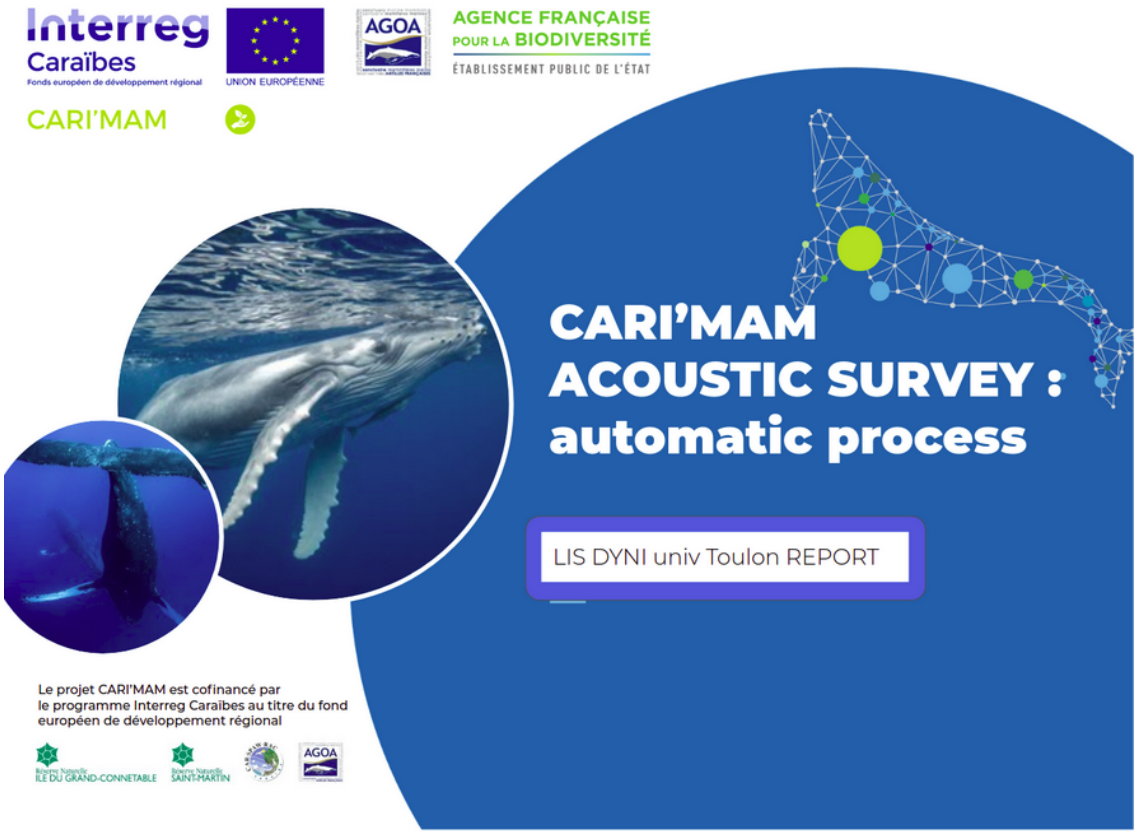
CARIMAM REPORT BIOACOUSTIC DATA PROCESSING

Hervé Glotin, Maxence Ferrari, Paul Best, Marion Poupard, Nicolas Thellier, Audrey Monsimer, Pascale Giraudet

CARIMAM REPORT

BIOACOUSTIC DATA PROCESSING



CNRS DYNI LIS univ. Toulon

Glotin Hervé
Ferrari Maxence
Best Paul
Poupard Marion
Thellier Nicolas
Monsimer Audrey
Giraudet Pascale

Funding Instrument: CARIMAM FEDER OFB
Beneficiary in Charge: DYNI, Lab. Information Systems (DYNI)

glotin@univ-tln.fr

Table of Contents

List of Figures

1 Introduction

This report summarizes the work carried out by the LIS towards the project goals and work plan of the project CARIMAM LOT2 LIS UTLN according to the axes below.

The Caribbean shows great heterogeneity in the protection of marine mammals. While some territories strictly regulate the approach of cetaceans, or carry out conservation and monitoring projects through marine protected areas, others still practice traditional hunting.

Launched in 2018, the CARI'MAM project aims to create a network of actors involved in the conservation of marine mammals and strengthen the skills of managers of marine protected areas. It would develop common management and evaluation tools on a Caribbean scale. To achieve these objectives, the structures involved share several areas of work:
Axis 1 - State of the art of legal knowledge and tools concerning marine mammals,
Axis 2 - Acquisition of knowledge,
Axis 3 - Strengthening the skills of managers,
Axis 4 -Development of common strategies for acoustic monitoring of species,
Axis 6 - Management plan for marine protected areas with a "marine mammal" responsibility,
Axis 7 - Communication and awareness.

In order to animate this network, international meetings were organized in order to allow the partners to present their structures and their projects, and to make them work collectively on common subjects through thematic workshops. COVID restrictions made difficult some work, but we succeeded due to the wish from all partners to bring CARIMAM to success.

At present, the network brings together 61 structures, coming from 28 territories of the Caribbean. This report presents the methods developed and distributed by LIS Dyni Toulon to process the 15 To of received data in october 2021, from the 20 deployed recording stations. More data are still arriving to LIS, and are continously processed for updated version of this report.

2 Data Management Plan

2.1 Real Recording Data Management Plan

The followed figures report the total recording effort at the time we write this report. Some recordings are still to be extracted from last placements in october and november, or in transfer to UTLN. So the tables will be updated in december 2021. The total effort, according to the data received by october 2021, in days, is given in the next graphics, showing an effort of recording up to 1120 days.

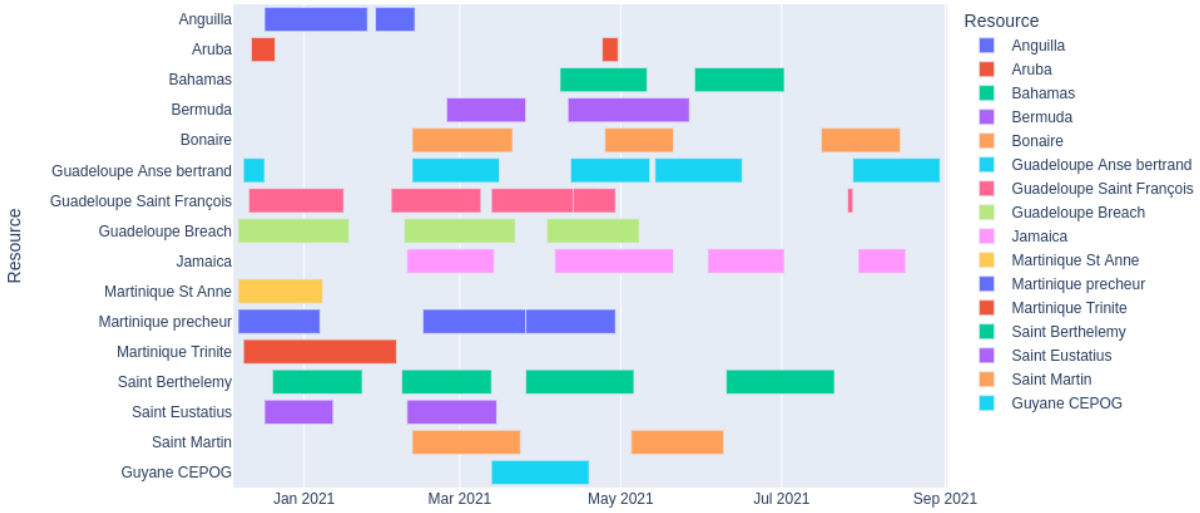


Figure 1: Summary of sessions recorded and received by the DYNI team up to the 5 october 2021.

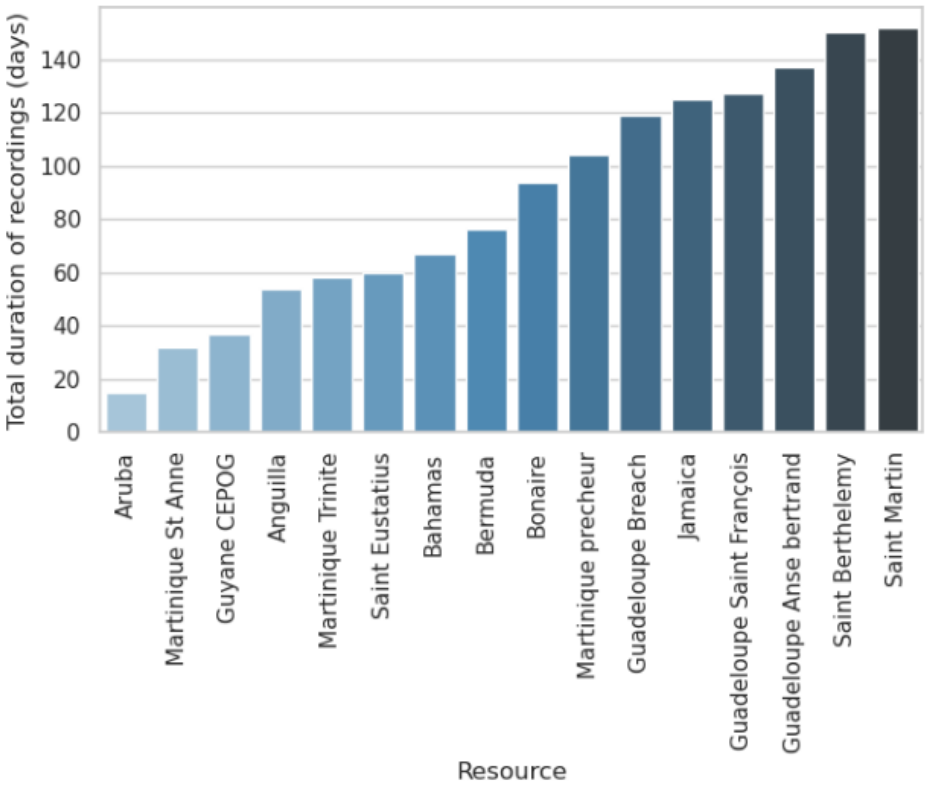


Figure 2: Total number of days recorded per session up to the 5 october 2021.

2.2 Repartition in space

We present below the recording time by station and period from Dec to March and April to October, on a spatial representation.

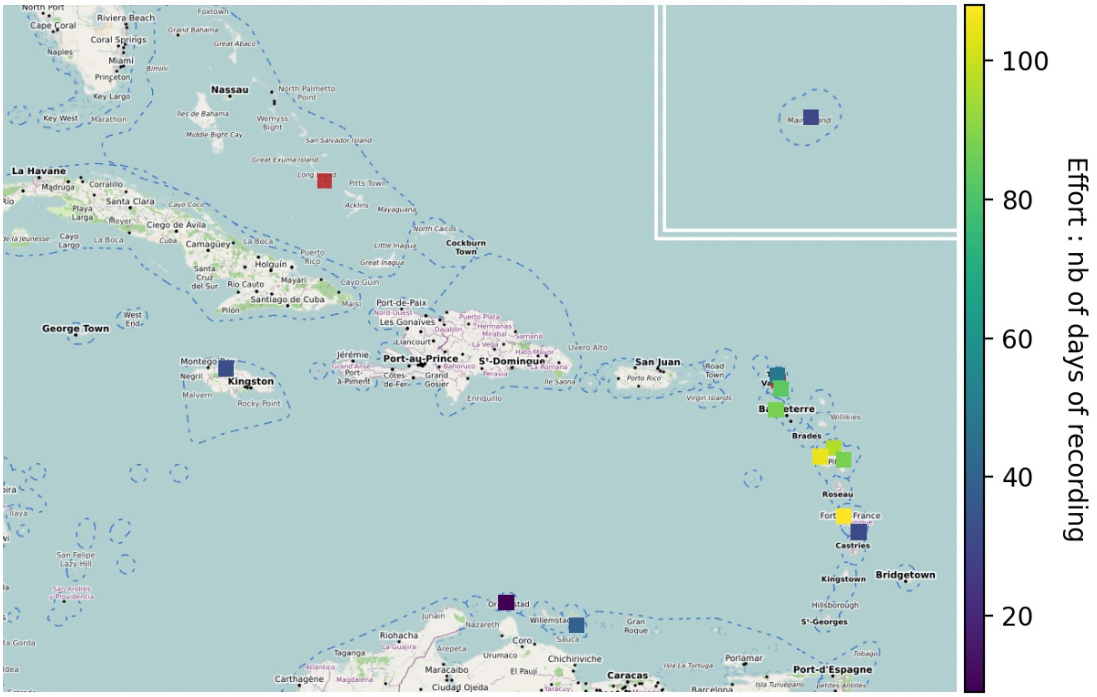


Figure 3: Effort map of the recording sessions, showing the number of days each station has been recording for the period T1: [Dec 2020 - March 2021]. Red means the station was not recording. Bermuda is at the top right.

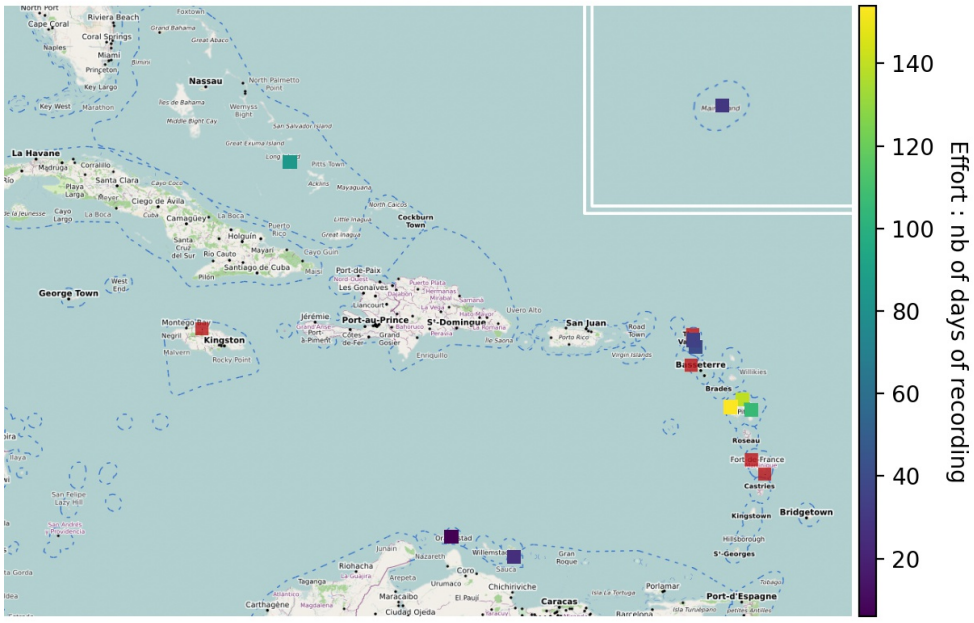


Figure 4: Effort map of the recording sessions, showing the number of days each station has been recording for the period T2: [April 2021 - October 2021]. Red means the station was not recording.

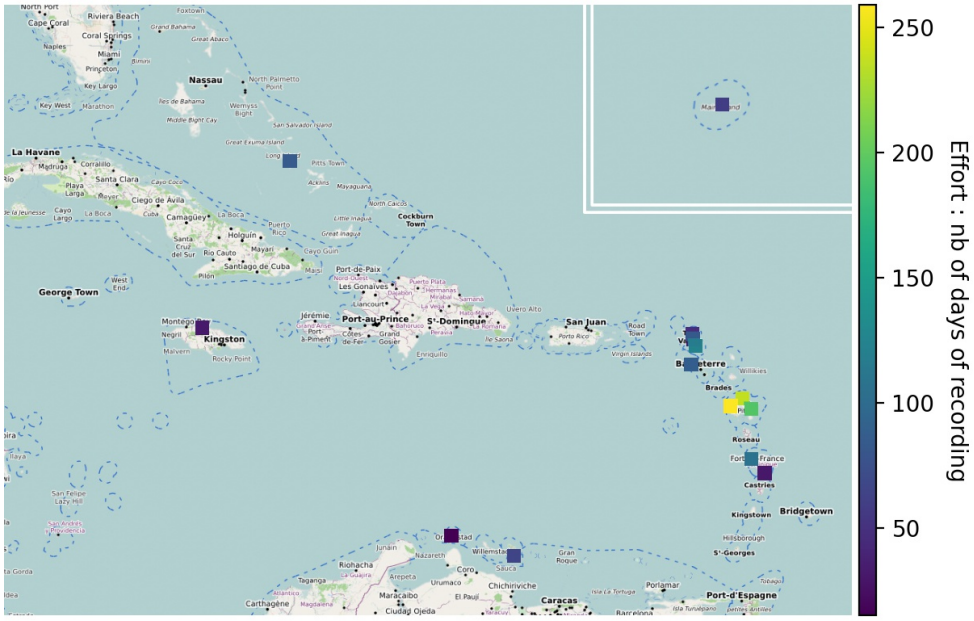


Figure 5: Total effort map of the recording sessions, showing the number of days each station has been recording from Dec 2020 to October 2021 (T1+T2).

2.3 Storage Data Management Plan

All received data are stored in a RED6 NAS and will be also copied into the NAS LIS in another site. They are restricted access sites. The data are also completely backup in each participant.

3 Work carried out by LIS

This section provides an overview of the work carried out and results achieved per **WP!** (**WP!**) up to mid October 2021 on the received data (mostly received in July, August). Some methods are still in computation, but the global framework is established.

3.1 WP01 Typical noise and Time-frequency spectrogram segmentation

Work Package Summary			
WP No.	01	*Title of WP*	Time-frequency spectrogram segmentation
Start	March	*End*	October
Participating Organisations			
• Participant: MF			

Goals
• Goal 1 : Detection of energy segment
• Goal 2 : Segmentation of patterns
• Goal 3 : Classification of patterns
• Goal 4 : Distribution of the code

Overview and Goals

The CARIMAM data are naturally noisy, due to reef proximity thus noises made by fish and shrimp. This noise cover weaker biological targets (Fig. 6).

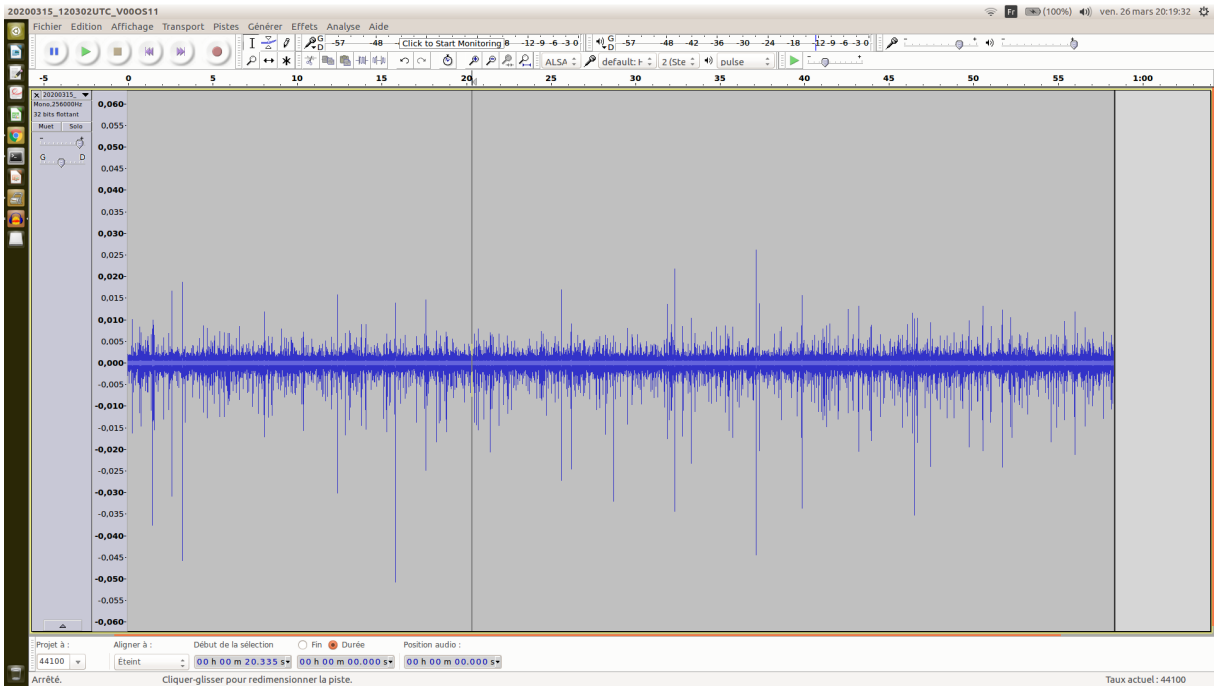


Figure 6: Example of waveform in Guadeloupe Breach of reef noise plus possible biosonar

Other sources of noises are some electronic noise arising mostly when the powering of the station becomes weak after several weeks, and other noise from the mooring line can be present. One issue with the electronic noise is sometimes its resemblance with biologic whistles, both in shape and in frequencies, overlapping with actual dolphin whistle.

Thus, this work package aims to enhance the biological whistles hidden behind clicks, stationary and narrow band noises.

Status

The two first task of this work package are done. The bibliography for the third task has been done, leaving the implementation of the classifier to be done. Exportation to HPC is in progress for extensive usage on the most important recordings.

Progress per Task

Task 1: Detection of energy segment

The detection of the energy segment is the backbone of this work project. It consist in finding the underlying structure of the spectrogram, made of time frequency paths and intersections of these paths creating a graph, where the vertices are the intersections, and the path segments between two intersection are the edges of the graph.

Task 2: Segmentation of patterns

From the graphs generated by the task 1, task 2 aim to find the paths in the graph by combing each segment. The challenge of this task is two allow overlapping paths (such as a dolphin whistle and an electronic whistle sharing the same time-frequency bins for a laps of time) to be correctly reconstructed, with their shared edges being associated to both. Path split in multiple part due to a low **SNR!** (**SNR!**) will also be reconnected if possible. Finally, the paths of harmonics and fundamental frequencies will be clustered together.

Task 3: Classification of patterns

The final task of this work project aims to classify the whistle based on the path clusters made by the other Work Packages. This is in progress and will run on HPC on the whole data collection that is not yet completely received.

Main Results and Achievements

The Fig.7, 8, 9 are representing the main steps of this process.

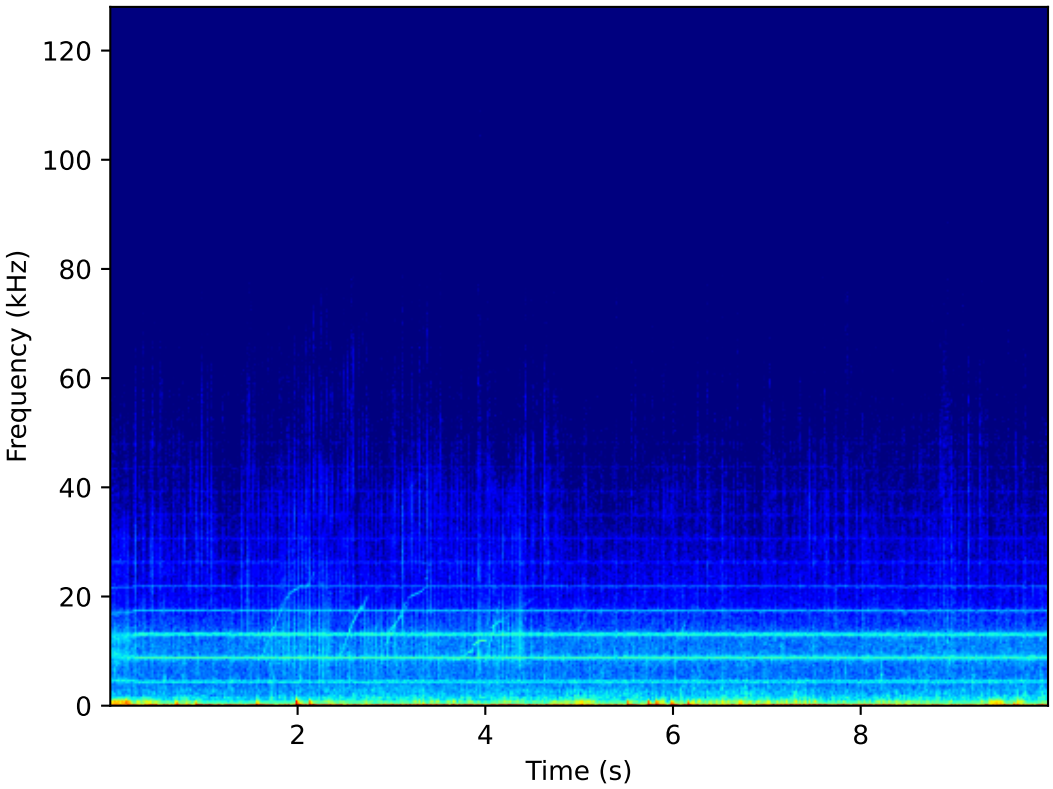


Figure 7: Spectrogram showing some weak SNR voicings overlapped with stationary noises.

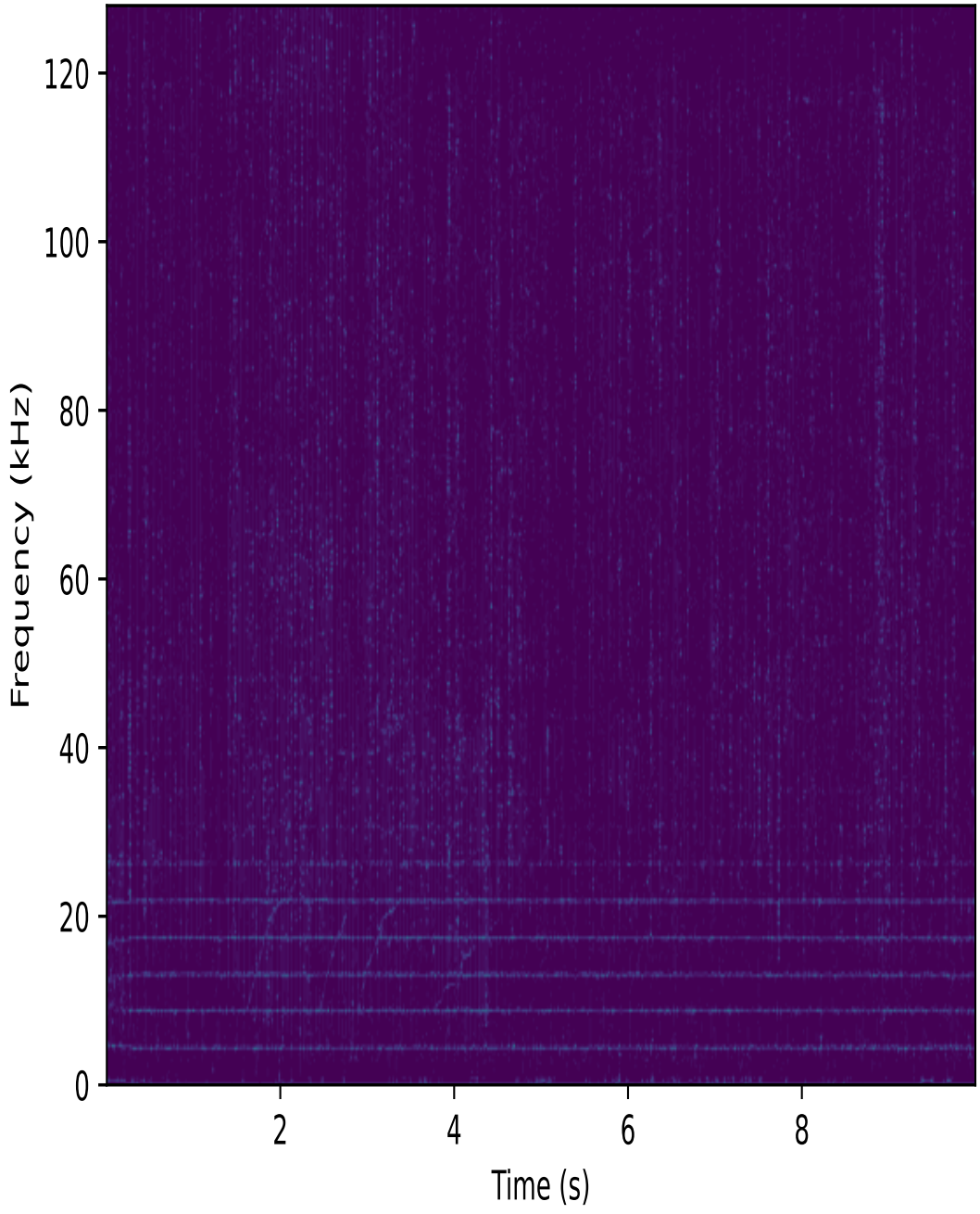


Figure 8: Energy segment detection of the previous spectrogram showing weak SNR voicing plus reef and stationary noises

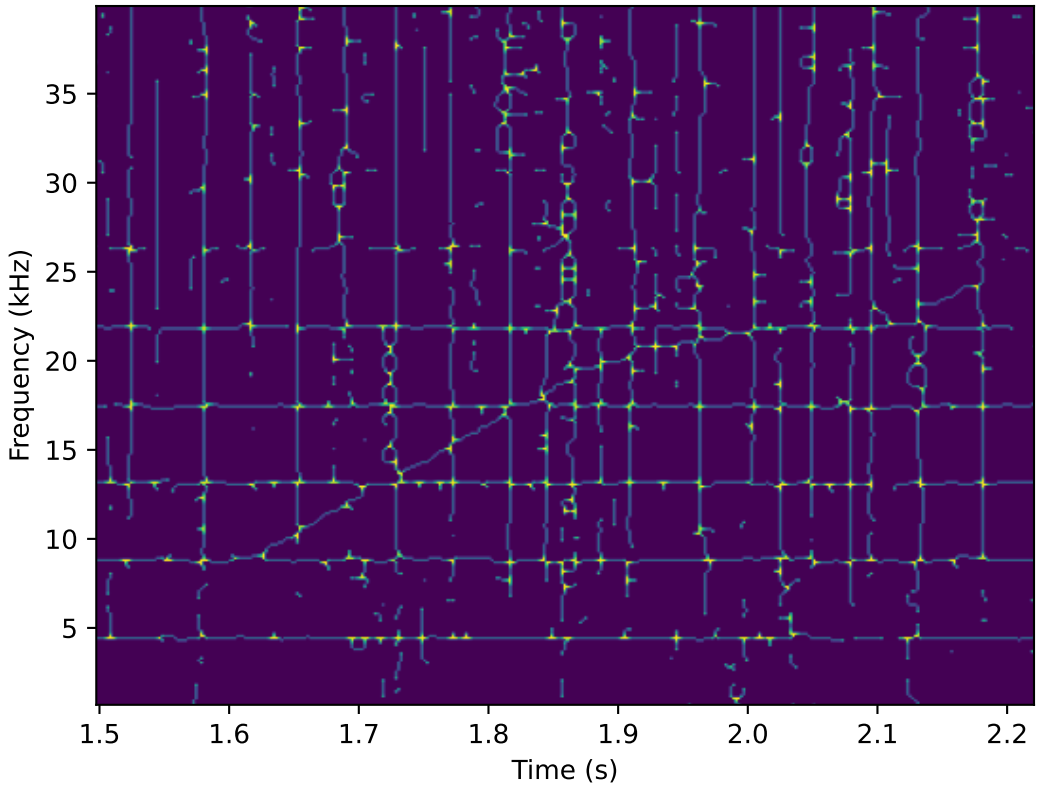


Figure 9: Zoom of the previous image, into the energy segment detection showing the voicing that has been tracked after the algorithm developed in this WP. The final pass of the algorithm will remove the noises (horizontal and vertical vertices), thus it will reveal the voicing.

3.2 WP02 Data exploration and clustering

Overview and Goals

Work Package Summary			
WP No.	02	*Title of WP*	Data exploration and clustering
Start	March	*End*	August 2022
Participating Organisations			
• Participants: PB, MF, AM, HG			

Goals
• Goal 1 : Discovery of acoustic singularities
• Goal 2 : Construction of a first train / test dataset
• Goal 3 : Clustering of detected event subsets in species
• Goal 4 : Clustering of detected event subsets over time
• Goal 5 : Clustering of detected event subsets over space
• Goal 6 : Clustering of detected event subsets over species time and space
• Goal 7 : Distribution of the code

Status An interface has been built to efficiently go through the data and annotate. First, spectrograms are computed for chunks of the signals (the size of the chunks is a parameter tuned by the annotator, depending on the target specie). Several spectrograms configurations were tuned for different target signals / species. To adapt the time and frequency resolution while preserving a standard spectrogram heigt, we resample the signal to a lower sampling rate before computing the spectrograms, all with a 1024 STFT window size and a 50% overlap.

Sampling frequency	Mel transform	Mel start frequency (Hz)	Target signals
2000	Yes	0	Very low frequency stationnary calls (Fin whales, Blue Whales)
16000	Yes	0	Low and Mid frequency stationnary calls (Humpack whales)
64000	Yes	2000	Mid frequency transitory clicks (Sperm whales)
256000	Yes	8000	High frequency transitory clicks (Delfinids, Ziphius)
256000	No	NA	High frequency transitory clicks (Delfinids, Ziphius, Kogia)

For each spectrogram configuration, we project the chunks of signals using the UMAP dimension reduction algorithm. We then cluster the latter projection using the DBSCAN algorithm. An interface was built to go through and validate those projections and the associated clustering. The user can see the projection (each point corresponding to a signal chunk), and the resulting clusters (each color corresponds to a cluster), and is able to click on a point to see its spectrogram see Fig. 10.

With such an interface in hand, the user can identify relevant clusters (ones that aggregate

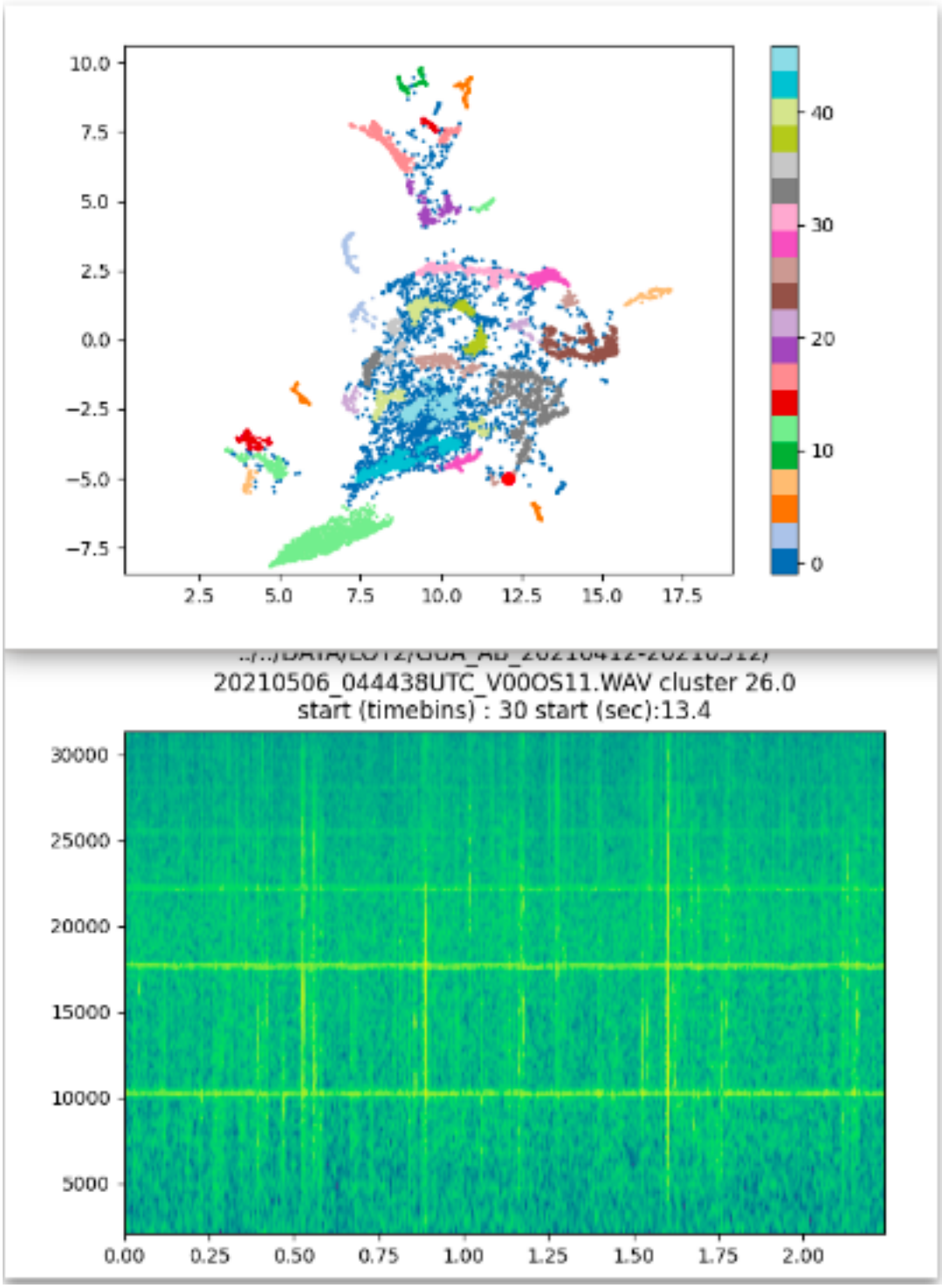


Figure 10: Interface for data exploration via projection and clustering of spectrograms. Each point in the projection (top) is a spectrogram (below).

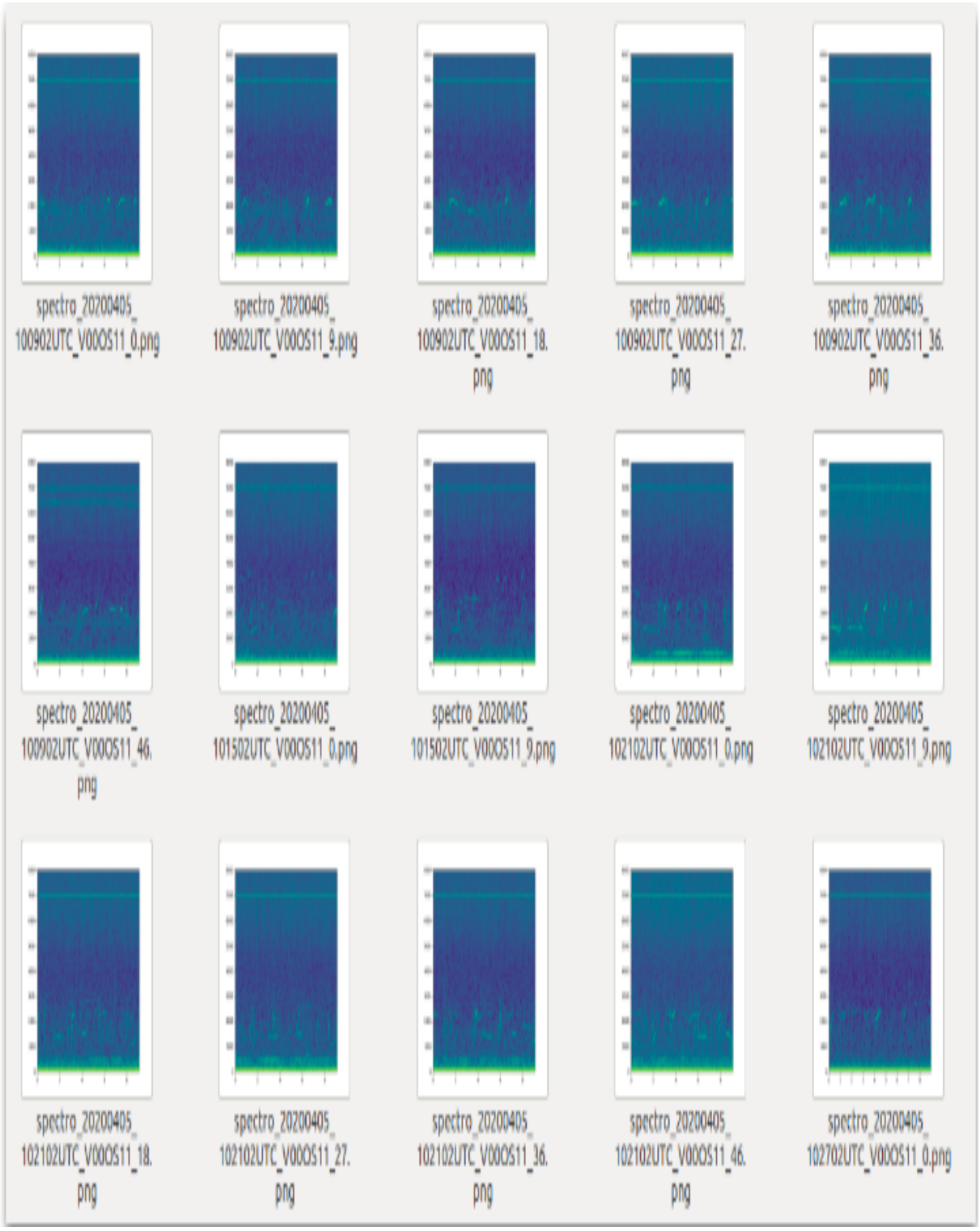


Figure 11: Samples extracted from the cluster containing mainly Humback Whale signals

similar signals of interest together). A program was then written to plot the spectrograms of the selected clusters for further annotation, via selection of .png files (see Fig. 11). The user can then sort out and filter potential misclassifications, and aggregate several clusters together, to form a database for training supervised models.

Progress per Task

Task 1: Humpack Whales

Clusters were found containing mainly humpback vocalizations (see Fig. 11). They were used to collect data and train a supervised classification model. Thousands of positive samples and negative samples were annotated with this technique.

Task 2: Dolphins

Despite extensive efforts in optimizing the representation of the data to segregate dolphin whistles together in an unsupervised manner, no cluster were found to do so. This phenomenon can be explained by several factors :

- Dolphin whistles are very rare (relatively to humpback whale calls for example). This is due to the fact that high frequency sounds travel a much smaller distance than low frequency sounds do.
- The clustering algorithm relies on a minimum amount of data points to be able to group them together
- The self noise of the sound card appeared to be quite diverse, and in shapes similar to what dolphin whistles are. The projections and clusters were thus overloaded with those artefacts, impeaching the good projection of the dolphin whistles.

Main Results and Achievements

The main results and achievements of this WP is that it allowed to discover regularities and singularities into the To of the recordings. It also allowed us to build an Active learning process to train CNN and learn few species.

3.3 WP03 Active learning

Overview and Goals

Work Package Summary			
WP No.	03	*Title of WP*	Active learning
Start	March	*End*	October
Participating Organisations			
• Participants: PB, MP, MF			

Goals
• Goal 1 Detect event of interest
• Goal 2 Assemble positive and negative data set, also from other recordings
• Goal 3 Detection of low frequencies, humpback and dolphins, then feedback Goal 2
• Goal 4 : Distribution of the code : Deliverable D2 : Software in Python GitHub to forward the data is available for CARIMAM users at https://gitlab.lis-lab.fr/paul.best/carimam_cnn. The Dyni team is available to help any users to install on local PC this system.

Status

Databases were built following the active learning pipeline (see Fig. 15). Convolutional neural networks were then trained on the binary classification task for each of the following types of signals : dolphin whistles and humpback whale vocalizations. The training revolves around the following architecture :

- Resampling (to 96kHz for dolphin whistles and 11kHz for humpback whale vocalizations)
- Standardization of the signal (substracting the mean and dividing by the standard deviation)
- Addition of brown noise with an SNR normaly distributed around 1dB (std of 2dB)
- Computation of the STFT, with a window sizes of 4096 and 512, and hop sizes of 1024 and 64 for dolphin whistles and humpback whale vocalizations respectively
- Mel transformation with 128 bands from 3kHz to 30kHz and with 64 bands from 300Hz to 3kHz for dolphin whistles and Humback whale vocalizations respectively
- Per Channel Energy Normalization (Wang, Getreuer, Hughes, Lyon, & Saurous, 2017)
- Convolutional neural network with an architecture inspired from (Grill & Schlüter, 2017)
- Binary cross entropy between the model prediction and the given label
- Backpropagation of the loss to the model's weights following stochastic gradient descent
- This process was conducted for 100 epochs

Progress per Task

Task 1: Detection of Humpback whales

The Humpback whales have been detected as described above, and then the model trained. The forward up to mid october on 60% of the received data show already interesting patterns (see section 4).
Humpback whales ranged throughout the Caribbean Sea during winter and spring (Swartz et al., 2000), Fig. 12 and Fig. 13 show Signal/Spectrogram of humpback whales during these periods.

Task 2: Detection of Dolphins A database of dolphin whistles was built from several databases available at our lab. Models were trained, forwarded on the CARIMAM dataset, until dolphin whistles were found and added to the training samples.

Active learning is a special case of machine learning in which a learning algorithm can interactively query a user (or some other information source) to label new data points with the desired outputs. It is helpful in the scenario where unlabeled data is abundant but manual labeling is expensive. In such case, active learning algorithms can actively query the user for labels. This way, new labels (verified by experts with knowledge about the desired outputs) are added at each iteration (see fig.15 for the pipeline), and a dataset weakly annotated can efficiently get strongly annotated to help better learn the model.

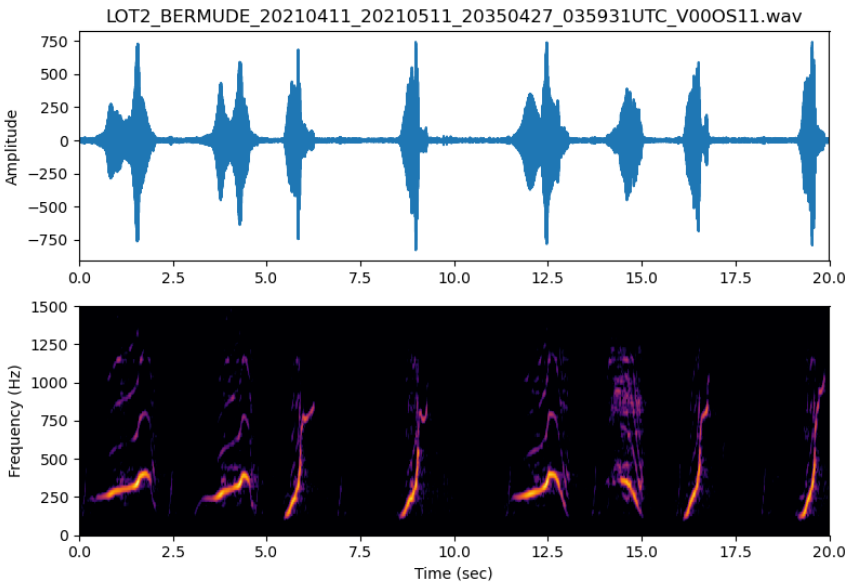


Figure 12: Example of 20 secondes of signal containing Humpback whales (Megaptera novaeangliae) vocalization for Bermude Station (LOT2)

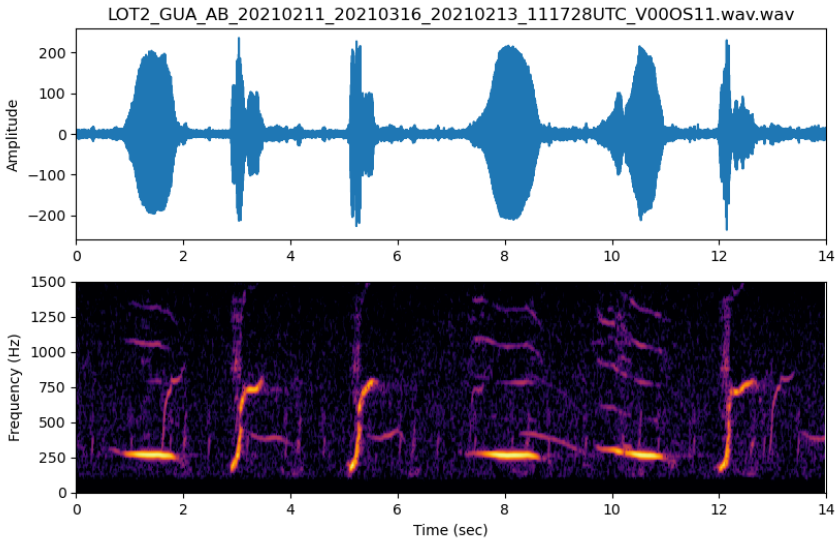


Figure 13: Example of 20 secondes of signal containing Humpback whales (Megaptera novaeangliae) vocalization for Guadeloupe Station (LOT2)

Task 3: Detection of low frequencies For the LF detection system, we used a previously trained model with an architecture similar than the one presented in (Best et al., 2020). The samples found in the CARIMAM data were not with a high enough SNR to assign them a source with full certainty. Thus, we did not further follow the active learning procedure.

Main Results and Achievements

The detections from the three models (low frequency, dolphin whistles, humpback whale vocal-

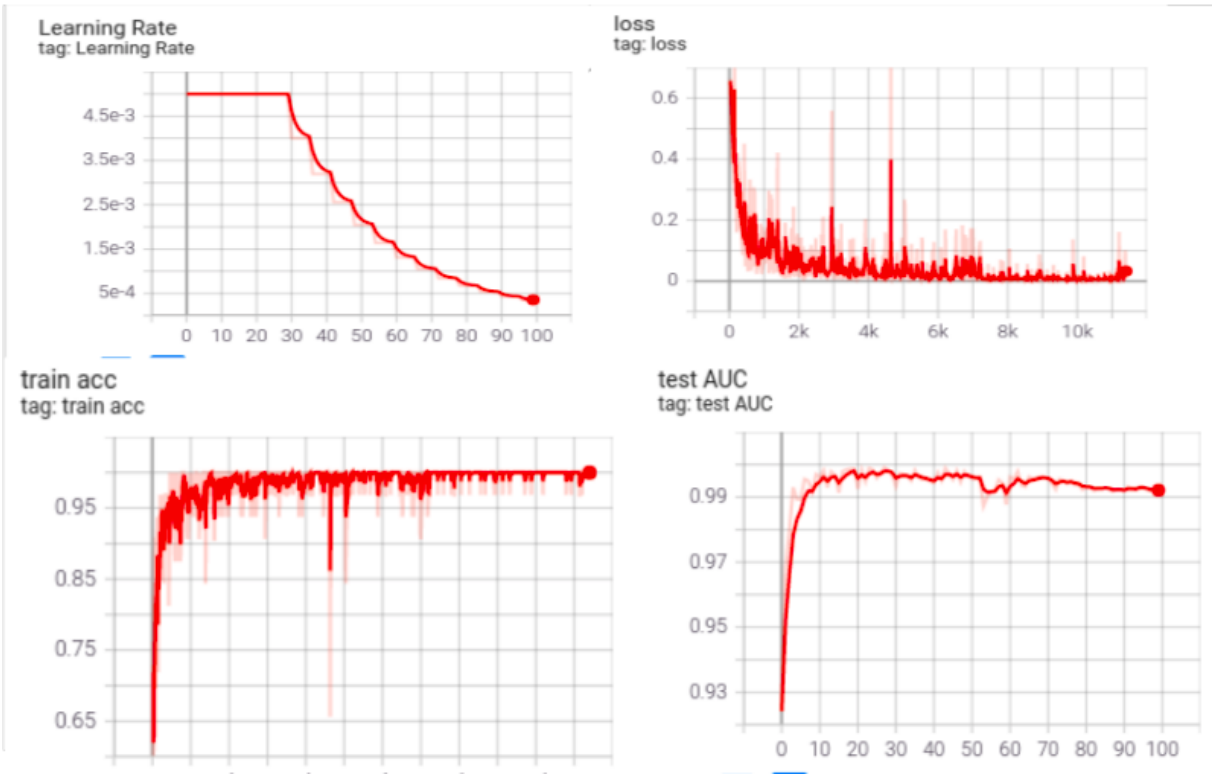


Figure 14: Training procedure for the dolphin whistle binary classification

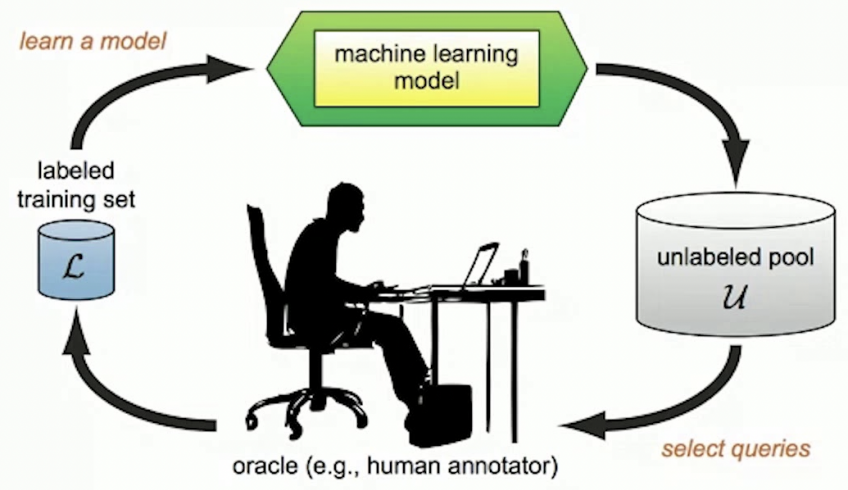


Figure 15: Active Learning Pipeline : the learning algorithm interactively query a user to label new data points with the desired outputs.

ization) are available at http://sabiod.lis-lab.fr/pub/carimam_detections/ as spectrograms pictures.

3.4 WP04 Clicks Detection and classification

Overview and Goals This WP aims first to detect clicks of odontoceti into the strong reef noise and second to inspect detections and hypotheses on the species. We started with the species that have been referenced by CARIMAM or our former projects. Up to now we have

been able to build from our own ressources the material on 8 odotoceti species. More are currently collected by CARIMAM subcontractors to complete it.

Work Package Summary			
WP No.	04	*Title of WP*	Biosonar detection by Deep learning *Start*
Automn 2019	*End*	October 2022	

Participating Organisations
• Participants: HG, MF, AM, PB, NT, MP

Goals
• Goal 1: build a reference data set to train Machine Learning solutions for click analyses of odontoceti of CARIMAM dataset Deliverable D1: corpus for the training is distributed by DYNI LIS UTLN on sabiod.fr and MADICS CNRS (http://sabiod.fr/pub/docc10) and similarly in the DATA challenge of ENS (https://challengedata.ens.fr/challenges/32).
• Goal 2: train model to detect / class odotonceties from their clicks
• Goal 3: forward using HPC on the 15 To of CARIMAM Deliverable D2 : Software in Python GitHub to forward the data is available for CARIMAM users at https://gitlab.lis-lab.fr/maxence.ferrari/carimam_docc10/-/tree/main. The Dyni team is available to help ay users to install on local PC this system and to produce pre-annotations to share with other users.
• Goal 4: refine the classifications.

Therefore we created a dataset made of clicks of various species present in the Caribbean. This dataset yet contains a third of the species that the CARI'MAM project aims to study. It allows us to build different machine learning approaches (semi- or fully automated analysis), as well as train deep learning models to solve the classification task of odontoceties into CARI'MAM. This dataset distributed to Carimam users as a benchmark for click classification in the DOCC10 (Dyni Odontocete Click Classification) challenge.[1] The next sections describe the construction of this dataset and the training of the neural networks.

3.4.1 Reference Biosonar dataset

To build this dataset large enough to train neural networks we gathered data from different sources: i) the 2018 DLCDE challenge[2], and ii) sperm whale clicks from the 2018 Sphyrna Odyssey expedition (Ferrari et al., 2019). These existing datasets contain long sequences of audio with rough annotations of the temporal regions with clicks. Our goal is to produce a set with individual clicks associated to a particular species. In this work we present our methodology to extract the clicks and label them with the species identity. We also present a preliminary analysis of the resulting corpus, a data split useful for benchmarking and a baseline deep learning model to classify the clicks. Even though our method to extract clicks and labels may induce some label noise, this is a situation encountered in a real scenario, thus increasing the ecological validity of the dataset. Furthermore this permits exploring the use of techniques specifically dealing with these issues, such as negative learning (Kim, Yim, Yun, & Kim, 2019; Fonseca et al., 2019). We thus decided to increase the number of samples, at the cost of a possible increase of mislabeling.

[1] https://challengedata.ens.fr/participants/challenges/32/
[2] http://sabiod.univ-tln.fr/DCLDE/challenge.html

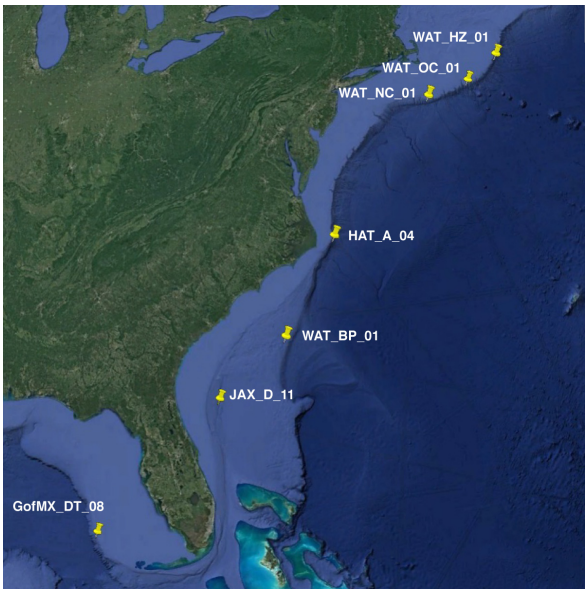


Figure 16: Recording locations of the 2018 DCLDE challenge used to train the detectors of CARIMAM

The high-frequency dataset from the 2018 DCLDE challenge consists of marked encounters with echolocation clicks of species commonly found along the US Atlantic Coast and in the Gulf of Mexico:

- Mesoplodon europaeus - Gervais' beaked whale
- Ziphius cavirostris - Cuvier's beaked whale
- Mesoplodon bidens - Sowerby's beaked whale
- Lagenorhynchus acutus - Atlantic white-sided dolphin
- Grampus griseus - Risso's dolphin
- Globicephala macrorhynchus - Short-finned pilot whale
- Stenella sp. - Stenellid dolphins
- Delphinid type A
- Delphinid type B

The goal is to identify the times at which echolocating individuals of a particular species approached the area covered by the sensors. Analysts examined the data in search of echolocation clicks and approximated the start and end times of acoustic encounters. Any period that was separated from another by five minutes or more was marked as a separate encounter. Whistle activity was not considered. Consequently, while the use of whistle information during echolocation activity is appropriate, reporting a species based on whistles in the absence of echolocation activity would be considered a false positive for this classification task.

The reference data were recorded at different locations in the Western North Atlantic and Gulf of Mexico as shown in Figure 16. In the accompanying table 1, we list the coordinates and depths of the various sites. These data were collected between 2011 and 2015, and the time period for each recording can be inferred directly from the data.

Project	Site	Deployment	Preamp	Lat N	Long W	Depth
WAT	HZ	1	734	41-03.7	66-21.1	850
WAT	OC	1	707	40-15.8	67-59.2	1100
WAT	NC	1	740	39-49.9	69-58.9	980
HAT	A	4	685	35-20.8	74-50.9	840
WAT	BP	1	810	32-06.4	77-05.7	945
JAX	D	11	681	30-09.0	79-46.2	800
GofMX	DT	8	638	25-32.3	84-37.9	1200

Table 1: DCLDE recording meta data

3.4.2 Enhancing weak labels of initial reference dataset

For each of the species contained in the initial DCLDE dataset previously described, the labels are the lapse of time indicating the presence of the corresponding species. The longest interval between two clicks in a segment can last up to 5 minutes. For CARIMAM, we considered these labels as weak, in the sense that they do not reflect precisely the timestamp of each click. As in CARIMAM we are interested in the detection and classification of the individual clicks, therefore annotations at a much finer temporal scale are required. These are the labels that we will refer to as *strong*.

In order to extract strong labels from the weak DCLDE 2018 weak labels, we first retain only energy components in the frequency ranges of the clicks by applying a bandpass filter. After this filtering step, we use a Teager-Kaiser (TK) filter (Kandia & Stylianou, 2006; Glotin, Caudal, & Giraudet, 2008) combined with a local maximum extractor having a half window length of 0.02 s, to obtain the position of all these clicks. Since most of the maxima will not be actual clicks but background noise, a median filter is used on the logarithms of these maxima to evaluate the background noise level. Any maxima above the noise level plus 0.5 dB are kept. Windows of 8192 samples are then extracted around these clicks.

We then proceed to label these maxima with the labels from the DCLDE challenge. If a click is in the interval of two or more weak labels, we assign it all of the corresponding labels. We also extract multiple acoustic features to curate the new DOCC10 dataset from mislabeled clicks. One must note that in the DCLDE data the clicks of all present species are not labeled. There may be segments labeled as containing a single species that contain clicks from other species that are not part of the DCLDE label set, such as sperm whales. We decided to use the spectral centroid as the feature to perform the final filtering, since it is the feature with which the outliers are better distinguishable from actual clicks. The spectral centroid is the weighted mean of the frequency, using the Fourier transform amplitude as weights.

The spectral centroid is however not useful to classify clicks on its own, as most of the DCLDE species will have clicks with similar spectral centroids, mainly in the range 30 kHz - 40 kHz. Thus it cannot be used to chose one label for clicks that have multiple labels.

The corpus has been completed by the 2018 Sphyrna Odyssey expedition. This set contains clicks from sperm whales, *Physeter Macrocephalus*. All the clicks are from a single sperm whale 3 hour encounter. These clicks were recorded at 300 kHz by a Cetacean Research C57 hydrophone and JASON sound card from SMIoT UTLN. The sperm whale clicks were detected using a detection process similar to the one used to create strong labels from the DCLDE dataset. We cross-correlated the signal with one period of a 12.5 kHz sine wave which acts

Label	Scientific name	Common name
Gg	Grampus griseus	Risso's dolphin
Gma	Globicephala macrorhynchus	Short-finned pilot whale
La	Lagenorhynchus acutus	Atlantic white-sided dolphin
Mb	Mesoplodon bidens	Sowerby's beaked whale
Me	Mesoplodon europaeus	Gervais' beaked whale
Pm	Physeter macrocephalus	Sperm whale
Ssp	Stenella sp.	Stenellid dolphins
UDA		Delphinid type A
UDB		Delphinid type B
Zc	Ziphius cavirostris	Cuvier's beaked whale

Table 2: Class labels

as a band-pass filter (bandwidth of echolocation clicks is 10–15 kHz (Madsen et al., 2002)). We then apply a Teager-Kaiser filter (Kandia & Stylianou, 2006; Glotin et al., 2008) and extract the local maxima in 20 ms windows (twice the largest inter-pulse interval of 10 ms (Abeille et al., 2014)). For each 1 minute audio file we compute the mean and standard deviation of the maxima values in decibels (dB), and only keep samples over three times the standard deviation (Pukelsheim, 1994). To incorporate them in DOCC10, we down-sampled the signal at 200 kHz to match the sampling rate of the DCLDE datset.

Since the data added contain a single unseen species, we are introducing a bias of high correlation between recording configuration, environment and the species label. However this can be seen as a usual approach to composing bioacoustics datasets for machine learning and will evidence the issues with such a method in the benchmark.

The final dataset consists of clicks centered in a window of 8192 samples. This was motivated by the possibility of analysing clicks in a window of 4096 samples while being able to offset this shorter window. The combination of DCLDE and Sphyrna Odyssey brought this new dataset to a total count of $134,080$, that we split into a training set of $113,120$ clicks and a test set of $20,960$ clicks for the DOCC10 challenge, which produces an approximately 85-15 split. The test set is balanced with 2096 clicks per class. For the challenge, the test set was split into a private test set (90%) and a public test set (10%). This split was done randomly, so that the classes are no longer perfectly balanced. The training set is also perfectly balanced with $11,312$ clicks per class. The species are the class as detailed in Table 2. We show in Fig. 17 examples of clicks contained in the DOCC10 dataset for each class (except for the sperm whale).

3.4.3 Convolutional Neural Net classification model

A large part of machine learning research is done on image classification (LeCun et al., 1989; Deng et al., 2009; Krizhevsky, Sutskever, & Hinton, 2012). When working on sounds, the usage of spectrograms or Mel-frequency cepstral coefficients (MFCC) allows one to convert these 1D signals into images, and use the state of the art techniques such as ResNet (He, Zhang, Ren, & Sun, 2016). Even if this trick is largely used in signal processing, it has the disadvantage of having a number of parameters that need to be tuned beforehand, such as the stride, the window size for the FFT, which will affect the time/frequency resolution. Not only choosing the right representation for each specific task is not obvious, but choosing the wrong parameters for these hand-crafted features might decrease the performance.

In bioacoustics, *bulbul* and *sparrow* (Grill & Schlüter, 2017), are two architectures using the

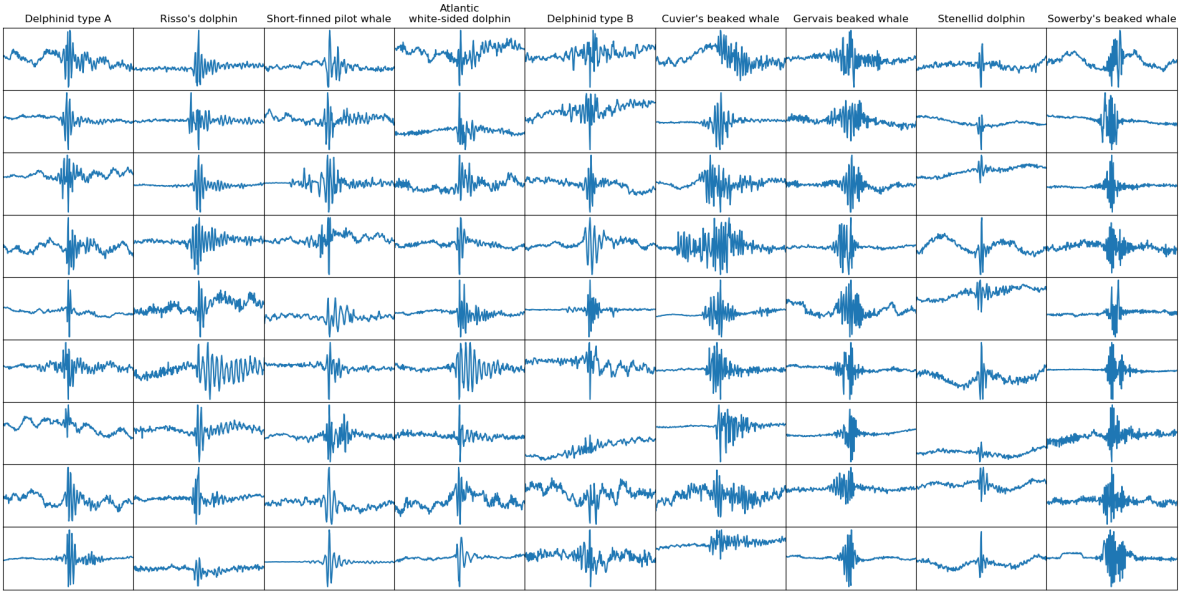


Figure 17: Zoom on the same examples of DCLDE test instances for each class (256 samples long)

STFT magnitude spectrograms that were made for the Bird audio detection challenge[3] and are nowadays used as the state of the art since *bulbul* won the challenge(Fairbrass et al., 2019; Poupard et al., 2019).

Instead of using 2D spectrograms, which are better for the analysis of chirps or stationary signals, we decided to learn directly from the raw signal, starting with convolution layers similarly to what is done in the study of ECG signals (Fukumori, Nguyen, Yoshida, & Tanaka, 2019; Kiranyaz, Ince, Abdeljaber, Avci, & Gabbouj, 2019). The advantage of a convolution layer over a dense layer is that it will force the learned filter to be invariant to a translation of the signal (LeCun, Bengio, et al., 1995). The multiple filters of a convolution layer will output multiple features per time step, which can be considered as a new dimension with one feature. Two-dimensional convolution can thus be used on this 2D signal, reducing the amount of parameters per layer amongst the other advantages of convolutions (Jacobsen, Oyallon, Mallat, & Smeulders, 2017). This can be done after the first layer, or after multiple 1D convolution layers (Huang & Leanos, 2018; Amodei et al., 2016). The operation can then be repeated to perform a 3D convolution. For convenience, we call this increase of dimension followed by a convolution, UpDim. This operation could then be repeated to increase the number of dimensions to 4D and more. However, usual deep learning libraries such as Tensorflow or PyTorch do not support convolution on tensors with more than 3 dimension (5 if the batch and feature dimension are taken into account).

Thus we apply our new operator UpDim in a CNN of 12 layers using the raw audio as an input. Windows of 4096 bins are extracted randomly from the 8192-wide samples, and random pink, white and transient noises are added to it, each having an independent amplitude distribution that is log-uniform (to be uniform in dB scale). The result is then normalised and given to the first layer of the CNN.

The first layers of this model are alternates of convolution and increase in dimension using our proposed UpDim operator.

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

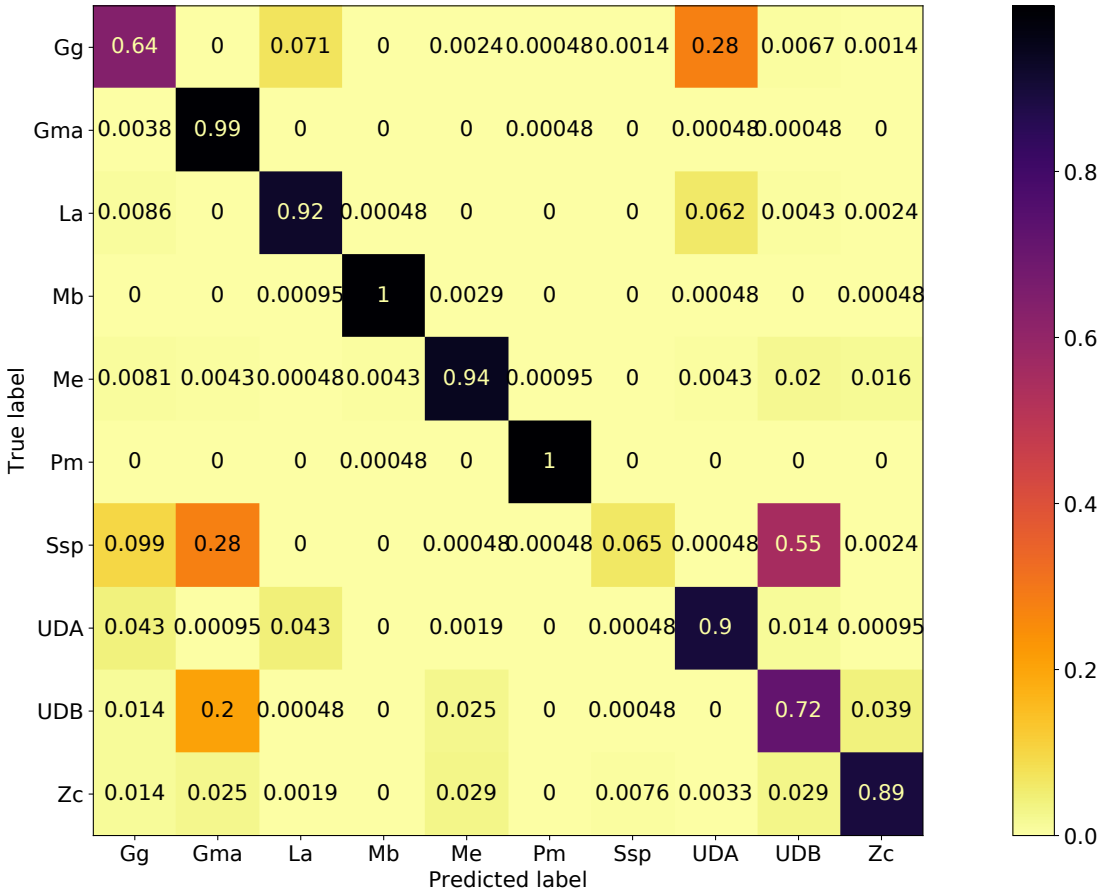


Figure 18: Confusion matrix on the test set

Main Results and Achievements

As this baseline was originally built for the CARI'MAM project, it was trained with an additional class, the noise class, which was trained with the artificial noise cited earlier. Hence the network topology has 11 classes instead of the 10 of the dataset. For the evalutation of the full DOCC10 test set, the logit of the noise class was dropped before the softmax. The confusion matrix is thus obtained by the prediction without the noise logit. Note that the confusion matrix on a test set which includes noise sample is the same as the one shown in this paper, with all the noise sample being classified as noise, and one Stenellid dolphin being classified as noise. The baseline obtains a MAP (mean Average Precision) of 77.12% and an accuracy of 71.13% on the full test set. On the public portion of the test set, the MAP is 77.68% and the accuracy is 70.52%.

Our model uses a modified version of a resnet that uses the UpDim principle (Ferrari, Glotin, Marxer, & Asch, 2020). The activation functions used are leaky ReLU. Batchnorm was also used after each convolution layer except the ones of the skip connections. The loss is the cross entropy with softmax (see (Ferrari et al., 2020) for details). The confusion matrix of this experiment with an accuracy of 80.62% is Fig. 18.

This DOCC10 dataset with strong labels for marine mammal transient classification has a total of 134,080 clicks for 10 species. Except for part of the test reserved for the scoring of the DOCC10 challenge that has been opened with this dataset, the dataset is publicly available as CARIMAM training dataset (Ferrari et al., 2020). According to the detections we get, we plan to include records from the CARI'MAM project into DOCC10. This augmented dataset will

increase the variety in the acoustic environment (the varioous reef noise in CARIMAM dataset) and the CNN on it will be more robust to unseen background noise.

In addition to their frequencies, the clicks have specific Inter Click Interval for each species (Tab. 3). We then merge the click detector max variance with basic statistics on ICI into a final decision that is presented into section 4.

Species	**Min Freq (kHz)**	**Max freq (kHz)**	**Duration (µs)**	**Centroide (kHz)**	**mean IPI (ms)**	**ICI (ms)**
Kogia sp	60	130	100 to 300		40	50 to 500
Orcinus Orca	20-30	40-60	100 to 250	50	8	
Mesoplodon europaeus	30	70	200		8	
Mesoplodon densirostris	25	55	250			
Ziphius cavirostris	16	60	175 to 300	40		400 to 500
Grampus griseus	10	150	30-50	60-90		
Pseudorca crassidens	30	117	30	62		
Peponocephala electra	13	24	586		2,47	
Stenella longirostris	40-60	120-140	9	70		
Stenella attenuata	40-60	120-140	18	70		
Tursiops truncatus	60	140	10-20			
Physeter macrocephalus	2.5	20	3000	15	2-10	20 to 2000
Globicephala macrorynchus			140			
Steno bredanensis	2.7	256	100-200			
Lagenodelphis hosei	?	?	?	?		

Table 3: Description of the train click and the cicks fro the species according to the current bibliography. (?) is not referenced according to our knowledge

4 Exploitation and Dissemination of Results

4.1 Detection results of the two groups : Humpback whales and dolphins

The Table 19 gives the count of the files with the detections per group ranging from December 2020 to October 2021.

station	startdate	enddate	nfiles	dolphin	LF(30Hz)	humpback	humpback_%detec
MART_PRECH	2020-12-03	2021-01-07	8262	0	3	92	1
GUA_BREACH	2020-12-07	2021-01-08	7551	9	7	110	1
MART_StANNE	2020-12-07	2021-01-08	7577	0	1637	258	3
GUA_AB	2020-12-09	2020-12-17	1736	2	31	25	1
GUA_SF	2020-12-11	2021-01-16	8602	0	0	19	0
ARUBA	2020-12-12	2020-12-21	520	2	0	6	1
BON	2020-12-17	2021-01-26	6648	5	3	13	0
StEUS	2020-12-17	2021-01-12	6169	3	10	61	0
StBARTH	2020-12-20	2021-01-23	8112	0	5	850	10
ANGUILLA	2021-01-28	2021-03-18	8619	1	0	7443	86
GUA_SF	2021-02-03	2021-03-09	8139	0	11	3079	37
StBARTH	2021-02-07	2021-03-13	7986	0	2	6845	85
GUA_BREACH	2021-02-08	2021-03-22	8177	3	2	6608	80
StEUS	2021-02-09	2021-03-15	8154	20	2	4384	53
JAM	2021-02-09	2021-03-14	7855	3	4	31	0
GUA_AB	2021-02-11	2021-03-16	7805	2	1	7684	98
MART_PRECH	2021-02-15	2021-03-21	8060	2	24	2174	26
BERMUDE	2021-02-24	2021-03-26	7147	7	7	5226	73
MART_PRECH	2021-03-26	2021-04-29	8153	2	11	2802	34
StBARTH	2021-03-26	2021-05-06	8075	0	0	7775	96
GUA_BREACH	2021-04-03	2021-05-08	8225	0	1	2891	35
BAHAMAS	2021-04-08	2021-05-11	7684	3	0	27	0
BERMUDE	2021-04-11	2021-05-11	7303	11	1	2751	37
GUA_AB	2021-04-12	2021-05-12	7003	5	0	5555	79
GUA_SF	2021-04-13	2021-04-29	3727	0	1	3365	90
ARUBA	2021-04-24	2021-04-30	551	1	4	6	1
BON	2021-04-25	2021-05-21	6179	20	6	14	0
StMARTIN	2021-05-05	2021-06-09	8042	0	1	4088	50
GUA_AB	2021-05-14	2021-06-16	7711	1	0	107	1
BAHAMAS	2021-05-29	2021-07-02	7876	9	0	2	0
GUA_BREACH	2021-06-03	2021-07-07	8194	2	0	213	2
StBARTH	2021-06-10	2021-07-21	16237	1	0	818	5
GUA_SF	2021-07-27	2021-07-28	365	0	0	3	0
GUA_AB	2021-07-28	2021-08-30	7926	8	0	7	0
GUA_BREACH	2021-08-01	2021-09-04	8071	1	0	129	1

Figure 19: Table of the detections by recording session, for each species

4.2 Method for spatial and time pooling on the map

To illustrate the trends of the dynamics of the distribution of these groups, we split the study into two periods : T1 goes from December 2020 to March 2021,T2 goes from April 2020 to October 2021. Both starting and ending months are included in the periods defined. In addition to splitting the results into 2 periods, they are also split into 2 species. We aimed to represent the detections repartition of Humpback whales and Dolphins. The unidentified Low Frequency (ULF) are suspected to be noises from the mooring.

The method used to plot those repartitions is described below :

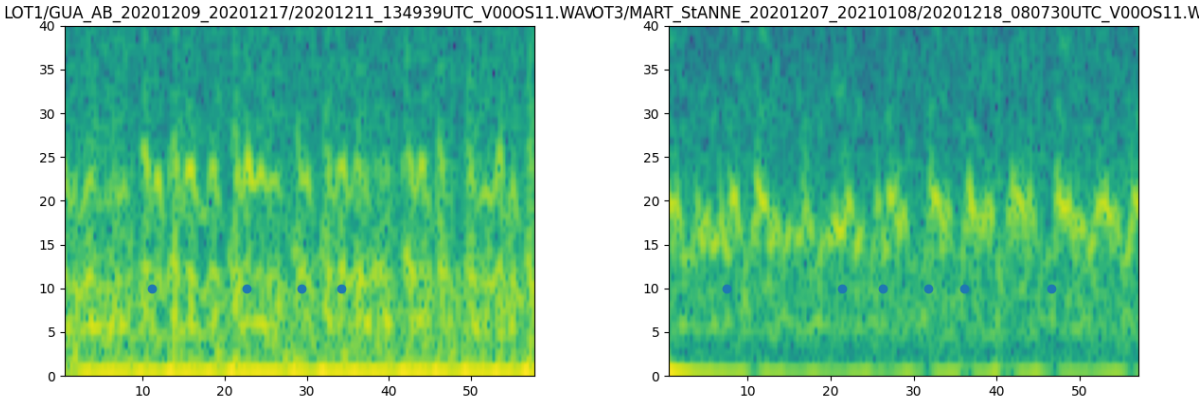
1. Inference of AI model to predict if species is present in each recording file of 1 min duration.
2. We merge the detection results in a table (see fig.19) and give, for each raw representing a recording session, the station concerned, number of files, period of the session, and the number of detections for each species.
3. We split the sessions according to periods T1 and T2 and merge results of several sessions within the same stations so we have a raw for each station in the final tables.
4. The ratio of detections is computed as : $r_{det} = n_{det}/n_{files}$
5. We then compute a 2D gaussian probability density function (pdf) weighted by r_{det}, centered and computed for each station and add those weighted pdf.
6. We then plot the log of this pdf to smooth big disparities of the number of detections between sessions, according to $pdf_{plot} = log(1e6.pdf + epsilon)$.

4.3 Spatial density detection results per group

The spatial and time illustrations suggest dynamics of the populations. More detailed detections, and per species of dolphins will follow. More stations and recordings will be received such that more consistent result will follow.

4.3.1 Unidentified Low Frequency group (infra sound)

We recorded low frequency bioacoustic signals from an unidentified source/specie. These events are Low Frequency (below 40Hz, see fig.20) and are currently investigated. Despite we don't know yet what those signals are from, we think it can be due to noise from the mooring.



(a) Signal recorded in Guadeloupe (Anse-Bertrand) *(b) Signal recorded in Martinique (Saint Anne)*

Figure 20: 1min records Spectrogram showing Unidentified Low Frequency (ULF) signals, in December 2020, from two different stations: Martinique then a week after in Guadeloupe. Abscissa in second, ordinate in Hz.

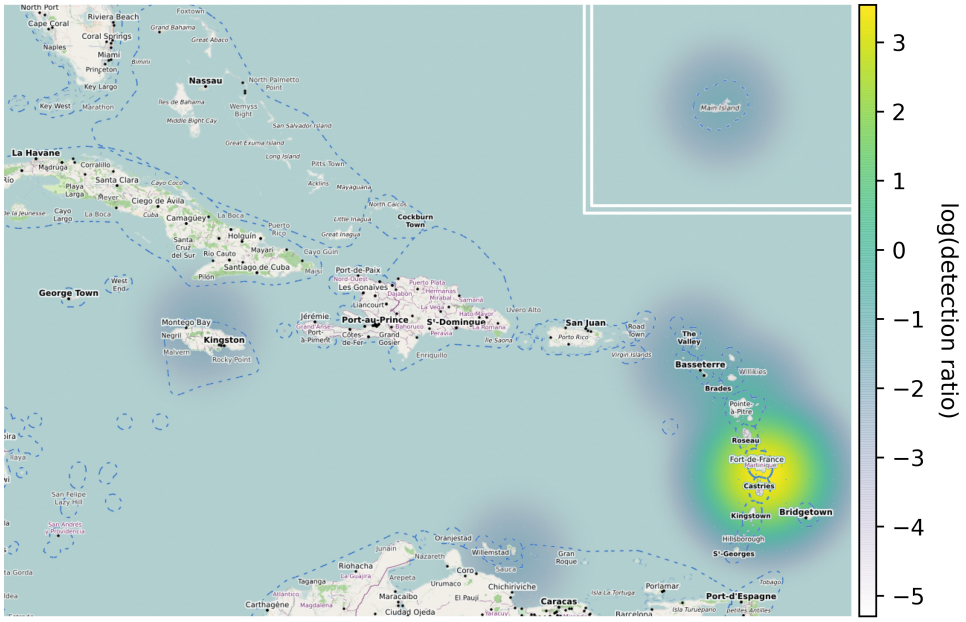


Figure 21: Repartition of ULF detection from Dec 2020 to March 2021.

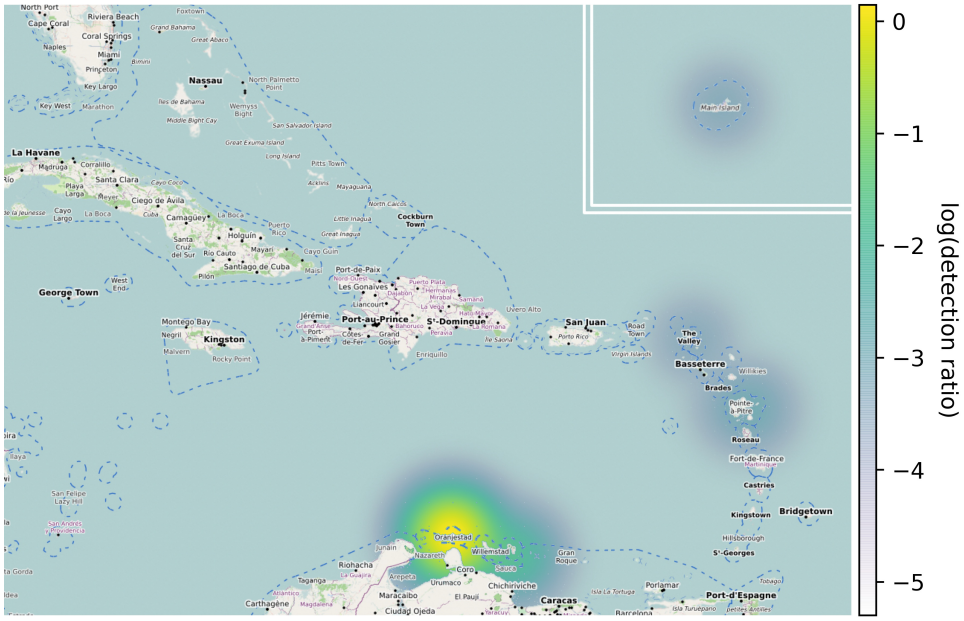


Figure 22: Repartition of ULF detections from April to October 2021.

4.3.2 Humpback whales group (medium sound)

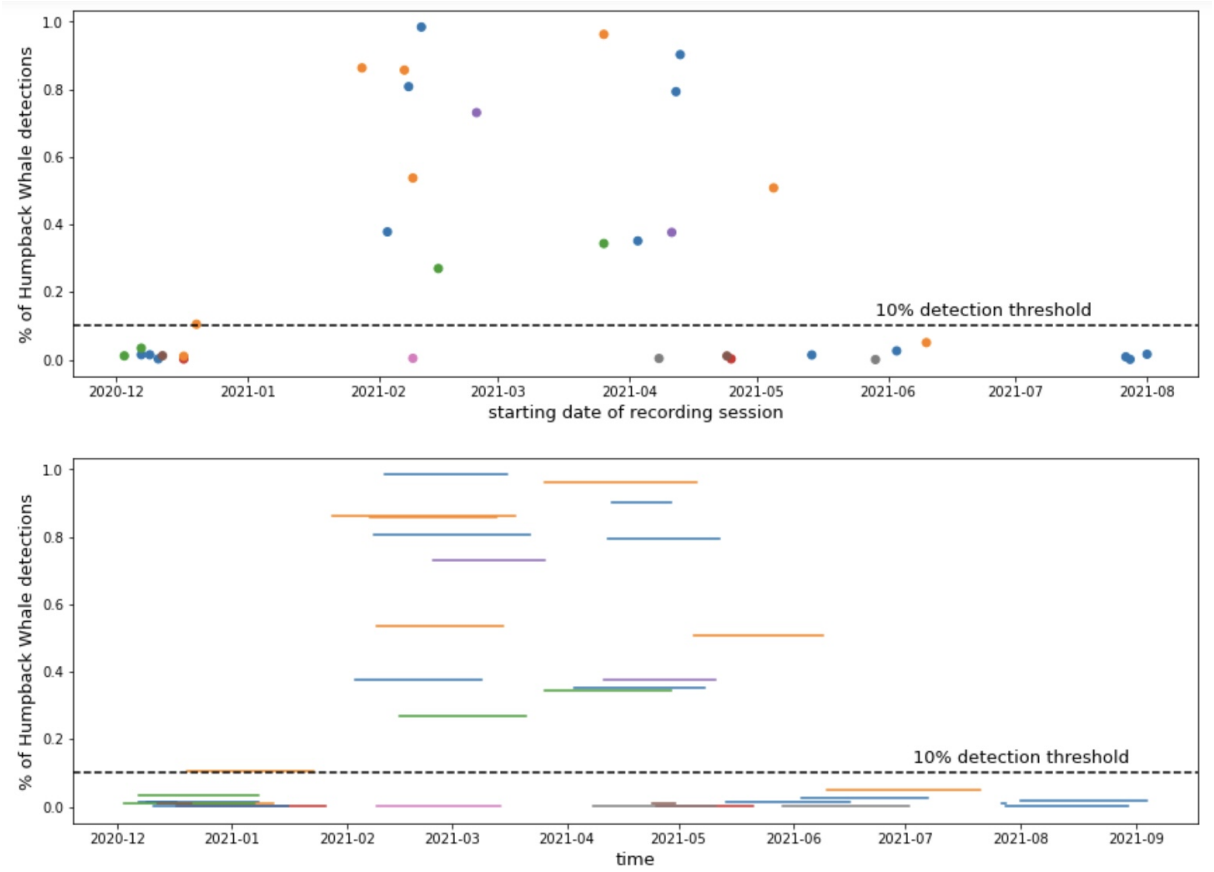


Figure 23: Chronologic repartition of Humpback Whales detections from December 2020 to october 2021. each color represents a unique recording station or a group of several ones (up to 3) close together.

The chronological repartition of humpback whales detection (fig.23) emphasizes the big presence peak of this specie during the February - March period. We plot a horizontal line denoting a 10% detection threshold and consider only what's above it. This is done to avoid wrong interpretations of the data corrupted by false positives which are unfortunately inevitable even with the best classifiers.

For Humpback Whales, because the dynamic is fast, we decided to split the chronological repartition in 3 periods. T1 goes from December 2020 to the end of January 2021. T2 goes from February to the end of March 2021. T3 goes from April to the end of May 2021.

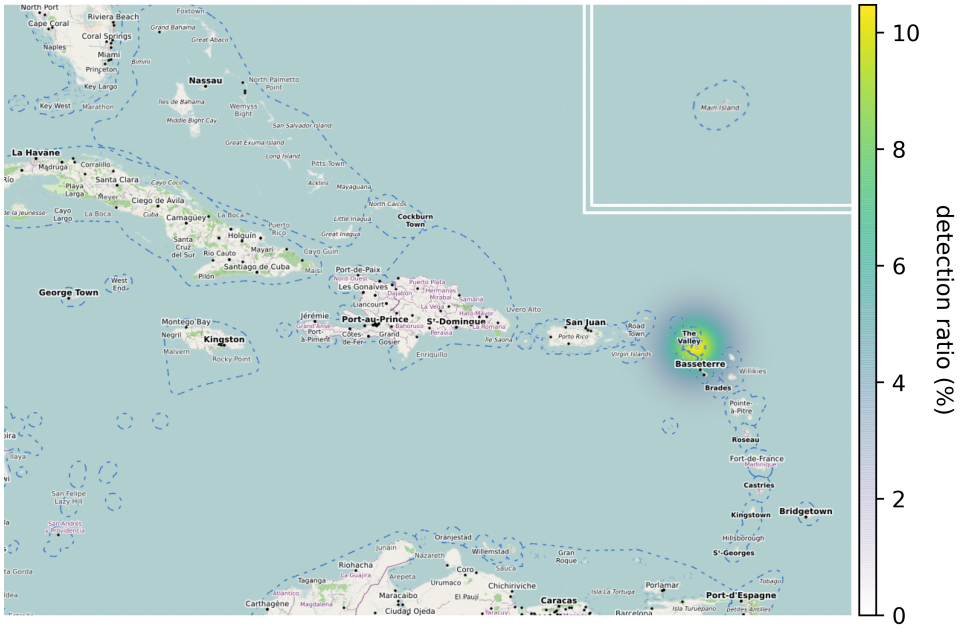


Figure 24: Repartition of Humpback Whales detection from December 2020 to the end of January 2021.

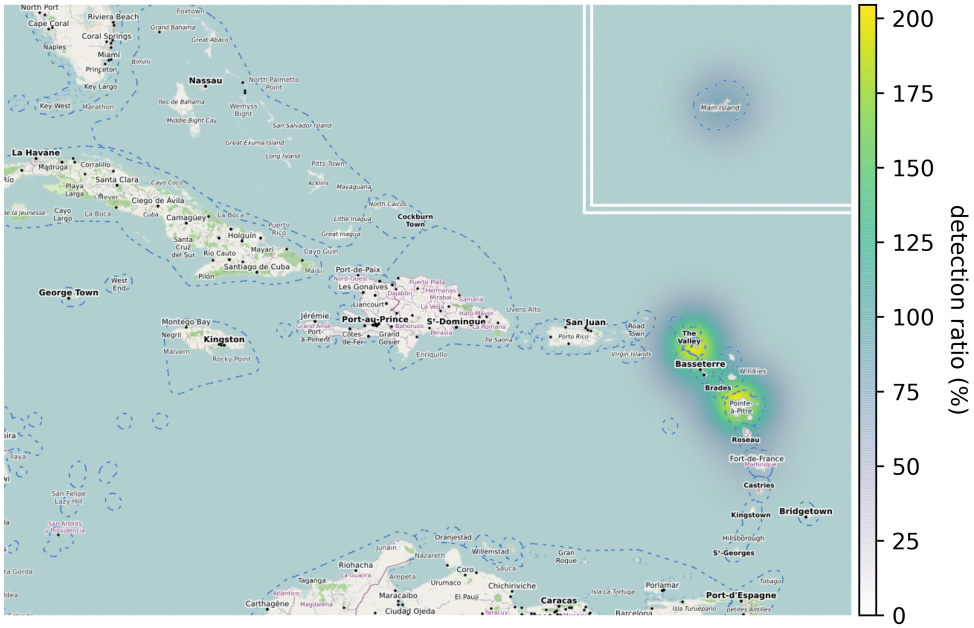


Figure 25: Repartition of Humpback Whales detection from February to the end of March 2021.

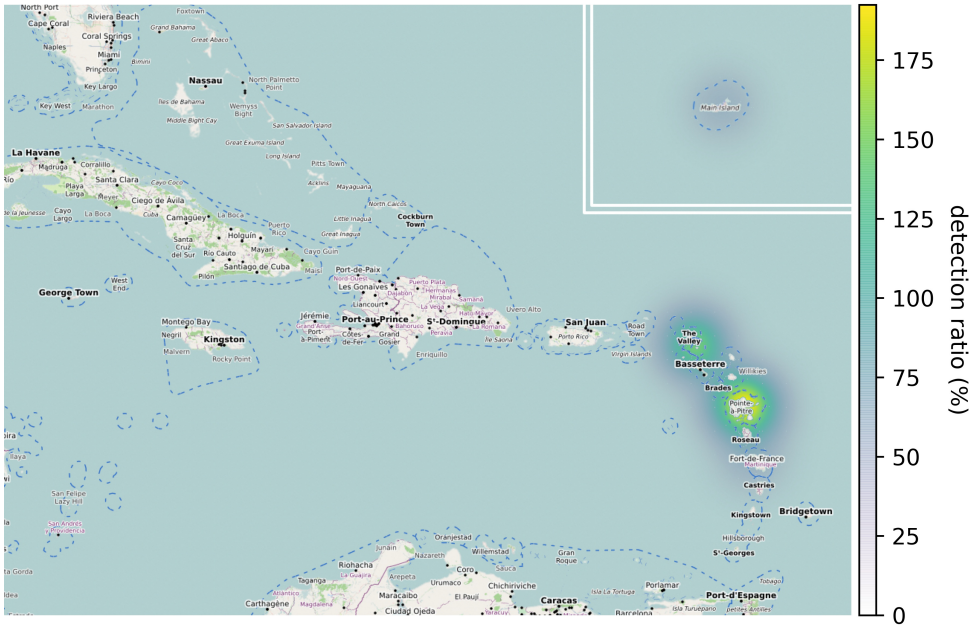


Figure 26: Repartition of Humpback Whales detection from April to the end of May 2021.

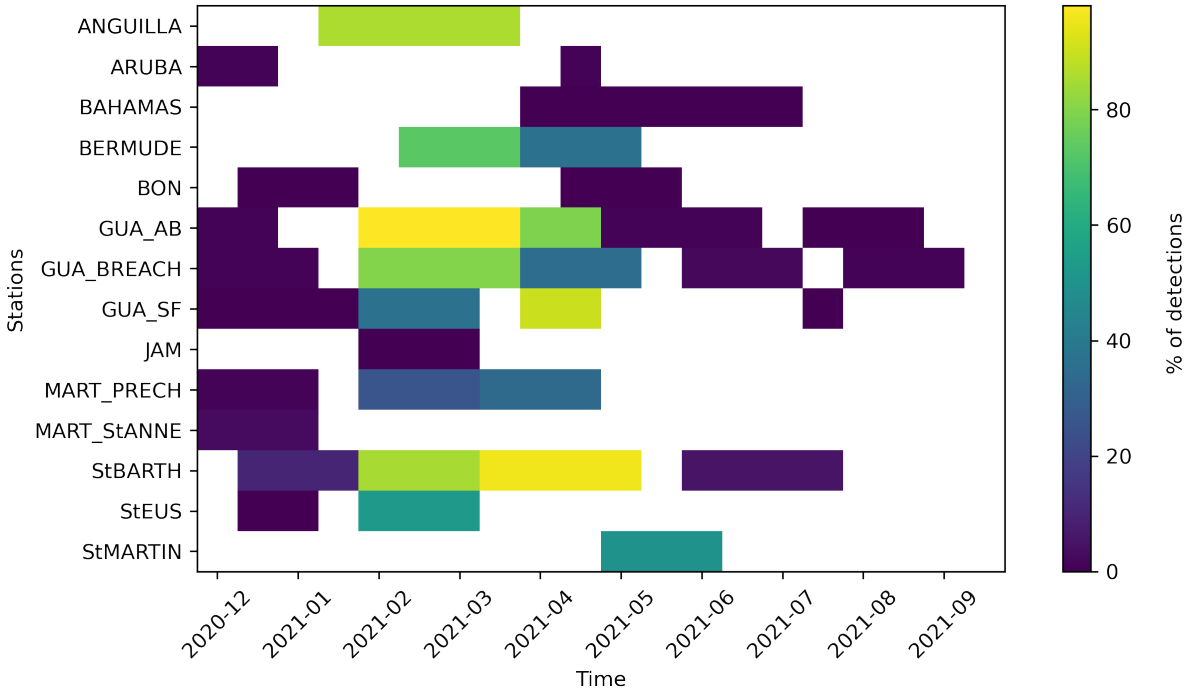


Figure 27: Time repartition of Humpback Whales detection for each recording station.

More detailed detections in time and space for humpback whales can be visited online on the interface we built for interaction with the users at : http://sabiod.lis-lab.fr/pub/CARIMAP/. An example is given in Fig. 28.

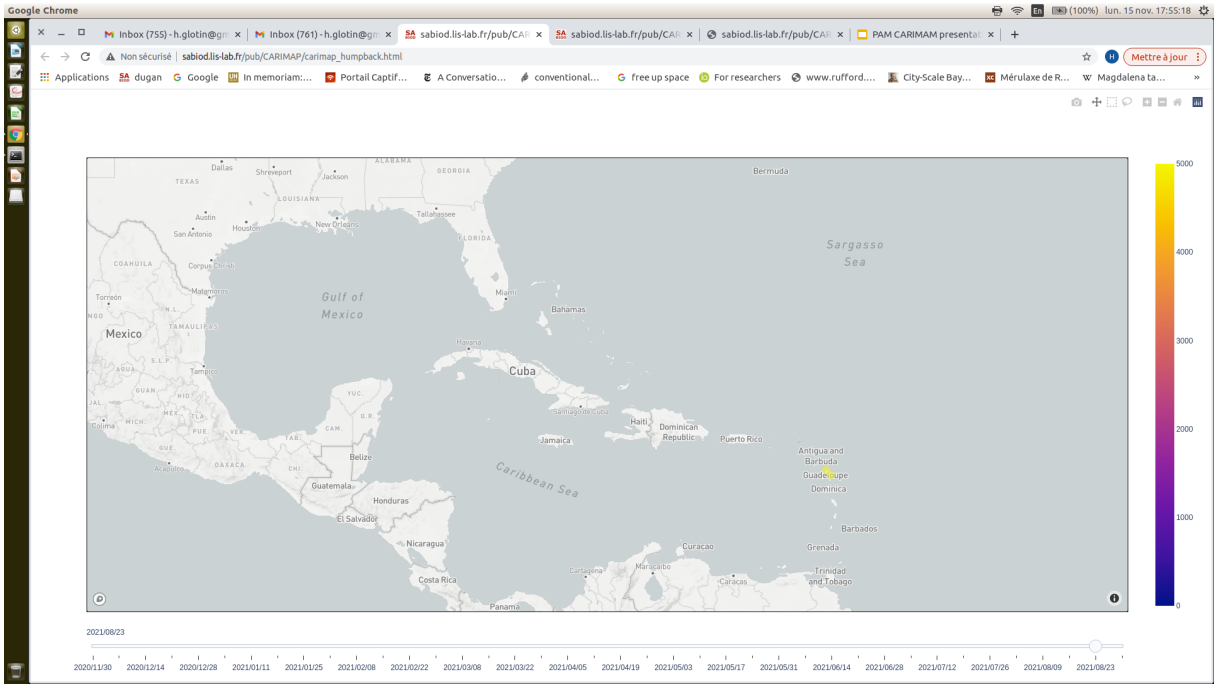


Figure 28: Time and space repartition of Humpback Whales detection for each recording that wan be visited to visual inspection at http://sabiod.lis-lab.fr/pub/CARIMAP/.

4.4 Dolphins group (ultra sound) : voicing detection

The dolphin detections are given from (1) the voicing detection (Work Package 3.3), and (2) completed by the click detection (from Work Package 3.4). We give in Fig. 29, 30 and 31 synthetic results for these voicing detections. The whole data set to be received will be processed accordingly and published to the users.

4.5 Dolphins group (ultra sound) : biosonar detection and classification

The intense reef noise (see Fig.6) can mask the biosonars of odontoceties. Then we used the DOCC10 prediction of biosonar to focus on most probable clicks. Percentiles of the 10 normalized logits of DOCC10 outputs over whole 10 sessions are shown Fig. 32. It shows the varying information in time and in session of the biosonar logits that may be due to displacement of the odontocetes. To go further, we clusterized some outputs of the 10 DOCC10 logits predictions and overlaped with the detected voicings from WP3. An example is given 33. These UMAP clustering in 2D over the 10 dimensionss of the DOCC10 outputs, over all the files of three sessions : JAMAICA 20210 202107, BONAIRE 20210211 131700UTC 20320321 and StMARTIN 20210211 103000 20210324 151700. The (X) labels are the detected dolphin voices. Colors of the clusters are groups of similar dates. We then explored as in WP3.3 the files close to the ones including voicings to build a simple statistical decision rule filtering the transient of the reef noise versus the biosonar. It uses results from WP3.4 on click detection checked accordingly to their temporal patterns (ICI). The ICI are smoothly changing in Biosonar, and can depend of the species, see and example illustrated in Fig. 35 and Tab.3.

We finally get hypotheses of dolphin species that is running on the whole CARIMAM dataset.

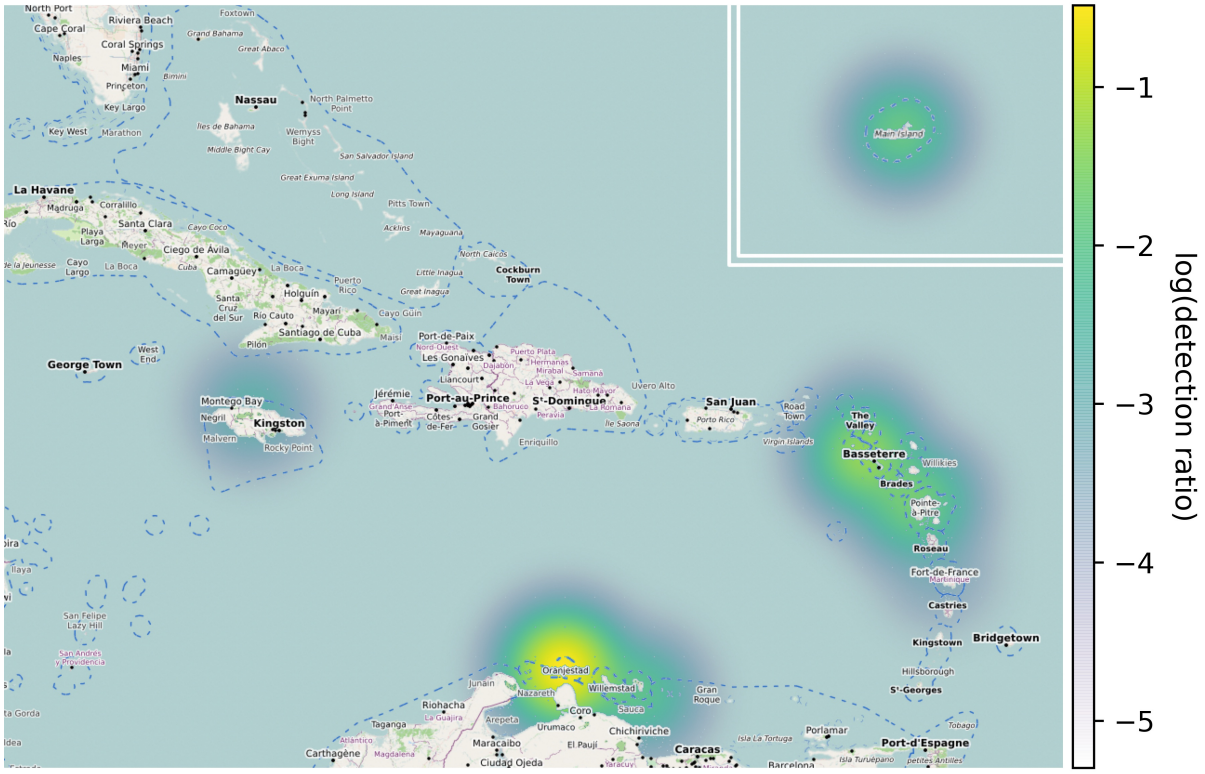


Figure 29: Repartition of Dolphins voicing detection from December 2020 to March 2021.

Maps of species detection are being edited online when the whole dataset will have been received.

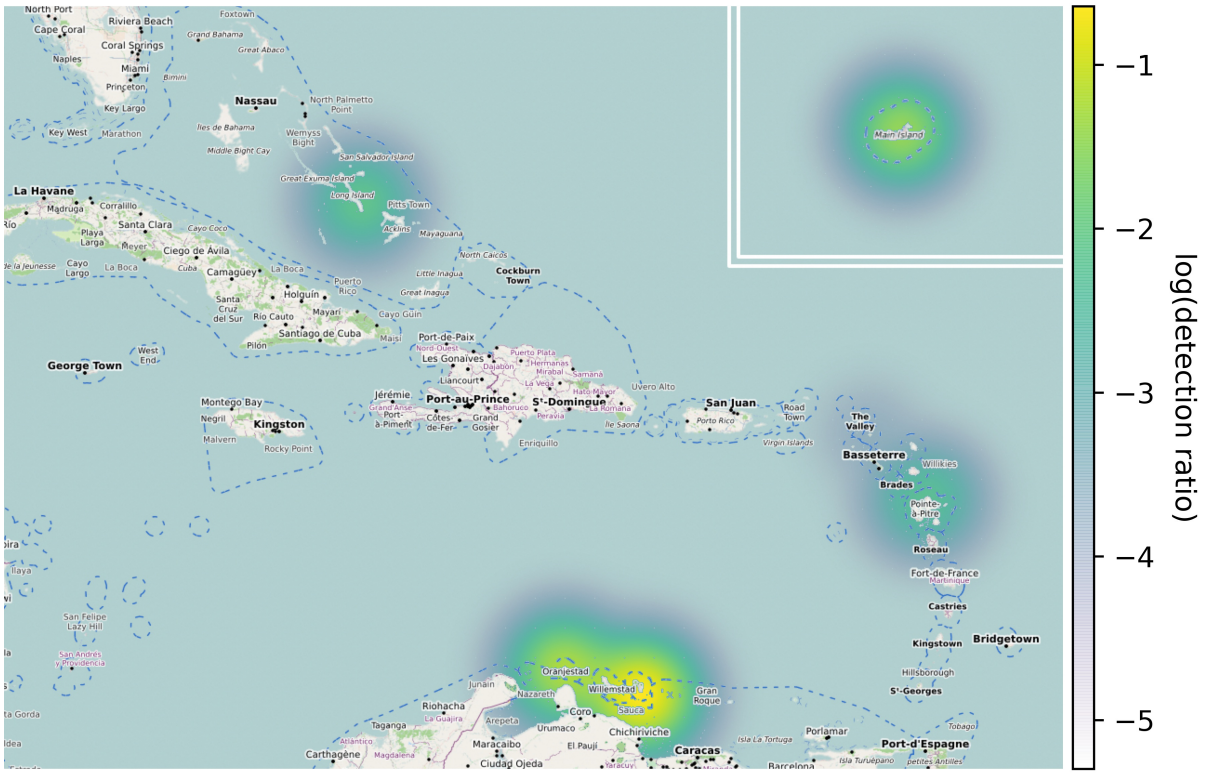


Figure 30: Repartition of Dolphins voicing detection from April to October 2021.

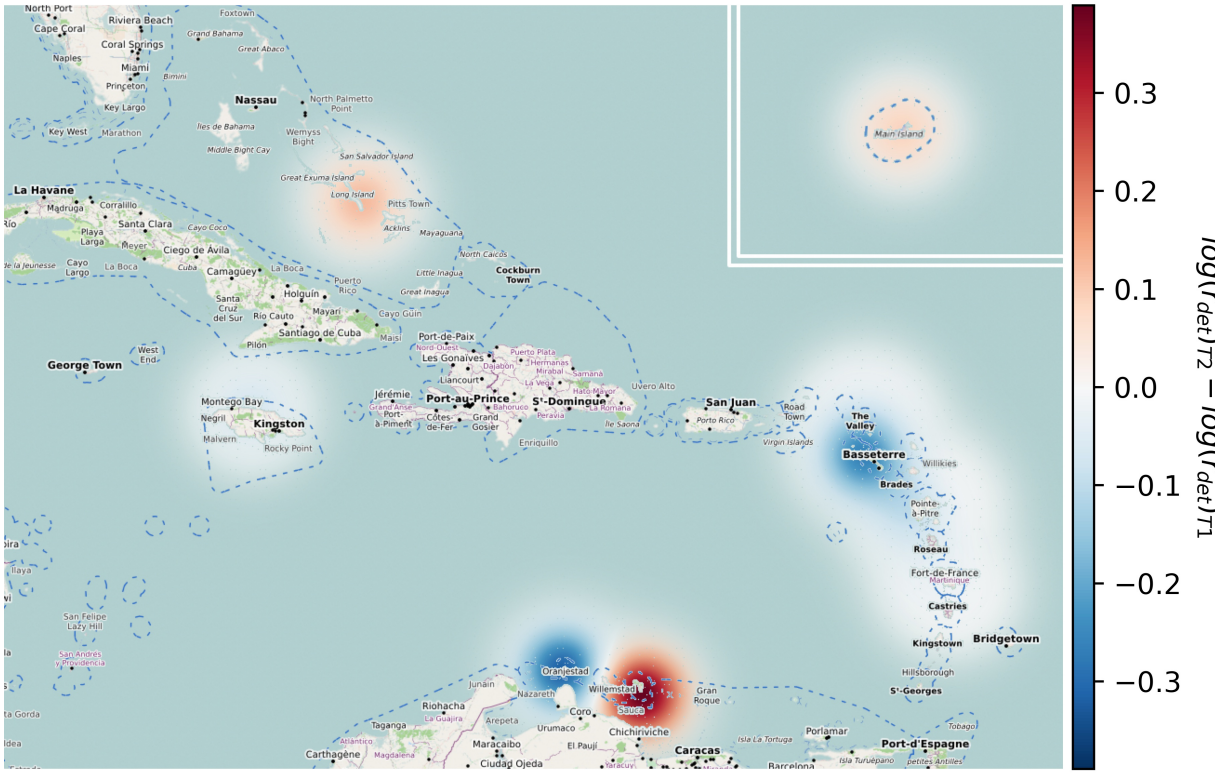


Figure 31: Difference of detection of Dolphins voicings between the 2 periods. Red means apparition and blue means disappearance.

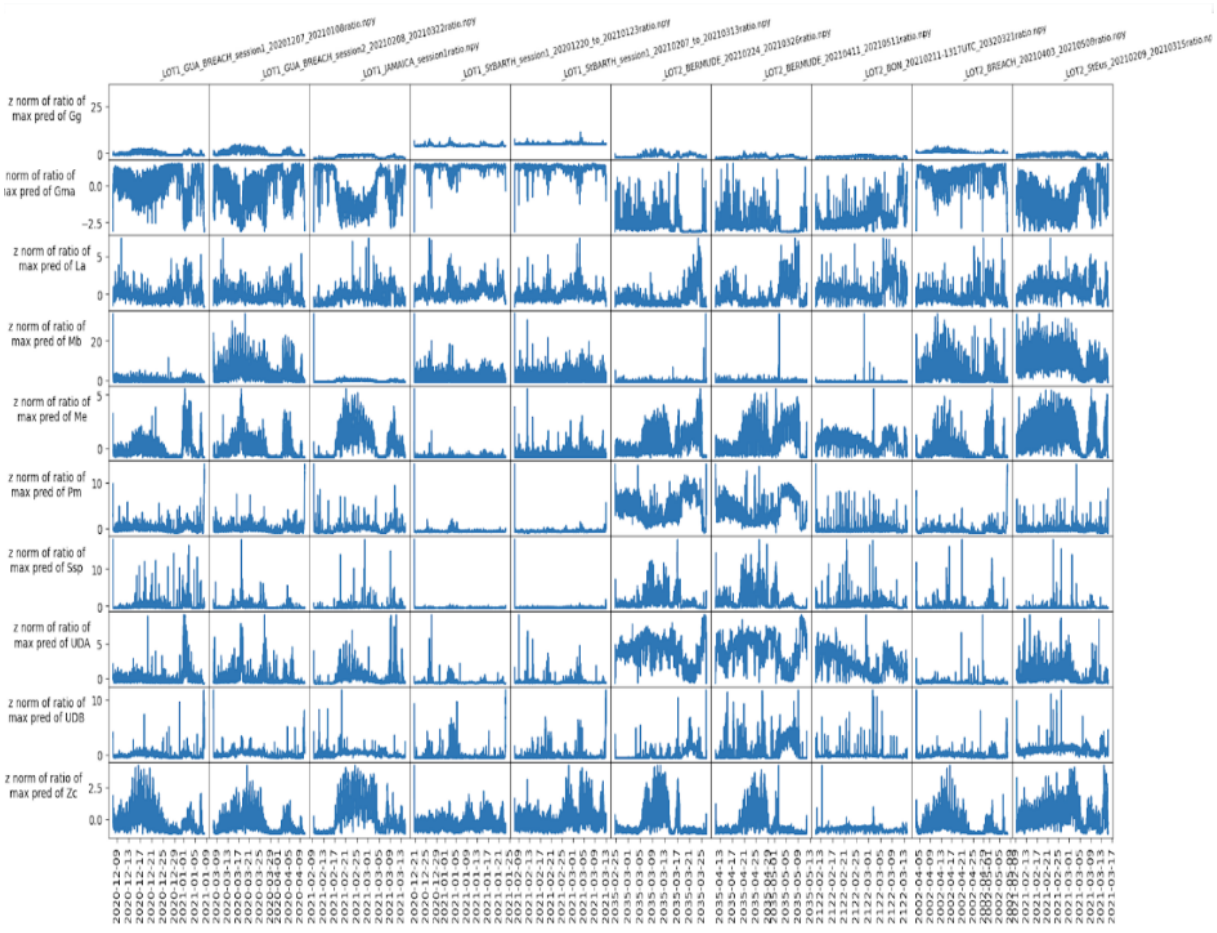


Figure 32: Percentiles of the 10 normalized logits of DOCC10 outputs over whole 10 sessions, showing the varying information in time and in session of the biosonar logits.

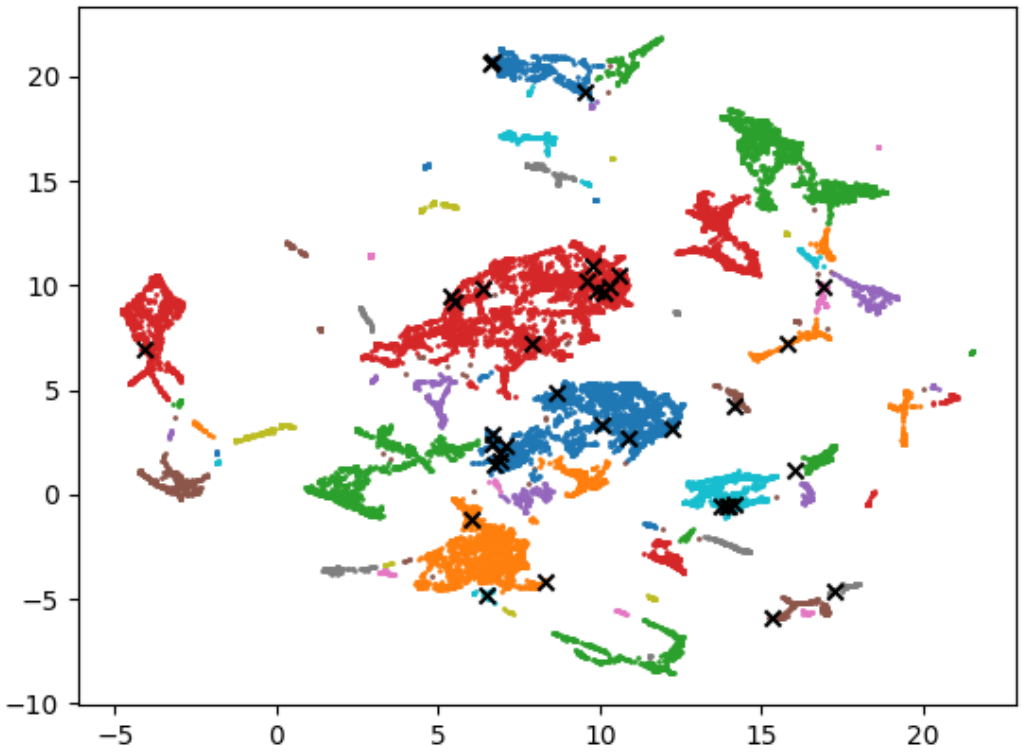


Figure 33: UMAP clustering in 2D over the 10 dimensions of the docc10 outputs, over all the files of three sessions : JAMAICA 20210 202107, BONAIRE 20210211 131700UTC 20320321 and StMARTIN 20210211 103000 20210324 151700. The (X) labels are the detected dolphin voices. Colors of the clusters are groups of similar dates. We then explored as in WP3.2 the files close to the ones including voicings to build a simple decision rule filtering the transient of the reef noise versus the biosonar.

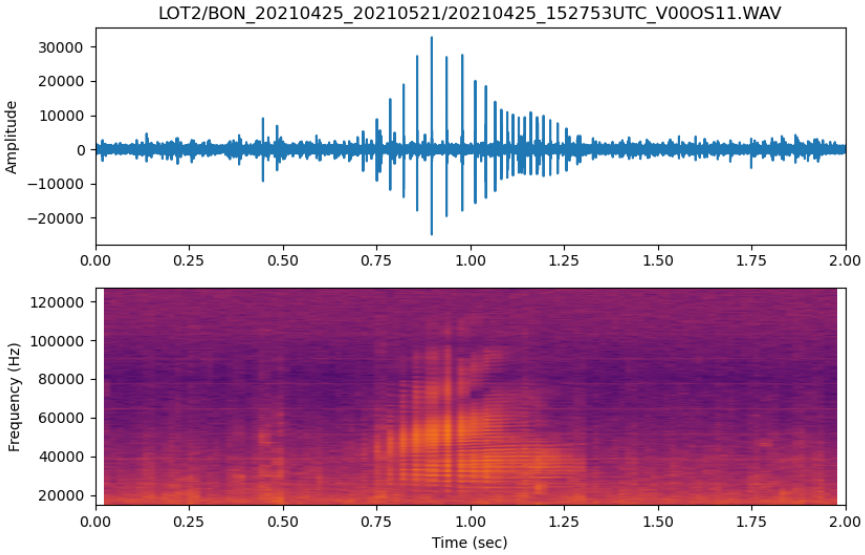


Figure 34: Example of a succession of clicks recorded in Bonaire.

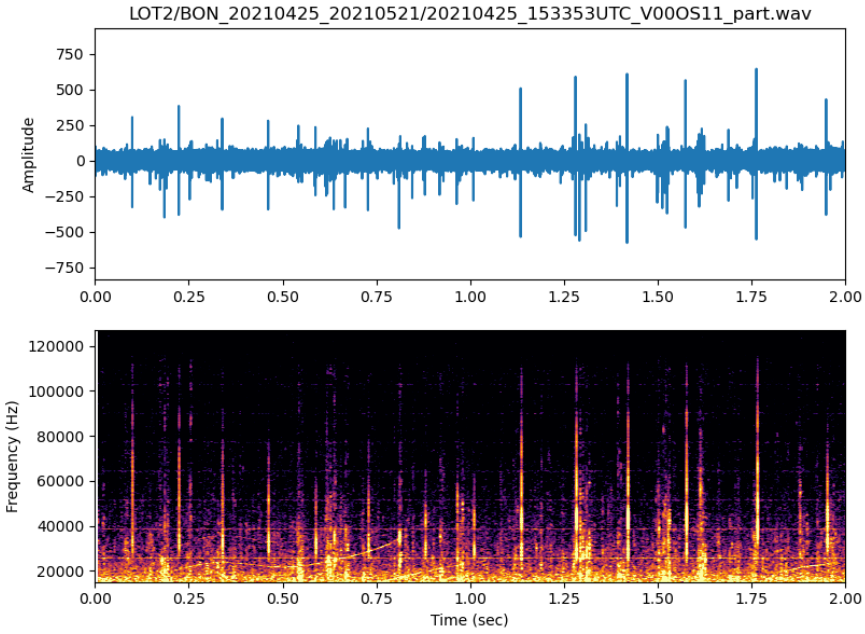


Figure 35: Example of a succession of clicks recorded in Bonaire.

5 Deviations

5.1 Data management of the recording to process

We received first recording in april / may 2021. Then most arrived in july august. This was later than expected. We received only a part of the expected recording. Last, 30% of the data base, is expected to arrive in october 2021. However we processed in 1 month 90% of the received data. We processed CNN at low frequency for 2 species, and high frequency for 10 species. We also work on denoising the noise of the reef. We processed 60% of the recording in wavelet transform to get complementary indices.

5.2 Data management of the reference recording

We received in late summer 2021 partial subset of the target species in reference recording from the responsible of this task in CARIMAM. Recordings with annotation of reference samples are still missing for 2/3 of the species. This is impacting our tasks, but we succeeded to run the first global architecture.

5.3 Machine learning task

The training of the model and transfer learning of our model is not fully implemented yet due to late and partial files sent from the project contractor responsible of recording and annotation of reference samples. However, from the preliminary training on two species, and the forward on ten odotoncetes, we demonstrate interesting preresults, that will build in next week new training set, crossing indices of wavelet and CNN representations.

5.4 Use of Resources

We had not issues on the ressource management. We have been invited by direction of CNRS to use CNRS HPC for wavelet transforms, high speed forward of the CNN and some correlations This allows to run in Jean-Zay national HPC in 2 weeks the process of the 1000 days of recordings.

6 Acknowledgments

We thank OFB for its trusts, Gerald Mannaerts (before Jeffrey Bernus). We thank all the collaborators over the whole observatory who did incredible recordings, despite the COVID. We received one by one 512 Go uSD card by snail mail in our physical box letter in La Garde. This was a challenging international work, we succeed. This mass of collected data is opening new horizons on the biodiversity of central Atlantic.

We thank PRNIA IN2SI program for its support on High Performance Computing, Audrey Monsimer, Gérald Dherbomez, and the CNRS INS2I direction staff Jamal Atif who allocated this support to this project.

This work on up to now 11 To of recordings, is a team work in Dyni LIS, that demanded long term and intense work of all the co-authors who collaborated in a great inspiration despite the number of issues that we tackled one by one and continue to.

References

Abeille, R., Doh, Y., Giraudet, P., Glotin, H., Prévot, J.-M., & Rabouy, C. (2014). Estimation robuste par acoustique passive de l'intervalle-inter-pulse des clics de physeter macrocephalus: méthode et application sur le parc national de Port-Cros. *Journal of the Scientific Reports of Port-Cros National Park*, *28*.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... others (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173–182).

Best, P., Marzetti, S., Poupard, M., Ferrari, M., Paris, S., Marxer, R., ... Glotin, H. (2020). Stereo to five-channels bombyx sonobuoys: from four years cetacean monitoring to real-time whale-ship anti-collision system. In *e-forum acusticum 2020.*

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Fairbrass, A. J., Firman, M., Williams, C., Brostow, G. J., Titheridge, H., & Jones, K. E. (2019). Citynet—deep learning tools for urban ecoacoustic assessment. *Methods in Ecology and Evolution*, *10*(2), 186–197.

Ferrari, M., Glotin, H., Marxer, R., & Asch, M. (2020, July). DOCC10: Open access dataset of marine mammal transient studies and end-to-end CNN classification. In *IJCNN.* Glasgow, United Kingdom. Retrieved from https://hal.archives-ouvertes.fr/hal-02866091

Ferrari, M., Poupard, M., Giraudet, P., Marxer, R., Prévot, J.-M., Soriano, T., & Glotin, H. (2019). Efficient artifacts filter by density-based clustering in long term 3d whale passive acoustic monitoring with five hydrophones fixed under an autonomous surface vehicle. In *Oceans 2019-marseille* (pp. 1–7).

Fonseca, E., Plakal, M., Ellis, D. P., Font, F., Favory, X., & Serra, X. (2019). Learning sound event classifiers from web audio with noisy labels. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 21–25).

Fukumori, K., Nguyen, H. T. T., Yoshida, N., & Tanaka, T. (2019). Fully data-driven convolutional filters with deep learning models for epileptic spike detection. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2772–2776).

Glotin, H., Caudal, F., & Giraudet, P. (2008). Whale cocktail party: real-time multiple tracking and signal analyses. *Canadian acoustics*, *36*(1), 139–145.

Grill, T., & Schlüter, J. (2017). Two convolutional neural networks for bird detection in audio signals. In *2017 25th european signal processing conference (eusipco)* (pp. 1764–1768).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Huang, J. J., & Leanos, J. J. A. (2018). Aclnet: efficient end-to-end audio classification cnn. *arXiv preprint arXiv:1811.06669*.

Jacobsen, J.-H., Oyallon, E., Mallat, S., & Smeulders, A. W. (2017). Multiscale hierarchical convolutional networks. *arXiv preprint arXiv:1703.04140*.

Kandia, V., & Stylianou, Y. (2006). Detection of sperm whale clicks based on the Teager–Kaiser energy operator. *Applied Acoustics*, *67*, 1144–1163.

Kim, Y., Yim, J., Yun, J., & Kim, J. (2019). Nlnl: Negative learning for noisy labels. In *Proceedings of the ieee international conference on computer vision* (pp. 101–110).

Kiranyaz, S., Ince, T., Abdeljaber, O., Avci, O., & Gabbouj, M. (2019). 1-d convolutional neural networks for signal processing applications. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 8360–8364).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Madsen, P.-T., Payne, R., Kristiansen, N., Wahlberg, M., Kerr, I., & Møhl, B. (2002). Sperm whale sound production studied with ultrasound time/depth-recording tags. *J. of Exp. Biology*, *205*(13), 1899–1906.

Poupard, M., Best, P., Schlüter, J., Symonds, H., Spong, P., & Glotin, H. (2019). *Large-scale unsupervised clustering of orca vocalizations: a model for describing orca communication systems* (Tech. Rep.). PeerJ Preprints.

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, *48*(2), 88–91.

Swartz, S., Clapham, P., Cole, T., Barlow, J., McDonald, M., Oleson, E., & Hildebrand, J. (2000). Locating and enumerating endangered humpback whales in the eastern caribbean with directional (difar) sonobuoys. *The Journal of the Acoustical Society of America*, *108*(5), 2540–2540.

Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., & Saurous, R. A. (2017). Trainable frontend for robust and far-field keyword spotting. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5670–5674).