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

Paul Nguyen Hong Duc[a)]

a) Sorbonne Universit, CNRS, Institut Jean Le Rond dAlembert, UMR 7190, Paris, France

Maëlle Torterotot[b)]

[b)] Universit de Brest, CNRS, Laboratoire Gosciences Ocan, Brest, France;
maelle.torterotot@univ-brest.fr

Flore Samaran[c)]

[c)] ENSTA Bretagne, CNRS, Lab-STICC, UMR 6285, Brest, France;
flore.samaran@ensta-bretagne.fr

Paul R. White[d)]

[d)] University of Southampton, ISVR, Southampton, United Kingdom; prw@isvr.soton.ac.uk

Odile Gerard[e)]

[e)] DGA-TN, Toulon, France; odile.gerard@intradef.gouv.fr

Olivier Adam[a),f)]

[a)] Sorbonne Universit, CNRS, Institut Jean Le Rond dAlembert, UMR 7190, Paris, France

[f)] Universit Paris-Saclay, CNRS, Institut des Neurosciences Paris-Saclay, Gif-sur-Yvette, France; olivier.adam@sorbonne-universite.fr

Dorian Cazau[c)]

[c)] ENSTA Bretagne, CNRS, Lab-STICC, UMR 6285, Brest, France;
dorian.cazau@ensta-bretagne.fr

Abstract

It is currently widely recognized that automated methods are crucial to help processing long-term recordings of marine bioacoustics. To evaluate the efficiency of such methods, it is essential to develop large-scale annotated datasets. However, besides being laborious and resource intensive, recent studies have

[*] Corresponding author
Email address: p.nguyenhongduc@gmail.com (Paul Nguyen Hong Duc[a)])
Underlined authors contributed equally to this work

Preprint submitted to Ecological Informatics *November 3, 2020*

suggested that such a task could also be highly subjective with the generation of annotator specific errors.

In this work, we investigate the question of inter-annotator agreement from a multi-annotator annotation campaign performed on a marine bioacoustics dataset. After providing quantitative evidence of inter-annotator variability, we investigate potential sources on both the user annotation practice and the annotation data and task to better understand why and how such variability occurs. Our study reveals that the acoustic event type, the Signal-to-Noise Ratio of the acoustic event and the annotator profile are three examples of critical factors impacting the annotation results of a multi-annotator campaign.

Keywords: DCLDE 2015 low frequency, multi-annotator agreement, marine bioacoustics, reliability

1. Introduction

Passive acoustic monitoring (PAM) is an effective and harmless way to evaluate biodiversity across large spatio-temporal scales.

In PAM, detection and classification methods are necessary steps to perform accurate acoustic surveys and improve knowledge of both marine and terrestrial ecosystems. To evaluate the performance of such methods, they have to be developed and tested on reference annotated datasets. Moreover, in the 21st century, supervised artificial intelligence (AI) techniques have become an essential part of detection and classification methods. Most of them rely on the amount and quality of annotated training data. The process of collecting annotations is thus the main bottleneck in building such methods.

The traditional approach to collect annotations in PAM most often involve bioacousticians (with different levels of expertise) who manually annotate the data. Such an approach is currently thought to be the most accurate one (*e.g.* in comparison to automatic labeling), and always serves as a reference (often referred to as ground truth) for further analysis. However, annotation in marine bioacoustics, besides being resource intensive, laborious and time consuming, is compounded by the intrinsic difficulty in discriminating underwater acoustic sources. Indeed, even experts recognize some inextricable ambiguities. For example, Baumgartner et al. (2019) observed that humpback whales sometimes produce an upsweep call that is hard to distinguish from a right whale upcall unless using a pitch track with a temporal context. Moreover, underwater soundscapes are composed of multiple sounds that can sometimes occur simultaneously, completely or partially masking each other (Clark et al., 2009; Erbe et al., 2016), which makes the identification of individual acoustic sources even more difficult. Overlaying killer whale harmonics and stationary boat noises at certain frequencies can also look very similar (Bergler et al., 2019). However, in spite of these constraints and uncertainties, some freely available annotated datasets do exist, such as the dataset from DCLDE workshop (Detection, Classification, Localization and Density Estimation of marine mammals using passive acoustics) workshop, which allow participants to directly compare algorithms and methodologies datasets[1]. But compared with the field of image recognition, daily visual objects (cats, dogs, chairs, etc.) are more familiar to human perception than marine mammal acoustic repertories. Consequently, experts in this field can generally label such data with a higher level of confidence and shorter annotation task completion time than in marine bioacoustics, and even commonly resort to crowdsourcing annotations to quickly size up their machine learning datasets.

Whatever the scientific fields, one of the best practices to set up an annotation campaign is to involve several independent annotators, who share a certain amount of data to be annotated so their results can be compared in an objective way. In such a context, inter-annotator agreement (Gwet, 2014) (also referred to as inter-rater reliability in the literature) is the extent to which human decisions coincide, or in other words it allows to measure the amount of consensus between a group of annotators. A high agreement means that the raters can be used interchangeably. If interchangeability is guaranteed, then the ratings of one annotator can be used with confidence, without asking which annotator provided the annotation (Gwet, 2014). Although most PAM studies mention the use of several annotators in their annotation protocol, they do not focus on this question of inter-annotator variability (*e.g.* Kirsebom et al. (2020) recognized the need for "more systematic and controlled investigation of the inter-annotator variability"). To the best of our knowledge, only Leroy et al. (2018) addressed this problem by measuring the inter- and intra-annotator variability in manually annotated Antarctic blue whale stereotyped calls. They revealed both a strong inter-annotator variability between two annotators (with less than 50% agreement between annotators), but also a poor agreement obtained with an annotator annotating the same audio segment twice. Otherwise, a few studies have evaluated the influence of annotation results on automated detection performance. Sirovic (2016) evaluated the impact of annotation variability on the performance of a spectrogram correlation detector. They found that annotator variability did not affect long-term trends in detection, but that it had an impact on the total number of detections and therefore on the call rate estimation. Torterotot et al. (2019) also reported the variability in recall and precision of a blue whale call automated detector determined by comparing the detector outputs with three different ground truths labeled by three different annotators. Overall, these studies highlighted the need for a more standardized approach for manual annotation and automatic detection evaluation to globally improve the comparability of PAM studies. This question has also often been addressed in the different DCLDE workshop editions (*e.g.* the DCLDE 2013 discussion panel emphasized the need for more exhaustive and reliable annotation campaigns based on consistent annotation protocols[2].). Our work

[1] `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`.

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

has also been highly motivated by related current investigations done in urban soundscapes. Recent studies Cartwright et al. (2017, 2019) have quantified the reliability/redundancy trade-off in crowdsourced airborne soundscape annotation, as well as examined the effect of various annotation campaign parameters (e.g. the number of classes to annotate, acoustic characteristics of the classes, the method for sound visualization, the annotator profiles, the soundscape complexity) on inter-annotator agreement. One interesting outcome of their work has been to estimate a minimal number of annotators to get a reliable annotation as a function of campaign parameters, ending up with a series of concrete suggestions to set up a campaign. Transposing such findings to marine bioacoustics would be highly valuable.

In this context, our work intends to pursue current efforts (Sirovic, 2016; Leroy et al., 2018; Torterotot et al., 2019) in better understanding inter-annotator agreement within collaborative annotation campaigns in marine bioacoustics. A new annotation campaign was performed on the DCLDE 2015 low frequency dataset, involving 6 annotators with different profiles, in addition to the two experts who originally annotated this dataset for the DCLDE 2015 challenge. Besides providing experimental evidence of inter-annotator variability, we also investigated potential sources explaining this variability, each source referring to either the dataset content (*e.g.* call type, Signal-to-Noise Ratio (SNR)) or to the annotator profile and behavior (*e.g.* average duration spent on annotating and annotator experience).

2. Material and methods

2.1. Annotation campaign

2.1.1. Dataset

The dataset used in this study is a subset of the DCLDE 2015 low frequency dataset. This dataset has been recorded with High-frequency Acoustic Recording Packages (HARP) deployed off the southern and central coast of California at different locations, spanning all four seasons, over 2009-2013. The data were decimated to a 2 kHz sampling frequency to provide a dataset for blue whale (*Balaenoptera musculus*) D-calls (Thompson, 1965) and fin whale (*Balaenoptera physalus*) 40-Hz calls (Watkins, 1981) identification (Fig. 1). On the occasion of the 2015 DCLDE challenge, a first annotation campaign with two annotators was performed, which consisted in annotating time intervals of D-calls and 40 Hz calls. The resulting annotation file made available[3] consists in the fusion by majority voting of these two individual annotations. Here, we only use a subset of acoustic data composed of 50 consecutive hours of the CINMS18B file, recorded within the Channel Islands National Marine Sanctuary and starting on the 23th June 2012. Table 1 sums up the dataset parameters.

[3] See `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`.

Table 1: Dataset parameters

DCLDE2015	LF
Sampling rate (kHz)	2
Annotated species	*Balaenoptera musculus* - blue whale D-calls Thompson (1965) *Balaenoptera physalus* - fin whale 40-Hz calls Watkins (1981)
Dataset size / Nb files	700 Mo / 563 files
Site [Year / Month]	CINMS_18_B_d06_120622_055731.d100.x.wav (2012 / 06)
File durations	first 50h split in 563 x 320s long audio files
Dates	Start: 2012-06-23 05:57:31 / End: 2012-06-25 07:56:51
Class count	'D-call': 719, '40-Hz': 156 from the DCLDE challenge annotations

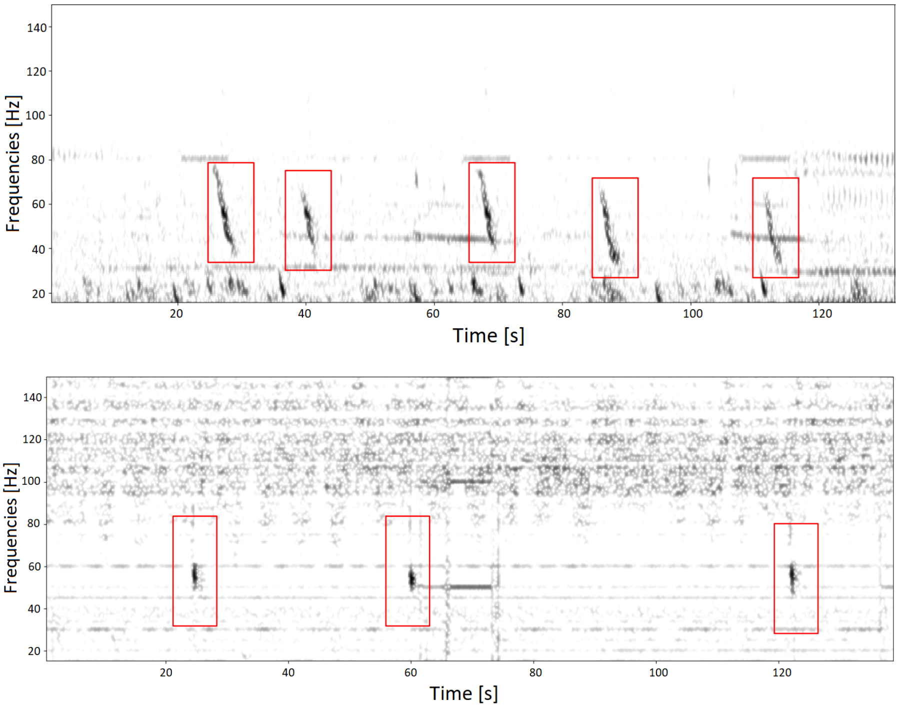

Figure 1: Examples of the two annotated call types: five blue whale D-calls (upper spectrogram in red boxes) and 3 fin whale 40-Hz calls (lower spectrogram in red boxes).

2.1.2. Annotation support and protocol

Annotation was performed through audio-visual inspection of spectrograms (*i.e.* FFT-based time-frequency representation of sounds). Table 2 sums up a set of parameters used to compute and display annotation spectrograms. They were chosen to fit at best the original annotation protocol except for a few differences. Indeed, our spectrogram contrast was empirically fixed based on the median of maximum values from filtered Power Spectral Density (PSD) in the [15-150] Hz frequency band. Furthermore, annotators could use a zoom up to 8x on the time axis (*i.e.* 40 second window) from the default duration of the spectrogram window set to 320 seconds (about 5 minutes). Listening to the recordings was allowed but not mandatory, with varying playing speeds from 0.25x to 4x.

Annotators were given instructions (see Supplementary Materials) with visual and aural examples of the sounds to annotate. They were gathered in a guide document, made available at all times during the campaign. They could refer to it at any time during the annotation process. To annotate a specific sound, they had to draw a time and frequency box around it as close as possible to the sound. Naturally, the annotators had only access to their own annotations to avoid any influence from other annotations.

Annotators could choose among 3 labels, tagged as follow: D-call (for blue whale D-calls), 40-Hz call (for fin whale 40-Hz calls) and Unknown call. Annotators were instructed to use this latter label in case of doubt between the two call types, but not to annotate a call type from an unknown source.

In the analysis, when not specified, 4 classes were used to compute the statistical metrics: “D-call”, “40-Hz”, “Unknown call” and “None”. This latter was used when an annotator identified a sound but no one else labeled it. For example, annotator A gives the D-call label to a sound, but the others do not label it. Annotator A label will thus be “D-call” and the other’s labels will be “None”.

Eventually, an annotated event was defined as an overlapping event by the following condition: if its midpoint fell within the time bounds of another annotation box Leroy et al. (2018). In the case where an annotator tagged two overlapping events, chronological order is kept in order to find corresponding labels for other annotators.

Table 2: Parameter description of the different APLOSE seed datasets.

Sample frequency (kHz)	2
Max → min display duration (s) / Zoom level number	320 → 40 / 4
nfft (samples)	4096
winsize (samples)	2000
overlap (percent)	90
Gain (dB)	35
Filtering frequency band (Hz)	[15-150]

2.1.3. Annotation software

For this study we developed our own open source annotation software named APLOSE. As it is a web-based annotation interface, APLOSE highly facilitates the setting up of collaborative campaigns because no transfer data to each participant is made as the data are on a server and it does not request that they install the same annotation software. Besides that, APLOSE has been deployed on the long term on dedicated web servers, and it now offers the capacity to anybody of easily updating its annotation campaigns, including the one of this current study.

2.1.4. Annotator profiles

A total of 6 annotators were enrolled in the campaign. We assume that these raters were capable of identifying the two calls. They can be gathered into 3 groups (cf Table 3): low-frequency whale sounds experts (annotators who have already annotated several hours of low-frequency mysticete calls, especially D-calls), bioacousticians (annotators who have already annotated several hours of cetacean sounds but not D-calls) and neophytes (annotators with no experience in underwater sound annotations). The annotators were volunteers and not compensated financially for their work. No quality check of the individual annotations was performed for this campaign. In addition to these 6 annotators, the annotations from the two DCLDE experts were used. However, only one set of annotations is available, as they only kept the common annotations (DCLDE_exp).

Table 3: Annotator profiles

Annotator	Expertise level
DCLDE_exp	Expert
A1	Bioacoustican
A2	Bioacoustican
A3	Expert
A4	Neophyte
A5	Bioacoustican
A6	Expert

2.2. Evaluation of inter-annotator agreement

Considering the difficulty of obtaining a “ground truth for underwater soundscape events, evaluation metrics rather qualify the relative level of agreement between annotators than the absolute annotator performance. The inter-annotator assessment is performed using only pseudo-presence observations resulting from the annotation process. In other words, only events annotated at least by one annotator are taken into account in the agreement evaluation. The agreement on pseudo-absence is not evaluated here.

To evaluate the inter-annotator agreement with our nominal data, the Fleiss κ score was computed (Fleiss, 1975; Zapf et al., 2016) as we considered all annotators as equally important. This metric corresponds to the proportion of agreement corrected for chance, scaled from -1 to +1, with a negative value indicating poorer than chance agreement, zero indicating exactly chance and positive values indicating better than chance agreement. Chance agreement happens when multiple annotators assign a similar label that is not directly dictated by the data. It occurs when an annotator does not know which label to give, and chooses one randomly. Mathematically, the Fleiss κ score is defined as (Fleiss, 1975) :

$$p_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} R_{ij} \left(R_{ij} - 1\right), \quad p_j = \frac{1}{Nn} \sum_{i=1}^{N} R_{ij} \tag{1}$$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} p_i, \quad \bar{P}_e = \sum_{j=1}^{K} p_j^2 \tag{2}$$

$$\kappa_{Fleiss} = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{3}$$

where n is the number of annotations per sample, p_j, the proportion of all assignments which were to the j-th category, p_i, the extent to which annotators agree for the i-th subject, $\bar{P}$ the mean of the p_i. If the annotators are in complete agreement then $\kappa = 1$. If there is no agreement among the annotators (other than what would be expected by chance) then $\kappa \leq 0$.

In order to compute the Fleiss κ score for a specific call type, all annotations for the call type to analyze are retrieved. Then if an annotator did not annotate with the same label, or did not annotate the audio segment at all, the label “None” is given for that annotator. As a consequence, for example, when computing the Fleiss κ score for D-calls, only two labels are possible: D-call or None.

2.3. Potential sources of inter-annotator agreement variability

To explain the agreement variability, we investigated the impact of several campaign factors on our inter-annotator agreement metrics.

2.3.1. Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio (SNR) was computed as in Torterotot et al. (2019). Audio files were passband filtered between 15Hz and 150Hz using a third-order Butterworth filter. For each 5-minute audio file, the noise power level was computed in the same frequency band, using the estimator presented in Socheleau et al. (2015). Each event’s power level was then compared to the matching noise power level of the file where it was identified. All values were then gathered in 4 categories of SNR: <= 0, (0, 5], (5, 10], >10.

2.3.2. Annotation durations

All along the annotation campaign, the time that each annotator spent to annotate each file was stored in the campaign log files. All retrieved durations for a task were rounded and then divided in two categories of files: the ones that contained at least one identified event and the others without any identified event. When durations were higher than 20 minutes, we removed these values, assuming that they correspond to a misuse of the annotation task (typically an annotator going on a break while leaving apart an ongoing annotation task).

2.3.3. Annotation clustering

In order to better characterize annotator behaviors (Kairam and Heer, 2016), the annotators were gathered into clusters using the Hamming distance (Hamming, 1950) as a measure of the distance between their annotations. Hierarchical agglomerative clustering (HAC) method was used with the single linkage method. Clusters were then formed from the computed hierarchical clustering based on the cophenetic distance between annotators. The threshold was set to 0.09 meaning that annotators within the same cluster have less than 127 different annotations. The threshold was set according to the minimum number of 40-Hz annotations (131). The multidimensional scaling was used to represent the annotators by preserving their Hamming distance. All computations were performed using Scikit-learn (Pedregosa et al., 2011).

3. Results

3.1. Quantitative evaluation of the inter-annotator agreement

3.1.1. Number of annotations

Fig. 2 shows the number of annotations per call type for each annotator. All of the annotators labeled more D-calls than fin whale 40-Hz calls. DCLDE_exp and A3 annotated about 700 D-calls, A2, A4 and A5 between 800 and 1000 and A1 and A6 more than 1000. For the 40-Hz call type, A4 identified more events (almost 400) than the other annotators (about 200). The label "Unknown call" was either used almost 200 times (by A2 and A6) or less than 50 times (A1, A3, A4, A5).

3.1.2. Agreement metric

Fleiss κ score reflects the reliability and agreement between annotators. For D-calls, the Fleiss κ score is around 0.4 showing a moderate agreement between annotators. For fin whale 40-Hz calls, this value is close to 0, meaning that annotators almost systematically disagreed on this label (see Fig. 3 left).

The average inter-annotator agreement slightly increases when the number of annotators increases until reaching a plateau. (see Fig. 3 right). Standard deviations of Fleiss κ are higher when comparing the agreement between groups of two people and decrease when the group size expands.

In the next section, we explore two categories of potential sources that can influence the inter-annotator agreement. The first one refers to the dataset

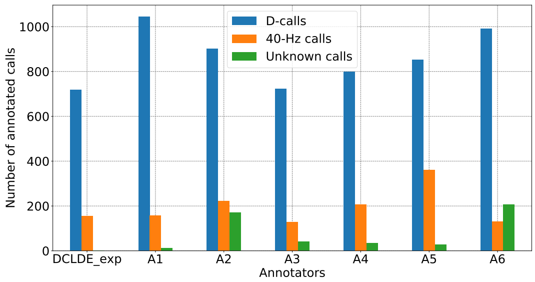

Figure 2: Number of annotated calls per label for each annotator

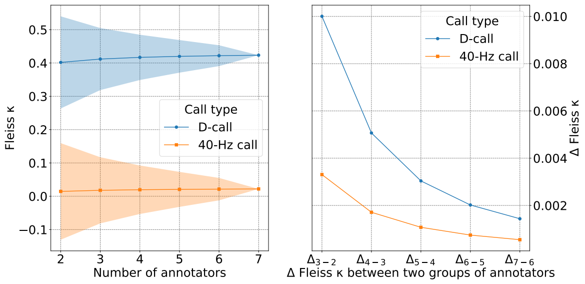

Figure 3: (left) Fleiss κ measure for all call types; the standard deviations are computed by comparing different groups of annotators. (right) Difference of averaged Fleiss measure between each value computed for a different number of annotators.

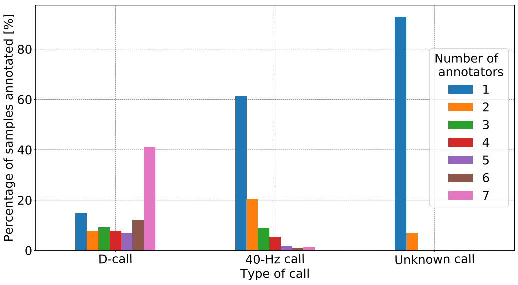

Figure 4: Number of calls annotated by at least N annotators.

itself (*e.g.* call type, SNR) and the second refers to the annotator profile and behavior.

3.2. Potential sources of inter-annotator agreement variability: from the dataset perspective

3.2.1. Call signatures and class labels

We first investigated how the call type impacts the inter-annotator variability. Fig. 4 shows that about 40% of the 1240 annotated D-calls were annotated by all annotators, while less than 20% were annotated by only one annotator. Most of the other D-calls were identified by at least two annotators (> 80%). Over 782 annotated fin whale 40-Hz calls, only 1% were unanimously annotated and more than 60% were annotated by only one annotator. The "Unknown" label was used sparingly, but the annotators never used it to identify the same event. 335 events were identified by at least one annotator as a D-call and by at least another one annotator as a fin whale 40-Hz call. Among those, 61 were identified by 6 annotators as a D-call and by only one as a fin whale 40-Hz call.

We propose a qualitative inspection of spectrograms that were marginally labeled. Fig. 5 represents some spectrograms annotated by all annotators (on the left) and by only one annotator (on the right). We chose to represent the spectrograms as they were visualized on the annotation interface. It is clear that events with salient acoustic features reached a more systematic consensus than those with less energy occurring in noisier time periods.

100 of the "Unknown call" labels were annotated by two annotators or less (Fig. 4). Some examples of labeled "Unknown calls" are represented in Fig. 6. These events exhibit a shape close to both D-calls or fin whale 40-Hz calls, and often occur in a noisy environment.

Finally, two examples of events that were annotated both as D-calls and fin whale 40-Hz calls by different annotators are represented in Fig. 7. To attribute

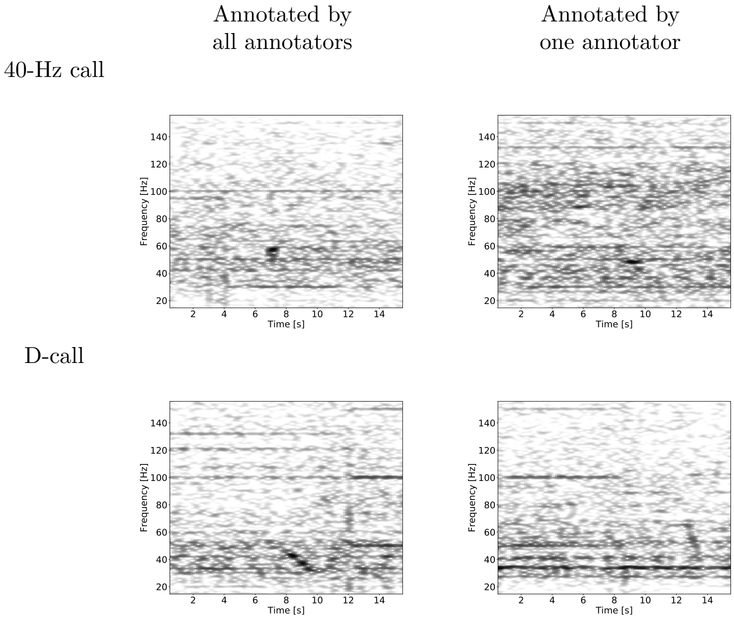

Figure 5: Examples of labeled calls annotated by all annotators (left panel) and calls annotated by only one annotator (right panel) for 40 Hz fin whale calls (top row) and D-calls (bottom row). They are represented in this figure as they appear in the APLOSE interface.

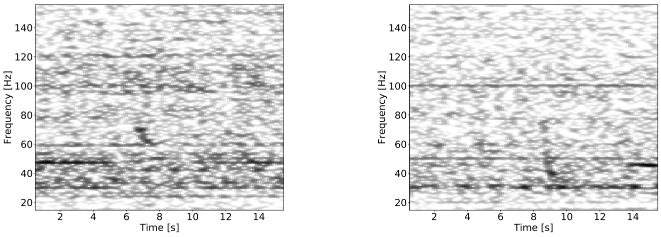

Figure 6: Spectrograms of two "Unknown calls" identified by at least one annotator. They are represented in this figure as they appear in the APLOSE interface.

a label, they seem to focus either on the downsweep shape (similar to a D-call) or on the part of the event with higher and shorter energy (similar to a fin whale 40-Hz call).

3.2.2. SNR

The SNR distribution is similar for the two call types, with a predominance of low SNR calls (Fig. 8). More than 50% of the annotated D-calls and 48% of the fin whale 40-Hz calls have an SNR between between 0 and 5 dB whereas more than 10% of D-calls and 40-Hz calls have a SNR > 10 dB.

Fig. 9 displays the number of events annotated by only one annotator and by all annotators as a function of SNR. The proportion of D-calls annotated by all annotators increases as the SNR increases. The proportion of D-calls annotated by one annotator is higher for negative SNR (up to 40%). For positive SNR, the proportion of calls annotated by only one annotator remains steady at around 20%. The proportion of fin whale 40-Hz calls annotated by all annotators is very scarce for all SNR bins.

Fig. 10 represent the Fleiss κ scores regarding the SNR distribution for D-calls. For D-calls, the Fleiss κ ranges from 0.36 for low SNR to 0.52 for high SNR. This means that agreement values are higher for high SNR D-calls. Moreover, the standard error of the agreement decreases with higher SNR calls.

For fin whale 40-Hz calls, the Fleiss κ ranges from about -0.05 for high SNR to 0.09 for low SNR. Consequently, the SNR does not seem to have an influence on the Fleiss κ for this call type.

Reliability did not appear to be to be significantly affected by the number of annotators rating each audio segment (Δ values < 10-2), and these findings are consistent across independent samples of annotators. However, despite the small absolute range of Δ values, we can still describe the extent to which our different inter-annotator agreement measure converges on a plateau-like region above a certain number of annotators.

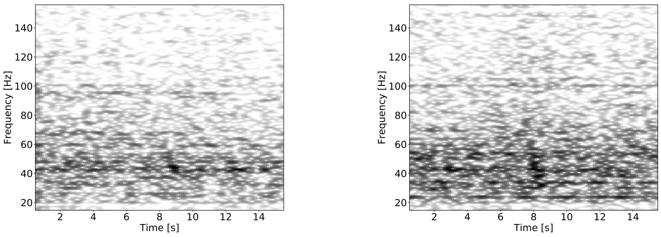

Figure 7: Examples of (left) events labeled as a fin whale 40-Hz call by 2 annotators and D-call by 2 annotators and (right) event labeled as a fin whale 40-Hz call by 2 annotators, a D-call by 3 annotators and as Unknown call by 1 annotator. They are represented in this figure as they appear in the APLOSE interface.

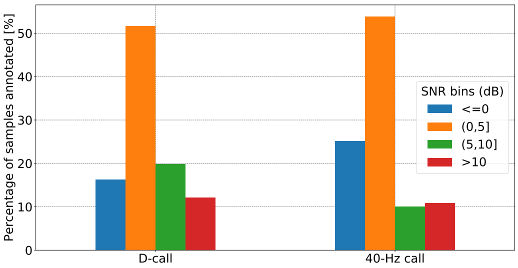

Figure 8: SNR distribution of (left) labeled D-calls and (right) labeled fin whale 40-Hz calls (right).

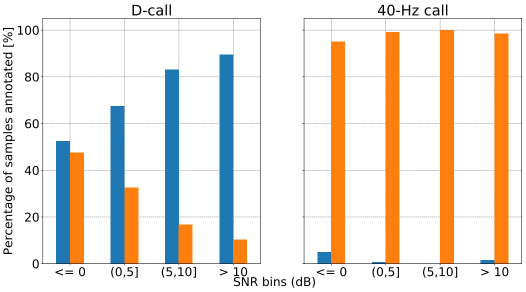

Figure 9: Number of samples of D-call (left) and 40-Hz (right) events per SNR category annotated by all annotators (blue) or by only one (orange).

In our experiments, a fast convergence towards this plateau was observed for D-calls and high SNR values, in comparison to lower SNR values with the same call type or when looking at values for the fin whale 40-Hz call type. These lower convergences reveal more difficult annotation tasks where the inter-annotator agreement will need an higher number of annotators to get stable.

3.3. Potential sources of inter-annotator agreement variability: regarding the annotator profile and behavior

3.3.1. Annotation duration

Fig. 11 shows the average duration spent by each annotator to label each 5-minute file. It compares average duration for files containing at least one annotation (in blue) versus files without any annotation (in orange) for each annotator. Also, because completion time was not available for the DCLDE experts we discarded them from this analysis. Overall, annotator A1 took longer to annotate the files than the other annotators, with respective median values of 100s and 40s. Unlike the other annotators, annotator A2 presents a few upper outliers with durations smaller than 150s. In general, the annotators took less time to annotate files they believed, mainly based on visual inspection, to contain no calls. For files with identified events, the median duration was about 30s higher than for files with no event. Also, minimal values for event and noise file are close ranging from 2s to 16s. However, annotator A1 still spends more time than the other on these files.

3.3.2. Annotator profile

Fig. 12 represents the results of our cluster analysis, in which we did not include the “Unknown call” label as it was not used by the DCLDE annotators. No clusters can be observed even if DCLDE_exp, annotator A3 and annotator A4

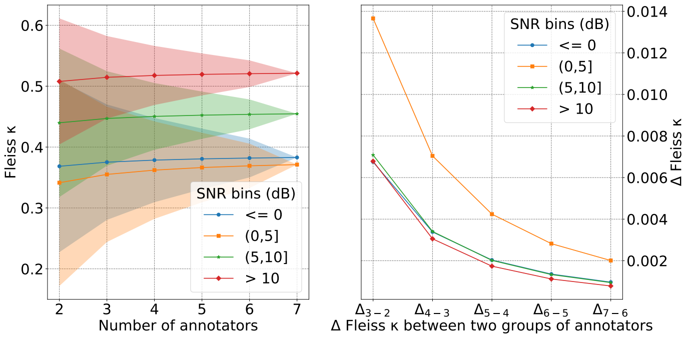

Figure 10: (left) Fleiss κ measure per SNR category for annotators for D-call and (right) the Δ measure between the number of annotators.

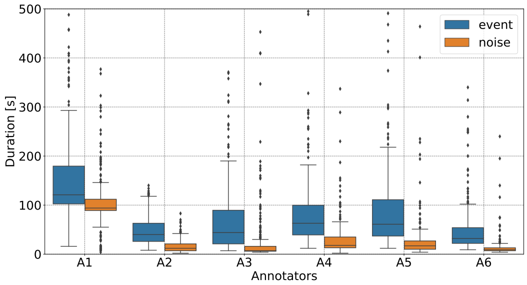

Figure 11: Duration of an annotation task for files containing events or not. Durations higher than 500s are not displayed.

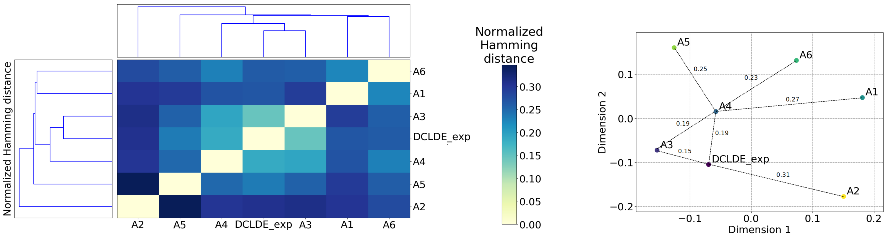

Figure 12: Left: HAC with heatmap representing Hamming distance metric between each pair of annotators. Hierarchical trees are shown on the upper left of the heatmap. Right: Divergence between annotators plotted using multidimensional scaling. The Hamming distance between some annotators was displayed on the dotted lines. "Unknown call" label was not taken into account in both figures.

are close (Hamming distance < 0.2). Between DCLDE_exp and A3, the Hamming distance was 0.15, meaning that they disagreed on about 212 annotations. The disagreement was slightly greater between DCLDE_exp and A4 (Hamming distance of 0.19 representing about 269 different annotations). However, the Fleiss κ scores were 0.7 and 0.63 revealing that a substantial inter-annotator agreement was found for the DCLDE_exp with A3 and DCLDE_exp with A4 pairs respectively. Two expert annotators (DCLDE and annotator A3) but also annotator A4 who is an amateur showed these substantial inter-annotator agreements. Annotator A6 who is also an expert in blue whale D-calls shows a divergent pattern from this group (DCLDE_exp, A3 and A4).

4. Discussion

Overall, our study brings further experimental evidence of inter-annotator variability in the annotation process of non-stereotyped blue and fin whale calls as already highlighted in previous works (Leroy et al., 2018). Although we observed similar behaviour patterns between the annotators such as the number of annotated D-calls and 40Hz calls (see Fig. 2), important discrepancies appeared when in-depth analysis of the annotation results was carried out.

Two quantitative metrics were first used to assess inter-annotator variability. The first one was the total number of annotated calls per annotator. This number was highly dependent on the annotator, with a maximal difference of 321 and 233 for D-calls and fin whale 40-Hz calls respectively (cf Fig. 2), while original DCLDE annotations reached a total of 719 and 156 calls respectively. Second, the inter-annotator agreement score Fleiss κ was computed, with values around 0.4 and 0 for D-calls and fin whale 40-Hz calls respectively. Following classical interpretation of this score (Landis and Koch, 1977), such values significantly reveal a medium to poor inter-annotator agreement. Also, theoretically, after a certain number of annotators, values should converge to a plateau-like region on the maximal inter-annotator agreement score reachable for a specific collaborative annotation campaign (Fleiss, 1975). Although this law was only partially verified experimentally in our study through the Δ Fleiss κ score (variations lower than 0.04, see Fig. 4), it can conceptually be used to estimate the minimal number of annotators that guarantees a maximal inter-annotator agreement, *i.e.* the highest annotation reliability one can expect from a given collaborative annotation campaign.

As a consequence, we further investigated the potential causes of inter-annotator variability. We first observed that this variability heavily depends on the call type. Over 782 events labeled as fin whale 40-Hz calls by at least one annotator, only 9 were labeled as fin whale 40-Hz calls by every annotator (1%). For the D-calls, this percentage reaches about 40 % of the annotated D-calls annotated by all annotators (cf Fig. 9). Furthermore, the "Unknown call" category is the one that displays the highest differences between the annotators with a ratio of 17 between the minimum and the maximum number of annotated calls (cf Fig. 4). Such trends might be explained by the similarity between the D-calls and fin whale 40-Hz call signatures, as also 305 samples were labeled with both labels, which means that the annotator could not clearly distinguish them.

The second cause that could explain the annotation differences between annotators is the salience of the calls, as partially measured by the SNR of the annotated calls. As expected, the agreement increases with SNR for D-calls as measured by the Fleiss κ score (Fig. 10), confirming the highest ambiguity between low SNR D-calls and noise. This was also observed for the detection of Antarctic blue whale Z calls (Leroy et al., 2018) and right whale contact calls (Urazghildiiev and Clark, 2007). This tendency was not observed for fin whale 40-Hz calls, for which the overall agreement (*i.e.* all SNR combined) is already poor.

Eventually, we also monitored the annotation profile by measuring the duration spent by each annotator on the annotation task (Fig. 11). The minimum duration for someone to listen to each audio file is 80 seconds (duration of the audio file 320 seconds divided by the maximum speed up ratio of 4, which has been almost always used). Except for annotator A1 (> 100s), the others took less than 80 seconds to annotate a file. Time spent on the annotation task reflects a certain behavior. For example, spending more time on the annotation task reflects probably that an annotator is more cautious. Overall, it is also interesting to note that the annotator profile does not correspond to annotation duration time, contradicting the intuition that less experienced annotators spend more time on each annotation display window.

The annotator profile is another source of inter-annotator variability investigated in this paper. Indeed, the results emphasized the subjectivity of the annotation task, which is mainly based on perception, and interpretation of the annotator. Especially, the "Unknown call" label may be more representative of the annotator "personality", as it reflects its overall level of confidence on this task, while the two other call type labels are more directly to the concrete skill of the annotator to discriminate both class independently from each other and from the background. Leroy et al. (2018) highlighted the effect of the annotator personality on their annotation "behavior". Indeed, one annotator annotated a lot of calls whereas the other tended to be more conservative and annotated less calls showing that the labeling behavior is related to subjectivity. Also, from our clustering analysis, we saw that an annotation pattern emerges with the annotators DCLDE / A3 and A4. Both DCLDE and annotator A3 were experts in the 40-Hz and D-call sounds. More surprisingly, annotator A4 presents a closer annotation pattern to the DCLDE / A3 while he is an amateur in underwater sounds. This confirms that even non-expert can provide high-quality annotation labels, which has already been observed in other research areas (Snow et al., 2008; Snel et al., 2012; Hantke et al., 2016). A similar result is reported in Rogers (2003), who describes two categories of expert bioacousticians. A splitter group tended to inventory each variant of a sound type as different sound types whereas a lumper group tended to regroup variants into single sound types.

Overall, these results have allowed us to identify a few guidelines on how to set up an annotation campaign in marine bioacoustics. First, an annotation campaign should involve more than one annotator, allowing for a minimal sanity check that informs us on the difficulty level of the annotation task with respect to the different call types to be annotated. In a machine learning context, it has already been recognized that evaluating and comparing algorithm performance on a poorly annotated dataset can lead to misinterpretations of their performance[4]). The number of annotators required to obtain robust annotations may vary with the difficulty of the task, but we suggest using at least two annotators, as it was enough to figure that the agreement for the 40 Hz calls annotation task waspoor. We realize that finding people willing to annotate acoustic data can be complicated and our study highlights that the participation of non-experts should not be excluded from underwater audio annotation. A second guideline we could formulate would be to employ inter-annotator agreement scores like Fleiss κ, envisioned as a standard, objective and absolute measure of the reliability of an annotation task, as well as of the minimal number of annotators required to maximize this reliability through the Δ Fleiss κ. This work focused on two similar calls that are hard to discriminate. We believe that increasing the number of sound types to identify might add confusion and ambiguity, especially if these sound types have similar time-frequency features. To keep the annotator focused on the task, we also recommend to set short duration annotation files. Having an open-ended question (identification of time and frequency of whale calls) makes it harder than having a closed task such as annotating a small audio sample (max 10s long) with predefined labels. At the end of each annotation campaign, a quality control Lee et al. (2018) could be set up with experts to review labels that divide annotators. In that way, both annotator quality and

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

ground truth inference strategies for label aggregation such as weighted voting could be determined. From our observations, it seems that annotators agree more on high SNR calls. Depending on the type of bioacoustic study, detecting only high vocalizations may be sufficient (*i.e.* to assess local presence of marine mammals). It would therefore be interesting to test the performance of the algorithms when they are trained only with high SNR calls vs when they are trained with all SNR calls.

Based on our study we now stress the need to systematically perform inter-annotation variability study prior to machine learning method development and validation, all the more so due to the renewed interest for the DCLDE challenge datasets (Socheleau and Samaran, 2018; Guilment et al., 2018; Shiu et al., 2020) that should now act as reference datasets for our community. Going further in this direction, we also champion the idea that such needs will require a new generation of more collaborative open tools. Our web-based annotation tool APLOSE (Nguyen Hong Duc et al., 2020), distributed freely to the community, is a first step in this direction from our part.

To the best of our knowledge, our study is the first effort in better understanding variability sources in collaborative annotation campaigns in marine bioacoustics, following preliminary works by Leroy et al. (2018). However, many other sources of variability remain to be investigated. For example, Cartwright et al. (2017) found that the complexity of a soundscape, in terms of number and overlapping level of sound sources, might affect the agreement. The annotation subjectivity can also arise from the annotator's previous annotation experience. For example, an expert and someone who sees a spectrogram for the first time probability will not annotate the same way.

5. Conclusion

In this study, we presented a new annotator subjectivity dataset of two cetacean call types. We have shown that the annotators in this dataset each have a distinct labeling behavior such as the time spent to identify acoustic events and their annotator profile. This last annotator characteristic showed that even beginners in labeling audio datasets could have a similar labeling behavior to experts. Furthermore, disagreements between annotators depend on the call type to annotate and their SNR. The large differences among annotator behaviors show that subjectivity plays a key role in annotating underwater sounds, which should be included into automatic classification systems of underwater sounds.

As a perspective for future works, note that our annotation campaign can still be joined by anybody, and the annotation results will be automatically updated. New annotation contributions from the community would allow to provide even stronger experimental evidence of our findings.

Acknowledgments

The authors acknowledge the Ple de Calcul et de Donnes Marines[5] (PCDM) for providing DATARMOR (storage, data access, computational resources, visualization, web-services, consultation, support services). The authors also would like to acknowledge Julie Beesau, Florent Boutonnet, Fabio Cassiano for their manual analyses of the dataset, the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase as well as Ana Širović, Simone Baumann-Pickering and the Scripps Institution of Oceanography for providing the DCLDE 2015 dataset.

Funding:. This work was supported by the CominLabs[6] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[7] and ISblue[8]. Paul Nguyen Hong Duc was supported by the the French Defence Procurement Agency (Direction Générale des Armements) and Maëlle Torterotot was supported by a PhD fellowship from the University of Brest.

Baumgartner MF, Bonnell J, Van Parijs SM, Corkeron PJ, Hotchkin C, Ball K, Pelletier LP, Partan J, Peters D, Kemp J, Pietro J, Newhall K, Stokes A, Cole TVN, Quintana E, Kraus SD. Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: System description and evaluation. Methods in Ecology and Evolution 2019;10(9):1476–89. doi:10.1111/2041-210X.13244.

Bergler C, Schrter H, Cheng RX, Barth V, Weber M, Noeth E, Hofer H, Maier A. Orca-spot: An automatic killer whale sound detection toolkit using deep learning. Scientific Reports 2019;9. doi:10.1038/s41598-019-47335-w.

Sirovic A. Variability in the performance of the spectrogram correlation detector for northeast pacific blue whale calls. Bioacoustics 2016;25(2):145–60.

Cartwright M, Dove G, Méndez Méndez AE, Bello JP, Nov O. Crowdsourcing multi-label audio annotation tasks with citizen scientists. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; CHI 19; 2019. p. 111. URL: https://doi.org/10.1145/3290605.3300522. doi:10.1145/3290605.3300522.

Cartwright M, Seals A, Salamon J, Williams A, Mikloska S, MacConnell D, Law E, Bello JP, Nov O. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations. Proc ACM Hum-Comput Interact 2017;29:21.

[5] http://www.ifremer.fr/pcdm

[6] https://www.cominlabs.u-bretagneloire.fr/

[7] https://www.afbiodiversite.fr/

[8] https://www.isblue.fr/about-us/

Clark C, Ellison W, Southall B, Hatch L, Van Parijs S, Frankel A, Ponirakis D. Acoustic masking in marine ecosystems: intuitions, analysis, and implication. Mar Ecol Prog Ser 2009;395(4):201–22.

Erbe C, Reichmuth C, Cunningham K, Lucke K, Dooling R. Communication masking in marine mammals: A review and research strategy. Marine Pollution Bulletin 2016;103(1):15 – 38. URL: http://www.sciencedirect.com/science/article/pii/S0025326X15302125. doi:https://doi.org/10.1016/j.marpolbul.2015.12.007.

Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975;31(3):651–9. URL: http://www.jstor.org/stable/2529549.

Guilment T, Socheleau FX, Pastor D, Vallez S. Sparse representation-based classification of mysticete calls. The Journal of the Acoustical Society of America 2018;144(3):1550–63. URL: http://asa.scitation.org/doi/10.1121/1.5055209. doi:10.1121/1.5055209.

Gwet KL. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.

Hamming RW. Error detecting and error correcting codes. Bell System Technical Journal 1950;29(2):147–60. doi:10.1002/j.1538-7305.1950.tb00463.x.

Hantke S, Marchi E, Schuller B. Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA); 2016. p. 2156–61. URL: https://www.aclweb.org/anthology/L16-1342.

Kairam S, Heer J. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. New York, NY, USA: Association for Computing Machinery; CSCW 16; 2016. p. 16371648. URL: https://doi.org/10.1145/2818048.2820016. doi:10.1145/2818048.2820016.

Kirsebom OS, Frazao F, Simard Y, Roy N, Matwin S, Giard S. Performance of a deep neural network at detecting north atlantic right whale upcalls. The Journal of the Acoustical Society of America 2020;147(4):26362646. URL: http://dx.doi.org/10.1121/10.0001132. doi:10.1121/10.0001132.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–74.

Lee W, Huang C, Chang C, Wu M, Chuang K, Yang P, Hsieh C. Effective quality assurance for data labels through crowdsourcing and domain expert collaboration. In: Bohlen M, Pichler R, May N, Rahm E, Wu SH, Hose K, editors. Advances in Database Technology - EDBT 2018. OpenProceedings.org; Advances in Database Technology - EDBT; 2018. p. 646–9. doi:10.5441/002/edbt.2018.75; 21st International Conference on Extending Database Technology, EDBT 2018 ; Conference date: 26-03-2018 Through 29-03-2018.

Leroy E, Thomisch K, Royer JY, Boebel O, Van Opzeeland I. On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions. The Journal of the Acoustical Society of America 2018;144:740–54. doi:10.1121/1.5049803.

Nguyen Hong Duc P, Torterotot M, Vovard R, Keribin E, Cazau D. APLOSE: a scalable web-based annotation tool for marine bioacoustics. Technical Report; 2020.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

Rogers TL. Factors influencing the acoustic behaviour of male phocid seals. Aquatic Mammals 2003;.

Shiu Y, Palmer K, Roch MA, Fleishman E, Liu X, Nosal EM, Helble T, Cholewiak D, Gillespie D, Klinck H. Deep neural networks for automated detection of marine mammal species. Scientific Reports 2020;10(1):1–12.

Snel J, Tarasov A, Cullen C, Delany SJ. A crowdsourcing approach to labeling a mood induced speech corpora. In: 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES. 2012. .

Snow R, OConnor B, Jurafsky D, Ng AY. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. USA: Association for Computational Linguistics; EMNLP 08; 2008. p. 254263.

Socheleau FX, Leroy EC, Pecci AC, Samaran F, Bonnel J, Royer JY, Carvallo Pecci A. Automated detection of Antarctic blue whale calls. Journal of the Acoustical Society of America 2015;138(5):3105–17. doi:10.1121/1.4934271.

Socheleau FX, Samaran F. Detection of Mysticete Calls : a Sparse Representation-Based Approach. Technical Report; 2018.

Thompson PO. Marine biological sound, west of San Clemente Island : diurnal distributions and effects on ambient noise level during July 1963. U.S. Navy Electronics Laboratory Report, 1965.

Torterotot M, Royer JY, Samaran F. Detection strategy for long-term acoustic monitoring of blue whale stereotyped and non-stereotyped calls in the Southern Indian Ocean. In: IEEE OCEANS. IEEE; 2019. p. 1–10.

Urazghildiiev IR, Clark CW. Detection performances of experienced human operators compared to a likelihood ratio based detector. The Journal of the Acoustical Society of America 2007;122(1):200–4. URL: http://asa.scitation.org/doi/10.1121/1.2735114. doi:10.1121/1.2735114.

Watkins WA. Activities and underwater sounds of fin whales [balaenoptera physalus]. In: Sci. Rep. Whales Res. Inst. volume 33; 1981. p. 83–117.

Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology 2016;16. doi:10.1186/s12874-016-0200-9.