June 26th 2022

Automatic classification of humpback whale *(Megaptera novaeangliae)* vocalization in the Caribbean

MASTER DEGREE REPORT

Student : **Stéphane Chavin**

*master **SDM** Interactions biotiques et Perturbations Anthropiques*

study year 2021-2022

duration of the internship: **6 months**
from **February** to the end of **July**

Laboratory supervisor : **Glotin Hervé** and **Poupard Marion** (LIS)

with P. Best, M. Ferrari and P. Mahé

University supervisor **: Molmeret Maëlle** *(Toulon University)*

UNIVERSITÉ DE TOULON

LIS LABORATOIRE D'INFORMATIQUE & SYSTÈMES

ACKNOWLEDGMENT

First of all, I would like to thank Mr. Glotin Hervé, teacher-researcher and head of the DYNI team at the LIS laboratory, and Mrs. Poupard Marion, post doc at the LIS laboratory, without whom this internship would not have taken place as well as the investors.

In fact, this research is co-financed by ANR-18-CE40-0014 SMILES and ANR-20-CHIA-0014-01 ADSIL national chair in AI for Bioacoustics.

A very big thank you to Paul Best, doctoral student in the DYNI team, for the great help he gave me on learning and perfecting my programming skills in the Python language, but also for providing me with his knowledge in machine learning.

In the same way, I would like to thank Maxence Ferrari for helping me solve so many computer and programming issues.

Thank you to the trainees with whom I have shared these last months, Pierangelo and Nicolas, for their good atmosphere and their advices.

A very special thanks to all the members of the Toulon LIS laboratory, for their technical support.

I would like to thank all the people who participated, directly or indirectly, in this internship and in particular the Port-Cros team for their logistical assistance during the installation of an underwater antenna in order to record sperm whales between Port-Cros and the Rayol area.

Finally, thank you to the teachers of the University of Toulon who did everything possible to ensure that the internships took place in the best conditions.

CONTENTS

TABLES AND FIGURES

Figure

Supplement figure

Supplement table

INTRODUCTION

Animal communication. The study of animal communication is an increasingly common science. Defined in several forms: visual, audible, chemical or touch, it is characterized by the emission by an individual (transmitter) of a signal towards another individual (receiver) generally of the same species (*Seyfarth and Cheney 2003, Bradbury and Vehrencamp 2011*). In nature, bees communicate with chemical signals, in the same way as ants are known to do with pheromones. However these also use dance as a visual method to communicate about food position, distance or new site to explore (*Khan et al. 2021*). In the marine environment, cephalopods have been extensively studied because of their changes in texture and color by contraction of the muscles of the dermis (*Hanlon et al. 1990*). Indeed it has often been shown that organisms communicate visually with predators, in particular with bright colors to signal toxicity and therefore avoid being hunted (*Blount et al. 2008*). In the oceans, fish and arthropods as example also communicate, but mainly by sound. Indeed the shrimps emit clicks with their claws (*Schmitz et al. 2000*) while fish produce grunts, with the rather well-known example of the grouper (*Bertucci et al. 2015*). However, one of the most studied songs remains on the terrestrial side with the birds with some papers dating from before the 20th century. The communication of birds has already been greatly studied even if many things remain to be discovered. It has thus been shown that the song of birds depends on the season (*Hiett and Catchpole 1982*), but that it also evolves over time (*Fowler 1896*). In particular, it is possible to recognize a species of bird only by its song (*Stowell et al. 2016*). It was only from the end of the 20th century that researchers became interested in the communication of animals in the oceans and more particularly the communication of cetaceans, including the humpback whale (*Whitlow 2018*). Marine mammals, and more precisely the infra-order *Cetacea*, are composed of two main micro-orders: *Mysticetes*, also called baleen whales, composed of right whales (*Balaenidae*), rorquals (*Balaenopteridae*), gray whales (*Eschrichtiidae*) and finally the pygmy whales (*Neobalaenidae*), as well as the toothed whales (*Odontoceti*), made up of dolphins and killer whales (*Delphinidae*), sperm whales (*Physeteridae*), beaked whales (*Ziphiidae*) and other families such as *Phocoenidae*, *Lipotidae*, *Pontoporiidae*. Among the baleen whales belonging to the rorqual family is the humpback whale (*Megaptera novaeangliae*), a species of size approaching 14 meters in length and present in a large part of the oceans of the world. Much studied for its long migration between the feeding zone, located at the poles, and the breeding zone, in tropical waters (*Johnson et al. 2022*), this species is also studied for its communication and in particular the complexity of its song (*R. Payne and McVay 1971*). As the sound produce underwater and propagate up to 5 times as in air due to the physical conditions of the water, scientist are able to collect those complex sounds.

Humpback whale song and its complexity. Cetaceans are well known for their sounds produced in water. Three main forms of sound are thus produced: clicks, often used for echolocation, the same way as bats do (*Suthers and Fattu 1973*) for a purpose of predation, particularly studied in sperm whales (*Madsen et al. 2002*) as well as whistles and vocalizations, more often associated with socialization between several members of the same species (*R. Mujalli et al. 2014*). Talking about the humpback whales (*Megaptera novaeangliae*), most males, called singers, have been shown to be capable of long song sessions during the breeding season (*R. Payne and McVay 1971*). These sounds have been recorded and studied resulting in the discovery of a particular structure. Indeed, the basic unit of a song is called a unit, it is a single vocalization lasting less than 3 seconds and spaced on both sides by silence. Its frequency is relatively low compared to certain cetaceans because these are between 15 Hz to 4,000 Hz depending on the area (*Fournet, Szabo, and Mellinger 2015*). It has been shown that this low frequency induce the sounds produced to travel up to several thousand kilometers underwater, depending on the water conditions (*Fisher and Simmons 1977*). The repetition of a few units forms sub-sentences which are also repeated over

time (*R. Payne and McVay 1971*). This repetition forms sentences which, put end to end, form a song. A song generally lasts several tens of minutes while a singing session, corresponding to a succession of songs, can last up to several hours without interruption (*R. Payne and McVay 1971*). Many studies about these songs have been done in an attempt to understand its origin and perhaps even its "purpose". The sound of humpback whales would be generated by the vibration of laryngeal vocal cords. This sound will then resonate in the nasopharynx and the laryngeal sac and can also be modulated (*Adam et al. 2013*). Although many similarities were found between the song of birds and the song of humpback whales (*Mercado and Perazio 2022*), it would seem that the latter is in reality very different in its way of evolving. However, the most probable hypothesis concerning its purpose would be the same, that is to say for the reproduction (*K. Payne 2000*). It has been demonstrated that a humpback whale present in a new environment with other congeners from a different group adapts its song to this new group (*Mercado 2022*). This could mean that the song would in fact have a purpose of localization in space. During a study, it was noticed that over time the song of these humpback whales varied (*Figure 1*). Indeed, each year the songs that were recorded in the breeding areas were different and it then suggest that the song of one year in a given area could be found in another area the following year due to horizontal transmission between regions (*E. C. Garland et al. 2011*). In spite of that, the question of song transmission remains difficult to analyze. Many scientists are still trying today to demonstrate that singing is transmitted between individuals in the same region through learning, which would correspond to a real notion of culture (*E. Garland and Mcgregor 2020*), while others think that environmental variables would induce humpback whale to change song and copy itself for the reproduction. Thus, younger whales would simply copy the song of a model humpback whale by attempting to make more personal modifications and according to its genetic predispositions (*Mercado 2021*).

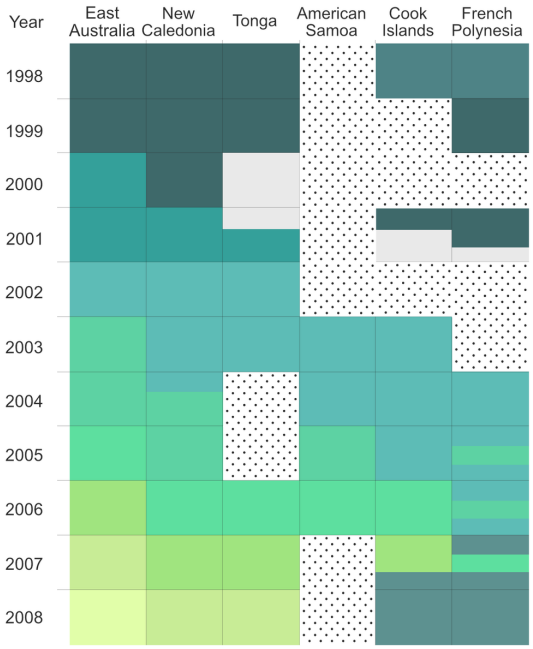


Figure 1. *Evolution of humpback whale song*

This figure show the evolution of humpback whale (Megaloptera novaeangliae) song, represented by a color, through time from 1998 to 2008 in 6 different regions. In this representation, we can see that the song evolve with a east to west gradient. Missing data are represented by dots filling case. (modified from Ellen C. Garland et al. 2011)

Deep neural network applications in bioacoustics. As stated above, humpback whales make a large migration during the year. Indeed, the feeding areas are located at the poles, where krill abounds due to currents and up wheeling, while the breeding areas are located in warm regions, at the level of the tropics. Thus, corridors are created with the main routes of these migrations (*Johnson et al. 2022*). The study of humpback whales is often complicated because of their way of life. In regions with extreme conditions for a large part of the year and capable of deep diving for long minutes (*Derville et al. 2020*), passive bioacoustics is a so-called non-invasive study method that has proven to be very effective. By capturing the signals emitted by the animal, it is indeed possible to know the species as well as its movement in the water column, and by analyzing the types of vocalizations, it is possible to deduce possible information regarding the origin of the individual and its comportement. However, there are different methods for processing acoustic signals. The first method is said to be manual by analysis and in particular listening to the recordings. This method

requires a lot of time but also a great ability to recognize the vocalizations of different species. In the study of certain taxa this method has long been used and is still a reference for the study of complex songs (*Rocha et al. 2015*). Over the past few years numerous studies have made it possible to collect hundreds or even thousands of hours of recording. Analyzing all of them manually would take many years. Whereas some automatic method was develop, the second method of analysis by computers appear. Called (semi) supervised method, with firstly a manual action then in a second time an automated action on machines. This automation can be done by training neural networks. Indeed, CNNs, for convolutional neural networks, are powerful computer tools capable of learning annotations by extracting features. In more common cases, such as the recognition of objects in photographs, the neural networks run through the image in two dimensions taking into account the colors, shapes, contrasts, then extract features according to different filters. The last learning layer allows the network to make a decision and therefore predict, with a more or less high confidence value, the object. Used on videos from underwater cameras to monitor fish populations for example (*Marini et al. 2018*), it can also be used on amateur images to detect plant species (*Lee et al. 2015*). This method can indeed be applied to bioacoustics. A spectrogram is a two-dimensional visualization of an audio recording. On this visualization we find frequencies on the y axis and time on the x axis, while the energy is expressed by a color gradient. Ergo, training a neural network on bioacoustic recordings is possible and as already been made. In this way, it is notably possible to detect cetacean songs and more particularly, it is possible to give the species of cetacean present in the recordings (*Poupard et al. 2022*). In the case of humpback whales, despite the fact that many studies still classify songs manually, some studies are moving towards automating the classification (*Heimlich et al. 2009*). Although sample annotation work for training can take some time, once well trained the model can work on hundreds of hours of recordings in just a few hours. From this, it is therefore entirely possible to envisage the creation of a computer model capable of detecting the vocalizations of humpback whales before classifying them by type and therefore reconstructing a song.

The CARI'MAM project. In this work, data from the Caribbean Marine Mammals Preservation Network (CARI'MAM) program were processed by focusing only on humpback whales. Indeed, this program funded by the European Union in the Caribbean Sea aims to develop the management of protected marine area for all marine mammals, but also to simplify their migration. Since the Caribbean is an important breeding area for humpback whales living in the North Atlantic (*Johnson et al. 2022*), the actors of this project are trying to preserve this area so that it remains that way. Indeed, this represents different challenges, in particular economic and ecological, especially since the important role of humpback whales in the global climate has recently been demonstrated, in particular by playing an important role in the life cycle of plankton (*Pershing et al. 2010*). The CARI'MAM project thus represents a network of shareholders bringing together the islands of Guadeloupe, Bermuda, Bahamas, Martinique, and its action with the LIS laboratory corresponds in particular to collecting numerous recordings from the seabed with the aim of identifying species present on the premises at different times of the year (*Glotin et al 2021*). The West Indies are an area of interest for the study of cetaceans and in particular the humpback whale. For several years, many researchers have been trying to record as many recordings as possible in this region of the world and this will allow the current recordings to be compared with those dating from several decades ago (*H. Winn and L. Winn 1978*).

Objectives. The objective of this work here is to automatically classify the vocalizations of humpback whales detected through hundreds of hours of recording. To reach this objectives, it will need to create a convolutional neural network which will have the task of automatically classifying. It will be first important to adapt a humpback whale detector and then manage to classify the different units.

https://www.car-spaw-rac.org/?Le-projet-CARIMAM

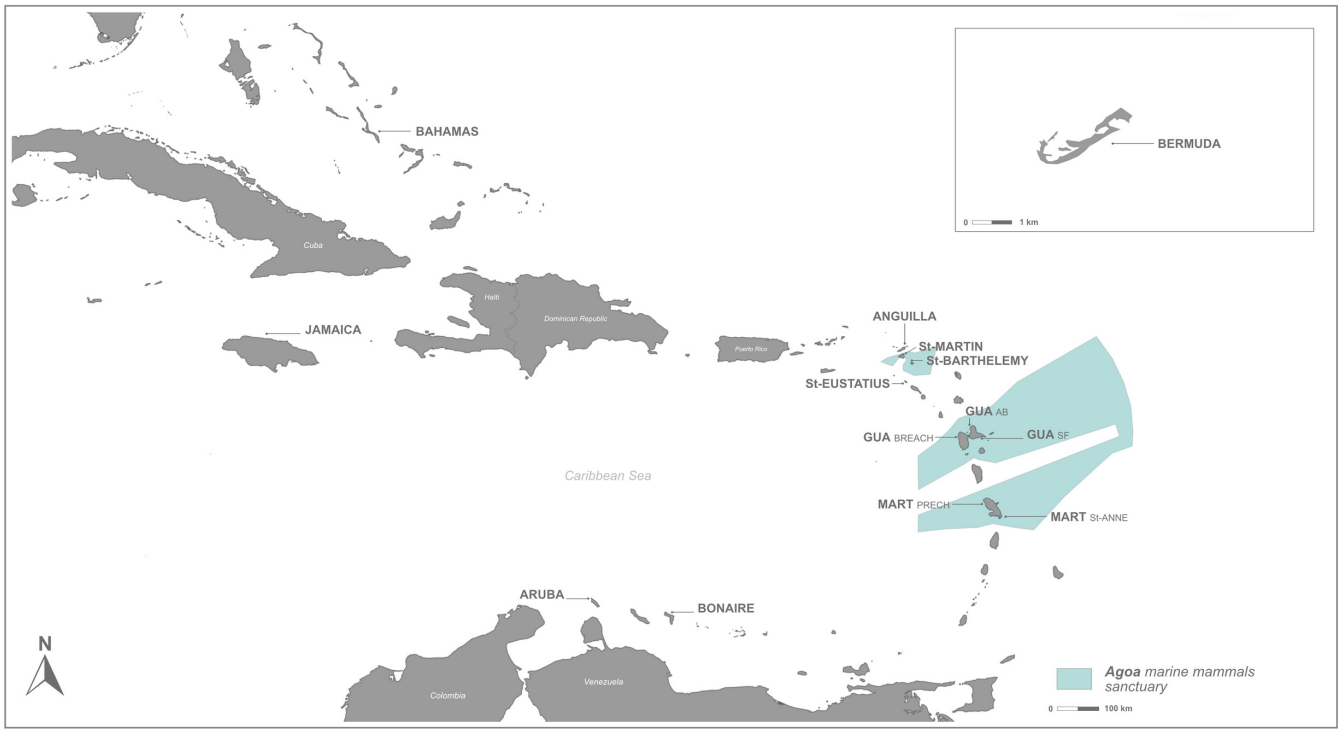


Figure 2. *Map of the Agoa sanctuary with all the recording stations*

In this map, we can see the different stations of the CARI'MAM project as GUA represent Guadeloupe and MART : Martinique. The Agoa marine mammals sanctuary is colored in blue.

MATERIAL AND METHODS

Data collection. The recordings were made in the Caribbean Sea on different stations in and out of the Agoa sanctuary. The Caribbean Sea is part of the Atlantic Ocean and is located east of Central America and has an area of 2,640,000 km ^{2}. Covering the entire exclusive economic zone of the French West Indies, i.e. Saint-Martin, Saint-Barthelemy, Guadeloupe and Martinique, the Agoa marine sanctuary, dedicated to marine mammals, was created in 2010. Breeding area for humpback whales but also an area where 35 different species of marine mammals live side by side including sperm whales (*Physeter macrocephalus*), pantropical spotted dolphin (*Stenella attenuata*), or short-finned pilot whale (*Globicephala macrorhynchus*), this marine protected area, located in the middle of the Caribbean, is a site of scientific interest. Regarding, the methodology followed by the CARI'MAM teams, initially the objective is to set up the hydrophones, then place them on site for 40 days before recovering them, make a copy of the files which will be then sent to the laboratory and finally repeat the operation. The hydrophones used for this mission was high-Frequency omnidirectional C75s connected to a HighBlue Mono recorder produced by SMIoT (*Scientific Microsystems for the Internet of Things*). The device has a total of 28 batteries and a storage of 512 Gb. The programming of the recordings has been made so that a 1 minute recording is triggered every 5 minutes, which allows the battery to be maintained for about thirty days while conserving storage space on the SD card. After choosing the site, the device is immersed in a 56 cm long sealed tube at a depth of about 20 m. The device is fixed to a mooring line, attached to the bottom by 20 kg lead weights acting as an anchor and held vertically by a surface buoy. The sampling frequency for the recordings was either 256 or 512 kHz. This high sampling frequency makes the recordings heavier in terms of storage capacity but makes it possible to obtain all the different sounds, from the very low frequency (*Reidenberg and Laitman 2007*) to the high frequency and also perturbation (*Wilcock et al. 2014*). In total, 15 different sites (*Figure 2*) were sampled and the data processed here concerns recordings from December 2020 to September 2021, which

https://www.univ-tln.fr/SMIoT.html

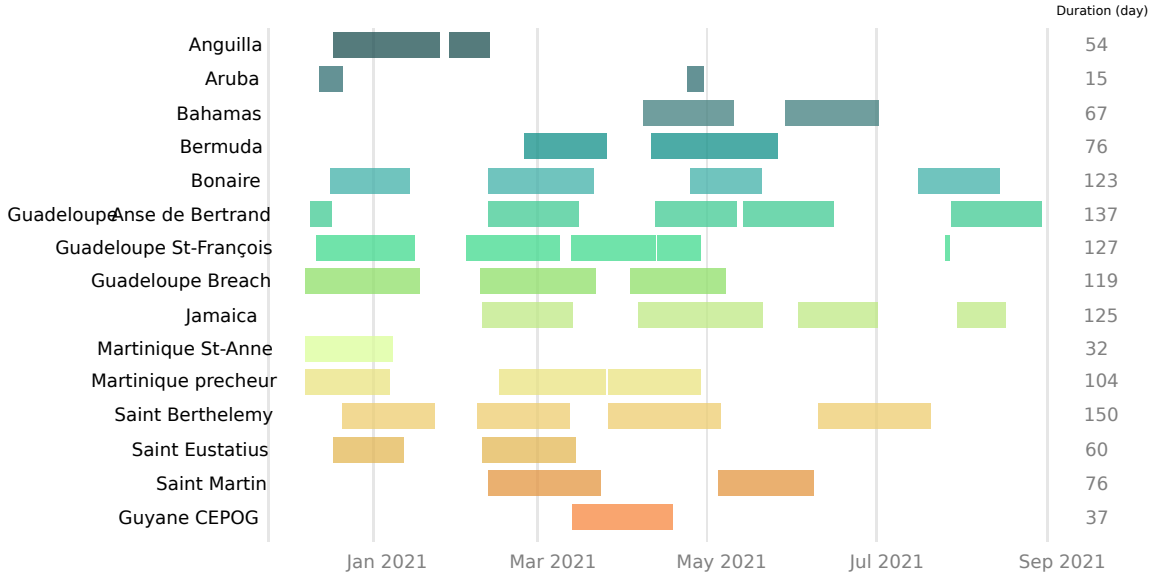


Figure 3. *Data collection dates organisation from December 2020 to September 2021*

This figure displays the recording effort per stations. In fact, bars represent the time in days corresponding to the recording time. Each color is assigned to a recording station. With 150 days, Saint Barthelemy is the first station regarding the recording duration while Aruba is the least one with only 15 days.

represents 1,302 days, i.e. more than 6,000 hours of recordings (*Figure 3*).

Creating a dataset. Since this work is based on neural network learning, it is first important to have annotated data. Indeed, as will be explained later in this report, neural networks need correctly labeled data to learn. As a result, there are several methods for creating a labeled dataset. First, the most common one corresponds to the annotation by hand of hundreds of vocalizations. In the case of this work, the hand annotation was done on Audacity. After having integrated the recording to be annotated in the audio processing software, a label is added to each vocalization. The result of this is a text file containing the position in time of the vocalization as well as its label. This method is easy but tedious. However, in the case of the classification of vocalizations this method can be complicated due to the complexity of defining whether or not one vocalization is very similar to another. The second method used in this work to create an annotated dataset was to use a tool for detecting regions of interest (ROI) in a spectrogram enable in the scikit-maad package in the Python language. To find out regions of interest in the spectrogram, the detection method starts by removing the background noise and then, with a double threshold technique, isolates the remaining regions of the spectrogram (see *Supp. Figure*). The result of this method is a file including all the ROIs of a recording with some parameters like minimum and maximum frequency as well as duration. This is a quicker and useful method in the case where the objective is not to classify but just to detect humpback whale vocalizations. In the case where obtaining a label for these annotations is expected, it is still possible to perform a dimensionality reduction and then a clustering on it, but in this study, this is not completely reliable. However, it is important when creating a dataset that it is representative of the data that will be forward. For humpback whales song, the most different vocalizations should be present in the dataset. Here, it was necessary to manually add several times vocalizations with low frequencies because they were not sufficiently represented in the dataset. During this work, these two methods were used in order to create a consistent set of data for training a humpback whale song detector. With the aim of increasingly improving the results of the training and therefore the efficiency of the model, the use of the predicted data during the different forwards was a way of improving the dataset. Actually, it firstly

example of recording from St Eustatius used for the study : http://sabiod.lis-lab.fr/pub/CARIMAM/SAMPLES/ST_EUS/

brings new annotations very quickly, after manual confirmation, and secondly it confirms the learning of the model because if the latter has correctly predicted a vocalization, by learning it, it will be able to predict others with even more confidence. This method was used when it was difficult to obtain a large number of values for training. After training the model with a small number of value, we run it on our data and correct its predictions by adding them to the training set. Data augmentation was also applied by adding noise to the recordings. Adding Brownian noise or by randomly removing parts of the spectrogram improve training on certain types of vocalization. The degradation of a recording by adding noise makes it possible to adapt the learning and make it focus on the vocalization itself and not the background noise.

Automatic detection of humpback whale vocalization. Automatic humpback whale vocalization detection was performed here with the use of a CNN. Indeed, trained on an annotated vocalization dataset, the neural network is normally able to detect vocalizations in a recording. To do this, an 8 convolution-layers network is set up based on *Schülter et al., (2017)*. As a reminder, a spectrogram is a time-frequency representation that shows how the spectral content of a signal varies with time, i.e. each time value is associated with frequencies values. Thus, as for the reading of an image, the convolution layer is characterized by the set of scalar products of the matrix of the nucleus, called kernel, and of the spectrogram on which the kernel moves horizontally and vertically. The organization of the model is done so that the first part is represented by a convolution layer followed by a batch normalization (*see Supp. Figure*), described as fixing the means and variances in inputs of each layer thus reducing the lag of the internal covariates (*Ioffe and Szegedy 2015*), then a activation function Leaky Rectified Linear Units (LeakyReLU) with negative slope value 0.01. The effect of this activation function is a non-linearity of the output due to the non-linearity of the input data. This therefore aims to improve the learning of the neural network (*Maas, Hannun, Ng, et al. 2013*). Then, the second convolution layer is followed by a batch norm then is marked by the presence of a layer named bi-dimensional maximum pooling layer (Maxpool2d) with the kernel size and the number of strides as parameters. This one aims to down-sample the input by taking only the highest value present in the kernel which moves along the input according to the value of stride. The first layer of Maxpool2d has a kernel size of 3x3 and a stride of 1 while the next one, after three convolutions has a size of 1x3 accompanied by a stride of 1x3 as well. This is done so that the input has become a line and no longer a multidimensional matrix. Thus, after the latter, the kernel sizes of the 3 following convolutions are set to 1x9 for the first one then 1x1 for the last two. Finally, the dropout layer, present at different places in the model after the activation function in particular, with a value of 0.5, has the effect of randomly zeroing some of the elements of the input. This method allows a better learning, avoid over-fitting and allows to accelerate the process (*Srivastava et al. 2014*).

Regarding the processing of the recordings in the case of the detection of humpback whale vocalizations, the choice turned to a so-called short-term Fourier transform (*J. Allen 1977*) seen as STFT, with a window size of 2,048 samples and a value of 64 samples between each transform (= hop size). While it is common to add a Mel-frequency cepstrum filter for spectrogram processing, which will calculate and change the frequency scale, it was not used here due to poor efficiency. Then, a logarithmic scale is applied to the spectrogram in order to enlarge the small frequency values without over-compressing the high frequencies. Indeed in the recordings, the song of humpback whales is in the range of 100 to 2,500 Hz and the recordings can go up to 512,000 Hz so in this case only the low frequencies are interesting. The dataset used includes 3,156 annotations, with at least 2,056 positives and 1,105 negatives (corresponding to noises such as boats, noises from sound card, etc). The objective of the negatives is that the model does not predict vocalizations only because it detect a sound on the spectrogram. So by annotating negatives, here CARI'MAM noises, it show the model what not to consider as humpback

whale vocalization. The data set is then divided into a training set and a test set. We therefore find a ratio of 2:1 between true positives and true negatives as well as a ration of approximately 9:1 between training and test. Annotations are grouped in a file including the path of the recording, the name of the recording and the position of the annotation (as well as the label : if it is a humpback whale or not). The recordings are then re-sampled at 11,025 Hz and a window is created 1.5 seconds on each side of the annotation, thus forming a spectrogram of 3 seconds with a frequency of 11,025Hz. During the training of the model on machines containing high-performance GPUs, a Stochastic Gradient Descent (SGD) optimizer (*Sutskever et al. 2013*) is set with a 0.9 momentum and a progressive learning rate is implemented. Indeed the latter has a great importance in the good learning of a model. As shown in the figure below (*Figure 4*) the loss is represented as a curve and we can see the learning steps. With too low value, the time to reach the local minimum of the cost function, i.e. the most optimized learning, will be very important, while if the step is too high, the latter will never be reached. By fixing a progressive learning, starting from 0.005 and following the function $lr = lr.0.9^{x}$ with x the number of epoch, the local minimum is reached more quickly. In the case of the detection of vocalizations, the metrics to measure the good functioning of the model are the average accuracy score and the value of the area under the ROC curve, noted AUC, i.e. the area under the curve of the rates of true positives (fraction of positives that are actually detected) versus false positive rate (fraction of negatives that are incorrectly detected). It is from these results for training that we can know the level of performance of a detector. With a kernel size of 3, a batch size of 10 and 25 epoch, the training loop takes around 10 minutes to complete. Once good values have been obtained, the test set is also passed through the model to see if the model is good on values that it has never seen and therefore knew which it could not train. After this step, the weights of the model are saved and then forwarded on all the data in order to obtain the predictions on all the records. Subsequently, after having obtained the lists of 1,038 values corresponding to the predictions for all the recordings, a threshold of 0.3 is set in order to only take into account the values above this threshold, thus potentially being humpback whales. A threshold that is too low will have the impact of detecting too much background noise while a threshold that is too high will pass through a lot of vocalizations. With about 10 predictions per second, however, only take the maximum value. In addition, since it is assumed that a vocalization generally lasts 1 second, the detector must not detect two vocalizations in the same second. For this, a distance is also added between each detection of 9 predictions, i.e. approximately 1 second. The result of this is a file containing, in the same way as the annotation file, a column corresponding to the path and one to the name of the recording, then a column corresponding to the confidence value of the detection (included between 0 and 1) and finally its position in the record.

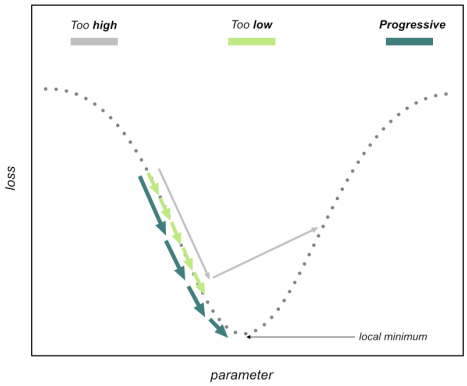


Figure 4. *Learning rate value impact on model training*

This figure show how the learning rate can induce an optimized model learning. With a too low learning rate, in green,, the optimization will take a long time, whereas a too high learning rate value, in grey, won't let the model reach a local minimum. In this work, a progressive learning rate, in blue, is used.

Dimensionality reduction. In machine learning, dimension reduction is very often used. Indeed when working with a high dimensions data, as well as humpback whale recordings like in this study, it is often preferable to reduce the dimensions of the data in order to process them more easily. Here at first, the use of an autoencoder made it possible to go from 128 x 128 input dimensions to a 16 dimensions' one. Indeed, an autoencoder aims to receive an input, process

it and reconstruct it, as best as possible, by reducing the dimensions. An autoencoder is a 2 parts unsupervised model: first part corresponds to the encoder, which will aim to transcribe the input then followed but a compressor will form a bottleneck (*see Supp. Figure*), while the last part corresponds to a decompressor and then the decoder which will aim to best recreate the input learned with the number of dimensions chosen, here 16. The learning rate used is once again progressive with a maximum of 0.003. The training is performed with a 64 sample batch size, on the previously predicted data. The metrics tracked are loss only. Once the training is good enough, all the data are passed through the autoencoder. While the detection have a 16 dimensions shape, they can therefore be analyzed in order to be classified by type of vocalization. To do this, a second method of dimensionality reduction is performed. Called *Uniform Manifold Approximation and Projection* (*McInnes, Healy, and Melville 2018*), this visualization method makes it possible to display the projections of detections in a two-dimensional space. Indeed a particular algorithm, UMAP preserves both the local structure and most of the global structure of the data which means that a projection of the data in space would make it possible to identify real distances between these data. In this work, the UMAP method was preferred on the t-sne because of the preservation of the global structure. Indeed by going from a 128 x 128 to a 16 dimensions, a lot of information has already been lost, so it is important to preserve the global and local structure of the different points for the classification step. The minimum distance (min-dist) value chosen for the UMAP representation is 0.0 as we are trying to classify and the number of neighbors (n-neighbors) was 24.

Classifier and sequences detector. As just stated above, the final step is the classification of vocalizations according to the different types present in the recordings. However, it's not easy to tell the difference with the naked eye without being used to recognizing the different units. So, a CNN was set up with a classifier role. In order to create the dataset for training the CNN, the UMAP method presented above was used to perform clustering. Given that we have a projection of the data in two dimensions, the clustering method HDBSCAN (*High Density-Based Spatial Clustering of Applications with Noise*) can be used to separate the groups formed on the projection. HDBSCAN clustering is based on two parameters: min-cluster-size and min-samples. Unlike the most common method DBSCAN which uses a fixed radius value around a point (core), here the radius increases according to the number of neighbors to be found (min-cluster-size). A value of 650 has been applied, while the value of min-sample has been set to 50. The value of min-sample has very little impact on the result. The most important is to choose an adapted value for min-cluster-size, based on the projection. In the case of min-cluster-size = 650, a circle will be drawn around the point so as to take at least 650 points around. For a second point, the principle will be the same then a calculation of distance called mutual accessibility is carried out according to the equation:

$$d_{\mathrm{mreach}-k}(a,b) = \max\{\mathrm{core}_k(a), \mathrm{core}_k(b), d(a,b)\}$$

where d(a,b) is the original metric distance between a and b. Thus, the points having a low accessibility distance from each other will be grouped into a cluster. However, it is possible that vocalizations are so close that they are not differentiated by clustering. During this work, it was therefore necessary to check all the clusters by hand and modify the errors. Subsequently, some clusters were added manually and by a Euclidean distance measurement method. Indeed, the annotation base of certain types, not having been considered in the clustering, was either created or amplified by taking a model vocalization and then calculating the 200 closest vocalizations in terms of Euclidean distance. A manual check was performed at each step to ultimately obtain a 5,886 annotations dataset. Since the detector has already passed over the data, it is not important to add true negatives because the detections are normally all vocalizations to be classified. Thus, a training set of 4,723 annotations was created, while the test set is 1,163 vocalizations. For the classifier dataset, an importance was given to the origin of the data.

Indeed, because the recordings come from several different stations, it is important that the classifier can classify the vocalizations of all these origins. The majority of the training data thus comes from recordings from Guadeloupe (3,026) and the test was carried out on all the islands. In addition, the test must be balanced on the types in training. For this, it is important that the number of vocalizations present in the training is not less than 250 but also not more than 450 (*see Supp. Table*). Speaking of an unbalanced game, a type of vocalization can be learned much more to the detriment of another and this will lead to an imbalance of the predictions. The CNN used for the classification is different from the one for detection. Instead of returning a list of values for each recording, the latter returns a type of vocalization for each detection. Moreover, even if the starting value of the learning rate and the evolutionary character is preserved, we are looking for a classifier and therefore the neural network must be trained to recognize objects etc. For this, the present CNN use a pre-trained Resnet18 network with a kernel size of 7, a stride of 2 and a padding of 3. The metrics selected to monitor its performance are accuracy and the F1 curve. Once again, after good training and test results, the model is forward on all detections. Finally, while we know the date of the recordings, the position in time of the vocalizations and their types, it is possible to determine the sequences, corresponding to the successions of vocalizations in the same recoding. By associating the date and time of the recording with the list of classified detections, by configuring at least 5 vocalizations to consider that the 1-minute recording includes a sequence, it is possible to obtain the sequences for the recordings containing songs.

Songs analysis. With the aim of detecting a possible evolution of the song over time or according to the different recording stations, a statistical method used in linguistics was used during this work. Indeed, given that the previously calculated results make it possible to obtain all the sequences recorded during the project, it is possible to process them in the form of n-grams. The idea of this method is to measure the occurrence of a repetition of n characters in a given sequence. In the case of the data of this work, a detected unit represents a character, assimilated to a word, and a recording represents a sequence, compared with a sentence when speaking of the linguistic study. After observing that humpback whales sing with a pattern of 3 or 6 to 7 units that they repeat over time, an n-gram of 7 and a one of 3 can be calculated. The number of different sequences is theoretically 7^{12} . After having calculated the occurrence of each of these sequences, if present in those detected, the values can therefore be compared and the sequences having the greatest occurrences can therefore be hypothetically considered as the most frequent. Given that a recording does not start at the beginning of the song but can sometimes start right in the middle of a song, it can be interesting to know the beginning of a song. To do this, in this work it was possible to take only the sequences having a first vocalization after 15 seconds of recording. Indeed it is very rare for a whale to pause for more than 15 seconds so if the first unit detected is after this time, it could mean the start of a song. Subsequently, by performing the n-grams method again, a typical starting sequence can be highlighted.

RESULTS

Humpback whale's distribution area in the Caribbean. The forward of all the data from the CARI'MAM project through the loop of the detector made it possible to detect a total of 1,026,235 vocalizations. As shown in the following figure (*Figure 5*), corresponding to 15 different stations for a period from January 2021 to September 2021, thus representing nearly 6,000 hours of recordings, once the data is normalized by the recording effort, i.e. the recording time in a station, we observe that 22.7% of the detections come from the Guadeloupe station Anse de Bertrand with more than 293,181 vocalizations, or more than 2,000 detections per day of recordings. In the list of the most active stations is also Bermuda with 14% of detections, Anguilla with 13.4%, Saint-Eustatius with 10.8% and Saint-Barthelemy with 10.7%. The station with the least detections corresponds to Aruba with only 51 detections.

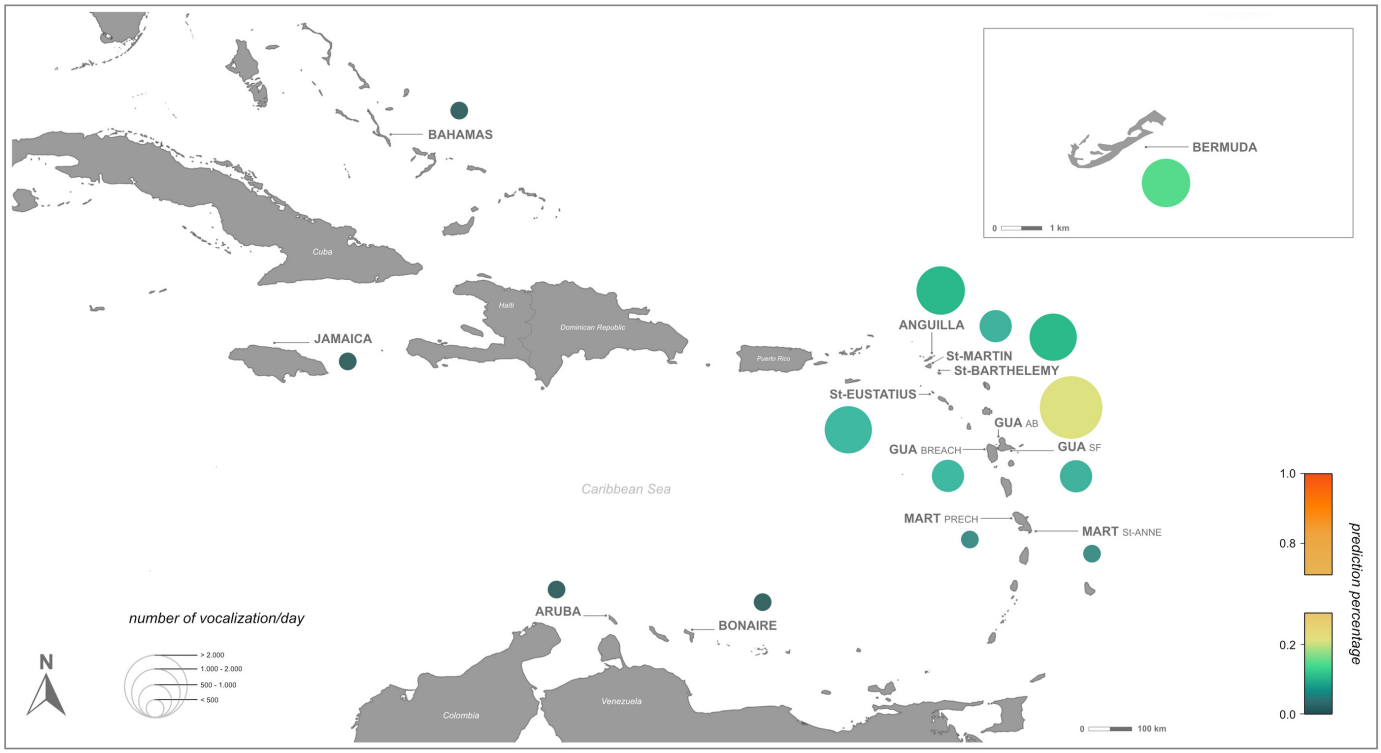


Figure 5. *Distribution area of humpback whale in the Caribbean*

In this map, the number of vocalization and the proportions are shown for the different stations. The data were normalized on the number of hours of recording to express the real proportions. GUA represent Guadeloupe and MART : Martinique.

However, when compared to the recording effort, the Jamaica station, with 125 days of recordings, i.e. 31 times more than Aruba (15 days only), Jamaica is last with less than 3 detections per day. When the detector detects very few vocalizations as in the present case, it is possible to consider a total absence of humpback whale songs in the recordings. Indeed, it is possible that the detector detected sounds that did not actually come from whales. As far as detection is concerned, with a training loss value of 0.01689, a mean Average Precision (mAP) of 0.9948, corresponding to the mean of the average precision (AP) of the n different classes :

$$mAP = \frac{1}{n}\sum_{k=n}^{k=1} AP_k$$

and a test AUC (*Area Under Curve*) reaching 0.9886 in just 7 epoch, it is possible to say that the learning of the training data went well and that the results are reliable.

Highlighting 12 different call types. As the objectives was to automatically classify humpback whales vocalization, it was necessary to determine which were the different units present in the recordings. To do so, after using the autoencoder to reduce the dimensions of the annotations, from 128 x 128 to 16 dimensions, it was possible to perform a UMAP map of the detections. After applying the HDBSCAN algorithm in order to bring out clusters, the figure presented (*Figure 6*) highlights the presence of 12 distinct clusters. These clusters were then extracted to verify the correct classification and then they were displayed. As shown in *Figure 7*, vocalisations have different criteria that can be used for identification: First we find the duration, some vocalizations can last more than 3 seconds ("long-moan"), while others are very short and less than 1 second ("yawp"); then, the second characteristic is the average frequency. As you can see on the different units, some are low in frequency ("low-freq", "growl") while others are slightly higher ("oop"); finally, one of the last identification parameters corresponds to the amplitude of the vocalization. While some start at low frequency and still end at low frequency without having made a significant fluctuation ("growl"), others start at low frequency and rise directly to almost 1 kHz ("dropplet"). To finish, the orientation of this amplitude can also be taken into account. When we look at the "whup" vocalization, we observe that the

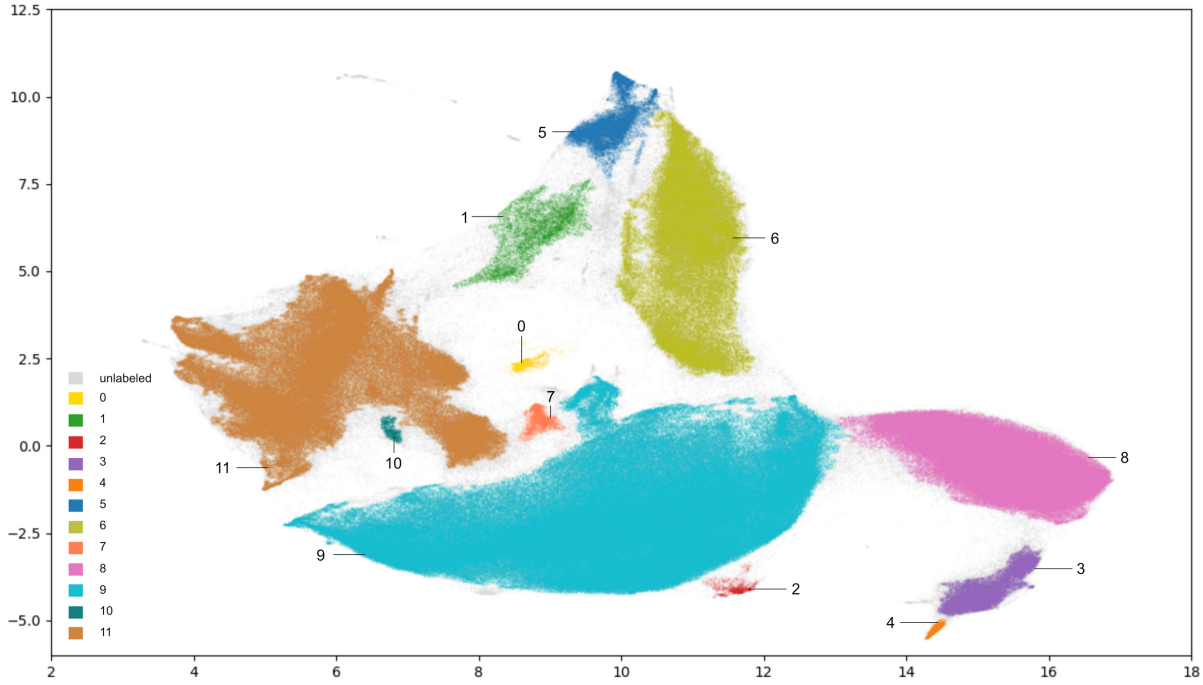


Figure 6. HDBSCAN clustering after Dimensionnality reduction

The umap was made on 958.079 vocalizations. UMAP parameter from the umap package are : min_dist = 0.0 and n_neighbors = 24 while HDBSCAN parameters are : min_cluster_size=650, min_samples=60. The unlabeled/noise detections are colored in grey while the 12 different clusters detected are displayed with colors.

frequency decreases over time while the frequency of the “low-freq” vocalization increases. These different characteristics make it possible to visually determine whether or not a vocalization is indeed part of a cluster. During this work, a particular effort was put on the detection of "swop" being an intermediary in terms of duration, amplitude and frequency between a "dropplet" and a "growl". When analyzing the origin and occurrence of these units in the different recordings, we saw that an unequal distribution can be observed (*Figure 8*). While "dropplet" represents nearly 20% of the units detected in the Martinique Saint-Anne station, the latter are absent from the recordings of Saint-Eustatius. Moreover, some units such as "fluct-3" are almost exclusively detected in the recordings of the Anguilla and Martinique Saint-Anne station. More generally, we can see that the "yawp" represent a significant proportion of the units detected in the majority of the stations with even nearly 60% on the Saint-Barthelemy station. Finally, in the two stations with the most detections, as seen previously (Guadeloupe Anse Bertrand and Bermuda), it would seem that the proportions between units present are similar (*see Supp. Table*) with, range from the most common to the least one, 21% ± 1.57 for the "fluct-2", 20% ± 0.48 of “fluct-1” and 20% ± 2.12 of “yawp”, then 9% ± 1.03 of “teepee”, 7% ± 1.26 of “whup”, and 6% ± 2.64 of “low-freq”, 5% ± 0.81 of “long-moan” , 4% ± 1 "dropple"t, 3% ± 0.27 "oop", 2% ± 0.19 "swop" and finally 1% ± 0.04 "growl". These are then extracted to create the train dataset for the classification.

Average classification scores. Once the classifier was trained to recognize the units presented previously and after 24 epoch and a training loss of 0.04, the test metrics showed an F1 value of 0.82 and an accuracy of 0.83. The accuracy is defined by the equation:

$$Accuracy = \frac{number\ of\ correct\ prediction}{number\ of\ prediction}$$

Here, with a value of 0.83, this means that during the test loops, 83% of the vocalizations to be classified were well classified. This gives a good indication on the learning of the model. In addition, the F1 curve is a metric combining both the precision seen previously and the recall, corresponding to the equation:

$$Recall = \frac{true\_positives}{true\_positives + false\_negatives}$$

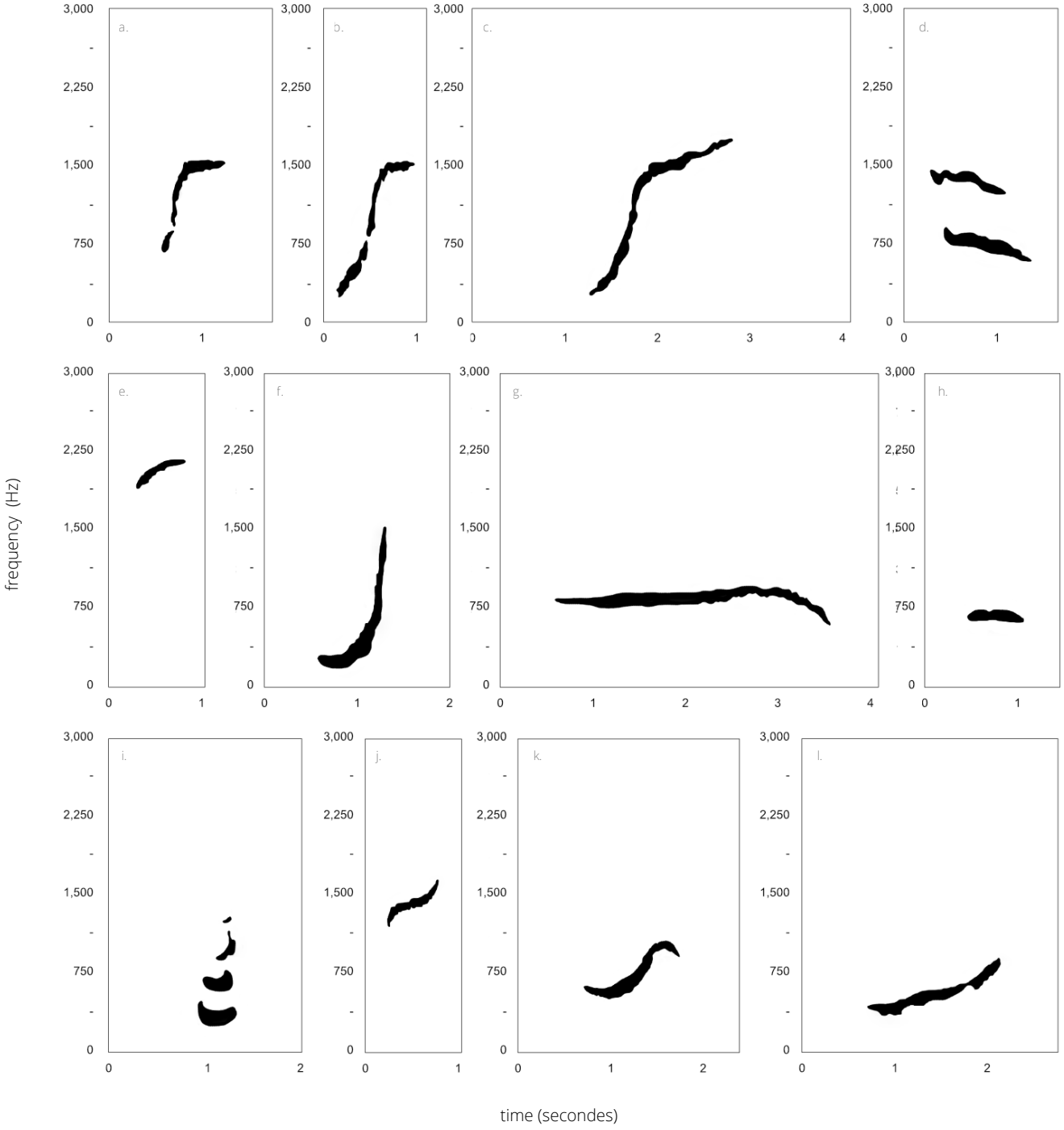


Figure 7. Caribbean Humpback whale's repertories

This figure show the 12 different call types that was detected in this work. These calls are named : (a) fluct_1; (b) fluct_2; (c) fluct_3; (d) whup; (e) oop; (f) droplet; (g) long_moan; (h) growl; (i) swop; (j) yawp; (k) teepee; (l) low_freq. Each spectrogram is represented with the frequency in the y axis and the time in the x axis. The longest vocalize is long_moan with more than 3 second wherase the shortest is growl and oop.
This figure as been made by extracting one units example on the forward data and then removing the background of the spectrogram and clearing the shape of the vocalize in Adobe Photoshop software.

Where as a reminder precision is :

$$Precision = \frac{true\_positives}{true\_positives + false\_positives}$$

Thus, F1 is calculated as :

$$F1score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

This signify that the mean of each classes precision and recall are high, then the F1 score will be high, while if one of the two parameters is low, then the score will be low. Finally, if both parameters are low, then the F1 score will be very low. Here, with a value of 0.82, we can therefore say that the two parameters are high, which means good learning of the model with high recall value and high precision value too. Thus, by displaying the confusion matrix *(Figure 9)* of the model in order to see the possible learning confusions, we can see that the lowest value is 0.64 for "fluct-2" while the maximum value has been reached for "growl" and "fluct -1". Indeed, the confusion matrix (*Figure 9*) provides information on the number of predictions of a labels compared to the real labels. In this way, we can therefore see that despite the fact that "oop" is well classified at around 77%, the model predicted a "yawp" in 23% of the cases. This figure is interesting because the errors highlighted are more often explainable. Indeed "low-freq" is well classified with a value of 79% but we can see that 19% of the predictions were considered as "long-moans". When we look at the figure of the different types of vocalizations (Figure 7), we see that the two units are quite similar in terms of duration and frequency, which can impact the learning of the model. The confusion between "fluct-2" (64%) with "fluct-1" (12%) and "fluct-3" (16%) can also be explained in the same way. In general, we can therefore see that on the data with a difficult test, the results are satisfactory. In order to confirm these, when the model was forward on the integrability of the detections, 200 spectrograms of each class were extracted to visually confirm the correct classification. By looking at the *Figure 10*, with the visual results of the classification, the figures represented correspond to the accuracy, i.e. the rate of good classification on the number of classified vocalizations. We therefore observe that the average rate is 73.6%, against 82% on the test data. Regarding the values per unit, we see that the minimum is 45% for the "swop" and the maximum 94% for low-freq. The "growls", which were well classified at 100% during the test, are now classified at 87% and the same for "fluct-3" which here drops to 80%. The "yawp", which were described earlier as the most

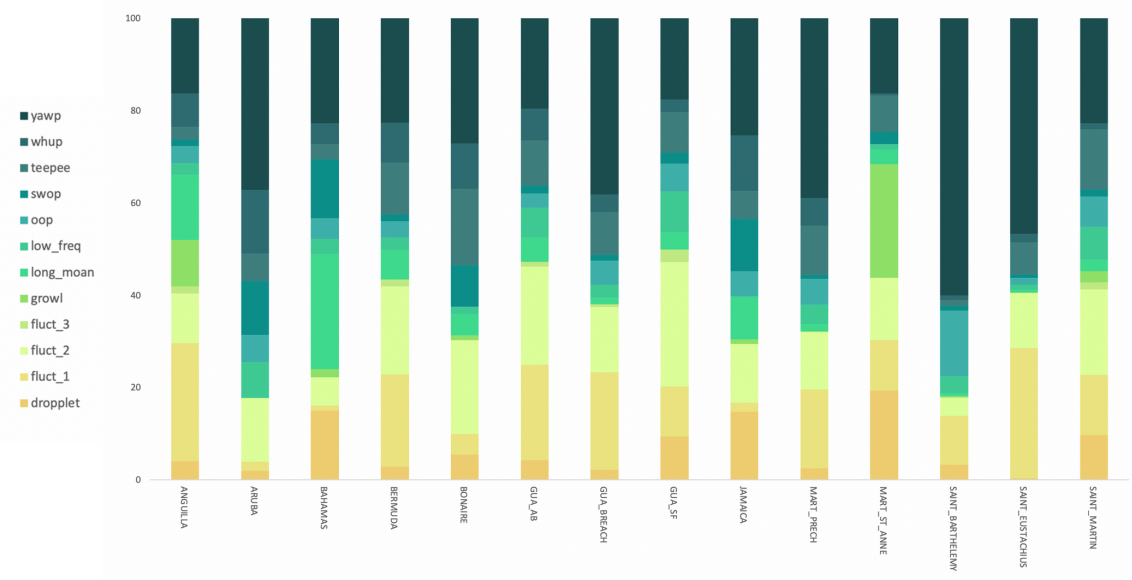


Figure 8. *Proportion of each unit in the recordings*

This figure is showing the proportion per recording station of each units. These were calculated by simply taking the number of vocalizations from one unit type in a station compared to the total number of vocalizations in that same station. Each stations are represented in the x axis.

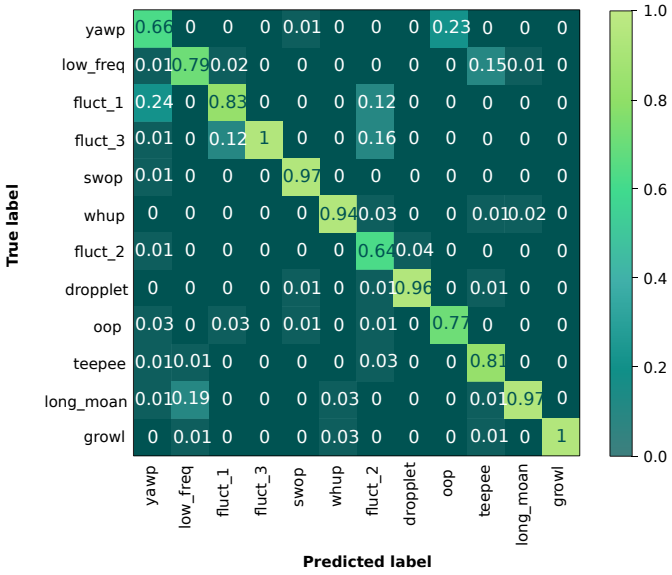


Figure 9. Test set confusion matrix

This figure show the results of confusion between labels just after the training and after 8 epoch. The color bar gradient express the value of good classification while 1 mean that 100% of the test annotations are well classified. The matrix is normalized on the predicted label columns.

common vocalizations in the recordings, are well classified at 73%.

An evolution in the song structure. As stated in the method, after having classified all the detections, a reconstruction of the sequences was attempted. As a result, a total of 48,267 sequences could be highlighted on all the recordings of the CARI'MAM project. Thus, if we represents 3 examples of sequences detected and classified during this work (*Figure 11*) we see on the first sequence a repetition of 11 dropplet types vocalizations then followed by the changing to the swop type. On the second part we can see a sequence composed with the same type of vocalization, named "oop", with a repetition of 4 and then 2 surrounded by long time pauses Finally, the last example shows a sequence characterized by the presence of 2 "whup" followed by a "fluct-2". This pattern is repeated 6 times in the recording despite the fact that the model only detected 4 of them correctly. This visualization shows that the pause times between each block of vocalization are generally very similar and therefore highlights the precise structuring of

Figure 10. *Classifier accuracy*

This figure display the average accuracy for each type of vocalize on the forward data. These results have been calculated with a visual method by extracting 200 samples of each labels and then count the correct classification.

humpback whale song. Since it is now possible to build the list and composition of all the recorded sequences, it is now possible as suggested to implement the n-grams method. In such a manner, as shown (*Figure 12*), we can see that when we study the sequences of 7 successives vocalizations in Guadeloupe in February 2021, i.e. when the humpback whales arrived in the region, out of the 3 Guadeloupe stations together, the most common sequence is a "fluct-1" repetition. However, when we compare the values of February with those of April (a), close to the end of the breeding period, the general pattern is different and the sequence which seems to be the most observed corresponds to a succession of 7 dropplets, as in

Figure 11. *Examples of detected and classified sequences*

By taking 3 different sequences from 3 different recording, sequences can be displayed with frequency as the y axis and time as the x axis using NFFT = 1024 and noverlap = 512 and using Adobe Photoshop software to remove the background and color the units. "dropplet are represented in light green while "swop" are in blue, "oop" in orange, "whup" in yellow and "fluct_2" in green. Pauses, corresponding in zero detection, are represented in grey.

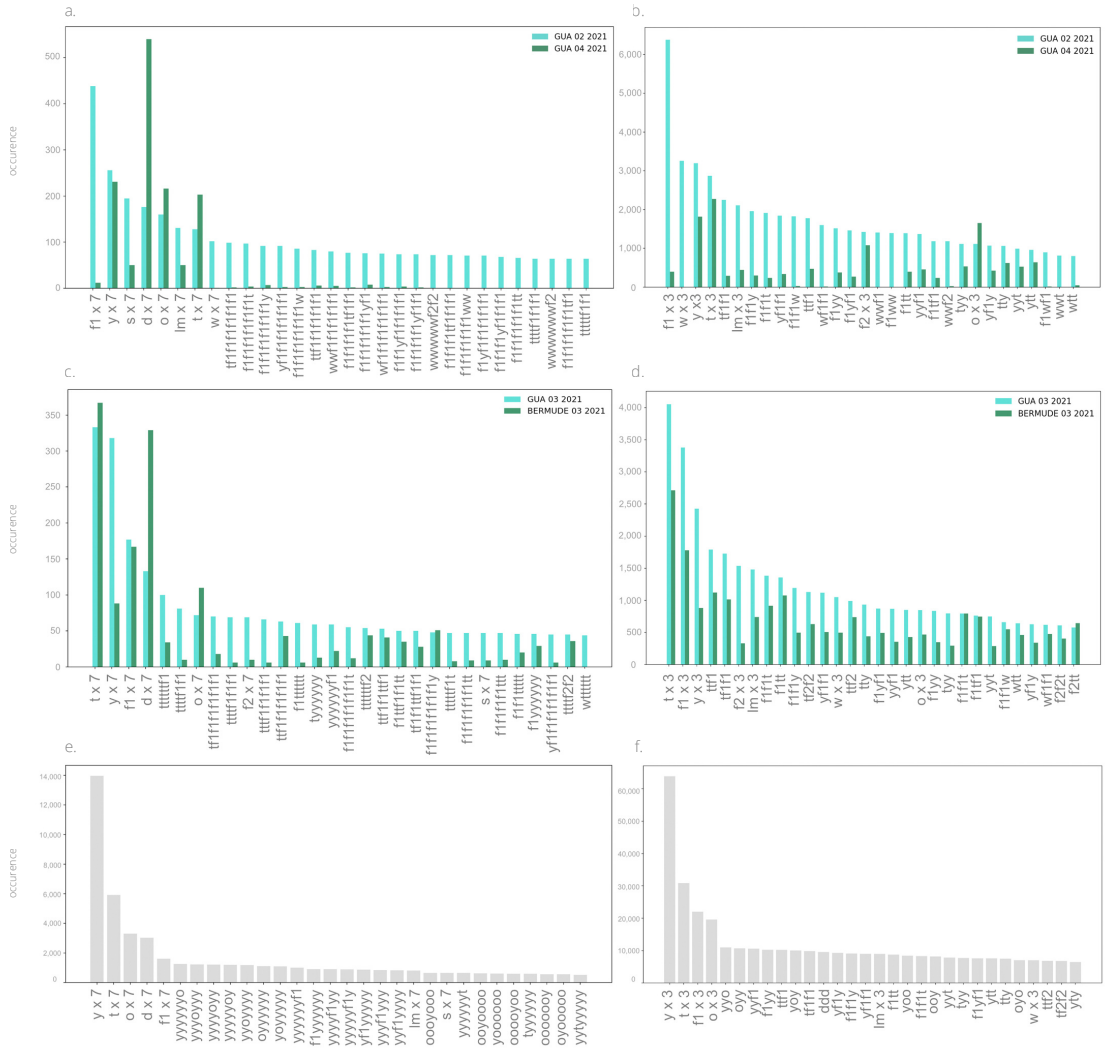


Figure 12. *Different n-grams comparing time and space effect on sequences occurence*

Based on linguistic analysis, n-grams are here associated to vocalize and sequences. Units names are shortened so that y = yawp; t = teepee; s = swop; d = dropplet; w = whup; lm = long_moan; lf = low_freq; o = oop; f1 = fluct_1; f2 = fluct_2; f3 = fluct_3 and g = growl. The comparison between February 2021 and April 2021 with 7 repetitions (a) and 3 repetitions (b) in the Guadeloupe station. Then the comparision between March 2021 in Guadeloupe and Bermuda with 7 repetitions (c) and 3 repetitions (d). Finally, the overall detection in all the recoding stations with 7 repetitions (e) and 3 repetitions (f).

the previous figure (*Figure 11.a*). When we decrease the number of repetitions from 7 to only 3 for the realization of the n-grams, we realize that the pattern seems to be more similar. However, when we look more closely, we can see that sequences containing "whups" are way less common in April than in February 2021 in the recordings of the 3 Guadeloupe stations. Finally, when comparing these sequences between regions, i.e. comparing the n-grams of Guadeloupe with those of Bermuda (*Figure 12.b*), two islands nearly 1,800 km away from each other, the results shows that the sequences of 3 successive vocalizations are partly similar in March 2021 despite lower occurrences and less use of some vocalizations type in Bermuda. Regarding the sequences of 7 vocalizations, the most common vocalizations are this time a succession of teepee but the other sequences are not all present in the recordings of the two stations. Taken together, these results suggest the hypothesis of song variation over time as well as variation in song composition by region. If we then display the most recorded sequences from all sessions of the project, over the full year, we can see that the n-grams of 3 and 7 show that the most commonly recorded is a succession of yawp-like vocalizations, described as the most detected vocalization in this work, followed by the succession of teepee. Unfortunately, by analyzing the n-grams of 7 (e), no sequence shows a complex song or with a pattern repetition. Indeed, we only observe that the n-grams of 7 corresponds to the repetition of vocalizations of all types.

DISCUSSION

The multiple results of this work are nevertheless debatable on various points. Indeed, despite the good scores of the different models, it is possible to discuss the reliability of the latter and their uses in the study of humpback whales around the world. Before that, it is first possible to talk about the impact of recording conditions as well as the choice of model parameters.

The impact of data collection and processing on detection and classification. The marine environment is an environment in which animals communicate largely through the emission of sound signals (*Haver et al. 2017*). In the case of humpback whales, this work has highlighted the presence of different types of vocalizations with a frequency range between 0.1 kHz and 3 kHz (*Figure 7*). It is first interesting to note the practical aspect of this frequency. First, low frequencies travel a longer distance (*Pace 2008*), which for very large animals can be useful because of the large range. Then, this frequency is used very little by other species of marine mammal, which avoids any disturbance in the reception of signals. The stationary signals, i.e. the whistles, of Risso's dolphins (*Grampus griseus*) for example are around 10 kHz, well above those of humpback whales (*Gannier et al. 2020*). However, at this low frequency are sounds emitted by boat engines, especially between 20 and 200 Hz (*Tyack and Janik 2013*). The presence of these disturbances at the same time as the presence of whales has shown to have a slight impact on individuals (*Villagra et al. 2021*), observed by the modification of behavior and in particular the speed of individuals. While studying a sound like in this work, if several different sound sources are on the same frequency band it can have a significant impact. Here, the recorders not being placed in areas without any human presence, many passages of boats have been recorded. While the sound of a passing boat creates temporary background noise, lasting a few minutes or even seconds, the background noise can in some cases be permanent, especially at reef level. The case for example of the GUA-BREACH station, deposited as its name suggests close to a reef zone, showed the impact of shrimp clicks which disturb the recordings and could therefore affect the performance of the model learning. In some studies, it has been shown that in the presence of continuous anthropogenic sounds, marine mammals, including humpback whales, would be able to avoid interference between signals by frequency modulation, thus making them pass above human origine signals (*Tyack and Janik 2013*). This can therefore put in difficulty the model which seeks to recognize a vocalization, usually at a precise frequency. It is therefore important to properly parameterize the model for learning. In addition, it is also important to avoid letting the training last too long. The longer the training is, the more the model will focus on the details of the training samples and therefore the less efficient the model will be on different samples. This phenomenon, called overfitting (*Dietterich 1995*), does not seem to have disturbed the results of the models in this work. However, if the training sample had not contained any background noise from the reef with only great vocalizations and the vocalizations of the tested sample did contain it, there is a good probability that the results would have been less good because of the disturbance caused by the background noise. Next, the choice of windows for learning is important. Indeed here a window of 3 seconds was chosen to ensure that the longest vocalizations ("long-moan") were taken into account. This size was used for the detector, but also to classify it. However, with a window that is too large, the risk is to take into account two vocalizations in the same window. It was for example observed in this work that when several individuals sing together, several vocalizations can be present in the window. This can therefore be a point of improvement for the future. Then, during this work, the choice of the architecture of the classifier led to several attempts. Starting with an architecture of Resnet50, the final choice was Resnet18 which seemed to show better performance even if sometimes minimal. However, it is difficult to explain how a less complex architecture could improve the learning of the classifier model. Regarding the detector, it seems that its operation is very good. Indeed, with scores close to an accuracy of 99.48%, it is possible to say that the model is robust in

terms of detecting humpback whales. Such a model had already been carried out, obtaining also very good results with an average precision value of 0.97 (*Allen et al. 2021*).

High variability in the units structure can induce errors in the classification. The major result of this work is of course the repertoire of 12 different types of vocalizations. This number is close to the one found in various studies carried out in other regions of the world (*Epp, M. E. Fournet, and Davoren 2022*). The names of the vocalizations in this work have been attributed in particular according to the work of *M. Epp, M. Fournet, and G. Davoren 2021, H. Winn and L. Winn 1978, Cusano et al. 2021*. As we described earlier, a vocalization is characterized by its duration, its frequency and its amplitude, thus, it is sometimes difficult to determine if a modulation of one of these parameters makes it a new unit or a sub-units. Here, by using a dimensionality reduction followed by a clustering, it made it possible to avoid a direct action of human, that is to say that the different vocalizations were highlighted by their measurable and not just visible differences. However, the structure of a vocalization remains complex and therefore this can complicate the work of classification (*Cholewiak, Sousa-Lima, and Cerchio 2013)*. The similarity between some vocalizations has in particular complicated this work, making the classification accuracy at only 83%. This score is already important as the task turns out to be complicated. Moreover, when we look at the classification errors, we realize that this is most generally done between similar vocalizations, fluct-1 with fluct-3, or fluct -2 with fluct-1 (*Figure 9*). These 3 vocalizations are indeed very similar in their structure, their times and their amplitude. With the naked eye it is sometimes difficult to distinguish them but since the clustering has separated them, this makes them 3 different units. In the literature, the classification is most often done manually by observing spectrograms, which makes comparison with the work presented here difficult. Despite the research carried out to compare these results with the literature, it would seem that the automation of the classification of humpback whale units, through the use of neural networks, is innovative. In order to improve the scores of our classifier, a greater diversity of vocalizations in the training must be added. Indeed, by showing the different possible shapes to the model during training, the latter will be better able to recognize them in the recordings. This was also achieved when it was found that the training set actually only corresponds to a minimal part of the diversity of the sample of detections. However, although the classification model has an interesting score, it cannot be used on all records worldwide.

Humpback whales who migrates in the Caribbean don't use the same units than in other migrating area. Although humpback whale song evolves over time, the song present in a given area still generally includes the same vocalizations (*Mercado III and C. E. Perazio 2021, Mercado 2022*). So much so that some of the vocalizations detected in this work has been highlighted in the same region as *Winn, H. E. 1978* did almost 40 years earlier. This is surely linked to the fact that the humpback whales present in a breeding area use the same songs and copy each other, therefore not bringing new units but rather a new structure in the song (*Mercado 2022)*. While our classify has highlighted 12 types of units in the Caribbean (*Figure 7*), it would seem that the repertoire of humpback whales is equivalent to nearly 60 units in the world. This number is obtained in particular by taking into account the sub-units, corresponding to the possible variations of the same units (*Pines 2018*) It is important to emphasize that it is possible that the annotations made for the detector may not contain all the units present in the recordings. Indeed, the large quantity of recordings makes the task of annotation complicated when there are variants and subunits that can be described as rare. Thus, launching our classification model set up here, with a good classification rate of 83% on only 12 units of the Caribbean, on a world scale would not make it possible to obtain a good classification of these 60 units. However, in the case of the Caribbean and with the data available, this work has nevertheless made it possible to highlight different phrases. However, it is possible to discuss in terms of number and proportion of vocalizations detected in the recordings. It would in fact seem that the "yawp" type vocalization is the most detected/common one in

recordings with up to 60% of detections at the Saint-Barthelemy station. When looking at this vocalization in the different sentences, it would appear to be used only in the form of repetition by group of 3 or 4 vocalizations. In a study by *V. Fournet et al. 2021*, concerning the repertoire of humpback whales from the northeast Newfoundland feeding area, we can see that this type of vocalization was predominant with in particular 52% of detections. However, as indicated, this concerns a feeding area and not a breeding ground, so there is no actual song (*Dunlop, Cato, and Noad 2008*). It is thus possible to hypothesize that these songs are not such songs and rather correspond to intra-specific communication defined as being signals of socialization (*Dunlop, Cato, and Noad 2008*). Indeed this is observed in birds including the example of female ducks (*Miller and Gottlieb 1978*) which use non-rhythmic sounds like a song and less complex, in particular to guide and call their young. Based on this hypothesis, this will explain the predominance of "yawp" in the recordings and the fact that these are also the most dominant in terms of occurrence in the sequences determined by the n-grams method. Regarding the latter, many difficulties were encountered in order to obtain interesting information. In particular, it was tried to obtain the beginning of a song, with the aim of reconstructing the entirety of a song. However, with the assumption that if the first vocalization of a recording is after the first 10 seconds of the latter, it is not possible to confirm whether the calculated beginning corresponds to the real beginning of a song or if the detector missed a vocalization present earlier.

Influence of time on humpback whale song. With the aim of highlighting the beginning of a song, the manual and visual method remains more suitable in the case of recordings of 1 minute every 5 minutes. Although it is possible to reconstruct the song in the manner of a DNA strand sequence as the Illumina method does for example (*Slatko, Gardner, and Ausubel 2018*), it is therefore difficult to do it automatically in this work because of the difficulty of finding the beginning of a song in a batch of recordings. Some studies propose processing the recordings by a method called recurrence-plot, which highlights the repetitions of songs in the same figure and thus makes it possible to observe an evolution over time (*Malige et al. 2021*). The results of this method seem to be promising. In our case and from all the observations made in this work, it is still possible to estimate that a modification has taken place in terms of occurrence of the detected sequences. Indeed, while the sequence most often detected corresponds to a succession of 7 "fluct-1" in February, the latter ranks 7th in terms of occurrence in April, behind a succession of "teepee", "oop" or even "droplet". Despite the fact that the precision of the detector and the classifier can still be questioned, it seems that this modification is an evolution of the song over time. The results of this work also suggested an evolution of song in terms of geography with sequence differences measured between two relatively distant regions (Guadeloupe and Bermuda). As mentioned in the introduction to this work, some papers have been able to highlight evolutions (*Mercado III and C. E. Perazio 2021, Ellen C Garland and McGregor 2020*), whether geographical or temporal. Indeed, with the hypothesis that individuals copy a model male (*Mercado 2022*), the impact of geography is significant in the sense that a group of individuals who do not encounter this model male will opt for another model that will not structure his singing the same way. However, given that the migration of humpback whales represents a very long journey, it is quite possible to imagine that an individual from group x met another individual from group y and that the latter imitate their songs, thus bringing diversity to the songs of isolated groups. This has been shown in particular in various studies (*Sousa-Lima 2005*). Concerning this work, it is therefore possible to estimate that the automatic detection, followed by an automatic classification and a detection of the sequences made it possible to highlight a local and temporal evolution of the song of the humpback whale in the breeding area of the Caribbean Sea and in particular in the two stations most frequented by humpback whales corresponding to Guadeloupe and Bermuda.

CONCLUSION

In order to conclude on this work, we have just seen that the semi-supervised learning method via convolutional neural network was reliable for detecting humpback whale vocalizations with great precision, but also to classify them by units. Indeed, after having used an autoencoder to reduce the dimensions of the data, the HDBSCAN clustering proved effective on a UMAP projection and made it possible to highlight the presence of 12 different units in the case of humpback whale songs in the Caribbean breeding area. These results made it possible, by comparing with the literature, to observe certain vocalizations in common with older papers in this same region. Through linguistic techniques, we have tried to highlight the sentences observed with the greatest occurrence. This seems to show an evolution of vocalizations in time and space with in particular a modification of the most used sequences. This observation was made by comparing the sequences between different remote islands but also by comparing the latter to a time interval of 2 months corresponding to firstly the arrival of the whales in the breeding area and secondly the end of the breeding period and therefore the departure of the latter.

BIBLIOGRAPHY

Adam, Olivier et al. (2013). "New acoustic model for humpback whale sound production". In: Applied Acoustics 74.10, pp. 1182–1190. doi: https://doi.org/10.1016/j.apacoust.2013.04.007.

Allen, Ann N. et al. (2021). "A Convolutional Neural Network for Automated Detection of Humpback Whale Song in a Diverse, Long-Term Passive Acoustic Dataset". In: Frontiers in Marine Science 8. doi: 10.3389/fmars.2021.607321.

Bertucci, Frédéric et al. (Aug. 2015). "Sound production by dusky grouper Epinephelus marginatus at spawning aggregation sites". In: Journal of Fish Biology 87, pp. 400–421. doi: 10.1111/jfb.12733.

Blount, Jon et al. (Dec. 2008). "Warning displays may function as honest signals of toxicity". In: Proceedings. Biological sciences / The Royal Society 276, pp. 871–7. doi: 10.1098/rspb.2008.1407.

Bradbury, J.W. and S.L. Vehrencamp (2011). Principles of Animal Communication. Sinauer. isbn: 9780878930456.

Cholewiak, Danielle M, Renata S Sousa-Lima, and Salvatore Cerchio (2013). "Humpback whale song hierarchical structure: Historical context and discussion of current classification issues". In: Marine Mammal Science 29.3, E312–E332.

Cusano, Dana et al. (Dec. 2021). "Socially Complex Breeding Interactions in Humpback Whales Are Mediated Using a Complex Acoustic Repertoire". In: Frontiers in Marine Science 8, p. 665186. doi: 10.3389/fmars.2021.665186.

Derville, Solène et al. (Mar. 2020). "Horizontal and vertical movements of humpback whales inform the use of critical pelagic habitats in the western South Pacific". In: Scientific Reports 10. doi: 10.1038/s41598-020-61771-z.

Dietterich, Tom (1995). "Overfitting and undercomputing in machine learning". In: ACM computing surveys (CSUR) 27.3, pp. 326–327.

Dunlop, Rebecca A, Douglas H Cato, and Michael J Noad (2008). "Non-song acoustic communication in migrating humpback whales (Megaptera novaeangliae)". In: Marine Mammal Science 24.3, pp. 613–629.

Epp, Mikala V, Michelle EH Fournet, and Gail K Davoren (2022). "Humpback whale call repertoire on a northeastern Newfoundland foraging ground". In: Marine Mammal Science 38.1, pp. 256–273.

Fisher, F. H. and V. P. Simmons (1977). "Sound absorption in sea water". In: The Journal of the Acoustical Society of America 62.3, pp. 558–564. doi: 10.1121/1.381574.

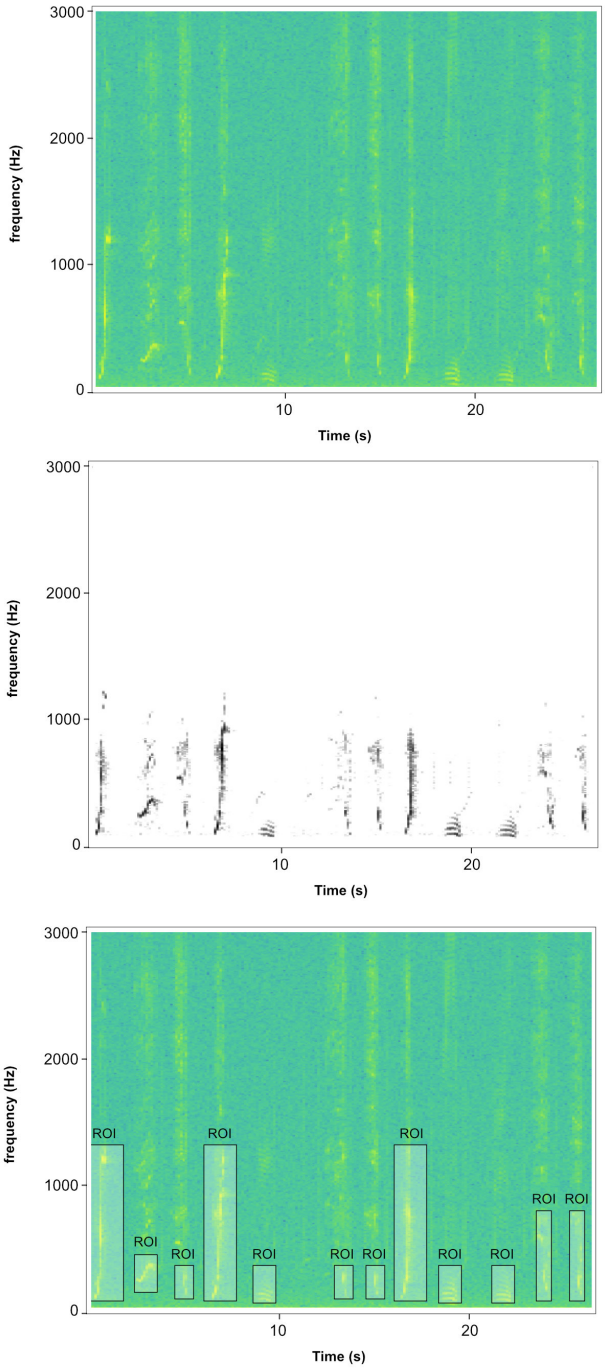
Fournet, Michelle, Andrew Szabo, and David Mellinger (Jan. 2015). "Repertoire and classification of non-song calls in Southeast Alaskan humpback whales (Megaptera novaeangliae)". In: The Journal of the Acoustical Society of America 137. doi: 10.1121/1.4904504.

Fowler, W. Warde (July 1896). "The Evolution of Bird-song". In: 54.1396, pp. 290–291. doi: 10.1038/054290a0.

Gannier, Alexandre et al. (2020). “Dolphin whistle repertoires around São Miguel (Azores): Are you common or spotted?” In: Applied Acoustics 161, p. 107169. issn: 0003-682X. doi: https://doi.org/10. 1016/j.apacoust.2019.107169.
Gao, Robert and Ruqiang Yan (Jan. 2006). “Non-stationary signal processing for bearing health monitoring”. In: IJMR 1, pp. 18–40. doi: 10.1504/IJMR.2006.010701.
Garland, Ellen and Peter Mcgregor (Sept. 2020). “Cultural Transmission, Evolution, and Revolution in Vocal Displays: Insights From Bird and Whale Song”. In: Frontiers in Psychology 11. doi: 10.3389/fpsyg.2020.544929.
Garland, Ellen C and Peter K McGregor (2020). “Cultural transmission, evolution, and revolution in vocal displays: insights from bird and whale song”. In: Frontiers in Psychology 11, p. 544929.
Garland, Ellen C. et al. (2011). “Dynamic Horizontal Cultural Transmission of Humpback Whale Song at the Ocean Basin Scale”. In: Current Biology 21.8, pp. 687–691. issn: 0960-9822. doi: https://doi.org/10.1016/j.cub.2011.03.019.
Glotin H., Ferrari M., Best P., Poupard M et al (2021), "CARIMAM Project Report 1: Bioacoustic data processsing". DYNI LIS. 2021.11, Ed. hal-03629286.
Grill, Thomas and Jan Schlüter (2017). “Two convolutional neural networks for bird detection in audio signals”. In: 2017 25th European Signal Processing Conference (EUSIPCO), pp. 1764–1768.
Hanlon, Roger et al. (Feb. 1990). “Physiological color change in squid iridophores - I. Behavior, morphology and pharmacology in Lolliguncula brevis”. In: Cell and tissue research 259, pp. 3–14. doi: 10.1007/BF00571424.
Haver, Samara M. et al. (2017). “The not-so-silent world: Measuring Arctic, Equatorial, and Antarctic soundscapes in the Atlantic Ocean”. In: Deep Sea Research Part I: Oceanographic Research Papers 122, pp. 95–104. issn: 0967-0637. doi: https://doi.org/10.1016/j.dsr.2017.03.002.
Heimlich, Sara et al. (May 2009). “Detecting humpback whale sounds in the Bering Sea: Confounding sounds in a cacophony of noise.” In: The Journal of the Acoustical Society of America 125, p. 2647. doi: 10.1121/1.4784132.
Hiett, Jane and Clive Catchpole (May 1982). “Song repertoires and seasonal song in the yellowhammer, Emberiza citrinella”. In: Animal Behaviour - ANIM BEHAV 30, pp. 568–574. doi: 10.1016/S0003-3472(82)80070-5.
Ioffe, Sergey and Christian Szegedy (2015). “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: pp. 448–456.
Johnson, Christopher M. et al. (Feb. 2022). Protecting Blue Corridors - Challenges and solutions for migratory whales navigating national and international seas. doi: 10 . 5281 / zenodo . 6196131. url: https://doi.org/10.5281/zenodo.6196131.
Khan, Momin et al. (June 2021). “Honey bees show dance pattern to communicate – A review”. In: World Journal of Biology and Biotechnology 6, p. 15. doi: 10.33865/wjb.006.02.414.
Lee, Sue Han et al. (Sept. 2015). “Deep-plant: Plant identification with convolutional neural networks”. In: pp. 452–456. doi: 10.1109/ICIP.2015.7350839.
Maas, Andrew L, Awni Y Hannun, Andrew Y Ng, et al. (2013). “Rectifier nonlinearities improve neural network acoustic models”. In: 30.1, p. 3.
Madsen, P et al. (Aug. 2002). “Sperm whale sound production studied with ultrasound time/depth-recording tags”. In: The Journal of experimental biology 205, pp. 1899–906. doi: 10.1242/jeb.205.13. 1899.
Malige, F. et al. (2021). “Use of recurrence plots for identification and extraction of patterns in humpback whale song recordings”. In: Bioacoustics 30.6, pp. 680–695. doi: 10.1080/09524622.2020.1845240.
Marini, Simone et al. (Sept. 2018). “Tracking Fish Abundance by Underwater Image Recognition”. In: Scientific Reports 8. doi: 10.1038/s41598-018-32089-8.
McInnes, Leland, John Healy, and James Melville (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. doi: 10.48550/ARXIV.1802.03426. url: https://arxiv.org/abs/1802.03426.
Mercado, Eduardo (2021). “Song Morphing by Humpback Whales: Cultural or Epiphenomenal?” In: Frontiers in Psychology 11. doi: 10.3389/fpsyg.2020.574403.
Mercado, Eduardo (May 2022). “The Humpback’s New Songs: Diverse and Convergent Evidence Against Vocal Culture via Copying in Humpback Whales”. In: Animal Behavior and Cognition 9, pp. 196–206. doi: 10 26451/abc.09.02.03.2022.
Mercado, Eduardo and Christina Perazio (Feb. 2022). “All units are equal in humpback whale songs, but some are more equal than others”. In: Animal Cognition 25. doi: 10.1007/s10071-021-01539-8.
Mercado III, Eduardo and Christina E Perazio (2021). “Similarities in composition and transformations of songs by humpback whales (Megaptera novaeangliae) over time and space.” In: Journal of Comparative Psychology 135.1, p. 28.

Miller, David B and Gilbert Gottlieb (1978). “Maternal vocalizations of mallard ducks (Anas platyrhynchos)”. In: Animal Behaviour 26, pp. 1178–1194. issn: 0003-3472. doi: https://doi.org/10.1016/00033472(78)90108-2.
Pace, Federica (2008). “Comparison of feature sets for humpback whale song classification”. PhD thesis. Doctoral dissertation, PhD dissertation, University of Southampton, UK.
Pace F., Benard F., Glotin H., Adam O., White P., "Subunit definition and analysis for humpback whale call classification", Applied Acoustics 71 (11), 1107-1112
Payne, K. (2000). “The progressively changing songs of humpback whales: a window on the creative process in a wild animal”. In: Origins of music. eds. N. L. Wallin, B. Merker, and S. Brown (Cambridge, MA: MIT Press), pp. 135–150. doi: 10.1121/1.381574.
Payne, RS and S McVay (1971). “Songs of humpback whales”. In: Science 173, pp. 585–597.
Pershing, A. et al. (Aug. 2010). “The Impact of Whaling on the Ocean Carbon Cycle: Why Bigger Was Better”. In: PloS one 5, e12444. doi: 10.1371/journal.pone.0012444.
Pines, Howard (2018). “Mapping the phonetic structure of humpback whale song units: extraction, classification, and Shannon-Zipf confirmation of sixty sub-units”. In: Proceedings of Meetings on Acoustics 35.1, p. 010003. doi: 10.1121/2.0000957.
Poupard, Marion et al. (Feb. 2022). “Passive acoustic monitoring of sperm whales and anthropogenic noise using stereophonic recordings in the Mediterranean Sea, North West Pelagos Sanctuary”. In:Scientific Reports 12. doi: 10.1038/s41598-022-05917-1.
Reidenberg, Joy and Jeffrey Laitman (June 2007). “Discovery of a low frequency sound source in Mysticeti (baleen whales): Anatomical establishment of a vocal fold homolog”. In: Anatomical record (Hoboken, N.J. : 2007) 290, pp. 745–59. doi: 10.1002/ar.20544.
Rocha, Luciana et al. (Jan. 2015). “An Evaluation of Manual and Automated Methods for Detecting Sounds of Maned Wolves (Chrysocyon brachyurus Illiger 1815)”. In: Bioacoustics In press. doi: 10.1080/09524622.2015.1019361.
Romero Mujalli, Daniel et al. (Apr. 2014). “Caracterización de Silbidos de Tursiops truncatus (Cetacea: Delphinidae) y su Asociación con el Comportamiento en Superficie”. In: Revista Argentina de Ciencias del Comportamiento.
Schmitz, Barbara et al. (2000). “A unique way of sound production in the snapping shrimp (Alpheus heterochaelis)”. In: The Journal of the Acoustical Society of America 108.5, pp. 2542–2542. doi: 10.1121/1.4743419.
Seyfarth, Robert M. and Dorothy L. Cheney (2003). “Signalers and Receivers in Animal Communication”. In: Annual Review of Psychology 54.1. PMID: 12359915, pp. 145–173. doi: 10.1146/annurev.psych.54.101601.145121.
Slatko, Barton E, Andrew F Gardner, and Frederick M Ausubel (2018). “Overview of next-generation sequencing technologies”. In: Current protocols in molecular biology 122.1, e59.
Sousa-Lima, Renata S (2005). “Songs indicate interaction between humpback whale (Megaptera novaeangliae) populations in the western and eastern South Atlantic Ocean”. In: Marine Mammal Science 21.3, pp. 557–566.
Srivastava, Nitish et al. (2014). “Dropout: a simple way to prevent neural networks from overfitting”. In: The journal of machine learning research 15.1, pp. 1929–1958.
Stowell, Dan et al. (Aug. 2016). “Bird detection in audio: a survey and a challenge”.
Roderick and James Fattu (Nov. 1973). “Mechanisms of Sound Production by Echolocating Bats”. In: Integrative and Comparative Biology 13. doi: 10.1093/icb/13.4.1215.
Sutskever, Ilya et al. (17–19 Jun 2013). “On the importance of initialization and momentum in deep learning”. In: Proceedings of Machine Learning Research 28.3. Ed. by Sanjoy Dasgupta and David McAllester, pp. 1139–1147.
Tyack, Peter L. and Vincent M. Janik (2013). “Effects of Noise on Acoustic Signal Production in Marine Mammals”. In: ed. by Henrik Brumm, pp. 251–271.
Villagra, Damian et al. (Jan. 2021). “Energetic Effects of Whale-Watching Boats on Humpback Whales on a Breeding Ground”. In: Frontiers in Marine Science 7. doi: 10.3389/fmars.2020.600508.
Whitlow, Au (2018). “History of bioacoustics research on aquatic and marine organisms in Hawaii”. In: The Journal of the Acoustical Society of America 143.3, pp. 1769–1769. doi: 10.1121/1.5035791.
Wilcock, William SD et al. (2014). “Sounds in the ocean at 1–100 Hz”. In: Annual review of marine science 6.1, pp. 117–140.
Winn, H. and Lizzy Winn (June 1978). “The song of the humpback whale Megaptera novaeangliae in the West Indies”. In: Marine Biology 47, pp. 97–114. doi: 10.1007/BF00395631.
Zandberg, L., Lachlan, R. F., Lamoni, L., & Garland, E. C. (2021). Global cultural evolutionary model of humpback whale song. Philosophical Transactions of the Royal Society B, 376(1836), 20200242.

SUPPLEMENT FIGURE



30 seconds spectrogram
to annotate

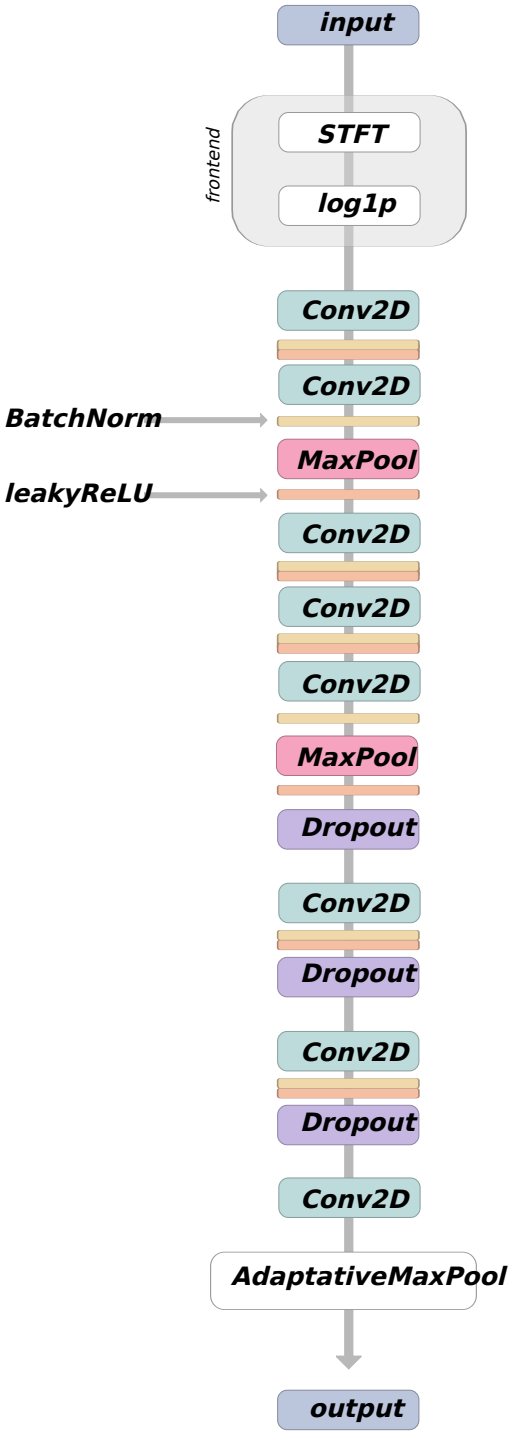
1

Removing the background
noise based on dB level

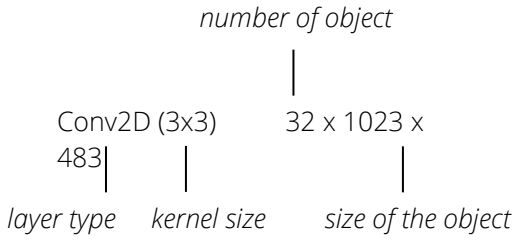
2

Double threshold that
isolates the remaining regions
of the spectrogram = ***ROI***

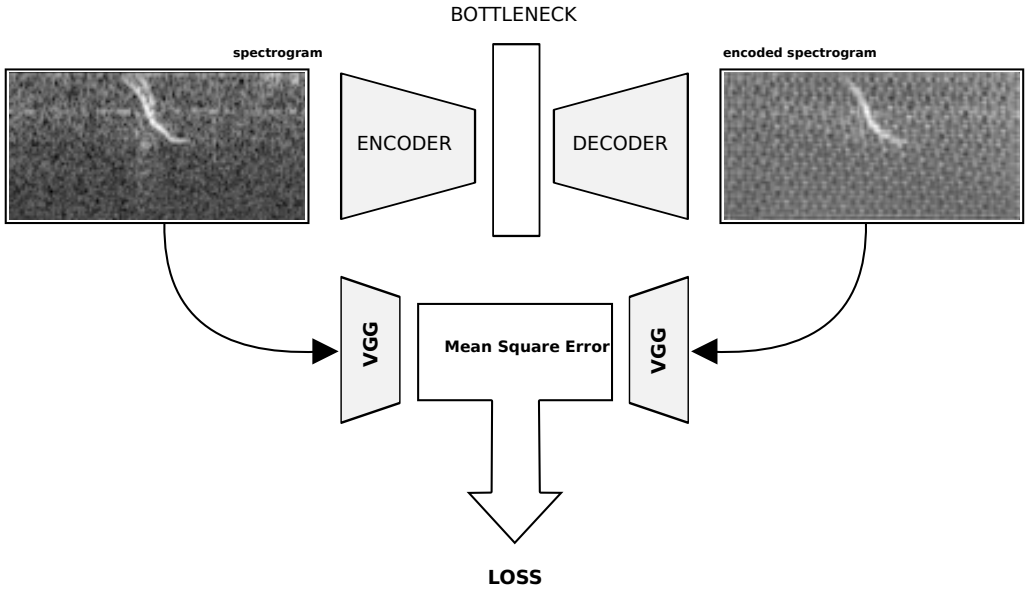
Supplement Figure 1. *ROIs detection method*



Input	1 x 1025 x 485
Conv2D (3x3)	32 x 1023 x 483
Conv2D (3x3)	32 x 1021 x 481
MaxPool (3x3)	32 x 340 x 160
Conv2D (3x3)	32 x 338 x 158
Conv2D (3x3)	32 x 336 x 156
Conv2D (3x16)	64 x 321 x 154
MaxPool (1x3)	64 x 321 x 51
Conv2D (9x1)	256 x 321 x 43
Conv2D (1x1)	64 x 321 x 43
Conv2D (1x1)	1 x 321 x 43
AdaptativeMaxPool	1 x 1 x 1



Supplement Figure 2. *Architecture of the CNN used for humpback whale detection with convolution parameter*



Supplement Figure 3. *Organisation of the autoencoder used for dimensionnality reduction*

	type	*number*
Test	*dropplet*	132
	fluct_1	94
	fluct_2	121
	fluct_3	63
	growl	94
	long_moan	103
	low_freq	85
	oop	96
	swop	98
	teepee	112
	whup	70
	yawp	95
Train	*dropplet*	411
	fluct_1	431
	fluct_2	378
	fluct_3	428
	growl	400
	long_moan	454
	low_freq	414
	oop	365
	swop	266
	teepee	389
	whup	395
	yawp	392

Supplement Table 1. *Test and train classifier dataset organization*

Supplement Table 2. *Number of different type vocalization found in the recordings from all stations*

	ANGUILLA	ARUBA	BAHAMAS	BERMUDA	BONAIRE	GUA_AB	GUA_BREACH
dropplet	2704	1	27	2785	63	**12419**	2353
fluct_1	**17492**	1	2	**20051**	53	**60724**	**22665**
fluct_2	7325	7	11	**19045**	236	**62332**	**15289**
fluct_3	1066	-	-	1476	1	2812	580
growl	6822	-	3	109	11	159	21
long_moan	9587	-	45	6398	54	**15370**	1536
low_freq	1717	4	6	2714	17	**18897**	2968
oop	2587	3	8	3465	2	9003	5729
swop	925	6	23	1524	104	5263	1356
teepee	1903	3	6	**11182**	193	**28430**	9861
whup	4934	7	8	8698	115	**20216**	4188
yawp	**11094**	19	41	**22652**	316	**57556**	**41093**

GUA_SF	GUYANNE_CEPOG	JAMAICA	MART PRECHEUR	MART St ANNE	SAINT BARTHELEMY	SAINT EUSTATIUS	SAINT MARTIN
7890	728	44	719	2646	4979	159	4396
9209	4	6	4955	1482	**16067**	**17281**	5895
22696	18	38	3623	1840	5957	7252	8370
2286	-	-	11	7	14	29	704
44	616	3	25	3355	483	34	1085
3067	418	27	437	433	815	329	1113
7555	4	1	1265	154	5689	681	3245
5024	65	16	1615	17	**21530**	985	2941
2060	309	34	246	353	1605	444	684
7399	7	18	3072	1074	1845	4216	5917
2261	62	36	1774	68	1479	1157	595
14890	86	76	**11308**	2218	**91115**	**28520**	**10304**

Supplement Table 3. Number of vocalizations of different type, normalized by the number of hours of recording, and proportion of each type according to the stations

	ANGUILLA		ARUBA		BAHAMAS		BERMUDA		BONAIRE		GUA_AB		GUA_BREACH	
	nb.	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*
dropplet	50,07	3,97	0,07	1,96	0,4	15	36,64	2,78	0,51	5,41	90,65	4,24	19,77	2,19
fluct_1	323,93	**25,66**	0,07	1,96	0,03	1,11	263,83	**20,03**	0,43	4,55	443,24	**20,71**	190,46	**21,06**
fluct_2	135,65	10,75	0,47	**13,73**	0,16	6,11	250,59	**19,03**	1,92	20,26	454,98	**21,26**	128,48	**14,2**
fluct_3	19,74	1,56	-	-	-	-	19,42	1,47	0,01	0,09	20,53	0,96	4,87	0,54
growl	126,33	10,01	-	-	0,04	1,67	1,43	0,11	0,09	0,94	1,16	0,05	0,18	0,02
long_moan	177,54	**14,07**	-	-	0,67	**25**	84,18	6,39	0,44	4,64	112,19	5,24	12,91	1,43
low_freq	31,8	2,52	0,27	7,84	0,09	3,33	35,71	2,71	0,14	1,46	137,93	6,45	24,94	2,76
oop	47,91	3,8	0,2	5,88	0,12	4,44	45,59	3,46	0,02	0,17	65,72	3,07	48,14	5,32
swop	17,13	1,36	0,4	**11,76**	0,34	**12,78**	20,05	1,52	0,85	8,93	38,42	1,8	11,39	1,26
teepee	35,24	2,79	0,2	5,88	0,09	3,33	147,13	**11,17**	1,57	**16,57**	207,52	9,7	82,87	9,16
whup	91,37	7,24	0,47	**13,73**	0,12	4,44	114,45	8,69	0,93	9,87	147,56	6,9	35,19	3,89
yawp	205,44	**16,28**	1,27	**37,25**	0,61	**22,78**	298,05	**22,63**	2,57	**27,12**	420,12	**19,63**	345,32	**38,18**

GUA_SF		GUYANNE_GEPOG_		JAMAICA		MART PRECH		MART St ANNE		SAINT BARTHELEMY		SAINT EUSTACHIUS		SAINT MARTIN	
nb.	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*	*nb.*	*%*
62,13	9,35	19,68	**31,42**	0,35	**14,72**	6,91	2,48	82,69	**19,39**	33,19	3,28	2,65	0,26	57,84	9,72
72,51	**10,91**	0,11	0,17	0,05	2,01	47,64	**17,06**	46,31	**10,86**	107,11	**10,6**	288,02	**28,29**	77,57	**13,03**
178,71	26,9	0,49	0,78	0,3	**12,71**	34,84	**12,47**	57,5	**13,48**	39,71	3,93	120,87	**11,87**	110,13	**18,5**
18	2,71	-	-	-	-	0,11	0,04	0,22	0,05	0,09	0,01	0,48	0,05	9,26	1,56
0,35	0,05	16,65	**26,59**	0,02	1	0,24	0,09	104,84	**24,58**	3,22	0,32	0,57	0,06	14,28	2,4
24,15	3,63	11,3	**18,04**	0,22	9,03	4,2	1,5	13,53	3,17	5,43	0,54	5,48	0,54	14,64	2,46
59,49	8,95	0,11	0,17	0,01	0,33	12,16	4,35	4,81	1,13	37,93	3,75	11,35	1,11	42,70	7,17
39,56	5,95	1,76	2,81	0,13	5,35	15,53	5,56	0,53	0,12	143,53	**14,2**	16,42	1,61	38,70	6,5
16,22	2,44	8,35	**13,34**	0,27	11,37	2,37	0,85	11,03	2,59	10,7	1,06	7,4	0,73	9,00	1,51
58,26	8,77	0,19	0,3	0,14	6,02	29,54	**10,57**	33,56	7,87	12,3	1,22	70,27	6,9	77,86	**13,08**
17,8	2,68	1,68	2,68	0,29	**12,04**	17,06	6,11	2,13	0,5	9,86	0,98	19,28	1,89	7,83	1,31
117,24	**17,65**	2,32	3,71	0,61	**25,42**	108,73	**38,93**	69,31	**16,25**	607,43	**60,11**	475,33	**46,69**	135,58	**22,77**

ABSTRACT

With nearly 1,302 days, i.e. more than 6,000 hours, of underwater recordings from the year 2021 in the Caribbean Sea, it was possible to develop a convolutional neural network with the role of detecting humpback whale (*Megaptera novaeangliae*) vocalization. The latter showed very good performance (mean Average Precision of 0.9948). Subsequently, with the aim of highlighting the different units composing the song of humpback whales during this given period of time, an autoencoder made it possible to reduce the dimensions of the recordings to 16 in order to perform a clustering. The use of the HDBSCAN method proved effective on a UMAP projection and made it possible to identify the presence of 12 different units. These 12 units were therefore learned by a second neural network, this time with the role of classifying (accuracy of 0.83). Finally, by studying song sequences and measuring their occurrence in the recordings, it was possible to assume an evolution in time but also geographically of the song of humpback whales during the reproduction period.

RÉSUMÉ

Avec près de 1.302 jours, soit plus de 6.000 heures, d'enregistrements sous-marin provenant de l'année 2021 en mer des Caraïbes, il a été possible de mettre au point un réseau de neurones à convolutions avec pour rôle la détection des vocalises de baleine à bosse (*Megaptera novaeangliae*). Ce dernier a montré de très bonnes performances (mAP de 0,9948). Par la suite, avec pour objectif de mettre en évidence les différentes units composants le chant des baleines à bosses durant cette période de temps donnée, un autoencoder a permis de réduire les dimensions des enregistrements à 16 afin d'effectuer un clustering. L'utilisation de la méthode HDBSCAN s'est montrée efficace sur une projection UMAP et a permis de révéler la présence de 12 units différentes. Ces 12 units ont donc été apprisses par un second réseau de neurones avec cette fois-ci un rôle de classifier (précision de 0,83). Enfin, par l'étude de séquences de chant et en mesurant leurs occurrences dans les enregistrements, il a été possible de supposer une évolution dans le temps mais également géographiquement du chant des baleines à bosse au cours de la période de reproduction.