Article

On the Importance of Passive Acoustic Monitoring Filters

Rafael Aguiar[1,*], Gianluca Maguolo[2], Loris Nanni[2], Yandre Costa[3] and Carlos Silla Jr. [1]

[1] Pontifícia Universidade Católica do Paraná
[2] University of Padua
[2] State University of Maringá
* Correspondence: aguiar.pr@gmail.com

Abstract: Passive acoustic monitoring (PAM) is a non-invasive technique to supervise the wildlife. Acoustic surveillance is preferable in some situation such as in the case of marine mammals, when the animals spend most of their time underwater, making it hard to obtain their images. Machine learning is very useful for PAM, for example, to identify species based on audio recordings. But some care should be taken to evaluate the capability of a system. We define PAM-filters as the creation of the experimental protocols according to the dates and locations of the recordings, aiming to avoid the use of the same individuals, noise and recording devices in both training and test sets. A random division of a database present accuracies much higher than accuracies obtained with protocols generated with PAM-filter. Although we use the animal vocalizations, in our method we convert the audio into spectrogram images, after that, we describe the images using the texture. Those are well-known techniques for audio classification, and they have already been used for species classification. Also, we perform statistical tests to demonstrate the significant difference between accuracies generated with and without PAM-filters with several well-known classifiers. The configuration of our experimental protocols and the database were made available online[1].

Keywords: PAM; Passive acoustic monitoring; audio classification; texture classification; PAM-filter; experimental protocols for audio classification; statistical tests.

1. Introduction

Techniques of Passive acoustic monitoring (PAM) are tools to automatic detect, localize and monitor animals [1]. Passive refers to the fact that the system is non-invasive, as it does not interfere with the environment. It is an acoustic system because the surveillance is done through audio signals. For example, a recording device connected to the internet could acquire data from an environment and send captured data to a classification system that identifies which species are nearby.

In the case of marine animals, the use of audio data might be preferred over image data [2]. The reasoning for that is because visual survey methods for some marine animals, such as whales, may detect only a fraction of the animals present in the area. This happens because visual observers can only see them during the very short period when they are on the surface, and also because visual surveys can be undertaken only during daylight hours and in relatively good weather.

Global warming, industrial fishing, oil spilling and other factors cause a lot of damage and changes in environment of marine mammals. In virtue of that, it is really important to keep marine mammals under supervision. A practical use of species identification is applied into the North Atlantic Right Whales, focusing in environmental conservation. Collisions between ships and these animals are one of the main threats

[1] https://figshare.com/articles/dataset/Database_of_spectrograms_of_marine_mammals/14068106

 doi:10.20944/preprints202104.0766.v1

to this species[2]. In order to avoid these collisions, in 2013 there was an international challenge[3] to automatically identify if a given audio contains or not vocalizations of such species. The data from the challenge was collected by floating buoys. This kind of recognition can help ships to change their route to avoid a possible collision. Beyond the challenge, there are other researches about North Atlantic Right Whales identification in the scientific literature [3–5]. In fact, the situation of this species is so critical that another challenge was proposed in 2015[4]. In this one, the database was composed of aerial images from these animals and the classification task was to identify each individual, to help researchers track health and general status of the individuals, focusing in conservation efforts.

Although PAM techniques and machine learning are useful in marine researches, there are two potential hazards concerning to machine learning we may note in systems related to species classification. The first one is the risk of using vocalizations of the same individuals both on training and test sets simultaneously, this may lead the system to be able to recognize the individuals instead of the species. Not only the vocalization of the same individual can bias the system, but even a characteristic noise presented in several samples of a class can be distinguishing from noises of other classes, which is the second hazard. That can happen when samples of the same class are recorded in the same location with the same devices, either the environment and the devices can create noise. Based on these issues and aiming on more reliable results for PAM systems, we propose the PAM-filter, which means trying to use the same individual always in the same set, whether training or test. In the database we use, it was possible to separate the individuals from the same class by location and date of record, trying our best to avoid the recognition of individuals and noise. Experiments were also conducted with a randomized version of the database, and the results are fairly disparate.

2. Materials and Methods

In this section, we describe the database used for experimentation and the protocols developed aiming to properly explore it. In addition, the theoretical framework is also described.

2.1. Watkins Marine Mammal Sound Database - WMMSD

Marine mammals are an informal group of animals that relies on a marine ecosystem for existence. According to taxonomy committee from The Society for Marine Mammalogy[5], marine mammals are classified in three orders and several families, genus, species and subspecies, some of them maybe are already extinct. In this work we use the Watkins Marine Mammal Sound Database[6] (WMMSD). The audio files were recorded from the 1940's to the 2000's, and the species in the audio files were identified by professional biologists.

The database is composed of almost 1,600 entire tapes. Each tape is composed of several minutes of recording, and they may contain vocalizations of several species. Smaller cuts of these long-length audio files are available in the website, usually with vocalizations of only one species. They are divided into two sections in the website, "all cuts" and "best cuts". "Best cuts" represents high-quality and low-noise cuts. "All cuts" contains all the audio files from the "best cuts" plus other ones, which are lower-quality and noisier.

We choose to use the "best cuts" for the classification task, as noise reduction and segmentation is not the focus of this work. "Best cuts" contains 1,694 audio files from

[2] https://ocean.si.edu/ocean-life/marine-mammals/north-atlantic-right-whale

[3] https://www.kaggle.com/c/the-icml-2013-whale-challenge-right-whale-redux

[4] https://www.kaggle.com/c/noaa-right-whale-recognition

[5] https://marinemammalscience.org/

[6] https://cis.whoi.edu/science/B/whalesounds/index.cfm

 doi:10.20944/preprints202104.0766.v1

32 species of marine mammals. There were 25 samples that contains vocalizations of more than one species, those were removed as we do not intend to handle multi-label classification neither audio segmentation.

The website also provides a meta-data file for each audio cut. The data contains additional information, such as the date and the location of the record. However, most meta-data files do not present information for all their fields. Several classes have must of their samples records in just a few locations and dates. It led us to suspect that they may contain samples of the same individual. Also, the noise pattern in such samples are homogeneous, using the metadata files it is possible to see that the cuts are extracted from the same long-length tapes.

2.2. PAM-filter

In another audio classification task, music genre classification, Flexer [6] defines the concept of "artist-filter", which means to have all samples of the same artist either in the training or test. The author noticed that experiments with samples from the same artists in training and test sets present higher accuracies and lower standard deviations, that suggests the music genre classification systems were learning the artist instead of the genre.

We consider that experiments with randomized sets of training and test may produce overestimated results of species classification, because the classifiers may be taking decisions based on underlying patterns. This is based on the information presented in the metadata files of the WMMSD, which indicates there are a lot of samples of the same individuals and, also, the same pattern of noise is presented in several samples of the same class. The concept of "artist-filter" led us to conceive the "PAM-filter", where we try to use the same individual's vocalization either in the training or test set, never in both them.

Figure 1 presents four different vocalizations of the species *Eubalaena glacialis*, all the spectrograms were generated with the same parameters. According to the metadata files, Figures 1a and 1b were extracted from the same long-length tape, recorded in the same day, with the same devices. It is discernible vertical lines that represents the same noise in the lower area of the Figures 1a and 1b. The vocalization in the Figure 1a is described as "grunt" and in the Figure 1b is described as "one long groan".

Figures 1c and 1d also share the same tape, date and devices. Both vocalizations are described as "moan". Although noise is not visible in vertical lines and neither reported in the metadata, a texture pattern resembling to "salt and pepper" is present all over the spectrogram, probably generated by the sound of the ocean.

2.3. Watkins experimental protocols

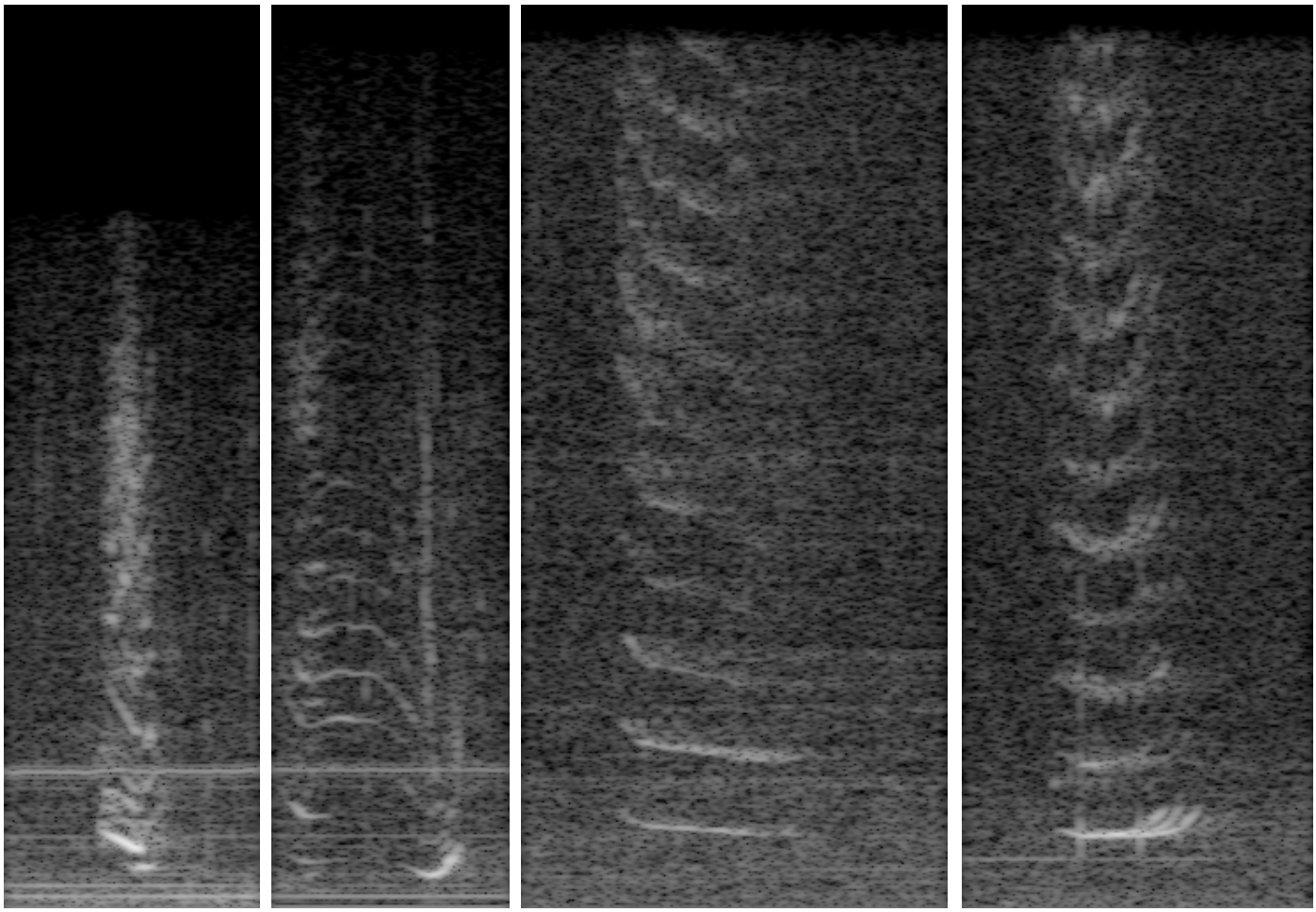
To investigate the impact of PAM-filter, we create three different protocols of the same database, with and without PAM-filter. They are called Watkins Experimental Protocols, WEP#1, WEP#2 and WEP#3, they are described as follow and their specifications are available online[7].

2.3.1. Watkins Experimental Protocol #1 (WEP#1): ten-fold cross-validation

The first protocol is defined without any concerns about PAM-filter. We use the database randomly divided into ten-fold cross-validation, classes with fewer than ten samples were removed. Table 1 presents the species, number of samples used in WEP#1, number of locations were the samples were recorded and the number of samples presented in each location.

[7] https://figshare.com/articles/dataset/Database_of_spectrograms_of_marine_mammals/14068106

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1



(a) Recorded in 1956. **(b)** Recorded in 1956. **(c)** Recorded in 1981. **(d)** Recorded in 1981.

Figure 1. Different vocalizations of *Eubalaena glacialis*.

2.3.2. Watkins Experimental Protocol #2 (WEP#2): training/test protocol

As it is listed in Table 1, some species, like *Balaenoptera acutorostrata*, hold all its records in the same place. It indicates the samples might contain the vocalizations of the same individual and, also, the noise pattern generated by the environment and the devices are usually similar. Such species were removed.

To create a experimental protocol using the PAM-filter, we scan the metadata files to separate the database according to locations, each location of each class is allocated exclusively in the training or the test set.

Also, in other cases, all samples from the same class were recorded in just a few different locations. In species where the samples belong to only two locations, the location with more sample goes to the training set and the second location goes to the test set. Classes with three or more locations were distributed trying to achieve 70% for training and 30% for test.

Table 2 presents the data used in WEP#2 and WEP#3. Classes with unfilled number of samples were not in the protocol. WEP#2 is composed of 24 classes, 908 samples for training and 412 for test.

2.3.3. Watkins Experimental Protocol #3 (WEP#3): two-fold cross-validation

Third and last protocol is two-fold cross-validation, it was created to use all the samples either for training and test, improving the reliability of the results. We could not increase the number of folds due to the number of locations, since eight classes have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1

Table 1. Composition of WEP#1: 31 species. Columns also present the number of locations were the samples were recorded, number of samples recorded in each location and number of samples per class.

Species	Number of locations	Samples by location	Number of samples in WEP#1
Balaena mysticetus	2	1; 49	50
Balaenoptera acutorostrata	1	17	17
Balaenoptera physalus	2	5; 45	50
Delphinapterus leucas	4	1; 6; 16; 27	50
Delphinus delphis	3	2; 15; 35	52
Erignathus barbatus	3	3; 9; 15	27
Eubalaena australis	2	7; 18	25
Eubalaena glacialis	4	3; 12; 19; 20	54
Globicephala macrorhynchus	4	5; 16; 18; 26	65
Globicephala melas	4	11; 12; 14; 28	65
Grampus griseus	3	1; 21; 45	67
Hydrurga leptonyx	1	10	10
Lagenodelphis hosei	1	87	87
Lagenorhynchus acutus	3	12; 12; 31	55
Lagenorhynchus albirostris	2	20; 37	57
Megaptera novaeangliae	3	1; 17; 46	64
Monodon monoceros	3	4; 10; 36	50
Odobenus rosmarus	3	1; 16; 21	38
Ommatophoca rossi	3	11; 19; 20	50
Orcinus orca	5	1; 2; 5; 8; 19	35
Pagophilus groenlandicus	1	47	47
Peponocephala electra	1	56	56
Physeter macrocephalus	6	2; 2; 2; 9; 12; 33	60
Pseudorca crassidens	2	11; 48	59
Stenella attenuata	2	11; 54	65
Stenella clymene	2	14; 49	63
Stenella coeruleoalba	4	8; 12; 27; 34	81
Stenella frontalis	1	58	58
Stenella longirostris	2	1; 113	114
Steno bredanensis	1	50	50
Tursiops truncatus	3	1; 10; 13	24
Number of samples			1,645

all their samples recorded in only two locations (see Table 1), increasing the number of fold to three would result in a major cut of classes.

However, a minor cut of classes was still necessary. For example, the class *Stenella longirostris* has recordings taken from two locations, one of them with one sample and the other one with 113. If it was used in cross-validation, 113 samples of a class would be tested in a model trained with just one sample of the class, an unfair task. Empirically, we decided to remove classes which hold fewer than ten samples in either one of the folds. Table 2 presents details of WEP#3. It uses 20 class. These two folds holds 542 and 539 samples each.

2.4. Theoretical Framework

In this section we present theoretical information about the experimental protocol used in our experiments. First, we describe how the audio signal was manipulated and the spectrograms generated. Than, we detail the feature extraction and, lastly, the classifiers trialed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1

Table 2. Data used in the WEPs#2 and #3. Classes and number of samples of the training and test set of the WEP#2 and the two folds of WEP#3. Unfilled cells mean the class was not used, 24 classes were used in WEP#2 and 20 in WEP#3.

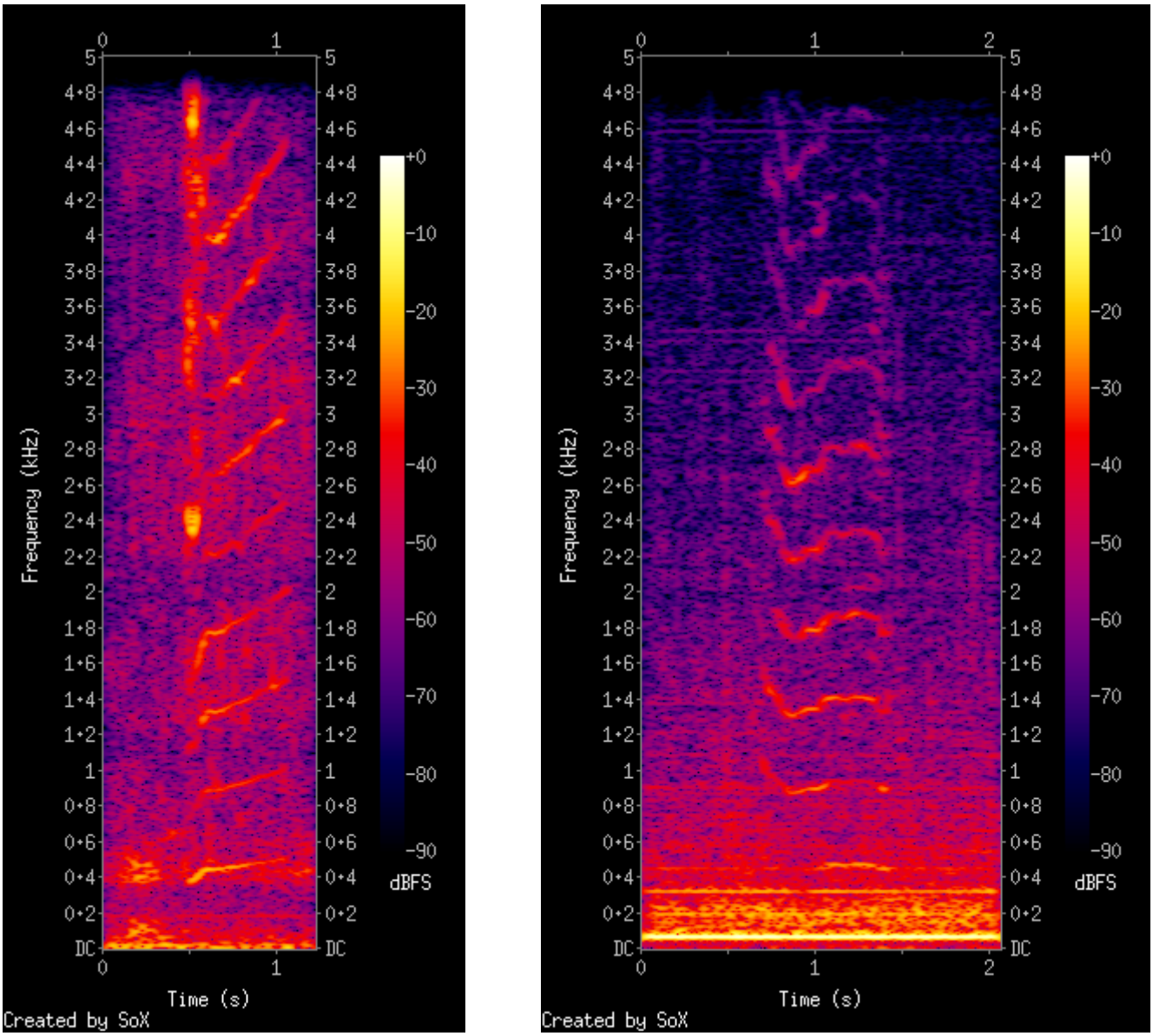
Species	Number of Samples			
	WEP#2		WEP#3	
	Train	Test	Fold 1	Fold 2
Balaena mysticetus	49	1	-	-
Balaenoptera acutorostrata	-	-	-	-
Balaenoptera physalus	45	5	-	-
Delphinapterus leucas	27	23	23	27
Delphinus delphis	35	17	35	17
Erignathus barbatus	15	12	15	12
Eubalaena australis	18	7	-	-
Eubalaena glacialis	31	23	31	23
Globicephala macrorhynchus	34	31	31	34
Globicephala melas	37	28	28	37
Grampus griseus	45	22	45	22
Hydrurga leptonyx	-	-	-	-
Lagenodelphis hosei	-	-	-	-
Lagenorhynchus acutus	31	24	31	24
Lagenorhynchus albirostris	37	20	37	20
Megaptera novaeangliae	46	18	46	18
Monodon monoceros	36	14	36	14
Odobenus rosmarus	21	17	21	17
Ommatophoca rossi	30	20	20	30
Orcinus orca	19	16	19	16
Pagophilus groenlandicus	-	-	-	-
Peponocephala electra	-	-	-	-
Physeter macrocephalus	33	27	33	27
Pseudorca crassidens	48	11	11	48
Stenella attenuata	54	11	11	54
Stenella clymene	49	14	14	49
Stenella coeruleoalba	42	39	42	39
Stenella frontalis	-	-	-	-
Stenella longirostris	113	1	-	-
Steno bredanensis	-	-	-	-
Tursiops truncatus	13	11	13	11
Sums	908	412	542	539

2.4.1. Signal

Several researches that address audio classification perform the feature extraction in the visual domain, typically using spectrogram images. Investigations have already been developed to handle tasks such as infant cry motivation [7], music genre classification [8] and music mood classification [9]. The visual domain has also been used with animal vocalizations, in tasks as species identification and detection [4,10].

Spectrograms are time-frequency representations of a signal and they can be plotted into an image. From a digital audio, a spectrogram can be generated using the Discrete Fourier Transform (DFT). It shows the intensity of the frequency values as time varies. An example of a spectrogram image of a vocalization of a marine mammal is presented in Figure 2. The X-axis represents time, the Y-axis represents information about the frequency and the Z-axis (i.e. color intensity of the image pixels) displays the intensity of the signal.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1



(a) *Delphinapterus Leucas.* **(b)** *Eubalaena glacialis.*

Figure 2. Examples of spectrograms of the WMMSD.

The spectrogram of the Figure 2a represents a vocalization of a *Delphinapterus Leucas*, the audio is a little bit more than one second long. Its metadata file describe it as a "moan", it was recorded at 1965, in the Coudres Island, Canada. The Figure 2b represents a vocalization of a *Eubalaena glacialis*, the audio is a bit longer than two seconds. The metadata also describes the vocalizations as "moan". It was recorded in the coast of Massachusetts, USA, in 1959.

2.4.2. Features

Texture is an important visual attribute in digital images. In case of spectrograms in particular, texture is a very prominent visual property. In this vein, the textural content of spectrograms has been used in several audio classification tasks, such as music genre classification [11], voice classification [12], birds species classification and whales recognition [4].

In [13], the authors propose the Local Binary Pattern (LBP). The texture of an image is described with a histogram. Each cell of the histogram holds the number of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1

occurrences of a binary pattern. One binary pattern is calculated for each pixel and is based in its neighbourhood. Equation 1 is computed for each pixel for the extraction of the LBP histogram.

$$LBP_{P,R} = \sum_{p=0}^{p-1} s(g_p - g_c)2^p, \tag{1}$$

the parameter P represents the number of neighbour pixels to be taken into account, and R stands for radius, the distance between the pixel and its neighbours. g_c and g_p stand for the gray level of the central pixel (i.e. the pixel for which the LPB is been calculated) and the gray level of one neighbour respectively. The function s is defined in the Equation 2.

$$s(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \tag{2}$$

An example of the computation of LBP in one pixel of an image is illustrated in Figure 3, considering the parameters $P = 8$ and $R = 1$.

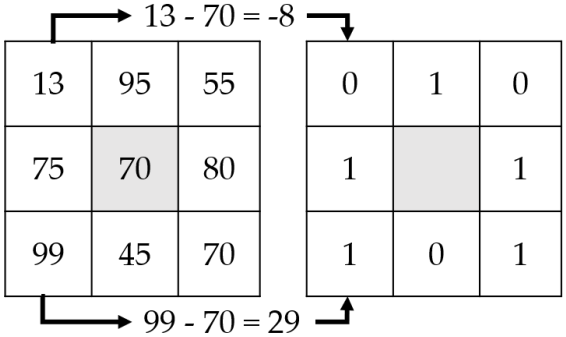


Figure 3. Example of computation of LBP in a gray scale image. Adapted from [14].

The binary pattern begins in the top left and goes clockwise. The pattern of the central pixel of the Figure 3 is defined as the sequence of bits

$$(01011011)_2.$$

It is possible to conceptualize the binary pattern as a decimal number, it is computed applying the sum of binary digits times their power of two (2^n)

$$(0 \times 128) + (1 \times 64) + (0 \times 32) + (1 \times 16) + (1 \times 8) + (0 \times 4) + (1 \times 2) + (1 \times 1) = 91_{10},$$

therefore, the decimal number of the pattern from the Figure 3 is 91.

Changes in the parameter P imply in changes in the number of features, eight neighbours binary described generate 256 possible patterns (2^8). LBP with $P = 8$ presents several non-uniform patterns, which are binary sequences that present more than two bitwise changes, for example, the binary pattern in Figure 3 is non-uniform, it presents six bitwise changes. Usually, LBP is used with $P = 8$, $R = 2$ and gathering together all the non-uniform patterns into just one feature, it results in 59 features. However, Ojala *et al.* [15] observed that non-uniform patterns do not contain fundamental properties of texture, and they suggest to sum up all the non-uniform patterns in the same feature.

2.4.3. Classifiers

To analyze the impact of the PAM-filter we tried several classifiers to observe their performances. The first classifier considered to this work is the well-known K-Nearest

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1

Neighbours (KNN). It is an instance-based algorithm, which means it does not produce a model, it only stores the instances of training. During test, the algorithm computes the distance between each test sample to each training sample. The prediction of a test sample is based on the classes of the nearest neighbours. Parameters of KNN are the distance metric and K, the number of neighbours. Other simpler classifiers used in our experiments are the Naïve Bayes (NB), a probabilistic classifier that is known for assuming independence between the features, and Decision Trees (DT), a tree composed of conditional statements created using information gain or impurity metrics.

Another classifier selected for research is the Support Vector Machine (SVM), a binary classifier proposed by Cortes and Vapnik [16]. It can be easily applied for multi-class problems using one of the following strategies: one-vs-one or one-vs-all. Three considerable benefits of SVM are the kernel method, maximum-margin hyperplane and soft margin. Kernel method consists in mapping the feature space into another feature space, dimensionally higher, where the data can be linearly separable. The maximum-margin hyperplane divides the data taking into account that the nearest points on each side is equally distant from the hyperplane. Soft margin is a technique to reduce overfitting, it treats the C samples nearest the margin as outliers and, with that, increases the distance between the classes. Parameters of SVM are the C and kernel function (which may have its own parameters).

Ensembles of classifiers based in decision trees are also used here. They are Bagging, Random Forest (RF), Extremely Randomized Trees (ERT), AdaBoost (AB) and Gradient Boosting (GB). The classifier Bagging builds several instances of classifiers from random subsets of the training database. RF combines the concept of Bagging but with the idea of random subsets of features per classifier. ERT is similar to Random Forest, the differences are in the choice of attributes and in the definition of cut-points, they are fully random. AB generates new classifiers by increasing the weight of samples that were wrongly classified by the previous models. Then, the outcome is obtained by the weighted predictions of all created models. GB is similar to AB, but the new classifiers are created using only the residual error from the previous classifier.

2.4.4. Deep Learning

To diversify the experiments of this work, we also executed deep learning tests. We use a pre-trained ResNet-50 (Residual Networks - 50 layers) [17] fine-tuned with the training samples. As it is common in convolutional neural networks, we use the spectrograms as input. The deep learning model was used here such a way that it provided features, in a non-handcrafted fashion, and it also performed the classification. So, in the deep learning experiments, the classifiers described in Section 2.4.3 and the features described in Section 2.4.2 are not applicable.

2.5. Experimental Methodology

The methodology of the non deep learning experiments of this work is illustrated in Figure 4. The first step is to acquire the database. WMMSD is available online, but the download is exclusively sample by sample. A crawler script was developed to download all the samples of the database and the metadata files. After that, as a pre-processing task, all the samples were converted to the same sample rate, 22050Hz. To accomplish this task we used LibRosa[8]

The audio files are converted to spectrogram images using the software SoX[9]. Features of LBP was extracted with the software library Scikit-Image[10]. The features from

[8] https://librosa.org/doc/latest/index.html

[9] http://sox.sourceforge.net/

[10] https://scikit-image.org/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1

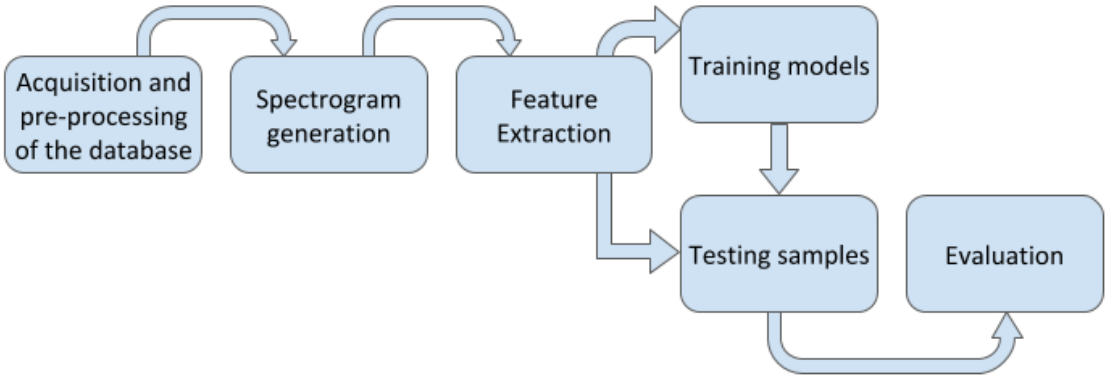


Figure 4. Illustration of the general methodology of this work.

the training sets were used to create a model with the library Scikit-Learn[11]. Features of the testing samples were, then, predicted by the model.

The experiments with deep learning are slightly different. A specific representation of it is presented in Figure 5. A pre-trained ResNet-50 were fine-tuned with the spectrograms of the training samples (the same generated in Figure 4). After that, the spectrograms of the testing samples are predicted with the neural network. The deep learning experiments were carried out using Matlab[12].

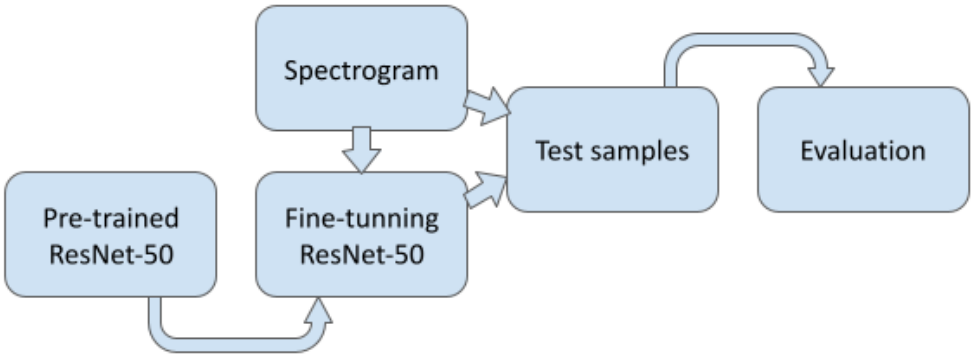


Figure 5. Illustration of the methodology of the deep learning experiments.

All the experiments were conducted in the three Watkins experimental protocols. WEP#1, the protocol composed of 31 classes, samples are randomly divided into ten-fold cross-validation, without concerns about using the same individuals or same noise pattern in training and test sets. WEP#2 and WEP#3, both with PAM-filter, WEP#2 has 24 classes and is a training and test protocol, WEP#3 holds 20 classes and it is a two-fold cross-validation protocol.

3. Results

Table 3 presents the results obtained with the three protocols presented in Section 2.3, WEP#1 without PAM-filter, and WEP#2 and #3 with PAM-filter. The best results of each experimental protocol were all found with deep learning.

The best results found using the WEP#1 was with deep learning and the classifier SVM, 78.10%σ2.73 and 64.62%σ3.49, respectively. The protocol holds 1,645 samples from 31 classes. It was randomly divided in ten folds for cross-validation, it does not apply PAM-filter.

[11] https://scikit-learn.org/stable/

[12] https://www.mathworks.com/products/matlab.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2021 doi:10.20944/preprints202104.0766.v1

Table 3. Experimental results in WMMSD. Features extracted with LBP and several different classifiers. Also, experiments with a deep learning architecture.

Classifiers	Parameters	Acc. and standard deviation (when applicable)		
		WEP#1 (31 classes)	WEP#2 (24 classes)	WEP#3 (20 classes)
NB	Gaussian Naive Bayes	38.78%σ3.76	18.45%	16.56%σ2.29
DT	Splitting with entropy	39.89%σ3.82	11.65%	12.39%σ0.47
	Splitting with Gini Impurity	43.25%σ2.80	10.44%	10.36%σ0.74
KNN	Manhattan distance, K=1	59.53%σ4.51	18.45%	17.02%σ1.76
	Manhattan distance, K=3	61.03%σ3.64	16.02%	17.29%σ3.46
	Manhattan distance, K=5	61.08%σ3.13	17.48%	17.38%σ4.38
	Manhattan distance, K=7	59.54%σ3.36	16.50%	17.38%σ4.90
	Manhattan distance, K=11	59.13%σ4.06	17.48%	16.09%σ4.12
	Euclidean distance, K=1	57.16%σ4.32	19.17%	17.20%σ2.03
	Euclidean distance, K=3	57.83%σ3.56	17.72%	17.20%σ3.33
	Euclidean distance, K=5	58.20%σ3.60	18.20%	18.86%σ5.16
	Euclidean distance, K=7	57.53%σ3.46	17.96%	17.94%σ5.42
	Euclidean distance, K=11	56.29%σ3.22	17.72%	17.38%σ4.39
SVM	Grid search	64.62%σ3.49	13.35%	13.69%σ1.36
Ensembles	AdaBoost	12.09%σ3.55	01.46%	09.62%σ0.22
	Bagging	50.75%σ3.72	13.11%	13.13%σ1.26
	Extremely Randomized Trees	46.91%σ4.66	16.02%	14.24%σ1.51
	Gradient Boosting	54.54%σ3.78	12.38%	11.84%σ0.31
	Random Forest	47.29%σ4.51	13.11%	15.07%σ2.43
Deep Learning	ResNet-50	78.10%σ2.73	24.27%	21.38%σ1.82

On the other hand, the protocols with PAM-filter present much lower results, with or without deep learning. The best results with WEP#2 was 24.27% and 19.17%, it is a train/test protocol. Although it has fewer samples than WEP#1, 1,320, it also hold fewer classes, 24. These results was achieved with deep learning and the KNN classifier.

The two-fold cross-validation protocol with PAM-filter, WEP#3, also got its best results with deep learning and KNN (but with a different number of neighbours). The best accuracies were 21.38%σ1.82 and 18.86%σ5.16. Whereas WEP#3 contains fewer samples than WEP#1, 1081, it also deals with fewer classes, 20.

4. Discussion

First, it is important to reiterate the potential of ResNet-50. One deep learning architecture outperformed all the other classifiers, and this architecture were not trained from scratch. The initial weights were generated with a general propose image database [13] and only the fine-tuning was carried out with vocalizations of marine mammals in the spectrogram format.

In general, the results indicate that it is more likely to achieve better accuracies without PAM-filter, since WEP#1 present accuracies much higher than the other two protocols, WEP#2 and WEP#3, which apply PAM-filter. Therefore, the outcome corroborates with the hypothesis that PAM based in machine learning can be biased by individuals, noise and/or devices used in the recording.

Further, we performed the Friedman statistical test [18,19] in the accuracies obtained with each combination of protocol and classifiers. The test compares the means of at least three samples, it is similar to ANOVA (Analysis of variance), but non-parametric. In our case, the samples are the protocols of the database, 2 with PAM-filter and the other one without it. The null hypothesis of the Friedman test stands that there is no

[13] http://www.image-net.org/

 doi:10.20944/preprints202104.0766.v1

difference between the samples. The result of the Friedman test indicated that at least one of the samples (protocols) are significantly different from another, with $\alpha < 0.05$.

At this point, to find which one is different from the others, we execute the Wilcoxon signed rank test, in pairs of protocols. Since there are three samples, we also apply Bonferroni correction to avoid the statistical error Type 1. So, now, $\alpha_b < 0.167$. The null hypothesis of the Wilcoxon test argues that there is no significant difference between the pairs. The hypothesis was retained only with the pair WEP#2 and WEP#3, in both comparisons with WEP#1 it was rejected.

5. Conclusions

PAM systems based in machine learning can be used to support several different application tasks. However, the evaluation protocol is a critical point, that must be carefully crafted, not to perpetrate wrong assumptions that could compromise the system as a whole.

Unfortunately, wildlife databases with information such as dates, locations, individuals and devices used in the recordings are not easily found. But our results suggest that they must be used to create appropriated experimental protocols.

Author Contributions: Conceptualization, L.N., Y.C. and C.S.Jr.; methodology, R.A., G.M., L.N. and C.S.Jr.; software, R.A. and G.M.; validation, R.A.; formal analysis, R.A.; investigation, R.A., G.M., L.N., Y.C. and C.S.Jr.; resources, R.A., G.M., L.N., Y.C. and C.S.Jr.; data curation, R.A. and G.M.; writing—original draft preparation, R.A.; writing—review and editing, L.N., Y.C. and C.S.Jr.; visualization, R.A, L.N., Y.C. and C.S.Jr.; supervision, L.N., Y.C. and C.S.Jr.; project administration, L.N., Y.C. and C.S.Jr.; funding acquisition, R.A., L.N. and C.S.Jr. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by Araucaria Fundation; National Council for Scientific and Technological Development (CNPq); Coordination of Superior Level Staff Improvement (CAPES) and National Council of State Research Support Foundations (CONFAP).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.14068106.

Acknowledgments: The authors are grateful to NVIDIA Corporation for supporting this research with the donation of a Titan XP GPU.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AB	AdaBoost
ANOVA	Analysis of Variance
DT	Decition Trees
DTF	Discrete Fourier Transform
ERT	Extreme Random Forest
GB	Gradient Boosting
KNN	K-Nearest Neighbours
LBP	Local Binary Pattern
NB	Naïve Bayes
PAM	Passive Acoustic Monitoring
ResNet-50	Residual Network - 50 layers
RF	Random Forest
SoX	SOund eXchange
SVM	Support Vector Machine
WEP	Watkins Experimental Protocol
WMMSD	Watkins Marine Mammal Sound Database

References

1. Bittle, M.; Duncan, A. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. Annual Conference of the Australian Acoustical Society 2013, Acoustics 2013: Science, Technology and Amenity, 2013, pp. 208–215.
2. Mellinger., D.; Barlow, J. Future directions for acoustic marine mammal surveys: Stock assessment and habitat use. Technical report, National Oceanic and Atmospheric Administration, 2003.
3. Freitas, G.K.; Costa, Y.M.G.; Aguiar, R.L. Using spectrogram to detect North Atlantic right whale calls from audio recordings. 2016 35th International Conference of the Chilean Computer Science Society (SCCC), 2016, pp. 1–6. doi:10.1109/SCCC.2016.7836034.
4. Nanni, L.; Aguiar, R.L.; Costa, Y.M.G.; Brahnam, S.; Silla Jr., C.N.; Brattin, R.L.; Zhao, Z. Bird and whale species identification using sound images. *IET Computer Vision* **2018**, *12*, 178–184. doi:10.1049/iet-cvi.2017.0075.
5. Nanni, L.; Costa, Y.M.G.; Aguiar, R.L.; Mangolin, R.B.; Brahnam, S.; Silla Jr., C.N. Ensemble of convolutional neural networks to improve animal audio classification. *EURASIP Journal on Audio, Speech, and Music Processing* **2020**, *2020*. doi:10.1186/s13636-020-00175-3.
6. Flexer, A. A Closer Look on Artist Filters for Musical Genre Classification. Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, 2007, pp. 341–344.
7. Felipe, G.Z.; Aguiar, R.L.; Costa, Y.M.G.; Silla Jr., C.N.; Brahnam, S.; Nanni, L.; McMurtrey, S. Identification of Infants' Cry Motivation Using Spectrograms. 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), 2019, pp. 181–186.
8. Nanni, L.; Costa, Y.M.G.; Aguiar, R.L.; Silla Jr., C.N.; Brahnam, S. Ensemble of deep learning, visual and acoustic features for music genre classification. *Journal of New Music Research* **2018**, *47*, 383–397.
9. Tavares, J.C.C.; Costa, Y.M.G. Music mood classification using visual and acoustic features. 2017 XLIII Latin American Computer Conference (CLEI), 2017, pp. 1–10.
10. Merchan, F.; Guerra, A.; Poveda, H.; Guzmán, H.M.; Sanchez-Galan, J.E. Bioacoustic Classification of Antillean Manatee Vocalization Spectrograms Using Deep Convolutional Neural Networks. *Applied Sciences* **2020**, *10*, 3286. doi:10.3390/app10093286.
11. Costa, Y.M.G.; Oliveira, L.; Koerich, A.; Gouyon, F. Music Genre Recognition Using Spectrograms. International Conference on Systems, Signals and Image Processing; , 2011.
12. Montalvo, A.; Costa, Y.M.G.; Calvo, J.R. Language identification using spectrogram texture. Iberoamerican Congress on Pattern Recognition. Springer, 2015, pp. 543–550.
13. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. Pattern Recognition, 1994, Vol. 1, pp. 582–585.
14. Lakshmiprabha, N.S. Face Image Analysis using AAM, Gabor, LBP and WD features for Gender, Age, Expression and Ethnicity Classification. *Computing Research Repository (CoRR)* **2016**.
15. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2002**, *24*, 971–987. doi:10.1109/TPAMI.2002.1017623.
16. Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297. doi:10.1007/BF00994018.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
18. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
19. Derrac, J.; García, S.; Molina, D.; Herrera, F. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation* **2011**, *1*, 3–18.