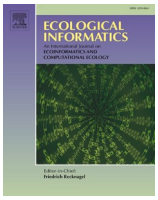
Ecological Informatics 81 (2024) 102642

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



From citizen science to AI models: Advancing cetacean vocalization automatic detection through multi-annotator campaigns

Gabriel Dubus [a,b,*], Dorian Cazau [b], Maëlle Torterotot [b], Anatole Gros-Martial [b,c,d], Paul Nguyen Hong Duc [e], Olivier Adam [a]

[a] Sorbonne University, CNRS, Institut d'Alembert UMR 7190, LAM, Paris, France
[b] ENSTA Bretagne, Lab-STICC, UMR CNRS 6285, Brest, France
[c] CEBC UMR 7372, CNRS, La Rochelle University, France
[d] Brest University, CNRS, Ifremer, IUEM, UMR6538 Geo-Ocean, Plouzané, France
[e] Centre for Marine Science and Technology, Curtin University, Bentley, WA 6102, Australia

ARTICLE INFO

Keywords:
Marine bioacoustics
Passive acoustic monitoring
Citizen science
Multi-annotation
Deep learning for automatic detection
Convolutional neural networks
Soft labeling

ABSTRACT

Continuous underwater Passive Acoustic Monitoring (PAM) has emerged as a strong tool for cetacean research. To handle the vast volume of collected data, it is essential to employ automated detection and classification methods. The recent advancement of deep learning, involving model training and testing, requires a large amount of labeled data. These labels are derived through the manual annotation of audio files often reliant on human experts. Based on an annotation campaign focusing on blue whale calls in the Indian Ocean involving 19 novice annotators and one expert in bioacoustics, this study explores the integration of novice annotators in marine bioacoustics research, through citizen science programs, which could drastically increase the size of labeled datasets and enhance the performance of detection and classification models. The analysis reveals distinctive annotation profiles influenced by the complexity of vocalizations and the annotators' strategies, ranging from conservative to permissive. To address the challenges of annotation discrepancies, Convolutional Neural Networks (CNNs) are trained on annotations from both novices and the expert. The results show variations in model performance. Our work highlights the importance of annotation guidelines encouraging a more conservative approach to improve overall annotation quality. In an effort to optimize the potential of multi-annotation and mitigate the presence of noisy labels, two annotation aggregation methods (majority voting and soft labeling) are proposed and tested. The results demonstrate that both methods, particularly when a sufficient number of annotators are involved, significantly improve model performance and reduce variability: the standard deviation of the area under PR and ROC curves fall under 0.02 for both vocalizations with 13 aggregated annotators, while it was at 0.17 and 0.21 for the Blue Whale Dcalls and 0.05 and 0.04 for the SEIO PBW vocalizations with all annotators separately. Moreover, these aggregation methods enable the training of models using non-expert annotations that achieve performance of models trained with expert annotations. These findings suggest that crowdsourced annotations from novice annotators can be a viable alternative to expert annotations.

1. Introduction

Continuous underwater Passive Acoustic Monitoring (PAM) conducted over extended periods has emerged as a pivotal tool for studying cetaceans as they rely on sound for essential activities and social interactions [Krause, 1987; Leroy et al., 2018a; Torterotot, 2020]. The amassed acoustic data furnishes invaluable insights across various dimensions of cetacean ecology, encompassing migration routes, population dynamics and behaviors [Courts et al., 2020; Yurk et al., 2002]. Yet, the colossal volume of acoustic data collected over the years calls for automated methodologies for detecting and classifying acoustic events. Recent strides in this arena have been made with the application of deep learning models, yielding efficacious outcomes [Shiu et al., 2020; Usman et al., 2020; Miller et al., 2022].

However, supervised methods, including deep learning approaches, demand a "ground truth" generated by human experts to create training

* Corresponding author at: Sorbonne University, ∂'Alembert, boîte 162, 4, Place Jussieu 75252, Paris, Cedex 05, France.
E-mail address: gabriel.dubus@dalembert.upmc.fr (G. Dubus).

https://doi.org/10.1016/j.ecoinf.2024.102642
Received 5 February 2024; Received in revised form 8 April 2024; Accepted 8 May 2024
Available online 14 May 2024

datasets and evaluate model performance. This ground truth is established through manual annotations of the audio data usually represented as spectrograms. Typically, one or two experts undertake this annotation process, but discerning unfamiliar acoustic features can prove more intricate than identifying commonplace objects like cats and dogs in images. Consequently, this process is laborious, time-intensive, and subject to human subjectivity [Nguyen Hong Duc et al., 2021a; Leroy et al., 2018b]. Additionally, assessing the quality of collected annotations is a challenge. To mitigate this challenge, the standard practice in underwater PAM studies is to annotate only a fraction of datasets [Usman et al., 2020; Miller et al., 2021; Solsona-Berga et al., 2020]. The repercussions of limited annotations manifest as difficulties in the model's ability to generalize due to the lack of reference data. This is especially pronounced given the existence of species-specific or geographical sound variations, variability in acoustic recording quality, and the wide range of non-standardized vocalizations.

Engaging citizen scientists within the marine bioacoustics research community emerges as a promising strategy to overcome annotation shortfalls [Shamir et al., 2014]. This approach allows access to an increased amount of annotated data for training and validating detection and classification models, which would enhance their capacity for generalization [Kosmala et al., 2016; McClure et al., 2020]. Nevertheless, several studies highlight inter-annotator variability, especially with novice annotators [Nguyen Hong Duc et al., 2021a; Leroy et al., 2018b; Dubus et al., 2023], often found in citizen sciences programs. As a consequence, training deep learning models on erroneous additions of labels could result in lower performance [Song et al., 2022; Frenay and Verleysen, 2014].

Based on an annotation campaign conducted on sounds emitted by two species of blue whales in the Indian Ocean, this paper presents a comparison of convolutional neural networks (CNN) trained from annotations of an expert and 19 novice annotators.

The first part of this study evaluates the inter-annotator variability with metrics generally used to evaluate detection models, which are precision and recall, calculated here between each pair of annotators.

In a second part, CNNs were trained based on the annotation from each annotator to assess the impact of the inter-annotator variability (i.e., the uncertainty regarding the noise on a labeled dataset) on models for automatic detection. All models are evaluated on datasets different from the one used for training, in terms of recording devices, geographical areas and annotator that produced the annotation for pseudo-ground truth. The performances of those models are then assessed by taking into account the agreement between the annotations of novices and those of the expert.

In order to reduce the variability due to the different novice annotators and increase the performance of models without a priori knowledge of the quality of the annotations produced by novices, two methods are proposed to aggregate the annotations: majority grouping and soft labeling. Ultimately, the objective of these methods is to ensure the quality of the aggregated annotation and the possibility of training high-performance models with it, without any regards on the quality of each individual annotation set.

Finally, multiple guidelines are proposed to manage citizen sciences for PAM studies.

2. Materials and methods

2.1. Datasets

Audio recordings used in this study originally came from three different datasets presented below. The first one, named AmStP here, is annotated by one expert and 19 novice annotators and is used as the development set for the training stage only. The second and third datasets, SWAMS and ElephantIsland2013 respectively, are used only as evaluation sets and were annotated by two different experts (and no novices), differents from the one that has annotated AmStP.

2.1.1. Development set – AmStP

The dataset annotated used for the present study is made of 762 h of audio signal recorded off Amsterdam and Saint Paul, two French sub-Antarctic islands in the Indian Ocean from February 28 to April 5, 2019 [Torterotot et al., 2022]. The acoustic signals were recorded continuously using a HTI92 WB hydrophone mounted on a SeaExplorer glider, sampled at 48 kHz and coded on 16 bits.

The whole dataset was manually annotated by an expert in bioacoustics, one of the authors of the original dataset publication. A hundred 10-min files downsampled at 250 Hz (16 h and 40 min) were kept for the annotation campaign with 19 novice annotators. They were asked to annotate South Eastern Indian Ocean Pygmy Blue Whales (SEIO PBW) vocalizations and blue whale's D-call type vocalizations (Dcall) [Torterotot et al., 2022]. Fig. 10 (in Appendix) presents the number of calls per species identified by each annotator. Each set annotated by an annotator was used as a different development set for CNNs. Two aggregating methods to constitute development sets for CNN training are also proposed and tested.

2.1.2. Evaluation sets - SWAMS and ElephantIsland2013

To evaluate the performance of the CNNs trained on different annotation sets, datasets containing other SEIO PBW vocalizations and blue whale's Dcall were chosen. Only a small portion of those datasets were used to reduce computational time, as hundreds of networks were trained and compared.

For the evaluation of the SEIO PBW vocalizations, 7 h of audio signals recorded during the OHASISBIO program were used [Royer, 2009; Torterotot et al., 2020]. These recordings were collected at the SWAMS site, located in the Indian Ocean's oceanic zone, between Kerguelen and Amsterdam Island. This dataset is called SWAMS in this paper. The recordings were made in March 2015 using a hydrophone deployed at a depth of 1000 m. The sampling rate was 240 Hz. A total of 102 vocalizations were manually annotated by a second expert.

For the evaluation of the models' performance on blue whale's Dcalls, we used the underwater acoustic dataset recorded during 2013, off Elephant Island, North of the Antarctic Peninsula [Miller et al., 2021]. A total of 16 h and 55 min have been randomly selected, including more than 600 annotated vocalizations annotated by a third expert. The sampling rate was 250 Hz.

Geographical positions of each site are reported on a map centered on Antarctica, Fig. 1.

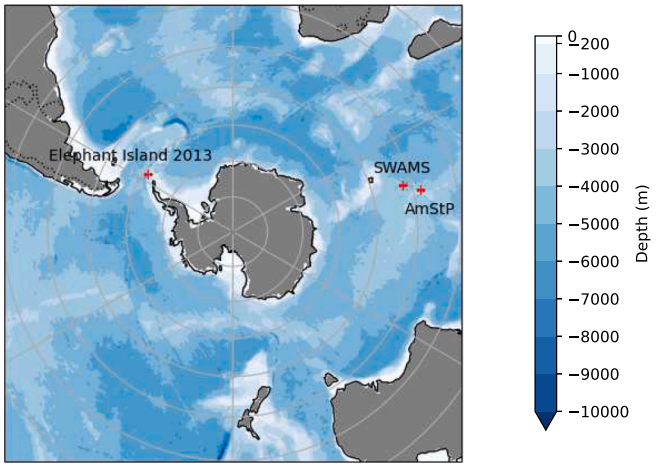


Fig. 1. Map of Antarctic underwater recording sites illustrating sites used in this study (red crosses). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

2.2. Annotation protocol

2.2.1. Multi-annotation campaign with novice volunteer annotators

In this study, 19 volunteer annotators were enlisted to provide annotations. They were recruited online, with the network of the association Astrolabe Expeditions.[1] The objective was to recognize SEIO PBW vocalizations and blue whale Dcalls. More specifically, they were asked to annotate only the third harmonic of the second part of the PWB calls (see Fig. 2). This part is known to be the loudest one in the SEIO PBW vocalization, with the better signal-to-noise ratio (SNR) and is therefore the easiest one to identify on spectrograms [Gavrilov et al., 2011]. All annotators in this campaign were novices: inexperienced in both PAM and underwater recording annotation. Prior to annotating acoustic recordings, they were provided a 1-h virtual training session. The objective was to give them general information about underwater soundscapes, to let them listen to cetacean vocalization samples and to introduce them to both time-frequency representations and the vocalizations of interest. When introducing the SEIO PBW and Dcalls, key features for their identification were presented, such as the time duration and the frequency bandwidth. The annotation campaign spanned a duration of one month. Throughout this period, annotators were encouraged to seek clarification by emailing our team all along the annotation phase.

2.2.2. Annotation platform for ocean sound explorers (APLOSE)

The web-based annotation platform APLOSE[2] [Nguyen Hong Duc et al., 2020; Keribin et al., 2024] was used to annotate acoustic events present in the audio recordings. Each user had login credentials, granting them access to the selected datasets, and a unique identifier (ID). They can visualize pre-computed spectrograms for visual inspection and listen to the acoustic recordings to confirm or not the presence of the vocalizations of interest. Playback controls allow users to play/pause the sound file and adjust playback speed (from 0.25× to 4×) to be able to hear the low frequency sounds studied in this work. A complete user guide for APLOSE can be found on our GitHub repository.[3] For each identified vocalization, annotators were asked to draw a time-frequency box around the identified sound to delimit its start/end times and frequencies. For each box, the annotators had to assign a class name from a given list corresponding to the sound they recognized. Finally, all annotations were continuously collected and automatically written to a downloadable CSV file, containing the name of the audio file, the ID of the annotator, the vocalization identified, the start and end time (relative and absolute), and the start and end frequencies of the time-frequency box. In order to mitigate the potential over-representation of the first annotations by influence, the platform ensured that annotators were unable to access annotations created by others.

2.2.3. Dealing with several novice annotators

Previous work assessing the inter-annotator variability, especially between novice annotators, shows that this leads to disagreement with the expert or systematic errors of annotation by novice annotators in labeling campaigns of underwater sounds in PAM studies [Dubus et al., 2023]. To reduce the disagreement between annotators in this work, a majority voting approach was used: a given sample was considered annotated by the group if more than half of the group annotated it. Therefore, a singular aggregated annotation was created from the novice annotations and compared to an expert annotation in Section 3.3. To consider that two annotations overlap, the overlap time needed to represent at least 20% of the total time-duration covered by the two boxes. This value was based on the histogram of all the time-overlap between boxes and allows to prevent the overlapping of two boxes drawn for different, but close, sound events. The coordinates of the shared annotation box were computed as an average of the coordinates of every annotator's box. A graphical example of the proposed method with three annotators is presented in Appendix (Fig. 11).

Detection models have been trained for different aggregated annotation sets, and compared with a model trained with the annotation of the expert.

2.3. Automatic detection using CNN

A CNN was designed and trained on different annotation sets. This type of model is already widely used for image classification and has also shown good results in bioacoustics for automatic detection of sound events [Shah et al., 2018].

2.3.1. Data preparation

The acoustic recordings were resampled at 250 Hz and cut into 50-s long sections to cover the frequency range and duration of the targeted vocalizations [Miller et al., 2021; McDonald et al., 2023; Nguyen Hong Duc et al., 2020]. They were then filtered by a high-pass filter at 5 Hz using a Butterworth filter at order 10 and normalized by the energy calculated on a sliding 1-h window centered on the sample. Power spectrograms were computed for each 50-s section with an analysis window of 512 samples (around 2.05 s) using a Hanning window, overlapped by 471 samples (1.89 s). The time and frequency resolutions of the spectrograms were respectively 0.16 s and 0.5 Hz, yielding 2D matrices of shape 292*256. Finally, each spectrogram was visualized in dB and thresholded between −20 dB and + 20 dB to optimize the visualization.

Each 50-s section was then characterized as a positive section (i.e., a section with a vocalization) or a negative section (i.e., a section without vocalization). As some annotations overlapped between two 50-s sections, a minimal threshold in the percentage of the annotation present within the 50-s section is set to avoid: (i) an incomplete but still consequent part of vocalization in negative sections, and (ii) positive sections which only contains a small part of an annotation and therefore of a vocalization. Above this threshold, the section is considered positive, otherwise, it is considered negative. Taking the margin between the vocalization on the spectrogram and the time-frequency boxes drawn around into account, this threshold is set at 20%.

The number of 50-s sections for each dataset is presented in Table 1. For the dataset AmStP, the percentage of positive sections is given as mean and standard deviation, computed on all annotations (19 novices and 1 expert). As the number of positive sections in the training set (dataset AmStP) for the blue whale's Dcall was particularly low, models trained on different ratios of positive/negative sections were tested from the annotation sets produced by the expert, by randomly removing negative sections from the testing set (undersampling method [Johnson and Khoshgoftaar (2019)]). Ultimately, a positive section ratio of 0.2 was practically determined to enhance the model's ability to learn from the data.

2.3.2. From multi-annotation to soft labels for training

Recent studies proposed that using explicit training models with soft labels may mitigate the disparity arising from human uncertainties and enhance overall generalization performance [Peterson et al., 2019; Du et al., 2022]. The annotation process usually employed in PAM studies does not provide probabilities for each annotation. From our annotation campaign, enough annotations were provided to transform the binary labels (*aka* Hard labels), to probabilities or soft labels for training purposes. For each sample, the soft labels used for the training is n/N with N the number of annotators that have checked the sample and n the number of annotators that have annotated it positively.

Fig. 3 presents three samples for the two distinct vocalizations types with the soft label associated. As expected, for both vocalizations,

[1] https://www.astrolabe-expeditions.org/
[2] https://osmose.ifremer.fr/app/
[3] https://github.com/Project-OSmOSE/osmose-app/blob/master/docs/user_guide_annotator.md

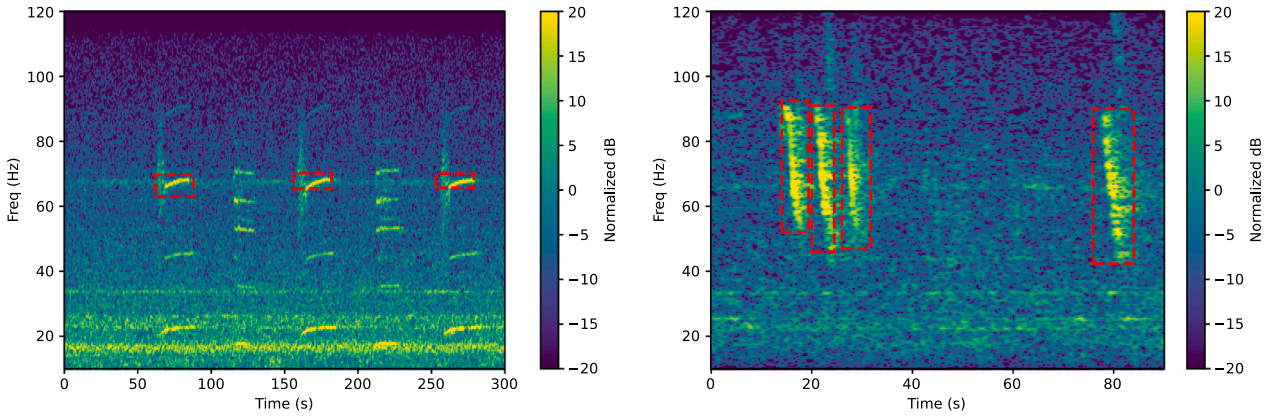


Fig. 2. Spectrograms of the two blue whale's vocalizations annotated during the campaign. Reds doted rectangles correspond to the annotation of the expert. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Datasets and number of positive samples used for each task in the proposed studies.

Sec	Method used	Number of annotators	Sound event to detect	Dataset used for training	Number of sample (percentage of positive)	Dataset used for evaluation	Number of sample (percentage of positive)	Number of trained models
3.2	Mono annotators Hard label	1	SEIO PBW	AmStP	1205 (45.66 +/− 5.82)	SWAMS	503 (30.81)	20 (1 from the expert, 19 from novices)
3.2	Mono annotators Hard label	1	Blue whale's Dcall	AmStP	323 (18.27 +/− 9.48)	Elephant Island 2013	1218 (49.91)	20 (1 from the expert, 19 from novices)
3.3	Multi-annotator majority voting	2–13	SEIO PBW	AmStP	1205 (41.55 +/− 3.31)	SWAMS	503 (30.81)	120 (10 for each size of aggregated subgroup)
3.3	Multi-annotator majority voting	2–13	Blue whale's Dcall	AmStP	323 (12.34 +/− 1.65)	Elephant Island 2013	1218 (49.91)	120 (10 for each size of aggregated subgroup)
3.4	Multi-annotators Soft labeling	13	SEIO PBW	AmStP	1205 (31.75 +/− 4.21)	SWAMS	503 (30.81)	10
3.4	Multi-annotators Soft labeling	13	Blue whale's Dcall	AmStP	323 (14.88 +/− 1.47)	Elephant Island 2013	1218 (49.91)	10

samples associated with a soft label of 1 correspond to isolated calls with a good SNR. The second example for SEIO PBW vocalization and Dcall, with a soft label of 0.68 and 0.78 respectively, shows samples with calls but also impulsive noises that could have misled some annotators. Finally, the third example for SEIO PBW vocalization presents the desired third harmonic of the first part of the call, with a low SNR, but also the second part of other calls and an impulsive noise, thus, the majority of novice annotators have missed this call. The last example of blue whale's Dcall has a very low SNR and impulsive noise too.

2.3.3. Architecture of the model used

Because the training sets for both cases were relatively small (Table 1), we implemented a CNN comprising three convolutional layers followed by three fully connected layers (Fig. 4). The training for each vocalization's type was done independently. In order to mitigate the impact of noisy labels on the robustness of the model, a dropout layer (with a dropout rate of 0.25) was incorporated to prevent overfitting [Jindal et al., 2017]. During the training process, we employed a binary cross-entropy loss function that computed the disparity between the annotated binary label and the network's output. The Adam optimizer was used as the gradient descent algorithm. The implementation was carried out using PyTorch [Paszke et al., 2019].

To ensure comparability across trained models, we maintained consistent hyperparameters across all training instances for each task (i.e., SEIO PBW vocalization and Dcall). Specifically, for SEIO PBW vocalization training and blue whale Dcall training, we set the batch sizes to 5 and 4, and the learning rates to 1e-4 and 1e-3, respectively. This decision was based on the fact that more samples were used for SEIO PBW vocalization training. The models were trained for 40 and 25 epochs respectively. However an early stopping method was used to keep the model weights before the models overfitted. The patience parameter is set at 10 epochs for all training phases.

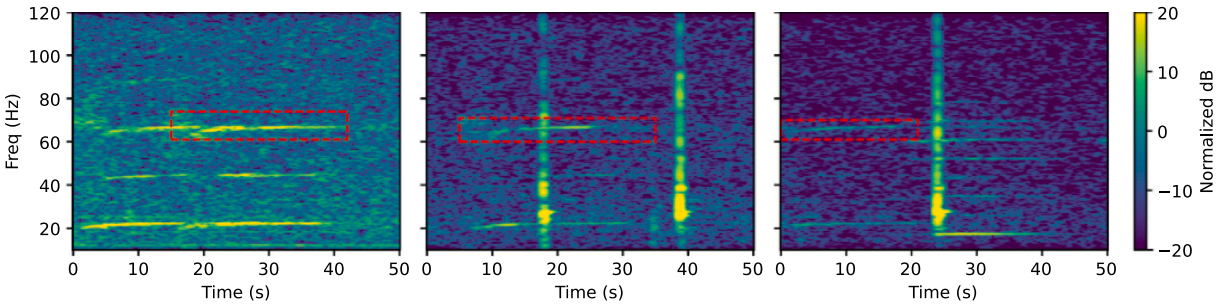
2.4. Evaluation metrics

Considering the absence of absolute ground truth for identifying audio events within underwater acoustic recordings, the annotation sets proposed by the annotators, expert or novice, are called pseudo ground truth in the present study.

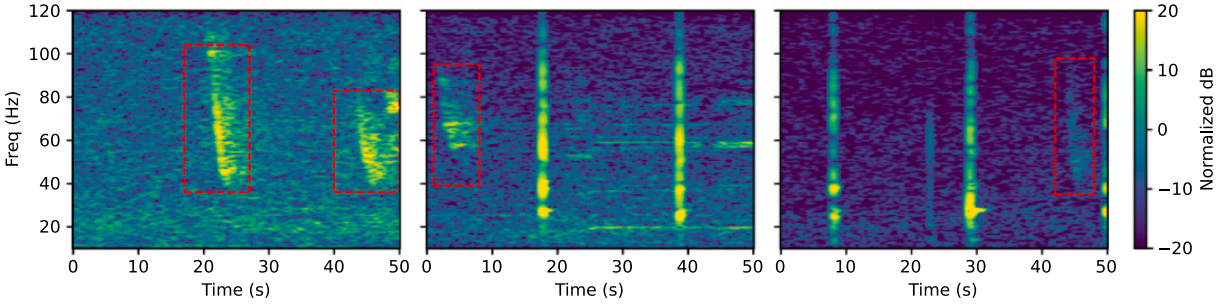
State-of-the-art metrics [Hildebrand et al., 2022] have been used to: 1. Compare a pair of annotators, with each one taking turns as the tester and the pseudo ground truth. 2. Evaluate the performance of the detection models with the annotations from an expert as pseudo ground truth.

For every sample subjected to analysis by either the CNN or an annotator, four possible outcomes are defined:

- True Positive (TP): A call is accurately detected and annotated.
- False Positive (FP): A call is detected, but there is no annotation.
- True Negative (TN): No call is detected and no call is annotated



(a) SEIO pygmy blue whales vocalizations. Soft labels (from left to right) : 1 - 0.68 - 0.28.



(b) Blue whale's Dcall type vocalizations. Soft labels (from left to right) : 1 - 0.78 - 0.25.

Fig. 3. Spectrograms of the two blue whale's vocalizations with different values of soft labels. All vocalizations presented have been annotated positively by the expert, displayed with red doted rectangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

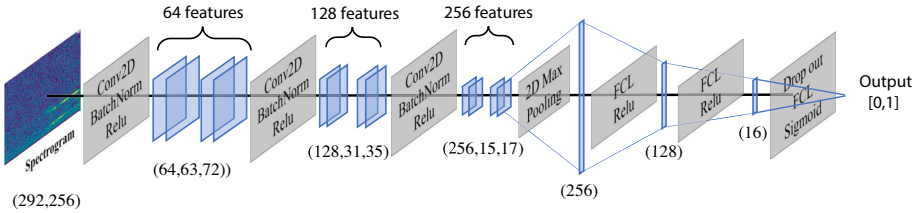


Fig. 4. Design of the CNN structure for automatic detection whale's vocalizations.

- False Negative (FN): No call is detected but a call was annotated

In the case of a comparison of two annotation sets produced by a given pair of annotators, A1 and A2, the false positives of A1 with A2 as pseudo ground truth, equal the false negatives of A2 with A1 as pseudo ground truth. Each pair of annotators, regardless of their expertise, is considered twice here, with each annotator once as the tester and once as the pseudo ground truth.

From these four potential outcomes, Precision, Recall, and False Positive Rate (FPR) are derived:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

The harmonic mean of recall and precision from Section 3.1, called F_1 score, has been used in Section 3.2 to represent with one value the agreement between a novice annotator and the expert:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Precision-Recall curve (PR) and Recall-False Positive Rate curve, usually called Receiver Operating Characteristic curve (ROC), are widely employed in the assessment of automated methods for detection and classification [Hildebrand et al., 2022; Best, 2022]. These curves are generated by adjusting the model's output threshold and subsequently calculating the aforementioned metrics for each threshold value.

Both representations were kept in Section 3.2 as (i) ROC representation is one of the most used representations in PAM for automatic detection of cetacean's low frequency vocalizations and (ii) Hildebrand et al. (2022) enlightened the potential overestimation of the performance using ROC representation due to the unbalanced ratio between FP and TN.

As 130 models were trained and compared in Section 3.3, the area under ROC and PR curves, respectively called AUC (Area Under ROC

Curve) and mAP (mean Average Precision), were computed, to facilitate the comparison. They are described as:

$$AUC = \int_0^1 Recall(x)dFPR(x) \tag{4}$$

$$mAP = \int_0^1 Recall(x)dPrecision(x) \tag{5}$$

with x the threshold applied on the output of the detection model.

3. Results

3.1. Precision and recall of novice annotators

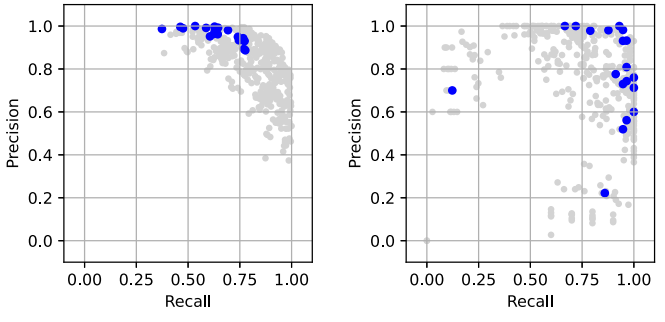
Based on the multi-annotators campaign managed on the dataset AmStP, Fig. 5 presents an analysis of the concordance among pairs of annotators, visualizing the precision and recall scores computed for each duo. Pairs wherein the expert annotations serve as references (or pseudo ground truth) are highlighted in blue: The more an annotator agrees with the expert, the more the corresponding blue dot point is close to the upper-right corner.

For the SEIO PBW vocalizations, results showed that all the novice annotators have good precision: the large majority of the sound events annotated by a novice have also been annotated by the expert. Their recall is lower, between 0.35 and 0.8 (but a large majority over 0.5), thus a substantial part of the vocalizations annotated by the expert were missed by the novices. Concerning blue whale's Dcall annotations, notable distinctions emerged for two annotators who displayed pronounced discordance with the expert. One of these annotators exhibited notably low recall (0.12), indicating a substantial omission of genuine Dcall sounds according to the expert's judgment. Conversely, the other annotator demonstrated low precision (0.22), implying excessive marking of false positives. In contrast, the remaining annotators appeared to be more closely aligned with the expert annotations.

3.2. Performance of the detection regarding the annotation set for training

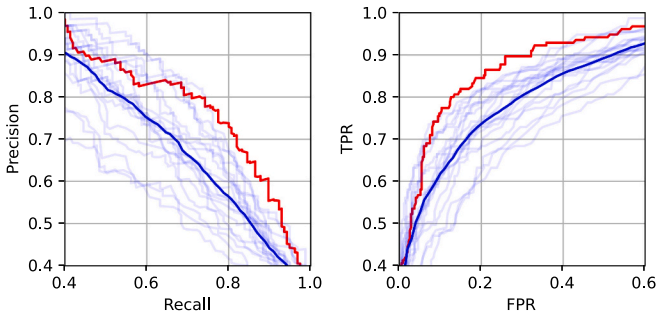
For each annotation provided (one based on the expert annotations and the other on the novices annotations), two models were trained on the dataset AmStP to detect SEIO PBW vocalizations and blue whale's Dcalls, and evaluated on the dataset SWAMS and Elephant Island 2013 respectively. ROC and PR curves computed using the expert annotations as pseudo ground truth for each vocalization are represented on Fig. 6.

As expected, the model trained with the expert annotations gets better results than the other models. The AUC and mAP metrics for the

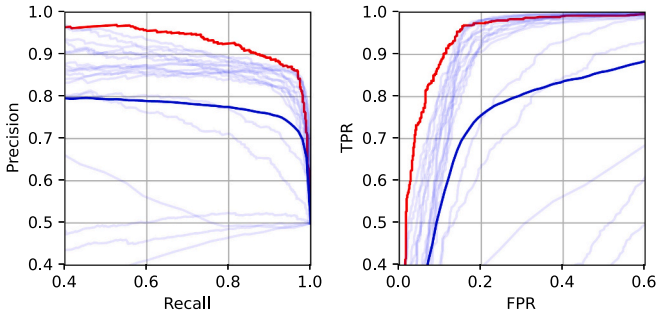


(a) SEIO PBW vocalizations (b) Blue whale's Dcall

Fig. 5. Precision and recall computed for each pair of annotators. Blue dots: results with expert considered as ground truth. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



(a) SEIO PBW vocalizations



(b) Blue whale s Dcall

Fig. 6. Performance of the models trained with all annotation sets provided from the annotation campaign. Red line corresponds to the model trained with the annotations from the expert. Blues lines with low opacity correspond to the model trained with the annotations from the novices. Blue line with high opacity is the mean curve for all novice lines.

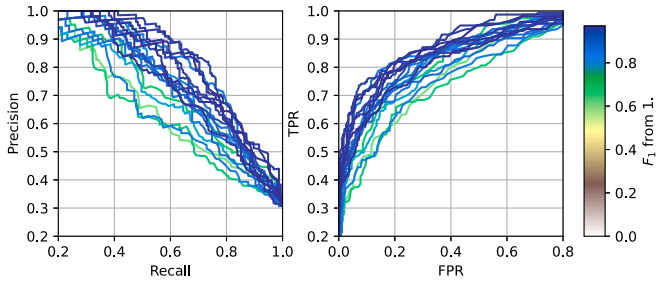
blues curves are 0.85 +/− 0.05 and 0.77 +/− 0.17 respectively, while they reach 0.90 and 0.85 with the model trained from the expert annotations. For the blue whale's Dcall, 13 models out of 19 trained from novice annotations showed performance close to the model trained on expert annotations with AUC and mAP metrics at 0.92 +/− 0.02 and 0.87 +/− 0.3 while the model trained with the expert got 0.96 and 0.94. However, 6 curves show very low performances in comparison to the others, with AUC and mAP of 0.56 +/− 0.21 and 0.57 +/− 0.17.

Performances of each model are then observed by considering the agreement between novice annotations and expert annotations. Fig. 7 presents the results of the models trained on novice annotation, by showing the F1 score of the annotation set computed from the measurements of recall and precision in the previous subsection, using the expert annotations as pseudo ground truth.

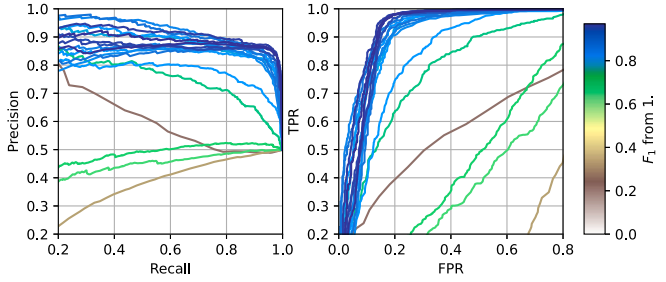
As expected, for both vocalizations, the models trained on the annotation sets closer to the expert annotation set show better performance. For the blue whale's Dcalls, the two annotation sets with systematic errors, corresponding to the two annotators who displayed pronounced discordance with the expert on Fig. 5b, did not allow the model to generalize. Moreover, the four other curves with low AUC and mAP values correspond to models trained on an annotation set with low precision but high recall by comparison to the expert one. All models based on annotation sets with high recall by comparison to the expert one yield performances close to the model trained with the expert annotation.

3.3. Performance of the detection regarding the annotation set for training - aggregated annotation

To remove noise in the novice annotation sets, a majority voting

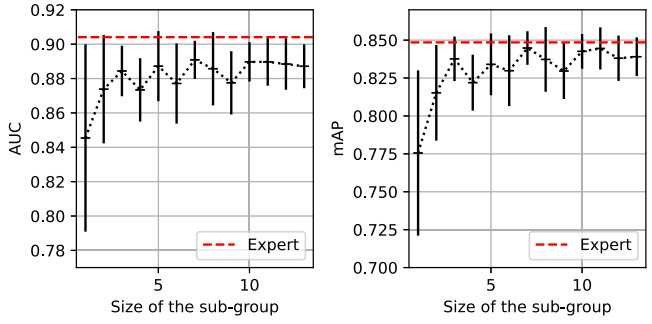


(a) SEIO PBW vocalizations

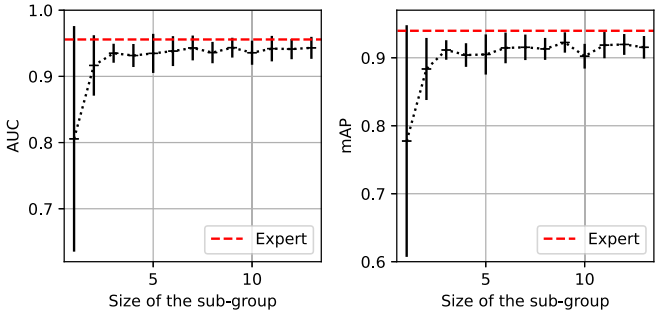


(b) Blue whales Dcall

Fig. 7. Performance of the models trained with all annotation sets of novices provided from the annotation campaign. The colour of the lines correspond to the F_1 score computed on the annotation set with the expert annotation as pseudo ground truth.



(a) SEIO PBW vocalizations



(b) Blue whale's Dcall

Fig. 8. Performance (AUC and mAP) of the models trained with aggregated annotation sets of novices using a major voting. The red dotted line corresponds to the performance of the model trained on the expert.

strategy was used to aggregate annotation sets. Thus, for different sizes of annotator subgroups (between 2 and 13), 10 random selections of novice annotators were realized, creating 10 distinct trainsets. A model was trained on the subgroup annotations from the development set AmStP and evaluated on the evaluation sets: SWAMS and ElephantIsland2013 for SEIO PBW and Dcall, respectively. Fig. 8 shows the mean and standard deviation of AUC and mAP metrics of those 10 models, in comparison with the performance of the model trained with the expert annotation. The results for the group size of 1 are the mean and standard deviation of the 19 models trained with the annotation of the novices, one by one, presented in the previous section.

An increase in both metrics with the number of novices is observed, indicating better performance. The aggregated novices produce on average a better annotation set than the majority of the novices alone. In the same way as the previous observation, the variation of performance for the SEIO PBW vocalizations is lower than the blue whale's Dcall.

Red dotted lines represent the performances of the model trained with the expert annotation set. A Krustkal-Wallis test is performed, using the performances of the largest subgroup size and the performance of the expert on Fig. 8, to evaluate the null hypothesis formulated as "the performance of the model trained with an expert annotation does not significantly differ from performance obtained with models trained on non-experts annotations". *P*-values computed for each metrics (AUC, mAP) are (0.11, 0.75) and (0.11, 0.21) for SEIO PW and Dcall respectively. The null hypothesis cannot be rejected. However, the performances of the model trained with the expert annotation are always higher than the mean performance obtained with models trained with non-expert annotations.

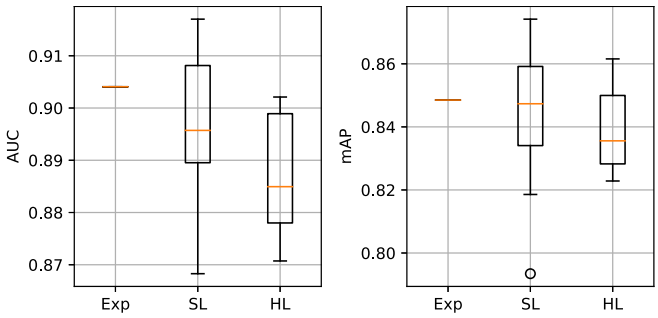
Although performances of the expert are not exactly reached, the improvement of AUC and mAP values and the reduction of the standard deviation with the subgroup size is observed. The evolution of the performance with the subgroup size seems to be asymptotic. A plateau is reached after a given size of the subgroup of novice annotators, around 5 and 10 for the SEIO PBW vocalizations and the blue whale's Dcall, respectively.

3.4. *Assessing the use of soft label from multi-annotation for the training*

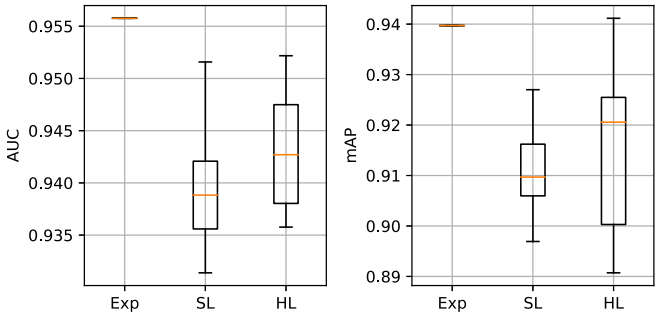
To assess the performance of the soft labeling aggregation method, it was benchmarked against the model trained using the expert annotations and the 10 models trained with the aggregated annotations of 13 annotation sets with hard labels (using the majority voting strategy). 10 models were trained using random selections of 13 annotator to build the soft labels. Fig. 9 presents the mean and standard deviation of the AUC and mAP metrics for all three cases (except for the model trained on the expert as only one annotation set was available).

The mean AUC and mAP are slightly better for the soft labeling methods in the case of the detection of SEIO PBW vocalizations with an increase of 0.01 for the mean of both metrics. Results are slightly lower for the blue whale's Dcalls with a decrease of 0.01 for the mean AUC and 0.005 for the mean mAP. Both methods give results close to the model trained with expert annotations for the detection of SEIO PBW: 0.849 and 0.904 for AUC and mAP respectively. The model trained on the expert annotation set produced better results for the Dcall with 0.94 for AUC and 0.956 for mAP.

A two-sample *Z*-test has been used for each metrics and each label to compare models trained with soft and hard labels. The null hypothesis formulated as "the distribution of the performance of the model trained with soft labels and the distribution of the performance of the models trained with hard labels has the same mean" cannot be rejected. *P*-values computed with each metrics (AUC, mAP) are (0.10, 0.38) and (0.22, 0.56) for SEIO PW and Dcall respectively. Moreover, any of the two methods provides best mean performances for both labels. Thus, no significant improvements between the soft and hard labels from 13 aggregated novice annotators are observed.



(a) SEIO PBW vocalizations



(b) Blue whale's Dcall

Fig. 9. Performance (AUC and mAP) of the models trained with expert annotation sets of 13 novices with soft labels (SL), aggregated annotation sets of 13 novices with hard labels (HL).

4. Discussion

Overall, our study provides new empirical results to better assess the impact of manual annotation variability on machine learning performance in the field of marine bioacoustics. A multi-annotator annotation campaign, including a variety of profiles from novices to experts, was set up with the task of annotating two blue whale call types in the Indian Ocean. Three hundred models were trained using different annotation sets and tested on two different datasets to assess the impact of inter-annotator discrepancies on model performance. This paper also presents a comprehensive comparison of two approaches (majority voting with hard labels and the creation of a soft labeled dataset) for aggregating annotations produced by novice annotators.

First, our study proposes to evaluate the inter-annotator variability in our campaign. We observe that there is notably less variability in the annotations of the novices for SEIO PBW vocalizations when compared to the annotations of blue whale's Dcalls. They exhibit a higher level of precision when compared to expert annotations used as the ground truth. This is indicative of the fact that SEIO PBW vocalizations tend to exhibit a more stereotypical and unique nature, compared to other numerical noise or biologic, anthropophonic and geophonic sounds in the area [Torterotot et al., 2019]. This specificity makes them less susceptible to confusion with other sound sources. In contrast, Dcalls are often mistaken for other types of short and impulsive sounds. It is noteworthy that two annotators appear to have misunderstood the task for the Dcall, as their recall and precision scores are significantly lower than the other annotators.

Two discernible annotation profiles in relation to the expert, became apparent as it has been reported in a previous study [Leroy et al., 2018a]. Firstly, the "conservative annotator" who limited their annotations to instances of high confidence. This approach yielded commendable precision while trading off some recall due to the potential oversight of less conspicuous calls. "Conservative" profiles are found above the $x = y$ line on Fig. 5. Secondly, the "permissive annotator" marked a broader range of shapes resembling the target call, leading to commendable recall but sacrificing precision due to an increased incidence of false positives. "Permissive" profiles are found under the $x = y$ line on Fig. 5. For the South Eastern Indian Ocean pygmy blue whale calls, all novice annotators exhibited a conservative annotation profile. This likely stems from the distinct stereotypical nature of these calls, making them less prone to confusion with alternative sound sources.

CNN models were trained using each annotation set to study the impact of inter-annotator variability on the performance of deep learning models. All trained models were evaluated on different datasets, in different areas with different recording devices and with multiple expert annotations.

Variations in performances are observed by comparing the model trained on the annotations from the expert and the 19 models trained on the annotations from the novices for both vocalization types. Novice annotators that produced annotations close to the expert got better performance in the ROC and Precision-Recall curves. Previous studies showed that the most reported consequence of label noise is a decrease in detection performance [Frenay and Verleysen, 2014; Shah et al., 2018]. Considering that the expert produced an annotation set closer to the ground truth than the annotation set from the novices, the latter can be considered noisier. Hence, models trained with the annotations from annotators for whom the results were farther from the expert get lower performance. This variability, due to the presence of non-correct annotations in the training set, put into question the use of CNNs applied directly to datasets where the relevance of annotations has not been evaluated.

However, results shows high performance (mean AUC and mAP above 0.75), and a good capacity for generalization of convolutional neural networks [Fonseca et al., 2019].

Fig. 6 shows that the gap between annotations from expert and novices produced more variation in the models' performances for blue whale's Dcalls than for SEIO PBW. Considering the difference in positive ratio between the two vocalizations (45.66 ± 5.82 and 18.27 ± 9.48 respectively. Table 1), label noise seems to be more detrimental under class imbalanced settings. This result has already been observed in image classification by Gu et al. (2023).

However, it has been observed that annotation sets used to train models qualified as "permissive annotators" get lower performance than the ones qualified as "conservative annotators" even with comparable F1 score with the annotations from the experts. Gu et al. (2023) also proposed that at the same ratio of noise on the training set, the model's performance can decrease differently regarding the type of noise. In this study, it seems that simple convolutional neural network models generalize better from the most obvious examples and the introduction of false positives deteriorates the performance more than the omission of a vocalization in the annotation set. For future annotation campaigns, this result might suggest that instructing annotators to adopt a more conservative approach could be beneficial.

To optimize the potential offered by multi-annotation and limit the potential error due to the addition of noisy labels in the training set, two grouping methods are used to assess the interest of crowdsourcing to improve the performance of the models.

In audio annotation tasks [Cartwright et al., 2019] and more specifically in PAM applied to cetaceans [Nguyen Hong Duc et al., 2021a; Dubus et al., 2023], it has been observed that an augmentation of the number of annotators increases the precision of the annotations and reduces the inter-annotator variability. Fig. 8 shows enhancements in AUC and mAP values and a decrease in standard deviation as the subgroup size increases. With the majority voting method, the systematic errors of some annotators are avoided. Similar results are observed in Wong et al. (2022), in a medical application of CNN for disease recognition, where the performance of models trained with the annotation of

Table 2
Manage a multi-annotator annotation campaign for manual annotations in PAM studies: a guideline.

Manage an annotation campaign for manual annotations in PAM studies: a guideline		
Before the campaign	Annotation platform	- Online, user-friendly and scalable - Provide uniform and non-editable spectrograms - Preserve annotator independence - Allows to listen sounds samples, with a playback control
	Upstream work on data	- Clear description of the task - Prior evaluation of the task difficulty - Propose a training session to to the annotators and present a catalogue of the targeted sounds
	Contact with the annotators	- Organize regular meetings - Answer possible questions
During the campaign	Annotations check:	- Favor a conservative annotation - Regularly evaluate the inter-annotator variability with κ metrics, or precision and recall computed for each pair of annotators. [Nguyen Hong Duc et al., 2021a; Dubus et al., 2023] - If the mean κ value is kept constant when adding new annotators and a large majority of the precision-recall values are in the upper right corner of the PR plot, the sufficient number of annotators is reached
	Contact with the annotators	- Summit a report to the annotators
After the campaign	Produce annotation set from the campaign	- Evaluate the inter-annotators variability to ensure the sufficient number of annotators is reached. - Pronounced discordance with the majority can be observed on the PR plot, they can be deleted - Use one of the 2 methods to aggregate all annotations proposed in this work: - majority grouping by sample (Section 2.2.3) - soft labeling (Section 2.3.2)

several annotators is higher than models trained with a single annotator. The performance trend with subgroup size appears to follow an asymptotic pattern, suggesting that adding more annotators for the aggregation will not increase the performance of the model. Similar results are observed in [Walter et al., 2022], where an increase in the number of annotators improves data quality. After a given number of raters (20 in their application of pointing on 2D images), the improvement got very small and is not worthwhile in relation to the costs.

The second aggregating method proposed produces results similar to the majority voting method. A similar result has been observed in an application with emotion recognition from images by [Fayek et al., 2016]. Gardiner et al. (2012) found that expert validation of data gathered through citizen science could be more cost-effective than traditional methods. Both methods proposed in this study, as they are applied with a sufficient number of annotators, allows us to use annotations produced from non expert annotators for the training of deep learnings models (classical CNN + Fully-connected layer model in this case) without expert validation, and nevertheless get detection performances close to that of a model that would be trained with expert annotations.

In general, even if the annotation of the two vocalizations presents different types of results, with different strategies of annotation (permissive and conservative), a high level of performance is reached after less than 10 annotators. Effectively overseeing annotation campaigns involving fewer than 10 annotators per sample seems to be feasible and has the potential to significantly boost the volume of annotated datasets available for training and testing automatic detection and classification models. Moreover, Kosmala et al. (2016) underlined the potential improvement of novice annotators in several citizen sciences programs, thus, the number of annotators per sample to ensure annotation quality could be expected to decrease as novice annotators will gain experience.

5. Managing an annotation campaign for manual annotations in PAM studies: A guideline

Based on the literature and on the present study, a guideline is proposed to manage annotation campaigns on PAM studies (Table 2).

This guideline starts with the key requirements for the design or selection of an annotation platform used to conduct a multi-annotator campaign. In the first part, propositions are made about the preparation of the campaign. The ergonomics of the annotation platform are important in the context of long-term campaigns: If the annotation platform is not user-friendly, annotators will become disengaged and cease annotating the files. As annotation campaigns generally aim to annotate large datasets, online access to the data is recommended. It is necessary for the procedure to be automatic, from the first observation of data by the annotators to the gathering of all annotations. Moreover, the platform must unify the data representation for all users, in order to reduce the variability. To reduce the over-representation of the first annotations, it is also important to keep all the annotations independent. We also strongly believe that the temporal context is an important factor for many underwater sound events: song context, periodic emission of vocalization [Madhusudhana et al., 2021]. Thus, the spectrogram should be proposed to the annotator in a chronological order, or be long enough to capture the long-term temporal context of targeted sound events. Furthermore, a large majority of the annotators used mainly the visual observation of the spectrogram rather than the sound. But the possibility of hearing the sounds was reported as extremely important as it increased the interest and motivation of annotators. Thus, it is important that the annotation platform gives the possibility to hear the sound. A playback control seems also important when the targeted sounds are outside the frequency range audible to humans.

In this kind of project, volunteer citizen scientists give free time to explore and annotate large datasets. Contact with the annotator is extremely important: (i) proposing a clear explanation about the aim of the annotation campaign, (ii) answering questions, (iii) organizing meetings throughout the campaign and share a report after. Those steps motivate the annotators and increase the educational potential of the campaign.

To ensure the relevance of the collected annotations, a clear description of the task is also extremely important: explaining clearly with examples the targeted sounds, if the time-frequency boxes have to be drawn on each vocalization, on only a part of the vocalization or around group of vocalzations. Then, a prior estimation of the task difficulty is of high interest for the management of an annotation campaign in terms of human resource planing. Indeed, such estimation will help in finding an initial guess of the number of annotators to start with, and even adapt this number to different time periods in case of a time-dependent estimation. Having said that, this estimation and its applications in setting an annotation campaign remains an open question for further investigation. Some previous works [Nguyen Hong Duc et al., 2021b; Dubus et al., 2023] already provided first evidence-based results confirming the intuition that certain acoustic features such as low SNR or the heterogeneity of some vocalizations could play an important role in complexifying the annotation process, but these results will have to be

scaled up to larger datasets and explanatory variable sets. The initial training of the annotators could be a source of bias, it has to be as representative as possible of the datasets proposed and in line with the difficulty of the task [Kosmala et al., 2016].

During the campaign, methods are proposed to assess inter-annotator variability and aggregate annotation sets. Those methods will help to define the minimal number of annotators needed [Dubus et al., 2023; Walter et al., 2022]. Increasing the number of annotators per sample could be helpful if the variability is too high. Finally, the two grouping methods: majority voting (Section 2.2.3) and soft labeling (Section 2.3.2) can also be used to produce relevant annotation sets for deep learning.

6. Conclusion

In this study, new approaches were explored to improve the accuracy and efficiency of cetacean vocalization detection using deep learning models. Leveraging multi-annotation campaigns involving both expert and novice annotators aimed to address the challenges associated with the scarcity of ground truth data in underwater PAM studies.

First, it was observed that annotator variability is influenced by factors such as the complexity of vocalizations and the annotation strategy (conservative or permissive). For South Eastern Indian Ocean pygmy blue whale calls, annotators exhibited a more conservative approach, resulting in high precision but slightly lower recall. Conversely, blue whale Dcalls, a non-stereotypical call with varied modulations, exhibited greater variability in annotations, with some annotators showing notably lower precision or recall. The results of the models trained on each annotator emphasize the importance of annotation guidelines that encourage a more conservative approach to improve annotation quality.

Furthermore, the study demonstrated the potential of crowdsourced annotations through a grouping method. Combining annotations from multiple novice annotators resulted in significant performance improvements, bringing detection models closer to the performance achieved with expert annotations. The results indicated that even with fewer than 10 annotators per sample, substantial enhancements in performance were attainable. This highlights the feasibility and effectiveness of crowdsourcing annotations to create larger and more diverse training datasets for cetacean vocalization detection models.

In the exploration of soft labeling, it was found that this approach provided a viable alternative to hard labeling when multiple annotators contributed to the annotations. While no significant improvements were observed over hard labeling, the soft labeling method consistently outperformed models trained on individual novice annotations and, in some cases, matched the performance of models trained with expert annotations.

In conclusion, this study underscores the potential of multi-annotation to advance the field of cetacean vocalization detection. By harnessing the collective efforts of novice annotators and optimizing annotation strategies, researchers can increase the quantity of annotated data and, thus, the capacity of generalization of deep learning models for detection. Ultimately, the findings provide valuable insights for future efforts in marine bioacoustics research and underline the importance of collaborative approaches in advancing our knowledge of underwater ecosystems and cetacean populations.

Data availability

Data will be made available on request.

Acknowledgments

The authors acknowledge all the annotators that gave their time in this manual annotation of our datasets. We particularly mention Julie Saidlitz from Astrolabe Expeditions[1] for driving with us the practical aspects of the annotation campaigns. The authors acknowledge the Ple de Calcul et de Données Marines5 (PCDM) for providing DATARMOR (storage, data access, computational resources, visualization, web-services, consultation, support services).

Appendix A. Appendix

A.1. Number of annotations per annotator and per label

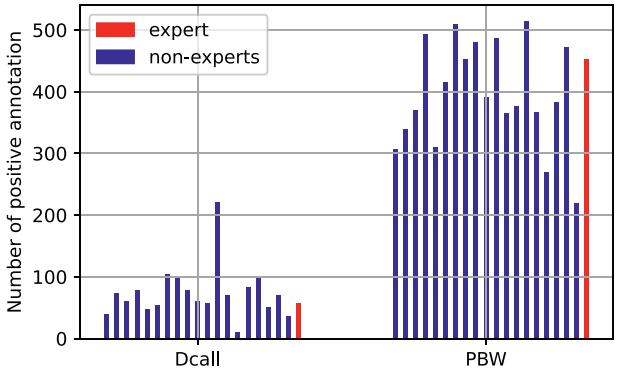


Fig. 10. Number of positive annotations per annotator and per label.

A.2. Schema of the aggregation methods by majority voting using three annotators

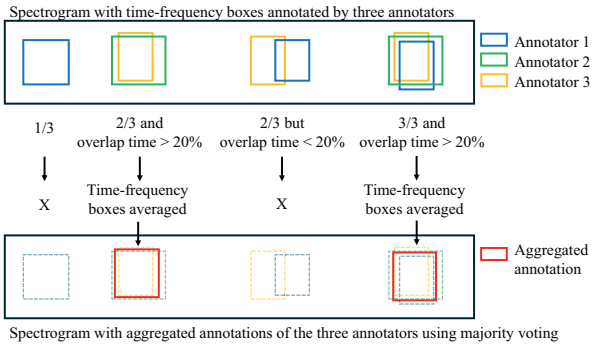


Fig. 11. Schema of the aggregation methods by majority voting using three annotators.

References

Best, P., 2022. Automated Detection and Classification of Cetacean Acoustic Signals. Universite de Toulon. URL: https://hal.science/tel-03826638.

Cartwright, M., Dove, G., Méndez Méndez, A.E., Bello, J.P., Nov, O., 2019. Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, Glasgow Scotland UK, pp. 1–11. https://doi.org/10.1145/3290605.3300522.

Courts, R., Erbe, C., Wellard, R., Boisseau, O., Jenner, K.C., Jenner, M.N., 2020. Australian long-finned pilot whales (*Globicephala melas*) emit stereotypical, variable, biphonic, multi-component, and sequenced vocalisations, similar to those recorded in the northern hemisphere. Sci. Rep. 10, 20609. URL: https://www.nature.com/articles/s41598-020-74111-y https://doi.org/10.1038/s41598-020-74111-y.

Du, R., Xie, S., Fang, Y., Hagino, S., Yamamoto, S., Moriyama, M., Yoshida, T., Igarashi-Yokoi, T., Takahashi, H., Nagaoka, N., Uramoto, K., Onishi, Y., Watanabe, T., Nakao, N., Takahashi, T., Kaneko, Y., Azuma, T., Hatake, R., Nomura, T., Sakura, T., Yana, M., Xiong, J., Chen, C., Ohno-Matsui, K., 2022. Validation of soft labels in developing deep learning algorithms for detecting lesions of myopic maculopathy from optical coherence tomographic images. Asia-Pacific J. Ophthalmol. 11, 227–236. https://doi.org/10.1097/APO.0000000000000466.

Dubus, G., Torterotot, M., Duc, P.N.H., Beesau, J., Cazau, D., Adam, O., 2023. Better quantifying inter-annotator variability: A step towards citizen science in underwater passive acoustics. In: OCEANS 2023 - Limerick, IEEE, Limerick, Ireland, pp. 1–8. URL: https://ieeexplore.ieee.org/document/10244502/ https://doi.org/10.1109/OCEANSLimerick52467.2023.10244502.

Fayek, H., Lech, M., Cavedon, L., 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. https://doi.org/10.1109/IJCNN.2016.7727250. pages: 570.

Fonseca, E., Plakal, M., Ellis, D.P.W., Font, F., Favory, X., Serra, X., 2019. Learning Sound Event Classifiers from Web Audio with Noisy Labels. URL: http://arxiv.org/abs/1901.01189 arXiv:1901.01189 [cs, eess, stat].

Frenay, B., Verleysen, M., 2014. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 25, 845–869. URL: http://ieeexplore.ieee.org/document/6685834/ https://doi.org/10.1109/TNNLS.2013.2292894.

Gardiner, M.M., Allee, L.L., Brown, P.M., Losey, J.E., Roy, H.E., Smyth, R.R., 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. Front. Ecol. Environ. 10, 471–476. https://doi.org/10.1890/110185.

Gavrilov, A.N., McCauley, R.D., Salgado-Kent, C., Tripovich, J., Burton, C., 2011. Vocal characteristics of pygmy blue whales and their change over time. J. Acoust. Soc. Am. 130, 3651–3660. URL: https://pubs.aip.org/jasa/article/130/6/3651/904284/Vocal-characteristics-of-pygmy-blue-whales-and https://doi.org/10.1121/1.3651817.

Gu, K., Masotto, X., Bachani, V., Lakshminarayanan, B., Nikodem, J., Yin, D., 2023. An instance-dependent simulation framework for learning with label noise. Mach. Learn. 112, 1871–1896. https://doi.org/10.1007/s10994-022-06207-7.

Hildebrand, J.A., Frasier, K.E., Helble, T.A., Roch, M.A., 2022. Performance metrics for marine mammal signal detection and classification. J. Acoust. Soc. Am. 151, 414–427. https://doi.org/10.1121/10.0009270.

Jindal, I., Nokleby, M., Chen, X., 2017. Learning Deep Networks from Noisy Labels with Dropout Regularization. URL: http://arxiv.org/abs/1705.03419 arXiv:1705.03419 [cs, stat].

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 27. https://doi.org/10.1186/s40537-019-0192-5.

Keribin, E., Morin, E., Vovard, R., 2024. APLOSE: A Scalable Web-Based Annotation Tool for Marine Bioacoustics - Public Repository. https://doi.org/10.5281/zenodo.10468000.

Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Front. Ecol. Environ. 14, 551–560. https://doi.org/10.1002/fee.1436.

Krause, B., 1987. Bioacoustics: Habitat Ambience and Ecological Balance. Whole Earth Review.

Leroy, E.C., Royer, J., Bonnel, J., Samaran, F., 2018a. Long-term and seasonal changes of large whale call frequency in the southern Indian Ocean. J. Geophys. Res. Oceans 123, 8568–8580. https://doi.org/10.1029/2018JC014352.

Leroy, E.C., Thomisch, K., Royer, J.Y., Boebel, O., Van Opzeeland, I., 2018b. On the reliability of acoustic annotations and automatic detections of Antarctic blue whale calls under different acoustic conditions. J. Acoust. Soc. Am. 144, 740–754. https://doi.org/10.1121/1.5049803.

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., Širović, A., Roch, M.A., 2021. Improve automatic detection of animal call sequences with temporal context. J. R. Soc. Interface 18, 20210297. https://doi.org/10.1098/rsif.2021.0297.

McClure, E.C., Sievers, M., Brown, C.J., Buelow, C.A., Ditria, E.M., Hayes, M.A., Pearson, R.M., Tulloch, V.J., Unsworth, R.K., Connolly, R.M., 2020. Artificial intelligence meets citizen science to supercharge ecological monitoring. Patterns 1, 100109. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666389920301434 https://doi.org/10.1016/j.patter.2020.100109.

McDonald, M.A., Mesnick, S.L., Hildebrand, J.A., 2023. Biogeographic characterisation of blue whale song worldwide: using song to identify populations. J. Cetacean Res. Manag. 8, 55–65. URL. https://journal.iwc.int/index.php/jcrm/article/view/702. 10.47536/jcrm.v8i1.702.

Miller, B.S., The IWC-SORP/SOOS Acoustic Trends Working Group, Miller, B.S., Stafford, K.M., Van Opzeeland, I., Harris, D., Samaran, F., Širović, A., Buchan, S., Findlay, K., Balcazar, N., Nieukirk, S., Leroy, E.C., Aulich, M., Shabangu, F.W., Dziak, R.P., Lee, W.S., Hong, J.K., 2021. An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors. Sci. Rep. 11, 806. URL. https://www.nature.com/articles/s41598-020-78995-8. https://doi.org/10.1038/s41598-020-78995-8.

Miller, B.S., Madhusudhana, S., Aulich, M.G., Kelly, N., 2022. Deep learning algorithm outperforms experienced human observer at detection of blue whale D-calls: a double-observer analysis. Remote Sens. Ecol. Conserv. https://doi.org/10.1002/rse2.297.

Nguyen Hong Duc, P., Torterotot, M., Cazau, D., Vovard, R., Keribin, E., 2020. Aplose: A Scalable Web-Based Annotation Tool for Marine Bioacoustics. Osmose Product Presentation. ENSTA Bretagne.

Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P.R., Gérard, O., Adam, O., Cazau, D., 2021a. Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. Eco. Inform. 61, 101185. URL. https://linkinghub.elsevier.com/retrieve/pii/S1574954120301357. https://doi.org/10.1016/j.ecoinf.2020.101185.

Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P.R., Gérard, O., Adam, O., Cazau, D., 2021b. Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. Eco. Inform. 61, 101185. URL. https://linkinghub.elsevier.com/retrieve/pii/S1574954120301357. https://doi.org/10.1016/j.ecoinf.2020.101185.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F.D., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc., p. 1

Peterson, J.C., Battleday, R.M., Griffiths, T.L., Russakovsky, O.. Human uncertainty makes classification more robust. https://doi.org/10.1109/ICCV.2019.00971.

Royer, J.Y., 2009. OHA-SIS-BIO - Observatoire Hydroacoustique. https://doi.org/10.18142/229.

Shah, A., Kumar, A., Hauptmann, A.G., Raj, B., 2018. A Closer Look at Weak Label Learning for Audio Events. URL. http://arxiv.org/abs/1804.09288. arXiv:1804.09288 [cs, eess].

Shamir, L., Yerby, C., Simpson, R., von Benda-Beckmann, A.M., Tyack, P., Samarra, F., Miller, P., Wallin, J., 2014. Classification of large acoustic datasets using machine learning and crowdsourcing: application to whale calls. J. Acoust. Soc. Am. 135, 953–962. https://doi.org/10.1121/1.4861348.

Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.M., Helble, T., Cholewiak, D., Gillespie, D., Klinck, H., 2020. Deep neural networks for automated detection of marine mammal species. Sci. Rep. 10, 607. URL: https://www.nature.com/articles/s41598-020-57549-y https://doi.org/10.1038/s41598-020-57549-y.

Solsona-Berga, A., Frasier, K.E., Baumann-Pickering, S., Wiggins, S.M., Hildebrand, J.A., 2020. DetEdit: a graphical user interface for annotating and editing events detected in long-term acoustic monitoring data. PLoS Comput. Biol. 16, e1007598 https://doi.org/10.1371/journal.pcbi.1007598.

Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G., 2022. Learning from Noisy Labels with Deep Neural Networks: A Survey. URL: http://arxiv.org/abs/2007.08199 arXiv:2007.08199 [cs, stat].

Torterotot, M., 2020. Traitement et analyse de données bioacoustiques dans l'océan Indien austral: application aux baleines bleues. Université de Bretagne Occidentale - Brest.

Torterotot, M., Royer, J.Y., Samaran, F., 2019. Detection strategy for long-term acoustic monitoring of blue whale stereotyped and non-stereotyped calls in the southern Indian Ocean. In: OCEANS 2019 - Marseille, IEEE, Marseille, France, pp. 1–10. URL: https://ieeexplore.ieee.org/document/8867271/ https://doi.org/10.1109/OCEANSE.2019.8867271.

Torterotot, M., Samaran, F., Stafford, K.M., Royer, J.Y., 2020. Distribution of blue whale populations in the southern Indian Ocean based on a decade of acoustic monitoring. Deep-Sea Res. II Top. Stud. Oceanogr. 179, 104874. URL: https://linkinghub.elsevier.com/retrieve/pii/S0967064520301247 https://doi.org/10.1016/j.dsr2.2020.104874.

Torterotot, M., Béesau, J., Perrier de la Bathie, C., Samaran, F., 2022. Assessing marine mammal diversity in remote Indian Ocean regions, using an acoustic glider. Deep-Sea Res. II Top. Stud. Oceanogr. 206, 105204. URL: https://linkinghub.elsevier.com/retrieve/pii/S0967064522001904 https://doi.org/10.1016/j.dsr2.2022.105204.

Usman, A.M., Ogundile, O.O., Versfeld, D.J.J., 2020. Review of automatic detection and classification techniques for cetacean vocalization. IEEE. Access 8, 105181–105206. URL: https://ieeexplore.ieee.org/document/9110497/ https://doi.org/10.1109/ACCESS.2020.3000477.

Walter, V., Kölle, M., Collmar, D., 2022. Measuring the wisdom of the crowd: how many is enough? PFG J. Photogram. Remote Sens. Geoinform. Sci. 90, 269–291. https://doi.org/10.1007/s41064-022-00202-2.

Wong, D.R., Tang, Z., Mew, N.C., Das, S., Athey, J., McAleese, K.E., Kofler, J.K., Flanagan, M.E., Borys, E., White, C.L., Butte, A.J., Dugger, B.N., Keiser, M.J., 2022. Deep learning from multiple experts improves identification of amyloid neuropathologies. Acta Neuropathol. Commun. 10, 66. URL: https://actaneurocomms.biomedcentral.com/articles/10.1186/s40478-022-01365-0 https://doi.org/10.1186/s40478-022-01365-0.

Yurk, H., Barrett-Lennard, L., Ford, J., Matkin, C., 2002. Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern Alaska. Anim. Behav. 63, 1103–1119. URL: https://linkinghub.elsevier.com/retrieve/pii/S0003347202930125 https://doi.org/10.1006/anbe.2002.3012.