(GIS) OSmOSE
Open Science meets
Ocean Sound Explorers

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

OSmOSE Working Report

Authorship This document was drafted by

- Paul Nguyen Hong Duc[1)]
- Maëlle Torterotot [2)]
- Dorian Cazau[3)]

belonging to the following institutes (at the time of their contribution): 1) Institut Jean le Rond d'Alembert, Sorbonne Universités, 2) IUEM, Université de Brest, 3) Lab-STICC, ENSTA Bretagne.

Document Review Though the views in this document are those of the authors, it was reviewed by a panel of acousticians before publication. This enabled a degree of consensus to be developed with regard to the contents, although complete unanimity of opinion is inevitably difficult to achieve. Note that the members of the review panel and their employing organisations have no liability for the contents of this document.

The Review Panel consisted of the following experts (listed in alphabetical order):

- Ronan Fablet[1)]

belonging to the following institutes (at the time of their contribution): 1) Lab-STICC, IMT Atlantique.

Last date of modifications June 11, 2020

Recommended citation Nguyen, P. et al. "Assessing inter-annotator agreement from collaborative annotation campaigns in marine bioacoustics", OSmOSE Working Report (version dating from June 11, 2020, distributed openly on `https://osmose.xyz/`)

Future revisions Revisions to this document will be considered at any time, as well as suggestions for additional material or modifications to existing material, and should be communicated to Dorian Cazau (dorian.cazau@ensta-bretagne.fr).

Acknowledgements We thank the Pôle de Calcul et de Données pour la Mer[1] from IFREMER for the provision of their infrastructure DATARMOR and associated services. We also would like to thank our main sponsors in this work: CominLabs[2] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[3] and ISblue[4]. The authors also would like to acknowledge the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase.

[1] `https://wwz.ifremer.fr/Recherche/Infrastructures-de-recherche/Infrastructures-numeriques/Pole-de-Calcul-et-de-Donnees-pour-la-Mer`

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

[3] `https://www.afbiodiversite.fr/`

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

Contents

Abstract

While it is now widely recognized that automated methods are crucial to help processing long-term recordings of marine bioacoustics, our community is still lagging behind in respect of developing large-scale annotated datasets used to supervise such methods. Besides being laborious and resource intensive, manual annotation might be also highly error prone, although such an issue currently lacks more systematic quantitative results to be fully addressed.

In this work, we study the question of inter-annotator agreement from collaborative annotation campaigns performed in marine bioacoustics. After providing quantitative evidence of inter-annotator variability, we investigated potential influences sources on both the user annotation practice and the annotation data and task to better understand why and how such a variability occurs. Our study reveals experimentally that the call type to be annotated, the SNR of the calls, and the annotator profile are three examples of critical factors impacting the annotation results from a collaborative campaign.

Chapter 1

Introduction

1.1 Context and Objectives

Automated methods of underwater sound recognition are a necessary step to develop our knowledge on the underwater environment that can improve ecosystems care. Supervised artificial intelligence techniques, which tend to exhibit the best recognition performance, highly rely on the amount and quality of annotated training data, and the process of collecting annotations is thus the main bottleneck in building such methods.

The traditional approach to collect annotations in the PAM scientific community has most often involved bioacousticians (more or less experts of the species to be identified) who manually annotate data. Although such an approach is currently thought to be the most accurate one (e.g. in comparison to automatic labelling), and has already allowed to build reference labelled datasets (e.g. DCLDE datasets `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`), such a work is well known to be resource intensive, laborious and time consuming.

Furthermore, in comparison to more traditional machine learning fields like computer vision, annotation in marine bioacoustics is compounded by the intrinsic difficulty in discriminating underwater acoustic sources, for which even experts may recognize some inextricable ambiguities (e.g. in the United States and Canadian east coasts in the discrimination between right and humpback whales, as humpbacks can sometimes make an upsweep call that is nearly impossible to distinguish from a right whale upcall (Baumgartner, 2019), or between overlaying killer whale harmonics and stationary boat noises at certain frequencies (et al., 2019)), whereas daily visual objects (cats, dogs, chair, etc) are more natural to human perception and then deemed for a more objective annotation.

Following this last statement, another drawback of manual annotation in marine bioacoustics is being error prone, with potential annotator-specific error bias for multiple-annotation campaigns. Inter-annotator agreement, or inter-rater reliability, is the extent to which human decisions coincide. Studying this agreement allows to measure the amount of consensus between a group of annotators. A high agreement means that both 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 this annotation (Gwet, 2014). To the best of our knowledge, although most PAM studies mention the use of several annotators in their annotation protocol, they generally only touch upon the question of inter-annotator variability, et al. (2018) being the only exception who fully addressed this problem. In this piece of work, the authors have measured the inter- and intra-analyst variability in manually annotated Antarctic blue whale stereotyped calls. They revealed both a strong inter-annotator variability between two analysts (with less than 50% agreement between analysts), but also a poor agreement obtained with a same person annotating the same audio segment twice.

To our point of view, such works are first concrete answer to recurrent criticism present in the discussion panels from the DCLDE workshop (e.g. Summary / Concluding remarks in `http://cetus.ucsd.edu/dclde/docs/pdfs/Wednesday/14-Gillespie.pdf,https://www.onr.navy.mil/reports/FY13/mbgilles.pdf`), highlighting the need for more exhaustive and reliable annotation campaigns based on consistent annotation protocols.

It is noteworthy that in other research areas, inter-annotator variability has been a well-studied issue. For example, regarding the current state of the art environmental labeling for language assessment, the LENA

system (Dongxin Xu and Gray, 2016), agrees with human annotators only 76% of the time on a four-way forced choice labeling task. While recent research in Music Retrieval Information address the issue of the subjectivity and the low agreement among different annotators (Flexer and Grill, 2016; Koops et al., 2019). Such works are especially critical to make informed choices on the annotation process settings used to build machine learning datasets. For example, Salamaon et al. (2014) set their maximum occurrence duration limit and sliding window size of their UrbanSound dataset based on a listening test (S. Chu and C.-C. Kuo), who found that 4 seconds were sufficient for subjects to identify environmental sounds with 82 % accuracy.

In this context, our work intends to pursue current efforts in better understanding inter-annotator agreement in marine bioacoustics. Our annotation campaign is composed of 8 annotators, plus two reference experts, on a whale vocal sound annotation task using the DCLDE 2015 LF dataset. Besides quantifying overall inter-annotation variability, we also investigated potential sources of parameters both on the annotator side (looking at the average duration spent on annotating and on the annotator profile) and on the annotation data side (looking at the call signatures and the SNR).

1.1.1 Related works

As already mentioned, within the field of marine bioacoustics, our study is an extended follow-up of et al. (2018)'s and M. Torterotot and Samaran (2019)'s works, which includes a higher number of annotator and more thorough sensitivity study on potential sources of inter-annotator variability. Outside our research field, our work has been highly motivated by current investigations done in urban soundscapes. (Cartwright et al., 2017) quantified the reliability/redundancy trade-off in crowdsourced airborne soundscape annotation by examining the effect of sound visualization (e.g. waveform or spectrogram) and acoustic characteristics (e.g. type and density of sound event) on the abilities of novice labelers to perform the audio annotation task accurately and efficiently. They also found that agreement increased with the number of annotators, and suggest that 5 annotators may be a reasonable choice for reliable annotation. They found that complex soundscape (loger, overlapping sound events) had a negative effect on agreement, which suggest that multiple participants are needed when working with complex under-water audio data. When collecting annotations from complex soundscapes, precision of the labels can be trusted, even if the annotations may be incomplete (higher precision than recall). There is a sound-class effect on annotation quality. To mitigate confusion errors, they advise to provide example recordings of sound classes to which the annotators could refer when needed. (Cartwright, 2019) on multi-label audio annotation: they found that annotators tended to over annotate when attending to one class (therefore, must be used when we want high recall) at a time and under annotate when asked to attend to many classes (therefore, must be used when we want high precision).

Chapter 2

Material and methods

2.1 Annotation campaign

2.1.1 Dataset

The dataset used in this paper is the low frequency evaluation dataset from DCLDE challenge 2015. The DCLDE 2015 dataset consists of data from multiple deployments of high-frequency acoustic recording packages (Wiggins and Hildebrand, 2007) deployed in the Southern California Bight. Originally data have been decimated to 1 and 1.6 kHz bandwidth and were selected to cover all four seasons and from multiple locations. For the purpose of this study, only 50 hour of data were used, taken from the first 50 hours of the CINMS_18_B_d06_120622_055731.d100.x.wav file, recorded in June 2012.

DCLDE2015	LF
Sampling rate (kHz)	2
Annotated species	Balaenoptera musculus - blue whale D calls Thompson and Laboratory (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	'Dcall': 719, '40-Hz': 156 from the DCLDE challenge annotations

On the occasion of the DCLDE challenge in 2015, a first annotation campaign with two analysts was performed, which consisted in annotating time intervals of Blue whale (*Balaenoptera musculus*) D calls (Thompson and Laboratory, 1965) and Fin whale (*Balaenoptera physalus*) 40 Hz calls (Watkins, 1981). The annotation file made available consists in the fusion by majority voting of the two individual annotations.

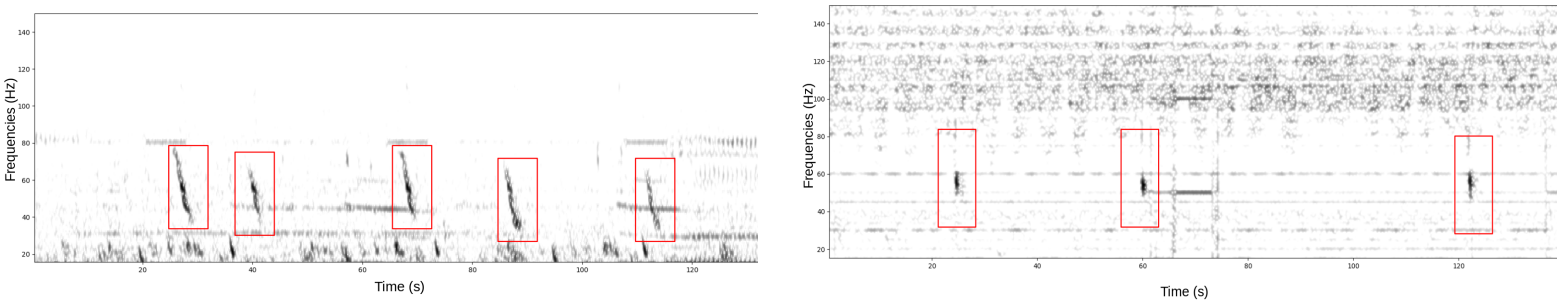

Figure 2.1: Example of calls: D calls (left) and 40 Hz calls (right).

2.1.2 Annotation support and protocol

As classically done, annotation was done both visually through time-frequency representations and aurally. Table 2.1 recaps a set of parameters used to compute and display annotation spectrograms. They were chosen to fit at best the original annotation protocol proposed by Sirovic et al. (?) Main discrepancies are that we used a special normalizing value, corresponding to the median of maximum values from filtered PSD in the [15-50] Hz frequency band.

Otherwise, the duration range of spectrogram window was set between 320 s to 40 s. Furthermore, annotators could use a zoom (only on time axis) from x1 to x8 (i.e. 320 s to 40 s) on the spectrogram, and an audio player with varying speed from

Annotators were given instructions (see Supplementary Materials) where visual but also aural examples of the sound to annotate were shown. They could refer to it at any time during the annotation process.

No annotator had access to the DCLDE experts' annotations. As a consequence, they could not adjust their annotations according to the DCLDE ones.

Analysts could chose between 3 labels, tagged as follow: DCALL (for Blue whale D call), 40-HZ (for fin whale 40 Hz call) and UNKNOWN CALL. The unknown label was used when an annotator was not sure if what he/she saw is a D call, a fin whale call or noise. This label is not used to tag sounds from unknown sources that appear in the spectrograms and that don't look like any of the calls that we're interested in. In the analysis, when not specified, 4 classes were used: Dcall, 40-Hz, Unknown call and when an annotator identify a sound but not the others, the former kept the label he/she gave and the latter were given the "None" label for that event.

Eventually, overlapping events were defined as if the midpoint of an event fell within the time bounds of the considered event, they were said to overlap (et al., 2018). In the case where a same annotator tagged two overlapping events, chronological order is kept in order to find corresponding labels for other annotators.

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]

Table 2.1: Parameter description of the different APLOSE seed datasets.

2.1.3 Annotation software

Our annotation campaign was conducted using the recently developed APLOSE system (). Note that our DCLDE2015LF campaign can still be joined by anybody, and annotation results automatically updated on the platform. To do so, please send a demand to the chief DCLDE2015LF campaign Paul Nguyen (p.nguyenhongduc@gmail.com), and he will give you access details. To our knowledge, these are the first attempts to provide to the marine bioacoustic community sustainable open collaborative annotation campaigns, which can be easily updated over a very long period of time by anybody. A first significant step towards crowd sourcing.

Furthermore, the csv result of our campaign has been made open source on APLOSE interface (even if you have not been enrolled as an annotator of the campaign), as well as a jupyter notebook[1] using panda operators processing this csv file, to make it easy for you to get started with such analysis.

[1]Download link:

Annotator	Expertise
A1	Bioacoustican
A2	Bioacoustican
A3	Expert
A4	Bioacoustician
A5	Amateur
A6	Expert

2.1.4 Annotator profiles

The annotators were volunteers and were not compensated financially for their annotations. No quality check of the annotators was set up for this campaign. It was opened to all citizen scientists. 6 annotators were found. We assume that these annotators were capable of identifying the two calls. They can be gathered into 3 groups: low-frequency whale experts (annotators have already annotated several hours of low-frequency mysticete calls, especially D calls), bioacoustisticians (they have already annotated several hours of cetacean sounds excpet Dcalls) and amateurs (no experience in underwater sound annotations).

2.2 Evaluation of inter-annotator agreement

Note that in our study we pretend we do not have any annotation ground truth, and our metrics rather qualify the relative level of agreement between annotators on a given annotation.

For the inter-agreement evaluation, 4 labels instead of 2 were used to compute metrics: “Dcall”, “40-Hz”, “Unknown call” and when an annotator annotated but the others don’t are considered as labels (label “None”). Expliciter le fait qu’on ne regarde que ce qui a été annoté au moins par 1 personne. On ne regarde pas l’agreement sur les ’none’.

2.2.1 Fleiss Kappa

To evaluate the inter-annotator agreement, the Fleiss Kappa [Fleiss et al., 1973], κ, score was computed. 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 raters assign a similar label that is not directly dictated by the data. In other words, it occurs when rater don’t know which label to give, and chose one randomly.

It can be interpreted as a measure of the difficulty of the task depending on the degree of agreement reached.

$$p_i = \frac{1}{n(n-1)} \sum_{j=1}^{K} R_{ij} \left(R_{ij} - 1\right), \quad p_j = \frac{1}{Nn} \sum_{i=1}^{N} R_{ij} \tag{2.1}$$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} p_i, \quad \bar{P}_e = \sum_{j=1}^{K} p_j^2 \tag{2.2}$$

$$\kappa_F = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2.3}$$

”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 raters agree for the i-th subject, $\bar{P}$ the mean of the p_i. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.” (From Wikipedia)

2.3 Annotation clustering

In order to characterize similar labeling behaviors, annotators were gathered into clusters Kairam and Heer (2016) using an Euclidean distance as a measure of the distance between annotators. Hierarchical agglomerative clustering (HAC) method was used with the Ward's minimum variance Ward (1963). Clusters were then formed from the computed hierarchical clustering based on the cophenetic distance between annotators. The threshold was set to 40 meaning that annotators in clusters have a cophenetic distance lower than 40. All computations were performed using scikit-learn package.

Chapter 3

Results

3.1 Quantitative evaluation of the inter-annotator agreement

3.1.1 Number of annotations

Fig. 3.1 shows the number of annotations per class for each annotator. Three groups can be distinguished for the D-call class based on the number of annotated events (DCLDE_exp and A3; A1 and A6; A2, A4 and A5, with about 700, more than 1000 and between 800 and 1000 annotated events respectively). For the 40-Hz class, A4 identified more events (almost 400) than the others annotators (about 200). Two groups in the annotators can be determined. 4 annotators used this label less than 100 times and 2 others about 200 times.

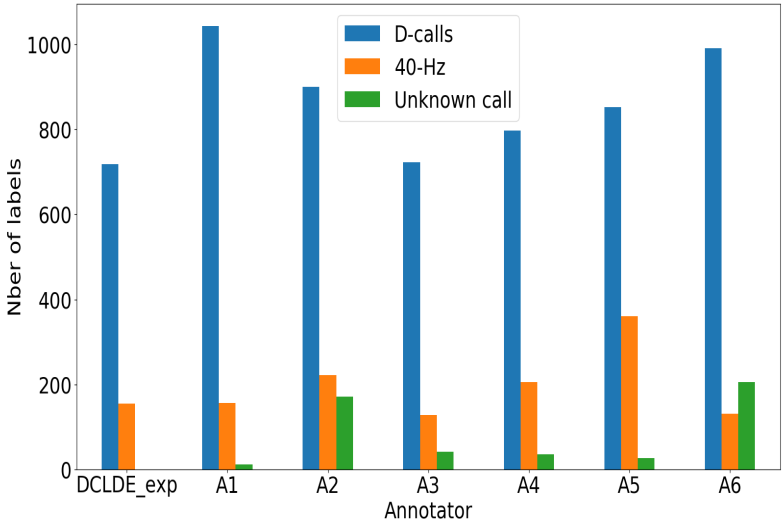

Figure 3.1: Number of annotations per class per annotator

3.1.2 Agreement metrics

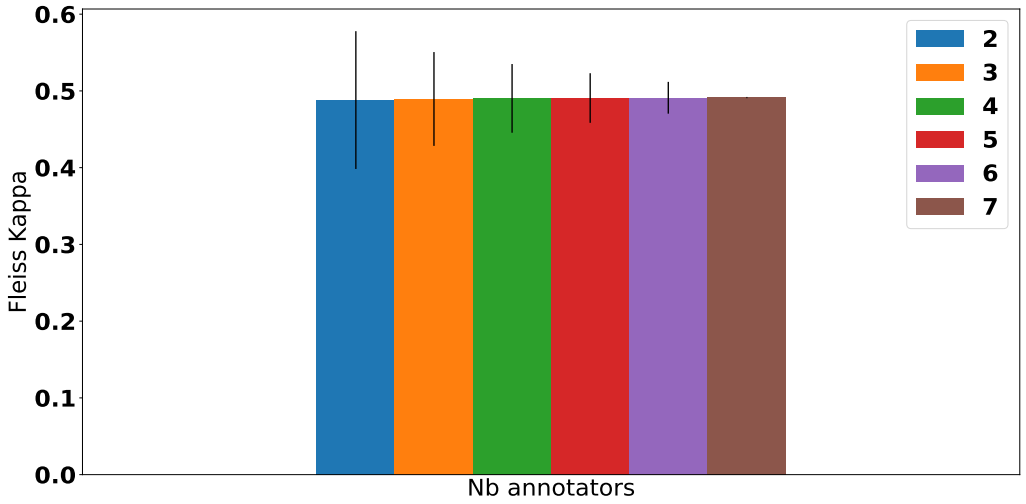

Figure 3.2: Reliability and inter-agreement of all annotators to identify only D-calls and 40-Hz pulses.

Fig. 3.2 shows the results on reliability and inter-agreement between annotators. Both metrics reveal evidence of variability as they both exhibit values significantly lower than 1, around 80 and 50 % respectively, regardless the number of annotators considered. MPA is more affected than Fleiss when increasing the number of annotators, going from almost 84% for 2 annotators to 81% for 7 annotators, while Fleiss remains constant at 0.48 from 2 to 7 annotators. The Fleiss Kappa measure reflects a passable inter-agreement between annotators.

- standard deviation of kappa higher with only 2 annotators revealing distinct clusters of a few persons

- with this metric there is supposed to have a low value bound for standard deviation , below which there are no visible inter-annotator differences , ie genre si low bound à partir de 10 annotateurs qu'on en ait 20 ou 30 alors ça n'apporte rien de plus, on a déjà tous les "comportements d'annotation"

- figure 3.2 : faire errorbar plot avec overall et les deux call types - figure 3.9 : faire aussi errorbar , there is a start of plateau-like curve pour dcall surtout sur les vocas avec plus gros snr , alors que dans 40hz on est en croissance perpétuelle , par contre faire gaffe au valeurs absolues de ces courbes!!

3.2 The reasons why variability, on the annotation data side

3.2.1 Call signatures and class labels

We first investigated how the class label (i.e. call type) impact the inter-annotator variability. In figure 3.3, we can observe that over 1241 annotated Dcalls, about half of them were annotated by all annotators, and less than 200 Dcalls were annotated by only one annotator. Most of the other Dcalls were identified by at least two annotators. Over 781 annotated 40Hz, only 9 were annotated by all annotators and more than 450 were annotated by only one annotator. Unknown calls were used sparingly but all annotators never used it to identify the same event. 335 events were identified by at least one annotator as a Dcall and by at least another one as a 40-Hz. Among these, 62 were identified by 6 annotators as a Dcall and by only one as a 40Hz. DCLDE, A4, A5, A6 identified a call as a 40-Hz while others identified it as a Dcall only 8 times, 6 times, 6 times and once respectively.

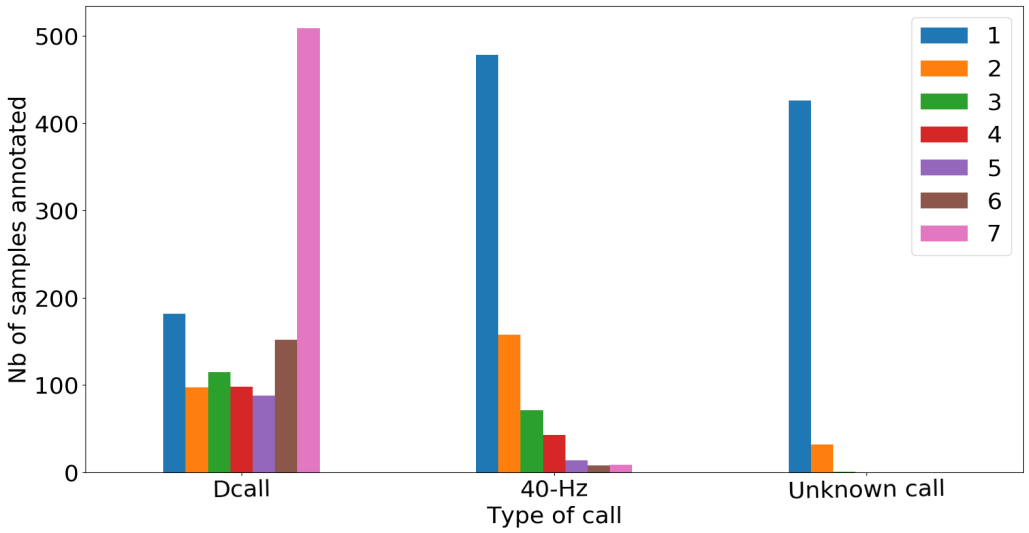

Figure 3.3: Number of calls identified by at least N annotators.

Similarity in call signatures may explain this unbalanced repartition when selecting the class label. Indeed, D-calls and the 40 Hz pulse are two similar call types characterized by a short downsweep time frequency shape, located in the same frequency bandwidth. Below, we propose a qualitative inspection of annotation spectrogram samples that were marginally labeled.

Figure 3.4 represents some spectrogram samples that were annotated by all annotators (on the left) and by only one annotator (on the right). As we will see in further details in the next section, events with salient acoustic features reached a more systematic consensus than those with less energy in noisy environment. Also, marginal labelling can be seen when the shape of a call differs from the typical call shape, as in graph ...

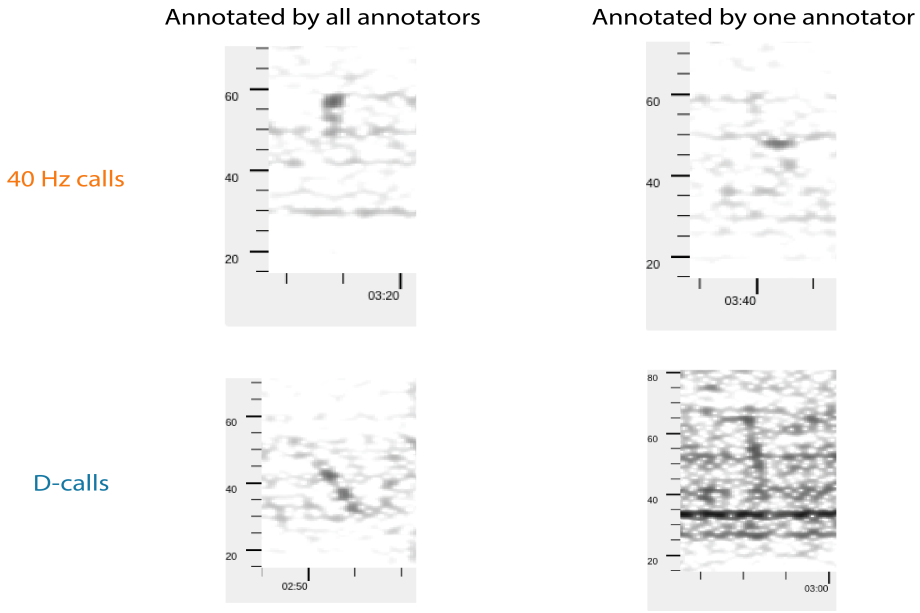

Figure 3.4: Examples of calls annotated by all annotators (left colum) and calls annotated by only one annotator (right column) for 40 Hz calls (top row) and D-calls (bottom row). X-axis is time in seconds and Y-axis is frequencies in Hz.

Some examples of labeled "Unknown calls" are represented in Figure 3.5. As expected, it is the class label revealing the most the ambiguity between the two call signatures.

Figure 3.5: Spectrograms of two Unknown calls identified by at least 1 annotator. X-axis is time in seconds and Y-axis is frequencies in Hz.

Finally, when crossing the class labels of D-calls and 40 Hz from all annotators, we saw that in total, 355 calls were labeled equally between the two classes, revealing a high ambiguity in properly identifying their call types. Two examples of these annotations are represented in Figure 3.6.

Figure 3.6: Spectrograms of two calls identified by (left) 2 annotators as 40-Hz, 2 annotators as Dcall and 3 did not annotate it (right) 2 annotators as 40-Hz, 3 annotators as Dcall, 1 as Unknown call and 1 did not annotate it. X-axis is time in seconds (interval between two ticks is 2 seconds) and Y-axis is frequencies in Hz (interval between two ticks is 2.5 Hz).

3.2.2 SNR

The distribution of the annotations according to their SNR is represented in Figure 3.7. More than 37% of the annotated D-calls and 46% of the 40Hz have a negative SNR whereas only 13% of D-calls and 8% of 40 Hz hve a SNR ¿ 10 dB. It suggests that the dataset is mostly composed of low SNR sounds.

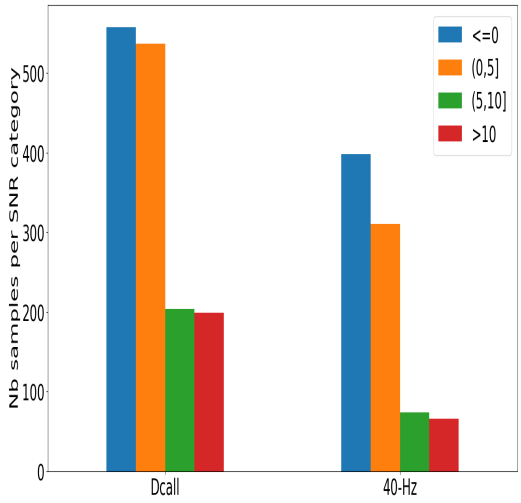

Figure 3.7: Distribution of the annotations according to their SNR, for D-calls (left) and 40 Hz (right). Each annotation is counted only once, even if it was annotated by multiple persons.

Figure 3.8 displays the number of samples annotated by only one person (orange) and by every annotator (blue) as a function of SNR. The number of D-calls annotated by only one person decreases as the SNR increases. The proportion of calls annotated by everyone increases with the SNR : 29% of negative SNR D-calls were labeled similarly and 45% of D-calls with a SNR ¿ 10dB were annotated by everyone. As for 40 Hz it is not possible to observe a trend in the proportion of calls annotated by everyone as a function of SNR, as only 9 calls were labeled as 40 Hz calls by every annotator.

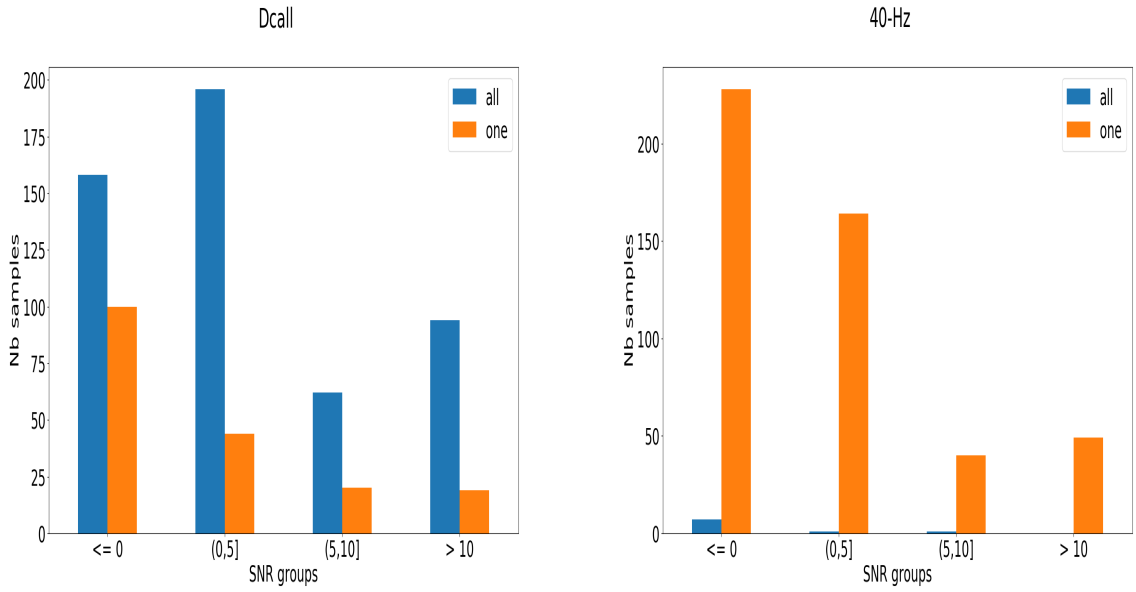

Figure 3.8: Nb of samples of Dcall (left) and 40-Hz (right) events per SNR category annotated by all annotators ('all') or by only one ('one').

Figure **??** represents MPA and Fleiss Kappa as a function of SNR. For D-calls, the agreement increases with the SNR, from [81% - 85%] for low SNR to [91-93 %] for high SNR. Moreover, the standard error of the agreement between multiple annotator decreases with higher SNR calls. On the contrary, the SNR does not seem to have a influence on the agreement for 40 Hz calls annotation.

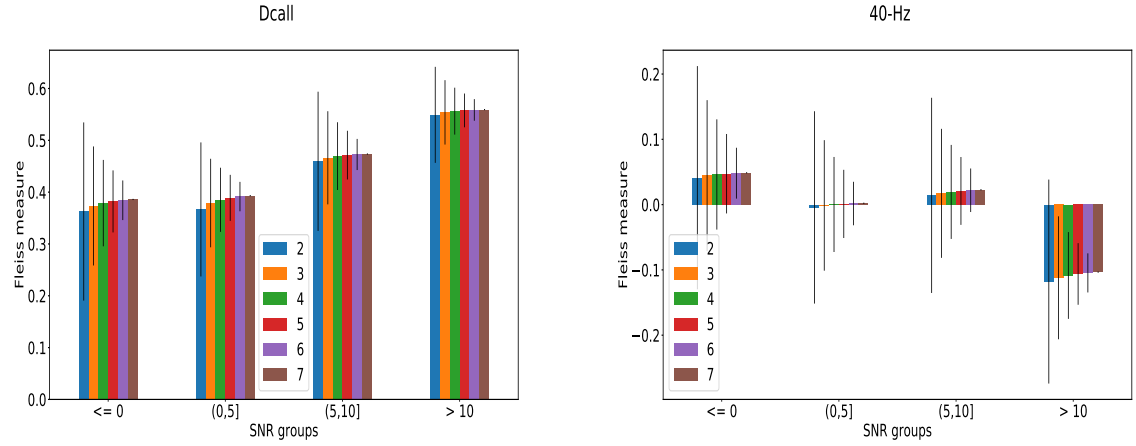

Figure 3.9: MPA and Fleiss measure per SNR category for expert annotators for Dcall class (left) for 40-Hz (right).

3.3 The reasons why variability, on the annotator side

3.3.1 Duration

Fig. 3.10 shows the average annotation time duration spent by each annotator to label each campaign file. Duration higher than 20 minutes were removed, assuming that they were not plausible candidates and rather corresponds to a misuse of our annotation tool (typically an annotator going on a break while leaving apart

an ongoing annotation task). Furthermore, as time duration were not available for the DCLDE experts we simply discarded them from this analysis.

Overall, results reveal a certain heterogeneity among the different annotators. When considering the 20 highest values of duration (from 744s to 14342s), A1 was found once, A3 4 times, A4 2 times, A6 3 times, A5 10 times and A2 was not found. Overall, A1 took longer to annotate the files than the other persons (median $>$ 100s). The other took less than 40 seconds (median levels). A2 almost doesn't present any upper outlier, or only very small one ($<$ 150 s).

Figure 3.11 compares average annotation duration time for files containing at least one annotation by one annotator, and files without annotation at all. In general, the annotators took less time to annotate files that they didn't think contained any call. For files with identified events, the median duration was about 30s higher than for files with no event. Minimal values for event and noise file are close ranging from 2s to 16s. However, A1 still spend more time than the other on these files.

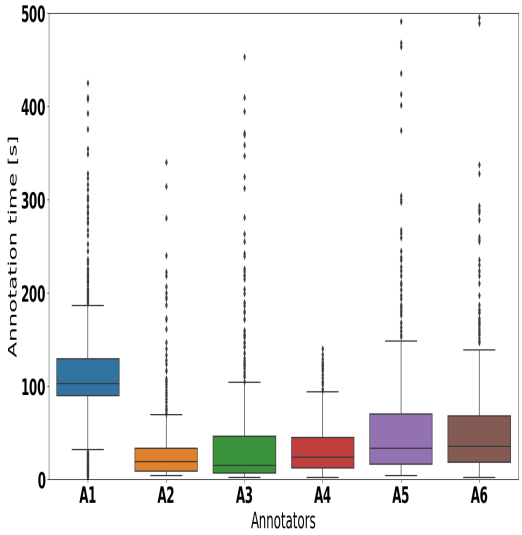

Figure 3.10: Boxplot of Annotation time duration for each annotator except DCLDE experts. Upper and lower box limits represent the 1st and the 3rd quartile values and the median is indicated by a horizontal inside the boxes. Dots outside the boxes are outliers.

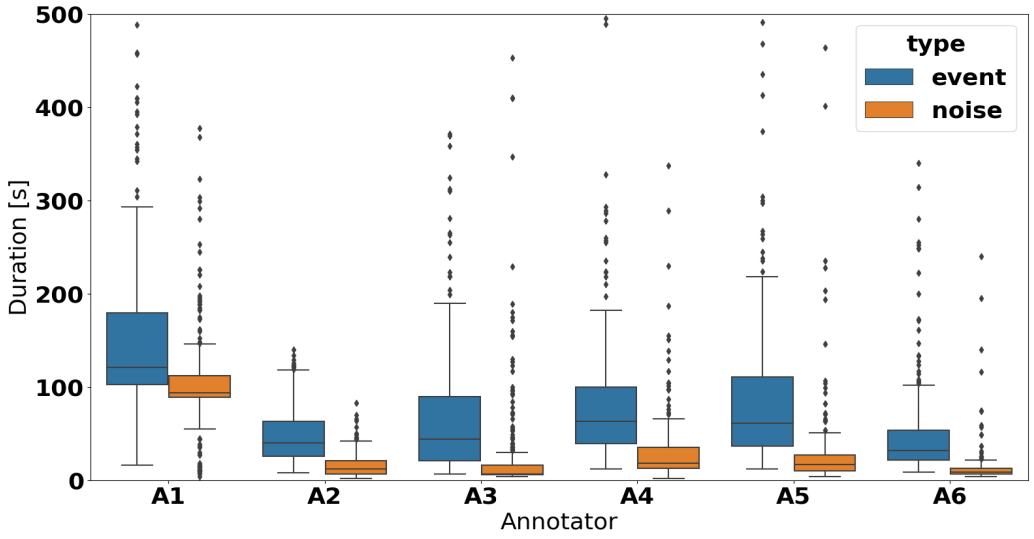

Figure 3.11: Duration of an annotation task for files containing events or not. Zoom on Y-axis, durations higher than 500s were not displayed.

3.3.2 Profile

Two clustering analysis were performed depending on if the unknown label was included in the analysis or not. Results are shown in Figs. 3.12 and 3.13. One main cluster of three annotators can be observed, with the DCLDE_exp, A3 and A4 annotators. In this subgroup, the MPA shows an almost perfect agreement between the three annotators and a substantial inter-annotator agreement with a Fleiss Kappa measure higher than 0.6, revealing a consistent annotation pattern for this cluster.

Having identified one specific cluster with consistent annotation patterns, its characteristics were examined. A cluster with a consistent annotation pattern could be defined by a MPA higher than 0.9 and a Fleiss Kappa measure greater than 0.7. The cluster is characterized by only expert annotations but A6 shows a divergent pattern from this cluster.

More surprisingly, A4 presents a closer annotation pattern to the DCLDE-A3 cluster than other annotators

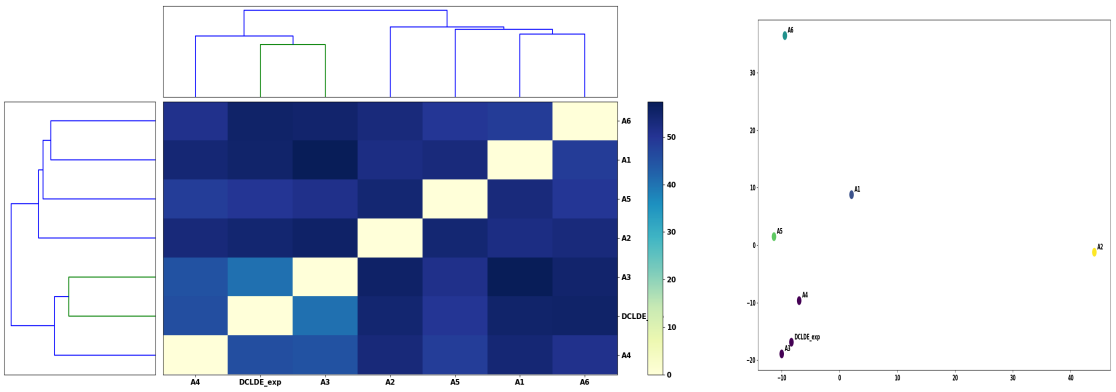

Figure 3.12: Left: HAC with heatmap representing Gower distance metric between each pair of annotators. Hierarchical trees are shown on the upper left of upper the heatmap. Right: Divergence between annotators plotted using principal components analysis (PCA). 'Unknown call' label was taken into account in both figures.

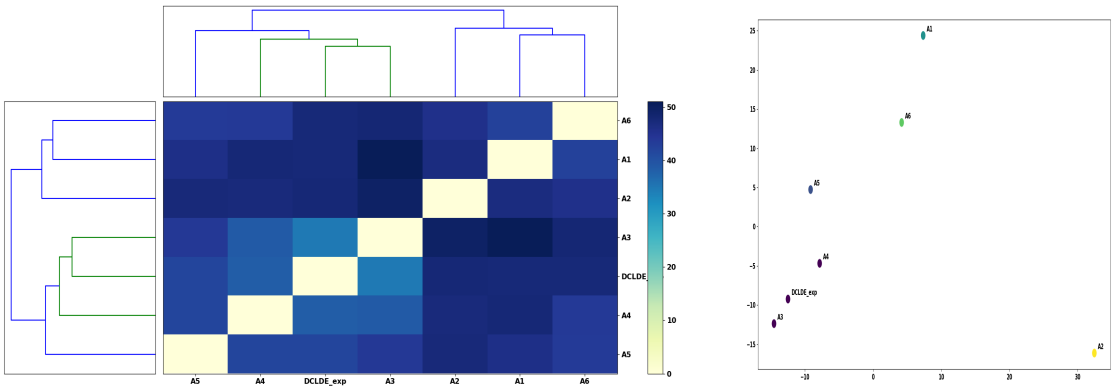

Figure 3.13: Left: HAC with heatmap representing Gower distance metric between each pair of annotators. Hierarchical trees are shown on the upper left of the heatmap. Right: Divergence between annotators plotted using principal components analysis (PCA). 'Unknown call' label was not taken into account in both figures.

A5 has a similar pattern to the DCLDE-A3-A4 cluster according to its distance to the other annotators of the group. Moreover, in Table **??**, even by adding A5 to the cluster, the annotation pattern is still consistent with a MPA measure of more than 0.8 and a Fleiss Kappa higher than 0.6.

Chapter 4

General discussion

Overall, our study brings further experimental evidence that confirm the presence of inter-annotator variability in the practice of whale vocal sound annotation. Although we found common trends between our annotators, like for example that more D-calls were annotated than 40Hz calls (see Figure 3.1), when looking more precisely at annotation results, important discrepancies can be highlighted. In our study, we first used three quantitative metrics to put them into evidence.

First we observed that the total number of annotations was dependent on the annotator, with a range value of to . Such a difference confirmed previous findings from et al. (2018), who observed in a similar task of annotating Antarctic blue whale Z-calls with two annotators, annotator numbers from less than 200 to more than 1200 calls of intermediate and low intensities.

The presence of inter-annotator variability in our campaign was further confirmed quantitatively with two classical metrics, namely MPA and Fleiss Kappa, which are widely used in various literature to assess inter-rater agreement (). The more annotators, the more different interpretations are added which can explain the decrease in the MPA score. This score should first decrease and then reach a lower bound if the annotation task is not too difficult. However, with only 7 annotators, it is hard to say if the lower bound has been reached. On the other hand, the Fleiss Kappa stagnated around 0.48 revealing a fair agreement corrected by chance. It enables to show the difficulty of the task to identify the Dcalls and 40-Hz at the same time. Even if agreement coefficients such as MPA and Fleiss Kappa suffer from several drawbacks (**?**), the inter-annotator agreement characterizes the level of consensus and uniformity in annotations. It is seen as a way to measure the reliability of annotated data and thus their validity for setting up machine learning methods [Feyisitan2017, Karen Fort 2011].

The second part of our study investigated the potential causes behind inter-annotator variability. We first observed that inter-annotation variability depends heavily on the class label. On 849 samples labeled as 40 Hz calls by at least one person, only 9 were labeled as 40 Hz by every analysts (1%). For the D-calls, this percentage is way higher with 34% of the annotated D-calls annotated by all analysts (see figure **??**). Furthermore, the "unknown call" category is the one that displays the most differences between the annotators with a ratio of 17 between the minimum and the maximum number of identified events. Such trends might be explained by the similarity between the D-calls and 40 Hz call signatures, as also 305 samples were labeled with both labels, which means that the annotator couldn't clearly distinguish them.

The second parameter that can explain the annotation differences between analyst is the salience of the calls, as partially measured by the SNR of the annotated samples. As expected, the agreement is greater for high SNR calls than for low SNR calls. These latter are easier to miss if the annotator doesn't pay great attention or they can also be misidentified for another call type or a noise. A similar results is observed on the detection of Antarctic blue whale calls (et al., 2018) and on the detection of right whale calls (Urazghildiiev and Clark, 2007). However, all of the analysts also agreed on low SNR calls, suggesting that this is not the only parameters that explain the annotation differences.

Considering now variability sources on the annotator side, previous results initiated us to investigate the annotator profile as another source of inter-annotator variability. Indeed, these results emphasized the subjectivity of the annotation task which is mainly based on perception and interpretation of the annotator. The "Unknown" 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. From our clustering analysis, we saw that an annotation pattern emerges with the annotators DCLDE / A3. The cluster is characterized by only expert annotations but A6 shows a divergent pattern from this cluster. More surprisingly, A4 presents a closer annotation pattern to the DCLDE-A3 cluster than other annotators confirming that even non-expert can provide good labels like in other research areas (Snow et al., 2008; Snel et al., 2012; Hantke et al., 2016). Furthermore, the highest number of D-calls may make orient our label choices to go more often to this class label instead of the most represented one. An unconscious orientation that might be highly annotator-specific. et al. (2018) highlighted the effect of the analyst personality on their annotation "behavior". Indeed, one analyst annotated a lot of calls whereas the other tended to be more conservative and annotated less calls. A similar thing is described in Rogers (2003) in which two categories of bioacoustician experts are discribed. The 'splitters' describe many different variants of a sound type while the 'lumpers' cluster the sound types to form groups that encompass the variations.

Here, we observed the annotator personality by monitoring their annotation practice. The minimum duration for someone to listen to each audio file is 80 seconds (duration of the audio file (320s) divided by the maximum speed up ratio (4)). Except for A1 (median value), the others took less than 80s to annotate a file. A1 seems to be more prone to visually and aurally inspect audio files. It seems that A1 was maybe more dedicated to the task. Spending more time on the annotation task reflects probably that an annotator is more cautious. The fact that A1 was also the person that annotated the highest number of D-calls and used the less the unknown label suggest that this person meticulously observed the data. It is interesting to note that the profile of the annotator doesn't match with their annotation duration time, as indeed, we would have expected that less experienced annotators would spend more time on the annotation task.

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 et al. (2018) in this direction. However, many other sources of variability remain to be investigated. For example, (Cartwright et al., 2017) find that the complexity of the soundscape might affect the agreement. The DCLDE database, contains a lot of impulsive noise, that look very similar to a fin whale 40 Hz call. The annotation subjectivity can also arise from the analyst's previous annotation experience. For example, an expert and someone who sees a spectrogram for the first time probablity won't annotate the same way. Even amongst experts, a person that is used to work with non-noisy recordings, containing very loud calls will probably tend to annotate only the louder calls whereas someone that is used to noisy environment will be very careful and annotate more faint calls. In our study, the analysts were all familiar with spectral representations, but at different extent, except A4 who was a complete novice.

A high inter-rater reliability coefficient also raise multiple questions : What raters are interchangeable ? Only the ones that were part of the study or can we extrapolate to other similar raters who may not have participated in the study ? if the raters agreed on the specific data that were rated, will they agree when rating similar data?

Chapter 5

Conclusion and perspectives

It would be interesting to investigate the inter-annotator variability in a more systematic and controlled manner than usually done in marine bioacoustics, but this is beyond the scope of the present study. (For example, it can be argued that the second expert may have been biased by prior knowledge of the labels proposed by the first expert and the DNN.)

in which human interpretation and perception are generally used as agents for setting up a ground truth prior to building labelled datasets dedicated to machine learning method development.

In theory, nothing should preclude this, and this observation further motivates our interest in studying how subgroups may hold interpretations which diverge from the expert labels in systematic, butdifferent, ways.

Bibliography

Baumgartner, M.F.e.a. (**2019**). "Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: system description and evaluation."

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

Cartwright, M.e.a. (**2019**). "Crowdsourcing multi-label audio annotation tasks with citizen scientists." ACM.

Dongxin Xu, U.Y. and Gray, S. (**2016**). "Reliability of the lenatm language environment analysis system in young children's natural home environment." Tech. rep.

et al., B. (**2019**). "Orca-spot: An automatic killer whale sound detection toolkit using deep learning."

et al., L. (**2018**). "On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions."

Flexer, A. and Grill, T. (**2016**). "The problem of limited inter-rater agreement in modelling music similarity." Journal of new music research, **45**, 239—251. URL `https://europepmc.org/articles/PMC5256035`.

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

Hantke, S., Marchi, E., and Schuller, B. (**2016**). "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification."

Janik, V.M. (**2005**). *Underwater acoustic communication networks in marine mammals* (Janik VM. . In: , editor. Animal communication networks. Cambridge University Press; 2005. p. 390–415.).

Kairam, S. and Heer, J. (**2016**). "Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks." pp. 1635–1646.

Koops, H., de Haas, W., Burgoyne, J., Bransen, J., Kent-Muller, A., and Volk, A. (**2019**). "Annotator subjectivity in harmony annotations of popular music." Journal of New Music Research, **48**, 1–21.

M. Torterotot, J.R. and 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, Marseille, France, 2019*. pp. 1–10.

Rogers, T.L. (**2003**). "Factors influencing the acoustic behaviour of male phocid seals." Aquatic Mammals.

S. Chu, S.N. and C.-C. Kuo, title = Environmental sound recognition with time-frequency audio features, v...n...p...y...p..I. (**????**).

Salamaon, J., Jacoby, C., and Bello, J. (**2014**). "A dataset and taxonomy for urban sound research." pp. MM'14, November 3–7, 2014, Orlando, Florida, USA.

Snel, J., Tarasov, A., Cullen, C., and Delany, S. (**2012**). "A crowdsourcing approach to labelling a mood induced speech corpus."

Snow, Rion, O'Connor, B., Jurafsky, D., and Ng, A. (**2008**). *Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. EMNLP'08*, pp. 254–263.

Thompson, P.O. and Laboratory, U.N.E. (**1965**). *Marine biological sound, west of San Clemente Island : diurnal distributions and effects on ambient noise level during July 1963* (San Diego, Calif. :U.S. Navy Electronics Laboratory,), 48 pp.

Urazghildiiev, I.R. and Clark, C.W. (**2007**). "Detection performances of experienced human operators compared to a likelihood ratio based detector." The Journal of the Acoustical Society of America, **122**, 200–204. URL `http://asa.scitation.org/doi/10.1121/1.2735114`.

Ward, J.H.J. (**1963**). "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association, **58**, 236–244.

Watkins, W.A. (**1981**). "Activities and underwater sounds of fin whales [balaenoptera physalus]."