram images. To fit the requirements of dimensions for the computer-vision architectures, they first extracted feature maps from raw audio waveforms, and they stacked them to form an image with 3 channels to be fed into the neural networks. For the melspectrogram images, they concatenated logmel-spectrograms and their first and second deltas to have a 3-channel image as inputs. The best model was the pre-trained ResNet101 model with logmel-spectrograms and their deltas that achieved about 100% to classify whale calls. The ResNet architecture was also the feature extractor in (Harvey, 2018; Kirsebom et al., 2020) to detect calls from different species.

Zhong et al. (2020) established a first benchmark with several computer-vision deep learning architectures. They compared VGG16, AlexNet, ResNet50 and DenseNet (Huang et al., 2017) to classify spectrograms of beluga calls. Finally, they found that the best model is the ensemble modeling of the 4 architectures by computing the weighted average of the 4 predictions. They achieved an accuracy of about 99%.

In the same vein, Shiu et al. (2020) also benchmarked 4 computer-vision architectures LeNet , BirdNET(Kahl et al., 2018), VGG, ResNet and a hybrid architecture made up of 1D convolutional layers with recurrent layers (gated recurrent layers) (Xu et al., 2017). All their deep learning architectures outperformed the approaches proposed for the DCLDE 2013 challenge. Moreover, they also presented the capacity of their architectures to generalize across different sites and periods.

3.3 Deep learning for odontocetes assessment using FFT-based soundscape features

Underwater soundscape studies still rely on manual annotations obtained from experts reducing their ability to scale. Consequently, researchers are looking for ways to take advantage of machine learning algorithms for automating natural sound processing (Lin and Tsao, 2019; Poupard et al., 2019; Roca and Van Opzeeland, 2019). (Lin and Tsao, 2019) have presented an efficient method to separate fish choruses in shallow environments and (Poupard et al., 2019) have proposed to use deep learning to recognize orca sounds in several years of audio recordings. When supervised, these techniques rely on manual analysis of the audio data which is resource-consuming, and it requires a high level of expertise which induces an observer bias (Leroy, 2018). One solution to reduce the amount of data to label is transfer learning (Yosinski et al., 2014) which is the use of pre-trained machine learning models for a specific task to capture intrinsic characteristics of a dataset into embeddings for another machine learning system that will be trained on a target dataset. In terrestrial environment, (Coban et al., 2020) used a VGGish embedding on 10-s long audio to classify audio segments into 8 soundscape contributors.

There has been little research about using low-level soundscape features as inputs to deep learning neural networks (NNs). Our aims were to 1) assess whether NNs are able to learn an embedding representing the presence of cetaceans by feeding into them a simple long-term representations, 2) evaluate the performance gains by adding biological a priori such as seasonal patterns, 3) determine if the knowledge transfer from one site to another could be favourable. Percentiles and root-mean square level were used to develop a supervised deep learning model to discriminate between audio recordings containing or not cetacean sounds.

A deep learning method was developed to assess the presence of cetacean sounds in underwater recordings based on long-term representation. Two datasets were used to eval-

uate the robustness of the implemented neural network. Saint-Pierre-and-Miquelon (SPM) audio data were retrieved at two different locations and were used for the training and testing of the NN. Audio recordings were also collected from the Fromveur Passage off the coast of the French Brittany. Both sites are prone to cetacean presence and anthropogenic activities.

The main goal of this analysis reported in this section is three-fold: i)detecting the presence of cetacean sounds in audio recordings based on underwater soundscape metrics; ii) investigating the benefits of adding biological a priori assumptions in a NN; iii) transferring knowledge from one site to another.

3.3.0.1 Material

The same SPM dataset as in the soundscape study is used. For the Fromveur dataset, passive acoustic data were collected using one Song Meter 3M (SM3M, Wildlife Acoustics Inc, Maynard, MA, USA), owned by the Parc Naturel Marin d'Iroise (PNMI, depending on the Office Française de la Biodiversité). The recorder was located in the channel between the Ouessant island and the Molene archipelago in the Parc National Marin d'Iroise. It was moored at a mean depth of 27 m. No gain or sensitivity were retrieved for this study. The acoustic survey was carried out from September 2017 to December 2017. This site was acoustically sampled for 20 mins followed by a pause of 40 mins (66% duty cycle, sampling rate of 256 kHz at 16 bits). This channel was sampled to study the presence of odontoceti, specially dolphins. The terms Fromveur or pnmi to name the Fromveur dataset are used interchangeably in the following.

Table 3.1 – Summary of the characteristics of recordings performed in each site for each year.

Recorder	Mooring period (MM/dd, hh)	Duty cycle (mins ON / mins OFF)	Use
2010			
AURAL-B	08/19, 18:00 - 11/02, 23:00	45/15	training
2011			
AURAL-A	04/23, 13:00 - 07/23, 12:00	30/30	testing
AURAL-B	04/25, 20:00 - 08/16, 18:00	30/30	training
2011-2012			
AURAL-A	10/15, 01:00 - 04/30, 18:00	17.03 / 42.57	testing
2017			
Fromveur	09/22 13:00 - 12/21, 15:00	20 / 40	transfer learning

3.3.0.2 Data analysis

Neural network (NN) inputs For each dataset, the PSD was determined by the Welch method (Welch, 1967) with 1024-point Hamming window, 50% overlap, based on 10-s temporal signal segments. The DC offset value was discarded. The time resolution is 10s and frequency resolution is 32 Hz. The framework proposed in OSmOSE report 2 was used to compute the Welch values. The data from the 10s segments were combined to obtain the 1st, 5th, 50th, 95th, 99th percentile levels and the RMS level. These features were then fed into the NN. No calibration and gains were retrieved for the Fromveur, uncalibrated Welch percentiles were computed. Inputs of the neural network was matrices

Dataset	Label	Training	Validation	Testing
SPM	Presence	622	156	286
	Absence	2989	747	6671
Fromveur baseline	Presence	401	93	767
	Absence	369	100	433

of size 512×6. 2010 and 2011 site B of SPM audio files were used as the training set while the testing set was made up of site A SPM files. This choice was motivated by the fact that more odontocetes were observed site B (soundscape study).

Training and testing were performed on the acoustic activity of odontocetes. Each audio file was given a label "presence" or "absence" when odontocete sounds were identified or not. This system does not require a high level of expertise in underwater sound identification and it is faster to annotate.

3.3.0.3 NN and baseline architectures

The focus of the paper was not to find the optimal parameters for a specific neural network architecture, the aim was to show differences between performances for different architectures.

NN Each percentile is seen as a sensor and 1D convolutional operations are performed on them. After flattening, 4 dense layers are added to perform the binary classification. The model was trained for 200 epochs with batch size set to 128. An Nadam optimizer (learning rate of $1e^{-3}$) was used. Reduction of the learning rate by 10% was introduced if the validation loss has not been decreased for 2 epochs. Each experiment is run using 10-fold cross validation based on the stratified subsets. A binary_crossentropy loss was used as the loss function. All implementations were performed in Keras.

All datasets were highly imbalanced (more files without cetacean sounds in SPM dataset but more with them in others), class weights were added to the Keras model to increase the accuracy of the tested models.

Light gradient boosting model (lgbm) Gradient boosting decision tree (GBDT) (Friedman, 2001) is a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability, achieving state-of-the-art performances in many machine learning tasks, such as multi-class classification (Li, 2012). The LightGBM [1] version from Ke et al. (2017) was used, which achieves similar performance than state-of-the-art GBDT but with an important reduction of training time.

Including seasonality pattern into NN The date of the recording was also fed into the NN when indicated. "Day", "month" and "hour" means that the day of the year, the month or the hour of the day was added as an input to the lgbm or NN network. Seasonal patterns were based on previous works ((Delarue et al., 2018) for "day" and "month" patterns and Cascao et al. (2020) for "hour"), a seasonality pattern was defined as a sine function with an offset to have values between -0.5 and 0.5. The "nots" indication means that no timestamp was used in the NN.

For the Fromveur dataset, the seasonal pattern was diurnal as dolphins are known to be in the surrounding of the Fromveur.

1. https://lightgbm.readthedocs.io/en/latest/

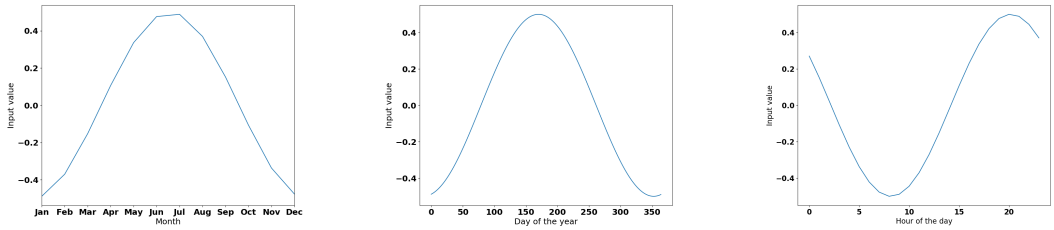


Figure 3.1 – (Left) Monthly pattern used for the NN architecture when training on the SPM dataset, (middle) daily pattern used for the NN architecture when training on the SPM dataset, (right) diurnal pattern for the Fromveur dataset.

Transfer learning For analyzing the cetacean presence in the Fromveur dataset, the pre-trained model weights on the SPM dataset were used as an initialization weights for the models trained on the Fromveur dataset. The main aim is to detect the presence of cetacean sounds, a custom loss in Keras was implemented to penalize false negatives to have less "absence" files classified into "presence".

Performance measures To evaluate the performance of our algorithms, the balanced accuracy was computed.

3.3.0.4 Key results

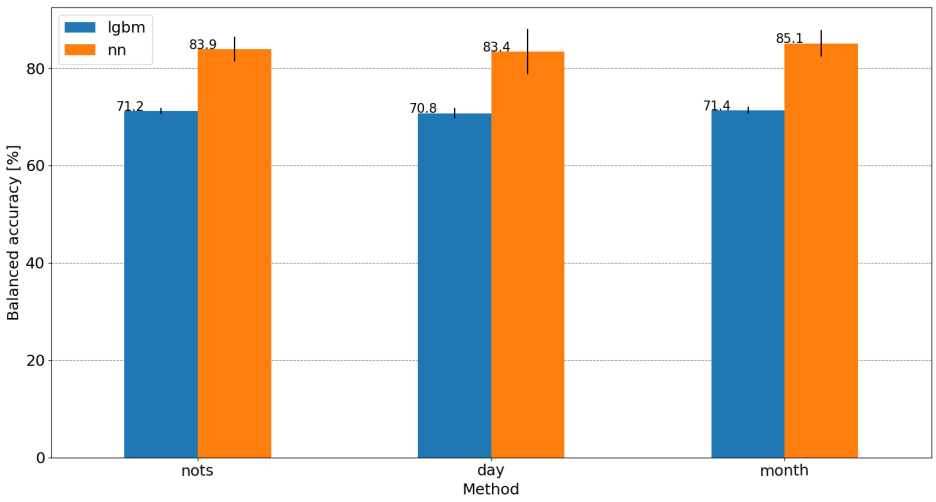


Figure 3.2 – Balanced accuracy for the lgbm and the NN architecture using or not seasonal pattern to help the algorithm to learn for the SPM dataset.

The lgbm method was outperformed by the NN in terms of balanced accuracy. The gap between these two methods reached more than 10% depending on the a priori added (cf Fig. 3.2). The balanced accuracy reaches more than 80% (cf Fig. 3.2) feeding only percentile and RMS levels into the NN.

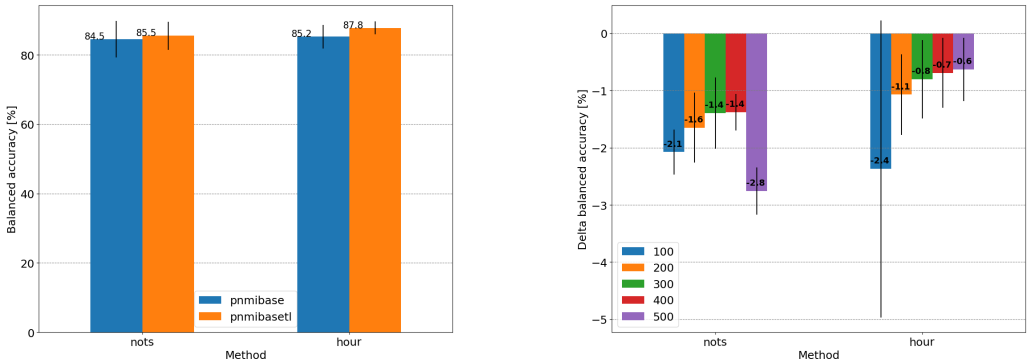


Figure 3.3 – (Left) Balanced accuracy for NN models for the Fromveur dataset: "pnmibase" is the NN trained from scratch whereas "pnmibasetl" is the NN that benefits from weight initialization of knowledge transfer (right) delta balanced accuracies between the mean balanced accuracy of the baselines on the left and the new computed balanced accuracies depending on the number of files used for training.

Adding the day of the year and the month as inputs to the NN architecture leads to a small variation in performances losing 0.5% and gaining 1.2%, respectively.

Knowledge transfer from one site with odontocete vocal activity and shallow waters to another shallow site containing delphinids is studied. It is set up as weight initialization for retraining on the new site data (cf Fig. 3.3 left).

Thanks to knowledge transfer, almost the same balanced accuracy as the systems without knowledge transfer is reached with less annotated data. By adding more data, balanced accuracy increased except for 500 files with no diurnal pattern. This raises two question about which data to feed into the deep learning method to faster and better train it and about the optimal parameters of the learning process (number of epochs, batch size, learning rate, optimizer) to find to have the best performances (cf Fig. 3.3 right).

3.3.0.5 Discussion

All results presented here are obtained for a specific task and specific datasets. It needs more thorough evaluations before drawing definitive conclusions about the different methods used.

With these first investigations, it has been shown that low-level features on long audio recordings (> 10 minutes) could be used as input of machine learning systems instead of using ecoacoustic indices (Roca and Van Opzeeland, 2019) as inputs which may be time-consuming (Lindseth and Lobel, 2018).

Moreover, a baseline with a lgbm method was used and outperformed by the neural network system in terms of detection performances. No optimization for both systems were used. The training time was similar for the same number of epochs. However, the neural network is less interpretable compared to the decision tree method. Then, a trade-off between interpretability and accuracy performances should be considered depending on the application but more and more research work addressed this black-box issue of neural networks (Fan et al., 2020).

The presented results also showed the importance of the inputs of the neural networks. By adding a single value for seasonal trends in biophony, the neural network performances

increased by almost 1%. Such result was not found for the lgbm method where all accuracies were similar. This was a first experiment in adding a priori information in neural networks to help the system in better classifying acoustic scenes containing biophony. The impact of such information still needs more investigations before drawing definitive conclusions but it is hoped that all the research accumulated on the marine mammals from its beginning could be used as added-value in the designing of neural networks.

Finally, the transfer learning (using pre-trained weights as initializations for a new neural network model on a new dataset) was found to be effective for sites with similar characteristics (depth and odontocete vocalizations).

3.3.0.6 Conclusion

Even if the presented results are only for our specific system (architecture, learning rates, number of epoch, batch size, loss detection task), it is believed that by optimizing all the parameters with brute force grid search, performances will increase. Thanks to the proposed Hadoop-Spark system, this step of optimization of a neural network could be performed efficiently and at scale thanks to the libraries that are compatible with both Python and Spark. It also showed that the big data system could be used for the computation of features to be fed into machine learning methods. The FFT-based features besides their usage in underwater soundscapes can be fed into deep learning methods to detect the vocal presence of cetacean in audio recordings. Moreover, knowledge transfer from one site to another with similar characteristics (depth, biologic sources) gave better results than training from scratch a new deep learning model. Here, soundscape metrics were averaged for each file. The use of small segments such as in DCASE challenge (10s-long audio recordings instead of 45min long and 16min long for SPM and Fromveur datasets respectively) could help recognizing which file contains cetacean sounds. The use of these features were used for the DCASE challenge and performed relatively well for classifying audio segments into 3 classes (Appendix H). However, splitting into small segments audio recordings involves spending more time to annotate if supervised deep learning are used. A solution would be to feed FFT-based features (combined to others that can be computed efficiently at scale) into unsupervised on-the-fly clustering method to gain annotation time. This short study also enables to introduce the second chapter with the involvement of cetacean experts to help deep learning methods to improve.

3.4 Data augmentation

Data augmentation can be seen as a particular case of transfer learning. Augmentation has a regularizing effect. Too much of this combined with other forms of regularization (weight L2, dropout, etc.) can cause the net to underfit.

3.4.1 Background

The data augmentation approach is motivated by the following principle: “The best way to make a machine learning model generalize better is to train it on more data. Of course, in practice, the amount of data we have is limited. One way to get around this problem is to create fake data and add it to the training set.” (from Goodfellow et al. (2016)).

Let’s remind that a classifier takes a complicated, high dimensional input x and summarize it with a single category identity y. This means that the main task facing a classifier is to be invariant to a wide variety of transformations. We can generate new (x,y) pairs

easily just by transforming the x inputs in our training set. Note that this approach is not as readily applicable to many other tasks. For example, it is difficult to generate new fake data for a density estimation task unless we have already solved the density estimation problem, but it is an effective technique for our specific classification problem of object recognition.

In data augmentation, the user of the algorithm encodes his or her prior knowledge of the task by specifying a set of transformations that should not alter the output of the network. The difference is that in the case of dataset augmentation, the network is explicitly trained to correctly classify distinct inputs that were created by applying more than an infinitesimal amount of these transformations. Also, on the contrary to analytical approach such as tangent propagation that only regularizes the model to resist infinitesimal perturbation, explicit dataset augmentation confers resistance to large perturbations and also works well with rectified linear units because different subsets of rectified units can activate for different transformed versions of each original input.

For example, images are high dimensional and include an enormous variety of factors of variation, many of which can be easily simulated. Operations such as translating the training images a few pixels in each direction can often greatly improve generalization, even if the model has already been designed to be partially translation invariant by using the convolution and pooling techniques of CNN. Many other operations such as rotating the image or scaling the image have also proven quite effective.

One must be careful with the "Class distribution invariance problem", i.e. not to apply transformations that would change the correct class. For example, optical character recognition tasks require recognizing the difference between "b" and "d" and the difference between "6" and "9", so horizontal flips and 180°rotations are not appropriate ways of augmenting datasets for these tasks. There are also transformations that we would like our classifiers to be invariant to, but which are not easy to perform. For example, out-of-plane rotation cannot be implemented as a simple geometric operation on the input pixels.

3.4.2 Key methods

Besides comparing traditional data augmentation methods for classification purposes, physical-based data augmentation methods by integrating underwater propagation models parametrized by the site characteristics (e.g. local sound speed in the water column) to describe the physical deformations of a source signal through the propagation medium.

A Pekeris waveguide is used to model sound propagation in shallow site while beam tracing models are used for deep-water site with Bellhop (Porter, 2011).

3.4.3 Key results

Simply re-balancing the original dataset during the training process through weight balancing should be a frist method to obtain accuracy gains for imbalance datasets. Despite its ease of implementation, it is barely reported in similar benchmarking reports, e.g. (Salamon and Bello, 2017). Directly augmenting unbalanced dataset proved to be efficient (Salamon and Bello, 2017).

Our most interesting result with data augmentation is how it allows to handle an explicit problem of generalization from non-deformed clean sounds to propagated distorted sounds, as formalized in our study by training models on the selecting the 150 samples with the highest SNR (SNR150 dataset) for each class while evaluating them on a non-overlapping sub-part of the original dataset. Indeed, we saw that augmenting the datasets

SNR150 allows to increase significantly their respective baselines up to 10% and even to reach similar performance than the best-performing one, namely weight balancing when augmenting the original dataset.

3.4.4 Using physical-based data augmentation to help deep learning system generalize

3.4.4.1 Key methods

For detection purposes with deep learning model, a noise class containing 7400 samples was generated by randomly extracting 3s of audio from SPM and DCLDE 2015 low frequency datasets. The deep learning model is trained with this noise class against the blue whale D-calls of the DCLDE 2015 low frequency dataset. All sounds are resampled to 2 kHz and they are passband filtered between 15Hz and 150Hz using a third-order Butterworth filter. The samples are then converted into a magnitude spectrogram (Shiu et al., 2020) using a window length of 1024 samples, a 4096-point Fast Fourier Transform and 95% overlap between the windows. The training set is made up of 4786 blue whale D-call and 7400 noise samples. No noise from the OHASISBIO dataset is extracted. The detection is performed on 9 files from the OHASISBIO dataset resulting in about 60hrs of audio test. Audio files of OHASISBIO covered different time seasons and sites. The 9 audio files are annotated by an expert in blue whale D-call. In these files, 1016 blue whale D-calls are identified.

To augment the database, physical-based augmentation is performed on the DCLDE 2015 low frequency D-calls. Site characteristics such as the mooring depth, bathymetry, and sound speed profile for the 4 seasons are included in the models. Random values are chosen for the whale distance from the recorder and All the recordings are done in deep-water sites and only beam tracing methods with Bellhop is used to generate deformations on the blue whale D-calls. 4 samples are generated using physical-based augmentation from the original dataset.

A ResNet with 18 layers is trained on the dataset with or without physical-based augmentation. Balanced accuracy reaches more than 99% to discriminate blue whale D-calls from noise. The best two models from the 5-fold cross-validation are used for detection on the OHASISBIO dataset.

3.4.4.2 Key results

For the model with no physical-based augmentation, the precision reaches 17% and the recall 86%. This system is more conservative and 86% of its detections are true positives. For the model with physical-based augmentation, the precision reaches 60% but the recall falls at 23%. Due to the larger training set and noisier blue whale D-calls synthetically generated, the system is more prone to false alarms. A trade-off between precision and recall should be considered depending on the final application.

All blue whale D-calls from the DCLDE 2015 low frequency dataset were used and augmented but as seen in the data augmentation study, it could be benificial to only extract high SNR blue whale D-calls and augment them.

Training samples are not time shifted unlike Kirsebom et al. (2020); Shiu et al. (2020). However, most the training samples are centered on the call to identify, it is believed that this lack of diversity in the positions of the call in a time window decreases the performance levels when using with a sliding window. In order to alleviate this effect, small step size in the sliding window can be chosen in spite of computer resources that can be saturated.

Indeed, start and end times of the identified blue whale D-calls by the deep learning model are the same meaning that only one time window is identified as a blue whale D-call. It is believed that by augmenting the dataset with time shift operations (to have samples at different locations in the time window), the precision and recall of the model will increase.

Moreover, each site has different ambient noise, it is believed that by generating pseudo absence from the site of interest and train a model with such noise will also increase the performance levels of detection. Whale calls are parsimonious in long-term dataset, it is unlikely to extract calls by randomly extracting noise.

Inference time is about 30s for a step size between sliding windows of 0.25s. This time is about 0.01 % of the total duration of the audio recording (about 6h30min). Consequently, trained deep learning systems are good candidates for real-time inference on embedded systems. However, for long calls, a 3s long buffer is considered in this work which is a big constraint to tackle depending on the audio sampling rate for embedding systems.

For this work, only a threshold of 0.5 and a time step between sliding detection window of 0.25s are tested. These parameters should be thoroughly analyzed to find the best trade-off between inference time and performance levels. Furthermore, more augmentation schemes should be assessed to determine the relevance of augmenting samples with physical-based methods which are time consuming.

3.4.5 Contributions to the doctoral research project

Building-upon physical principles of e.g. underwater sound propagation, and theoretical developments within well-established sub-fields such as production-based bioacoustics, seems a promising research avenue (Bianco et al., 2019).

Different augmentation schemes were evaluated in terms of classification performance of marine mammal vocal sounds, especially investigating the use of physically based data augmentation schemes to address the weak supervision and imbalance training set issue.

Physical-based augmentations are performed on an publicly available dataset to perform detection on new underwater sites for generalization purposes.

3.5 Multimodal deep learning

In this doctoral research project, we have mainly explored two “big” solutions to tackle at its source the challenges of weak supervision, i.e. providing original “knowledge basis” from which to train AI models, through Big Data (see chapter 1) and Big experts (see chapter 2). One other solution we have also investigated during this doctoral research project is Big Sensors, i.e. exploiting complementary information from different types of sensors.

3.5.1 Motivations

3.5.1.1 General

Learning from multimodal sources offers the possibility of capturing correspondences between modalities and gaining an in-depth understanding of natural phenomena. First, having access to multiple modalities that observe the same phenomenon may allow for more robust predictions. This has been especially explored and exploited by the AVSR (i.e. Audio-Visual Speech Recognition) community (Potamianos et al., 2003). Second, having access to multiple modalities might allow us to capture complementary information - something that is not visible in individual modalities on their own. Third, a multimodal

system can still operate when one of the modalities is missing, for example recognizing emotions from the visual signal when the person is not speaking (Andrew et al., 2013). In a few words, expected benefits from multimodal approaches include

— Bringing together complementary information

— Robustness in presence of noise in one modality

— Dealing with missing or unobserved data in one of the modalities

The multimodal approach shares a lot of its motivations with ensemble methods, which still bring impressive successes in various machine learning tasks. They have in common the capacity of putting together a high diversity of models, which ideally should be as accurate as possible while being as different as possible so that the different biases cancel each other out. This typically means using very different architectures or even different brands of machine-learning approaches.

Furthermore, one may argue that the more the heterogeneity of data and the noisier and the more complex the learning process will be, in particular due to the curse of dimensionality. Also, the more the data, the more difficult they are to annotate. But recent findings using deep learning methods with noisy and weak labels seems to indicate on the contrary that "the potential for enormous training sets can counterbalance these potential weaknesses in the labels" (Virtanen et al., 2017).

3.5.1.2 In marine ecology

Oceanic environments and marine animals are inherently difficult to study, and there are significant challenges involved with acquiring observational data at the scales and resolutions required to elucidate important conservation issues such as habitat preferences and the way that those preferences will be impacted by climate change (Ockendon et al., 2014). For example, Sveegaard et al. (2015) demonstrated the usefulness of the combination of a suite of methods to determine dynamic management borders of three harbour porpoise populations that have different conservation status, inhabit different habitats with different prey species and availability and face different anthropogenic pressures, and it is thus important to detect trends and manage them separately. To obtain this spatial definition of the management unit, they had to employ complementary monitoring methods such as genetics and morphometrics (to detect population differentiation and propose transition zones), tagging and passive acoustics.

In maritime surveillance, the satellite imaging can provide an exact instantaneous picture of the sea area at definite moments, and the passive acoustic system can provide a continuous track of boats that were detected by satellite (Bruno et al., 2010). Acoustics also show the ship track history. The ship tracking can be performed from the moment when the noise from the ship is detected.

3.5.1.3 Contributions to the doctoral research project

Our work consisted in developing different machine learning strategies able to fuse raw data from multiÂŋsource earth observation data. We were especially interested in the use of deep learning-based strategies, including multi-task and transfer learning, which are well suited to perform end-to-end co-training of heterogeneous modeling approaches. We demonstrated the usefulness of our methods on the two following scientific case studies, which have both been published:

In the following a concise summary of our study on cetacean distribution modeling (CDM), which is the most relevant in the context of our doctoral research project.

The first results reached in this work were to compile and manage a multi-source dataset useful for whale habitat modeling, starting from scratch. At least three types of data are quite sensitive: whale presence observations, passive acoustics and AIS data. For the first one, more than 12 different collaborative partnerships and data sharing agreements have been made to compile the dataset of fin whale presence-only records (see table A1 in the manuscript), for a total of 2575 over 12 years, making it one of the biggest dataset ever compiled for this species. Administrative procedures have also been engaged to make our dataset publicly available, so that a larger community of data scientists can potentially contribute.

Regarding modeling frameworks for CDM, in a nutshell, the CDM study consisted in investigating the usefulness of deep learning-based strategies such as multi-task and transfer learning regarding the extrapolation problem in CDM, leveraging their capacity to perform end-to-end co-training of heterogeneous learning tasks, namely a stochastic presence-background classification task and a deterministic rule-based model in our case. We demonstrate in this study superior performance with our hybrid models than expert models and conventional data-driven, taking fin whales of Mediterranean Sea as a study case. Although this is preliminary work, our close collaboration with J-N. Druon [2] (from European Commission), initiated within this project, is already a first evidence of interest of the Ecology community for our work.

To the best of our knowledge, no machine learning-based multimodal data fusion framework for any DCL tasks applied to marine mammals, but more quantitative comparisons of independent audio and visual observations. PAM is generally used to compute presence/absence indexes and cue rates, further explained by environmental variables. Rayment et al. (2017) compared detection results from visual and acoustic observations, and used GAM to evaluate the factors that most influenced the different observation modalities. In their work Baumgartner et al. (2019), whale occurrence was evaluated on a daily basis within particular radii of the buoy for the aerial survey observations (within 20-60 km in 10 km increments), and within particular time intervals before the start of the aerial survey for the near real-time passive acoustic observations from the buoy (within 12-72 hours in 12 hours increments). The log odds ratio test evaluates the ratio of the odds of acoustic detection when a species is visually present to the odds of acoustic detection when a species is visually absent. The log odds ratio was evaluated using a logistic regression between the near real-time passive acoustic observations (dependent variable) and the visual observations (independent variable). In addition to comparing daily occurrence estimates, they also used logistic regression to assess whether the probability of detecting a species during an aerial survey was related to the percentage of near real-time tally periods scored as "detected" within 12-72 hours prior to the start of the survey. Based on CPOD outputs, Thompson et al. (2014) performed GEE analysis to evaluate the best predictors (e.g. distance to coast, sediment, depth, ...) of dolphin presence likelihood in all PAM sites. They use the best fitting models to predict the likelihood of dolphin presence in all grid cells across the study area. Similarly, they modelled visual observations with a classification tree. To get a detection likelihood per species over all the spatial area, they multiplied the model outputs together, as PAM do not provide species identification. Fujioka et al. (2014) compared a visual sighting dataset and a stationary presence-type PAM dataset for bowhead whales (*Balaena mysticetus*) and explored seasonal and diel variation in sightings and acoustic detections to investigate how the PAM data could complement visual surveys and fill temporal gaps.

2. https://ec.europa.eu/jrc/en/person/jean-no%C3%ABl-druon

3.6 Highlights & Summary

In this chapter, we describe another solution to our initial issue on the assessment of marine mammal presence from underwater passive acoustics with deep learning-based methods in a weakly-supervised (but) big data context by involving cetacean experts:

— Transfer learning (architectures from other research communities, pre-trained networks) is performed to improve the accuracy of deep learning models for a specific task (T_3^1).

— Transfer learning from one site to another is also performed to evaluate the ability of a deep learning system to generalize to unseen audio data recorded on other sites (T_3^2).

Conclusion

Doctoral research project summary

This doctoral research project aims at revisiting the questions of how obtaining the "knowledge basis" from which to train AI models, and how developing models that better comply with small amount of training data and/or partially labeled data. UPA monitoring has been widely deployed for the analysis of the marine environment leading to a huge volume of acoustic data to be processed. Consequently, developing automatic analysis tools is needed. However, most of the automatic analysis methods rely on annotated data which are costly. For that purpose, this work proposes tools for the automatic analysis of marine mammal acoustic presence in UPA recordings considering the huge volume of data to process and the lack of annotated data to train automatic methods.

The first question about obtaining "knowledge basis" has been addressed through two "big" sources: raw data and expert annotators. To deal with these "big" sources, two tools are presented in the two first chapters of this thesis. The first tool (1) enables to efficiently and at scale compute FFT-based metrics of long-term PAM datasets that are at the core of PAM studies (raw data processing). It relies on specific big data frameworks Apache Hadoop and Spark. It is demonstrated that using these frameworks reduces the computation time of the FFT-based metrics compared to the traditional PAM tools. In Chapter 2, an open source and collaborative annotation platform (expert annotators), APLOSE, is presented. It is compared to the existing tools and used to perform a collaborative annotation campaign on an existing PAM dataset. An exploratory analysis about the variability of automatic DC system performances depending on the reference dataset used for training and testing is presented.

Regarding the second question about the development of AI models that comply with small amount of annotated data (weak supervision) , transfer learning is used. Either computer-vision state-of-the-art architecture trained from scratch or pre-trained deep learning model are used in this work. Moreover, data augmentation is also used to tackle this weak supervision challenge. A new type of data augmentation is proposed based on underwater sound propagation models depending on the site characteristics. This new data augmentation is used to transfer the knowledge from one PAM database to another. Finally, multimodal deep learning is used to assess marine mammal presence based on visual sights. The relevance of using multiple supervisory signals is demonstrated. All AI methods are developed with the aim of selecting the most relevant samples to help the deep learning model learn better and faster.

Automatic tools proposed in this work target an a posteriori analysis of the environment. Indeed, it is believed that a huge amount of PAM data has not been processed yet in the different laboratories due to its resource intensive requirements. Having the right tools may unlock new analysis of the collected long-term data. The work proposed is exploratory and fell within applied research. It aims at transferring knowledge from differ-

ent research areas to PAM. For that purpose, interdisciplinary and collaboration between different partners is also at the core of this work. Multiple working reports are publicly available to help researchers interested in this work to use it.

Future directions

Big data system

Regarding the big data system development: this work proposes an efficient and scalable big data system for computing FFT-based features which are at the core of most analysis in the PAM community such as underwater soundscape studies, annotation of audio recordings and development of automatic methods. This system is at its early stages and needs more features such as ecoacoustic indices which are resource intensive. Moreover, in order to better understand the long-term contributions of biophony, anthrophony and geophony at a specific underwater site, it should be interesting to join spatio-temporally audio with other data such as automatic identification system of ships or environmental time series like wind speed. Still in this idea of improving soundscape analysis, it should be interesting to develop the variability axis for soundscape parameters to be able to take better decisions in underwater conservation programs. Making all implementations available to the user also helps in open sciences to make soundscape studies reproducible and to help design standards for their analysis. More efficient frameworks can be implemented in other programming languages but it is believed that the proposed big data system enables each researcher willing to use an efficient tool without a huge background of the used technology. Another feature to add to the big data system is the feature computation over time periods that exceed the audio file duration.

For assessing marine mammal vocal presence in audio recordings, it should be interesting to consider on-the-fly clustering methods of FFT-based features. This will help analyze the large amount of audio data without the need of a resource intensive annotation process. This will also allow the user to know time periods with no acoustic presence of marine mammals. It will help generate pseudo noise for supervised detection methods. Unlocking the potential to rerun in small time a lot of scenarios, it should be interesting to migrate all the deep learning methods on such frameworks to enable hyperparameters optimization but also to try different resolutions for automatic method inputs.

Annotation expertise

A first axis of development of APLOSE should be to be connected to a distributed backend system to efficiently compute the tiles, to integrate the hydrophone or system calibration in the computations of the tiles, and to quickly search for a more relevant maximum value for the normalization stage by normalizing the entire audio dataset, and set nominal contrast values around this value before starting a campaign.

A second development axis of APLOSE should be to integrate more features (e.g. a confidence index that the annotator sets to show how confident he/she was when the acoustic event is identified) and an active learning process and evaluate the reduction of the annotation effort thanks to it. In the same vein, it should be interesting to enable the training of one or multiple neural networks on already annotated data of a previous/current campaign, predict the remaining audio data and assess how this method helps reduce the annotation effort but also which neural network influenced the most the annotator.

Moreover, collaborative datasets with multiple annotations should be created to serve as reference in the PAM research. It should also be interesting to propose best practices

for manual annotation process following annotation campaigns. Moreover, the findings proposed in this work based on the collaborative annotation campaign could be adjusted depending on the dataset.

Finally, multi-annotated datasets could help a deep learning model better learn audio representations (curriculum learning Lotfian and Busso (2019)) by first training it with the samples with the highest agreement and then with the smallest.

AI models

All the experiments in this work are presented for specific cases and more thorough investigations should be considered before drawing definitive conclusions about transfer learning and data augmentation techniques.

Moreover, it should be interesting to consider synthetic underwater soundscape with physical-based a priori to train neural networks and help them generalize for new sites.

One type of sound has been addressed in this work, but marine mammal produce variable ones and more investigations should be carried out to train a neural network on a multi-species dataset. In the same vein, multi-class (multiple calls) and multi-label (overlapping acoustic events) DC systems should be investigated instead of binary ones.

Finally, new AI methods should be investigated such as generative adversarial networks, variational autoencoders for data augmentation, and new active learning methods should be proposed to select the most relevant samples for better and faster training AI models. This should help target the most relevant samples to annotate and consequently significantly reduce the annotation process.

Publication list

This thesis consists of several publications, listed in the following, along with their section locations of the thesis in which they appear, or partially appear. The full papers are also included below.

Chapter 1

Peer-reviewed journal article

- **Article 1 Nguyen Hong Duc, P.**, Cazau, D., White, P. R., Gérard, O., Detcheverry, J., Urtizberea, F. and Olivier Adam, O. and The marine soundscape off the North Atlantic French Saint-Pierre-et-Miquelon Archipelago (in revision), *Submitted to Applied Acoustics*

Peer-reviewed conferences

- **Conference 1 Nguyen Hong Duc, P.**, Degurse, A., Adam, O., White, P. R., Gérard, O., Fablet, R. and Cazau, C., A scalable Hadoop/Spark framework for general-purpose analysis of high volume passive acoustic data, IEEE OCEANS 2019, Marseille, France

Working reports

- **Report 1** Cazau, D., **Nguyen Hong Duc, P.** et al. (2019) "Pushing the standards forward in Underwater Passive Acoustics processing for both theory and code" working report , arXiv:1902.06659 (https://arxiv.org/abs/1902.06659)
- **Report 2 Nguyen Hong Duc, P.**, Cazau, D. et al. (2019) "Achieving basic processing of UPA data "at scale - with speed"" OSmOSE working report , arXiv:1903.06695 (https://arxiv.org/abs/1903.06695)
- **Report 3 Nguyen Hong Duc, P.**, Cazau, D., et al., Inform, Compute, Visualize, Estimate: a notebook-basedprocessing chain for Underwater Passive Acoustics, OSmOSE working report (distributed openly on https://osmose.xyz/

Chapter 2

Peer-reviewed journal article

- **Article 2 Nguyen Hong Duc, P.**, Torterotot, M., Samaran, F., White, P. R., Gérard, O., Olivier Adam, O. and Cazau, D., Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics, *Accepted in Ecological Informatics*

Working reports

- **Report 4 Nguyen Hong Duc, P.**, Cazau, D., et al., APLOSE: a scalable web-based annotation tool for marine bioacoustics, OSmOSE working report (distributed openly on https://osmose.xyz/
- **Report 5 Nguyen Hong Duc, P.**, Cazau, D. et al., Assessing inter-annotator agreement from collaborative annotation campaigns in marine bioacoustics, OSmOSE working report (distributed openly on https://osmose.xyz/)

Chapter 3

Peer-reviewed conferences

- **Conference 2** Boittiaux, C., **Nguyen Hong Duc, P.**, Longépé, N., Pensieri, S., Bozzano, R. and Cazau, D. "Multi-modal deep learning models for ocean wind speed estimation", *Accepted in ECML/PKDD 2020 conference (https://ecmlpkdd2020.net/)*

In internal review

- **Report 1 Nguyen Hong Duc, P.**, Adam, O., Gérard, O., White, P. R., and Cazau, D., Data augmentation for marine mammal vocal sound classification: from naive to physically-based schemes
- **Report 2** Cazau, D., **Nguyen Hong Duc, P.**, Druon, J-N., Matwins, S. and Fablet, R., Multimodal deep learning for cetacean distribution modeling. Case study of fin whales (*Balaenoptera physalus*) in the western Mediterranean Sea

Technical report for challenge

- Technicalreport**Nguyen Hong Duc, P.**, Cazau, D., Adam, O., Gérard, O. and White, P. R., Acoustic scene classification using long-term and fine-scale audio representations, DCASE Challenge 2020, task 1 B.

Seminar

- Presentations in a Data Science meet'up Brest (Brest, France): Atelier Machine Learning, Alloha. (10/10/2018, 24/09/2019)
- Presentation in a Data Science meet'up Brest (Brest, France): Ocean & Big Data (27/09/2018)
- Presentation SERENADE, Brest, France (17/10/2018)
- Presentation for the User Day Pôle de Calcul et de Données pour la Mer, Ifremer, Brest (17/05/2018), https://wwz.ifremer.fr/pcdm/L-animation/Journees-utilisateurs/Journee-2018
- Presentation UK-FR PhD student day, DSTL-DGA, Porsthmouth, UK.

Annex

Appendix A

Working report 1

OSmOSE report 1

(GIS) OSmOSE
Open Science meets
Ocean Sound Explorers

Achieving intensive computation of low-level descriptors "at scale - with speed" in Underwater Passive Acoustics

OSmOSE Working Report

Abstract

In the big data era of observational oceanography, passive acoustics datasets are becoming too high volume to be processed on local computers due to their processor and memory limitations. As a result there is a current need for our community to turn to cloud-based distributed computing. We present a scalable computing system for FFT (Fast Fourier Transform)-based features (e.g., Power Spectral Density) based on the Apache distributed frameworks Hadoop and Spark. These features are at the core of many different types of acoustic analysis where the need of processing data at scale with speed is evident, e.g. serving as long-term averaged learning representations of soundscapes to identify periods of acoustic interest. In addition to provide a complete description of our system implementation, we also performed a computational benchmark comparing our system to three other Scala-only, Matlab and Python based systems in standalone executions, and evaluated its scalability using the speed up metric. Our current results are very promising in terms of computational performance, as we show that our proposed Hadoop/Spark system performs reasonably well on a single node setup comparatively to state-of-the-art processing tools used by the PAM community, and that it could also fully leverage more intensive cluster resources with a almost-linear scalability behaviour above a certain dataset volume.

Authorship This document was drafted by

- Paul Nguyen Hong Duc[1)]
- Alexandre Degurse[2]
- Joseph Allemandou[3]
- Dorian Cazau[4)]

belonging to the following institutes (at the time of their contribution): 1) Institut Jean le Rond d'Alembert, Sorbonne Universités, 2) IUEM, Université de Brest, 3) JoalTech, 4) Lab-STICC, ENSTA Bretagne.

Document Review Though the views in this document are those of the authors, it was reviewed by a panel of acousticians before publication. This enabled a degree of consensus to be developed with regard to the contents, although complete unanimity of opinion is inevitably difficult to achieve. Note that the members of the review panel and their employing organisations have no liability for the contents of this document.

The Review Panel consisted of the following experts (listed in alphabetical order):

- Ronan Fablet[1)]

belonging to the following organisms / research institutes (at the time of their contribution): 1) Lab-STICC, IMT Atlantique.

Last date of modifications October 31, 2020

Recommended citation Nguyen, P. et al. "Achieving intensive computation of low-level descriptors "at scale - with speed" in Underwater Passive Acoustics", OSmOSE working report (version dating from October 31, 2020, distributed openly on https://osmose.xyz/)

Future revisions Revisions to this document will be considered at any time, as well as suggestions for additional material or modifications to existing material, and should be communicated to Dorian Cazau (dorian.cazau@ensta-bretagne.fr).

Acknowledgements We thank the Ple de Calcul et de Donnes pour la Mer[1] from IFREMER for the provision of their infrastructure DATARMOR and associated services. We also would like to thank our main sponsors in this work: CominLabs[2] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[3] and ISblue[4]. The authors also would like to acknowledge the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase.

Related publications The content of this document has been partially used in the following publications:

- Nguyen, P., Degurse, A., Allemandou, J., Adam, O., Gerard, O., White, P. R., Fablet, R. and Cazau, D. (2019) "A scalable Hadoop/Spark framework for general-purpose analysis of high volume passive acoustic data", IEEE OCEANS 2019, Marseille, June 17-20.

[1] https://wwz.ifremer.fr/Recherche/Infrastructures-de-recherche/Infrastructures-numeriques/Pole-de-Calcul-et-de-Donnees-pour-la-Mer
[2] https://www.cominlabs.u-bretagneloire.fr/
[3] https://www.afbiodiversite.fr/
[4] https://www.isblue.fr/about-us/

1 Introduction

1.1 Context

Technological progress in observational oceanography gave rise to a two-tiered system in which major strategic investments have been put primarily in data acquisition rather than in data management and processing plans. As a result, there is currently a huge gap between in-situ small-scale data acquisition and a more integrated global knowledge that could be directly used in operational oceanography research and by decision-making managers. A good example of scientific community facing these difficulties is the underwater Passive Acoustic Monitoring (PAM) one, which investigates biological (e.g, whale census) and human (e.g., ship noise monitoring) activities, as well as physical processes (e.g, wind speed and rainfall estimation), in the ocean. Specifically, due to the development of cabled observatories that now provide virtually unlimited power for high bandwidth, continuous data acquisition, and the increase of storage capacity and life battery of temporary recorders, the volume of datasets to process has become larger and larger. For instance, the PerenniAL Acoustic Observatory in the Antarctic Ocean (PALAOA) observatory has been recording quasi-continuously the underwater soundscape of the Southern Ocean since 2005 Boebel et al. (2006), generating about 140 GB per day Kindermann et al. (2008), and the Ocean Network Canada has collected more than 300 TB of PAM data in their database Biffard et al. (2018). In France, governmental agencies like Service Hydrographique et Océanographique de la Marine (SHOM) and Agence Française de la Biodiversité (AFB) are also experiencing similar challenges of processing large volume of data in the Directive Cadre Stratégie pour le Milieu Marin (DCSMM) context, where anthropogenic ambient noise analysis and marine mammal census have to be performed on a long-term continuous effort.

Several projects have started to address the question of processing high volume PAM data more efficiently by adopting distributed computing systems. A distributed computing system can be simply defined as a "system whose components are located on different networked computers (or nodes), which communicate and coordinate their actions by passing messages to one another" [5]. Each computer has its own multiprocessor structure and memory. This makes it good for redundant storage and availability, durability. In contrast, local systems based on a single node, as usually used in the PAM community, all processors may have access to a shared memory to exchange information between processors, like when performing multiprocessor parallel computing. Among well-known distributed environments within the big data space, Apache Spark has become a prominent player. Initially developed in 2012 at the AMPLab at UC Berkeley, Spark is an "open-source distributed general-purpose cluster-computing framework, also providing an interface for programming entire clusters with implicit data parallelism and fault tolerance" [6].

1.2 Contributions

In this paper, we wish to share these efforts by proposing a scalable computation chain for FFT (Fast Fourier Transform)-based features based on the Hadoop and Spark frameworks. These features (e.g., full frequency band Sound Pressure Levels, SPL) are at the core of many different types of acoustic analysis where the need of processing data at scale with speed is evident. For example, these features often serve as long-term averaged learning representations (among them, the well-known Long-Term Spectral Average, LTSA) of soundscapes to identify periods of acoustic interest Erbe et al. (2015); Merchant et al. (2015), which is done either manually or with image-based pattern recognition methods Frasier et al. (2018). Such applications, namely fast automatic content report and interactive annotation of large datasets, need fast and scalable computations of the features to be performed efficiently. Furthermore, LTSA generation relies on several processing parameters (e.g., analysis window size) that can highly modify event-specific averaged patterns and reduce the interpretability of LTSA Hawkins et al. (2014). To better assess this variability, systematic comparative testing of different parameter sets need to be carried out, which also requires intensive computing.

In addition to develop a new scalable Hadoop/Spark system for high performance computing of FFT-features, we also provide a computational benchmark comparing our system against three other computing systems based on different programming languages (Scala, Matlab and Python) in standalone executions, using execution time as evaluation metric. We also evaluate the system scalability in its distributed configuration.

[5] From https://en.wikipedia.org/wiki/Distributed_computing.
[6] From https://en.wikipedia.org/wiki/Apache_Spark.

1.3 Related works and projects

Big data analytics in the UPA community is only in its early stages, but a few major projects have already been launched over the world, from which this work has drawn inspiration and methods. In a nutshell, our proposed system shares with projects like QUT Ecoacoustics and Raven X the view that high speed data processing using parallel and distributed computing is an effective way to deal with large sound archives. In the following, we briefly review such projects and describe our original contributions from them.

1.3.1 With distributed computing systems

Works at University of Tasmania The team of James Montgormery at University of Tasmania has already used Spark Streaming and Hadoop Distributed File System for routine processing (filtering and spectrogram analysis) of large terrestrial datasets Thudumu et al. (2016). Using Hadoop Distributed File System (HDFS) as a distributed storage system, their system resulted in better runtime performance in comparison to standalone execution, with approximately 78% reduction in the execution time. Later, the same team has developed a pipeline of different preprocessing operations to reduce concurrent noise sources in audio recordings, running 21.76 times faster with 32 cores over 8 virtual machines than a serial process Brown et al. (2018).

Our proposed system differs essentially from their more recent work in our use of an off-the-shelf system to achieve parallelisation while they constructed a bespoke master-slave model to achieve a lower level control over data in order to maximise efficiency.

QUT Ecoacoustics QUT Ecoacoustics Research Group[7] designed an open source bioacoustic workbench called QutBioacoustics [8] to provide access to large-scale ecoacoustic datasets. The website successfully allows random-access to any of the ingested audio data. They also advocated colleagues to provide more practical descriptions of on-going research with complex applications are needed. In 2014, their infrastructure contains twelve machines (dual Intel Xeon E5-2665, 32 virtual cores, 256GB DDR3 RAM, 3TB SCSI Raid, dual 1Gb Ethernet) designed to address the needs of researchers working with data that is impractical to process on their personal computers.

Regarding distributed computing system, they have developed a master-slave model with data distribution, where a master node first splits large audio files into smaller chunks, creates a list of work tasks that are distributed over the nodes and eventually aggregates each node output. Such a system has been developed for acoustic event detection and bioacoustic spectral indices Truskinger et al. (2014), improving average execution time by 24x for a 5 instance, 32 thread distributed cluster over a single threaded process. As in Brown et al. (2018), they constructed a bespoke master-slave model to achieve a lower level control over data.

Raven X The Matlab-based Raven-X toolset[9] is to integrate high-performance computing (HPC) technologies and bioacoustics data-mining capabilities, developed by the Center for Conservation Bioacoustics at Cornell Lab.

Raven X differentiates from our approach as they used Matlab Parallel Computing Toolbox and Matlab Distributed Computing Server to develop their system, and focus more on detection and classification algorithms for whale species recognition Dugan et al. (2016). Their most improved process (classifier-based detection) was 6.57 faster for an 8-node server over a serial process.

1.3.2 Other systems and tools

Automated Remote Biodiversity Monitoring Network (ARBIMON) It isa web-based platform used for storage, management, annotation and analysis of audio recordings[10]. All the servers are linked to field audio recorders that send their data to the a local computer station. According to Aide et al. (2013),

[7] https://research.ecosounds.org/

[8] https://github.com/QutBioacoustics/baw-server

[9] https://www.birds.cornell.edu/ccb/raven-x/, see also https://fr.mathworks.com/company/user_stories/cornell-bioacoustics-scientists-develop-a-high-performance-computing-system-for-analyzing-big-data.html.

[10] https://www.sieve-analytics.com/arbimon

other audio files from different recorders can be imported and analyzed on the web-based platform. All the data can be analyzed on the platform. It proposes several features such as the visualisation and annotation of the audio time series transformed into spectrograms, soundscape analysis and specie call identification with Hidden Markov Models. The whole audio processing tool from acquisition to analysis can be performed with this platform and several studies already performed soundscape analysis thanks to it (). Efforts were made to propose tutorials and user guides for beginners. One marine mammal project is under development with this platform. It requires an account but it is free.

Bioacoustica This project[11] proposes an online repository and analysis system for scientific recordings of wildlife sounds. It is hoped that this archive of field and laboratory recordings will become a useful resource for those working on acoustic identification and monitoring whether they work by ear or are developing automated techniques (Baker et al., 2015).

Meridian MERIDIAN (Marine Environmental Research Infrastructure for Data Integration and Application Network)[12] mainly located at Dalhousie University in Nova Scotia is a multi-institutional consortium of ocean researchers, computer and data management professionals that are developing a research data infrastructure to consolidate and support the national and international ocean acoustic and vessel tracking (AIS Automatic Identification System) community. One of the primary goals of MERIDIAN are to develop open-source software solutions for data analysis and visualization, and to assist the community in the use of data science technologies to discover, access, analyze and visualize UPA data. In this PhD, on the big data processing side, the proposed system shares the key points of accessing, analyzing and visualizing UPA data at scale with speed.

IMOS The IMOS Acoustic Data Viewer[13] provides access to sea noise recordings from the IMOS Australian National Mooring Network (ANMN) Passive Acoustic Observatories facility. This infrastructure has been used in research soundscape studies such as (McCauley et al., 2017).

Triton This tool[14] developed by SCRIPPS aimed to quickly review a large dataset via an easy to use graphical user interface (GUI), and also provide a basis for additional features and enhancements (e.g. event detection and classification algorithms). Another software was developed on top on Triton: Silbido (Roch et al., 2011) which aims at automatically detect and extract whistles.

Other tools Other tools exist to explore soundscapes and a complementary and more exhaustive list of those mentioned above can be found in Mooney et al. (2020).

2 Methods

2.1 Proposed Hadoop/Spark-based system

Our proposed distributed computing system, based on the Apache Hadoop and Spark frameworks, is shown in Figure 1.

2.1.1 System overview

Hadoop is responsible for distributed data storage and resource management (including job scheduling/monitoring) accross multiple nodes of a cluster, relying respectively on Hadoop Distributed File System (HDFS) and Yet Another Resource Negotiator (YARN). The main function of HDFS is to divide a data file (e.g. 45-mins long / 169 MB for audio files) into smaller blocks of specified size (default is 64 MB). Each block is processed by one map process and map processes run in parallel. In HDFS, the master is the NameNode

[11] http://bio.acousti.ca/
[12] https://meridian.cs.dal.ca
[13] https://acoustic.aodn.org.au/acoustic/
[14] http://cetus.ucsd.edu/technologies_Software.html

which manages the filesystem namespace and logs all modifications and the state of the filesystem. It communicates all the information about the content of a filesystem to the DataNodes, corresponding to different machines where the HDFS blocks are locally located. Regarding YARN, the Application Master (here the Spark Driver as we will see after) negotiates resources with the ResourceManager, which is responsible for granting containers corresponding to the resources allocated (container Hadoop = N cores + M GB RAM). Containers are then supervised locally by the NodeManagers. Globally, as represented by the dashed arrows in figure 1, both Hadoop components HDFS and YARN communicate with other machines through a master-slave model as follows: NameNode ↔ DataNodes for HDFS and ResourceManager ↔ NodeManagers for YARN.

Spark focuses on processing data in parallel across a cluster. When used in conjunction with Hadoop, the Spark Driver organizes the completion of the jobs across the cluster of executors by interacting with the ResourceManager and the NodeManagers. Jobs are performed across the worker nodes (CPUs (cores) and allocated memory) using Stages and Tasks. Two main components of the Spark architecture are:

- Spark Driver: this is the Application Master in our workflow. It tracks all the operations by executors. Moreover, this entity parses the code, and serializes the byte level code across the executors. Any computation is actually done at the local level by each of them. Furthermore, the Driver aims to plan all the computation in the cluster with Directed Acyclic Graph (DAG). Once a DAG is created, it represents a job which is divided in stages. Then, each stage is carried out as tasks. Finally, the Driver handles fault tolerance of all performed operations;
- Spark Executors: represent processes running in the containers in a cluster. One or more executors could be in each worker node and multiple tasks can be run in a single executor.

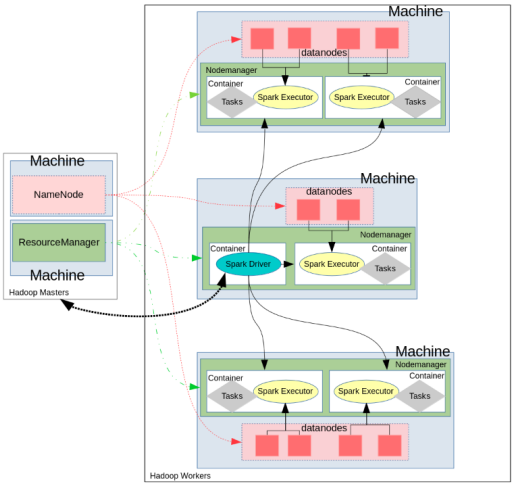


Figure 1: Implementation diagram of the proposed Hadoop/Spark system.

2.1.2 Implementation details

The three parameters *num-executors* (number of executors requested), *executor-cores* (number of tasks the executor can run simultaneously), and *executor-memory* (controlling the executor heap size) play a key role in performance of the Spark system as they control the amount of CPU and memory the application

gets. The parameters of the nominal configuration, called SparkSys, have been set as follows: *executor-cores=3*, *num-executors=8* and *executor-memory=11.5GB*. A more optimized version, which corresponds roughly to the best parameter setup in the balance between system performance and resource allocation, and called SparkSysOpti, have also been tested, with parameters: *executor-cores=3*, *num-executors=17* and *executor-memory=5.5GB*. In particular, it was observed that HDFS client has trouble with a great number of concurrent threads, and that it achieves full write throughput with 5 or less tasks per executor.

When using N node clusters, one node is used as a master and remaining N-1 as slaves. Indeed, when a Spark application is run using a cluster manager like YARN, several daemons will run in the background like NameNode, Secondary NameNode, DataNode, JobTracker and TaskTracker. Thus, while specifying *num-executors*, we need to make sure that we leave aside enough cores (typically 1 core per node) for these daemons to run smoothly. Furthermore, the programming language Scala (version 2.11.8) has been used to implement the tested workflow (described in Sec. 2.2.2), and we also used the multithreaded FFT library JTransforms [15], written in Java. Note that Scala is a programming language that has flexible syntax as compared to other languages like Python or Java, and Apache Spark itself is written in Scala as it is more scalable on Java Virtual Machine.

2.2 Experimental setup

2.2.1 Infrastructure

All our numerical experimentations have been performed on the DATARMOR infrastructure (`http://www.ifremer.fr/pcdm`), belonging to IFREMER. Each node is composed of an Intel Xeon 2X CPU E5-2680 v4 (28c / 56t), 128 GB DDR3, i.e. up to 56 cores (28 $\times$ 2 hyperthreaded CPUs) and 128 GB RAM per node. The multi-node Hadoop-Spark cluster of the SparkSys system was also deployed on DATARMOR. Up to 16 nodes were used to test different distributed configurations o this system. Each node of the cluster runs recent versions of Hadoop and Spark, i.e. 2.8 and 2.4.0, respectively, within the SUSE environment.

Note that the infrastructure architecture of Datarmor may be under-optimal for computational performance based on Hadoop/Spark frameworks, especially for high volume data, as each node does not have its own hard drive, making the data pass through different pipes with limited I/O throughput instead of being read and written locally. Furthermore, there is no dedicated access to node resources, which are instead shared among users (especially in terms of I/O throughput).

2.2.2 Workflow & dataset

The workflow used for the FFT-based feature computations is based on classical PAM analysis blocks (see e.g. Merchant et al. (2015) for background information), including three main steps: short-term FFT analysis (e.g., 32 ms), feature computations and feature integration over longer time segments (e.g., 1 min). Three FFT-based features have been computed: pwelch spectra, Third Octave Levels (TOL) and Sound Pressure Levels (SPL). It is noteworthy that we performed two independent segmentations at different time resolutions to cope with the minimal time window expected in TOL features, set to 1s according to ISO and ANSI standards ISO (1975); ANSI (2009). Also, all the results were sorted by time order and saved in JSON files. A complete description of this workflow, including both theory and implementation details, is available OSmOSE (2019). The two parameter sets used for our experiments are listed in table 1.

The dataset used to evaluate computational performance of computing systems is a real underwater PAM dataset recorded at 32,768 Hz near the archipelago of Saint-Pierre-et-Miquelon over the last three months of the year 2010. It consists of 1807 45-min long wav files for a total volume of 320 GB, each file being 169 MB.

2.2.3 Tested systems

Table 2 describes the three different computing systems tested. SparkSys has already been described in Sec. 2.1. It was benchmarked for computational performance against three other systems based on the Scala (v 2.11.8), Matlab (v 2016b) and Python (v 3.5) programming languages, respectively called ScalaSys, MatlabSys and PythonSys.

[15] https://github.com/wendykierp/JTransforms

Parameter	Set 1	Set 2
nfft	256	1024
windowOverlap	128	0
windowSize	256	1024
recordSizeInSec	1	30
Processing load	2700 / 0.06 / 691213	90 / 1.92 / 86401

Table 1: Parameter sets of the FFT-related variables in the workflow: nfft (number of points in FFT), windowOverlap (number of overlapping samples in consecutive windows), windowSize (number of samples of short-term analysis windows), recordSizeInSec (number of samples of longer time segments over which periodograms are averaged). At the line Processing load we report the : number of integration segments per audio file / volume (in MB) of each segment / number of analysis windows per segment.

For all systems, we tried at best to comply with some "best practices in programming", drawing from template-like codes that are widely used in the PAM community (e.g., the *PAMGuide* toolbox by Merchant et al. (2015)) for the Matlab implementation, and in the data scientist communities (e.g., the Scipy toolbox, https://www.scipy.org/) for the Scala and Python implementations. Double-precision floating-point format has been used in all three implementations. Multiple unitary tests have been performed on the core features of the workflow, and the outputs were cross-validated with a root mean square error below 10^{-16} [16].

In their nominal configurations, ScalaSys is the exact same Scala code as in SparkSys but without the Spark and Hadoop connections, using instead parallel collections included in the Scala standard library for multi-threaded processing, set with a default value of 24 threads. MatlabSys runs parallel Matlab code on a single node using the Parallel Computing Toolbox with 24 workers (equivalent to threads). Due to logistic reasons, we have not tested the use of Matlab Distributed Computing Server in a multiple node setup. Also, on a single node, PythonSys uses Scipy as backend to compute acoustic features, and the Multiprocessing library to run computations in parallel. A distributed version of PythonSys based on the Dask library is being developed and will be reported in future publications.

System	Language (version)	Parallel framework	Distributed framework
SparkSys / SparkSysOpti	Scala (2.11.8)	Pseudo-distributed Spark	Hadoop / Spark
ScalaSys	Scala (2.11.8)	Standard parallel collections	-
MatlabSys	Matlab (2016b)	Parallel Computing Toolbox	-
PythonSys	Python (3.5)	Scipy/Multiprocessing	-

Table 2: Nominal configurations of the tested computing systems. The parallel frameworks are used in a single node setup, while the distributed framework is used for a multiple node setup (only tested for SparkSys).

2.3 Evaluation

To evaluate computational performance of our different systems, execution time was assessed. We paid attention that software launch was not included in the execution time computation. Two different types of experimentations have been run, corresponding either to a parallel (single node) or to a distributed (multiple node) setup, as described in the following. We have also performed sensitivity studies for each system individually, which allows to assess performance variability around the nominal configurations given by table 2. To determine fluctuation in execution times, each tested system configuration has been executed 3 times, then the execution times were averaged over these 3 executions, and the standard deviations were computed.

2.3.1 Single node experimentation

As Spark also supports a pseudo-distributed local mode, we first benchmarked the three computing systems (see table 2) executed in a single node mode, over a linear increase of workloads, from 0.169 GB (1 wav file)

[16] Cross-validation tests can be reproduced following our codes here https://github.com/Project-ODE/FeatureEngine-benchmark/blob/master/run-tests.sh

to 16.9 GB (100 wav files).

2.3.2 Multi-node experimentation

In a second experimentation, we evaluated the scalability of our Hadoop/Spark system using the speed up metric (also referred to as improvement rate in Brown et al. (2018)), which corresponds to the reduction of execution time due to running a fixed workload using an increased number of hardware processors. A linear increase of workloads, from 16.9 GB (50 wav files) to 270.4 GB (1600 wav files), has been used. Note that, comparatively, workloads used for scalability analysis in the literature appear to be very small, e.g. in Thudumu et al. (2016) and Brown et al. (2018) data volumes of less than 5 GB are used. The other systems have not been tested in their distributed configurations.

3 Results & Discussion

3.1 Single node experimentation

3.1.1 Benchmarking of systems

Figures 2 and 3 compare computational performance of the different computing systems in a single node setup, representing execution time (in mins) against workload (GB) for parameter sets 1 (on the left) and 2 (on the right). SparkSysOpti and MatlabSys perform quite similarly, especially for parameter set 1, both largely outperforming ScalaSys and SparkSys, and outperformed by PythonSys. For example, for parameter set 1 and a workload of 16.9 GB (i.e. 100 wav files of our dataset), it takes 2.5 minutes of computation time for SparkSysOpti and MatlabSys, which is more than twice as faster than ScalaSys and SparkSys but more than twice as slower than PythonSys. The difference between SparkSysOpti and MatlabSys is a bit more pronounced on the second set of parameters, although the performance slopes remain quite similar, while PythonSys still outperforms them. Standard deviation values of execution times are relatively minor comparatively to computational gain: 4 s (± 3.2) for SparkSys, 2.5 s (± 2.2) for ScalaSys and 8 s (± 4.8) for MatlabSys.

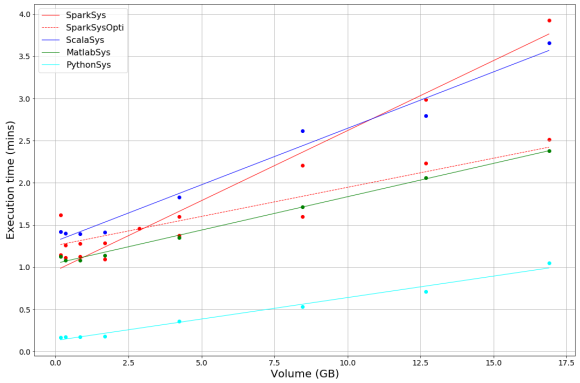


Figure 2: Execution time (mins) against workload (GB) for parameter set 1.

In definitive, SparkSys performs reasonably well in its standalone mode, with performance similar to the Scala-only version. The slight optimization of its core parameters done with SparkSysOpt allows it to scale up easily and reach MatlabSys performance. This result is particularly interesting as it reveals, although the expected advantage of Apache Spark technology is to scale out processing over several nodes, that our system is also valuable on a single-node architecture, which is the most common computer architecture within the PAM community.

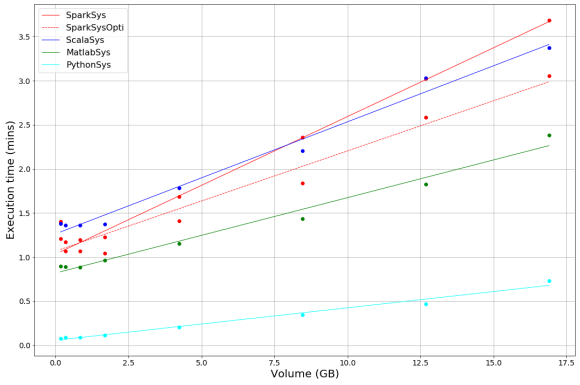


Figure 3: Execution time (mins) against workload (GB) for parameter set 2.

As a reference, we also provide in figure 4 the execution times of each version in a single node / single CPU mode, which corresponds to the minimal computer resource we can set.

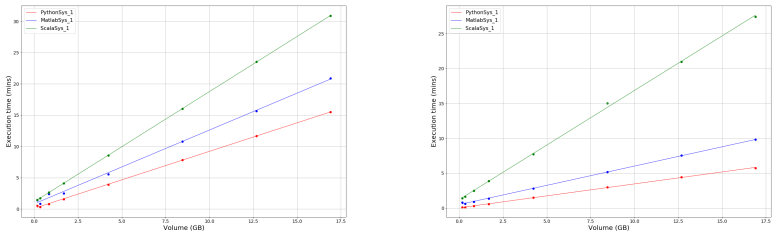


Figure 4: Execution time (mins) against workload (GB) for parameter sets 1 (on the left) and 2 (on the right).

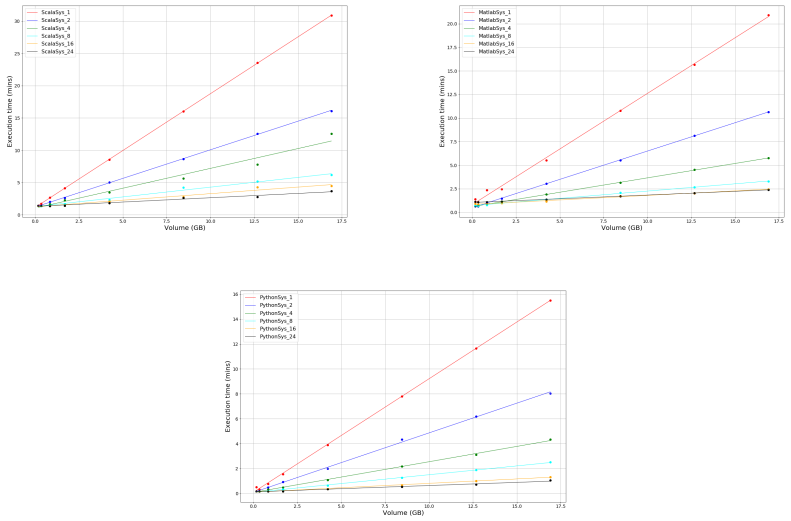


Figure 5: Execution time (mins) against workload (GB) for parameter set 1.

3.1.2 Sensitivity studies of systems

Figure 5 and 6 compare computational performance of each system using different number of threads, written after the underscore in the system names (e.g. "MatlabSys_2" uses two threads).

3.2 Multiple node experimentation

Figures 7 and 8 represent the speed up metric of SparkSysOpti relatively to MatlabSys, as a function of the number of nodes for different workloads (in GB) for parameter sets 1 and 2, respectively. These results show that above 200 GB, an almost linear increase in speed is achieved. Indeed, as the workload increases, speed up linearizes towards the ideal case of scalability represented by the dashed black curve. For example, for parameter set 1 with a 33 GB workload, execution time only decreases by 3 when going from 1 to 4 cluster nodes, and further increasing the number of nodes up to 16 nodes does not decrease this time correspondingly. On the contrary, with a 300 GB workload, a decrease of execution time by almost 12 is observed over the increase of cluster nodes from 1 to 16. As this result has been obtained without specific optimization process, e.g. adapting the number of executors to the split length of audio file, it is very promising for further development towards a more general-purpose cloud-based analytics engine. It is noteworthy here that high-level frameworks like Hadoop / Spark highly facilitate access and democratize the use of distributed computing. On the contrary, frameworks like MPI (Message Passing Interface) would likely need more complex hand-code and fine tuning (e.g. manually setting chunking size, worker task and their synchronization), and require programming skills that are often well beyond competences of most computational scientists and researchers Dunner et al. (2017).

As expected, our proposed system SparkSysOpti does not perform well for small-volume datasets (approximately below 250 GB in our case), as a lot of executors are made available for a small number of tasks, resulting in a lot of unused resources. One way to boost scalability here would be simply to reduce the granularity of computations, i.e. reduce the Hadoop block size so that more tasks are created and more executors

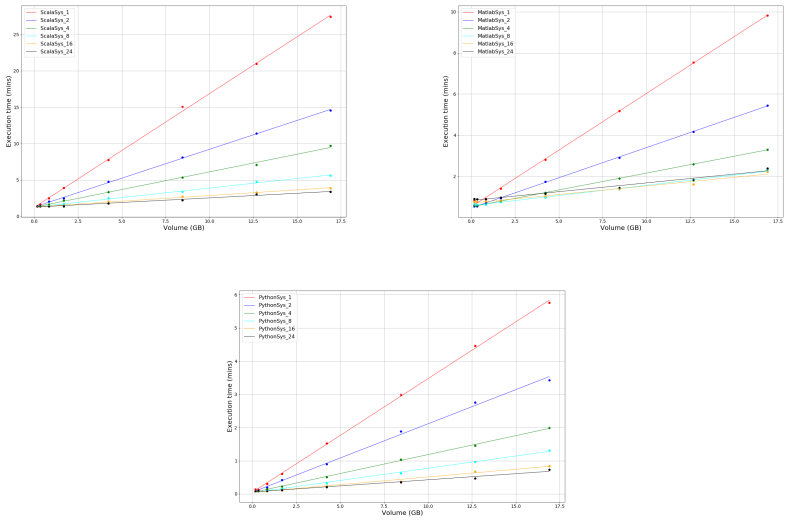


Figure 6: Execution time (mins) against workload (GB) for parameter set 2.

work simultaneously. Similarly, running executors with too much memory often results in excessive garbage collection delays, while running tiny executors (with a single core and just enough memory needed to run a single task, for example) throws away the benefits that come from running multiple in a single JVM.

Also, the different scalability behaviours observed across the different parameter sets can easily be explained. In set 1, the system scalability is stronger than set 2 as the number of operations to be performed (i.e. FFT computations and integration over segmentation windows) per executor for each wav is more important, as we can see at the last line of table 1, resulting in a smaller IO response time for the system (i.e. smaller waiting times for the workers).

4 General discussion

Overall, in addition to this capacity of leveraging complex analytics, we believe that Hadoop and Spark should help to reshape the big data landscape in the field of PAM research for at least three other reasons. First, Spark is able to capture fairly general computations and facilitates the implementation of iterative algorithms, e.g. used for the training algorithms of machine learning systems[17], which now play an important role in most PAM applications (e.g. for whale detection and classification, see the DCLDE workshops). It also facilitates the implementation of interactive/exploratory data analysis (i.e., the repeated database-style querying of data), especially through its SQL-compliant query capability allowing user-defined functions that leverage any general-purpose function to apply to the data columns (e.g. to rank or aggregate rows of data over a sliding window). Such computational functionalities, made here at scale with speed, are now crucial in the context of big ocean data where PAM metrics are processed conjointly with multiple heterogeneous time series from other sensors. As a result, although we focus in this work on simple FFT-based descriptor computations, we envision our Apache Hadoop/Spark big data ecosystem growing as a general-purpose analysis system useful

[17]Spark's machine learning library MLlib, made interoperable with NumPy

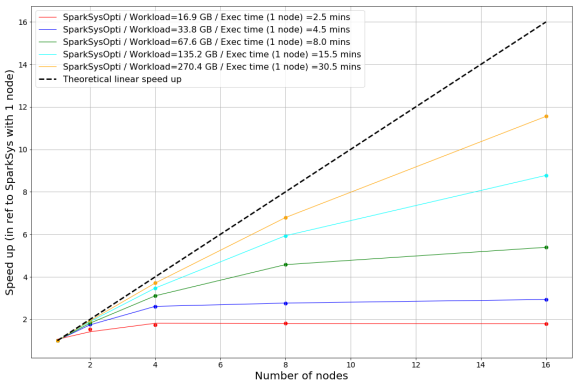


Figure 7: Speed up metric of SparkSysOpti relatively to MatlabSys against number of cluster nodes for parameter set 1 and for different workloads (GB).

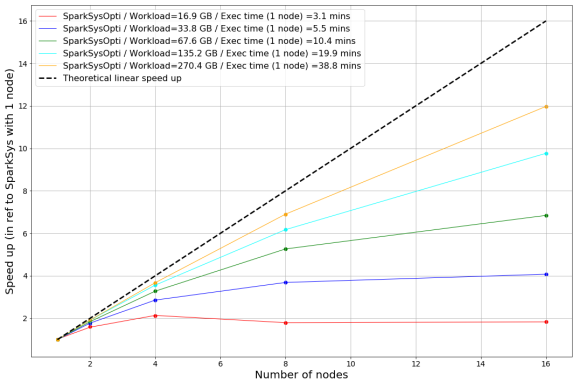


Figure 8: Speed up metric of SparkSysOpti relatively to MatlabSys against number of cluster nodes for parameter set 2 and for different workloads (GB).

for many different types of PAM analysis. Third, numerous efforts have been made so far to outline some best practices for PAM processing (Robinson et al., 2014; Merchant et al., 2015), in the hope of boosting standardization and interoperability. On the contrary to expensive proprietary softwares, we believe that open source software like Apache Spark will strongly contribute to this dynamic, and we would encourage computational scientists and researchers to leave behind them "academic" codes that are too often made unreproducible, unbuildable, undocumented, unmaintained and backward incompatible.

References

Aide, T.M., Corrada Bravo, C., Campos Cerqueira, M., Milan, C., Vega, G., and Alvarez, R. (**2013**). "Real-time bioacoustics monitoring and automated species identification." PeerJ, **1**, e103.

ANSI (**2009**). "Specification for octave-band and fractional-octave-band analog and digital filters." In *ANSI (2009) ANSI/ASA S1.11-2004 (R2009).*

Baker, E., Price, B.W., Rycroft, S.D., and Hill, J. (**2015**). "Bioacoustica: a free and open repository and analysis platform for bioacoustics." Database (Oxford).

Biffard, B., Morley, M., Hoeberechts, M., Rempel, A., Dakin, T., Dewey, R.K., and Jenkyns, R. (**2018**). "Adding value to big acoustic data from ocean observatories: Metadata, online processing, and a computing sandbox." J. Acoust. Soc. Am., **144:3**, 1956–1956.

Boebel, O., Kindermann, L., Klinck, H., Bornemann, H., Plotz, J., Steinhage, D., Riedel, S., and Burkhardt, E. (**2006**). "Acoustic observatory provides real-time underwater sounds from the antarctic ocean." EOS, **87**, 361–372.

Brown, A., Garg, S., and Montgormery, J. (**2018**). "Scalable preprocessing of high volume bird acoustic data." arxiv.

Dugan, P., Klinck, H., A. Roch, M., and Helble, T. (**2016**). "Raven x high performance data mining toolbox for bioacoustic data analysis."

Dunner, C., Parnell, T., Atasu, K., Sifalakis, M., and Pozidis, H. (**2017**). "Understanding and optimizing the performance of distributed machine learning applications on apache spark." IEEE International Conference on Big Data (Big Data).

Erbe, C., Verma, A., McCauley, R., Gavrilov, A., and Parnum, I. (**2015**). "The marine soundscape of the perth canyon." Progress in Oceanography, **137**, 38–51.

Frasier, K.E., Roch, M.A., Hodge, L.E.A., Wiggins, S.M., and Hildebrand, J.A. (**2018**). "Machine learning methods to guide odontocete echolocation insights from large datasets." DCLDE 2018 Official Program, June.

Hawkins, R., Miksis-Olds, J., Bradley, D.L., and Smith, C. (**2014**). "Periodicity in ambient noise and variation based on different temporal units of analysis." In *Proc. Meet. Acoust.* vol. 17.

ISO (**1975**). "Iso 266-1975 (e): Acoustics–preferred frequencies for measurements."

Janik, V.M. (**2005**). *Underwater acoustic communication networks in marine mammals* (Janik VM. . In: , editor. Animal communication networks. Cambridge University Press; 2005. p. 390415.).

Kindermann, L., Boebel, O., Bornemann, H., Burkhardt, E., Klinck, H., Opzeeland, I.and Plotz, J., and Seibert, A.M. (**2008**). "A perennial acoustic observatory in the antarctic ocean, computational bioacoustics for assessing biodiversity." Proceedings of the international expert meeting on IT-based detection of bioacoustical patterns.

McCauley, R.D., Thomas, F., Parsons, M., Erbe, C., Cato, D., Duncan, A., Gavrilov, A., Parnum, I., and Salgado-Kent, C. (**2017**). "Developing an underwater sound recorder: The long and short (time) of it..." Acoustics Australia, **45**.

Merchant, N.D., Fristrup, K.M., Johnson, M.P., Tyack, P.L., Witt, M.J., Blondel, P., and Parks, S.E. (**2015**). "Measuring acoustic habitats." Methods in Ecology and Evolution, **6**, 257–265.

Mooney, T.A., Di Iorio, L., Lammers, M., Lin, T.H., Nedelec, S.L., Parsons, M., Radford, C., Urban, E., and Stanley, J. (**2020**). "Listening forward: approaching marine biodiversity assessments using acoustic methods." Royal Society Open Science, **7**, 201287. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsos.201287.

OSmOSE (**2019**). "Theory-plus-code documentation of depam." Tech. rep., OSmOSE report 1, arXiv:1902.06659.

Robinson, S.P., Lepper, P.A., and Hazelwood, R.A. (**2014**). "Good practice guide for underwater noise measurement." Tech. Rep. Guide No. 133: 95 pp., National Measurement Office, Marine Scotland, The Crown Estate, NPL Good Practice.

Roch, M.A., Scott Brandes, T., Patel, B., Barkley, Y., Baumann-Pickering, S., and Soldevilla, M.S. (**2011**). "Automated extraction of odontocete whistle contours." The Journal of the Acoustical Society of America, **130**, 2212–2223. URL https://doi.org/10.1121/1.3624821.

Thudumu, S., Garg, S., and Montgomery, J. (**2016**). "B2p2: A scalable big bioacoustic processing platform." In *IEEE 18th International Conference on High Performance Computing and Communications.*

Truskinger, A., Cottman-Fields, M., Eichinski, P., Towsey, M., and Roe, P. (**2014**). "Practical analysis of big acoustic sensor data for environmental monitoring." In *IEEE Fourth International Conference on Big Data and Cloud Computing, pp. 91-98.*

Appendix B

Working report 2

OSmOSE report 2

(GIS) OSmOSE

Open Science meets
Ocean Sound Explorers

Pushing the standards forward in Underwater Passive Acoustics processing for both theory and code

OSmOSE Working Report

Abstract

In the Big Data era, the community of PAM faces strong challenges, including the need for more standardized processing tools across its different applications in oceanography, and for more scalable and high-performance computing systems to process more efficiently the everly growing datasets. In this work we address conjointly both issues by first proposing a detailed theory-plus-code document of a classical analysis workflow to describe the content of PAM data, which hopefully will be reviewed and adopted by a maximum of PAM experts to make it standardized. Second, we transposed this workflow into the Scala language within the Spark/Hadoop frameworks so it can be directly scaled out on several node cluster.

Authorship This document was drafted by

- Paul Nguyen Hong Duc[1)]
- Dorian Cazau[2)]

belonging to the following institutes (at the time of their contribution): 1) Institut Jean le Rond d'Alembert, Sorbonne Universités, 2) Lab-STICC, ENSTA Bretagne.

Document Review Though the views in this document are those of the authors, it was reviewed by a panel of acousticians before publication. This enabled a degree of consensus to be developed with regard to the contents, although complete unanimity of opinion is inevitably difficult to achieve. Note that the members of the review panel and their employing organisations have no liability for the contents of this document.

The Review Panel consisted of the following experts (listed in alphabetical order):

- Ronan Fablet[1)]

belonging to the following organisms / research institutes (at the time of their contribution): 1) Lab-STICC, IMT Atlantique.

Last date of modifications October 31, 2020

Recommended citation Nguyen, P. et al. "Pushing the standards forward in Underwater Passive Acoustics processing for both theory and code", OSmOSE working report (version dating from October 31, 2020, distributed openly on https://osmose.xyz/)

Future revisions Revisions to this document will be considered at any time, as well as suggestions for additional material or modifications to existing material, and should be communicated to Dorian Cazau (dorian.cazau@ensta-bretagne.fr).

Acknowledgements We thank the Ple de Calcul et de Donnes pour la Mer[1] from IFREMER for the provision of their infrastructure DATARMOR and associated services. We also would like to thank our main sponsors in this work: CominLabs[2] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[3] and ISblue[4]. The authors also would like to acknowledge the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase.

Related publications The content of this document has been partially used in the following publications:

- Nguyen, P., Degurse, A., Allemandou, J., Adam, O., Gerard, O., White, P. R., Fablet, R. and Cazau, D. (2019) "A scalable Hadoop/Spark framework for general-purpose analysis of high volume passive acoustic data", IEEE OCEANS 2019, Marseille, June 17-20.

[1] https://wwz.ifremer.fr/Recherche/Infrastructures-de-recherche/Infrastructures-numeriques/Pole-de-Calcul-et-de-Donnees-pour-la-Mer
[2] https://www.cominlabs.u-bretagneloire.fr/
[3] https://www.afbiodiversite.fr/
[4] https://www.isblue.fr/about-us/

Introduction

1 Context

Measured noise levels in Passive Acoustic Monitoring (PAM) are sometimes difficult to compare because different measurement methodologies or acoustic metrics are used, and results can take on different meanings for each different application, leading to a risk of misunderstandings between scientists from different PAM disciplines. For reasons of comparability, and since it is cumbersome to define each term every time it is used, some common definitions are needed for acoustic metrics.

In the hope of boosting standardization and interoperability, numerous efforts have already been made to outline some best practices regarding PAM both as an ocean observing measure and as a STIC discipline. Robinson et al. (2014) provided a full technical report of best practices, reviewed by a comitee of experts. Merchant et al. (2015) provided a comprehensive overview of PAM methods to characterize acoustic habitats, and released an open-source toolbox both in R and Matlab with a theoretical document.

2 Contributions

In the same vein, our work addresses the need for a common approach, and the desire to promote best practices for processing the data, and for reporting the measurements using appropriate metrics.

We release a new open source end-to-end analytical workflow for description and interpretation of underwater soundscapes, along with the present document. We outline the following contributions

- this workflow has been implemented in **three different computer languages**: Matlab, Python and Scala. These three implementations perfectly match in regards to the unitary tests done on core functions, with rms error below 10^{-16}, and to the data processing operations and end-user functionalities and results. Note also that in these implementations we try at best to fit with "the best practices in programming" from the DCLDE community in Passive Acoustic Monitoring, for the Matlab implementation, and with the web community and data scientists, for the Scala implementation. These different versions of the workflow have been released on github under a GNU licence;
- in this document, we aligned the lines of codes with their corresponding theoretical signal processing definitions, so as to **fill at best the gap between theory and code**;
- the Scala implementation of the workflow allows for a **direct and transparent scaling out of data processing** over a CPU cluster using the Hadoop/Spark frameworks, allowing for significant computational gain.

As stated in the preamble, this workflow has been collaboratively elaborated, co-developed and reviewed by a research team gathering more than 2 PAM experts over 2 different institutes. Thus, it should provide a reliable value of standardization. Also, during all our work, we built at best on similar works in order to avoid replicating previous efforts. In table 1, we list the different source codes on which we have relied to implement our workflow. In reference to these sources, we systematically highlighted agreements and disagreements with their implementations (and theoretical explanations when present) in the paragraphs named "Discussion", discussed them in regards to each of these different sources and thus justified the choices made for our own implementation.

Eventually, note that reported codes in this document are not representative of their real implementation structure (e.g. in terms of functions), but we rather focus on reporting the essential code lines that implement litteraly each equation and theoretical points.

3 Overview

As shown in figure 1, our workflow is composed of the following blocks

- pre-processing (Sec. 3);

Code source	Language	Main functions used	References
Package scipy v-1.0.0	Python	stft.py / spectrogram.py / welch.py	https://www.scipy.org/
Matlab 2014a	Matlab	spectrogram.m / pwelch.m	MathWorks
pamGuide	R / Matlab	PAMGuide.R	Merchant et al. (2015)

Table 1: Details of codes reviewed.

- segmentation (Sec. 5.3);
- feature computation and integration (Sec. 7.3);

Note that we have two different time scales for data analysis:

- first scale (see Section 5.3): for feature computation in short-term analysis windows of length "windowSize";
- second scale (see Section 7.3): for feature integration in longer time segments, applied when $segmentSize > windowSize$.

Note that when $segmentSize <= windowSize$, these time scales are similar and only one segmentation is performed.

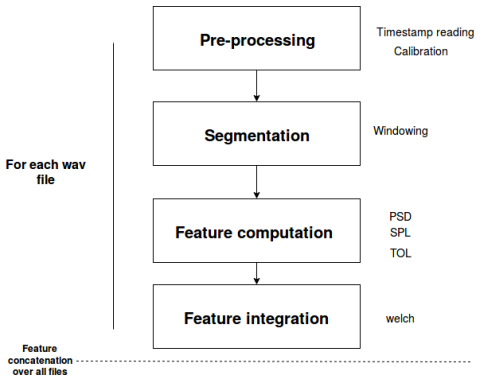


Figure 1: Diagram block.

The implemented acoustic metrics are (selected among the list in (Robinson et al., 2014, Sec. 2.1.2))

- **PSD** Power Spectral Density;
- **TOL** Third-Octave Levels;
- **SPL** Sound Pressure Level

Pre-processing

4 Timestamp reading

4.1 Theory

The CSV file must contain (at least) the following columns:

- filename: "Example0_16_3587_1500.0_1.wav"
- start_date: "2010-01-01T00:00:00Z"

The workflow first imports the list of filenames and only process corresponding audio files. Thus, an audio file not referred into the csv file will not be processed. Note that this metadata organization corresponds to the raw format of several manufacturers of recorders such as AURAL.

4.2 Matlab code

Correspondences with theory Reading the list of filenames from csv is performed at line 3. The structure of audio file metadata is enforced at lines 5-9. No more detailed explanations needed.

```
fid = fopen('../../test/resources/metadata/Example_metadata.csv');
metadataHeader = textscan(fid,'%q_%q', 1, 'delimiter', ',');
metadata = textscan(fid,'%q_%q','delimiter',',');
fclose(fid);
wavFiles = struct(...
    'name', string(metadata{1}),...
    'fs', [1500, 1500],...
    'date', string(metadata{2})...
);
```

Discussion No sources, custom code.

4.3 Python code

Correspondences with theory Reading the list of filenames from csv is performed at line 8. The structure of audio file metadata is enforced at lines 1-7. No more detailed explanations needed.

```
FILES_TO_PROCESS = [{
    "name": file_metadata[0],
    "timestamp": parse(file_metadata[1]),
    "sample_rate": 1500.0,
    "wav_bits": 16,
    "n_samples": 3587,
    "n_channels": 1
} for file_metadata in pd.read_csv(METADATA_FILE_PATH).values]
```

Discussion No sources, custom code.

5 Audio reading and calibration

5.1 Theory

Initially, xin is a digital (bit-scaled) audio signal recorded by the hydrophone, such that the amplitude range is -$2^{N_{bit}-1}$ to $2^{N_{bit}-1}$-1. A first calibration operation is to convert this signal into a time-domain acoustic pressure signal (also called pressure waveform, in Pa, as defined by the International System of Units) as follows:

$$xin = \frac{xin}{10^{\frac{S}{20}}} \quad [Pa] \tag{1}$$

where S is the calibration correction factor corresponding to the hydrophone sensitivity (typically in dB ref 1 V/ μ Pa, with negative values for underwater measurements). Note that it is possible to correct for the variation in the sensitivity with frequency if the hydrophone is calibrated over the full frequency range

of interest [IEC 60565 2006]. When this factor is frequency dependent, it must be applied within spectral features (see eq 10, 16 and 17 in (Merchant et al., 2015, Appendix 1)).

5.2 Matlab code

Correspondences with theory Eq. 1 is performed in line 2.

```
rawSignal=audioread(strcat(wavFileLocation, wavFileName),'double');
calibratedSignal = rawSignal * (10 ^ (calibrationFactor / 20));
```

Discussion Used in the function PG_Waveform.m from PAMGuide (Merchant et al., 2015, eq. 21).

5.3 Python code

Correspondences with theory Eq. 1 is performed in line 2.

```
sound, sample_rate = self.sound_handler.read()
calibrated_sound = sound / 10 ** (self.calibration_factor / 20)
```

Discussion No sources, custom code.

Segmentation

6 Case where $segmentSize > windowSize$

6.1 Theory

We call segmentation the division of the time-domain signal, x, into $segmentSize$-long segments. The s^{th} segment is given by

$$segment^s[n] = xin[n + mN] \tag{2}$$

where N is the number of samples in each window, $0 \leq n \leq N\text{-}1$ (Prentice Hall Inc, 1987) and $0 \leq s \leq S$. For each audio file, a certain number of segments S is obtained, and the last truncated one is removed.

We then perform a short-term division of each segment $segment$ into $windowSize$-long windows, which may be overlapping in time. The m^{th} window is given by

$$xin^m[n] = segment[n + (1 - r)mN] \tag{3}$$

where N is the number of samples in each window, $0 \leq n \leq N\text{-}1$ (Prentice Hall Inc, 1987), r is the window overlap and M is the number of windows in a segment. The last truncated short-term window is removed. A window function is then applied to each data chunk. Denoting the m^{th} windowed data chunk $xin^{(m)}_{win}[n]$

$$xin^{(m)}_{win}[n] = \frac{w[n]}{\alpha} xin^{(m)}[n] \tag{4}$$

where w is the window function over the range $0 \leq n \leq N\text{-}1$, and α is the scaling factor, which corrects for the reduction in amplitude introduced by the window function (Cerna and Harvey, 2000).

Discussion This section has been drawn from (Merchant et al., 2015, Supplementary Material). However, we introduce two successive levels of segmentation, integration-level and short-term window-level, where the second is imbricated into the first one. We follow here the order of segmentations as they appear in numerical implementations, making explicit the truncation problem when $windowSize$ is not an integer multiple of $segmentSize$, which is not transparent in the paragraph of (Merchant et al., 2015, Supplementary Material, sectin 6.4).

7 Case where *segmentSize <= windowSize*

7.1 Theory

In this case, only the short-term segmentation into analysis windows is performed (ie eq. 3 and 4), only now the segment is seen as the full audio file, so that M (in eq. 3) is the number of windows into the complete audio file. Likewise, the last truncated short-term window is removed.

Discussion This section has been drawn from (Merchant et al., 2015, Supplementary Material) without any modifications.

7.2 Matlab code

Correspondences with theory After variable initialization (lines 1-3), eq. 2 is done at line 8 and eq. 4 at line 13. The scaling factor α is included in the variable windowFunction.

```
segmentSize = fix(segmentDuration * fs);
nSegments = fix(wavInfo.TotalSamples / segmentSize);
windowFunction = hamming(windowSize, 'periodic');

% going backwards to have the right struct size allocation of results
for iSegment = nSegments-1 : -1 : 0

    signal = calibratedSignal(1 + iSegment*segmentSize : (iSegment+1) * segment Size);

    nPredictedWindows = fix((length(signal) - windowOverlap) / (windowSize -            windowOverlap));

    % grid whose rows are each (overlapped) segment for analysis
    segmentedSignalWithPartial = buffer(signal, windowSize, windowOverlap, 'nodelay');

    segmentedSignalWithPartialShape = size(segmentedSignalWithPartial);

    % remove final segment if not full
    if segmentedSignalWithPartialShape(2) ~= nPredictedWindows
        segmentedSignal = segmentedSignalWithPartial(:, 1:nPredictedWindows);
    else
        segmentedSignal = segmentedSignalWithPartial;
    end

        % multiply segments by window function
        windowedSignal = bsxfun(@times, segmentedSignal, windowFunction);


    %% FEATURE COMPUTATION
```

Discussion Drawn from the function pwelch.m in Matlab 2014a.

7.3 Python code

Correspondences with theory After variable initialization (line 1), eq. 2 is done at line 2-4 and eq. 4 at lines 5-6. The scaling factor α is included in the function win.

```
nSegments = sound.shape[0] // self.segmentSize
segmentedSound = numpy.split(sound[:self.segmentSize * nSegments], nSegments)
for iSegment in range(nSegments):

        signal=segmentedSound[iSegment]
        shape = (nWindows, windowSize)
    strides = (nWindows * signal.strides[0], signal.strides[0])
    windows = np.lib.stride_tricks.as_strided(signal, shape=shape, strides=strides)
    windowedSignal = windows * windowFunction
```

%% FEATURE COMPUTATION

Discussion Adapted from the function spectrogram in scipy, modifications only done to make this code suitable for our variable names.

Feature Computation

8 PSD (Power Spectral density)

8.1 Theory

The Discrete Fourier Transform (DFT) of the m^{th} segment $X^{(m)}(f)$ is given by

$$X^{(m)}(f) = \sum_{n=0}^{N-1} xin_{win}^{(m)}[n] e^{\frac{-i2\pi fn}{N}} \tag{5}$$

The power spectrum is computed from the DFT, and corresponds to the square of the amplitude spectrum (DFT divided by N), which for the m^{th} segment is given by

$$P^{(m)}(f) = |\frac{X^{(m)}(f)}{N}|^2 \tag{6}$$

where $P^{(m)}(f)$ stands for the power spectrum. For real sampled signals, the power spectrum is symmetrical around the Nyquist frequency, $Fs/2$, which is the highest frequency which can be measured for a given Fs. The frequencies above $Fs/2$ can therefore be discarded and the power in the remaining frequency bins are doubled, yielding the single-sided power spectrum

$$P^{(m)}(f') = 2.P^{(m)}(f') \tag{7}$$

where $0 < f' < fs/2$. This correction ensures that the amount of energy in the power spectrum is equivalent to the amount of energy (in this case the sum of the squared pressure) in the time series. This method of scaling, known as Parseval's theorem, ensures that measurements in the frequency and time domain are comparable. The power spectral density PSD (also called mean-square sound-pressure spectral density) is defined by:

$$PSD(f', m) = \frac{P^{(m)}(f')}{B\Delta f} \quad [\mu\text{Pa}^2\ /\text{Hz}] \tag{8}$$

where $\Delta f = fs/2N$ is the width of the frequency bins, and B is the noise power bandwidth of the window function, which corrects for the energy added through spectral leakage:

$$B = \frac{1}{N} \sum_{n=0}^{N-1} (\frac{w[n]}{\alpha})^2 \tag{9}$$

Note that a spectral density is any quantity expressed as a contribution per unit of bandwidth. A spectral density level is ten times the logarithm to the base 10 of the ratio of the spectral density of a quantity per unit bandwidth, to a reference value. Here the power spectral density level would be expressed in units of dB re 1 μPa^2 /Hz.

Discussion This section has been integrally drawn from (Merchant et al., 2015, Supplementary Material) without any modifications.

8.2 Matlab code

Correspondences with theory Eq. 5 is performed at lines 6-7. Eq. 6 is performed at lines 8. Eq. 7 is performed at lines 9.

```
if (mod(nfft, 2) == 0)
        spectrumSize = nfft/2 + 1;
else
    spectrumSize = nfft/2;
end
twoSidedSpectrum = fft(windowedSignal, nfft);
oneSidedSpectrum = twoSidedSpectrum(1 : spectrumSize, :);
powerSpectrum = abs(oneSidedSpectrum) .^ 2;
powerSpectrum(2 : spectrumSize-1, :) = powerSpectrum(2 : spectrumSize-1, :) .* 2;
psdNormFactor = 1.0 / (fs * sum(windowFunction .^ 2));
powerSpectralDensity = powerSpectrum * psdNormFactor;
welch = mean(powerSpectralDensity, 2);
```

Discussion Drawn from the function pwelch.m in Matlab 2014a.

8.3 Python code

Correspondences with theory Eq. 5 is performed at lines 1-3. Eq. 6 is performed at lines 4-7. Eq. 7 is performed at lines 8-13. Eq. 8 is performed at lines 14-16.

```
rawFFT = np.fft.rfft(windowedSignal, nfft)
vFFT = rawFFT * np.sqrt(1.0 / windowFunction.sum() ** 2)
periodograms = np.abs(rawFFT) ** 2
vPSD = periodograms / (fs * (windowFunction ** 2).sum())
vWelch = np.mean(vPSD, axis=0)
```

Discussion Adapted from the function spectrogram in scipy, with modifications only done to make this code suitable for our variable names.

9 TOL (Third-Octave Levels)

9.1 Theory

Center frequencies can be computed in base-two and base-ten. In our computations, only base-ten exact center frequencies were used. It has to be noted that the nominal frequency is not the exact value of the corresponding center frequency. Readers are referred to Wikipedia (2018) and ISO standards to have the first center frequencies of the TOLs. Center frequencies of the TOLs can be calculated as follow:

$$toCenter = 10^{0.1*i} \tag{10}$$

with i the number of the TOL. In order to determine the bandedge frequencies of each TOL, ANSI and ISO standards give the following equations:

$$\begin{aligned} lowerBoundFrequency &= toCenter \div tocScalingFactor \\ upperBoundFrequency &= toCenter \times tocScalingFactor \end{aligned} \tag{11}$$

with $toCenter$ the center frequency of the TOL and $tocScalingFactor = 10^{0.05}$. From (Merchant et al., 2015, Appendix 1) and Richardson et al. (1995), a TOL is defined as the sum of the sound powers within all 1-Hz bands included in the third octave band (third octave band). Mathematically, according to (Merchant et al., 2015, Supplementary Material), it can be expressed as:

$$TOL(toCenter) = 10log_{10}(\frac{1}{p_{ref}^2} \sum_{f=lowerBoundFrequency}^{f=upperBoundFrequency} \frac{P(f)}{B}) - S(toCenter) \quad (12)$$

For computational efficiency, TOLs are computed by summing the frequency bins of the power spectrum that are included in a TOL. In ISO (1975) and **?** standards, filters with specific characteristics should be designed to compute TOLs with the time-domain signal. For what concerns TOL units, Richardson et al. (1995) and (Merchant et al., 2015, Supplementary Material) disagree about units. For Richardson et al. (1995), correct units are dB re 1 μPa whereas for (Merchant et al., 2015, Supplementary Material), TOL units are dB re 1 μPa or dB re 1 μPa^2 or dB. Note that for accurate representation of third-octave band levels at low frequencies, a long snapshot time is required (sufficient accuracy at 10 Hz requires a snapshot time of at least 30 seconds).

9.2 Matlab code

Correspondences with theory All these conditions are to be met in order to follow the ISO and ANSI standards. TOL are computed for a second and Nyquist frequency cannot be exceeded. Moreover, we have chosen to start our TOL computations with the TOB at 1Hz. However, we are aware that the TOBs under 25 Hz lead to inaccurate computations (Mennitt and Fristrup, 2012). This can be easily modified in that condition $if(lowFreqTOL < 1.0)$.

```
if (length(signal) < sampleRate)
    MException('tol:input', ['Signal incompatible with TOL computation, '...
    'it should be longer than a second.'])
end

if (length(windowFunction) ~= sampleRate)
    MException('tol:input', ['Incorrect windowFunction for TOL, '...
    'it should be of size sampleRate.'])
end

if (lowFreqTOL < 1.0)
    MException('tol:input', ['Incorrect lowFreq for TOL, '...
    'it should be higher than 1.0.'])
end

if (highFreqTOL > sampleRate/2)
    MException('tol:input', ['Incorrect highFreq for TOL, '...
    'it should be lower than sampleRate/2.'])
end

if (lowFreqTOL > highFreqTOL)
    MException('tol:input', ['Incorrect lowFreq, highFreq for TOL, '...
    'lowFreq is higher than highFreq.'])
end
```

After the normalized power spectrum computation, the TOL calculation is done. Eq. 10 and Eq. 11 are done in the following code:

```
tobCenters = 10 .^ ((0:59) / 10);

tobBounds = zeros(2, 60);
tobBounds(1, :) = tobCenters * 10 ^ -0.05;
tobBounds(2, :) = tobCenters * 10 ^ 0.05;
```

We chose to set the TOB centers in order to ba as close as possible to the Scala workflow to have a consistent benchmark. However, in PAMGuide, the TOB centers are set according to the frequency range

set by the user. The 59th TOB center corresponds to about 794328 Hz which is much more greater than standard sampling rate of hydrophones. It has to be noted that this value can also be easily modified.

Eq. 12 is done in the following code:

```
% Find indices of the TOB
inRangeIndices = find((tobBounds(2, :) < sampleRate / 2)...
    & (lowFreqTOL <= tobBounds(2, :))...
    & (tobBounds(1, :) < highFreqTOL));
% Convert indices to match those in the spectrum
tobBoundsInPsdIndex = zeros(2, length(inRangeIndices));
tobBoundsInPsdIndex(1, :) = fix(tobBounds(1, inRangeIndices(1):inRangeIndices(end)) * (nfft / sampleRate));
tobBoundsInPsdIndex(2, :) = fix(tobBounds(2, inRangeIndices(1):inRangeIndices(end)) * (nfft / sampleRate));

tol = zeros(1, length(inRangeIndices));
% Compute TOL
for i = 1 : length(inRangeIndices)
    tol(i) = sum(sum(...
        normalizedPowerSpectrum(1+tobBoundsInPsdIndex(1, i) : tobBoundsInPsdIndex(2, i), :)...
    , 1));
end

tol = 10 * log10(tol);
```

Eq. 10 is done with the or loop in the following code:

```
% Calculate centre frequencies (corresponds to Eq. 4.6 in the User doc and 13 in PAMGuide tutorial)
for i = 2:nband                 %calculate 1/3 octave centre
    fc(i) = fc(i-1)*10^0.1;     % frequencies to (at least) precision
end                             % of ANSI standard
```

Eq. 11 is done at lines 2 and 3:

```
% Calculate boundary frequencies of each band (EQUATIONS 14-15 in PAMGuide tutorial and 4.7 in User doc)
fb = fc*10^-0.05;               %lower bounds of 1/3 octave bands
fb(nfc+1) = fc(nfc)*10^0.05;    %upper bound of highest band (upper
                                %   bounds of previous bands are lower
                                %   bounds of next band up in freq.)
if max(fb) > hcut               %if highest 1/3 octave band extends
    nfc = nfc-1;                %   above highest frequency in DFT,
end
```

Eq. 12 is done in the following code:

```
% Calculate 1/3-octave band levels (corresponds to EQUATION 16 in PAMGuide tutorial and 4.8 in the User doc)
    P13 = zeros(M,nfc);             %initialise TOL array

    for i = 1:nfc                   %loop through centre frequencies
        fli = find(f >= fb(i),1,'first');   %index of lower bound of band
        fui = find(f < fb(i+1),1,'last');   %index of upper bound of band
        for q = 1:M                 %loop through DFTs of data segments
            fcl = sum(Pss(q,fli:fui));%integrate over mth band frequencies
            P13(q,i) = fcl ;         %store TOL of each data segment
        end
    end
    if ~isempty(P13(1,10*log10(P13(1,:)/(pref^2)) <= -10^6))
        lowcut = find(10*log10(P13(1,:)/(pref^2)) <= -10^6,1,'last') + 1;
                                    %index lowest band before empty bands
                                    % at low frequencies
        P13 = P13(:,lowcut:nfc);        %remove empty low-frequency bands
    end
        a = 10*log10((1/B)*P13/(pref^2))-S; %TOLs
clear P13
clear Pss

%% Construct output array
A = 10*log10(mean(10.^(double(a)./10))); % Mean aggregation depending on the length of integration windows
```

Discussion Dranw from PAMGuide (Merchant et al., 2015).

9.3 Python code

Correspondences with theory All these conditions are to be met in order to follow the ISO and ANSI standards as in Matlab codes.

```
# We're using some accronymes here:
#   toc: third octave center
#   tob: third octave band
        if nfft is not int(sample_rate):
            Exception(
                "Incorrect fft-computation window size ({})".format(nfft)
                + "for TOL (should be higher than {})".format(sample_rate)
            )

        self.lower_limit = 1.0
        self.upper_limit = max(sample_rate / 2.0,
                               high_freq if high_freq is not None else 0.0)

        if low_freq is None:
            self.low_freq = self.lower_limit
        elif low_freq < self.lower_limit:
            Exception(
                "Incorrect low_freq ({}) for TOL".format(low_freq)
                + "(lower than lower_limit {})".format(self.lower_limit)
            )
        elif high_freq is not None and low_freq > high_freq:
            Exception(
                "Incorrect low_freq ({}) for TOL".format(low_freq)
                + "(higher than high_freq {}".format(high_freq)
            )
        elif high_freq is None and low_freq > high_freq:
            Exception(
                "Incorrect low_freq ({}) for TOL".format(low_freq)
                + "(higher than upper_limit {}".format(self.upper_limit)
            )
        else:
            self.low_freq = low_freq

        if high_freq is None:
            self.high_freq = self.upper_limit
        elif high_freq > self.upper_limit:
            Exception(
                "Incorrect high_freq ({}) for TOL".format(high_freq)
                + "(higher than upper_limit {})".format(self.upper_limit))
        elif low_freq is not None and high_freq < low_freq:
            Exception(
                "Incorrect high_freq ({}) for TOL".format(low_freq)
                + "(lower than low_freq {})".format(high_freq)
            )
        elif low_freq is None and high_freq < self.lower_limit:
            Exception(
                "Incorrect high_freq ({}) for TOL".format(high_freq)
                + "(lower than lower_limit {})".format(self.lower_limit)
            )
        else:
            self.high_freq = high_freq

        # when wrong low_freq, high_freq are given,
        # computation falls back to default values
```

```
        if not self.lower_limit <= self.low_freq\
                < self.high_freq <= self.upper_limit:

            Exception(
                "Unexpected exception occurred - "
                + "wrong parameters were given to TOL"
            )

        self.sample_rate = sample_rate
        self.nfft = nfft

        self.tob_indices = self._compute_tob_indices()
        self.tob_size = len(self.tob_indices)
```

Eq. 10 and Eq. 11 are done in the following code:

```
    def _compute_tob_indices(self):
        max_third_octave_index = floor(10 * log10(self.upper_limit))

        tob_center_freqs = np.power(
            10, np.arange(0, max_third_octave_index + 1) / 10
        )

        all_tob = np.array([
            _tob_bounds_from_toc(toc_freq) for toc_freq in tob_center_freqs
        ])

        tob_bounds = np.array([
            tob for tob in all_tob
            if self.low_freq <= tob[1] < self.upper_limit
            and tob[0] < self.high_freq
        ])

        return np.array([self._bound_to_index(bound) for bound in tob_bounds])

    def _bound_to_index(self, bound):
        return np.array([floor(bound[0] * self.nfft / self.sample_rate),
                         floor(bound[1] * self.nfft / self.sample_rate)],
                        dtype=int)

    def _tob_bounds_from_toc(center_freq):
    return center_freq * np.power(10, np.array([-0.05, 0.05]))
```

Eq. 12 is done in the following code:

```
    def compute(self, psd):
        third_octave_power_bands = np.array([
            np.sum(psd[indices[0]:indices[1]]) for indices in self.tob_indices
        ])
        return 10 * np.log10(third_octave_power_bands)
```

Discussion To our knowledge, this is is the first Python version of a TOL computation under the ISO and ANSI standards.

10 Sound Pressure Levels

10.1 Theory

Sound Pressure Level (SPL), actually the broadband SPL here, is computed as the sum of PSD over all frequency bins, that is

$$SPL = 10log_{10}(\frac{1}{Bp_{ref}^2}\sum_{f=1}^{nfft} P(f)) \quad (13)$$

with P the single-sided power spectrum (eq. 7), $p_{ref} = 1\mu$ Pa, and B the noise power bandwidth of the window function (B=1.36 for a Hamming window).

Discussion This section has been integrally drawn from (Merchant et al., 2015, Supplementary Material, eq. 17) without any modifications.

10.2 Matlab code

Correspondences with theory Eq. 13 is performed at lines 1

```
SPL = 10*log10(mean(vPSD_int))
```

Discussion No source code has been found for this implementation.

10.3 Python code

Correspondences with theory Eq. 13 is performed at lines 1

```
spl = numpy.array([10 * numpy.log10(numpy.sum(welch))])
```

Discussion No source code has been found for this implementation.

Feature integration

Feature integration is performed in the case where $segmentSize > windowSize$. Note that the timestamp associated with each segment corresponds to the absolute time of the first audio sample in each segment.

11 Welch

11.1 Theory

When averaging noise, it is necessary first to square the data (since sound pressure has both positive and negative excursions, the unsquared data will tend to average to zero). Therefore, the noise values are most often stated as mean square values, or in terms of root mean square (RMS) values. The Welch method (Welch, 1967) simply consists in time-averaging the M PSD from each segment. The resulting representation consists of the mean of M full-resolution segments averaged in linear space.

Note that many other averaging operators (eg median) can be used as detailed in (Robinson et al., 2014, Sec. 5.4.4).

11.2 Matlab code

Correspondences with theory The averaging of PSD is done at the end of each loop (line 4, algorithm 3.2.2).

```
vWelch = mean(vPSD, 2)
```

Discussion No source code has been found for this implementation. Note that Matlab uses a "datawrap" technique that time-averages analysis window and computes only one single FFT in each segment.

11.3 Python code

Correspondences with theory The averaging of PSD is done at the end of each loop (line 4, algorithm 3.2.3).

```
vWelch = np.mean(vPSD, axis=0)
```

Discussion This code has been drawn from the welch function of the scipy package.

References

Cerna, M. and Harvey, A. (**2000**). "The fundamentals of fft-based signal analysis and measurements." Application Note 041. Tech. rep.

ISO, I.S. (**1975**). "Iso 266-1975 (e): Acoustics–preferred frequencies for measurements."

Merchant, N.D., Fristrup, K.M., Johnson, M.P., Tyack, P.L., Witt, M.J., Blondel, P., and Parks, S.E. (**2015**). "Measuring acoustic habitats." Methods in Ecology and Evolution, **6**, 257–265.

Prentice Hall Inc, N., ed. (**1987**). *Marple, S.L.* (Digital Spectral Analysis with Applications).

Richardson, W.J., Greene, C.R., Malme, C.I., and Thomson, D.H. (**1995**). *Marine Mammals and Noise* (Greeneridge Sciences Inc., Editor(s): W. John Richardson, Charles R. Greene, Charles I. Malme, Denis H. Thomson, , Academic Press), chap. ACOUSTIC CONCEPTS AND TERMINOLOGY, pp. 15–32.

Robinson, S.P., Lepper, P.A., and Hazelwood, R.A. (**2014**). "Good practice guide for underwater noise measurement." Tech. Rep. Guide No. 133: 95 pp., National Measurement Office, Marine Scotland, The Crown Estate, NPL Good Practice.

Wikipedia (**2018**). URL https://en.wikipedia.org/wiki/Octave_band#Base_10_calculation.

Appendix C

Working report 3

OSmOSE report 3

(GIS) OSmOSE
Open Science meets
Ocean Sound Explorers

Inform, Compute, Visualize, Estimate: a notebook-based processing chain for Underwater Passive Acoustics

OSmOSE Working Report

Abstract

The recent and ever-growing volume of Underwater Passive Acoustics data to be processed rises the need for the development of new data-driven methodologies and software tools aimed at enabling the wider ocean sciences community to take full advantage of these big acoustic data. These new methods have to be promoted through effective computational workflows and reproducible research practices in ocean sciences so that as many people as possible get involved. In this current report, we focus on the question of how making our developed tools accessible to the largest possible number of persons. Our solution essentially rests on the development of end-to-end general-purpose processing chains, using web-based interactive tools called Jupyter notebooks.

Authorship This document was drafted by

- Paul Nguyen Hong Duc[1)]
- Dorian Cazau[2)]

belonging to the following institutes (at the time of their contribution): 1) Institut Jean le Rond d'Alembert, Sorbonne Universités, 2) Lab-STICC, ENSTA Bretagne.

Document Review Though the views in this document are those of the authors, it was reviewed by a panel of acousticians before publication. This enabled a degree of consensus to be developed with regard to the contents, although complete unanimity of opinion is inevitably difficult to achieve. Note that the members of the review panel and their employing organisations have no liability for the contents of this document.

The Review Panel consisted of the following experts (listed in alphabetical order):

- Ronan Fablet[1)]

belonging to the following organisms / research institutes (at the time of their contribution): 1) Lab-STICC, IMT Atlantique.

Last date of modifications October 31, 2020

Recommended citation Nguyen, P. et al. "Inform, Compute, Visualize, Estimate: a notebook-based processing chain for Underwater Passive Acoustics", OSmOSE working report (version dating from October 31, 2020, distributed openly on https://osmose.xyz/)

Future revisions Revisions to this document will be considered at any time, as well as suggestions for additional material or modifications to existing material, and should be communicated to Dorian Cazau (dorian.cazau@ensta-bretagne.fr).

Acknowledgements We thank the Ple de Calcul et de Donnes pour la Mer[1] from IFREMER for the provision of their infrastructure DATARMOR and associated services. We also would like to thank our main sponsors in this work: CominLabs[2] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[3] and ISblue[4]. The authors also would like to acknowledge the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase.

[1] https://wwz.ifremer.fr/Recherche/Infrastructures-de-recherche/Infrastructures-numeriques/Pole-de-Calcul-et-de-Donnees-pour-la-Mer
[2] https://www.cominlabs.u-bretagneloire.fr/
[3] https://www.afbiodiversite.fr/
[4] https://www.isblue.fr/about-us/

1 Context, motivations and objectives

Advancements in instrumentation over the last decade have greatly expanded our capability to collect underwater passive acoustic (UPA) data over long time spans for ocean monitoring. However, much of these new data remain significantly under-utilized. This acoustic data deluge now rises the need for the development of new data-driven methodologies and software tools aimed at enabling the wider ocean sciences community to take full advantage of these big acoustic data. These new methods have to be promoted through effective computational workflows and reproducible research practices in ocean sciences so that as many people as possible get involved.

In this report, we are describing our efforts to progress down that road through our OSmOSE research group. This group gathers several marine acousticians from different sub-fields, from noise pollution to bioacousticians, passing by marine meteorologists. Our starting point was to choose an appropriate centralized platform to store all raw and generated data, and to host our deployed tools and running analysis.

In previous working reports Nguyen and Cazau (2020) and Nguyen et al. (2020), we already described our distributed computing architecture hosted on Datarmor (IFREMER infrastructure) and associated computational performance. In this current report, we focus on the question of how making our tools accessible by the largest possible number of persons interested in processing UPA data. Our technological answer, detailed in Sec. 2, was the adoption of web-based interactive tools called Jupyter. Besides facilitating access to distributed processing chains, our tools also provide to the user some meta-analysis information such as the average computational time of a given analysis, as well as automatic safeguards to avoid misuses of cluster resources. Our second answer, detailed in Sec. 4, was to identify in a collaborative way common issues from the very bottom of our different standard processing chains. We ended up with the following different processing blocks:

- **inform** : provide, access and analyse all metadata ;
- **compute** : perform intensive computation of low-level "soundscape" descriptors ;
- **visualize** : visualize low-level "soundscape" descriptors ;
- **estimate** : run and benchmark machine learning models for detection and classification of acoustic sources.

2 Backgound on Jupyter notebooks

After a general introduction on Jupyter notebook, we describe two advantages more specific to our UPA community.

2.1 Jupyter notebook : what is it ?

IPython is a Python library that was originally meant to improve the default interactive console provided by Python and to make it scientist-friendly. In 2011, its first release, the IPython Notebook, was introduced. This web-based interface to IPython combines code, text, mathematical expressions, inline plots, interactive figures, widgets, graphical interfaces, and other rich media within a standalone sharable web document.

In 2014, the IPython developers announced the Jupyter project, an initiative created to improve the implementation of the Notebook and make it language-agnostic by design. The name of the project reflects the importance of three of the main scientific computing languages supported by the Notebook: Julia, Python, and R.

As Jupyter notebooks are accessible on the web, meaning they can be easily shared with and reproduced by colleagues. Within a same infrastructure, a code can even be directly connected to the data to be processed, and thus reproduce results in a few clicks.

General-purpose advantages of Jupyter notebook are[5]:

[5]The following has been partially drawn from https://hub.packtpub.com/10-reasons-data-scientists-love-jupyter-notebooks/.

- Language independent: The architecture of Jupyter is language independent. The decoupling between the client and kernel makes it possible to write kernels in any language.
- Easy to create kernel wrappers: Jupyter brings a lightweight interface for kernel languages that can be wrapped in Python. Wrapper kernels can implement optional methods, notably for code completion and code inspection.
- Stress-free Reproducible experiments: Jupyter notebooks can help you conduct efficient and reproducible interactive computing experiments with ease. It lets you keep a detailed record of your work. Also, the ease of use of the Jupyter Notebook means that you dont have to worry about reproducibility; just do all of your interactive work in notebooks, put them under version control, and commit regularly. Dont forget to refactor your code into independent reusable components.
- Effective teaching-cum-learning tool: The Jupyter Notebook is not only a tool for scientific research and data analysis but also a great tool for teaching. An example is IPython Blocks - a library that allows you or your students to create grids of colorful blocks.
- Interactive code and data exploration: The ipywidgets package provides many common user interface controls for exploring code and data interactively.

2.2 Advantage 1 for UPA : easier access to cluster resources

Facing our Big Data challenges, we claim from our own use-case experiences that Jupyter notebooks can be a very interesting gateway to the distributed resources hosted on a cluster for the least technically skilled scientists among us.

Put it another way, notebook is simply a way to avoid console commands. Indeed, for example, hybrid Jupyter notebooks can be developed, which allows for both interactive and distributed batch processing (i.e. through the launching of pbs jobs) within a same environment. Only high-level cluster job (e.g. number of nodes) and processing (e.g. analysis time window) parameters can be set by users, and which can be further tested numerically to have an estimation of used resources and computational time, to make sure that there will not be overuse resources. Such a process makes distributed computing totally transparent for them.

2.3 Advantage 2 for UPA : interactive automatic reporting

The Jupyter Notebook is a web-based interactive environment with the idea of having “all features in one place”, as it combines code, rich text, images, videos, animations, mathematical equations, plots, maps, interactive figures and widgets, and graphical user interfaces, into a single document. Furthermore, it is easy to convert through the nbconvert tool, which converts notebooks to other formats such as HTML and PDF. Another online tool, nbviewer, allows us to render a publicly-available notebook directly in the browser.

Thus, in definitive, the Jupyter Notebook is globally well suited to interactive scientific computing and data analysis, and especially to the compilation of ready-to-be distributed reports which can be endlessly reproduced, modified and updated. In the next section we will describe our current processing chain of four successive modules, each having its own notebook. But one other great advantage of notebooks is to be highly modular, in the sense that any notebook code cell is self-sufficient and can then be executed on its own. This property allows for example anyone to make his own tailor-made notebook, which can be a wrap-up notebook of some results generated along our processing chain. Thus, modularity in notebooks offers an interesting solution to build specialized application-oriented content reports based on more general-purpose processing blocks.

3 Overview of OSmOSE notebooks

Figure 1 presents an overview of OSmOSE backend processing services, which are described in the following.

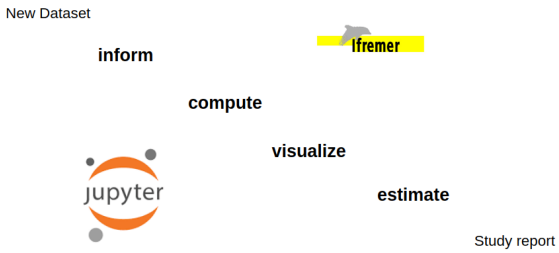


Figure 1: Overview of OSmOSE backend processing services.

3.1 Inform

This first module aims to provide the minimal amount of information on a new dataset. The **inform.ipynb** notebook will :

- check that all data have been properly pre-formatted as expected by the platform ;
- automatically extract from raw data remaining metadata (e.g. sample frequency);
- perform a quality control over both raw files and metadata. This step is able to detect you some anomalies in your campaign recording, such as missing or partially recorded audio files;
- visualize all OSmOSE datasets on a world map.

3.2 Compute

This module allows for an intensive computation of the low-level acoustic descriptors classically used in soundscape analysis (Merchant et al., 2015). It also includes joining operations with auxiliary data. A detailed description of the processing chain and an empirical evaluation of its computational performance has been written elsewhere Nguyen and Cazau (2020) and Nguyen et al. (2020). The **compute.ipynb** notebook allows you to configure :

- your set of analysis parameters (e.g. nfft). You can also provide a list of several values for a given parameter (e.g. segment_duration = [1,5,10,30]), as many jobs will be automatically launched;
- your joining operations if any ;
- your job parameters (e.g. number of cluster nodes), which have a direct impact on computational performance.

In return, the **compute.ipynb** notebook will :

- list you all the configurations of analysis parameters that have already been computed for a given dataset ;
- perform a battery of tests ensuring that your analysis is feasible (e.g. window size should be smaller than your segment duration) ;
- give you an estimate of your job computational time and used resources (i.e. CPU and memory) (see Annex A1 for further details). Note that a job cannot be launched in case of excessive used resources, and the notebook will ask you to re-configure your job with suggestions on which parameters and values ranges to adjust.

Once the job is launched, your notebook **compute.ipynb** and Jupyter Hub session can be closed.

Technical notes The job launcher will start a pbs job to process the dataset. The job will be put in the pbs queue and may not start immediately depending on the usage of the cluster. Using a large number of nodes will delay the start of the job. All computed features are stored in json files in `/home/datawork-osmose/dataset/\textit{dataset\_id}/results/soundscapes/features\_*` where the folder names `/features*` contain the parameter values as follows: `features\_segduration\_winsize\_pctoverlap\_nbernodes\_nberexecSpark_withjoin`.

3.3 Visualize

This module provides you with different visualization methods of low-level descriptors computed previously, as classically used in soundscape analysis (Merchant et al., 2015). The **visualize.ipynb** notebook will :

- plot four different soundscape visualization methods, namely long-term spectrograms, averaged spectra, TOL, concurrently with auxiliary data ;
- give you an estimate of average computational time of this analysis.

3.4 Estimate

This module allows you to estimate which of your audio files contain a significant acoustic activity, a task called acoustic scene detection, and if so which acoustic sources are present, a task called acoustic event classification.

The **estimate.ipynb** notebook thus allows you to :

- use off-the-shelf methods for scene / event detection / classification on your dataset. For the moment, all implemented methods 1) take as input data the acoustic features pre-computed in the previous module, and 2) must be supervised by some external training data, i.e. to be used your dataset must be provided with a file relating your audio filenames and target sources ;
- write your own custom methods and benchmark it with our off-the-shelf methods.

4 Using OSmOSE notebooks

4.1 Pre-requirements

We will guide you step-by-step to be able to upload your dataset in the OSmOSE workspace, located on the infrastructure Datarmor.

4.1.1 Step 0: Datarmor account

You need to have an account on Datarmor, that will provide you with

- extranet logins (username + password)
- intranet logins (username + password)

To get an account, please contact your IT assistant if you are from an institute partner of Datarmor, otherwise contact dorian.cazau@ensta-bretagne.fr.

4.1.2 Step 1: set your Datarmor home directory

When connecting to Datarmor, you are automatically directed to your home workspace. Just copy the source folder for analysis codes in your home directory. Just be sure that you have correct permission rights here. Following the documentation [6], they need to be set as follows:

[6] https://w3z.ifremer.fr/intraric/Mon-IntraRIC/Calcul-et-donnees-scientifiques/Datarmor-Calcul-et-Donnees/Datarmor-calcul-et-programmes/Datarmor-sur-le-web/JupyterHub

- $HOME/ → 700 / rwx—— : using the command >> chmod -R 700 .
- $HOME/.ssh → 700 / rwx——
- public keys (inside $HOME/.ssh/*.pub) → 644 / rw-r–r– : using the command >> chmod -R 644 ./.ssh/*.pub
- private keys (inside $HOME/.ssh) → 600 / rw——- : using the command >> chmod -R 600 ./.ssh

4.1.3 Step 2: connect to Datarmor Jupyter-Hub

All analytics will be done through Jupyter notebooks hosted in the Jupyter-Hub of Datarmor, through the following steps

1. go to https://domicile.ifremer.fr/dana-na/auth/url_default/welcome.cgi and enter your extranet logins
2. select Jupyter on the portal and press Start My Server
3. you have the choice between different job profiles, it will depend on what you want to do:
   - if you want to INFORM and/or COMPUTE features (see Sec. 3.1 and 3.2), select **Datarmor - 1 core, 2GB RAM, 2 hours**;
   - if you want to VISUALIZE and/or ESTIMATE features (see Sec. 3.3 and 3.4), select **Datarmor - 8 cores, 16 GB RAM, 2 hours**
4. after selecting one, enter your intranet logins and you are now on the Jupyter-Hub of Datarmor!

4.1.4 Step 3: copy and paste Notebook folders in your datahome

All our processing services have been written as a suite of notebooks present in `/home/datawork-osmose/notebookSuite_v`. You need to copy the last present version into your home workspace.

4.2 FAQ

4.2.1 How using notebooks on existing OSmOSE datasets ?

To open a notebook, just double click on it. If your notebook kernel is not automatically set to allohaEnv (see Fig. 2), change it in Kernel → Change Kernel → Python [conda env:allohaEnv]

Figure 2: Notebook kernel need to be set to allohaEnv.

Except that, the use of our notebooks should be straight forward for users with basic skills in Python. Below we list a few tips for complete beginners, but the best solution for you might be that you globally increase your programming skills in Python!

1. We did our best to hide from you most of our codes, so you can easily identify the **parts of the notebook that you have to modify**: basically watch everything in red and in green (except the code lines with an import..) ;
2. Be careful to stop running notebooks once you are done (on the Jupyter Hub main page → click tab Running → click Shutdown) ;
3. Avoid running several notebooks simultaneously !

ANNEX

A1. Predicting used resources and computational time

Several job parameters have an important impact on used cluster resources, especially memory (GB) and CPU load (%), on consequently to computational time, as illustrated in figures 3 - 5. The number of nodes and of Spark executors, as well as segment duration, are the most impacting parameters. Also note that these relations are quite dependent on the dataset to be processed.

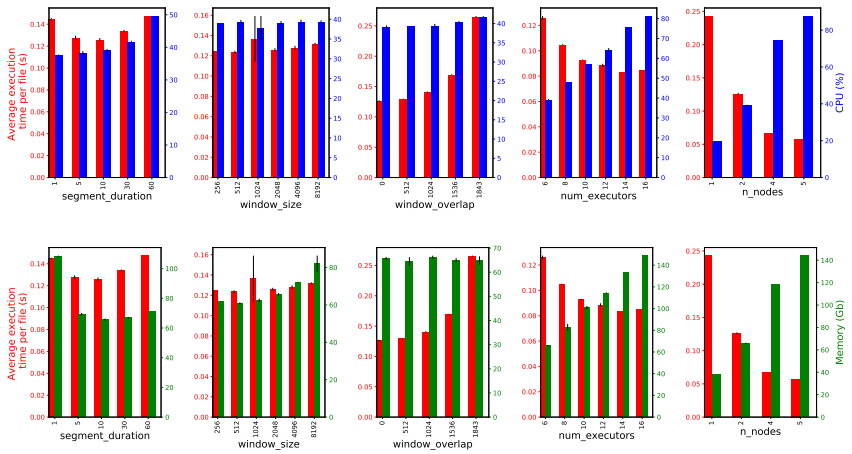


Figure 3: Average computational time per file (in s) against used cluster resources, namely CPU load (%) and memory (in GB) for the dataset chagos.

Our prediction of used resources and computational time results from a multi-linear regression model taking as inputs four job meta-information, namely the volume of a segment (in MB), the number of segments per file, the number of Spark executors per node, and the number of nodes, which have been normalized by z-score. Two models have been trained for each dataset, taking as output either the overall computational time (in s) or CPU load (in %) as output. Predictive performance have been estimated from a stratified 5-fold cross validation, and resulting average and standard deviation are displayed in figure 6. Although the model is quite simple and not optimized, relative error on both prediction tasks remain below 20 %, which is quite satisfactory to meet our current requirements of acting as a safeguard (avoiding misuses of cluster) and of providing approximate information on the processing behaviour.

References

Merchant, N.D., Fristrup, K.M., Johnson, M.P., Tyack, P.L., Witt, M.J., Blondel, P., and Parks, S.E. (**2015**). "Measuring acoustic habitats." Methods in Ecology and Evolution, **6**, 257–265.

Nguyen, P. and Cazau, D. (**2020**). "Pushing the standards forward in underwater passive acoustics processing for both theory and code." Tech. rep., OSmOSE working report (version dating from October 31, 2020, distributed openly on https://osmose.xyz/).

Nguyen, P., Degurse, A., Allemandou, J., and Cazau, D. (**2020**). "Achieving intensive computation of low-

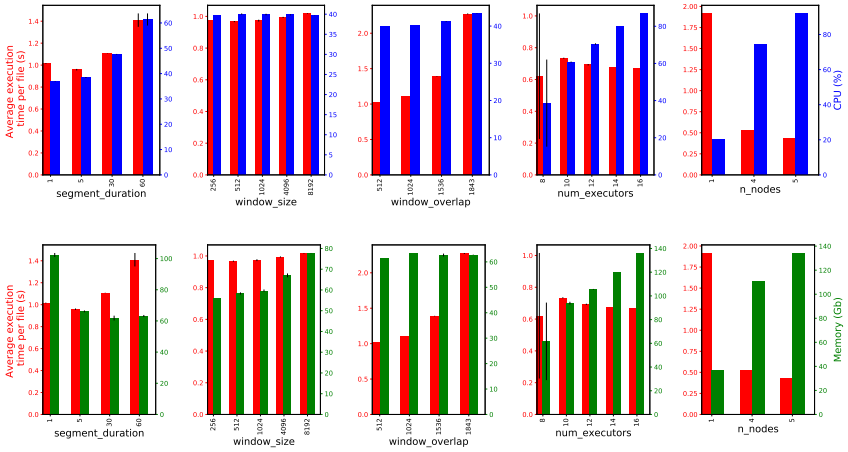


Figure 4: Average computational time per file (in s) against used cluster resources, namely CPU load (%) and memory (in GB) for the dataset SPMAuralB2010.

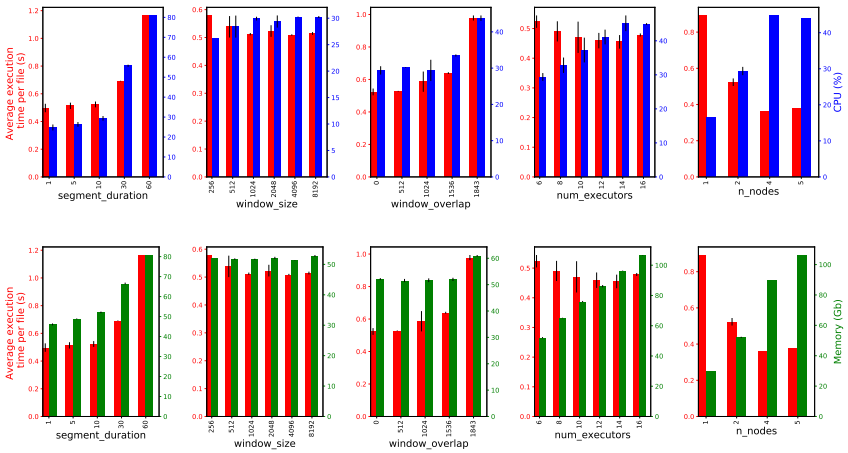


Figure 5: Average computational time per file (in s) against used cluster resources, namely CPU load (%) and memory (in GB) for the dataset argoLOV.

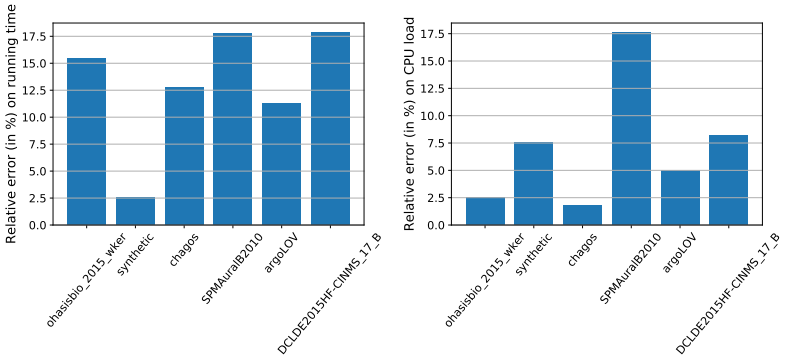


Figure 6: Relative error in the prediction of running time (on the left) and CPU load (on the right) for the different OSmOSE datasets.

level descriptors “at scale - with speed” in underwater passive acoustics.” Tech. rep., OSmOSE working report (version dating from October 31, 2020, distributed openly on https://osmose.xyz/).

Appendix D

Working report 4

OSmOSE report 4

(GIS) OSmOSE
Open Science meets
Ocean Sound Explorers

APLOSE: a scalable web-based annotation tool for marine bioacoustics

OSmOSE Product Presentation

Authorship This document was drafted by

- Paul Nguyen Hong Duc[1)]
- Maëlle Torterotot [2)]
- Romain Vovard[3)]
- Erwan Keribin[4)]
- Dorian Cazau[5)]

belonging to the following institutes (at the time of their contribution): 1) Sorbonne Universités, 2) IUEM, Université de Brest, 3) Freelance developer at Élan Créateur, 4) Freelance developer, 5) Lab-STICC, ENSTA Bretagne.

Document Review Though the views in this document are those of the authors, it was reviewed by a panel of acousticians before publication. This enabled a degree of consensus to be developed with regard to the contents, although complete unanimity of opinion is inevitably difficult to achieve. Note that the members of the review panel and their employing organisations have no liability for the contents of this document.

The Review Panel consisted of the following experts (listed in alphabetical order):

- Ronan Fablet[1)]

belonging to the following organisms / research institutes (at the time of their contribution): 1) Lab-STICC, IMT Atlantique.

Last date of modifications June 9, 2020

Recommended citation Nguyen, P. et al. "APLOSE: a scalable web-based annotation tool for marine bioacoustics", OSmOSE Product Presentation (version dating from June 9, 2020, distributed openly on `https://osmose.xyz/`)

Future revisions Revisions to this document will be considered at any time, as well as suggestions for additional material or modifications to existing material, and should be communicated to Dorian Cazau (dorian.cazau@ensta-bretagne.fr).

Document and code availability This document has been made open source under a Creative Commons Attribution-Noncommercial-ShareAlike license (CC BY-NC-SA 4.0). All associated codes have also been released in open source and access under a GNU General Public License and are available on github (`https://github.com/Project-ODE`).

Acknowledgements We thank the Pôle de Calcul et de Données pour la Mer[1] from IFREMER for the provision of their infrastructure DATARMOR and associated services. We also would like to thank our main sponsors in this work: CominLabs[2] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[3] and ISblue[4]. The authors also would like to acknowledge the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase.

[1] `https://wwz.ifremer.fr/Recherche/Infrastructures-de-recherche/Infrastructures-numeriques/Pole-de-Calcul-et-de-Donnees-pour-la-Mer`
[2] `https://www.cominlabs.u-bretagneloire.fr/`
[3] `https://www.afbiodiversite.fr/`
[4] `https://www.isblue.fr/about-us/`

Contents

Abstract

Lately, underwater passive acoustics has been used extensively for a variety of purposes, from monitoring of marine mammals populations to surveillance of human activities. Recent technological improvements have made it possible to increase the frequency range and autonomy of acoustic recorders, resulting in an enormous amount of data. Manual inspection of these huge datasets is unmanageable and a need for automated recognition, detection and classification methods has emerged. However, in order to train and test the performance of these algorithms, data subsets must still be annotated by human operators. To date, the difficulty of creating large amounts of accurately annotated data has been a major obstacle to build accurate underwater sound recognition algorithms.

In this paper, we present APLOSE, an open-source, web-based yet scalable tool which should highly facilitate collaborative annotation campaigns in marine bioacoustics. This would eventually lead to the creation of reference datasets that could be used to build robust, low-bias detection and classification algorithms.

Chapter 1

Introduction

1.1 Context

A variety of animals produce species-specific acoustic signals, including marine mammals (Richardson et al., 1995), fish (Anorim, 2006), crustaceans (Versluis et al., 2006)... Acoustic analysis has become a standard method in studies of animal vocal communication, and manual detection of acoustic cues was initially the common practice. However, advances in recording hardware speeds, battery life and data storage capacity have increased the rate of acoustic data accumulation to a point where reliance on manual analysis has become unmanageable.

Automated detection and classification algorithms have become necessary for the analysis process. These algorithms provide more consistent and comparable estimates throughout a study period and across studies when processing long-term time series. They are less prone to bias than human analysts, and can be quantified more objectively. However, they cannot be used without supervision, and typically require performance evaluation or correction at some point in the processing pipeline. For instance, labelled datasets are used for training and evaluation of machine learning models, and misclassifications may need to be quantified or corrected (Marques, 2013). Manual review remains also an important part of the process for additional scientific insights, since analysts are best able to judge the context-dependent nature of biological data. Such a supervision often goes through a manual annotation process by one or several analysts. Obtaining such expert annotations is resource intensive and laborious, especially for long recording campaigns as a reliable annotation often needs a careful listening of sounds. Overall, either the lack or quality level of annotated datasets are now frequently criticized (e.g. Leroy (2018), see last slide from `http://cetus.ucsd.edu/dclde/docs/pdfs/Wednesday/14-Gillespie.pdf`), preventing our community to comply with the best practices in machine learning development, with for example the construction of sustainable reference benchmarking datasets.

1.2 Motivations and objectives

In its essence, our project fulfills the general objective of developing an open source and collaborative tool for annotating long-term passive acoustic data, underpinned by the motivation of increasing annotation result and method sharing and visibility. By encouraging a data sharing culture amongst the passive acoustics community, as well as facilitating data discussion and multi-user support, such collaborative environment like APLOSE should help performing better research in less time (Lowndes et al., 2017).

This combination of free and open tools, infrastructure and a collaborative environment has been highlighted as key requirements to lead to more informed conclusions and management decisions, as already observed in similar projects (e.g., Whalenet[1], the Narragansett Bay Coyote StudyWhalenet[2], the Information System for the Analysis and Management of Ungulate Data (Cagnacci and Urbano, 2008)). This

[1] `http://whale.wheelock.edu/Welcome.html`
[2] `http://www.theconservationagency.org/coyote.html`

increased synergy may also reduce the need to collect further or new data and can lead to academic as well as financial gains (Huettmann, 2005; Boulton et al., 2012).

More specifically to our community, at least long-term three research avenues motivated our work. The first general one is to speed up the process of large marine bioacoustics dataset annotation while improving annotation quality, by providing a facilitator tool for collaborative annotation. Indeed, such a collaborative approach can potentially reduce time of a given annotation task by distributing it to multiple annotators, while some overlapping parts of their annotation tasks can be used to cross-validate them. More labels with higher-quality will allow for sure the development of more reliable machine learning algorithms, especially through the creation of larger benchmarking datasets (Piczak, 2015). A second motivation was to be able to generate analytical supports to help understanding annotators' behaviour by capturing the ambiguity that its content might produce in relation to the defined ontology, quantifying individual biases and better characterizing what makes the quality of audio annotations (as already done in other communities like e.g. in urban soundscapes (Cartwright, 2019)). It is known that low-quality human annotations indirectly contribute to the creation of inaccurate or biased learning systems. A third motivation would be to be able to handle large-scale crowd-sourcing campaigns. Crowdsourcing has been shown to be a viable alternative to conventional labelling paradigms to rapidly collect the mass of annotations needed to leverage new data sources. Crowd-sourcing programs increasingly thrive in the marine realm (e.g. Southern Ocean Research Partnership[3], OBSenMER[4], HappyWhale[5]), although it is much less favourable than the terrestrial one where year-round participative surveys can be easily organized, e.g. for birds[6]. Whilst crowdsourcing has many positive aspects including efficiency and cost reduction, the online recruitment of anonymous annotators brings new and different issues especially in relation to the annotation quality.

In line up with these motivations, our development objective was to offer an easy way to annotate complex bioacoustics sound events within dynamic multisource soundscapes, with the requirements of being collaborative, user-friendly, scalable (adapted to large scale datasets), flexible (creation of fitting ontologies for the multiple tasks it can be applied to, easy integration of new apps), as well as open source.

1.3 Related works and contributions

1.3.1 From neighboured communities

To date, a variety of tools have been created for the annotation of audio events, e.g. the famous Praat[7] and HAT[8] for human speech, MUCOSA (P. Herrera and Fabra., 2005) and Sonic Visualizer (C. Cannam and Bello., 2006) for music or more general-purpose tools like ASAnnoatation[9] (Bogaards, 2008) which also provides low-level feature information (e.g. pitch content). These tools allow for a wide variety of analysis, like annotation and generation of music metadata at different abstraction levels including a collaborative annotation subsystem, as well as analysis and feature extraction applications, for MUCOSA. Also, tools to deal with multimedia data have also been developed, such as ELAN (P. Wittenburg and Sloetjes., 2006), a full-featured and complex tool that allows the annotation of both audio and video.

All these tools share the property of being only locally executable. It is only more recently that annotation tools started to get online, particularly impelled by the open-source release of key software components. Among them WaveSurfer (Sjolander and Beskow, 2000)[10], a tool initially designed for speech annotation but deliberately made flexible and extensible to different tasks. Based on it, CrowdCurio[11] is a JavaScript web interface for the annotation of audio events that uses and extends the Regions plugin of wavesurfer.js, including useful features such as labeled regions or the possibility to switch the sound visualization between its waveform and its spectrogram. CrowdCurio has already been adopted in the ISMIR community with Melendez-Catalan et al. (2017)'s BAT interface, allowing the user to label different audio sub-regions with

[3] http://www.marinemammals.gov.au/sorp/sightings
[4] https://www.obsenmer.org/pages/presentation
[5] https://happywhale.com/home
[6] https://www.citizenscience.gov/ebird-bird-data/#
[7] http://www.fon.hum.uva.nl/praat/
[8] http://www.speech.kth.se/hat/
[9] http://recherche.ircam.fr/anasyn/ASAnnotation/
[10] http://www.speech.kth.se/wavesurfer/
[11] https://github.com/CrowdCurio/audio-annotator

overlapping events by giving a salience rate of the different recognized sounds in that segment. Ontologies and cross-annotation are also featured with this tool.

More advanced machine learning oriented development have also been proposed. I-SED (Kim and Pardo, 2017) is a web-based tool for speech recognition involving a machine learning method to help the labeler spotting audio segments similar to the target sound to find. Then, the annotator decides whether or not accepting them. It helps reducing the annotation time. SoundScape (Krijnders and Andringa, 2009) also uses machine learning, and it allows the annotation of specific time-frequency regions of the spectrogram.

1.3.2 Within the bioacoustics community

In the bioacoustics community, similar development trends can be observed. Highly specialized tools have been developed such as the famous Raven Pro software[12]. It is a prevalent software program for the acquisition, visualization, measurement, and analysis of sounds, used by many bioacousticians to annotate their datasets (e.g. Leroy (2018)). The Pro version (Raven Pro), costs from $50 up to $800 dollars depending on the license type and terms. It allows the user to visualise the sound as a waveform and/or a spectrogram. Multiple parameters may be modified by the user, such as the spectrogram windows' sizes, the contrast, the colorbar. Presets can be created, saved and downloaded with custom parameters. The user can also zoom in and out and play the sound at different rates. The annotations are made by drawing a time/frequency box (or time only if the annotations are done on the waveform) around the acoustic event of interest. Annotation are then stored in an 'Annotation Table' with the start and end time and upper and higher frequency values of each "box". Multiple annotation tables can be filled at the same time when annotation is performed on numerous call types simultaneously. The Lite version (Raven Lite) is free and provides the basic functions of Raven Pro, but don't allow for advanced and customizable control of spectrogram parameters and advanced sound measurements and annotation. A more generic open source software is Audacity[13], offering also many features for sound annotation. The user needs the dataset in local to be able to perform audio transformation and annotation.

Like for the other communities, these softwares are not distributed as online interfaces, impeding noticeably collaborative works on a same dataset. Only recently first web-based appear, such as Koe[14], a web-based application for classifying and analysing animal vocalizations (Fukuzawa, 2019). Koe offers bulk-labelling of units via interactive ordination plots and unit tables, as well as visualization and playback, segmentation, measurement, data filtering/exporting and new tools for analysing repertoire and sequence structure - in an integrated environment.

1.3.3 Contributions

Following these efforts, our work aims to provide a new web-based annotation interface dedicated to marine bioacoustics datasets, able to answer some limitations of existing tools in our community. In table 1.1, we list the main non-functional requirements of APLOSE, and below we further discuss how they are addressed by these tools. Note that we left aside more functional features that are generally shared in the system design of these tools, such as exploring the data interactively using smooth zoom and scrolling, providing time/spectral annotation box, spectrogram contrast, audio player with speed variation, user-defined labels... Many of these features were not present on the original CrowdCurio version on which APLOSE is based, although they are crucial to allow for a more flexible annotation practice. Those will be further described and illustrated for APLOSE in Sec. 2.1.

Accessibility

In this first and perhaps most important non-functional requirement, we qualify the capacity of an annotation system to "easily integrate any potential campaign participant". Any web-based technology like APLOSE or Koe will perfectly answer this requirement as it simply comes down to the need of having a web browser on its computer. This should be highly beneficial as sharing of data or granting access to a web application is

[12] `http://ravensoundsoftware.com/software/raven-pro/`
[13] `https://www.audacityteam.org/`, see manual at `https://manual.audacityteam.org/man/spectrogram_view.html`
[14] `https://koe.io.ac.nz/#!`

much easier than sharing data directly, and then avoids the situation where every annotator has to manipulate/install data and tools locally, which may raise complex logistic problems, even more with high-volume datasets. More anecdotally maybe, this requirement will facilitate annotation to anyone working across multiple computers and devices, or not having local administrative access on their computer, which may also become tricky when dealing with proprietary and non-inter-operable softwares.

Scalability

Here, we qualify the capacity of an annotation system to "easily integrate a high number of campaign participants, even if that number shall increase during the campaign". In other words, scalability is the property of a system to handle a growing amount of work by adding resources to the system (Bondi, 2000), and globally to facilitate performance requirement regardless the work load. One dimension (sometimes referred to as **administrative scalability**) on which measuring it is the user ability to access the system for an increasing number of organizations or users. Then, an effective scalability allows to enlarge the previous requirement of accessibility to a very large number of annotators even during a given annotation campaign. This second requirement now allows for the creation of real-time collaborative environment where multiple users can access the same project and can interact with each other.

One of the biggest most valuable and predominant features of modern cloud-based services, as used by APLOSE and Koe, is simplified scalability, being often the primary requirement of IT environments. Especially, horizontal scaling, delivering both performance along with storage capacity, allows a total workload volume to be aggregated over the total number of nodes and latency is effectively reduced.

Performance

Through the performance requirement, we qualify the capacity of an annotation system to "offer a good user experience, even if that experience shall involve high-volume datasets". To have a good user experience the response time of the user interface has to be less then a second, over which users tend to consider them as waiting time[15]. As we shall see in further details in Sec. 3.1, performance becomes critical for certain annotation use cases that need a higher volume of annotation data per displayed window, like for long-term scene classification or context-based event annotation, as waiting times tend to drastically increase with the size of the displayed duration, up to getting too high for a satisfying user experience. For example, a sound file as distributed in the last DCLDE2015HF workshop will take on average more than 20 seconds to be opened on audacity using a standard professional laptop. Here performance can be linked to the notion of **computational scalability** which is the ability of processing more data in a reasonable amount of time. Put in other terms, the processing for visualization purposes should be able to handle large amount of acoustic data with the same performance as for smaller dataset.

APLOSE was developed to be able to handle larger volumes of annotation data than currently done without scarifying performance, thanks to two modern web-based technologies, namely cloud-based services and tile-based rasterization, similarly to the Google Pattern Radio system[16]. First, as already mentioned, cloud-based offers a simplified horizontal scaling, which is ideal for workloads that require reduced latency and optimized throughput as for annotation interface. Koe shares with APLOSE this capacity, while other tested softwares are not cloud-based. Second, as we describe in more details in Sec. 2.2.3, the tiling technology used in APLOSE provides an efficient solution for annotation practices requiring higher volumes of data than currently done, as all spectrograms are pre-computed before the annotation campaign and serve to the client as compressed images of smaller duration tiles (i.e. segments). On the contrary, Koe performs a client-side spectrogram computation which highly limits the processing and rendering of large annotation datasets. But note that this way, Koe achieves full interactivity with browsed data using vector data directly in the browser.

Extensibility

Through the extensibility requirement, we qualify the capacity of an annotation system to "be easily extended to meet new needs while inducing a minimal level of efforts to implement the extension". There are many different dimensions behind this notion, such as cost-effectiveness in terms of both software licensing and

[15] See for example `https://www.nngroup.com/articles/response-times-3-important-limits/` for more details.
[16] `https://patternradio.withgoogle.com/`

hardware. Ideally, a system should be freely distributed with copyleft licenses which guarantee that future versions of the software will remain free and publicly available. This is an important factor for smaller research bodies, consultancies and conservation organisations who have limited resources to dedicate towards software purchasing (Tufto and Cavallini, 2005). Concerning hardware, extending an annotation system so it can meet increasing demands on the number of annotators or volume of data to be processed can induce severe extra-cost if the system architecture has not integrated such extension from scratch.

Once again, one major benefits of cloud computing, as used by APLOSE and Koe, is precisely cost-effectiveness, as it allows a processing activity to grow without making any expensive changes in the current setup thanks to scalability in the cloud, reducing significantly the cost and effort implications of storage and processing resource growth, compared with hosting the system locally. In practical terms, for APLOSE, we just have to commission additional virtual machines to scale out to a larger amount of annotators or data. For what concerns free and open source distribution, as for BAT (Melendez-Catalan et al., 2017), our software development is based on open source software components like CrowdCurio. This is shared by most annotation tools distributed (at least) in the research community, excepting Raven Pro X (although a free version exists).

With a more programming language point of view, extensions can pass through the addition of new functionalities or through the modification of existing ones without impairing existing system functions[17]. Such extensions are then highly facilitated by the use of high-level programming languages like Python and Javascript instead of low-level ones like C++, making them more user-friendly for less technically skilled users. To maximise software sustainability and extensibility, in APLOSE we exclusively used such high-level programming frameworks, and no additional software, add-on packages or plugins are required to run, visualise or export the raw, filtered or analysed data other than the Web browser (e.g. Edge, Firefox and Chrome). On the contrary, tools like Audacity, which are heavily based on C++ to optimize performance, are less suitable to easy system extension.

One last dimension of extensibility would be interoperability, with the aim of having one general purpose application to access and browse through data or even to do data analytic tasks on datasets. Here, again, web applications are well suited for interoperability as they are easier to maintain, update and develop than directly on the operating system installed tools. However, note that this requirement is going along with the performance requirement as one has to deal with the trade-off having a cross-platform application not specialized for one operating system and its lack of performance optimizations available in modern web browsers.

Configurability

At last, through the configurability requirement, we qualify the capacity of an annotation system to "offer a large panel of user-configurable functionalities dedicated to bioacoustic annotation". Raven Prox X is undoubtedly the most exhaustive annotation tool available today towards bioacoustics applications, while Audacity is the less fit-on-purpose.

Although priorities for this first version of APLOSE were more focus on accessibility and performance, we still propose basic campaign management tools like setting up an annotation campaign. We also allow the user to switch between different pre-configured spectrogram resolutions. Contrasting with other softwares, APLOSE is less flexible in the choice of spectrogram resolutions during a campaign, that needs to be pre-set at the campaign creation. However, its resulting advantage is to be highly performing regardless these resolutions and to propose a smooth resolution switching, while in other softwares each demand of a different resolution will ask for the re-computation of the displayed spectrogram, each time occurring a waiting time that can be unreasonably high when dealing with large volume of annotation data.

One original APLOSE feature is to be able to freeze annotation parameters during a campaign, which we found very interesting e.g. to perform parameter-specific sensitivity study. This is done via configurable permissions and access rights which are provided by the data owner to other Aplose users, and to our knowledge is not available in existing tools. Another original feature of APLOSE is the capacity of following the progress of an annotation campaign.

[17] `https://en.wikipedia.org/wiki/Extensibility`

Tools	**Features / Requirements**				
-	Accessibility	Scalability	Performance	Extensibility	Configurability
Raven Pro	X	X	X	X	√
Audacity	X	X	X	X	X
Koe	√	√	X	√	X

Table 1.1: Schematic comparison between desired APLOSE non-functional requirements and related existing features from three different annotation tools of bioacoustics community.

Chapter 2

System overview

2.1 On the user side

In this section we propose to the reader a short tour of APLOSE functionalities on the user side.

2.1.1 Preparing a campaign

User profiles We define two types of user profiles based on their role during the annotation process: the **administrator** (or campaign leader) and the **annotator**. The former is the person who will create the annotation project, upload the data, define the annotation tasks/ontology (i.e. which calls or acoustic event will be annotated) and the campaign parameters (e.g. spectrogram parameters, number of zoom levels, name of the labels...), as well as whether annotators can modify these parameters or not.

The annotators are the persons who use the interface only to annotate the data. Most commonly, they are invited to participate to a campaign with login details and the APLOSE url. They log in with a specific user name and password and then have access to the list of tasks they have to annotate, as illustrated in figure 2.1.

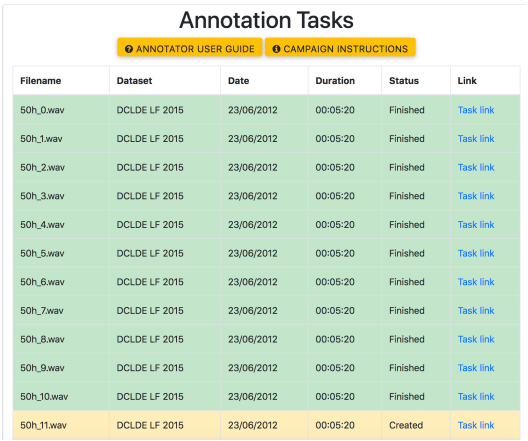


Figure 2.1: Annotation tasks window. The files that have already been annotated appear in green, the other in yellow.

Preparatory data In order to launch an annotation campaign, 3 CSV files are required. These files contain metadata about the the recording campaign and audio data such as recording sites, campaign name, campaign period, sampling rate, duty cycle and others. An exhaustive list of metadata to provide and templates of these files are available on github `https://.....` Each annotation campaign has a unique name and a specific annotation task with predefined labels. A list of predefined annotators is also created. At the time of writing, the number of annotators cannot be modified once the campaign is launched. Moreover, the following parameters have to be pre-defined before launching the campaign :

1. largest duration of a displayed spectrogram ;
2. number of zoom levels, which will define the smallest duration of a displayed spectrogram ;
3. spectrogram configurations with different time and frequency resolutions.

One needs to provide the full set of spectograms to be displayed during the campaign and associated wav files. Note that only WAV audio files are supported for now. Further computational details on spectrogram pre-computing will be provided in Sec. 2.2.2.

2.1.2 During the campaign

Interface description As shown in Figure 2.2, the main page of the site allows to annotate a spectrogram visualization of a sound file. The spectrogram is labelled with time and frequency axes. Several control panels and buttons are available for the annotators to improve their user experience.

1. allows to change the way the spectrogram is displayed. With the select list, the user can choose the way the spectrogram was generated among available settings (nfft, winsize, overlap). A zooming feature is available by two means:
   - Clicking / tapping the buttons on the control panel: the spectrogram is centered on the progress bar
   - Scrolling over the spectrogram with a mouse or a touch pad: the spectrogram is center on the cursor position

   The spectrogram only zooms on time (no zoom on frequency). The zoom is discrete: each zoom level offers pre-computed spectrograms, meaning levels are decided by the creator of the dataset.
2. The sound file is playable by clicking on the play / pause button under the spectrogram. A thin black playback bar is displayed over the spectrogram (not shown on the Figure 2.2). Moreover, the speed at which the sound file is played can be changed from a select list displayed (only on Firefox) next to the play button. Available speeds are 0.25x, 0.5x, 1x, 1.5x, 2x, 3x and 4x. There is no pitch correction so modifying the playback rate also modifies the frequency. Listening to low frequency sounds is allowed thanks to this specific feature.
3. The **Submit & load next recording** button works this way:
   - If several annotations are not tagged, it selects the first one, display an error message and stay on this task
   - If all annotations are tagged (or if no annotation has been created), it saves them for this task, and loads the next available task
   - If there is no next available task, the user is sent back to the task list for this campaign.

   If this task has been annotated and submitted previously, the application will load and display previous annotations. Like new ones, these annotations can be modified or deleted.
4. To create an annotation, click on the spectrogram and drag over the area containing the feature. On click / tap release, the annotation is created and selected: it appears in **Selected annotation block** (4) and in the **Annotations list** (6), both below the spectrogram. Overlap annotations are permitted.

Moreover, annotated event can be listened to (at speed rate set in (2)) by clicking the play button at the upper left of the time-frequency box. It also can be easily removed with the close button at the upper right of the annotation box. A selected annotation block gives precise details about the annotation: start and end time, min and max frequency (4). It also list available tags (from the dataset). To tag / untag the annotation, press the matching button in (4) among the different choices set by the campaign manager (here "Dcall", "40-Hz", "Unknown call"). An annotation must have one and only one tag.

5. All the annotations created by the user for the current task are listed in the annotations list block, sorted by start time. Start / end times, min / max frequencies and the tag are ordered in row in this table. Clicking on an annotation selects it (it appears it the selected annotation block and can be tagged).

6. are links to the user manual for the annotation tool and the campaign instructions given by the campaign manager.

The red button **Back to campaign** at the upper right of the interface leaves the current task and saves the work in progress.

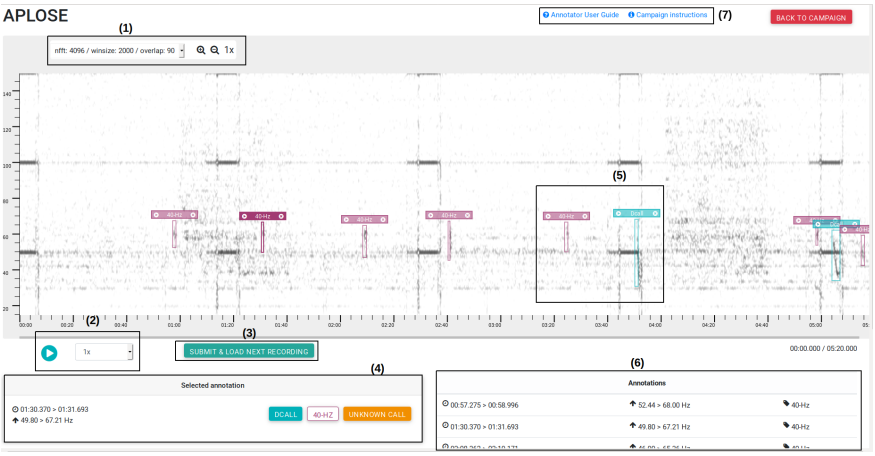


Figure 2.2: Layout of the APLOSE tool

Progress monitoring In the *Annotation Campaign* tab, by clicking on the campaign's name, a dashboard is displayed. It describes the campaign and details the progress of each annotator on this campaign. Administrators have access to a dashboard from where they can perform all these tasks. They can also annotate the data from their account as well as monitor the progress of the annotators and download the annotations as csv files. Annotators can check their own progression within each dataset, and modify or add annotations to the spectrograms. Once every audio file is annotated, the annotator can always go back and edit its annotations until the campaign is ended by the administrator.

2.1.3 Ending a campaign

The *Download CSV results* button creates a CSV file of all the annotations for this campaign. Each line of the file is an annotation, with the following columns:

- **dataset**: dataset short name

- **filename**: task file name
- **start_time**: start time of the annotation box
- **end_time**: end time of the annotation box
- **start_frequency**: min frequency of the annotation box
- **end_frequency**: max frequency of the annotation box
- **annotation**: the tag for the annotation
- **annotator**: the user who did this annotation

2.2 On the development side

2.2.1 Key components

APLOSE is an open-source, web-based tool programmed in JavaScript with React (for the front application) and Node.js (for the feature data API) libraries. It has been dockerized for an easy deployment on servers. The front-end part is heavily inspired by the extended version of wavesurfer used in the CrowdCurio project. Mozilla Firefox is full-featured whereas Google Chrome does not support the sound playback function yet. APLOSE source codes can be found on Github at `https://github.com/Project-ODE/` with a GNU GPL v3.0 license. The repository contains the source code, documentation about the installation and a user guide for annotation task. At the time of writing, the APLOSE system is hosted temporarily on server infrastructure within the OVH private company, with ongoing negotiations to transfer it permanently to the IFREMER server infrastructure named DATARMOR.

2.2.2 Spectrogram generation

Our spectrogram generation method lays on the distributed computing framework detailed in OSmOSE (2019), which include both standardized signal processing definitions and a distributed computing architecture that outperforms classical computing systems in processing high volume at scale with speed. A stand-alone Python custom code has also been written for local generation of spectrograms, available on GitHub `https://github.com/ixio/ODE-Scripts`. This code performs classical audio pre-processing operations such as filtering, amplification to raw audio signals to generate appropriate spectrograms, after pre-segmenting original files into smaller files with the desired duration as spectrogram displayed duration.

Note that current version of APLOSE does not allow to apply hydrophone or system calibration, and spectra are displayed in relative dB scaled to give a 0 dB maximum, such that amplitudes have negative dB values. Regarding this question, one interesting feature of the APLOSE version connected to a distributed back-end like OSmOSE (2019) is to quickly search for a more relevant maximum value (e.g. in a pre-defined biologically-relevant frequency band) with which normalizing the entire audio dataset, and set nominal contrast values around this value before starting a campaign.

2.2.3 Tile-based rasterization

Tiled server side rasterization on the other hand does not have any size limitations beyond your server farms capacity. And even that is not a hard limit as you can either pre-generate all tiles, spending whatever time necessary to do so, or generate tiles and cache them on the fly depending on your specific data and usage patterns. Tile-based rendering solutions have been initially developed for geospatial map applications, allowing to pan and zoom over the whole world. Loading one huge world image and just zoom and pan on it would lead to either bad image qualities on higher zoom levels or allocating too much memory and taking huge amounts of transmission time when using high resolution images. To avoid both problems the world image is split into tiles and stored in a tile tree structure, which typically starts with zoom level zero, which includes every geometry; zoom level 1 has four tiles and every higher zoom level doubles the amount of tiles.

Our requirement of handling large acoustic datasets naturally motivated this solution of pre-rendered raster images in our work. Similarly, the initial acoustic recording (zoom level 0) is sliced into tiles according to the number of zoom levels requested, as illustrated in figure 2.3. Individual spectrograms are computed for all tiles and then concatenated to provide a whole audio spectrogram at each zoom level. Our tile tree structure is a simple directory tree and PNG files on a hard disk.

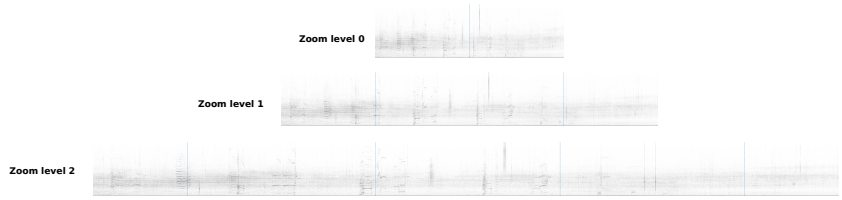


Figure 2.3: Illustration of the different zoom levels in the tile tree.

Chapter 3

Experimental evaluation and use cases

In this section, we first provide short experimental evidence on the performance feature/requirement mentioned in table 1.1 (Sec. 3.1), then we present the seed demo where anyone can test APLOSE on different datasets (Sec. 3.3), and finally we present in broad lines an ongoing research study case about inter-annotator variability using APLOSE (Sec. 3.2).

3.1 Experimental evidence for APLOSE performance

As already mentioned, APLOSE was built with the essential requirement of "offering a good-user experience, even if that experience shall involve high-volume datasets". In the following, we describe a short user-experiment demonstrating that this requirement is better met in APLOSE than in existing softwares, using Raven Pro X as comparative software.

We conducted two annotation campaigns with the DCLDE2015HF dataset (see presentation in Sec. 3.3) using parameters described in table 3.1. The first set of parameters (i.e. max display duration of 300 s, nfft = winsize = 4096, overlap = 90 %) corresponds to a "event-wise annotation setup at fine temporal scale", while the second one (i.e. max display duration of 1080 s, nfft = winsize = 1024, averaging_factor = x 10) corresponds to a "scene-wise annotation setup for long-term". With its high sample frequency of 200 kHZ, the DCLDE2015HF dataset is a good example of high-volume datasets in passive acoustics, where each 5 minute long audio file weights 114 MB.

Experiments were performed using a personal computer MacBookPro 2.7 Ghz i7 16 Go RAM. We must be very clear that these results do not intend to report reliable quantitative metrics but rather rough magnitude of order on waiting times that are significantly perceived by a software user. The measured loading times in Raven Pro X are around 20 s and 6 s per annotation window, respectively for the two campaign setups. Figure 3.1 provides more loading times as a function of the displayed spectrogram duration. Naturally, loading time of the spectrogram increases both with the display duration and the overlap ratio, impacting directly the number of Fast Fourier Transform to be performed, which is the operation with the highest computational cost in a spectrogram display. And of course, the higher the sampling frequency, the longer it takes to compute and load the spectrogram for a given display duration. Comparatively, APLOSE is by construction insensitive to these parameters, which only impact the creation phase of a campaign but not its progress. For these two sets of parameters, APLOSE exhibits waiting times inferior to 1 s between each annotation window, depending mainly on the user internet bandwidth.

3.2 Research study cases

In this section, we describe an ongoing research study aiming to better quantify and understand inter-annotator variability within collaborative annotation campaigns, so as to illustrate the potential of APLOSE of helping researchers to answer more fundamental questions in marine bioacoustics. In the following we briefly describe the annotation campaign setup and illustrative achieved results, and we also provide details on how joining this campaign and reproducing current analytical results.

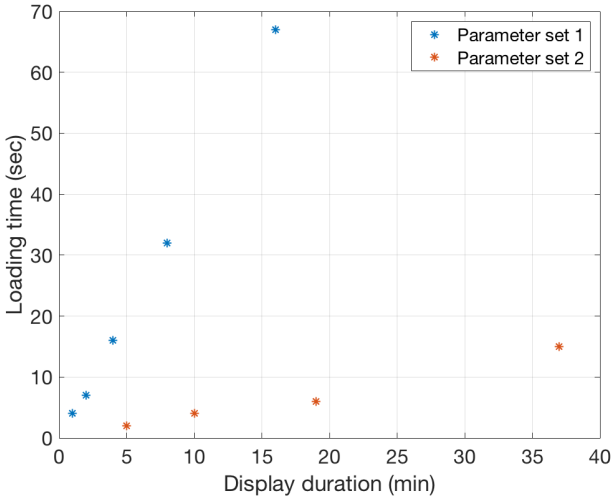


Figure 3.1: Loading times in Raven Pro X using two annotation campaign setups on the DCLDE2015HF as described in table 3.1.

3.2.1 Annotation campaign setup

Dataset The publicly available DCLDE2015LF dataset has been used (see presentation in Sec. 3.3). A 50h-long audio sampled at 2 kHz was splitted into 5min20s long smaller files. The dataset contained 563 small files to annotate. Annotators were asked to identify: D-calls (ref) and 40-Hz pulses (ref). If they have doubt about one of the two calls to find they could use a "Unknown call" label.

Campaign preparation Several parameters were set initially by an expert committee. The contrast and resolution of the displayed spectrograms, audio files were high-pass filtered at 15 Hz and low-pass filtered at 150 Hz, and then pre-amplified with a 35 dB gain, spectrograms were normalized based on the mean of maximum values of PSD from all wav.

Annotating process Annotators were given instructions (see Supplementary Materials) where visual but also aural examples of the sound to annotate were shown. They could refer to it at any time during the annotation process. Annotators could use a zoom from x1 to x8 on the spectrogram, speed up or decrease sounds. Annotations could only be made by visualizing and listening to spectrograms. Waveforms were not displayed for this annotation process.

3.2.2 Some achieved results

Assessing the inter-annotator agreement plays a key role to build more reliable and valid annotated datasets in the underwater passive acoustics community. Here, validity can be defined as the correctness of annotations (i.e "ground truth" or "gold label"). However, in underwater acoustics, there is no ground truth as annotations rely on perception and interpretation of annotators. In order to approximate golden labels, reliability of annotations is measured. Reliability can be defined as a measure of how consistent an annotation is across. Annotations are reliable if their agreement is high. A high reliability is a prerequisite for validity.

Number of annotations per annotator This is the number of acoustic events identified by each annotator. Noisy annotators (randomly annotate data) can be identified if their number of annotations is very different to others (cf Fig. 3.3).

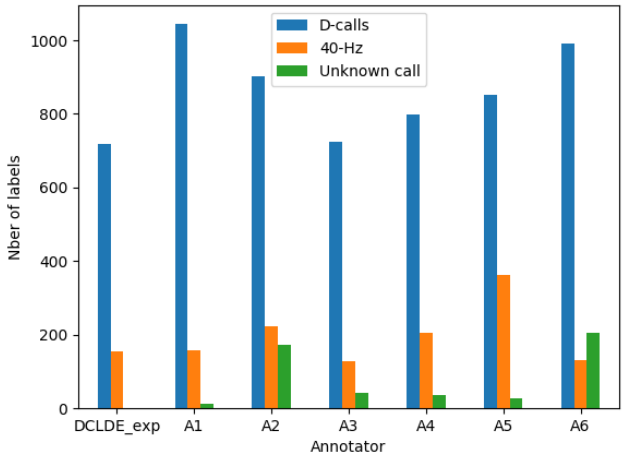


Figure 3.2: Number of annotations per annotator and labels.

Annotation task duration This is the time spent for annotating an audio file. At the end of an annotation campaign, a CSV file containing annotation times for each file per annotator is retrieved.

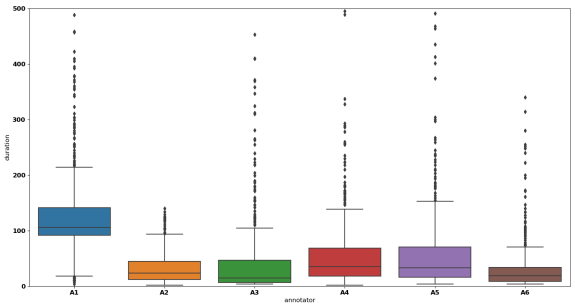


Figure 3.3: Boxplots on duration of an annotation task per annotator.

3.2.3 Joining and updating our DCLDE2015LF campaign

Leveraging the scaling out capacity of APLOSE (see Sec. 1.3.3), our DCLDE2015LF campaign can be joined by anybody, and annotation results automatically updated. To do so, please send a demand to the chief DCLDE2015LF campaign Paul Nguyen (p.nguyenhongduc@gmail.com), and he will give you access details. To our knowledge, these are the first attempts to provide to the marine bioacoustic community sustainable open collaborative annotation campaigns, which can be easily updated over a very long period of time by anybody. A first significant step towards crowd sourcing.

Furthermore, the csv result of any APLOSE campaigns is openly distributed directly on the web interface (even if you have not been enrolled as an annotator of the campaign). We alos distribute a jupyter notebook[1] that computes basic summary statistics on this csv file using python-based panda operators to make it easy for you to get started with such analysis.

3.3 Demonstration version of annotation campaigns

As a last evaluation step of APLOSE, and in complement to the DCLDE2015LF re-annotation campaign described above, we now provide access details to demonstration versions of ongoing annotation campaigns so that interested readers can directly experiment our tool. Three publicly available sub-datasets, described below, are used, exhibiting different sample frequency ranges and spectrogram parameter sets. These parameters are detailed in table 3.1, and have been chosen to cover different annotation use cases, e.g. from fine-scale event annotation to scene annotation over long-term averaged spectrograms. Access link to these demonstrations is `https://demo-aplose.osmose.xyz/` (login: dc@test.ode / pwd: password).

DCLDE2015LF The 2015 Low-Frequency DCLDE dataset (DCLDE2015LF) was recorded with High-frequency Acoustic Recordings Packages deployed off the southern and central coast of California at different locations, spanning all four seasons, over 2009-2013 period (see the full dataset documentation at `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`). The three different recorders (CINMS site B, DCPP sites A and C) were resampled at 2000 Hz, and exhibit different depths of 600 , 65 and 1000 m, respectively. As this dataset was used in the DCLDE international challenge on detection and classification of marine mammals in 2015, it has already been annotated by two independent experts, with a total of 5211 strong labels (i.e. with start and end times of events) over 2 whale species classes that are highly unbalanced: blue whale D calls (4796 samples) and fin whale 40 Hz calls (415 samples).

OrcaSound Our second demonstration dataset is composed of recordings from the open source project Orcasound[2]. Centered within the summertime habitat of the endangered southern resident killer whales, Orcasound Lab is a good place to listen for orcas as well as ships passing through Haro Strait and boats traveling along the west side of San Juan Island.

We ingested in APLOSE one of their test set already annotated by Scott Veirs, that will be re-annotated by 5 other regional experts with the goal of reaching consensus on a label for each SRKW signal (calls only for now, not whistles or clicks). The test set is listed in the orcadata wiki under "Intermediate signal:noise ratio" and is 1/2 hour from 5th July 2019.

DCLDE2015HF The 2015 High-Frequency DCLDE dataset (DCLDE2015HF) was recorded at the same locations as the low-frequency one[3]. For this study, only the November 2009 SOCAL R campaign was analyzed. The mooring depth was 1200m. This dataset was also used in the DCLDE international challenge on detection and classification of marine mammals in 2015. We do not know which annotation guidelines and protocol was used to annotate it. During this campaign, acoustic encounters ("Any period of an animal echolocation that was separated from another one by five minutes or more was marked as a separate encounter. Whistle activity was not considered") from Sperm whale (5 acoustic encounters), Cuvier's beaked whale (2 acoustic encounters) and unidentified odontocete (1 acoustic encounter) were identified.

[1] `https://github.com/ixio/ODE-Scripts`
[2] `https://www.orcasound.net/portfolio/orcasound-lab-hydrophone/`
[3] See the full dataset documentation at `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`.

Parameters	**DCLDE2015LF**	**orcasound**		**DCLDE2015HF**	
Sample frequency (kHz)	2	44.1		200	
Max → min display duration (s) / Zoom level number	**320** → 40/3	**60** → 3.75/4		**300** → 9.375/5	**1080** → 33.75/5
Original file volume (MB)	4.6	10.1		114	412
nfft (samples)	4096	4096	4096	4096	1024
winsize (samples)	2000	1024	4096	4096	1024
overlap (percent)	90	90	0	90	0
averaging factor	0	0	0	0	x 10
APLOSE tile volume (MB)	3.7	27.2	3.2	70.9	26

Table 3.1: Parameter description of the different APLOSE seed datasets. We also provide a comparison between the volumes in MB of original wav files and APLOSE annotation files, containing the pre-filtered audio files to be listened (wav files) to and the pre-computed spectrogram images png files) to be visualized. Note that the maximal spectrogram duration corresponds to the original wav file duration.

Chapter 4

Conclusions & Future directions

In this report, we have introduced the open-source web-based tool APLOSE whose main features are to allow for collaborative annotation campaigns on larger scales, both in terms of annotator numbers and data volume to be processed, than what can currently be done. Potential use cases are to help standardizing annotation processes and to build cross-validated reference datasets for the UPA community. We have demonstrated such needs with its evaluation by highlighting some disagreement between marine bioacoustics experts on a test case study.

We have not mentioned yet that such challenging goals were only made possible through a collaborative partnership between ocean researchers and software engineers, setting up locally at the Technopole Brest Iroise (French Brittany). In the future, we consider automating all preprocessing tasks before launching the annotation campaigns but also the inclusion of active learning methods to reduce the cost of the labelling budget.

Bibliography

Anorim, M.C.P. (**2006**). *Communication in Fishes* (Collin S.P., Moller P., Kapoor BG), chap. Diversity of sound production in fish, p. 71–105.

Bogaards, N.e.a. (**2008**). "Introducing asannotation: a tool for sound analysis and annotation." ICMC.

Bondi, A.B. (**2000**). "Characteristics of scalability and their impact on performance." In *Proceedings of the second international workshop on Software and performance – WOSP '00. p. 195. doi:10.1145/350391.350432. ISBN 158113195X.*

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., and Laurie, G.e.a. (**2012**). "Science as an open enterprise." Tech. rep., The Royal Society Science Policy Centre report 02/12. The Royal Society: London.

C. Cannam, C. Landone, M.S. and Bello., J.P. (**2006**). "The sonic visualizer: A visualization platform for semantic descriptors." In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR-06), pages 324–327, 2006.*

Cagnacci, F. and Urbano, F. (**2008**). "Managing wildlife: a spatial information system for gps collars data." Environ Model Software. 23:957-9.

Cartwright, M.e.a. (**2019**). "Crowdsourcing multi-label audio annotation tasks with citizen scientists." ACM.

Fukuzawa, Y.e.a. (**2019**). "Koe: Web-based software to classify acoustic units and analyse sequence structure in animal vocalizations." Methods Ecol Evol. 2020;11:431–441.

Huettmann, F. (**2005**). "Databases and science-based management in the context of wildlife and habitat: toward a certified iso standard for objective decision-making for the global community by using the internet." J Wildl Manag. 69:466–72.

Kim, B. and Pardo, B. (**2017**). "I-sed: an interactive sound event detector." In *IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces, p. 553-557.*

Krijnders, D. and Andringa, T. (**2009**). "Soundscape annotation and environmental source recognition experiments in assen (nl)." In *Inter Noise.*

Leroy, E.e.a. (**2018**). "On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions." J. Acoust. Soc. Am.

Lowndes, J., Best, B., and Scarborough, C.e.a. (**2017**). "Our path to better science in less time using open data science tools." Nat Ecol Evol 1, 0160.

Marques, T.A.e.a. (**2013**). "Estimating animal population density using passive acoustics." Biol. Rev., **88**, 287–309.

Melendez-Catalan, D., Molina, E., and Gomez, E. (**2017**). "Bat: An open-source, web-based audio events annotation tool." In *Web Audio Conference WAC-2017, August 21–23, 2017, London, UK.*

OSmOSE (**2019**). "Theory-plus-code documentation of depam." Tech. rep., OSmOSE report 1, arXiv:1902.06659.

P. Herrera, J. Massaguer, P.C.F.G.M.K.N.W. and Fabra., U.P. (**2005**). "Mucosa: a music content semantic annotator." Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR-05), 2005.

P. Wittenburg, H. Brugman, A.R.A.K. and Sloetjes., H. (**2006**). "Elan: A professional framework for multimodality research." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), pages 1556–1559, 2006.*

Piczak, K.J. (**2015**). "Esc: Dataset for environmental sound classification, 23rd acm international conference on multimedia, brisbane, australia, oct. 2015, pp. 1015–1018."

Richardson, W.J., Greene, C.R., Malme, C.I., and Thomson, D.H. (**1995**). *Marine Mammals and Noise* (Greeneridge Sciences Inc., Editor(s): W. John Richardson, Charles R. Greene, Charles I. Malme, Denis H. Thomson, , Academic Press), chap. ACOUSTIC CONCEPTS AND TERMINOLOGY, pp. 15–32.

Sjolander, K. and Beskow, J. (**2000**). "Wavesurfer - an open source speech tool." INTERSPEECH, volume 4, pages 464–467.

Tufto, J. and Cavallini, P. (**2005**). "Should wildlife biologists use free software?" Wildl Biol. 2005;11:67–76.

Versluis, M., Schmitz, B., von der Heydt, A., and Lohse, D. (**2006**). "How snapping shrimp snap: Through cavitating bubbles." American Association for the Advancement of Science, **289**, 2114–2117.

Appendix E

Article

Soundscape study

An overview of the Saint-Pierre-et-Miquelon marine soundscape.

Paul Nguyen Hong Duc
Institut d'Alembert
Sorbonne University
Paris, France
p.nguyenhongduc@gmail.com

Olivier Adam
Institut d'Alembert
Sorbonne University
Paris, France
olivier.adam@upmc.fr

Paul R. White
ISVR
University of Southampton
Southampton, UK
prw@isvr.soton.ac.uk

Odile Gerard
DGA Toulon
DGA Techniques Navales
Toulon, France
odile.gerard@intradef.gouv.fr

Dorian Cazau
Lab-STICC
Institut Mines Télécom Atlantique
Plouzané, France
dorian.cazau@imt-atlantique.fr

***Abstract*—This paper presents the first analysis of the Saint-Pierre-and-Miquelon (SPM) archipelago soundscape. Acoustic recordings were sampled at different time periods to highlight seasonal variations over two years. A standard soundscape analysis workflow was used to compute acoustic metrics such as Power Spectral Density (PSD), third-octave levels and sound pressure level (SPL). Results show that the SPM marine soundscape is made up of three main sound sources. Several cetacean species were encountered in the audio recordings such as sperm whales (which were never seen by visual observations, except the stranding of a male individual in 2014), humpback and blue whales. Seasonality was observed in their sound activity. Moreover, an increase of marine traffic was noticed over the two years. Finally, weather conditions, especially the high intensity of rain, came acoustically masking the study of the SPM soundscape. This study can help to regulate human activities, for example to set up new shipping roads to avoid collision with cetaceans during their migration.**

***Index Terms*—underwater soundscape, passive acoustic monitoring, Saint-Pierre-et-Miquelon**

I. Introduction

Underwater "soundscapes" [37], [31] are a valuable source information as they are the result of complex interactions between biophony, geophony and anthrophony at various spatio-temporal scales. Characterizing an underwater soundscape can provide an insight of the situation of the ocean ecosystem of a particular place. Describing the soundscape and then understanding relationships between these different types of sounds may help to protect marine ecosystems from anthropogenic activities which are increasing over the years. This marine anthropogenic sound pollution is a threat for the fauna, especially for marine mammals [43]. Indeed, sound is a key element in marine life. For example, marine mammals perceive and generate sounds for foraging, social interactions, mating, escaping from predators and also navigating [13]. Moreover, it has been shown that marine mammals are affected by the deterioration of the sound underwater environment, mainly due to the increasing maritime traffic [33], [39].

Underwater soundscapes can bring complementary information about life activities from underwater or aerial visual observations [44]. Sometimes because of the access to the ocean (due to bad weather conditions, lack of observers, or too much expensive outcomes to organise field missions), Passive Acoustic Monitoring (PAM) becomes an attractive alternative. Moreover, marine spatio-temporal trends can be identified at different scales and acoustic variations in this environment can be spot with PAM [24].

Soundscape ecology is a new emerging field of research but it requires relevant but also accurate soundscape assessments as basics. Several methods has been employed to characterize the underwater soundscapes. Acoustic indices were first used in the assessment of terrestrial soundscapes [47] but they were also used for marine soundscapes [29], [36], [7]. However, these indices seem to be well-designed for coral reef [21] but not for all marine environments [15]. Researchers views are divided on the relevance on the use of acoustic indices in marine soundscapes and it is suggested to not use them alone but to correlate them with other descriptive metrics such as long-term spectrograms [35]. Nevertheless, a new kind of analysis is being proposed: characterizing soundscapes thanks to sparse coding methods such as Principal Latent Component Analysis (PLCA) [15] or with Nonnegative Matrix Factorization (NMF) [25]. These methods should take into account time and frequency variations whereas acoustic indices can only catch one of these two separately.

Furthermore, to enhance the long-term soundscape studies, other environmental variables were provided in the analysis. Trends in sound level across years and seasonality patterns could be revealed by correlating descriptive metrics such as SPL and PSD with wind speed [12], [2], [27], [38], [40], moon phase [46], temperature [8], [40], day / night variations [22], [18], [40], ship traffic [48], [19], [40], [38]. In addition to these environmental variables, inter-sites studies were

PhD is granted by the Direction Générale des Armements (DGA).

carried out and different seasonality or sound patterns were spotted [45], [32], [40], [41], [5], [26]. Principal Component Analysis (PCA) [4], [42] but also different statistical tests such as Krusker-Wallis test, Mann-Whitney U tests were used to provide these in-depth analysis. Nonetheless, it is not explicitly described how environmental variables are fitted to acoustic ones except for [40] who describe their process by interpolation to adjust environmental variables to acoustic data.

We deployed a semi-permanent acoustic observatory off the St Pierre-et-Miquelon Archipelago (French islands close to Newfoundland, Canada) for 3 reasons: firstly, this geographic site is the hot-spot for all different cetacean species, especially the migratory species which come to eat from June to October, including humpback whales, fin whales and blue whales. Secondly, the weather conditions (cold, fog, rain) are not appropriate for boat trips and to collect recurrent visual observations, even at the sea surface. Thirdly, human activities are present, including fisheries and marine traffic to the US and Canada countries, and will potentially dramatically increase in the next decade (project to build a hub to optimically organize the marine trafic in the St Lawrence channel). Our acoustic observatory is used to provide the presence of vocal cetaceans off St Pierre and Miquelon islands and also to inform about the anthropogenic sounds.

Many underwater soundscapes has been studied but none about the SPM underwater soundscape has been published yet. This work is the first analysis of the SPM underwater soundscape. The aim of this study was to (i) describe the soundscape of SPM over 4 different periods in 2010, in 2011, 2011 and 2012, and in 2015 (ii) investigate the interactions between wind speed, rainfall, the presence of ships and cetaceans by season and year.

II. Materials & Methods

A. Study Area

The archipelago of Saint-Pierre-and-Miquelon (SPM) (4647N 5610W) is located at the South of the Canadian island Newfoundland. The archipelago consists of eight islands, the two largest being Saint-Pierre and Miquelon. The last is composed of two peninsulas linked together by a sandy tombolo. The SPM archipelago is of great interest for cetacean monitoring as it is situated near several migrating corridors. Indeed, several cetacean species were visually surveyed until today. Table 1 shows the different observed species.

TABLE I: Visually recognized cetacean species (w. stands for whale and d. for dolphin).

Mysticete	odontocete
Minke w.	Sperm w.
Fin w.	Beluga w.
Blue w.	Pilot w.
Humpback w.	Killer w.
Sei w.	Porpoise
	Pygmy sperm w.
	White-Sided d.
	Striped d.
	Common d.
	Bottlenose d.

B. Acoustic data

The acoustic data was collected thanks to two Autonomous Underwater Recorder for Acoustic Listening, Model-2 (AURAL-M2, by Multi-Electronique Inc., Rimouski, QC, Canada), owned by the SPM Frag-îles association. These recorders were located at two sites of the archipelago: one was at the North, the other at the South as shown in Fig. 1.

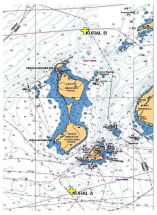


Fig. 1: Deployment locations.

The AURAL-A was moored at a mean depth of 60.5 m whereas the AURAL-B was anchored at a mean depth of 59 m. These two recorders were fitted with HTI-96-MIN hydrophone (High Tech Inc., Long Beach, MS, USA) with a sensitivity response of -165 dB re V/uPa and set with a 22 dB gain. The acoustic surveys were carried out during two different seasons to see if there were differences in the SPM soundscape according to the seasonality. In 2010, the devices were set to record for 45 mins continuously followed by a pause of 15 mins (75 % of duty cycle, sampling rate of 32,768 Hz at 16 bit) and measurements took place in late summer and fall. Then, in 2011, the duty cycle was modified and set to 50 % and the equipment was deployed in late spring and summer. The other configuration settings have been left unchanged. A summary is given in Table 2.

TABLE II: Summary of the characteristics of recordings performed in each site for each year.

Recorder	Mooring period (MM/dd, hh)	Duty Cycle (min on / min off)
2010		
AURAL-A	08/19, 5 p.m - 11/02, 11 p.m	45/15
AURAL-B	08/19, 5 p.m - 11/02, 11 p.m	45/15
2011		
AURAL-A	04/23, 1 p.m - 07/23, 12 p.m	30/30
AURAL-B	04/25, 8 p.m - 08/16, 6 p.m	30/30
2011-2012		
AURAL-A	10/15, 1 a.m - 04/30, 6 p.m	17.03 / 42.57
2015		
AURAL-A	05/17, 1 a.m - 11/02, 0 a.m	20.03 / 39.57
AURAL-B	05/03, 1 a.m - 11/01, 5 a.m	33.03 / 26.57

C. Environmental data

Wind speed data were included in this study to describe the soundscape and assess the possible interactions of natural phenomena to acoustic levels. Wind speed (red lines on LTSA) and daily summaries precipitation (mauve lines on LTSA) data were downloaded from the NOAA database (https://www.ndbc.noaa.gov/). Data were retrieved from the 44235 weather buoy and the SBM00071805 onshore station for wind speed and precipitation, respectively. The daily mean of wind speed is computed and matched to the Welch values. When auxiliary data is unavailable, they do not appear on plots.

D. Data analysis

Quantitative methods to characterize the spatial, temporal and spectral contents of the SPM underwater soundscape were used. The Power Spectral Density (PSD) was determined with the Welch method [14] with 1024-point Hamming window, 0% overlap, with 10-s temporal signal segments. As a consequence, time resolution is 10s and frequency resolution is 32 Hz. PSDs were then aggregated by days for LTSA and percentile representations. Full frequency band Sound Pressure Level (FFB SPL) was also computed at each site for each year. In order to remove the unidentified noise from the analysis, SPL in several bands were computed. The chosen frequency bands are 2-4kHz, 4-8kHz and 8-16kHz. PSD percentiles [12], [11], [19] were computed to describe statistically the noise levels at different frequency bins.

[Nguyen et al., 2019] workflow was used to compute Welch periodograms. Moreover, the statistical variability of underwater acoustic events over the frequency range of 016,384 Hz was illustrated as power spectral density (PSD) percentile graphs for each month of deployment if the month contains more than 5 days. As a consequence, November percentile plot across the campaigns are not shown.

III. Results

A. Unidentified strumming noise

It has to be noted that a low-frequency strumming sound was observed in the LTSA and needs to be taken into account in the analysis of acoustic levels. It is assumed that this noise is the result of shaking ropes / wires, moving chains / metal joints / batteries. This assumption is doubtful as the temporal and frequency representations mismatched with other observed mooring noises but strong winds seem to influence the apparition of such a noise. In this study, the frequency band altered by this noise ranges from 0 Hz to 3 kHz. It is described as the unidentified noise in the following. This noise was really strong in 2010 due to strong winds and perhaps currents. The other years of the study, batteries were placed in a different manner in order to reduce their movements in the recorder to attenuate the noise.

B. Overview

The marine soundscape in the SPM archipelago is rich in sound activity produce by various and concurrent acoustic sources. Whales and dolphins are the primary contributors to the biophony and the overall ambient noise level between 15 to 16,384 Hz for many months of the different years. The anthrophony in our recordings is represented only by vessels travelling from and to SPM. The geophony was described by wind-dependent noise and precipitation. The LTSA shown Fig. **??**, **??** are made of PSD of sea noise averaged over a day. It shows the key acoustic sources of SPM during the years 2010, 2011, 2012 and 2015. A biological activity is regularly seen at 6,000-16,384 Hz from August to November at each site. We expect this is diverse marine mammals. Even though the LTSA temporal and frequency resolutions are too rough to discriminate well contributions of each specie and characterize distinct vocalizations and calls, it gives an overview of the underwater soundscape. Nevertheless, the presence of several species of whales and dolphins is evident from the prominent narrow-band noise at 6,000-16,384 Hz, which is constituted by clicks, vocalizations and calls. The 14,000 Hz peak of rain [6], [3] is observed on several years especially on site A. LTSAs from site A show similar patterns for 2010 and 2015 from mid-summer to early fall in high frequencies.

The statistical variability of PSD levels was estimated and depicted as power spectral density percentile. These monthy percentile figures showed similar patterns according to the recording period demonstrating seasonality in sound emission in high frequencies. The same hump in August is retrieved in site A in 2010 and 2015 (cf Fig.). Major contributors to the underwater soundscape were annotated. Ships emit various frequencies depending on their propeller blade rate, engine tones and overtones ranging from 100 Hz to 7,000 Hz in our results (peaks at 100, 200, 350, 550, 700 Hz and a contribution to ripples from 3,000 Hz to 7,000 Hz). In 2010, peaks near 1,000 Hz are attributed to the unidentified noise. Humpback whales are supposed to be an active sound source for several months according to the season with the presence of a hump near 300 Hz in different percentiles. Wind-dependent noise is seen at mid-frequencies, 200 - 2,000 Hz for all months in the 1st and 5th percentiles. Between 2,000 and 16,384 Hz, the soundscape is dominated by odontocetes. Dolphins and whales generate clicks, whistles and strong tones in these frequencies. There is a hump in the 95th and 99th percentiles at 7,000 Hz every year at both sites. It never gets quieter than this level. It is thought to be rain. On site A, all months are affected by vessel noise as seen on the 95th and 99th hump at 100 Hz. This trend is not observed on site B which seem to be more sheltered from marine traffic.

C. Geophony

1) Wind speed contribution: On the 2010/09/20, for the two sites, saturation was observed. This could be due to the end of the Hurricane Igor (2010/09/08 - 2010/09/21). Resulting swell, rainfall and wind saturate the underwater recorders. It has to be noted that wind records were retrieved from a station further than the SPM archipelago of the hurricane's eye which implies stronger winds on the archipelago. Vice and versa, winds could be overestimated for other dates. It was not the

only storm in 2010 that could have affected SPM recorders during the deployment (e.g, Earl hurricane).

Fig. show the Pearson correlation coefficient for wind speed with different SPL bands and its statistical distribution for each year. Most of the deployment periods, winds reach greater speeds than 50 km/h which makes it a good candidate for a major contributor in the underwater soundscape as the moorings depth is about 60m. 2,000 - 4,000 Hz and 4,000 - 8,000Hz SPL bands are strongly correlated to wind speed. The Pearson coefficient was greater than 0.7 for each year at both sites except in 2011 it reached less than 0.65 for site A. Recordings were visually checked to understand why the correlation shrunked in 2011 on site A. As a result, more ships were encountered in the audio files. Ship noise is present in almost all 2011 recordings of site A. It is more surprising that PSD levels do not reveal this tendency. On site B, wind has a major impact as seen on the humps from 200-2,000 Hz in the 50th, 95th, 99th percentiles.

2) Precipitation contribution: The SPM archipelago witnesses sparse rainfall that often occurred at the same time of strong winds. However, These two variables are not correlated according to the low Pearson correlation coefficient (< 0.4). Moreover, SPL bands is not strongly correlated to rainfall with a coefficient less than 0.4.

D. Anthrophony

Human pressure on the underwater soundscape was only the result of different type of vessels. No AIS data were retrieved for this study. Indeed, this data are expensive and they are not reliable because not all vessels are required to have AIS systems (Canada's Navigation Safety Regulations (SOR/2005 - 134)) Nonetheless, audio recordings were visually inspected to detect the presence of ships. Private boats, ferries and fishing ships were visually and aurally observed. The southern AURAL recorded much more ship noise than the northern. Two situations are met in presence of vessels: passing ones on small duration or the presence of a ship in several audio recorders. In the latter, boats can either navigate slowly or being on site for professional activities which requires reducing speed or alternates on/off motor engines. There is more maritime traffic from April to mid-August than from mid-August to October.

E. Biophony

All extracted sounds from the datasets were aurally and visually checked. No detector was used in this study. As a consequence, this section is a brief overview of species that can be encountered near the SPM archipelago with specific spectrograms (plotted in Audacity).

1) Blue whales: Blue whales generate stereotyped calls which are used to discriminate sub-specie groups in different geographic locations. For example, Antarctic Blue whale vocalization [30], [28] differs from those encountered near SPM.

2) Humpback whales: Humpback whales produce complex tonal and pulsed sounds between 20 Hz and $>$ 15 kHz [12]. Fig. ... shows a distant humpback song spanning frequencies from about 50 Hz to 2 kHz, with the more intense part at 100400 Hz. Some sound units of this are repeatedly emitted and organized in the same order in the reproduction phase. Here, this type of songs were encountered in the audio recordings. Humpback whales sounds are typical for social and foraging activities.

3) Sperm whales: Sperm whales only produce clicks with peak energy at about 5,000, 7,000 and 11,000 Hz. These are easily identifiable with their time duration and peak frequencies. Creaks were also recorded corresponding to foraging activity (Fig. ...). In our recordings, it seems by visually and aurally inspection that the northern site is more visited than the southern from May to October. It is usually identified as the hump near 7,000 Hz in percentile plots.

4) Killer whales: A large repertoire of sounds are emitted by killer whales such as whistles, clicks [17], [49]. It is not surprising to see such a specie in the northern hemisphere with other marine mammals [1]. Fig ... depict some killer whales' calls found in the datasets.

5) Dolphins: They are the primary contributors to the SPM soundscape in high frequencies with whistles and clicks. Groups were often seen in audio recordings when high energy was recognized on LTSA. Boats were also found at the same time.

IV. Discussion

In this paper, four years of SPM passive acoustic recordings were investigated for principal sound contributors and SPL were correlated to meteorological data (wind and rain), underwater marine mammal presence, and anthropogenic activity (maritime traffic). Spatial distribution of the recorders enable to compare the two sites at spatio-temporal scales.

A. Geophony

While no weather buoy was placed above our recorders, climate data were retrieved from a weather station at more than 70 km for the southern site and about 40 km from the northern. Correlations of wind speed with underwater noise were evident at 200 - 8,000 Hz. Daily summary precipitation was weakly correlated. The Ocean Atlantic is a really active cyclone zone in summer which implies strong winds and more precipitation at this season. Specifically, SPM is near the intersection of the Gulf Stream and the Labrador current and it is on the way of one channel of the Labrador current. When the two currents meet, storms can be generated. Underwater noise is an ideal climate archive and can be exploited as an economical and efficient proxy for monitoring seaward weather [9], [34].

Our recorders were located at shallower water (about 60m deep) which only provide local-scale climate patterns. However, at both SPM sites, Earl hurricane was spotted in 2010, for several days. Wind was the major abiotic source in the SPM underwater soundscape for both sites and clearly a major contributor in site B.

B. Biophony

Ocean currents bring nutrients essential for underwater ecosystem. Plantkons develop near the meeting point of the

cold and warm currents and are carried by the west channel of the Labrador current in the Gulf of St Lawrence which makes it an attractive place for marine mammals and fishes in mid-summer to early fall which is expressed by peaks and large humps in percentile plots. Outside this period, marine mammals are present intermittently as seen by the small humps with low amplitude on percentile plots.

Several mysticete and odontocete species were aurally an visually identified in audio recordings. SPM archipelago seems to be a a transit area for feeding before going to cold waters. Major contributors are delphinides (whistles and clicks from 8,000 Hz to 16,384 Hz) for both sites and sperm whales for the northern site (clicks at 7,000 Hz).

In order to have a better insight of the presence of mysticetes in our recordings, 1-Hz daily-mean spectrograms and percentiles were computed for the 10 Hz - 500 Hz to identify specific patterns (not shown in this study). However, no seasonality was spotted. Low frequencies were polluted by shipping noise and geophony. On percentile graphs, several humps corresponding to different mysticetes. These plots only revealed possible presence of mysticetes. Then, we aurally and visually checked our assumptions. For both sites, visitation patterns from 2,000 Hz to the Nyquist frequency are very similar from year to year at a given month.

In site A, odontocete activity peak was observed from mid-summer to early fall. PAM can be used alone or or as a complement to boat and airplane-based visual observations, strandings, whale-vessel collisions for biological applications such as marine mammal population abundance. This is a mean to perform continuous monitoring over long periods and at different spatial scales of biological activities if there is any.

C. Anthrophony

The SPM archipelago is a place of intensive human pressure on underwater ecosystems. The acoustic data analysis demonstrated that variations in underwater ambient noise at two sites near SPM are driven by near and distant seasonal vessel traffic. The issue of the interactions between ships and marine mammals is well-known near SPM [10] and may affect cetacean in various manners [20], [23], [11], [16].

Seasonality was observed across the different recording campaigns in ship noise. There is more ship traffic in spring to mid-summer. Ferries are not the most disturbing anthropogenic activity as their travels are weekly from SPM to Fortune port but it could match the fishing season dates.

By comparing 100 Hz humps of the two sites, site A is more exposed to vessel noise than site B. Indeed, the former is near a maritime lane whereas the latter is less prone to vessel traffic. Site B is a more sheltered area for marine mammals. If biophony looks similar from year to year for a specific month for both sites, vessel noise near 100 Hz seems to increase or to be steady over the years for a specific month.

Conclusion and Future work

Although computations of the daily-averaged long-term representation, SPL and percentiles is straightforward, they proved to be good descriptive tools to characterize long-term underwater soundscapes. This paper presents the first underwater soundscape characterization of the archipelago of SPM over several years. It also presents the acoustic characterization of underwater noise levels in this site. This acoustic environment is dominated by biological (marine mammals), geological (the Atlantic Ocean is an intensive cyclone zone in the summer) and anthropogenic (ships) sounds, which show seasonality. Mid-summer until early fall SPM underwater soundscape is more likely to be made up of extreme weather and marine mammal activity whereas spring until beginning of summer soundscape of anthropogenic activity. These results show that SPM archipelago is a hot-spot for several cetaceans during summer but human pressure is experienced by them at the same period. Future studies need to be undertaken to i) analyze the variability of underwater soundscapes depending on duty cycle and window size to compute FFT ii) detect acoustic events in these datasets to have a better insight of the visitation pattern in the SPM archipelago.

Acknowledgment

The authors acknowledge the Pôle de Calcul et de Données Marines (PCDM) for providing DATARMOR storage, data access, computational resources, visualization, web-services, consultation, support services (http://www.ifremer.fr/pcdm). PhD is granted by the Direction Générale des Armements.

References

[1] THOMAS A. JEFFERSON, PAM J. STACEY, and Robin Baird. A review of killer whale interactions with other marine mammals: predation to coexistence. *Mammal Review*, 21:151 – 180, 12 1991.

[2] Heidi Ahonen, Kathleen M. Stafford, Laura de Steur, Christian Lydersen, ystein Wiig, and Kit M. Kovacs. The underwater soundscape in western fram strait: Breeding ground of spitsbergen's endangered bowhead whales. *Marine Pollution Bulletin*, 123(1):97–112, October 2017.

[3] Tahani Alsarayreh and Len Zedel. Quantifying snowfall rates using underwater sound. *Atmosphere-Ocean*, 49(2):61–66, 2011.

[4] G. Bazile Kinda, Yvan Simard, Cdric Gervaise, Jrome I. Mars, and Louis Fortier. Under-ice ambient noise in eastern beaufort sea, canadian arctic, and its relation to environmental forcing. *The Journal of the Acoustical Society of America*, 134(1):77–87, 2013.

[5] Frdric Bertucci, E Parmentier, Cecile Berthe, Marc Besson, Anthony Hawkins, Thierry Aubin, and David Lecchini. Snapshot recordings provide a first description of the acoustic signatures of deeper habitats adjacent to coral reefs of moorea. *PeerJ*, 5:e4019, 11 2017.

[6] L BJRN. Underwater rain noise: Sources, spectra and interpretations. *http://dx.doi.org/10.1051/jp4:19945225*, 04, 05 1994.

[7] D Bohnenstiehl, R Lyon, Olivia Caretti, Shannon Ricci, and David B. Eggleston. Investigating the utility of ecoacoustic metrics in marine soundscapes. *Journal of Ecoacoustics*, 2:R1156L, 09 2018.

[8] DelWayne R. Bohnenstiehl, Ashlee Lillis, and David B. Eggleston. The curious acoustic behavior of estuarine snapping shrimp: Temporal patterns of snapping shrimp sound in sub-tidal oyster reef habitat. *PLOS ONE*, 11(1):e0143691, January 2016.

[9] D. Cazau, J. Bonnel, and M. Baumgartner. Wind speed estimation using acoustic underwater glider in a near-shore marine environment. *IEEE Transactions on Geoscience and Remote Sensing*, 57(4):2097–2106, April 2019.

[10] Clment Chion, Dominic Lagrois, Jrme Dupras, Samuel Turgeon, Ian H. McQuinn, Robert Michaud, Nadia Mnard, and Lael Parrott. Underwater acoustic impacts of shipping management measures: Results from a social-ecological model of boat and whale movements in the st. lawrence river estuary (canada). *Ecological Modelling*, 354:72 – 87, 2017.

[11] Alexander Gavrilov Shyam Madhusudhana Christine Erbe, Robert McCauley and Arti Verma. The underwater soundscape around australia. In *Proceedings of ACOUSTICS 2016*, 2016.

[12] Robert McCauley Alexander Gavrilov Iain Parnum Christine Erbe, Arti Verma. The marine soundscape of the perth canyon. *Progress in Oceanography*, 2015.

[13] Christopher Clark, William Ellison, Brandon Southall, L T. Hatch, Sofie Van Parijs, Adam Frankel, and Dimitri Ponirakis. Acoustic masking in marine ecosystems: Intuitions, analysis, and implication. *Marine Ecology Progress Series*, 395:201–222, 12 2009.

[14] Peter D. Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *Audio and Electroacoustics, IEEE Transactions on*, 15:70 – 73, 07 1967.

[15] Alice Eldridge, Michael Casey, Paola Moscoso, and Mika Peck. A new method for ecoacoustics? toward the extraction and evaluation of ecologically-meaningful soundscape components using sparse coding methods. *PeerJ*, 4:e2108, June 2016.

[16] James J. Finneran, Carolyn E. Schlundt, Randall Dear, Donald A. Carder, and Sam H. Ridgway. Temporary shift in masked hearing thresholds in odontocetes after exposure to single underwater impulses from a seismic watergun. *The Journal of the Acoustical Society of America*, 111(6):2929–2940, 2002.

[17] John Ford. Acoustic behaviour of resident killer whales (orcinus orca) off vancouver island, british columbia. *Canadian Journal of Zoology*, 67:727–745, 01 1989.

[18] Lauren Freeman and Simon Freeman. Rapidly obtained ecosystem indicators from coral reef soundscapes. *Marine Ecology Progress Series*, 561, December 2016.

[19] Ignacio Gendriz and L.R. Padovese. Underwater soundscape of marine protected areas in the south brazilian coast. *Marine pollution bulletin*, 105, 02 2016.

[20] C. Gomez, J.W. Lawson, A.J. Wright, A.D. Buren, D. Tollit, and V. Lesage. A systematic review on the behavioural responses of wild marine mammals to noise: the disparity between science and policy. *Canadian Journal of Zoology*, 94(12):801–819, 2016.

[21] Sydney A. Harris, Nick T. Shears, and Craig A. Radford. Ecoacoustic indices as proxies for biodiversity on temperate reefs. *Methods in Ecology and Evolution*, 7(6):713–724, 2016.

[22] M. B. Kaplan, Aran Mooney, Jim Partan, and A. R. Solow. Coral reef species assemblages are associated with ambient soundscapes. *Marine Ecology Progress Series*, 533, August 2015.

[23] Vronique Lesage, Cyrille Barrette, Michael C. S. Kingsley, and Becky Sjare. The effect of vessel noise on the vocal behavior of belugas in the st. lawrence river estuary, canada. *Marine Mammal Science*, 15(1):65–84, 1999.

[24] Ashlee Lillis, Francesco Caruso, T. Aran Mooney, Joel Llopiz, DelWayne Bohnenstiehl, and David B. Eggleston. Drifting hydrophones as an ecologically meaningful approach to underwater soundscape measurement in coastal benthic habitats. *Journal of Ecoacoustics*, 2:STBDH1, 2018.

[25] Tzu-Hao Lin, Yu Tsao, Yu-Huang Wang, Han-Wei Yen, and Sheng-Shan Lu. Computing biodiversity change via a soundscape monitoring network. In *PNC 2017 Annual Conference and Joint Meetings*, November 2017.

[26] Sarah A. Marley, Chandra P. Salgado Kent, Christine Erbe, and Iain M. Parnum. Effects of vessel traffic and underwater noise on the movement, behaviour and vocalisations of bottlenose dolphins in an urbanised estuary. *Scientific Reports*, 7(1):13437, October 2017.

[27] Delphine Mathias, Cdric Gervaise, and Lucia Di Iorio. Wind dependence of ambient noise in a biologically rich coastal area. *The Journal of the Acoustical Society of America*, 139(2):839–850, February 2016.

[28] Mark Mcdonald, John A Hildebrand, and Sarah Mesnick. Biogeographic characterisation of blue whale song worldwide: Using song to identify populations. *Journal of Cetacean Research and Management*, 8, 01 2006.

[29] Martin B. MacDonnell J. McPherson, C. and C. Whitt. Examining the value of the acoustic variability index in the characterisation of australian marine soundscapes. 2016.

[30] David K. Mellinger and Christopher W. Clark. Blue whale (balaenoptera musculus) sounds from the north atlantic. *The Journal of the Acoustical Society of America*, 114(2):1108–1119, 2003.

[31] Jennifer L. Miksis-Olds, Bruce Martin, and Peter L. Tyack. Exploring the ocean through soundscapes. *Acoustics Today*, 14(1):26–34, 2018.

[32] CA Radford & AG Jeffs MK Pine. Eavesdropping on the kaipara harbour: characterising underwater soundscapes within a seagrass bed and a subtidal mudflat. *Journal of Marine and Freshwater Research*, 2015.

[33] Tasker M.L., Mats Amundin, Michel Andr, Hawkins T, Lang W, Merck T, Scholik-Schlomer A, Jonas Teilmann, Frank Thomsen, Stefanie Werner, and Manell Zakharia. Marine strategy framework directive - task group 11 report - underwater noise and other forms of energy,. Technical report, 04 2010.

[34] Sara Pensieri, Roberto Bozzano, Jeffrey Nystuen, Emmanouil Anagnostou, Marios Anagnostou, and Renzo Bechini. Research article underwater acoustic measurements to estimate wind and rainfall in the mediterranean sea. *Advances in Meteorology*, 15:1–18, 04 2015.

[35] Arief Ahmad Zailani Hatta Philippe Blondel. Acoustic soundscapes and biodiversity comparing metrics, seasons and depths with data from the neptune ocean observatory offshore british columbia. In *UACE2017 - 4th Underwater Acoustics Conference and Exhibition*, 2017.

[36] Nadia Pieretti, Marco Lo Martire, A Farina, and Roberto Danovaro. Marine soundscape as an additional biodiversity monitoring tool: A case study from the adriatic sea (mediterranean sea). *Ecological Indicators*, 83:13–20, 12 2017.

[37] Bryan C. Pijanowski, Almo Farina, Stuart H. Gage, Sarah L. Dumyahn, and Bernie L. Krause. What is soundscape ecology? an introduction and overview of an emerging new science. *Landscape Ecology*, 26(9):1213–1232, Nov 2011.

[38] R. L. Putland, R. Constantine, and C. A. Radford. Exploring spatial and temporal trends in the soundscape of an ecologically significant embayment. *Scientific Reports*, 7(1):5713, July 2017.

[39] Rosalyn L. Putland, Nathan D. Merchant, Adrian Farcas, and Craig A. Radford. Vessel noise cuts down communication space for vocalizing fish and marine mammals. *Glob Change Biol*, 24(4):1708–1721, April 2018.

[40] Miriam Romagosa, Irma Casco, Nathan D. Merchant, Marc O. Lammers, Eva Giacomello, Tiago A. Marques, and Mnica A. Silva. Underwater ambient noise in a baleen whale migratory habitat off the azores. *Frontiers in Marine Science*, 4:109, 2017.

[41] Sharon L. Nieukirk Haru Matsumoto Robert P. Dziak Jennifer L. Miksis-Olds Samara M. Haver, Holger Klinck. The not-so-silent world: Measuring arctic, equatorial, and antarctic soundscapes in the atlantic ocean. *DeepSea Research I*, 2017.

[42] Marc Holderied Andrew N. Radford Gael Lecellier Craig Radford David Lecchini Sophie L. Nedelec, Stephen D. Simpson. Soundscapes and living communities in coral reefs: temporal and spatial variation. *Mar Ecol Prog Ser*, 524:125135,, 2015.

[43] BRANDON L. SOUTHALL, ANN E. BOWLES, WILLIAM T. ELLISON, JAMES J. FINNERAN, ROGER L. GENTRY, CHARLES R. GREENE JR., DAVID KASTAK, DARLENE R. KETTEN, JAMES H. MILLER, PAUL E. NACHTIGALL, W. JOHN RICHARDSON, JEANETTE A. THOMAS, and PETER L. TYACK. Marine mammal noise-exposure criteria: Initial scientific recommendations. *Bioacoustics*, 17(1-3):273–275, 2008.

[44] E. Staaterman, M. B. Ogburn, A. H. Altieri, S. J. Brandl, R. Whippo, J. Seemann, M. Goodison, and J. E. Duffy. Bioacoustic measurements complement visual biodiversity surveys: preliminary evidence from four shallow marine habitats. *Mar Ecol Prog Ser*, 575:207–215, 2017.

[45] E Staaterman, Aaron Rice, David Mann, and Claire Paris. Soundscapes from a tropical eastern pacific reef and a caribbean sea reef. *Coral Reefs*, 32, 06 2013.

[46] Erica Staaterman, Claire Paris, Harry A. DeFerrari, David Mann, Aaron Rice, and Evan K. DAlessandro. Celestial patterns in marine soundscapes. *Marine Ecology Progress Series*, 508:17–32, 08 2014.

[47] Jrme Sueur, Sandrine Pavoine, Olivier Hamerlynck, and Stphanie Duvail. Rapid acoustic survey for biodiversity appraisal. *PLOS ONE*, 3(12):1–9, 12 2009.

[48] S. Viola, R. Grammauta, V. Sciacca, G. Bellia, L. Beranzoli, G. Buscaino, F. Caruso, F. Chierici, G. Cuttone, A. DAmico, V. De Luca, D. Embriaco, P. Favali, G. Giovanetti, G. Marinaro, S. Mazzola, F. Filiciotto, G. Pavan, C. Pellegrino, S. Pulvirenti, F. Simeone, F. Speziale, and G. Riccobene. Continuous monitoring of noise levels in the gulf of catania (ionian sea). study of correlation with ship traffic. *Marine Pollution Bulletin*, 121(1):97 – 103, 2017.

[49] Rebecca Wellard, Christine Erbe, Leila Fouda, and Michelle Blewitt. Vocalisations of killer whales (orcinus orca) in the bremer canyon, western australia. *PLOS ONE*, 10(9):1–26, 09 2015.

ANNEX

Appendix F

Cetacean distribution modeling study

CDM study

Multimodal deep learning for cetacean distribution modeling.

Case study of fin whales (*Balaenoptera physalus*) in the western Mediterranean Sea.

Cazau, D.[(1)], Nguyen Hong Duc, P.[(2)], Druon, J-N.[(3)], Matwins, S.[(4)], Fablet, R.[(5)]

(1) Lab-STICC, UMR CNRS 6585, ENSTA Bretagne, France
(2) Sorbonne Université, Institut Jean Le Rond d'Alembert, Paris, France
(3) European Commission - Joint Research Centre
(4) Faculty of Computer Science, Dalhousie University, Canada
(5) Lab-STICC, UMR CNRS 6585, IMT Atlantique, France

November 1, 2020

Contents

Abstract

Accurate predictions of Cetacean Distribution Modeling (CDM) are essential to their conservation but are limited by statistical challenges and a paucity of data, making far from complete our knowledge of mobile marine species distributions and densities, and underlining the need for more generalizable CDM models across large ocean basins. CDM approaches typically rely on inferred species-environment relationships to derive predictions. They often fail in capturing generic properties of cetacean habitats, which would be valid within a broader range of environmental conditions that characterize the surveyed regions.

This paper aims at investigating the usefulness of deep learning based strategies including multi-task and transfer learning regarding the extrapolation problem in CDM. We leveraged their capacity to perform end-to-end co-training of heterogeneous learning tasks, namely a stochastic presence-background classification task and a deterministic rule-based model in our case. This new approach has been experimented through the study case of fin whales in the western Mediterranean Sea, using concurrently a highly heterogeneous dataset of whale presence-only records and the index outputs of the Feeding Habitat Occurrence model. Using a new metric called TPSH (True Positive rate per unit of Surface Habitat), our preliminary quantitative results show that our multi-task learning model improves both the feeding habitat model by 3 % and data-driven models by more than 5 %. We also found that these results could be further improved by adopting more optimal thresholds as observed from ROC curves, e.g. the multi-task learning model could reach gains up to 10 % in the True Positive Rate while maintaining its habitat spatial spreading. These trends in results have been further supported by the use of two other independent datasets that forced models to generalize beyond their training dataset of species-environment relationships. Another finding of our study is that increasing the number of background samples (a.k.a pseudo-absences) so as to approximate the average spatial size of the feeding habitat model also proved to be an efficient yet simple modeling technique to reduce habitat spatial spreading while boosting performance accuracy of models (gains up to 5 % for data-driven models).

1 Introduction

1.1 Context

1.1.1 Cetacean Distribution Models (CDM)

Cetacean Distribution Models (CDM) have become a prominent technique in conservation biogeography and are increasingly used as prediction tools for environmental change forecasts (Franklin, 2010). An accurate knowledge of the spatial distribution of species is actually of crucial importance for many concrete scenarios including the ocean management, the preservation of rare and/or endangered species, the surveillance of alien invasive species, the measurement of human impact or climate change on species, etc. Especially, a current need has raised for larger datasets of CDM data at broader scales to achieve for example more comprehensive risk assessment from plastics debris abundance (Fossi and et al., 2017), and implement spatially explicit conservation measures for Ecologically and Biologically Significant Marine Areas (Dunn, 2014).

A common modeling approach in CDM relies on the niche theory, which links environmental processes with ecological ones to select a limited set of predictors and improve model interpretability. In other words, CDM relies on some behavioural assumptions that cetaceans do not move randomly, but that their presence is for example linked to the abundance of their food source in their habitats, which in turn depends on physical, biogeochemical and biological factors over distance and time. The goal of CDM is to infer the spatial distribution of a given cetacean species based on a function that takes a matrix of ocean variables (e.g. temperature, currents, salinity, and chlorophyll levels) of a given location as input and outputs an estimate of the abundance or a presence index of the species.

1.1.2 Fin whales in the Mediterranean sea

Our study will focus on the western Mediterranean sea. The Mediterranean Sea is unique among large sea basins because it constitutes a miniature ocean with contrasted physical, climatic and biological characteristics (Bethoux, 1999), and supports a highly diverse marine fauna including large mobile animals such as cetaceans (Coll, 2010; Notarbartolo di Sciara et al., 2016). The Mediterranean Sea biodiversity is undergoing profound alterations as high levels of anthropogenic pressures synergistically interact with the effects of climate change (Lejeusne and Pérez, 2010; Micheli, 2013). Despite these marked anthropogenic pressures, the collection of systematic data to assess marine animal abundances and responses to these stressors has been heterogeneous throughout the Mediterranean Sea, reflecting the uneven distribution of funding for population monitoring (Mannocci, 2018). Cetacean population monitoring represents no exception: line transect survey programs to estimate cetacean abundances have been implemented mostly by European countries, in the north-western and central Mediterranean (Laran, 2017; Panigada, 2017b; Mannocci, 2018). The wide range of some cetaceans across the Mediterranean Sea (e.g., fin whale and sperm

whale (Aïssi, 2008; Panigada, 2017a)), combined with their vulnerability to the multiple anthropogenic pressures (Panigada, 2006; David et al., 2011; Castellote et al., 2012; Cañadas and Vazquez, 2014), stress the need to develop predictive models to map their densities throughout the Mediterranean Sea.

One particular endangered species is the fin whale (*Balaenoptera physalus* L. 1758), which is the only commonly observed mysticete in the Mediterranean Sea. Its presence has been documented since ancient times (Notarbartolo di Sciara et al., 2003) but because of its pelagic distribution, this species is among the least known of all cetaceans in the Mediterranean Sea. The degree of protection of the fin whale is important. It has recently been declared as an endangered species by the IUCN and Mediterranean fin whales are of high priority for the ACCOBAMS to better understand their population structure in order to assess their conservation status. Based on the information available, there are fewer than 10.000 mature individuals and they are subject to ongoing threats that may be causing a decline, population trends are still unknown (Reeves R. et al., 2006). Also, fin whales are among the fastest cetaceans in the world, that can sustain speeds between 37 km/h (23 mph) and 41 km/h (25 mph) and bursts up to 46 km/h (29 mph)[1], and can cover > 150 km daily, making its behaviour sensitive to meso-scale oceanographic features. Individuals aggregate during the summer in the feeding grounds of the North-Western Mediterranean Sea (Druon et al., 2012; Panigada, 2017a).

1.1.3 Current challenges of CDM

Gap analyses of mobile marine species observation data have led to identification of geographic areas and seasons with important knowledge gaps (Mannocci, 2018). For example, a global gap analysis of line transect surveys used to derive abundance estimates and habitat-based density models of cetaceans has revealed large geographic gaps in the Southern Hemisphere and seasonal gaps in non-summer months(Kaschner and Harris, 2012). Especially, Mediterranean waters characterised by comparatively warmer temperatures, lower productivity and higher eddy activity were poorly surveyed for cetaceans. This raised the prospect that cetacean density models fitted to environmental covariates would have to be extrapolated in order to provide predictions for the entire Mediterranean Sea.

This problem is all the more critical as a fundamental assumption in CDM calibration is the unbiased sampling of available conditions in the environmental space (Pearce and Boyce, 2006), whereas opportunistic data, not collected following strict random or systematic sampling methods (Yackulic, 2012), are generally used. Presence-only CDMs are particularly vulnerable to sampling bias, when compared to presence-absence CDMs, since background points are generally randomly drawn across the area of interest instead of representing the environmental conditions of the sampled area (Phillips, 2009; Yackulic, 2012). The resulting predictions from these naïve SDMs reflect the joint distributions

[1] From `https://animaldiversity.org/site/accounts/information/Balaenoptera_physalus.html`

of species probability of presence and sampling effort (Elith, 2011; Guillera-Arroita, 2015). Note however that absence for fin whales in particular can include substantial missed presence.

Hence, extrapolation in environmental space can lead to highly uncertain predictions because species-environment relationships are unknown in unsampled and/or extreme environments, i.e. unreliable extrapolations (Elith and Leathwick, 2009; Elith, 2010). Furthermore, the selected model strongly influences predictions, particularly when predicting outside the sampled range of the covariate data (Pearson, 2006; Mainali, 2015). In definitive, current CDM methods generally fail in capturing generic properties of cetacean habitats (Mannocci et al., 2015), which would be valid within a broader range of environmental conditions that characterize the surveyed regions.

1.2 Motivations, objectives & methods

The starting point of our work is that niche models are natural candidates to capture generic properties of habitats as they are directly built on fundamental ecological links between an animal behaviour and its environment (Panigada, 2017a; Druon, 2019). Having said that, the accuracy of such models in predicting whale occurrences might also benefit from more data-driven strategies to learn new relationships that may explain such occurrences regardless a specific animal behaviour.

In this current work, we address such a modeling hypothesis through the use of deep learning based strategies, including multi-task and transfer learning, which are well suited to perform end-to-end co-training of heterogeneous modeling approaches. To do so, we will employ two well-known CDM approaches, namely 1) the stochastic presence-background classification task, based on heterogeneous whale sightings and background sampling, and 2) the deterministic rule-based models (also referred to as niche models), based on *a priori* expert knowledge and a pre-calibration of rule parameters with whale sightings. Our new approach then aims to better fuse these two approaches. It will be experimented through the study case of Fin Whales in the Western Mediterranean Sea.

1.3 Related works

Within the CDM literature, our study will draw from two broad families of methods that are rarely processed together. A first one is the niche model, where observation data are processed as a "whole" to provide summary statistics over large spatio-temporal scales used to calibrate parameters of explicit "behavioral rule". For example, it is known that chlorophyll-a fronts are productive oceanic features that attract large predators like fin whales, a knowledge that is at the core of Druon et al. (2012)'s favourable feeding habitat of fin whale (more details will be provided at Sec. 4.1.1). But note that the favourable feeding habitat is centred on the productivity fronts and not strictly on the punctual environment of all observations. Fin whales are hypothesized to search for these fronts for

feeding, they are therefore supposed to be often nearby, searching for them, except when migrating and reproducing. A second family of CDM methods involve pure data-driven approaches, where different statistical algorithms (e.g. Derville (2018) benchmark GAM, SVM and MaxEnt methods) are first used to learn specific patterns in the environmental variables related to the presence of a whale or to its "supposed" absence. Then, such an algorithm can then predict a presence vs background probability at new temporal and spatial points. On the one hand, expert models will tend to be biased towards a too restricted whale-environment relationships. On the other hand, data-driven approaches may involve large sampling biases leading to complex extrapolation problems like already mentioned above. In some way, our work aims at proposing a new possible CDM framework that consists in better combining knowledge both from rule-based expert models and data-driven learning.

Conceptually, our work connects to the literature of multimodal machine learning (Baltrusaitis et al., 2017). These models are generally trained end-to-end from large amount of data with multiple modalities can be trained and aim to extract a joint representation, which is a more compact representation integrating all modalities. Doing so, they able to learn complex decision boundaries that other approaches struggle with. For example, the model can easily generate one modality from another modality via this representation if it can obtain the joint representation properly. Deep neural networks are very actively explored in multimodal fusion (Ngiam et al., 2011). They have been used to fuse information for audio-visual emotion classification (Wöllmer et al., 2010), gesture recognition (N. Neverova and Nebout., 2016), affect analysis (S. E. Kahou, 2016), and video description generation (Jin and Liang., 2016). While the modalities used, architectures, and optimization techniques might differ, the general idea of fusing information in joint hidden layer of a neural network remains the same. In the marine realm, our work also shares some ideas with Rao et al. (2017), who used multimodal deep learning for autonomous underwater vehicles (AUV) to build a semantic mapping (i.e., understanding) of unseen or unfamiliar benthic habitats.

With the recent revival popularity of DNNs (Hinton et al., 2006), a few studies have begun to explore the use of DNNs for Species Distribution Modeling, a very close discipline to CDM. One layered-NN's had already been tested in the context of species distribution modeling (Thuiller, 2003). More recently, Chen (2017) jointly modeled the distribution of multiple bird species while also simultaneously learning the shared habitat preferences among species. Their DMSE (Deep Multi-Species Embedding) framework considers two heterogeneous contextual information feature sets (environmental features and interspecies relationships). It uses a deep neural network to extract high-level feature from environmental covariates and it couples the environmental and species embeddings into a predictive multi-species distribution model. They show that their model better deals with the fusion of heterogeneous multi-scale features, as the environmental features used in the model describe habitat characteristics at a much coarser spatial resolution than that of the inter-species interactions, this model formulation can be seen as a multi-scale approach that shares in-

formation at coarse scales while simultaneously allowing fine-scale variabilities between species. Botella et al. (2018) applied DNNs on plant distribution modeling. They show how mutualizing model features for many species prevent deep NN to overfit and finally allow them to reach a better predictive performance than the MAXENT baseline. They also show that spatially extended environmental patterns contain relevant extra information compared to their punctual values, and that species generally have a highly autocorrelated distribution in space.

2 Observation data

2.1 Cetacean sightings data

2.1.1 Overview

As detailed in table 4, a total of 12 survey cetacean sightings, plus 2 datasets of e-tagging data, were included in the full dataset. Data were sourced both from the public domain (e.g., OBISSEAMAP) and from personal dataset obtained upon demand (e.g., ferry-Cotte2009 and etagging-Cotte2009). Sightings will be processed conjointly to e-tagging data regardless their distinct natures, making this dataset of presence-only data of fin whales highly heterogeneous. Having said that, both types of presence data are well balanced in our dataset. Also, note that several animals might have been spotted at the time and location of our sightings data, but we did not take into consideration this information in our analysis. The complete dataset compiled from these sources provided a total of $n = 1479$ samples.

2.1.2 Spatial and temporal distribution

Figure 1 shows the geographical locations of all presence data by the background density plot, at the exception of the data from etaggingCotte2009 and ACCOBAMS2019 campaigns, plotted respectively as orange and yellow crosses for the need of a generalization study (descried later in Sec. 4.2). We can observe that most presence data are mainly located in the North Western Mediterranean Sea. Figure 2 shows the temporal distributions of presence data over years and months, mainly concentrated in autumn and spring, and with two years containing a high number of observations in 2012 and 2015. Once again we superimpose on the FWO_all data distribution the two datasets etaggingCotte2009 and ACCOBAMS2019. In table 4 of annex A1, we further describe our presence data for each campaign.

2.1.3 Data cleaning

Some e-tagging data were filtered out based on animal speed, discarding too high speed values that may correspond to a migration behaviour or unrealistic

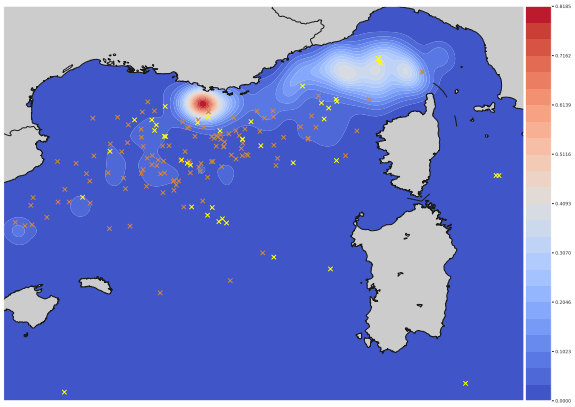
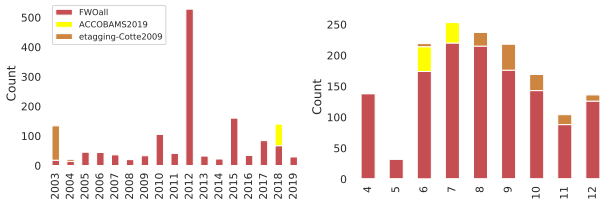


Figure 1: Spatial distribution of fin whale occurrence data represented as a density plot. Orange and yellow crosss represent the geolocations of sightings from our two independent test datasets, respectively FWO_Cotte2009 and FWO_ACCOBAMS2019.



(a) Year (b) Month

Figure 2: Temporal distributions of FWO over years and months.

values. Indeed, the use of e-tagging gives the opportunity to estimate approximately the animal mean speed based on the elapsed time between successive locations. An average speed above 8 km/h per day can be considered as a migratory behaviour from one location to another rather than a foraging behaviour. For example, Figure 3 represents the histogram of average daily whale speed from the etaggingCotte2009 dataset. The proportion of non-foraging e-tag observations is around 10 %. A more elaborated method presented in Panigada (2017a) and not used here is the area-restricted search behaviour which is characterised by higher turning angles and lower autocorrelation values in direction and speed to maximise searching effort in the most profitable areas.

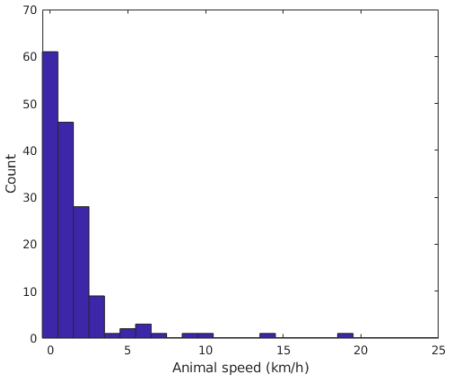


Figure 3: Histogram of animal speed (km/h) using the cotte2009 dataset. Here only 10 % of speed values exceed 8 km/h, above which a non-foraging behaviour can be hypothesized.

2.1.4 Background sampling (a.k.a "pseudo-absences")

As we do not have sampling effort from all campaigns, we used a background sampling method that do not need them, namely random sampling. It consists in generating a first set of naive background points (i.e. unbiased) taken at random within the whole study area and over all the months from the year 2013, except at the occurrence localities of the target species. We followed past studies (Barbet-Massin et al., 2012; Cerasoli et al., 2017) in taking a number of random background sampling points equal to the number of presence records, which have already been shown for example to provide a good accuracy in boosted regression tree models.

Environmental predictor		Variable name	Source
Bathymetry	Depth (m)	*Bathy*	General Bathymetric Chart of the Oceans (`http://www.gebco.net/`)
Biological	Chlorophyll-a concentration (mg m^{-3})	CHL	SeaWiFS and MODIS-Aqua sensors (`https://oceandata.sci.gsfc.nasa.gov/`)
	Chlorophyll-a gradient (mg $m^{-3} km^{-1}$)	$gradCHL$	Processed by JN Druon from CHL

Table 1: Candidate environmental predictors of cetacean habitat models.

For what concerns the number of background samples, it will be set to $n \times \alpha_{pa}$, where α_{pa} is a multiplicative factor that will be set either to 1 or 3. This latter value is used to unbalance the number of background samples comparatively to presence ones, as classically done in presence/absence model (Barbet-Massin et al., 2012), and was chosen following summary statistics on the percentage of pseudo-absences in FHO maps of size 49 $\times$ 49, which has a 25.6 % mean and 16.4 % standard deviation over the $n = 1479$ patches.

2.2 Environmental data

To describe the characteristics of pelagic habitats, we gathered and compiled geographic rasters containing the pixel values of 3 different environmental descriptors, as described in table 1, and which were chosen to mirror the information received by the FHO's model (Sec. 4.1.1). They have different spatial resolutions and nature of pixel values, but a common cover all over the western Mediterranean sea. Furthermore, biological variables have been sourced from the same publicly available archive, thus reducing the variability that may result from sourcing data from a large number of primary sources.

2.2.1 Bathymetry

The water depth was taken from the General Bathymetric Chart of the Oceans (GEBCO) of the British Oceanographic Data Centre, with a spatial resolution of 1 arc-minute grid (ca. 1.85 km at the latitude of interest; GEBCO[2]) and interpolated to the 4.6 km selected grid. Note that in contrast to other variables, bathymetric features (e.g., coastal areas, continental slopes, submarine canyons and seamounts) are all static.

2.2.2 Biological data

Several studies have shown that feeding fin whales and other rorqual whales are often located in areas of Chlorophyll-a fronts (Cotté and Guinet, 2009), where many zooplankton species are abundant. The concentration of small and large zooplankton in convergence areas, such as Chlorophyll-a fronts, is known to attract higher trophic level predators, leading to the assemblage of a complete pelagic food web (Olson et al. 1994). Chlorophyll-a fronts were also shown to be related to mesozooplankton biomass (Druon, 2019). Chlorophyll-a fronts

[2] `www.bodc.ac.uk/products/bodc_products/gebco/`

are thus seen as continuous productive features of organic matter efficiently assimilated by the food chain. It results that the vicinity of chlorophyll fronts together with medium Chlorophyll-a content (CHL in mgChl.m^{-3}) are assumed to be indicative of fin whale foraging niches in the Mediterranean Sea. The geo-projected product CHL was selected from MODIS-Aqua, mapped daily (at the scale of the processes involved, i.e. within 12 h) with a medium resolution of 4.6 km for MODIS. The gradient of Chlorophyll-a ($gradCHL$ in mgChl.m^{-3}.km^{-1}) is computed following (Druon, 2019, Step 1, Chlorophyll-a data, Material and Methods).

3 Proposed frameworks

In this section, we detail the frameworks used in our study, starting with with a description of input data before providing a complete model description and implementation details.

3.1 Input data

First, rasters of all environmental variables were aggregated to match the same 0.5 ° latitude × 0.5 ° longitude grid, corresponding to the 4.6 km resolution of MODIS, using built-in join operations in Python on panda dataframes. This resolution is the one used in whale habitat maps from Druon et al. (2012). Then, input data of our models take the generic form of a collection of image-like patches that are stacked in a design tensor $X \in \mathbb{R}^{n \times patchSize \times patchSize \times d}$. The k^{th} patch $X^k_{:,:,j}$ of descriptor j is a raster of size $(patchSize, patchSize)$ cropped[3] from the global raster of each environmental descriptors, where $patchSize$ is uneven to have a squared patch. If we denote r_{lon} and r_{lat} the spatial resolutions in longitude and latitude of global raster, the spatial extent of $X^k_{:,:,j}$ is $(r_{lon} \times patchSize, r_{lat} \times patchSize)$.

In our study, we have $r_{lat} = r_{lon} = 4.6$ km. All input tensors are set to have a $patchSize = 49$ pixels (i.e. ≈ 225 km), and gather one or several of the variables: $[Bathy, CHL, gradCHL, FHO]$, where FHO is the output of a model-based CDM that will be detailed in Sec. 4.1.1. In the following we list the different design tensors used in our analysis:

- X_{fwp} resulted from the sampling of $n = 1479$ patches with the variables $[Bathy, CHL, gradCHL]$ (d=3) centred at fin whale presence locations. We also defined X_{fhop} the tensor containing the FHO variable at fin whale presence locations;
- X_{pa} resulted from the sampling of $n \times \alpha_{pa}$ patches with the variables $[Bathy, CHL, gradCHL]$ (d=3) at background locations, as defined by the random sampling strategy described in Sec. 2.1.4;

[3] Using the function extract_patches_2d in package sklearn.

- X_{fho} resulted from the sampling of $40,000$ patches with the variables $[Bathy, CHL, gradCHL]$ (d=3) from the complete space area, excluding all patches containing FWO and/or at least one NaN value. Furthermore, FHO median values over each patch have been used to stratify the sampling into the five categories $0, [0; 0.3], [0.3; 0.4], [0.4; 0.6], [> 0.6]$ so as to get a representative dataset of the entire panel of FHO values. This design tensor will be used to perform supervised learning on FHO and evaluate sensitivity of results on the training dataset size ;

All resulting patches have been further post-processed with the following operations. Temporal interpolation was first performed, which consisted in applying a pixel-wise 5-days interpolation on NaN values of CHL and $gradCHL$ variables, under the hypothesis that these variables are persistent over this time duration. Bathymetric values corresponding to the land have been set to NaN, and patches containing more than half of NaN values were discarded.

3.2 Model overview

Our models aim to predict a probability presence p_k from an environmental patch k sampled at both presence and pseudo-absence locations of fin whales, and eventually from outputs of the expert model FHO (described in Sec. 4.1.1), that is $p_k = \mathcal{F}([X_{fwp}, X_{pa}]^k, X^k_{fhop})$, where the nonlinear function approximator $\mathcal{F}$ will be modelled as one or several neural network(s). The considered models use different learning strategies that can be broadly divided into:

1. **data-driven** (DD), where models are trained performing a binary classification from whale sightings and background samples;

2. **multi-task learning** (MTL) (a.k.a parallel transfer learning), where a multi-output model performs a joint learning of both fin whale observations and FHO processed respectively as binary classification and regression tasks;

3. **transfer learning** (TL) (a.k.a sequential transfer learning), which applies a two-step learning procedure: 1) starting by learning from FHO maps only (using MD models), and then 2) fine-tuning the model with fin whale observations.

Figure 4 provides a schematic overview of the multi-task learning and transfer learning schemes we use. Note that the difference between them lies in the way they transfer the knowledge: the shared information among all tasks is learned jointly in multi-task learning, while the knowledge in general transfer learning is transferred from one to the other. The proposed models, listed in table 2, are named following this template: *learnStrat_archi*, where *learnStrat* refers to one of the four learning strategies just described, and *archi* corresponds to the core architecture of the model we further detail in Sec. 3.3.

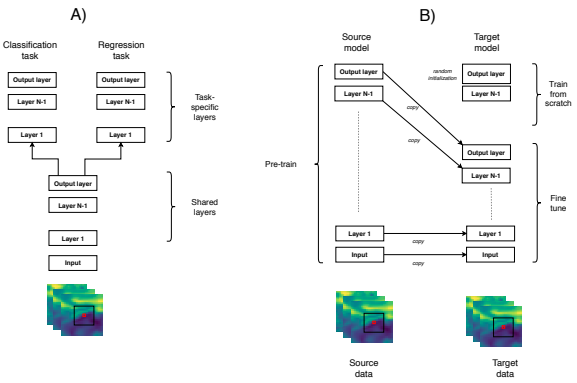


Figure 4: Schematic overview of A) multi-task learning and B) transfer learning approaches.

			Tasks	
Model	**Learning strategy**	**Architecture**	**Regression**	**Binary classification**
DD_MLP	Data-driven based on FWO and background samples	MultiLayer Perceptron	X	√
MTL_MLP//MLP	Multi-task learning performing jointly regression and classification tasks	Two MultiLayer Perceptron run in parallel	√	√
TL_MLP+MLP	Transfer learning from a regression task on FHO to a classification task on FWO and background samples with fine-tuning	Two MultiLayer Perceptron run sequentially with fine tuning	√	√

Table 2: Details of proposed models.

3.3 Implementation details

DD_MLP This model is a simple Multi Layer Perceptron (i.e. a feed-forward artificial neural network) that performs a binary classification task to predict a fin whale presence of pseudo-absence. It consists of a simple stack of 3 fully connected layers with relu activations, each having 16 hidden units, and followed by a fourth single-unit layer with a softmax activation and a binary cross-entropy as loss function defined as

$$L_{cla} = -\frac{1}{N}\sum_{k=1}^{K}(y_k log(p_k) + (1-y_k) log(1-p_k) \tag{1}$$

where $K = n \times (1 + \alpha_{pa})$ is the number of samples, y_k is the ground truth of whale presence (i.e. 0 or 1) in patch k, and p_k is the predicted probability of whale presence.

MTL_MLP//MLP This model performs multi-task learning by jointly running a regression task and a binary classification task. As illustrated in graph 1) of figure 4, the model follows a single-input multi-output architecture, using the first three layers of the DD_MLP model as the shared network branch, on which we added two parallel fourth layers that are trained jointly through a combination of their loss functions, as follows:

$$L_{mtl} = w_{reg} * L_{reg} + w_{cla} * L_{cla} \tag{2}$$

where L_{reg} is the function loss of the regression task, defined as a mean absolute error:

$$L_{reg} = -\frac{1}{k}\sum_{k=1}^{K}|p_k - y_k| \tag{3}$$

and w_{reg} and w_{cla} are two weights to balance the two task-specific losses, p_k is the predicted FHO value and y_k is its corresponding ground truth taken from X^k_{fho}. From preliminary numerical experiments, we set these weithgs as 0.6 and 0.4, respectively for w_{reg} and w_{cla}, which best satisfied both accuracy performance in our results and model optimization.

TL_MLP+MLP This last model is a case of (sequential) transfer learning where we first built a knowledge base pre-trained on a FHO regression task. To do so, the exact same MLP architecture as DD_MLP was used with a sigmoid (as FHO values are unitary) activation in its last layer and a mean absolute error loss function, and performing a 5-fold cross validation using X_{fho} tensors as inputs and the corresponding FHO values as targets. This knowledge base was then transferred to our binary classification task by first adding a fully connected classifier on top of it, and secondly by fine tuning an intermediate dense layer (namely the dense_layer_2, see model summary in Annex A.2) of MD_MLP, i.e. jointly training the new added classifier and this block while freezing the other layers.

Common details All models output a presence probability through their softmax activation in last layers. They have been developed using Keras functional API and Tensorflow backend [4]. Our network is trained end-to-end using RMSProp optimizer (Tieleman and Hinton, 2012). We decreased the learning rate accross iterations, typically from 1e-3 to 1e-6. Note that all models involving FWO data are relatively shallow networks with low capacity, in order to prevent overfitting the small amount of FWO data. All the experiments were run under workstations with a single GPU (Nvidia GTX 1080 and GTX 1080 Ti), resulting in computational times for model learning up to 10 minutes for the models with the highest capacity.

4 Experimental setup

4.1 Baseline methods

4.1.1 FHO (Feeding Habitat Occurrence) model

Druon's favourable feeding habitat of fin whale (simply referred to FHO in the following) is a daily indicator relating whale feeding behavior to environmental variables from satellite remote sensing. The feeding habitat is mostly related to the occurrence of chlorophyll-a fronts that are detected by satellite sensors of ocean colour. Chlorophyll-a fronts are mesoscale features such as ocean eddies or meandering currents, generating upwelling (divergence) and downwelling (convergence), and that persist long enough (i.e. weeks to months) to potentially sustain zooplankton production (Labat, 2009; Druon, 2019) prior to becoming hotspots for feeding at higher trophic levels. These productive oceanic features have then been shown to attract large predators like fin whales (D'Amico, 2003; di Sciara and Panigada, 2016; Panigada, 2017a), but also e.g. Atlantic bluefin tuna (Druon, 2016) and skipjack tuna (Druon et al., 2017). The FHO model provides a daily unitary index of the fin whale's preferred feeding habitat based on the distribution of horizontal chlorophyll-a gradients (green histogram) at the species' locations, represented in figure 5. 2318 fin whale sightings have been used to calibrate this model version (some are present in table 4), with the following numerical parameters

- $gradCHL_{min} = 0.000628$ (in mgChl.m^{-3}.km^{-1}), equal to the 25^{th} percentile;
- $gradCHL_{int} = 0.0041346$ (in mgChl.m^{-3}.km^{-1}), extracting from a linear fit of the cumulated distribution of $gradCHL$ values located at CS (see figure 5 for details);
- $CHL_{min} = 0.111$ (in mgChl.m^{-3}), equal to the 3^{th} percentile;
- $CHL_{max} = 0.4196$ (in mgChl.m^{-3}), equal to the 97^{th} percentile;

[4] `https://keras.io/,https://www.tensorflow.org/`

- $bathy_{min} = 90$ (in m), which is an expert value.

where the variables $gradCHL$ and CHL correspond to chlorophyll-a fronts and chlorophyll-a, respectively. In more details, a preferred range for CHL content (i.e. upper and lower values) is first set based on percentile values. The $gradCHL_{min}$ values that defined a daily habitat value of zero (lower value of the orange line segments) corresponds to the percentile 25^{th}, while the slope between the zero and one values was defined by the maximum slope of the cumulative distribution (green dashed line), as reflecting by the spectrum of CHL gradients of interest to fin whales. A minimum water depth of 90 m excluding the inner continental shelf. In contrast to Druon (2019), in this model no clustering has been used to estimate min max values of chlorophyll range values as the environment and whale behaviour was considered as simple enough.

For $gradCHL_{min} \leq gradCHL \leq gradCHL_{max}$ and $CHL_{min} \leq CHL \leq CHL_{max}$, we formulate FHO as

$$FHO = 1 + \frac{ln(gradCHL) - ln(gradCHL_{int})}{ln(gradCHL_{int}) - ln(gradCHL_{min})} \tag{4}$$

Also note that, Druon et al. (2017) arbitrarily set FHO to zero when it is inferior to 0.3, as a way to filter out sighted whales that would follow a behaviour different to foraging. Areas meeting the daily requirements of the habitat index can then be integrated over time to create seasonal suitability maps of feeding habitat. This time composite is expressed as a frequency of occurrence of daily habitat values, i.e. the sum of the daily habitat values (from 0 to 1) over the number of days for which the habitat suitability index was effectively estimated.

To be able to benchmark the FHO model on our task of whale presence/absence prediction, we made the following assumptions: 1) the FHO index is similar to a presence probability $P_{FHO}^{(p)}$ for patch p, 2) a presence is defined for habitat values higher than 0.5, that is

$$\begin{cases} P_{FHO}^{(p)} = 1 & \text{if } X_{fwFHO}(p, patchSize, patchSize, ind_FHO) \geq 0.5 \\ P_{FHO}^{(p)} = 0 & otherwise \end{cases} \tag{5}$$

where $patchSize$ is the considered spatial extent of the patch and $countpercent$ is a mathematical operator that return the total percentage of observations within T_{FHO} that respect the constraint.

4.1.2 GAM

We also implemented a GAM (Generalized Additive Model), as it has been widely used for whale occurrence prediction (e.g. Derville (2018)). We used the pyGAM implementation[5], which has been written mainly based on the R mgcv package[6] by Wood (2006). As we deal with a binary classification task, logisticGAM is used, featuring a Binomial error distribution and a logit link.

[5] `https://pygam.readthedocs.io/en/latest/`

[6] `https://cran.r-project.org/web/packages/mgcv/mgcv.pdf`

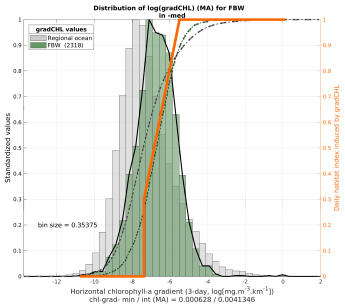


Figure 5: Definition of the daily favourable feeding habitat (orange line segments) for fin whale based on the distribution of horizontal chlorophyll-a gradients (green histogram) at the species' locations. These chlorophyll-a gradients represent small and large productive surface front features, as detected by MODIS-Aqua sensor. The cumulative frequencies and distributions of horizontal chlorophyll-a gradients in the regional Oceans are also plotted for comparison.

Like for our NN-based models, we did not perform hyperparameter tuning of these models as it is generally done in CDM literature (e.g. (Derville, 2018)), except for quick optimization on the kernel selection.

4.2 Training and test sets

We built four different datasets following the different train/test splitting as detailed in table 3. Globally, cross-validation has been systematically used to evaluate the stability of models, e.g. by observing no overfitting between the loss on the validation vs test set during cross-validation, and to report standard deviation along the median values. But note that it was not used to validate choices of hyper-parameters, as no in-depth optimization has been performed.

The dataset FWO_all gathers almost all of our cetacean occurrence data except those from Cotte2009 and ACCOBAMS2019, which have both been used as two independent datasets to test generalization capacity of our models. For these two datasets, Leave-One-Group-Out (LOGO) has been naturally employed, with the campaign names of Cotte2009 and ACCOBAMS2019 as group information to encode domain specific pre-defined cross-validation folds. We kept the same number of training samples as for the FWO_all dataset.

Back to figure 1, we can observe that most presence data are mainly located in North Western Mediterranean Sea, while geolocations from FWO_Cotte2009 and FWO_ACCOBAMS2019 data are located further west, at a few hundreds

of km from the hot spot of FWO_all within the Pelagos sanctuary. Temporally, figure 2 shows that data from the two datasets FWO_Cotte2009 and FWO_ACCOBAMS2019, although they have been collected mostly in the same season as FWO_all, their years of collection are at the extremities of FWO_all temporal range. Thus, these two datasets propose interesting cases of generalization both in space and time.

Note that the FHO model has not been re-trained in these different experiments, but that most of our whale sightings, except the two datasets FWO_Cotte2009 and FWO_ACCO2019, are shared with the training dataset for its development and calibration. Also, our FWO dataset has been mostly sampled in an area and time period favourable for fin whale feeding behavior, in rather good agreement with the FHO model assumptions.

For evaluation of our models, several metrics were used. As we use pseudo-absences, we cannot obtain an unbiased estimation of false positives from our binary classification results. Instead, we only use test patches sampled from fin whale presence locations, and report the True Positive rate per Unit of favourable Surface Habitat through the following TPSH metric

$$TPSH = \frac{1}{K}\sum_{k=1}^{K}(\mathbb{1}_\lambda(p_k) \times (1 - SSH_k)) \quad [\text{in } \%] \tag{6}$$

where $\mathbb{1}(p_k)^\lambda$ is the indicator function (i.e. $\mathbb{1}_\lambda$(x)=1 for $x \geq \lambda$ and 0 otherwise), p_k is the model output probability for patch k, K the number of test samples, and SSH_k is the Spreading of Surface Habitat of pixels in patch k reduced to the unitary interval, that is

$$SSH_k = \sum_{n=1}^{N_k} \frac{\mathbb{1}_\lambda(p_k(n))}{(N_k - Nnan_k)} \quad [\text{in } \%] \tag{7}$$

where N_k the total number of pixels per patch, i.e. $49^2 = 2401$, and N_{nan} the number of NaN values in the patch.

Basically, the metric $TPSH$ has positive contributions only from true positive estimation of the model, and increases inverse proportionally to its number of pseudo false positives over a 49×49 pixel patch. This metric thus includes an implicit evaluation of false positive rates, yet it does not consider an explicit count of absences correctly predicted or not as in the TSS metric used in presence/absence problem, and defined as the sum of sensitivity and specificity minus one, the sensitivity being the proportion of presences correctly predicted, and while the specificity is the proportion of absences correctly predicted (Barbet-Massin et al., 2012).

The threshold λ is set to 0.5 by default, but we also report in our study ROC curves by exploring the value range [0.35 ; 0.65] with a 0.05 step. From such curves, the euclidean distance D to the top-left corner of the ROC curve for each λ can be computed, where lower values will show better performance than larger ones.

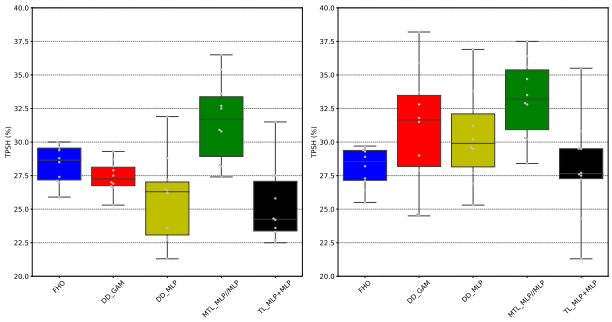
Datasets	Train samples	Test samples	Splitting procedure
FWO_all	1331	148	10-fold CV on FWO
FHO_TL	32000	8000	Stratified 10-fold CV
FWO_Cotte2009	1331	123	10-fold LOGO using Cotte2009 as test group
FWO_ACCOBAMS2019	1331	44	10-fold LOGO using ACCOBAMS2019 as test group

Table 3: Details of test / training sets for the different experimental datasets. CV stands for cross-validation.

5 Experiments

We first benchmarked our models on the task of presence-absence prediction of the whale sightings from the FWO_all dataset. Results are displayed in figure 6 for $\alpha_{pa} = 1$ (subfigure (a)) and $\alpha_{pa} = 3$ (subfigure (b)), where we used the TPSH metric with $\lambda = 0.5$. Starting with subfigure (a), and considering the margins of error, we can observe that our proposed model MTL_MLP//MLP performs the best in predicting fin whale presence, improving FHO of 3 % on average, and data-driven models (namely, DD_GAM and DD_MLP) of more than 5 %. From subfigure (b), increasing the number of background samples improves all models, with greater benefits for the two data-driven models, e.g. up to 5 % for DD_MLP, and to a lesser extent the MTL_MLP//MLP and TL_MLP+MLP models, e.g. with a gain of 1 % for MTL_MLP//MLP. This also globally increases dispersion of output models. This result is in line with CDM literature, as it is known that a random selection of pseudo-absences is recommended when high specificity is valued over high sensitivity (e.g. for reserve planning) (Barbet-Massin et al., 2012). Furthermore, when comparing our two different learning strategies, clearly the knowledge from FHO is better exploited in the multi-task learning MTL_MLP//MLP than in the sequential transfer learning TL_MLP+MLP models, where in this latter the classification fine-tuning seems to overwhelm the FHO knowledge base as this model does not even reach FHO performance.

To provide further insight into the trade-off between True Positive Rate and Surface Habitat Spreading inside the TPSH metric, we represented in figure 7 the ROC curves obtained by thresholding from 0.35 to 0.65 the probability output of each model. Naturally, we reach the same trends in results as with the TPSH metric, but we can now better visualize in subfigure (a) how data-driven models are quite limited in decreasing the spatial extent of their habitat estimation, demonstrating their inability in properly isolating locally-defined environmental features related to fin whales occurrences. From subfigure (b), benefits from increasing the α_{pa} parameter can be simply described as translating all ROC curves on the left, plus forming a plateau-shaped slope on certain models like MTL_MLP//MLP and DD_MLP for reasonably small SSH values ($\leq$ 35 %). This means that this mechanism of unbalancing presence-absence classes during training allows to reduce habitat spatial spreading while boosting performance accuracy of models(gains up to 5 % for data-driven models), which are promising properties towards better CDM models. We can also observe how this augmentation makes the crossing between the ROC curves of



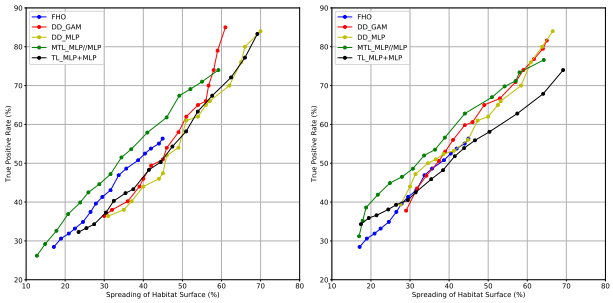
(a) $\alpha_{pa} = 1$ (b) $\alpha_{pa} = 3$

Figure 6: Boxplots of TPSH metric (in %) to assess model performance on a presence-absence classification task on the FWO_all dataset. Remember that coefficient α_{pa} is used to unbalance the number of background samples w.r.t to presence ones, as detailed in the background sampling strategy in Sec. 2.1.4.

MTL_MLP//MLP and data driven models occurring for smaller SSH values.

It is now interesting to visualize how the proposed models deviate from original FHO maps. To do so, we used for each model the training set from the best performing fold, i.e. with the highest TPSH metric, to predict maps of whale presence-absence over 49 × 49 pixel patches (i.e. ≈ 225 km × 225 km), represented in figure 10. From these figures, we can better describe the modeling behaviour of MTL_MLP//MLP, which basically tends to slightly spread around FHO patterns, thus capturing in finer details salient mesoscales structures than the data-driven models, while being also able to expand a bit its habitat surface over new areas by using the learnt local relationships between presence data and its surrounding environment. From subfigure d), it is also interesting to visualize how models extrapolate beyond FHO patterns so as to better reach visual sighting locations based on the learnt relationships from data. Although salient structures from these extrapolations are quite similar, once again the TL_MLP//MLP model performs the best in terms of narrowing its spatial prediction of habitats.

Our last experiment consisted in evaluating generalization capacity of our models using two independent campaigns of whale sightings, i.e. none of which have been seen during model training. Note that in the previous experiment, some presence data from a same campaign could appear both in training and test sets, limiting the evaluation of model generalization. Results are displayed in figure 9, where we used the TPSH metric with a 0.5 threshold probability and $\alpha_{pa} = 3$, respectively on subfigures (a) and (b) for the datasets etagging-

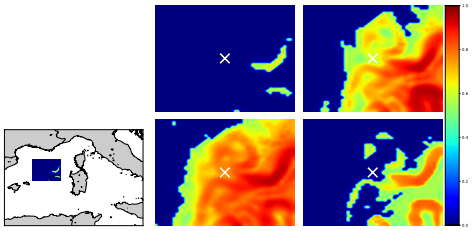


(a) $\alpha_{pa} = 1$ (b) $\alpha_{pa} = 3$

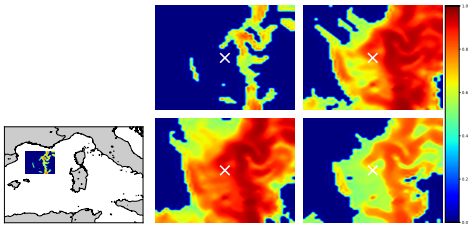
Figure 7: ROC curves to assess model performance on a presence-absence classification task on the FWO_all dataset. Note that the most interesting cases in terms of use are when the habitat surface is below about 35 % as corresponding to relatively high discriminant habitat.

Cotte2009 and ACCOBAMS2019. Performance results on the first dataset etaggingCotte2009 are quite similar to those obtained for the FWO_all dataset (subgifure (b) in figure 9), and degrade significantly on the second ACCOBAMS2019 dataset, showing also higher dispersion values on all models. Explanations of such a result can be initiated with the respective multivariate density distributions of these two datasets in comparison to FWO_all (see figure 10 of annex A2), which reveal stronger similarities between FWO_all and etaggingCotte2009 than with ACCOBAMS2019, especially through the $gradientCHL$ variable. Note that this was expected considering the time periods of both campaigns, where a higher primary productivity is expected in early winter (etaggingCotte2009 time period) than in spring and summer (ACCOBAMS2019 time period) which are more transitional seasons.

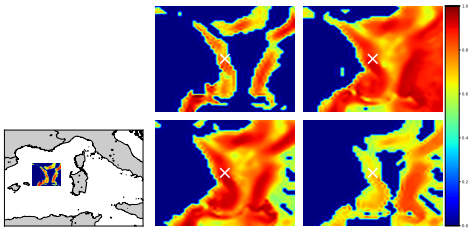
When comparing model results in more details, we can also remark that FHO and MTL_MLP//MLP models exhibit the best performance on the TPSH metric, with performance gaps up to 5 % in comparison to data-driven averaged performance, confirming our initial hypothesis that expert information on whale behaviour represent a good knowledge base for model generalization, which can be further improved by a concurrent learning of more local species-environment relationships from *in-situ* presence data. Eventually, regarding this generalization question, note that we privileged neural network architectures with low capacity models so as to better emphasize gains from the FHO knowledge in MTL_MLP//MLP and TL_MLP+MLP models, and likely improve model generalization.



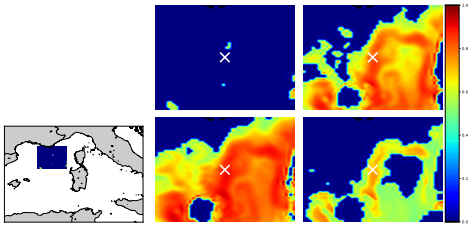
(a) On 28th March, 2015



(b) On 10th September, 2012

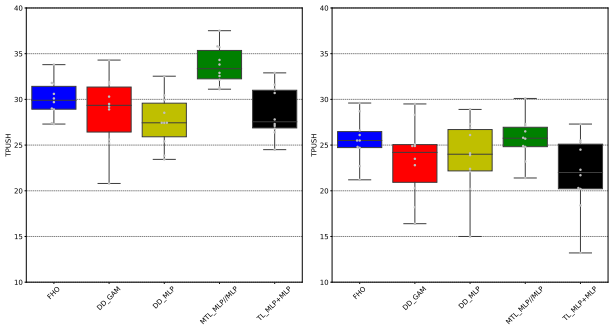


(c) On 23th June, 2010



(d) On 23th August, 2003

Figure 8: Predicted presence probability maps of size 49 × 49 pixel (i.e. ≈ 225 km × 225 km) for the different models: A) FHO, B) GAM, C) TL_MLP+MLP and D) TL_MLP//MLP. The white cross points to the location of a whale presence observation.



(a) Dataset etaggingCotte2009 (b) Dataset Accobams2019

Figure 9: Percentage of missed events on a presence-absence classification task on the FWO_all dataset using the TPSH evaluation metric with a 0.5 threshold probability and a random background sampling process with $\alpha_{pa} = 3$.

6 Conclusion and perspectives

In this study we have proposed a deep learning based strategy, i.e. a neural network model plus a learning scheme, able to perform concurrently pixel-based classification and regression tasks over both whale sightings and pre-computed expert maps. Such an approach has revealed a promising capacity in enhancing models w.r.t the extrapolation problem, and allows to obtain more refined and delimited likely occurrence areas of whales in contrary to data-driven models, which is a crucial requirement for operational management applications dedicated to marine mammal protection and conservation.

Future directions of our work will investigate more complex spatio-temporal CDM modeling. Indeed, although this is not the case of the feeding habitat model used in our study, current environmental covariables in CDM are mostly restricted to localized, single-pixel values alone (or average at fixed spatio-temporal resolutions), preventing them to capture larger contextual information that may be of importance for whale habitat modelling. Indeed, the precise time and spatial scales of raster data relevant to whale habitat modeling is still an open question. The integration of contextual information with multi-scale features can be done conveniently in some deep learning architectures, e.g. by feeding convolutional neural networks with high-resolution spatial patches of raster data around each spotted whale. In particular, end-to-end training of networks should be able to fuse multi-scale features from both large-scale contextual information and fine-grained details, and discover themselves the most

optimal space scales for habitat modelling, which are difficult to set *a priori*.

Acknowledgements

We would like to thank our main sponsors in this work: CominLabs[7] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[8] and ISblue[9]. We are also very grateful for all providers of cetacean visual sightings and e-tagging data, namely: Cotté, C. and Guinet, C., the associations Souffleurs d'Ecume and GECEM, Pelagis Observatoire and Thetys institutes.

Annex

A1. Details on our dataset of fin whale presence-only data

Table 4 provides details on our dataset of fin whale presence-only data. Also note that we have initiated administrative procedures so this dataset can be made publicly available.

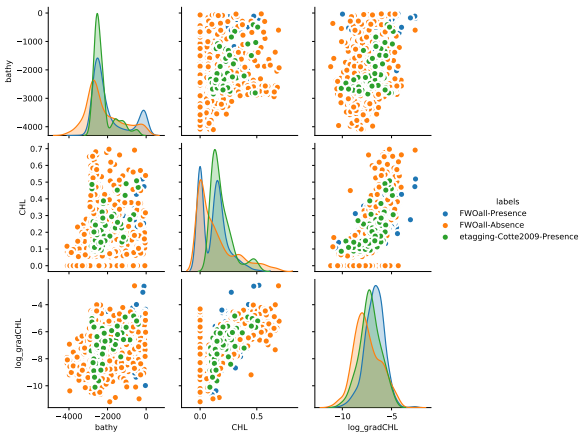
Campaign name	Sample number	Longitude (min-max)		Latitude (min-max)		Dates (min-max)	
ACCOBAMS2019	44	3.56	10.50	38.24	43.57	06/15/2018	07/26/2018
Atlantide	19	5.55	5.98	42.70	42.98	06/26/2018	08/19/2018
Espace Mer	18	6.13	8.16	42.67	43.03	06/24/2018	09/21/2018
Fossi2017_table1	3	8.02	10.07	41.35	43.55	09/11/2014	09/14/2014
GECEM	142	5.39	10.88	42.26	43.03	05/06/2018	10/30/2016
OBISSEAMAP	86	4.27	10.90	40.77	44.41	05/07/2007	12/27/2011
OceanCare_20012014	22	5.93	7.53	42.76	43.53	06/12/2009	09/09/2011
PelagisObservatoire_20042009	6	4.39	5.19	42.89	43.12	08/08/2008	09/03/2008
ThetysAerial_20092011	81	5.91	11.90	40.58	44.01	06/17/2010	08/02/2009
ThetysShipboard_19862012	232	6.66	15.74	38.12	44.10	05/19/2011	09/25/2007
WW_SouffleursdEcume_2012_2017	93	5.30	8.87	42.13	43.32	06/09/2017	10/29/2016
etagging-Cotte2009	123	1.40	9.28	39.40	43.36	04/19/2004	12/18/2003
etagging-Panigada2017	569	2.59	11.97	38.02	43.96	04/09/2015	12/31/2012
ferry-Cotte2009	41	3.88	7.10	38.10	42.90	05/01/2007	12/12/2006

Table 4: Details on our dataset of fin whale presence-only data.

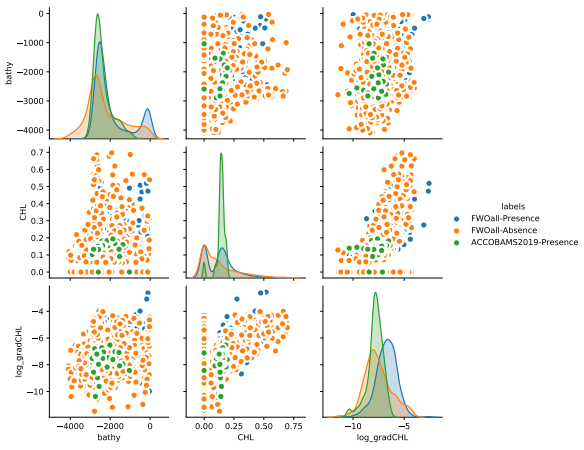
A2. Multivariate density distributions of fin whale presence-only data

Figure 10 represents multivariate density distributions and scattering plots of FWO_all training presence and background data, as well as both independent test datasets FWO_Cotte2009 and FWO_ACCOBAMS2019.

[7] `https://www.cominlabs.u-bretagneloire.fr/`
[8] `https://www.afbiodiversite.fr/`
[9] `https://www.isblue.fr/about-us/`



(a) etagging-Cotte2009



(b) ACCOBAMS2019

Figure 10: Multivariate density distributions and scattering plots of training presence and background data, as well as of both independent test datasets.

References

Aïssi, M.e.a. (**2008**). "Large-scale seasonal distribution of fin whales (balaenoptera physalus) in the central mediterranean sea." J. Mar. Biol. Assoc. U. K.

Baltrusaitis, T., Ahuja, C., and Morency, L.P. (**2017**). "Multimodal machine learning: A survey and taxonomy." arXiv.

Barbet-Massin, M., Jiguet, F., Albert, C.H., and Thuiller, W. (**2012**). "Selecting pseudo-absences for species distribution models: how, where and how many?" Methods in Ecology and Evolution.

Bethoux, J.P.e.a. (**1999**). "The mediterranean sea: a miniature ocean for climatic and environmental studies and a key for the climatic functioning of the north atlantic." Prog. Oceanogr., **44**, 131– 146.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. (**2018**). "A deep learning approach to species distribution modelling." Multimedia Tools and Applications for Environmental, Biodiversity Informatics, Springer, pp.169-199.

Cañadas, A. and Vazquez, J.A. (**2014**). "Conserving cuvier's beaked whales in the alboran sea (sw mediterranean): Identification of high density areas to be avoided by intense man-made sound." Biol. Conserv. 178, 155–162.

Castellote, M., C.C.W..L.M.O.A., behavioural changes by fin whales (Balaenoptera physalus) in response to shipping, and airgun noise. Biol. Conserv. 147, (**2012**).

Cerasoli, F., Iannella, M., D'Alessandro, P., and Biondi, M. (**2017**). "Comparing pseudo-absences generation techniques in boosted regression trees models for conservation purposes: A case study on amphibians in a protected area." PLoS ONE.

Chen, D.e.a. (**2017**). "Deep multi-species embedding." arxiv.

Coll, M.e.a. (**2010**). "The biodiversity of the mediterranean sea: estimates, patterns, and threats." PloS One.

Cotté, C. and Guinet, C. (**2009**). "Scale-dependent habitat use by a large free-ranging predator, the mediterranean fin whale." Deep-Sea Research I 56 (2009) 801–811.

David, L., Alleaume, S., and Guinet, C. (**2011**). "High risk areas of collision between fin whales and ferries in the north-western mediterranean sea." J. Mar. Anim. Their Environ. 4, 17–28.

Derville, S.e.a. (**2018**). "Finding the right fit: Comparative cetacean distribution models using multiple data sources and statistical approaches." Diversity and Distributions. 2018;24:1657–1673.

di Sciara, G. N., C.M.D.J.N. and Panigada, S. (**2016**). "Chapter three-fin whales, balaenoptera physalus: At home in a changing mediterranean sea?" Adv. Mar. Biol. 75, 75–101 (2016).

Druon, J.N. (**2019**). "Satellite-based indicator of zooplankton distribution for global monitoring." Scientifc Reports Nature.

Druon, J.N., Chassot, E., Murua, H., and Lopez, J. (**2017**). "Skipjack tuna availability for purse seine fisheries is driven by suitable feeding habitat dynamics in the atlantic and indian oceans." Front. Mar. Sci. 4:315. doi: 10.3389/fmars.2017.00315.

Druon, J.N., Panigada, S., David, L., Gannier, A., Mayol, P., Arcangeli, A., Cañadas, A., Laran, S., Di Meglio, N., and Gauffier, P. (**2012**). "Potential feeding habitat of fin whales in the western mediterranean sea: an environmental niche model." Mar Ecol Prog Ser, **464**, 289–306.

Druon, J.N.e.a. (**2016**). "Habitat suitability of the atlantic bluefin tuna by size class: An ecological niche approach." Prog. Oceanogr. 142, 30–46 (2016).

Dunn, D.C. (**2014**). "The convention on biological diversity's ecologically or biologically significant areas: origins, development, and current status." Mar. Policy, 49 (2014), pp. 137-145.

D'Amico, A., B.A.Z.P.C.S.N.E.P.N.e.a. (**2003**). "Qualitative correlation of marine mammals with physical and biological parameters in the ligurian sea. ieee j. ocean. eng. 28, 29–43. doi: 10.1109/joe.2002.808206."

Elith, J. and Leathwick, J.R. (**2009**). "Species distribution models: ecological explanation and prediction across space and time." Annu. Rev. Ecol. Evol. Syst. 40, 677–697.

Elith, J.e.a. (**2011**). "A statistical explanation of maxent for ecologists. –." Divers. Distrib. 17: 43–57.

Elith, J., K.M..P.S. (**2010**). "The art of modelling range-shifting species." Methods Ecol. Evol., 1, 330–342.

Fossi, M.C. and et al. (**2017**). "Plastic debris occurrence, convergence areas and fin whales feeding ground in the mediterranean marine protected area pelagos sanctuary: A modeling approach." Front Mar. Sci., **4:167**.

Franklin, J. (**2010**). "Moving beyond static species distribution models in support of conservation biogeography." Diversity and Distributions, **16**, 321–330.

Guillera-Arroita, G.e.a. (**2015**). "Is my species distribution model fit for purpose? matching data and models to applications." Global Ecol. Biogeogr. 24: 276–292.

Hinton, G.E., Osindero, S., and Teh, Y. (**2006**). "A fast learning algorithm for deep belief nets." Neural Computation, **18**, 1527–1554.

Jin, Q. and Liang., J. (**2016**). "Video description generation using audio and visual cues." In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pages 239–242. ACM.*

Kaschner, K., Q.N.J.J.R.W.R. and Harris, C.M. (**2012**). "Global coverage of cetacean line-transect surveys: status quo, data gaps and future challenges." . PloS One 7, e44075 (2012).

Labat, J.P.e.a. (**2009**). "Mesoscale distribution of zooplankton biomass in the northeast atlantic ocean determined with an optical plankton counter: relationships with environmental structures." Deep Sea Res. Part Oceanogr. Res. Pap. 56, 1742–1756 (2009).

Laran, S.e.a. (**2017**). "Seasonal distribution and abundance of cetaceans within french waters- part i: The north-western mediterranean, including the pelagos sanctuary. deep sea res. part ii 141, 20–30 (2017)."

Lejeusne, C., C.P.P.M.C.B.C.F. and Pérez, T. (**2010**). "Climate change effects on a miniature ocean: the highly diverse, highly impacted mediterranean sea." . Trends Ecol. Evol. 25, 250–260.

Mainali, K.P.e.a. (**2015**). "Projecting future expansion of invasive species: Comparing and improving methodologies for species distribution modeling." Glob. Change Biol. 21, 4464–4480.

Mannocci, L., Monestiez, P., Sptiz, J., and Ridoux, V. (**2015**). "Extrapolating cetacean densities beyond surveyed regions: habitat-based predictions in the circumtropical belt." Journal of Biogeography, **42**, 1267–1280.

Mannocci, L.e.a. (**2018**). "Assessing cetacean surveys throughout the mediterranean sea: a gap analysis in environmental space." Scientifc Reports, **8:3126**.

Micheli, F.e.a. (**2013**). "Cumulative human impacts on mediterranean and black sea marine ecosystems: Assessing current pressures and opportunities. plos one 8, e79889 (2013)."

N. Neverova, C. Wolf, G.T. and Nebout., F. (**2016**). "Moddrop: adaptive multimodal gesture recognition." In *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8):1692– 1706, 2016.*

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. (**2011**). "Multimodal deep learning." In *Proc. 28th Int. Conf.Mach. Learn., pp. 689–696.*

Notarbartolo di Sciara, G., Castellote, M., Druon, J.N., and Panigada, S. (**2016**). "Fin whales, balaenoptera physalus: At home in a changing mediterranean sea?" Adv Mar Biol. 2016;75:75-101. doi: 10.1016/bs.amb.2016.08.002. Epub 2016 Sep 24.

Notarbartolo di Sciara, G., Zanardelli, M., Jahoda, M., and Panigada, S.and Airold, S. (**2003**). "The fin whale balaenoptera physalus (l. 1758) in the mediterranean sea." Mammal Review 33:105-150.

Panigada, S.e.a. (**2006**). "Mediterranean fin whales at risk from fatal ship strikes. mar. pollut. bull. 52, 1287–1298 (2006)."

Panigada, S.e.a. (**2017**a). "Satellite tagging of mediterranean fin whales: working towards the identification of critical habitats and the focussing of mitigation measures." Scientifc Reports, **7: 3365**.

Panigada, S.e.a. (**2017**b). "Supplementary information : Satellite tagging of mediterranean fin whales: working towards the identification of critical habitats and the focussing of mitigation measures." Scientifc Reports, **7: 3365**.

Pearce, J.L. and Boyce, M.S. (**2006**). "Modelling distribution and abundance with presence-only data. –." J. Appl. Ecol. 43: 405–412.

Pearson, R.G.e.a. (**2006**). "Model-based uncertainty in species range prediction." J. Biogeogr. 33, 1704–1711 (2006).

Phillips, S.J.e.a. (**2009**). "Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data." Ecological Applications, 2009, vol. 19, no. 1, p. 181-197.

Rao, D., De Deuge, M., Nourani-Vatani, N., Williams, S.B., and O, P. (**2017**). "Multimodal learning and inference from visual and remotely sensed data." The International Journal of Robotics Research,Vol. 36(1) 24–43.

Reeves R., N.d.S.G.e..T.s., distribution of cetaceans in the Black Sea, and Mediterranean Sea. IUCN Centre for Mediterranean Cooperation, Malaga, S..p. (**2006**).

S. E. Kahou, X. Bouthillier, P.L.C.G.V.M.K.K.S.J.P.F.m.Y.D.N.B.L.e.a. (**2016**). "Emonets: Multimodal deep learning ap- proaches for emotion recognition in video." In *Journal on Multimodal User Interfaces, 10(2):99– 111.*

Thuiller, W. (**2003**). "Biomod–optimizing predictions of species distributions and projecting potential future shifts under global change." Global change biology, 9(10), 1353–1362.

Tieleman, T. and Hinton, G. (**2012**). "Lecture 6.5- rmsprop: Divide the gradient by a running average of its recent magnitude." COURSERA: Neural networks for machine learning, 4(2):26-31.

Wood, S.N. (**2006**). "Generalized additive models: an introduction with r."

Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., and Narayanan, S. (**2010**). "Context-sensitive multi-modal emotion recognition from speech and facial expression using bidirectional lstm modeling." In *Proc. INTERSPEECH 2010, Makuhari, Japan, pages 2362–2365.*

Yackulic, C.B.e.a. (**2012**). "Presence-only modelling using maxent: when can we trust the inferences?" Methods Ecol. Evol. 4: 236–243.

Appendix G

Data augmentation for marine mammal vocal sound classification: from naive to physically-based schemes.

Data augmentation methods

Data augmentation for marine mammal vocal sound classification: from naive to physically-based schemes.

Paul Nguyen Hong Duc,
Paul R. White, Odile Gerard, Olivier Adam, Dorian Cazau

November 3, 2020

Contents

Abstract

Especially, data driven and physically-based schemes offered promising results as a tool for annotation aid and data discovery in the practical application case where only a small set of high-SNR clean vocalization is available.

1 Intro

1.1 Context

In most data science domains, deep learning methods have outperformed state-of-the-art methods. However, the main drawback of deep learning methods is the huge amount of labelled data they need to supervise the learning of their networks, and such requirement often limits their applications to many Earth sciences (Reichstein, 2019), including marine bioacoustics.

Indeed, bioacoustics datasets are often scarce and small. To put this into perspective, note that Mobysound, one of the most well known published dataset of annotated vocalizations of marine mammals, have only a few thousands of annotated sound samples for approximately 20 classification classes corresponding to different species, whereas in the vision computer community a database like ImageNet (a database of human-labelled images, with labels like, for example, "cat" or "dog" (Deng et al., 2009)) gathers more than 3.2 millions images in 2014. It is noteworthy that in urban acoustics, manually annotated datasets for sound event detection with strong labels are also very limited in size (e.g. the TUT Sound Events 2016 development set is 78 minutes long), although considerable efforts are now made (e.g., Piczak (2015); Mesaros et al. (2016)). Then, large, labelled bioacoustics datasets barely exist, not only because of the sizes of the datasets involved, but also owing to the conceptual difficulty in labelling datasets; for example, determining that an image depicts a cat is much easier than determining that a recording contains a vocalization, given that vocalizations are contingent on multiple parameters specific to sources (e.g., level, distance to hydrophone) and environment (e.g., underwater propagation) and can also change according to the methods used to collect and analyse the data, and that there are not enough labelled cases for training a machine learning system. A recent comparative study on individual annotation processes has revealed this potential error-prone human bias (Leroy, 2018). Furthermore, annotation tasks are laborious and time-consuming and can affect the quality of annotation over time.

Another source of complexity with annotated bioacoustics datasets is that some species are often under-represented in the labels, creating imbalanced datasets. A dataset is called imbalanced if it contains many more samples from one class than from the rest of the classes. Datasets are unbalanced when at least one class is represented by only a small number of training examples (called the minority class) while other classes make up the majority. In this scenario, classifiers can have good accuracy on the majority class but very poor

accuracy on the minority class(es) due to the influence that the larger majority class. It also creates a problem of making a validation or test sample as its difficult to have representation across classes in case number of observation for few classes is extremely less.

Marine bioacoustics then shares with other geoscientific disciplines (Reichstein, 2019) the need to develop machine learning methods to learn from few labelled examples and unbalanced classes. One modern approach to tackle this issue is data augmentation, which basically consists in augmenting the number of training audio files by producing other files each with slight variations from the original. Generally, by augmenting the dataset, robustness of created models is improved because more variance and diversity is fed into machine learning algorithms. In other words, by training the model on the additional deformed data, the hope is that the network becomes invariant to these deformations and generalizes better to unseen data. Object recognition is a classification task that is especially amenable to this form of dataset augmentation because the class is invariant to so many transformations and the input can be easily transformed with many geometric operations. Classifiers can for example benefit from random translations, rotations, and in some cases, flips of the input to augment the dataset. Data augmentation is also an efficient way to integrate system knowledge into a machine learning framework, expanding a training dataset for undersampled domains with physical constraints (Xie et al., 2018). For instance, it was reported that selecting vocal tract length perturbation warping factors from a limited set of perturbation factors was better (Jaitly and Hinton, 2013; Cui et al., 2015).

Over the last years, data augmentation has been widely used in conjunction with deep learning architectures as among most machine learning models they may benefit the most of increasing dataset training size and its diversity (). While deep learning based approaches have already been employed in the field of marine bioacoustics (e.g., for), more extensive explorations of such approaches have been precisely limited by the small size of labelled datasets ().

1.2 Contributions

In this work, we wish to explore the use of data augmentation strategies within deep learning frameworks to improve classification performance on small size whale vocalization datasets with unbalanced classes. We compare three different data augmentation strategies, ranked from the most "naive" (i.e., less informed) to the most informed problem-specific ones. First, naive augmentations are low-level deformations of audio signal properties, such as noise (classically white, blue or pink) corruption, time stretching, pitch shifting, dynamic range compression, generally parametrized within value ranges arbitrarily set. Second, data-driven augmentation uses generative models directly learnt from non-labelled real data. Using real noise distribution might result in more realistic augmentations as it is well known that in underwater acoustics the assumption of noise to be additive Gaussian is seldom valid in ocean environments (Wenz, 2005; Panaro et al., 2012; Rahmati et al., 2014). Third, physically-based aug-

mentation integrates underwater propagation models parametrized by the site characteristics (e.g. local sound speed in the water column) to describe the physical deformations of a source signal through the propagation medium. These three data augmentation strategies have been implemented on top of different classifiers, and benchmarked w.r.t their performance in classifying low-frequency whale vocalizations from the DCLDE 2015 LF challenge.

1.3 Related works

In the following, we review related research studies that have employed at least one of the three data augmentation strategies described above to audio processing problems.

First, the naive approach has been widely used over the past years, following what has been done in image processing. Thus, while images have been rotated, cropped.. sounds have been augmented through low-level transformations like time stretching, pitch shifting, dynamic range compression.. as demonstrated by the different classification models proposed in the DCASE challenges[1]. Overall, according to Sakashita and Aono (2018) and Himawan (2018), such transformations applied blindly on sound samples did not offer significant improvement in DCASE classification tasks.

Wei (2018) showed that naive augmentation (i.e. pitch shifting, time stretching and adding background noise) does not show much improvement in audio tagging, and explore the sample mixed data augmentation for the domestic audio tagging task, including mixup, sample pairing and extrapolation, achieving a state-of-the-art of equal error rate of 0.10 on DCASE 2016 task4 dataset.

Comparatively, data-driven and physically-based approaches have been much less explored, at the exception of Ferguson (2016) in an underwater context, where average power spectral density have been computed from real data and used to generate new background noise samples. Also, denoising autoencoders have been proved to be most effective when the noise injected during training matches the actual noise in the data distribution (Xie et al., 2012).

2 Data augmentation schemes

Table 1 details the different data augmentation schemes and their parameter ranges tested.

2.1 Naive

Below we provide implementation details on data augmentation techniques from our first naive approach.

[1]See DCASE challenge webpages here `http://dcase.community/challenge2019/`.

	Data augmentation schemes	Acronym	Description	Parametrization
Naive	Pitch-shift	PS	Pitch raised or lowered while keeping its duration the same	12-valued linear variation between -3 and 3 values (by semitones)
	Time-stretching	TS	Audio signal speed up or slowed down without modifying the pitch	12-valued linear variation between 0.7 and 1.35
	Dynamic Range Compression	DRC	The audio signal dynamic range is compressed randomly.	Compression range between 0.75 to 1.25 of the original dynamic range
Data	Noise Corruption	NC	Adding a background noise learnt from real data to the audio sample.	Random mixing coefficient sampled uniformly between 0.7 and 1.35
Physics	Pekeris waveguide model	phys	Sound transformation through a Pekeris waveguide model with fixed seabed properties.	See table 2.
	Pekeris waveguide model	phys-seabed	Sound transformation through a Pekeris waveguide model with variable seabed properties.	See table 2.

Table 1: Tested different data augmentation schemes and their parameter ranges.

Pitch-shift (PS) The pitch of the audio sample is raised or lowered while keeping its duration the same. Each audio sample was pitch shifted by -3 to 3 values (by semitones);

Time-stretching (TS) The audio signal is speed up or slowed down without modifying the pitch. For instance, an offset of several seconds can be added to samples (Piczak, 2017). In our work, each audio sample was speed up or slowed down by 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.1, 1.15, 1.2, 1.25, 1.3, 1.35 % of the original duration;

Dynamic Range Compression (DRC) The audio signal dynamic range is compressed randomly within a range of 0.8 to 1.2 the original dynamic range. We used open-source codes from Panotti (`https://github.com/drscotthawley/panotti`) to implement this deformation;

Furthermore, following Salamon and Bello (2017), all these augmentations were applied directly to the audio signal prior to converting it into the input representation used to train the network (log-mel-spectrogram). The MUDA library (`https://muda.readthedocs.io/en/latest/deformers.html`) was used to compute pitch shift and time stretching deformations.

2.2 Data-based

Drawing from Ferguson (2016)'s data augmentation method, we have developed a data-based augmentation technique called Noise Corruption (NC), consisted in adding to each original audio sample a background noise learnt from real data. All 5s-long background noise samples have been aurally and visually checked to verify that they do not contain any short-term acoustic events.

Using MUDA library, from each background noise signal, $n\_samples$ clips are randomly extracted and mixed with the input audio with a random mixing coefficient sampled uniformly between $weight\_min$ and $weight\_max$, respectively set to 0.7 and 1.35. Only one sample per noise signal was used (n_samples = 1). 45 noise samples were extracted from CINMS recordings, 67 from DCPP_A and

17 from DCPP_C. For each noise corruption augmentation, only noise samples from the same site of the audio to modify were used.

2.3 Physically-based

This third and last data augmentation approach aims to constraint augmentations with some model-based physical knowledge specific to our recording environment. We are then interested in modeling acoustic propagation in the environment assuming that the acoustic channel can be modeled as a linear time-invariant space-variant filter using a Pekeris waveguide (Pekeris, 1945) and beam tracing model using Bellhop (Porter, 2011).

Local depth of the water column at the hydrophone locations were extracted using the GEBCO 2D bathymetric maps. All audio samples are first bandpass filtered between 15 and 150 HZ, corresponding roughly to the frequency range of the whale vocalizations. All model parameters are summed up in table 2.

	Variable	**Name**	**Numerical values**
Channel	Density in water	ρ_w	1000
	Density in seabed	ρ_s	Random integers in [1800:2800]
	Sound speed in water	c_w	1500
	Sound speed in seabed	c_s	Random integers in [1700:2200]
	Water depth	D_w	100
Source / Reception	Source depth	d	Random integers in [15:200]
	Reception depth	d_r	[600,65,1000]
	Distance source / reception	r	Random integers in [2000:5000]

Table 2: Parameters of the Pekeris model and numerical values for the LF-DCLDE2015 dataset.

3 Experimental set-up

3.1 Original dataset

The 2015 low-frequency DCLDE dataset was recorded with High-frequency Acoustic Recordings Packages deployed off the southern and central coast of California at different locations, spanning all four seasons, over 2009-2013 period (see the full dataset documentation at `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`). The sampling rate is 2000 Hz for recorders Channel Islands National Marine Sanctuary (CINMS, site B in figure 1) and 3200 Hz for DCPP (Diablo Canyon Power Plant, sites A & C in figure 1)) recorders, and depths of 600 , 65 and 1000 m, respectively. As a consequence, we resampled DCCP recordings to 2000 Hz. As this dataset was used in the DCLDE international challenge on detection and classification of marine mammals in 2015, it has already been annotated by two independent experts, with a total of 5211 strong labels (i.e. with start and end times of events) over 2 whale species classes that are highly unbalanced: blue whale D calls (4796 samples) and fin whale 40 Hz calls (415 samples).

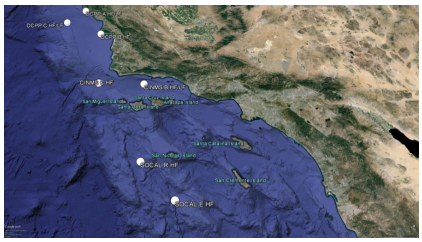


Figure 1: Recorder sites of 2015 low-frequency DCLDE dataset (taken from `http://www.cetus.ucsd.edu/dclde/datasetDocumentation.html`).

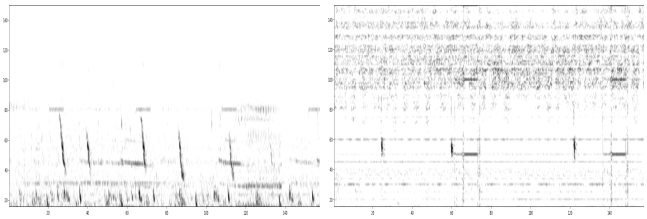


Figure 2: Spectrograms of sounds to identify: Dcall in the 40-80 Hz frequency band (left), 40-Hz pulse in the 40-60 Hz frequency band (right). Y axis: frequencies in Hz, X axis: time in seconds. APLOSE platform was used to display these spectrograms with 4096 Fast Fourier Transform points and an overlap of 90%.

3.2 Classification model

Description A ResNet is a deep neural network using skip connections or short-cuts which jumps over some layers (He et al., 2015; Schaetti, 2018). The motivation behind skipping layers in ANN is to avoid the well known problem of vanishing gradients using activation from a previous layer until the next one has learned its weights. Usually ResNet architectures are designed with a hundreds of layers because its structure enables to successfully deal with the vanishing gradient issue when training a neural network with a lot of layers. ResNet models won several competition in computer vision. In our study, only 18 layers were stacked in the ResNet to avoid overfitting as the training set is not very large. It was trained from scratch to handle the size of the melspectrogram images (110×90) instead of the initial shape of 224×224. Our implemented

version is based on existing open source codes[2].

Implementation details Spectrograms were extracted from audio samples. Both the FFT size and the window size between two frames were set to 2000 with 90% overlap. Inputs of the neural network was images of size 180×40. All audio samples were extracted in a way that each was 5s long.

The number of epochs and the batch size were set to 20 and 128 respectively when using Resnet18 and CNN architecrures. Default parameters were kept for training the LightGBM classifier.

3.3 Dataset rebalancing

While data augmentation can be directly applied to unbalanced datasets, such as our original dataset (named *original*), we also built two other non-augmented datasets which have been rebalanced (i.e. each class has the same number of samples) using weight balancing and a sample selection criterion based on the highest signal-to-noise ratio (SNR). While the first one (named *weightBal*) is used as a benchmarking approach that does not use data augmentation to deal with unbalanced datasets, the second one (named *highestSNR_N*) is used as a real-life case study in marine bioacoustics where a labelled dataset might contain only very clean and high-level vocal sounds.

3.3.1 Weight balancing

Weight balancing consists in setting different class weights in the loss function of our models so as to over-penalize the miss-classification of the under-represented class compared to other classes. Using cross entropy as loss function, it can be modified for this purpose as follows. Let y_{ik} be 1 if k is the true class of data point i, and be 0 otherwise, and $y'_{ik} \in (0, 1]$ be the corresponding model estimation. The original cross-entropy can be written as:

$$H_y(y') = -\sum_i \sum_{k=1}^{K} w_k y_{ik} log(y'_{ik}) \tag{1}$$

This is called class weight balancing. We set the class weights to Bm: 1 and Bp: 11.57 during the training phase of the model (class_weight parameter in Keras). This means that 1 occurrence of Bp weighs 11 Bm instances in the loss function. Note that rebalancing with data augmented oversampling and class weighting are equivalent. Indeed, copying the samples of a class 3X is equivalent to assigning a 3X weight to the class. However, the weighting is better from storage and computational point of view since it avoids working with a larger data-set, but keep in mind that this is not a data augmentation technique.

[2] `https://github.com/keras-team/keras-contrib/blob/master/keras_contrib/applications/resnet.py`

3.3.2 Highest signal-to-noise ratio (SNR) criterion

Here we followed the idea that to achieve more realistic augmentations we will need vocal sounds that have not been already too distorted during the propagation and acquisition process. Such sounds should correspond to the ones presenting the highest SNR, i.e. the closest ones to the hydrophone (following the hypothesis of a constant source level).

Then, in this rebalancing approach, a same number of audio samples was selected per class based on their SNR. This ratio was calculated following Socheleau et al. (2015).

The number of samples per class N was set to 50 and 150, building respectively the datasets *highestSNR_50* and *highestSNR_150*.

3.4 Evaluation procedures

Evaluation of all our experimentations is done within a 5-fold cross validation, using the four datasets described above, namely *original*, *weightBal*, *highestSNR_50*, *highestSNR_150*. In each fold, we first randomly sample 80 % of the original dataset to build a training dataset and use the remaining 20 % ones as a test dataset, for both classes. Naturally, we split into test and train sets before applying augmentation techniques to keep only real non-augmented samples in test set. During the test phase, we compute balanced accuracy (Brodersen et al., 2010) on the overall confusion matrix, as obtained by summing individual confusion matrices across all folds (leading to the posterior of the algorithm as a whole)[3], using the implementation from sklearn[4] package. This metric basically avoids that our classifier is biased towards the more frequent class, yielding too optimistic accuracy estimate. In an extreme case, the classifier might assign every single test case to the large class, thereby achieving an accuracy equal to the fraction of the more frequent labels in the test set. Note that using such a metric, although the training sets *original* and *weightBal* are unbalanced, they still exhibit a pure chance level accuracy of 50 %. Table 2 shows the details of sample number per species (Bm (D-calls) / Bp (40 Hz calls)) and per dataset during the training phase of our experiments.

In all our experiments, we apply an augmentation ratio of 2, i.e. for example in the augmented version of Bp (40 Hz calls): $332 + 332\times\ 2 = 996$, and each augmentation has been applied individually.

[3]Note that it can also be defined as the average of recall obtained on each class.

[4]`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html`.

Dataset	Baselines	Augmented
original	3836 / 332	11505 / 996
weightBal	3836 / 332	-
highSNR_50	50 / 50	150 / 150
highSNR_150	150 / 150	450 / 450

Table 3: Number of training samples in each fold of the cross-validation for the species Bm (D-calls) / Bp (40 Hz calls), and for the different evaluation datasets both as baselines (i.e. non-augmented) and augmented versions.

4 Experimentations

Our experimentations have been designed around the two following questions:

- Does data augmentation improve non-augmented baselines ?
- Do physically-based augmentation techniques perform better than naive ones ?

4.1 Does data augmentation improve non-augmented baselines ?

Figure 3 shows balanced classification accuracy for our different baselines, which do not use any augmentation technique (i.e. model names without subscript), plus models with weight balancing (i.e. model names with "wei") and those using the underSet training set (see table 3). Best classification performance are obtained with weight balancing, bringing balanced accuracy gains up to 20 %. A trivial explanation is that due to re-balancing during training, a same number of samples of each class is used and consequently these methods will tend to highly increase performance on the minority class while slightly decrease performance on the majority class, resulting in a positive gain for the overall accuracy. Furthermore, both the random undersampling strategy (i.e. resnet18_underset) and SNR50 drastically decrease performance from the unbalanced original dataset, revealing that there is a high possibility that the data we are either randomly deleting or removing based on their too small SNR actually contain important information about the predictive classes.

4.2 Do physically-based augmentation techniques perform better than naive ones ?

Indeed, one more specific questions we wanted to investigate was whether more site-specific and model-based augmentation techniques would bring better performance gains in comparison to more naive and blind techniques. In the following we review the different classes of augmentation techniques, contrasting their respective performance.

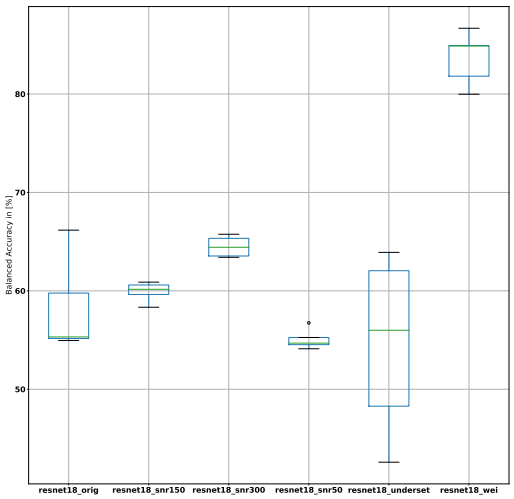


Figure 3: Balanced classification accuracy for our different baselines using the model resnet_18, which do not use any augmentation technique (i.e. model names without subscript), plus models with weight balancing (i.e. model names with "wei") and those using the underSet training set (see table 3).

Globally, naive techniques such as time stretching and pitch shifting proved to have quite similar behaviours, bringing either consistent negative gains (e.g. from figure 13) or inconsistent contributions w.r.t the number of augmentation ratio (e.g. from figures 6 and 7). We can explain this by the fact that 40Hz pulses are rather short sounds compared to Dcalls. Then, with time stretching, we will tend to add confusion between the two by losing the invariance of the distribution of the 40 Hz class. Also, both sounds have similar frequency ranges (between 40Hz-70Hz for 40Hz pulses and about 30-90Hz for Dcalls (Sirovic, 2014)), and so most likely the sound transformations brought by pitch shifting, applied similarly to both classes, do not allow for a discriminative learning.

For what concerns, the data-driven technique NC offered various performance depending on the augmentation scenario. It offers interesting positive gains up to 4 % for underset, contrasting here with all other data augmentation techniques, but negative ones for SNR50.

Eventually, for the SNR150 dataset, the two physically-based augmentations, namely phys and phys-seabed, are the only ones to consistently bring positive gains for both the $\times$ 2 and $\times$ 12 augmentation ratios. Otherwise, they brought negative gains for the underset dataset (see figure 13) and better improvements on the orig dataset (see figure 4), compared to to the other augmentation techniques. Consequently, it seems important that such physically-based augmentations are applied to clean samples, i.e. high SNR, as already noisy signals would likely be unrealistically degraded.

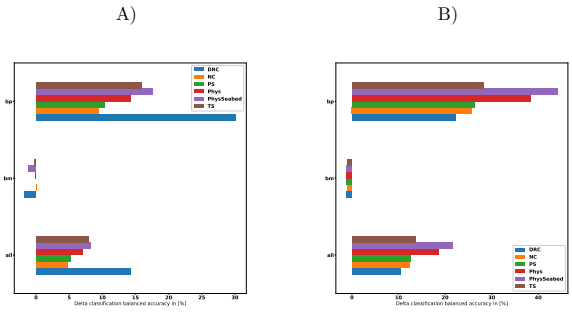


Figure 4: Delta balanced accuracy for the orig augmentation datasets using the resnet18 model, w.r.t to the resnet18_orig baseline, using augmentation ratios of $\times$ 2 (A) and 12 (B).

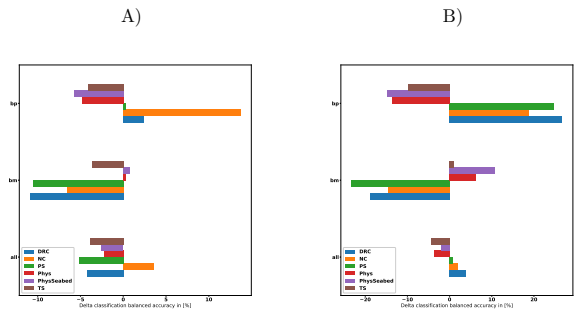


Figure 5: Delta balanced accuracy for the underset augmentation dataset using the resnet18 model, w.r.t to the resnet18_underset baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

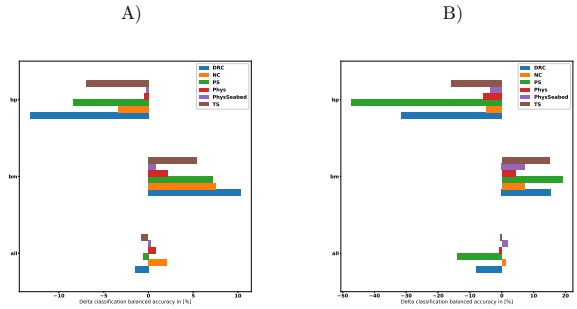


Figure 6: Delta balanced accuracy for the SNR50 augmentation dataset using the resnet18 model, w.r.t to the resnet18_SNR50 baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

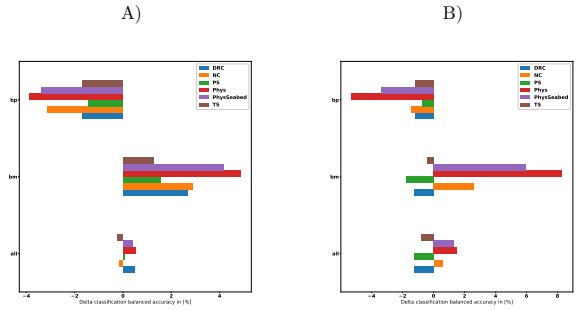


Figure 7: Delta balanced accuracy for the SNR150 augmentation dataset using the resnet18 model, w.r.t to the resnet18_SNR150 baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

5 General discussion

In this study we investigated whether data augmentation could bring classification improvement on an unbalanced two-species acoustic dataset. To do so, the augmentation process will have to enrich diversity of training samples in each class so the classifier can better generalize, but without changing the class sample distributions, which would make samples from the different classes more likeable and so more difficult to separate during classification.

All baseline systems had a very low balanced accuracy score. One reason could be that 5s long is too long compared to the mean duration of a fin whale 40 Hz call (about 1s). Another one is the hyperparameters of the network that were not tuned. The number of epochs was maybe too low and the learning rate too high.

Overall, we have first shown from the baseline performance that significant performance gains were obtained using data augmentation on both the *original* dataset and the reduced datasets *highSNR_50* and *highSNR_150*.

However, note that similar gains were obtained by simply re-balancing the *original* dataset during the training process through weight balancing (*weightBal* dataset). Despite its ease of implementation, it is barely reported in similar benchmarking reports, e.g. Salamon et al. (2017). Also, in contrast to Salamon and Bello (2017), directly augmenting unbalanced dataset did not prove to be efficient.

Our most interesting result with data augmentation is how it allows to handle an explicit problem of generalization from non-deformed clean sounds to prop-

agated distorted sounds, as formalized in our study by training models on the SNR* datasets while evaluating them on a non-overlapping sub-part of the original dataset. Indeed, we saw that augmenting the datasets SNR50 and SNR150 allow to increase significantly their respective baselines up to 10 % for SNR150 and even to reach similar performance than the best-performing one, namely weight balancing, for orig. As these SNR* datasets have a much smaller training dataset than orig and wei datasets (i.e. 3836 and 332 samples respectively for the Bm and Bp calls), and as they are also more quickly and easily annotated [5], then such augmentation strategy might be used as an efficient annotation assisting tool by avoiding exhaustive annotation campaigns. Annotators would focus on the most salient sounds, and then the machine could recommend new samples to annotate from the learning of the augmented dataset.

Another question we investigated in this study is to know whether the *a priori* complexity (in terms of modeling information) of the data augmentation scheme matter. To do so we compared on a same benchmark three different schemes, namely naive, data-driven and physically-based ones.

One interesting future research avenue here would be to optimize the choice of augmentation techniques w.r.t the class sample distribution features, and the parametrization of the augmentation ranges. Past studies (Salamon and Bello, 2017) have shown that the use of data augmentation could be improved by applying class-conditional data augmentation, after observing that the performance of the model for each sound class is influenced differently by each augmentation set. In the same vein, recommendations are given with the generation of synthetic soundscapes that can be not plausible (e.g. using SCAPER (Salamon et al., 2017)), and so the soundscape parameters need to be chosen conscientiously and as a function of the specific domain application. Then, augmented datasets must not go too far from their original data generation distribution. In other words, adding deformations to existing ones might lead to severe unrealistic signatures, which explained why data augmentation proved to be so efficient, especially with physically-based schemes, on the SNR* datasets where most vocal sounds are clean and robust patterns without transformations.

6 Conclusion

In a near future, we will look at other scenarios of applications, such as multi-class classification including background noise, and see we could more optimally combine the different ingredients of a data augmentation, namely the augmented features, the augmentation method and the model according to these different applications. Another future direction will be to artificially create vocalisations from a vocal generator model depending of the considered specie and deform it through a propagation medium.

Eventually, going back to our real-life study cases, data augmentation could

[5] As they exhibit robust patterns with a high SNR, which are generally rapidly catchable by manual annotators and make good agreement between different annotators (avoiding e.g. the typical hesitation of picking or not a vocalization when it is too deformed).

be a way to avoid manually labelling long PAM time series. An experiment should be carried out to approximately determine the minimum number of samples needed to achieve the same classification performances with the whole dataset.

A1. Results for CNN

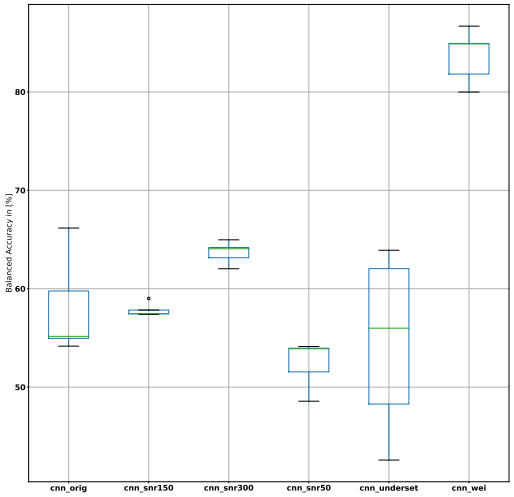


Figure 8: Balanced classification accuracy for our different baselines using the model CNN, which do not use any augmentation technique (i.e. model names without subscript), plus models with weight balancing (i.e. model names with "wei") and those using the underSet training set (see table 3).

A2. Results for LGB

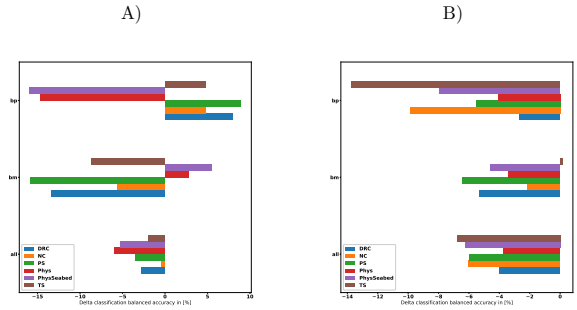


Figure 9: Delta balanced accuracy for the underset augmentation dataset using the cnn model, w.r.t to the cnn_underset baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

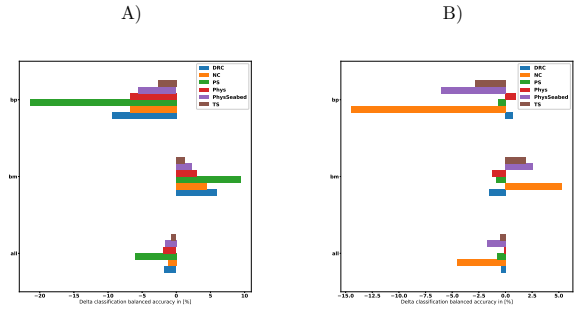


Figure 10: Delta balanced accuracy for the SNR50 augmentation dataset using the CNN model, w.r.t to the cnn_SNR50 baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

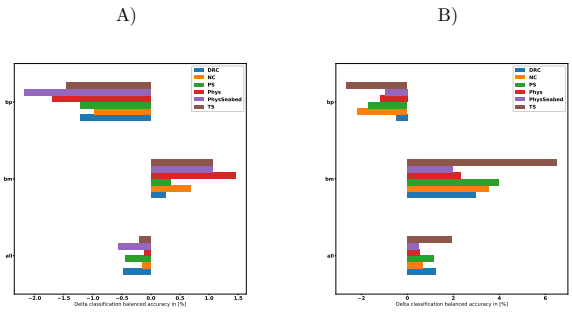


Figure 11: Delta balanced accuracy for the SNR150 augmentation dataset using the CNN model, w.r.t to the cnn_SNR150 baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

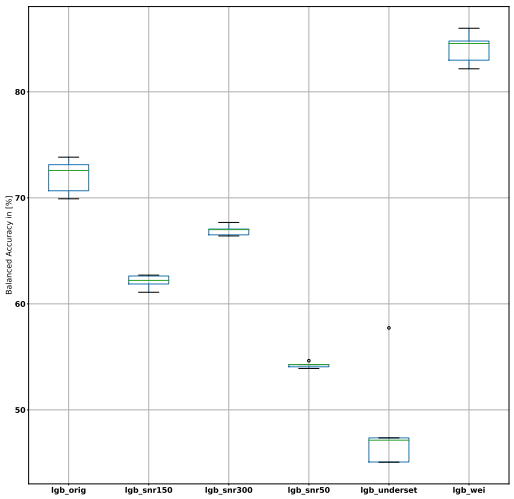


Figure 12: Balanced classification accuracy for our different baselines using the model LGB, which do not use any augmentation technique (i.e. model names without subscript), plus models with weight balancing (i.e. model names with "wei") and those using the underSet training set (see table 3).

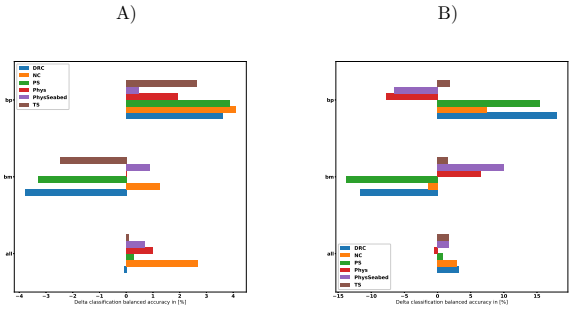


Figure 13: Delta balanced accuracy for the underset augmentation dataset using the LGB model, w.r.t to the lgb_underset baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

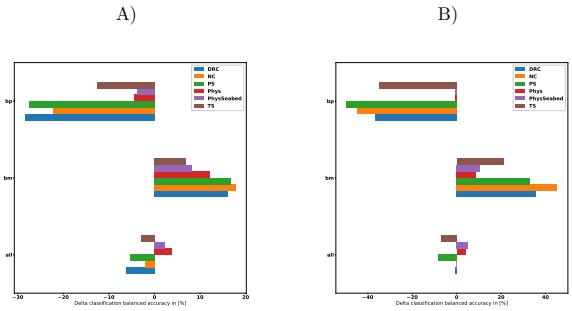


Figure 14: Delta balanced accuracy for the SNR50 augmentation dataset using the LGB model, w.r.t to the lgb_SNR50 baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

A) B)

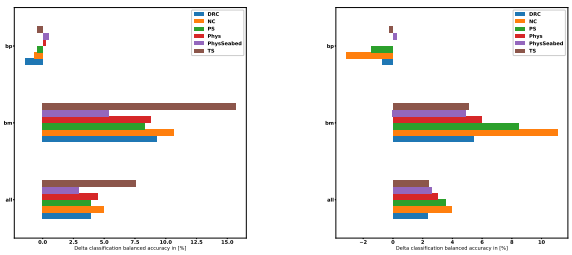


Figure 15: Delta balanced accuracy for the SNR150 augmentation dataset using the LGB model, w.r.t to the LGB_SNR150 baseline. On the left graph, an augmentation ratio of $\times$ 2 has been used, and a ratio of $\times$ 12 on the right graph.

References

Brodersen, K.H., Ong, C.S., Stephan, K.E., and Buhmann, J.M. (**2010**). "The balanced accuracy and its posterior distribution." In *International Conference on Pattern Recognition*.

Cui, X., Goel, V., and Kingsbury, B. (**2015**). "Data augmentation for deep neural network acoustic modeling." In *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP);23(9):1469–1477*.

Deng, J., Wei Dong, Richard Socher, L.J.L.K.L., and Fei-Fei., L. (**2009**). "Imagenet: A large-scale hierarchical image database." In *In Proc. of the IEEE Cnference on Computer Vision and Pattern Recognition. IEEE, 248–255*.

Ferguson, E.L.e.a. (**2016**). "Convolutional neural networks for passive monitoring of a shallow water environment using a single sensor." arxiv.

He, K., Zhang, X., Ren, S., and Sun, J. (**2015**). "Deep residual learning for image recognition." In *arXiv:1512.03385v1*.

Himawan, I.e.a. (**2018**). "3d convolution recurrent neural networks for bird sound detection."

Jaitly, N. and Hinton, G. (**2013**). "Vocal tract length perturbation (vtlp) improves speech recognition. in: Proc. icml workshop on deep learning for audio, speech and language. 2013, ."

Leroy, E.e.a. (**2018**). "On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions." J. Acoust. Soc. Am.

Mesaros, A., Fagerlund, E., Hiltunen, A., Heittola, T., and Virtanen, T. (**2016**). "Tut sound events 2016, development dataset." URL `Availableonline: http://dx.doi.org/10.5281/zenodo.45759`.

Panaro, J.S., Lopes, F.R., Matos, L.J., and Barreira, L.M. (**2012**). "Empirical noise model and likelihood metrics for underwater acoustic communications." In *Proc. of the IEEE Conference on Underwater Communications Networking (UComms), Sestri Levante, Italy*.

Pekeris, C. (**1945**). *Theory of propagation of explosive sound in shallow water*.

Piczak, K.J. (**2015**). "Esc: Dataset for environmental sound classification, 23rd acm international conference on multimedia, brisbane, australia, oct. 2015, pp. 1015–1018."

Piczak, K.J. (**2017**). "The details that matter: Frequency resolution of spectrograms in acoustic scene classification." In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*. pp. 103–107.

Porter, M.B. (**2011**). "The bellhop manual and user's guide: Preliminary draft." Heat, Light, and Sound Research, Inc., La Jolla, CA, USA, Tech. Rep, **260**.

Rahmati, M., Pandey, P., and Pompili, D. (**2014**). "Separation and classification of underwater acoustic sources." In *Underwater Communications and Networking (UComms), 2014*. pp. 1–5.

Reichstein, M.e.a. (**2019**). "Deep learning and process understanding for data-driven earth system science." Perspective Nature.

Sakashita, Y. and Aono, M. (**2018**). "Acoustic scene classification by ensemble of spectrograms based onadaptive temporal divisions."

Salamon, J. and Bello, J.P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification." IEEE Signal Proc. Letters.

Salamon, J., MacConnell, D., Cartwright, M., Li, P., and Bello, J.P. (**2017**). "Scaper: A library for soundscape synthesis and augmentation." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 15-18, 2017, New Paltz, NY*.

Schaetti, N. (**2018**). "Character-based convolutional neural network andresnet18 for twitter author profiling." In *Notebook for PAN at CLEF 2018*.

Sirovic, A.e.a. (**2014**). "Bryde's whale calls recorded in the gulf of mexico." Marine Mammal Science, 30(1): 399–409.

Socheleau, F.X., Leroy, E., Carvallo Pecci, A., Samaran, F., Bonnel, J., and Royer, J.Y. (**2015**). "Automated detection of antarctic blue whale calls." The Journal of the Acoustical Society of America, **138**, 3105–3117. URL `http://scitation.aip.org/content/asa/journal/jasa/138/5/10.1121/1.4934271`.

Wei, S.e.a. (**2018**). "Sample mixed-based data augmentation for domestic audio tagging." In *arXiv:1808.03883v1*.

Wenz, G.M. (**2005**). "Acoustic ambient noise in the ocean: spectra and sources." The Journal of the Acoustical Society of America, vol. 34, no. 12, pp. 1936–1956.

Xie, J., Xu, L., and Chen, E. (**2012**). "Image denoising and inpainting with deep neural networks." In *Advances in Neural Information Processing Systems, pp. 341–349*.

Xie, Y., Franz, E., Chu, M., and Thuerey, N. (**2018**). "tempogan: a temporally coherent, volumetric gan for super-resolution fluid flow." Preprint at https://arxiv.org/ abs/1801.09710.

Appendix H

DCASE Challenge technical report.

ACOUSTIC SCENE CLASSIFICATION USING LONG-TERM AND FINE-SCALE AUDIO REPRESENTATIONS

Technical Report

Paul Nguyen Hong Duc[1*], *Dorian Cazau*[2], *Olivier Adam*[1],
Odile Gerard[3], *Paul R. White*[4]

[1] Institut d'Alembert, Sorbonne Universite, Paris, France
p.nguyenhongduc@gmail.com, olivier.adam@sorbonne-universite.fr
[2] ENSTA Bretagne, Lab-STICC, Brest, France,
dorian.cazau@ensta-bretagne.org
[3] DGA-TN, Toulon, France, odile.gerard@intradef.gouv.fr
[4] University of Southampton, ISVR, Southampton, UK, P.R.White@soton.ac.uk

ABSTRACT

Audio scene classification (ASC) is an emerging filed of research in different scientific communities such as urban soundscape characterization or bioacoustics. It has gained visibility and relevance with open challenges especially with the benchmark dataset and evaluation from DCASE. This paper present our deep learning model to address the ASC task of the DCASE 2020 challenge edition. The model exploits multiple long-term and fine-scale audio representations as inputs of the neural network. Each representation is fed into a different network. The audio embedding of each branch are fused before a Multi-Layer Perceptron to predict the final classes.

***Index Terms*—** ASC, task1b, RMS level, third octave levels, sonic atmosphere features, ensemble method

1. INTRODUCTION

Sound plays a key role in our perception of urban environments. Acoustic scenes classification (ASC) can be essential when visual information is not or partially available. ASC aims at classifying acoustic scenes into predefined classes. For the DCASE 2020 challenge edition, 2 subtasks were proposed to the participants. This report focuses on subtask B which is a classification of 3 acoustic scenes acquired in 12 European cities with the same recording device.

In the DCASE 2020 challenge task 1b, a new taxonomy is introduced. The goal of the challenge is to classify acoustic scenes into three classes: indoor, outdoor and transportation. Moreover, it is required that the neural network size should not exceed 500 KB.

A special focus is put on the audio embedding to meet the model size requirements of the task 1b. By reviewing last DCASE challenge edition, most used audio representations in ASC are spectrogram-like ones and sometimes raw audio waveform. The three classes of the task allow more flexibility on the choice of audio representation. In this report, we use long-term representations combined with sonic atmosphere features and log-mel spectrograms. The following section will cover a detailed explanation of the feature extraction process, system architecture, results, and conclusion.

[*] PhD is granted by the the French Defence Procurement Agency (Direction Générale des Armements).

2. FEATURE EXTRACTION

The AARAE Matlab toolbox was used [1] to compute the interaural cross correlation coefficient of the filtered spectrum and the Leq sound level defined in [1] were using the stereo recordings. Furthermore, audio segments were converted to mono using the librosa Python package [2]. Eight timbral characteristics (hardness, depth, brightness, roughness, warmth, sharpness, boominess, reverb [3]) were computed for each second and then averaged over the whole audio recording. All these features enabled to describe the sonic atmosphere of the acoustic scene with only a few number of features. They are named sonic atmosphere features in the following. Log-mel spectrograms were also extracted with 64 bands. The analysis frame was set to about 85 ms (50% hop size). This enabled to have a low temporal resolution representation of the audio signals. This will help to describe the soundscape at finer temporal and frequency scales than other computed features.

Finally, two long-term representations were used. The Power Spectral Density (PSD) was determined by the Welch method [4] with 1024-point Hamming window, 50% overlap, based on 1s temporal signal segments. As a consequence, the time resolution is 1s and frequency resolution is 46.8 Hz. The root-mean square (RMS) level was then computed. This feature will help to have an overview of the dominant frequencies in the acoustic scene. Furthermore, third octave band levels (TOL) were also evaluated on each second of the 10s-long audio clips. Other 1/n octave bands were tried but 1/3 ones give better results in our experiments. The workflow used to compute RMS level and TOL follows that of [5].

Even if most of the proposed features are extracted from the audio spectrum and the information contained in such representations may be redundant, the objective is to help the model in order to reduce its complexity.

Four different inputs are fed into neural networks. There are matrices of size 10×34 and 512×1 for the TOL and RMS level inputs respectively, and an array of length 10 (for the sonic atmosphere features) are fed into a dense neural network. The log-mel spectrograms are stored in a 64×265 matrix.

In order to increase the number of training samples and to make the model more robust to new data, mixup data augmentation technique is used [6].

3. SYSTEM ARCHITECTURE

Three different models (cf Fig. 1) are trained. Their predictions are averaged to make the final decision (cf Fig. 1D). The averaging ensemble method aims at combining different models to improve predictive performance from any individual model. The variety of features fed into the different models and the different model architectures can improve the ability of the ensemble to generalize to unseen data.

3.1. Model 1 (M1)

Log-mel spectrograms, TOL and sonic atmosphere features are fed into a three branches neural network (cf Fig. 1). Both log-mel spectrograms and TOL are inputs of a 2D convolutional layers. The embedding of the latter is flattened at the end while a global average pooling is performed on the TOL embedding. The sonic atmosphere features are fed in a multi-layer perceptron (MLP) with only two dense layers. The log-mel spectrogram branch of the neural network is inspired by the baseline with a reduction of the input size of the melspectrograms. TOL branch and the sonic atmosphere embedding aim at helping the model to capture acoustic scene time and frequency variations based on several seconds. This model has 47,911 non-zero parameters.

3.2. Model 2 (M2)

The RMS level, TOL and the sonic atmosphere features are the inputs of this three branch model. 1D convolutional layers with different dilatation rates are applied to the RMS level. This enables the network to learn relations between other frequencies than TOL at a low computational cost. The TOL is modified with three Gated Recurrent Units (GRU) to learn different temporal relations on the audio spectrum. These layers are equivalent to Long-Short Term Memory cells but with less computational complexity. In this model, the total number of non-zero parameters is 29,117.

3.3. Model 3 (M3)

Only log-mel spectrograms and TOL are fed into a fully convolutional and a recurrent network respectively. Both models are characterized by a low complexity. However, in the final ensemble model, it weights about a third of the total number of non-zero parameters. It contains 45,465 non-zero parameters.

3.4. Training parameters shared by all models

All experiments were completed with Keras [7] with a Tensorflow backend [8] on a Google Colab GPU environment [9]. All models were trained for 200 epochs in batches of 32 samples. A reduction of the learning rate for each model is set up if the validation loss did not decrease since 3 epochs. An early stopping was used to stop the training and to avoid overfitting.

Class label	Baseline	Best ensemble (M1+M2+M3)	Best ensemble (M1+M2)
indoor	82.0	**86.4**	86.2
outdoor	88.5	**96.1**	95.9
transportation	91.3	**94.7**	**94.7**
Average Acc.	87.3	**92.4**	92.3
Model size	450 KB	478.5 KB	**300.9 KB**

Table 1: Results on the development dataset for our two systems compared to baseline. Characters in bold are the best accuracy (acc.) for each row or the smaller model size.

4. RESULTS

4.1. Dataset

The dataset for task 1b is the TAU Urban Acoustic Scenes 2020 3Class. This subtask addresses acoustic scene classification problem. An audio recording is classified into three different: indoor, outdoor and transportation. These classes represent the place where the recording took place. The dataset consists of 10-seconds stereo audio clips (sampling rate of 48 kHz) from 10 acoustic scenes. In total, 40 hours of audio recording was available as the development dataset.

4.2. Results

For both proposed systems, the baseline is outperformed (cf Table 4.2). The contribution of M3 is limited. Adding this model to the ensemble only improves by 0.1 % the macro-average accuracy while its number of non-zero parameters is about more than 1.5 times higher than M3 ones.

The reason of why some acoustic scene are misclassified was investigated. For example, the misclassification of indoor scenes occurs when the acoustic scene is either really quiet or when there is a specific noise such as the clatter of metro doors.

5. CONCLUSION

In this paper, 2 ensemble models are tried to improve the accuracy of the acoustic scene classification. Long-term but also fine-scale audio representations were combined as inputs to the neural networks. Averaging was considered for the ensemble. The results showed an increase in the classification accuracy as compared to the baseline for both proposed low complexity systems.

6. REFERENCES

[1] D. Cabrera, D. Jimenez, and W. Martens, "Audio and acoustical response analysis environment (aarae): A tool to support education and research in acoustics," *INTERNOISE 2014 - 43rd International Congress on Noise Control Engineering: Improving the World Through Noise Control*, 01 2014.

[2] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, , R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss,

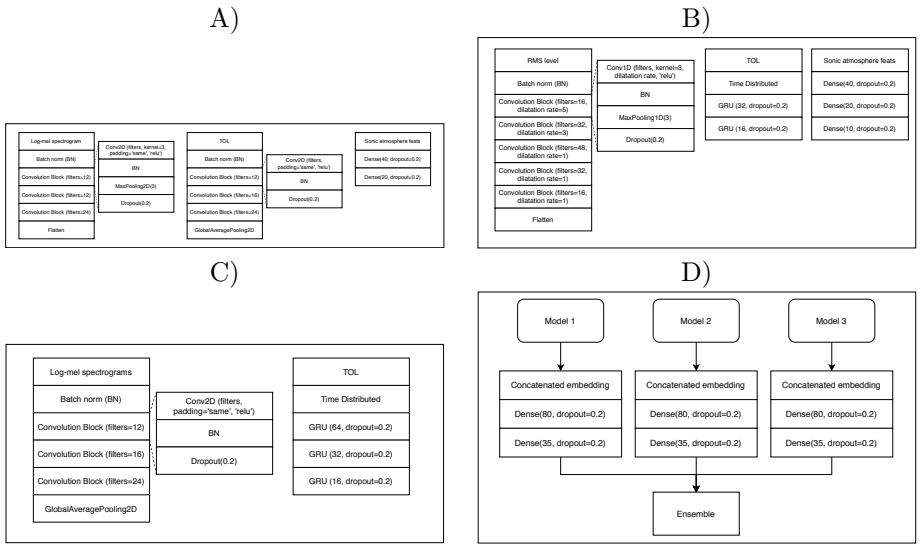


Figure 1: Model graphs. A) Model 1 (M1), B) Model 2 (M2), C) Model 3 (M3) and D) Average ensemble

"librosa/librosa: 0.7.2," Jan. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3606573

[3] A. Pearce, S. Safavi, T. Brookes, R. Mason, W. Wang, and M. Plumbley, "Deliverable d 5.8: Release of timbral characterisation tools for semantically annotating non-musical content," 2019. [Online]. Available: https://www.audiocommons.org/

[4] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 15, pp. 70 – 73, 07 1967.

[5] P. Nguyen Hong Duc, A. Degurse, J. Allemandou, O. Adam, P. R. White, O. Gerard, R. Fablet, and D. Cazau, "A scalable hadoop/spark framework for general-purpose analysis of high volume passive acoustic data," 2019.

[6] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412

[7] F. Chollet *et al.*, "Keras," https://github.com/fchollet/keras, 2015.

[8] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[9] ttps://research.google.com/colaboratory/faq.html.

Bibliography

Adam, O., Cazau, D., Gandilhon, N., Fabre, B., Laitman, J.T., and Reidenberg, J.S. (**2013**). "New acoustic model for humpback whale sound production." Applied Acoustics, **74**, 1182–1190. URL http://www.sciencedirect.com/science/article/pii/S0003682X13000819.

Ahonen, H., Stafford, K.M., de Steur, L., Lydersen, C., Wiig, Ã., and Kovacs, K.M. (**2017**). "The underwater soundscape in western fram strait: Breeding ground of spitsbergen's endangered bowhead whales." Marine Pollution Bulletin, **123**, 97–112. URL http://www.sciencedirect.com/science/article/pii/S0025326X17307543.

Aide, T.M., Corrada Bravo, C., Campos Cerqueira, M., Milan, C., Vega, G., and Alvarez, R. (**2013**). "Real-time bioacoustics monitoring and automated species identification." PeerJ, **1**, e103.

Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (**2013**). "Deep canonical correlation analysis." (PMLR, Atlanta, Georgia, USA), vol. 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255. URL http://proceedings.mlr.press/v28/andrew13.html.

Attenberg, J. and Ertekin, Å. (**2013**). *Class Imbalance and Active Learning* (John Wiley and Sons, Ltd), chap. 6, pp. 101–149. ISBN 9781118646106. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118646106.ch6.

Au, W.W.L., Pack, A.A., Lammers, M.O., Herman, L.M., Deakos, M.H., and Andrews, K. (**2006**). "Acoustic properties of humpback whale songs." J. Acoust. Soc. Am., **120**, 1103–1110.

Bailey, H., Senior, B., Simmons, D., Rusin, J., Picken, G., and Thompson, P. (**2010**). "Assessing underwater noise levels during pile-driving at an offshore windfarm and its potential effects on marine mammals." Marine pollution bulletin, **60**, 888–97.

Balke, S., AbeÃ§er, J., Driedger, J., Dittmar, C., and MÃijller, M. (**2016**). "Towards Evaluating Multiple Predominant Melody Annotations in Jazz Recordings." In *Proceedings of the 17th International Society for Music Information Retrieval Conference* (ISMIR, New York City, United States), pp. 246–252.

Barchiesi, D., Dimitrios, G.D., Stowell, D., and Plumbley, M. (**2015**). "Acoustic scene classification: Classifying environments from the sounds they produce." Signal Processing Magazine, IEEE, **32**, 16–34.

Barlow, J. and Gisiner, R. (**2006**). "Mitigating, monitoring and assessing the effects of anthropogenic sound on beaked whales." Journal of Cetacean Research and Management, **7**, 239–249.

Baumgartner, M.F., Bonnell, J., Van Parijs, S.M., Corkeron, P.J., Hotchkin, C., Ball, K., Pelletier, L.P., Partan, J., Peters, D., Kemp, J., Pietro, J., Newhall, K., Stokes, A., Cole, T.V.N., Quintana, E., and Kraus, S.D. (**2019**). "Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: System description and evaluation." Methods in Ecology and Evolution, **10**, 1476–1489. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13244.

Baumgartner, M.F. and Fratantoni, D.M. (**2008**). "Diel periodicity in both sei whale vocalization rates and the vertical migration of their copepod prey observed from ocean gliders." Limnology and Oceanography, **53**, 2197–2209.

Bergler, C., Schröter, H., Cheng, R., Barth, V., Weber, M., Nöth, E., Hofer, H., and Maier, A. (**2019**). "Orca-spot: An automatic killer whale sound detection toolkit using deep learning." Scientific Reports, **9**.

Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S., and Gruber, D.F. (**2019**). "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics." Scientific Reports, **9**. URL http://www.nature.com/articles/s41598-019-48909-4.

Bertucci, F., Parmentier, E., Berthe, C., Besson, M., Hawkins, A., Aubin, T., and Lecchini, D. (**2017**). "Snapshot recordings provide a first description of the acoustic signatures of deeper habitats adjacent to coral reefs of moorea." PeerJ, **5:e4019**.

Bianco, M.J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M.A., Gannot, S., and Deledalle, C.A. (**2019**). "Machine learning in acoustics: Theory and applications." The Journal of the Acoustical Society of America, **146**, 3590–3628. URL https://doi.org/10.1121/1.5133944.

Biffard, B., Morley, M., Hoeberechts, M., Rempel, A., Dakin, T., Dewey, R.K., and Jenkyns, R. (**2018**). "Adding value to big acoustic data from ocean observatories: Metadata, online processing, and a computing sandbox." J. Acoust. Soc. Am., **144:3**, 1956–1956.

Bittle, M. and Duncan, A. (**2013**). "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring." Annual Conference of the Australian Acoustical Society 2013, Acoustics 2013: Science, Technology and Amenity, pp. 208–215.

Bjørnø, L. (**2003**). "Features of underwater acoustics from aristotle to our time." Acoustical Physics, **49**, 24–30.

Blondel, P. and Hatta, A.A.Z. (**2017**). "Acoustic soundscapes and biodiversity – comparing metrics, seasons and depths with data from the neptune ocean observatory offshore british columbia." In *UACE2017 - 4th Underwater Acoustics Conference and Exhibition*.

Boebel, O., Kindermann, L., Klinck, H., Bornemann, H., Plotz, J., Steinhage, D., Riedel, S., and Burkhardt, E. (**2006**). "Acoustic observatory provides real-time underwater sounds from the antarctic ocean." EOS, **87**, 361–372.

Bohnenstiehl, D., Lyon, R., Caretti, O., Ricci, S., and Eggleston, D.B. (**2018**). "Investigating the utility of ecoacoustic metrics in marine soundscapes." Journal of Ecoacoustics, **2**, R1156L.

Bohnenstiehl, D.R., Lillis, A., and Eggleston, D.B. (**2016**). "The curious acoustic behavior of estuarine snapping shrimp: Temporal patterns of snapping shrimp sound in sub-tidal oyster reef habitat." PLoS ONE, **11**, e0143691. URL https://doi.org/10.1371/journal.pone.0143691.

Bolgan, M., Amorim, M.C., Fonseca, P., Di Iorio, L., and Parmentier, E. (**2018**). "Acoustic complexity of vocal fish communities: a field and controlled validation." Scientific Reports, **8**.

Bolgan, M., O'Brien, J., Winfield, I., and Gammell, M. (**2016**). "An investigation of inland water soundscapes: Which sonic sources influence acoustic levels?" vol. 27.

Bouffaut, L. (**2019**). "Detection and classification in passive acoustic contexts: Application to blue whale low-frequency signals." Ph.D. thesis.

Bradfer-Lawrence, T., Gardner, N., Bunnefeld, L., Bunnefeld, N., Willis, S.G., and Dent, D.H. (**2019**). "Guidelines for the use of acoustic indices in environmental research." Methods in Ecology and Evolution, **10**, 1796–1807. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13254.

Bradley, D.L. (**2008**). "Underwater sound and the marine mammal acoustic environment: a guide to fundamental principles."

Brown, J.C., Smaragdis, P., and McGregor, A.N. (**2010**). "Automatic identification of individual killer whales." J. Acoust. Soc. Am., **128**, 93–98.

Bruno, M., Chung, K.W., Salloum, H., Sedunov, A., Sedunov, N., Sutin, A., Graber, H., and Mallas, P. (**2010**). "Concurrent use of satellite imaging and passive acoustics for maritime domain awareness." In *2010 International WaterSide Security Conference*. pp. 1–8.

Buscaino, G., Ceraulo, M., Pieretti, N., Corrias, V., Farina, A., Filiciotto, F., Maccarrone, V., Grammauta, R., Caruso, F., Giuseppe, A., and Mazzola, S. (**2016**). "Temporal patterns in the soundscape of the shallow waters of a mediterranean marine protected area." Scientific Reports.

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., and Xie, Z. (**2018**). "Deep learning and its applications in biomedicine." Genomics, Proteomics and Bioinformatics, **16**, 17–32. URL http://www.sciencedirect.com/science/article/pii/S1672022918300020.

Cascao, I., Lammers, M., Prieto, R., Santos, R., and Silva, M. (**2020**). "Temporal patterns in acoustic presence and foraging activity of oceanic dolphins at seamounts in the azores." Scientific Reports, **10**.

Cauchy, P., Heywood, K.J., Merchant, N.D., Queste, B.Y., and Testor, P. (**2018**). "Wind Speed Measured from Underwater Gliders Using Passive Acoustics." Journal of Atmospheric and Oceanic Technology, **35**, 2305–2321. URL https://doi.org/10.1175/JTECH-D-17-0209.1.

Cazau, D. and Adam, O. (**2016**). "A plca model for detection of humpback whale sound units." Advances in Applied Acoustics (AIAAS), **5**, 1–5.

Cazau, D., Adam, O., Laitman, J.T., and Reidenberg, J.S. (**2013**). "Understanding the intentional acoustic behavior of humpback whales: a production-based approach." J. Acoust. Soc. Am., **134**, 2268–2273.

Cazau, D., Bonnel, J., Jouma'a, J., Le Bras, Y., and Guinet, C. (**2017**). "Measuring the marine soundscape of the indian ocean with southern elephant seals used as acoustic gliders of opportunity." Journal of Atmospheric and Oceanic Technology, **34**, 207–223.

Chapelle, O., Schlkopf, B., and Zien, A. (**2010**). *Semi-Supervised Learning* (The MIT Press), 1st edn. ISBN 0262514125.

Chen, H., Sun, H., Junejo, N.U.R., Yang, G., and Qi, J. (**2019**). "Whale vocalization classification using feature extraction with resonance sparse signal decomposition and ridge extraction." IEEE Access, **7**, 136358–136368.

Chollet, F. (**2017**). "Xception: Deep learning with depthwise separable convolutions." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1800–1807.

Chu, S., Narayanan, S., and Kuo, C.C. (**2009**). "Environmental sound recognition with time-frequency audio features." IEEE Trans. on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1142–1158, Aug. 2009., **17**, 1142–1158.

Clark, C., Ellison, W., Southall, B.L., Hatch, T., Parijs, S.V., Frankel, A., and Ponirakis, D. (**2009**). "Acoustic masking in marine ecosystems: Intuitions, analysis, and implication." Marine Ecology Progress Series, **395**, 201–222.

Coban, E.B., Pir, D., So, R., and Mandel, M.I. (**2020**). "Transfer learning from youtube soundtracks to tag arctic ecoacoustic recordings." In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 726–730.

Cohn, D., Atlas, L., and Ladner, R. (**1994**). "Improving generalization with active learning." Mach. Learn., **15**, 201–221. URL https://doi.org/10.1023/A:1022673506211.

Danka, T. and Horvath, P. (**2018**). "modAL: A modular active learning framework for Python." URL https://github.com/modAL-python/modAL.

Deecke, V.B., Ford, J.K.B., and Spong, P. (**1999**). "Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (orcinus orca) dialects." J. Acoust. Soc. Am., **105**, 2499–2507.

Dekeling, R.P.A., Tasker, M.L., Van der Graaf, S., Ainslie, M., Andersson, M., André, M., Borsani, J., Brensing, K., Castellote, M., Cronin, D., Dalen, J., Folegot, T., Leaper, R., Pajala, J., Redman, P., Robinson, S., Sigray, P., Sutton, G., Thomsen, F., and Young, J.V. (**2014**). "Monitoring guidance for underwater noise in european seas, part i: Executive summary." Tech. rep.

Delarue, J., Kowarski, K., Maxner, E., MacDonnell, J., and Martin, B. (**2018**). "Acoustic monitoring along canada's east coast: August 2015 to july 2017."

Delfour, F. and Marten, K. (**2001**). "Mirror image processing in three marine mammal species: killer whales (orcinus orca), false killer whales (pseudorca crassidens) and california sea lions (zalophus californianus)." Behavioural Processes, **53**, 181 – 190. URL http://www.sciencedirect.com/science/article/pii/S0376635701001346.

Demchenko, Y., Membrey, P., Grosso, P., and Laat, C. (**2013**). "Addressing big data issues in scientific data infrastructure."

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., and Fei-Fei, L. (**2009**). "Imagenet: A large-scale hierarchical image database." In *In Proc. of the IEEE Cnference on Computer Vision and Pattern Recognition. IEEE, 248–255.*

Dugan, P., Ponirakis, D., Zollweg, J., Pitzrick, M., Morano, J., Warde, A., Rice, A., Clark, C., and Van Parijs, S. (**2011**). "Sedna - bioacoustic analysis toolbox." pp. 1–10. ISBN 978-1-4577-1427-6.

Dumbill, E. (**2012**). "What is big data? - an introduction to the big data landscape." URL http://radar.oreilly.com/2012/01/what-is-big-data.html.

Dunner, C., Parnell, T., Atasu, K., Sifalakis, M., and Pozidis, H. (**2017**). "Understanding and optimizing the performance of distributed machine learning applications on apache spark." IEEE International Conference on Big Data (Big Data).

Eaton, J., Gaubitch, N., Moore, A., and Naylor, P. (**2015**). "The ace challenge - corpus description and performance evaluation."

Eldridge, A., Casey, M., Moscoso, P., and Peck, M. (**2016**). "A new method for ecoacoustics? toward the extraction and evaluation of ecologically-meaningful soundscape components using sparse coding methods." PeerJ, **4**, e2108. URL https://doi.org/10.7717/peerj.2108.

Erbe, C. (**2011**). "Underwater acoustics: Noise and the effects on marine mammals."

Erbe, C., Dunlop, R., Jenner, K., Jenner, M.N., Parnum, I., Parsons, M., Rogers, T., and Salgado Kent, C. (**2017**). "Review of underwater and in-air sounds emitted by australian and antarctic marine mammals." Acoustics Australia, **45**, 1–63.

Erbe, C., Verma, A., McCauley, R., Gavrilov, A.N., and Parnum, I. (**2015**). "The marine soundscape of the perth canyon." Progress in Oceanography, **103**.

Etter, P. (**2012**). "Advanced applications for underwater acoustic modeling." Advances in Acoustics and Vibration, **2012**.

Fan, F., Xiong, J., and Wang, G. (**2020**). "On interpretability of artificial neural networks."

Farcas, A., Thompson, P.M., and Merchant, N.D. (**2016**). "Underwater noise modelling for environmental impact assessment." Environmental Impact Assessment Review, **57**, 114–122.

Farina, A. and Gage, S. (**2017**). *Ecoacoustics: The Ecological Role of Sounds.* 1–336 pp.

Ferrari, M., Glotin, H., Marxer, R., and Asch, M. (**2020**). "DOCC10: Open access dataset of marine mammal transient studies and end-to-end CNN classification." In *IJCNN* (Glasgow, United Kingdom). URL https://hal.archives-ouvertes.fr/hal-02866091.

Figueroa-Mata, G. and Mata-Montero, E. (**2020**). "Using a convolutional siamese network for image-based plant species identification with small datasets." Biomimetics, **5**, 8.

Fleiss, J.L. (**1975**). "Measuring agreement between two judges on the presence or absence of a trait." Biometrics, **31**, 651–659. URL http://www.jstor.org/stable/2529549.

Fonseca, E., Pons, J., Favory, X., Font, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. (**2017**). "Freesound Datasets: A Platform for the Creation of Open Audio Datasets." In *Proceedings of the 18th International Society for Music Information Retrieval Conference* (ISMIR, Suzhou, China), pp. 486–493. URL https://doi.org/10.5281/zenodo.1417159.

Foote, J. (**1999**). "An overview of audio information retrieval." Multimedia Systems, **7**, 2–10.

Frasier, K.E., Roch, M.A., Hodge, L.E.A., Wiggins, S.M., and Hildebrand, J.A. (**2018**). "Machine learning methods to guide odontocete echolocation insights from large datasets." DCLDE 2018 Official Program, June.

Freeman, L. and Freeman, S. (**2016**). "Rapidly obtained ecosystem indicators from coral reef soundscapes." Marine Ecology Progress Series, **561**.

Friedman, J.H. (**2001**). "Greedy function approximation: a gradient boosting machine." Annals of statistics, pages 1189–1232.

Frisk, G.V. (**2012**). "Noiseonomics: The relationship between ambient noise levels in the sea and global economic trends." Scientific Reports.

Fujioka, E., Soldevilla, M.S., Read, A.J., and Halpin, P.N. (**2014**). "Integration of passive acoustic monitoring data into obis-seamap, a global biogeographic database, to advance spatially-explicit ecological assessments." Ecological Informatics, **21**, 59–73. URL http://www.sciencedirect.com/science/article/pii/S1574954113001258. Ecological Acoustics.

Fukuzawa, Y., Webb, W.H., Pawley, M.D., Roper, M.M., Marsland, S., Brunton, D.H., and Gilman, A. (**2020**). "Koe: Web-based software to classify acoustic units and analyse sequence structure in animal vocalizations." Methods in Ecology and Evolution, **11**, 431–441. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13336.

Gaetz, W., Jantzen, K., Weinberg, H., Spong, P., and Symonds, H. (**1993**). "A neural network method for recognition of individual orcinus orca based on their acoustic behaviour: phase 1." In *Proceedings of OCEANS '93*. pp. I455–I457 vol.1.

Garland, E.C., Castellote, M., and Berchok, C.L. (**2015**). "Beluga whale (delphinapterus leucas) vocalizations and call classification from the eastern beaufort sea population." The Journal of the Acoustical Society of America, **137**, 3054–3067. URL https://doi.org/10.1121/1.4919338.

Gemba, K., Nannuru, S., and Gerstoft, P. (**2019**). "Robust ocean acoustic localization with sparse bayesian learning." IEEE Journal of Selected Topics in Signal Processing, **13**, 49–60.

Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., and Ritter, M. (**2017**). "Audio set: An ontology and human-labeled dataset for audio events." In *Proc. IEEE ICASSP 2017*.

Gendriz, I. and Padovese, L.R. (**2016**). "Underwater soundscape of marine protected areas in the south brazilian coast." Marine pollution bulletin, **105**.

Gillespie, D., Mellinger, D., Gordon, J., Mclaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X.Y., and Thode, A. (**2009**). "Pamguard: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans." The Journal of the Acoustical Society of America, **125**, 2547.

Glorot, X., Bordes, A., and Bengio, Y. (**2011**). "Domain adaptation for large-scale sentiment classification: A deep learning approach."

Goodfellow, I.J., Bengio, Y., and Courville, A. (**2016**). *Deep Learning*.

Gordon, J., Gillespie, D., Potter, J., Frantzis, A., Simmonds, M., Swift, R., and Thompson, D. (**2003**). "A review of the effects of seismic surveys on marine mammals." Marine Technology Society Journal, **37**, 16–34.

Guilment, T., Socheleau, F.X., Pastor, D., and Vallez, S. (**2018**). "Sparse representation-based classification of mysticete calls." The Journal of the Acoustical Society of America, **144**, 1550–1563. URL http://asa.scitation.org/doi/10.1121/1.5055209.

Hakkani-Tür, D., Riccardi, G., and Gorin, A. (**2002**). "Active learning for automatic speech recognition." In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. vol. 4, pp. IV–3904–IV–3907.

Hantke, S., Marchi, E., and Schuller, B. (**2016**). "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (European Language Resources Association (ELRA), Portorož, Slovenia), pp. 2156–2161. URL https://www.aclweb.org/anthology/L16-1342.

Harper, M. (**2015**). "The automatic speech recogition in reverberant environments (aspire) challenge."

Harris, P., Sotirakopoulos, K., Robinson, S., Wang, L., and Livina, V. (**2019**). "A statistical method for the evaluation of long term trends in underwater noise measurements." The Journal of the Acoustical Society of America, **145**, 228–242.

Harris, S.A., Shears, N.T., and Radford, C.A. (**2016**). "Ecoacoustic indices as proxies for biodiversity on temperate reefs." Methods in Ecology and Evolution, **7**, 713–724. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12527.

Harvey, M. (**2018**). "Acoustic detection of humpback whales using a convolutional neural network." Tech. rep. URL ai.googleblog.com/2018/10/acoustic-detection-of-humpback-whales.html(2018).

Hatch, L., Clark, C., Van Parijs, S., Frankel, A., and Ponirakis, D. (**2012**). "Quantifying loss of acoustic communication space for right whales in and around a u.s. national marine sanctuary." Conservation biology : the journal of the Society for Conservation Biology, **26**.

Haver, S.M., Fournet, M.E.H., Dziak, R.P., Gabriele, C., Gedamke, J., Hatch, L.T., Haxel, J., Heppell, S.A., McKenna, M.F., Mellinger, D.K., and Van Parijs, S.M. (**2019**). "Comparing the underwater soundscapes of four u.s. national parks and marine sanctuaries." Frontiers in Marine Science, **6**, 500.

Haver, S.M., Klinck, H., Nieukirk, S.L., Matsumoto, H., Dziak, R.P., and Miksis-Olds, J.L. (**2017**). "The not-so-silent world: Measuring arctic, equatorial, and antarctic soundscapes in the atlantic ocean." Deep–Sea Research I.

Hawkins, R.S., Miksis, J.L., and Smith, C.M. (**2014**). "Variation in low-frequency estimates of sound levels based on different units of analysis." J. Acoust. Soc. Am. 135 (2).

He, K., Zhang, X., Ren, S., and Sun, J. (**2015**). "Deep residual learning for image recognition." CoRR, **abs/1512.03385**. URL http://arxiv.org/abs/1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778.

Heaton, J.B., Polson, N.G., and Witte, J.H. (**2017**). "Deep learning for finance: deep portfolios." Applied Stochastic Models in Business and Industry, **33**, 3–12.

Heenehan, H., Stanistreet, J.E., Corkeron, P.J., Bouveret, L., Chalifour, J., Davis, G.E., Henriquez, A., Kiszka, J.J., Kline, L., Reed, C., Shamir-Reynoso, O., Védie, F., De Wolf, W., Hoetjes, P., and Van Parijs, S.M. (**2019**). "Caribbean sea soundscapes: Monitoring humpback whales, biological sounds, geological events, and anthropogenic impacts of vessel noise." Frontiers in Marine Science, **6**, 347. URL https://www.frontiersin.org/article/10.3389/fmars.2019.00347.

Hendrycks, D., Mazeika, M., Wilson, D., and Gimpel, K. (**2019**). "Using trusted data to train deep networks on labels corrupted by severe noise."

Hermannsen, L., Beedholm, K., Tougaard, J., and Madsen, P. (**2014**). "High frequency components of ship noise in shallow water with a discussion of implications for harbor porpoises (phocoena phocoena)." The Journal of the Acoustical Society of America, **136**, 1640.

Hey, T. (**2014**). "Beyond open access to open data." URL http://hdl.handle.net/2142/47423.

Hildebrand, J., spain, G., Roch, M., and Porter, M. (**2010**). "Glider-based passive acoustic monitoring techniques in the southern california region."

Hildyard, A. (**2001**). *Endangered Wildlife and Plants of the World*, vol. 12.

Howe, B.M., Arbic, B.K., Aucan, J., Barnes, C.R., Bayliff, N., Becker, N., Butler, R., Doyle, L., Elipot, S., Johnson, G.C., Landerer, F., Lentz, S., Luther, D.S., MÃijller, M., Mariano, J., Panayotou, K., Rowe, C., Ota, H., Song, Y.T., Thomas, M., Thomas, P.N., Thompson, P., Tilmann, F., Weber, T., and Weinstein, S. (**2019**a). "Smart cables for observing the global ocean: Science and implementation." Frontiers in Marine Science, **6**, 424. URL https://www.frontiersin.org/article/10.3389/fmars.2019.00424.

Howe, B.M., Miksis-Olds, J., Rehm, E., Sagen, H., Worcester, P.F., and Haralabus, G. (**2019**b). "Observing the oceans acoustically." Frontiers in Marine Science, **6**, 426. URL https://www.frontiersin.org/article/10.3389/fmars.2019.00426.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K.Q. (**2017**). "Densely connected convolutional networks." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2261–2269.

Ilhan, H. and Amasyali, M. (**2014**). "Active learning as a way of increasing accuracy." International Journal of Computer Theory and Engineering, **6**, 460–465.

ISO (**1975**). "Iso 266-1975 (e): Acoustics–preferred frequencies for measurements."

Ivie, P. and Thain, D. (**2018**). "Reproducibility in scientific computing." ACM Comput. Surv., **51**. URL https://doi.org/10.1145/3186266.

Jarvis, S., DiMarzio, N., Morrissey, R., and Moretti, D. (**2008**). "A novel multi-class support vector machine classifier for automated classification of beaked whales and other small odontocetes." Canadian Acoustics, **36**, 34–40. URL https://jcaa.caa-aca.ca/index.php/jcaa/article/view/1988.

Johnson, J. and Khoshgoftaar, T. (**2019**). "Survey on deep learning with class imbalance." Journal of Big Data, **6**, 27.

Kahl, S., Wilhelm-Stein, T., Klinck, H., Kowerko, D., and Eibl, M. (**2018**). "Recognizing birds from sound - the 2018 birdclef baseline system." arXiv preprint arXiv:1804.07177.

Kairam, S. and Heer, J. (**2016**). "Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks." In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (Association for Computing Machinery, New York, NY, USA), CSCW '16, pp. 1637–1648. ISBN 9781450335928. URL https://doi.org/10.1145/2818048.2820016.

Kaplan, M.B., Mooney, A., Partan, J., and Solow, A.R. (**2015**). "Coral reef species assemblages are associated with ambient soundscapes." Marine Ecology Progress Series, **533**.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.Y. (**2017**). "Lightgbm: A highly efficient gradient boosting decision tree." In *Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 3149-3157.*

Kholghi, M., Phillips, Y., Towsey, M., Sitbon, L., and Roe, P. (**2018**). "Active learning for classifying long-duration audio recordings of the environment." Methods in Ecology and Evolution, **9**, 1948–1958. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13042.

Kinda, B.G., Le Courtois, F., and Stéphan, Y. (**2017**). "Ambient noise dynamics in a heavy shipping area." Marine Pollution Bulletin, **124**, 535 – 546. URL http://www.sciencedirect.com/science/article/pii/S0025326X17306124.

Kinda, B.G., Simard, Y., Gervaise, C., Mars, J.I., and Fortier, L. (**2013**). "Under-ice ambient noise in eastern beaufort sea, canadian arctic, and its relation to environmental forcing." The Journal of the Acoustical Society of America, **134**, 77–87.

Kindermann, L., Boebel, O., Bornemann, H., Burkhardt, E., Klinck, H., Opzeeland, I.and Plotz, J., and Seibert, A.M. (**2008**). "A perennial acoustic observatory in the antarctic ocean, computational bioacoustics for assessing biodiversity." Proceedings of the international expert meeting on IT-based detection of bioacoustical patterns.

Kinoshita, K., Delcroix, M., Gannot, S., Habets, E.A.P., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., and Yoshioka, T. (**2017**). *The REVERB Challenge: A Benchmark Task for Reverberation-Robust ASR Techniques*

(Springer International Publishing, Cham), pp. 345–354. ISBN 978-3-319-64680-0. URL https://doi.org/10.1007/978-3-319-64680-0_15.

Kirsebom, O.S., Frazao, F., Simard, Y., Roy, N., Matwin, S., and Giard, S. (**2020**). "Performance of a deep neural network at detecting north atlantic right whale upcalls." The Journal of the Acoustical Society of America, **147**, 2636–2646. URL https://doi.org/10.1121/10.0001132.

Klein, E. (**1968**). "Underwater sound and naval acoustical research and applications before 1939." The Journal of the Acoustical Society of America, **43**, 931–947.

Koch, G., Zemel, R., and Salakhutdinov, R. (**2015**). "Siamese neural networks for one-shot image recognition." In *ICML deep learning workshop*. vol. 2.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (**2012**). "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems* (Curran Associates Inc., Red Hook, NY, USA), NIPS'12, pp. 1097–1105.

Küsel, E.T., Mellinger, D.K., Thomas, L., Marques, T.A., Moretti, D., and Ward, J. (**2011**). "Cetacean population density estimation from single fixed sensors using passive acoustics." The Journal of the Acoustical Society of America, **129**, 3610–3622.

Lasky, M. (**1977**). "Review of undersea acoustics to 1950." The Journal of the Acoustical Society of America, **61**, 283–297.

Lee, W., Huang, C., Chang, C., Wu, M., Chuang, K., Yang, P., and Hsieh, C. (**2018**). "Effective quality assurance for data labels through crowdsourcing and domain expert collaboration." In *Advances in Database Technology - EDBT 2018*, edited by M. Bohlen, R. Pichler, N. May, E. Rahm, S.H. Wu, and K. Hose (OpenProceedings.org), Advances in Database Technology - EDBT, pp. 646–649. 21st International Conference on Extending Database Technology, EDBT 2018 ; Conference date: 26-03-2018 Through 29-03-2018.

Leroy, E. (**2017**). "Surveillance acoustique des baleines bleues antarctique dans l'océan indien austral : traitement, analyse et interprétation." Ph.D. thesis.

Leroy, E.e.a. (**2018**). "On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions." J. Acoust. Soc. Am., **144**, 740–754.

Lewis, D.D. and Catlett, J. (**1994**). "Heterogeneous uncertainty sampling for supervised learning." In *Machine Learning Proceedings 1994*, edited by W.W. Cohen and H. Hirsh (Morgan Kaufmann, San Francisco (CA)), pp. 148–156. ISBN 978-1-55860-335-6. URL http://www.sciencedirect.com/science/article/pii/B978155860335650026X.

Li, P. (**2012**). "Robust logitboost and adaptive base class."

Li, P., Liua, X., Palmer, K., Fleishman, E., Gillespie, D., Nosal, E.M., Shiu, Y., Klinck, H., Cholewiak, D., Helble, T., and Roch, M. (**2020**). "Learning deep models from synthetic data for extracting dolphin whistle contours."

Lillis, A., Caruso, F., Mooney, T.A., Llopiz, J., Bohnenstiehl, D.R., and Eggleston, D.B. (**2018**). "Drifting hydrophones as an ecologically meaningful approach to underwater soundscape measurement in coastal benthic habitats." Journal of Ecoacoustics, **2**. URL https://www.veruscript.com/a/STBDH1/.

Lin, T.H., Fang, S.H., and Tsao, Y. (**2017**a). "Improving biodiversity assessment via unsupervised separation of biological sounds from long-duration recordings." Scientific Reports, **7**, 4547.

Lin, T.H. and Tsao, Y. (**2019**). "Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval." Remote Sensing in Ecology and Conservation.

Lin, T.H., Tsao, Y., Wang, Y.H., Yen, H.W., and Lu, S.S. (**2017**b). "Computing biodiversity change via a soundscape monitoring network." In *PNC 2017 Annual Conference and Joint Meetings*. vol. abs/1708.02002. URL http://arxiv.org/abs/1708.02002.

Lindseth, A. and Lobel, P. (**2018**). "Underwater soundscape monitoring and fish bioacoustics: A review." Fishes, **3**, 36.

Lotfian, R. and Busso, C. (**2019**). "Curriculum learning for speech emotion recognition from crowdsourced labels." IEEE/ACM Transactions on Audio, Speech, and Language Processing, **27**, 815–826.

Lowndes, J., Best, B., and Scarborough, C.e.a. (**2017**). "Our path to better science in less time using open data science tools." Nat Ecol Evol 1, 0160.

Mann, D., Hill-Cook, M., Manire, C., Greenhow, D., Montie, E., Powell, J., Wells, R., Bauer, G., Cunningham-Smith, P., Lingenfelser, R., DiGiovanni, Jr, R., Stone, A., Brodsky, M., Stevens, R., Kieffer, G., and Hoetjes, P. (**2010**). "Hearing loss in stranded odontocete dolphins and whales." PLoS ONE, **5**, 1–5. URL https://doi.org/10.1371/journal.pone.0013824.

Manocha, P., Badlani, R., Kumar, A., Shah, A., Elizalde, B., and Raj, B. (**2018**). "Content-based representations of audio using siamese neural networks." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 3136–3140.

Marley, S.A., Kent, C.P.S., Erbe, C., and Parnum, I.M. (**2017**). "Effects of vessel traffic and underwater noise on the movement, behaviour and vocalisations of bottlenose dolphins in an urbanised estuary." Scientific Reports, **7**, 13437. URL https://doi.org/10.1038/s41598-017-13252-z.

Marques, T.A.e.a. (**2013**). "Estimating animal population density using passive acoustics." Biol. Rev., **88**, 287–309. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12001.

Mathias, D., Gervaise, C., and Iorio, L.D. (**2016**). "Wind dependence of ambient noise in a biologically rich coastal area." The Journal of the Acoustical Society of America, **139**, 839–850. URL https://doi.org/10.1121/1.4941917.

Mazhar, S., Ura, T., and Bahl, R. (**2007**). "Vocalization based individual classification of humpback whales using support vector machine." In *OCEANS 2007*. pp. 1–9.

McDonald, M.A., Hildebrand, J.A., and Wiggins, S.M. (**2006**). "Increases in deep ocean ambient noise in the northeast pacific west of san nicolas island, california." The Journal of the Acoustical Society of America, **120**, 711–718.

McKenna, M.F., Katz, S.L., Wiggins, S.M., Ross, D., and Hildebrand, J.A. (**2012**). "A quieting ocean: Unintended consequence of a fluctuating economy." The Journal of the Acoustical Society of America, **132**, EL169–EL175.

McPherson, C., Martin, B., MacDonnell, J., and Whitt, C. (**2016**). "Examining the value of the acoustic variability index in the characterisation of australian marine soundscapes."

Melekhov, I., Kannala, J., and Rahtu, E. (**2016**). "Siamese network features for image matching." In *2016 23rd International Conference on Pattern Recognition, ICPR 2016* (IEEE, United States), pp. 378–383. International Conference on Pattern Recognition, ICPR ; Conference date: 04-12-2016 Through 08-12-2016.

Mellinger, D. and Clark, C. (**2006**). "Mobysound: A reference archive for studying automatic recognition of marine mammal sounds." Applied Acoustics, **67**, 1226–1242.

Mellinger, D.K. (**2002**). "Ishmael : 1.0 user's guide ; ishmael : integrated system for holistic multi-channel acoustic exploration and localization." In *NOAA technical memorandum OAR PMEL 120*.

Mellinger, D.K. and Clark, C.W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation." J. Acoust. Soc. Am., **107**, 3518–3529.

Mercado, E., Schneider, J., Pack, A.A., and Herman, L.M. (**2010**). "Sound production by singing humpback whales." J. Acoust. Soc. Am., **127**, 2678–2691.

Merchan, F., Guerra, A., Poveda, H., GuzmÃąn, H., and Sanchez Galan F, J. (**2020**). "Bioacoustic classification of antillean manatee vocalization spectrograms using deep convolutional neural networks." Applied Sciences, **10**, 3286.

Merchant, N., Blondel, P., Dakin, T., and Dorocicz, J. (**2012**). "Averaging underwater noise levels for environmental assessment of shipping." The Journal of the Acoustical Society of America, **132**, EL343–EL349.

Merchant, N., Brookes, K., Faulkner, R., Bicknell, A., Godley, B., and Witt, M. (**2016**). "Underwater noise levels in uk waters." Scientific Reports, **6**.

Merchant, N.D., Fristrup, K.M., Johnson, M.P., Tyack, P.L., Witt, M.J., Blondel, P., and Parks, S.E. (**2015**). "Measuring acoustic habitats." Methods in Ecology and Evolution, **6**, 257–265.

Merchant, N.D., Pirotta, E., Barton, T.R., and Thompson, P.M. (**2014**). "Monitoring ship noise to assess the impact of coastal developments on marine mammals." Marine Pollution Bulletin, **78**, 85–95.

Mesaros, A., Heittola, T., and Virtanen, T. (**2016**). "Tut database for acoustic scene classification and sound event detection." In *2016 24th European Signal Processing Conference (EUSIPCO)*. pp. 1128–1132.

Miksis-Olds, J.L., Martin, B., and Tyack, P.L. (**2018**). "Exploring the ocean through soundscapes." Acoustics Today, **14**, 26–34.

Mooney, T.A., Di Iorio, L., Lammers, M., Lin, T.H., Nedelec, S.L., Parsons, M., Radford, C., Urban, E., and Stanley, J. (**2020**). "Listening forward: approaching marine biodiversity assessments using acoustic methods." Royal Society Open Science, **7**, 201287. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsos.201287.

Muir, T.G. and Bradley, D.L. (**2016**). "Underwater acoustics: A brief historical overview through world war ii." Acoustics today, **12**.

Munger, L., Mellinger, D., Wiggins, S., Moore, S., and Hildebrand, J. (**2005**). "The performance of spectogram correlation in detecting right whale calls in long-term recordings from the bering sea." Canadian Acoustics, **33**.

Murray, S.O., Mercado, E., and Roitblat, H. (**1998**). "The neural network classification of false killer whale (pseudorca crassidens) vocalizations." The Journal of the Acoustical Society of America, **104 6**, 3626–33.

Mydlarz, C., Salamon, J., and Bello, J.P. (**2017**). "The implementation of low-cost urban acoustic monitoring devices." Applied Acoustics, **117**, 207–218. URL http://www.sciencedirect.com/science/article/pii/S0003682X1630158X. Acoustics in Smart Cities.

Nanaware, S., Shastri, R., Joshi, Y., and Das, A. (**2014**). "Passive acoustic detection and classification of marine mammal vocalizations." pp. 493–497.

Nanni, L., Costa, Y., Aguiar, R., Mangolin, R., Brahnam, S., and Silla, C. (**2020**). "Ensemble of convolutional neural networks to improve animal audio classification." EURASIP Journal on Audio, Speech, and Music Processing, **2020**.

Ness, S. (**2013**). "The orchive: A system for semi-automatic annotation and analysis of a large collection of bioacoustic recordings." Ph.D. thesis.

Niu, H., Gerstoft, P., and Ozanich, E. (**2017**). "Source localization in an ocean waveguide using supervised machine learning." The Journal of the Acoustical Society of America, **142**.

NRC (**2003**). *Ocean Noise and Marine Mammals* (Committee on Potential Impacts of Ambient Noise in the Ocean on Marine Mammals, National Research Council).

Nystuen, J.A. and Selsor, H.D. (**1997**). "Weather classification using passive acoustic drifters." Journal of Atmospheric and Oceanic Technology, **14**, 656–666. URL https://doi.org/10.1175/1520-0426(1997)014<0656:WCUPAD>2.0.CO;2.

Ockendon, N., Baker, D.J., Carr, J.A., White, E.C., Almond, R.E.A., Amano, T., Bertram, E., Bradbury, R.B., Bradley, C., Butchart, S.H.M., Doswald, N., Foden, W., Gill, D.J.C., Green, R.E., Sutherland, W.J., Tanner, E.V.J., and Pearce-Higgins, J.W. (**2014**). "Mechanisms underpinning climatic impacts on natural populations: altered species interactions are more important than direct effects." Global Change Biology, **20**, 2221–2229. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.12559.

Pace, F., Benard, F., Glotin, H., Adam, O., and White, P. (**2010**). "Subunit definition and analysis for humpback whale call classification." Applied Acoustics, **71**, 1107–1112. URL http://www.sciencedirect.com/science/article/pii/S0003682X10001234. Proceedings of the 4th International Workshop on Detection, Classification and Localization of Marine Mammals Using Passive Acoustics and 1st International Workshop on Density Estimation of Marine Mammals Using Passive Acoustics.

Pan, S.J. and Yang, Q. (**2010**). "A survey on transfer learning." IEEE Transactions on Knowledge and Data Engineering, **22**, 1345–1359.

Patris, J., Malige, F., Glotin, H., Asch, M., and Buchan, S.J. (**2019**). "A standardized method of classifying pulsed sounds and its application to pulse rate measurement of blue whale southeast pacific song units." The Journal of the Acoustical Society of America, **146**, 2145–2154.

Payne, R. and McVay, S. (**1971**). "Songs of humpback whales." Science, **173**, 585–597.

Pensieri, S., Bozzano, R., Nystuen, J., Anagnostou, E., Anagnostou, M., and Bechini, R. (**2015**). "Research article underwater acoustic measurements to estimate wind and rainfall in the mediterranean sea." Advances in Meteorology, **15**, 1–18.

Phillips, Y.F., Towsey, M., and Roe, P. (**2018**). "Revealing the ecological content of long-duration audio-recordings of the environment through clustering and visualisation." PLoS ONE, **13**, 1–27. URL https://doi.org/10.1371/journal.pone.0193345.

Picciulin, M., Colla, S., Pranovi, F., Malavasi, S., Fiorin, R., and Bolgan, M. (**2016**). "The soundscape of a mussel farm: Biophony and man-made noise levels." Proceedings of Meetings on Acoustics, **27**, 010016. URL https://asa.scitation.org/doi/abs/10.1121/2.0000268.

Piczak, K.J. (**2015**). "Esc: Dataset for environmental sound classification, 23rd acm international conference on multimedia, brisbane, australia, oct. 2015, pp. 1015–1018." URL https://doi.org/10.1145/2733373.2806390.

Pieretti, N., Farina, A., and Morri, D. (**2011**). "A new methodology to infer the singing activity of an avian community: The acoustic complexity index (aci)." Ecological Indicators, **11**, 868–873. URL http://www.sciencedirect.com/science/article/pii/S1470160X10002037.

Pierretti, N., Martire, M.L., Farina, A., and Danovaro, R. (**2017**). "Marine soundscape as an additional biodiversity monitoring tool: A case study from the adriatic sea (mediterranean sea)." Ecological Indicators, **83**, 13–20.

Pijanowski, B.C., Farina, A., Gage, S.H., Dumyahn, S.L., and Krause, B.L. (**2011**). "What is soundscape ecology? an introduction and overview of an emerging new science." Landscape Ecology, **26**, 1213–1232. URL https://doi.org/10.1007/s10980-011-9600-8.

Pine, M.K., Radford, C.A., and Jeffs, A.G. (**2015**). "Eavesdropping on the kaipara harbour: characterising underwater soundscapes within a seagrass bed and a subtidal mudflat." Journal of Marine and Freshwater Research.

Porter, M.B. (**2011**). "The bellhop manual and user's guide: Preliminary draft." Heat, Light, and Sound Research, Inc., La Jolla, CA, USA, Tech. Rep, **260**.

Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A.W. (**2003**). "Recent advances in the automatic recognition of audiovisual speech." Proceedings of the IEEE, **91**, 1306–1326.

Poupard, M., Best, P., Schluter, J., Prevot, J.M., Symonds, H., Spong, P., and Glotin, H. (**2019**). "Deep learning for ethoacoustics of oreas on three years pentaphonie continuous recording at orcalab revealing tide, moon and diel effects." pp. 1–7.

Putland, R.L., Constantine, R., and Radford, C.A. (**2017**). "Exploring spatial and temporal trends in the soundscape of an ecologically significant embayment." Scientific Reports, **7**, 5713. URL https://doi.org/10.1038/s41598-017-06347-0.

Qian, K., Zhang, Z., Baird, A., and Schuller, B. (**2017**). "Active learning for bird sound classification via a kernel-based extreme learning machine." The Journal of the Acoustical Society of America, **142**, 1796–1804.

Racah, E., Beckham, C., Maharaj, T., Ebrahimi Kahou, S., Prabhat, M., and Pal, C. (**2017**). "Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc.), pp. 3402–3413. URL http://papers.nips.cc/paper/6932-extremeweather-a-large-scale-climate-dataset-for-semi-supervised-detection-localiz pdf.

Radhakrishnan, R., Divakaran, A., and Smaragdis, A. (**2005**). "Audio analysis for surveillance applications." In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.* pp. 158–161.

Rahul, M.V., Ambareesh, R., and Shobha, G. (**2017**). "Siamese network for underwater multiple object tracking." In *Proceedings of the 9th International Conference on Machine Learning and Computing* (Association for Computing Machinery, New York, NY, USA), ICMLC 2017, pp. 511–516. ISBN 9781450348171. URL https://doi.org/10.1145/3055635.3056579.

Rakotomamonjy, A. and Gasso, G. (**2014**). "Histogram of gradients of time-frequency representations for audio scene detection." Tech. Rep. 1.

Rayment, W., Webster, T., Brough, T., Jowett, T., and Dawson, S. (**2017**). "Seen or heard? a comparison of visual and acoustic autonomous monitoring methods for investigating temporal variation in occurrence of southern right whales." Marine Biology, **165**, 1–10.

Reeves, E., Gerstoft, P., Worcester, P.F., and Dzieciuch, M. (**2017**). "Arctic soundscape measured with a drifting vertical line array." The Journal of the Acoustical Society of America, **141**, 3524–3524. URL https://doi.org/10.1121/1.4987426.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat (**2019**). "Deep learning and process understanding for data-driven earth system science." Nature, **566**, 195–204.

Riccardi, G. and Hakkani-Tur, D. (**2005**). "Active learning: theory and applications to automatic speech recognition." IEEE Transactions on Speech and Audio Processing, **13**, 504–511.

Richardson, W.J., Greene, C.R., Malme, C.I., and Thomson, D.H. (**1995**). *Marine Mammals and Noise* (Greeneridge Sciences Inc., Editor(s): W. John Richardson, Charles R. Greene, Charles I. Malme, Denis H. Thomson, , Academic Press), chap. ACOUSTIC CONCEPTS AND TERMINOLOGY, pp. 15–32.

Riera, A., Ford, J., and Chapman, R. (**2013**). "Effects of different analysis techniques and recording duty cycles on passive acoustic monitoring of killer whales." The Journal of the Acoustical Society of America, **134**, 2393–404.

Robinson, S.P., Lepper, P.A., and Hazelwood, R.A. (**2014**). "Good practice guide for underwater noise measurement." Tech. Rep. Guide No. 133: 95 pp., National Measurement Office, Marine Scotland, The Crown Estate, NPL Good Practice.

Roca, I. and Van Opzeeland, I. (**2019**). "Using acoustic metrics to characterize underwater acoustic biodiversity in the southern ocean." Remote Sensing in Ecology and Conservation.

Roch, M.A., Soldevilla, M.S., Hoenigman, R., Wiggins, S.M., and Hildebrand, J.A. (**2008**). "Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes." Canadian Acoustic, **36**, 41–47.

Rogers, T.L. (**2003**). "Factors influencing the acoustic behaviour of male phocid seals." Aquatic Mammals.

Roma, G., Janer, J., and Herrera, P. (**2012**). "Active learning of custom sound taxonomies in unstructured audio data." In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval* (Association for Computing Machinery, New York, NY, USA), ICMR '12. ISBN 9781450313292. URL https://doi.org/10.1145/2324796.2324872.

Romagosa, M., Cascao, I., Merchant, N.D., Lammers, M.O., Giacomello, E., Marques, T.A., and Silva, M.A. (**2017**). "Underwater ambient noise in a baleen whale migratory habitat off the azores." Frontiers in Marine Science, **4**, 109. URL https://www.frontiersin.org/article/10.3389/fmars.2017.00109.

Roman, J. and McCarthy, J.J. (**2010**). "The whale pump: marine mammals enhance primary productivity in a coastal basin." PLoS One.

Royer, J.Y. (**2009**). "Oha-sis-bio - observatoire hydroacoustique."

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L. (**2015**). "Imagenet large scale visual recognition challenge." Int. J. Comput. Vision, **115**, 211–252. URL https://doi.org/10.1007/s11263-015-0816-y.

Salamon, J. and Bello, J.P. (**2017**). "Deep convolutional neural networks and data augmentation for environmental sound classification." IEEE Signal Processing Letters, **24**, 279–283. URL http://dx.doi.org/10.1109/LSP.2017.2657381.

Samuel-Rhoads, Y., Morreale, S., Clark, C., Greene, C., and Richmond, M. (**2005**). "Underwater, low-frequency noise in a coastal sea turtle habitat." The Journal of the Acoustical Society of America, **117**, 1465–72.

Sayigh, L., Daher, M.A., Allen, J., Gordon, H., Joyce, K., Stuhlmann, C., and Tyack, P. (**2016**). "The watkins marine mammal sound database: An online, freely accessible resource." Proceedings of Meetings on Acoustics, **27**, 040013. URL https://asa.scitation.org/doi/abs/10.1121/2.0000358.

Schaar, M.V.D., Delory, E., Català, A., and André, M. (**2007**). "Neural network-based sperm whale click classification." Journal of the Marine Biological Association of the United Kingdom, **87**, 35–38.

Schaetti, N. (**2018**). "Character-based convolutional neural network andresnet18 for twitter author profiling." In *Notebook for PAN at CLEF 2018*.

Serizel, R., Turpault, N., Shah, A., and Salamon, J. (**2020**). "Sound event detection in synthetic domestic environments." In *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing* (Barcelona, Spain). URL https://hal.inria.fr/hal-02355573.

Settles, B. (**2012**). *Active Learning Synthesis Lectures on Artificial Intelligence and Machine Learning* (Morgan and Claypool).

Shabangu, F. and Charif, R.A. (**2020**). "Short moan call reveals seasonal occurrence and diel calling pattern of crabeater seals in the weddell sea antarctica." Bioacoustics.

Shabangu, F., Yemane, D., Stafford, K., Ensor, P., and Findlay, K. (**2017**). "Modelling the effects of environmental conditions on the acoustic occurrence and behaviour of antarctic blue whales." PLoS ONE, **12**.

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, S., Tyack, P., Samarra, F., Miller, P., and Wallin, J. (**2014**). "Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls." The Journal of the Acoustical Society of America, **135**, 953–962.

Shi, X., Dou, Q., Xue, C., Qin, J., Chen, H., and Heng, P.A. (**2019**). *An Active Learning Approach for Reducing Annotation Cost in Skin Lesion Analysis*. URL http://arxiv.org/abs/1909.02344.

Shiu, Y., Palmer, K., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (**2020**). "Deep neural networks for automated detection of marine mammal species." Scientific Reports, **10**, 1–12.

Shuyang, Z., Heittola, T., and Virtanen, T. (**2017**). "Active learning for sound event classification by clustering unlabeled data." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 751–755.

Siddagangaiah, S., Li, Y., Guo, X., and Yang, K. (**2015**). "On the dynamics of ocean ambient noise: Two decades later." Chaos: An Interdisciplinary Journal of Nonlinear Science, **25**, 103117.

Simonyan, K. and Zisserman, A. (**2015**). "Very deep convolutional networks for large-scale image recognition." ICLR.

Sinha, V., Rao, S., and Balasubramanian, V. (**2018**). "Fast dawid-skene: A fast vote aggregation scheme for sentiment classification." arXiv: Machine Learning.

Sirovic, A. (**2016**). "Variability in the performance of the spectrogram correlation detector for northeast pacific blue whale calls." Bioacoustics, **25**, 145–160.

Smith, L.V., McMinn, A., Martin, A., Nicol, S., Bowie, A.R., Lannuzel, D., and van der Merwe, P. (**2013**). "Preliminary investigation into the stimulation of phytoplankton photophysiology and growth by whale faeces." Journal of Experimental Marine Biology and Ecology, **446**, 1–9.

Snel, J., Tarasov, A., Cullen, C., and Delany, S.J. (**2012**). "A crowdsourcing approach to labeling a mood induced speech corpora." In *4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals*.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. (**2008**). "Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, USA), EMNLP '08, pp. 254–263.

Socheleau, F.X., Leroy, E., Carvallo Pecci, A., Samaran, F., Bonnel, J., and Royer, J.Y. (**2015**). "Automated detection of antarctic blue whale calls." The Journal of the Acoustical Society of America, **138**, 3105–3117. URL http://scitation.aip.org/content/asa/journal/jasa/138/5/10.1121/1.4934271.

Socheleau, F.X. and Samaran, F. (**2018**). "Detection of Mysticete Calls : a Sparse Representation-Based Approach." Tech. rep.

Southall, B.L., Bowles, A.E., Ellison, W.T., Finneran, J.J., Gentry, R.L., Jr., C.R.G., Kastak, D., Ketten, D.R., Miller, J.H., Nachtigall, P.E., Richardson, W.J., Thomas, J.A., and Tyack, P.L. (**2008**). "Marine mammal noise-exposure criteria: Initial scientific recommendations." Bioacoustics, **17**, 273–275. URL https://doi.org/10.1080/09524622.2008.9753846.

Staaterman, E., Ogburn, M.B., Altieri, A., Brandl, S., Whippo, R., Seemann, J., Goodison, M., and Duffy, J. (**2017**). "Bioacoustic measurements complement visual biodiversity surveys: Preliminary evidence from four shallow marine habitats." Marine Ecology Progress Series, **575**.

Staaterman, E., Paris, C., DeFerrari, H.A., Mann, D., Rice, A., and D'Alessandro, E.K. (**2014**). "Celestial patterns in marine soundscapes." Marine Ecology Progress Series, **508**, 17–32.

Staaterman, E., Rice, A., Mann, D., and Paris, C. (**2013**). "Soundscapes from a tropical eastern pacific reef and a caribbean sea reef." Coral Reefs, **32**. URL http://link.springer.com/article/10.1007/s00338-012-1007-8/fulltext.html.

Stanistreet, J.E., Nowacek, D.P., Read, A.J., Baumann-Pickering, S., Moors-Murphy, H.B., and Van Parijs, S.M. (**2016**). "Effects of duty-cycled passive acoustic recordings on detecting the presence of beaked whales in the northwest atlantic." The Journal of the Acoustical Society of America, **140**, EL31–EL37.

Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., and Plumbley, M. (**2015**). "Detection and classification of acoustic scenes and events." IEEE Transactions on Multimedia, **17**, 1733–1746.

Sueur, J., Farina, A., Gasc, A., Pieretti, N., and Pavoine, S. (**2014**). "Acoustic indices for biodiversity assessment and landscape investigation." Acta Acustica united with Acustica.

Sueur, J., Pavoine, S., Hamerlynck, O., and Duvail, S. (**2009**). "Rapid acoustic survey for biodiversity appraisal." PLoS ONE, **3**, 1–9. URL https://doi.org/10.1371/journal.pone.0004065.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (**2017**). "Revisiting unreasonable effectiveness of data in deep learning era." In *2017 IEEE International Conference on Computer Vision (ICCV)*. pp. 843–852.

Sveegaard, S., Galatius, A., Dietz, R., Kyhn, L., Koblitz, J.C., Amundin, M., Nabe-Nielsen, J., Sinding, M.H.S., Andersen, L.W., and Teilmann, J. (**2015**). "Defining management units for cetaceans by combining genetics, morphology, acoustics and satellite tracking." Global Ecology and Conservation, **3**, 839–850. URL http://www.sciencedirect.com/science/article/pii/S2351989415000384.

Tasker, M.L., Amundin, M., André, M., Hawkins, T., Lang, W., Merck, T., Scholik-Schlomer, A., Teilmann, J., Thomsen, F., Werner, S., and Zakharia, M. (**2010**). "Marine strategy framework directive - task group 11 report - underwater noise and other forms of energy,." Tech. rep.

Thomas, M., Martin, B., Kowarski, K., Gaudet, B., and Matwin, S. (**2019**). "Marine mammal species classification using convolutional neural networks and a novel acoustic representation."

Thomisch, K., Boebel, O., Zitterbart, D., Samaran, F., Van Parijs, S., and Van Opzeeland, I. (**2015**). "Effects of subsampling of passive acoustic recordings on acoustic metrics." The Journal of the Acoustical Society of America, **138**, 267.

Thompson, P.M., Brookes, K.L., and Cordes, L.S. (**2014**). "Integrating passive acoustic and visual data to model spatial patterns of occurrence in coastal dolphins." ICES Journal of Marine Science, **72**, 651–660. URL https://doi.org/10.1093/icesjms/fsu110.

Thompson, P.O. (**1965**). *Marine biological sound, west of San Clemente Island : diurnal distributions and effects on ambient noise level during July 1963* (U.S. Navy Electronics Laboratory Report), 1–42 pp.

Torterotot, M., Royer, J.Y., and Samaran, F. (**2019**). "Detection strategy for long-term acoustic monitoring of blue whale stereotyped and non-stereotyped calls in the Southern Indian Ocean." In *IEEE OCEANS* (IEEE), pp. 1–10. ISBN 9781728114507.

Trowitzsch, I., Taghia, J., Kashef, Y., and Obermayer, K. (**2019**). "Nigens general sound events database." URL https://doi.org/10.5281/zenodo.2535878.

Truskinger, A., Cottman-Fields, M., Eichinski, P., Towsey, M., and Roe, P. (**2014**). "Practical analysis of big acoustic sensor data for environmental monitoring." In *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*. pp. 91–98.

Urazghildiiev, I.R. and Clark, C.W. (**2006**). "Acoustic detection of north atlantic right whale contact calls using the generalized likelihood ratio test." J. Acoust. Soc. Am. 120, 1956-1963.

——— (**2007**). "Detection performances of experienced human operators compared to a likelihood ratio based detector." The Journal of the Acoustical Society of America, **122**, 200–204. URL http://asa.scitation.org/doi/10.1121/1.2735114.

Usman, A., Ogundile, O., and Versfeld, J. (**2020**). "Review of automatic detection and classification techniques for cetacean vocalization." IEEE Access, **PP**, 1–1.

Van Trees, H.L. (**1968**). *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory* (John Wiley & Sons).

Viola, S., Grammauta, R., Sciacca, V., Bellia, G., Beranzoli, L., Buscaino, G., Caruso, F., Chierici, F., Cuttone, G., D'Amico, A., Luca, V.D., Embriaco, D., Favali, P., Giovanetti, G., Marinaro, G., Mazzola, S., Filiciotto, F., Pavan, G., Pellegrino, C., Pulvirenti, S., Simeone, F., Speziale, F., and Riccobene, G. (**2017**). "Continuous monitoring of noise levels in the gulf of catania (ionian sea). study of correlation with ship traffic." Marine Pollution Bulletin, **121**, 97–103. URL http://www.sciencedirect.com/science/article/pii/S0025326X1730423X.

Virtanen, T., Plumbley, M., and Ellis, D. (**2017**). *Computational Analysis of Sound Scenes and Events*. 1–422 pp.

Wang, Y., Yao, Q., Kwok, J.T., and Ni, L.M. (**2020**). "Generalizing from a few examples: A survey on few-shot learning." ACM Comput. Surv., **53**. URL https://doi.org/10.1145/3386252.

Watkins, W.A. (**1981**). "Activities and underwater sounds of fin whales [balaenoptera physalus]." In *Sci. Rep. Whales Res. Inst.* vol. 33, pp. 83–117.

Welch, P.D. (**1967**). "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms." Audio and Electroacoustics, IEEE Transactions on, **15**, 70–73.

Williams, R., Wright, A., Ashe, E., Blight, L., Bruintjes, R., Canessa, R., Clark, C., Cullis-Suzuki, S., Dakin, D., Erbe, C., Hammond, P., Merchant, N., O'Hara, P., Purser, J., Radford, A., Simpson, S., Thomas, L., and Wale, M. (**2015**). "Impacts of anthropogenic noise on marine life: Publication patterns, new discoveries, and future directions in research and management." Ocean and Coastal Management, **115**, 17–24. Making Marine Science Matter: Issues and Solutions from the 3rd International Marine Conservation Congress.

Wu, S., Wu, Y., Cao, D., and Zheng, C. (**2019**). "A fast button surface defect detection method based on siamese network with imbalanced samples." Multimedia Tools and Applications, **78**.

Xu, Y., Kong, Q., Huang, Q., Wang, W., and Plumbley, M.D. (**2017**). "Convolutional gated recurrent neural network incorporating spatial features for audio tagging." In *2017 International Joint Conference on Neural Networks (IJCNN)*. pp. 3461–3466.

Yang, J., Riser, S.C., Nystuen, J.A., Asher, W.E., and Jessup, A.T. (**2015**). "Regional rainfall measurements using the passive aquatic listener during the spurs field campaign." Oceanography, **28**. URL https://doi.org/10.5670/oceanog.2015.10.

Yang, L., Zhang, Y., Chen, J., Zhang, S., and Chen, D.Z. (**2017**). "Suggestive annotation: A deep active learning framework for biomedical image segmentation." CoRR, **abs/1706.04737**. URL http://arxiv.org/abs/1706.04737.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (**2014**). "How transferable are features in deep neural networks?" CoRR, **abs/1411.1792**. URL http://arxiv.org/abs/1411.1792.

Zapf, A., Castell, S., Morawietz, L., and Karch, A. (**2016**). "Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate?" BMC Medical Research Methodology, **16**.

Zhang, L., Wang, D., Bao, C., Wang, Y., and Xu, K. (**2019**). "Large-scale whale-call classification by transfer learning on multi-scale waveforms and time-frequency features." Applied Sciences, **9**, 1020.

Zhang, Y., Pardo, B., and Duan, Z. (**2018**). "Siamese style convolutional neural networks for sound search by vocal imitation." IEEE/ACM Transactions on Audio, Speech, and Language Processing, **PP**, 1–1.

Zhang, Z. and Schuller, B. (**2012**). "Active learning by sparse instance tracking and classifier." In *Confidence in Acoustic Emotion Recognition, in Proc. INTERSPEECH 2012*.

Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., and Brewer, A. (**2020**). "Beluga whale acoustic signal classification using deep learning neural network models." The Journal of the Acoustical Society of America, **147**, 1834–1841.

Zimmer, W.M.X. (**2011**). *Passive acoustic monitoring of cetaceans.* (Cambridge University Press.).