Machine/Deep Learning categorisation of sub-kilohertz Arctic soundscapes

Jonathan Cleverly[1], Philippe Blondel[1], Hanne Sagen[2], Espen Storheim[2], and Matthew Dzieciuch[3]

[1]University of Bath, Bath, UK
[2]Nansen Environmental and Remote Sensing Center, Bergen, Norway
[3]Scripps Institution of Oceanography, San Diego, USA

Contact: Jonathan Cleverly, Department of Physics, University of Bath, Claverton Down, Bath, United Kingdom, BA2 7AY, jgc60@bath.ac.uk

Abstract: *Arctic soundscapes are being modified by climate change, which is greatly amplified in the region. Cryophony (sounds from sea ice processes) will become more variable as ice floes become increasingly fragmented. These changes to the sea ice will also result in shifts in temporospatial patterns of marine mammal vocalisations and anthropogenic sounds. These markers of the state of the Arctic Ocean are monitored using passive acoustic technologies, however there are still no standard practices for exploring soundscapes in this region. Here we investigate Machine/Deep Learning (ML/DL) approaches for categorising deep-water Arctic soundscapes. Recordings from hydrophone moorings deployed along the Nansen Basin during the "Coordinated Arctic Acoustic Thermometry Experiment" (CAATEX, 2019-2020) have been considered for this study. We utilise AVES (Animal Vocalisation Encoder based on Self-Supervision) to identify sounds within recordings with broad descriptors. Training datasets for ML/DL algorithms usually consider a broader frequency range (beyond 20 kHz), but it is not always feasible to use these higher sampling rates, thus the robustness of algorithms requires testing with lower sample rate data (976 Hz here), where the frequency content of sounds is not always fully recorded. To study multiple sound sources at once (e.g. whale songs, anthropogenic sounds), we consider longer context windows ('snippets') of 120 seconds, currently seldom considered in acoustic ML problems. These techniques will be crucial for avoiding current bottlenecks in data processing, in particular in the Arctic, enabling more in-depth studies for marine mammal conservation and industrial regulation.*

Keywords: *Arctic, Nansen Basin, Soundscape Analysis, Machine Learning, Remote Sensing*

1. INTRODUCTION

The study of soundscapes is a crucial part of marine ecosystem management [1]. However, methodologies for assessing biodiversity and human impacts in temperate waters are not necessarily applicable to polar waters. Cryophony leads to significant variability in soundscapes [2], raising issues in using traditional methods, such as comparing third-octave bands.

The Nansen Basin of the Arctic Ocean remains understudied with regard to soundscapes. While under-ice hydrophone moorings have been successfully deployed in the region [3], there are still no universally effective strategies to automatically identify and separate features in recordings. ML/DL techniques have grown in popularity for detecting specific sounds within datasets, but these require large amounts of labelled data, which can be time-consuming to produce. Arctic marine mammal vocalisations and tones from sea ice can vary significantly in frequency and duration. Additionally, the acoustic 'surface duct' of ice-covered waters results in sounds warping and attenuating greatly during propagation. Hence, it is perhaps more suitable to first study soundscapes with a holistic approach, rather than searching for specific time-frequency signatures.

Here, we have considered a pretrained ML model, AVES [4], to process recordings taken during CAATEX [5] and categorise individual snippets of recordings by broad features, such as anthropophony, biophony and cryophony, then by a second set of category-specific descriptors. Locations of hydrophone moorings used throughout CAATEX are shown in Figure 1, as well as sea ice area fraction in April 2020 across the region. By developing a robust soundscape classifier, larger networks of hydrophones deployed on moorings (e.g. [6]) can be processed efficiently and automatically, freeing up valuable time for more focussed, source-specific research.

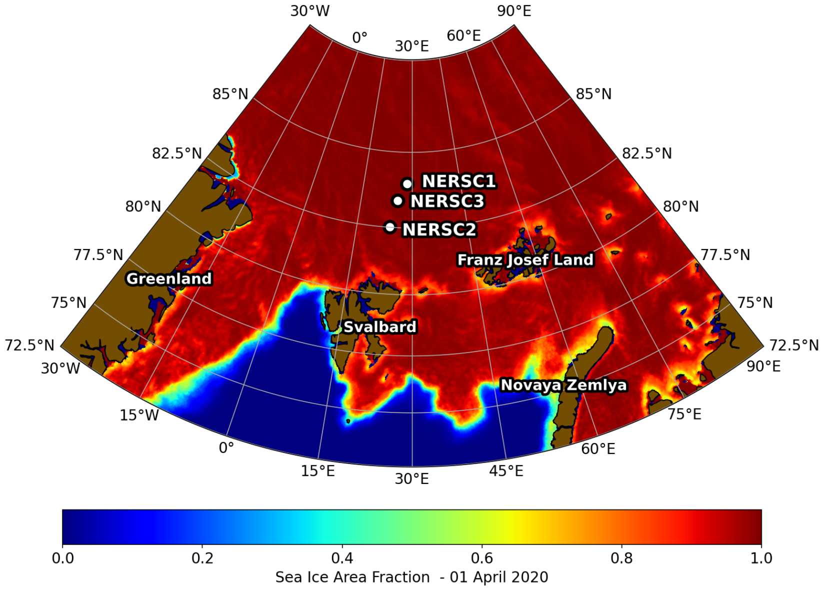

Figure 1: Location of moorings used during CAATEX. Sea ice area fraction in April 2020 taken from the Copernicus programme [7,8]. Map created using Cartopy [9].

2. DATA AND METHODOLOGY

2.1. SOUND PROCESSING AND ANNOTATION PROCESS

CAATEX recordings were taken between September 2019 and September 2020, every 12 hours for 45 minutes and 12 seconds, with a sample rate of 976 Hz. Only recordings from the shallowest hydrophone on each mooring were considered here, with depths ranging between 55-85 m. Recordings were truncated into 120-s snippets for input into AVES, often encompassing several acoustic sources. Partial annotation of the NERSC2 dataset was completed manually, selecting 3,000 snippets at random and labelling them by the following categories: ‘Anthro’ (anthropophony, i.e. sounds of human origins), ‘Bio’ (biophony, i.e. sounds associated with marine life, like whale vocalisations), ‘Cryo’ (cryophony, i.e. sounds associated with ice processes), ‘Geo’ (geophony, i.e. earthquakes) and ‘Noise’ (everything else). A second tier of annotations was completed to further categorise soundscape features, with biophonic activity split into five further categories, whereas anthropophonic and cryophonic activities were split into two categories each, transients and tones, for the 3,000 annotations (Table 1).

For manual inspection, spectrograms were computed with a segment length of 0.5 s with 50% overlap and Hann window convolution. Throughout the CAATEX deployment, a thermometry signal was transmitted every 3 days, dominating the soundscape for the first 18 minutes of recording time, so snippets covering this time were omitted from the study.

Category	Sub-category	Description	No. labels
Anthro	0	Consistent tones, e.g. shipping	209
	1	Impulsive sounds, e.g. seismic airguns (in concurrence with tones)	54
Bio	0	Bearded seal mating calls (distinctive trilled downsweeps)	308
	1	Repetitive downsweeps, terminating above 300 Hz	495
	2	Low frequency tones, below 100 Hz	53
	3	Mixed, few non-overlapping calls	637
	4	Mixed, many overlapping calls	213
Cryo	0	Impulsive ice cracking	283
	1	Tones from ice shearing or hums	59
Geo	0	Seismic activity	21
Noise	0	Absence of any identifiable or discernible signals	668

Table 1: Summary of the sound categories used in this study, with their respective numbers.

2.2. IMPLEMENTATION OF AVES

AVES is a transformer-based ML algorithm built primarily for encoding animal vocalisations [4], but here we utilise it for categorising sounds of man-made and geophysical origin as well. Based on the HuBERT model for human speech [10], AVES is pretrained on the large unlabelled audio datasets: AudioSet, FSD50K and VGGSound, covering a wide-range of sounds, both in air and water, and processing their waveform representations.

A recent *torchaudio*-based version of AVES has been released as a standalone Python package, *esp-aves*, and it has been used for this study [11]. Example configurations for AVES are available online and we selected *aves-base-all.torchaudio* ('AVES-all') as it had been trained on all segments segments in the pretraining datasets.

The *AVES-all* model consists in a 12-layer Convolutional Neural Network encoder with a 768-unit transformer for feature extraction. For both the 'Category' and 'Sub-category' runs of this study, the *AVES-all* model was trained using 1,000 of the annotated recording snippets, selected randomly, and then tested against the remaining 2,000. In both instances, the model was trained up to 30 epochs, with a default learning rate of 10^{-3} and batch size of 50 (i.e. 50 recordings read into the model at a time). Afterwards, the model was applied to the entire NERSC2 dataset (training snippets omitted) to observe temporal patterns of each sound source.

3. RESULTS AND DISCUSSION

3.1. TRAINING AND TESTING *AVES* WITH BROAD CATEGORY LABELS

The performance of *AVES-all* after 30 train/test epochs with broad category labels is presented in Figure 2. This model was able to recognise biophony with the highest success for all metrics (accuracy, precision, recall, Intersection over Union (Jaccard Index) and F1 score). The high number of samples for this category should however be compared with the numbers for other categories, although it is justified by the wide variety of vocalisations observed throughout the dataset. Snippets containing geophonic signals were labelled with moderate success despite the low sample size, although a significant number of snippets in the category were labelled as containing biophony, resulting from vocalisations concurrent with seismic events.

Most snippets featuring cryophony were correctly labelled, but the numbers of false positives and negatives were high. False positives may have occurred due to clipping within a snippet, appearing as a single intense transient above all other levels. Sometimes cryophony and biophony were coincident, resulting in false positives/negatives between the two categories.

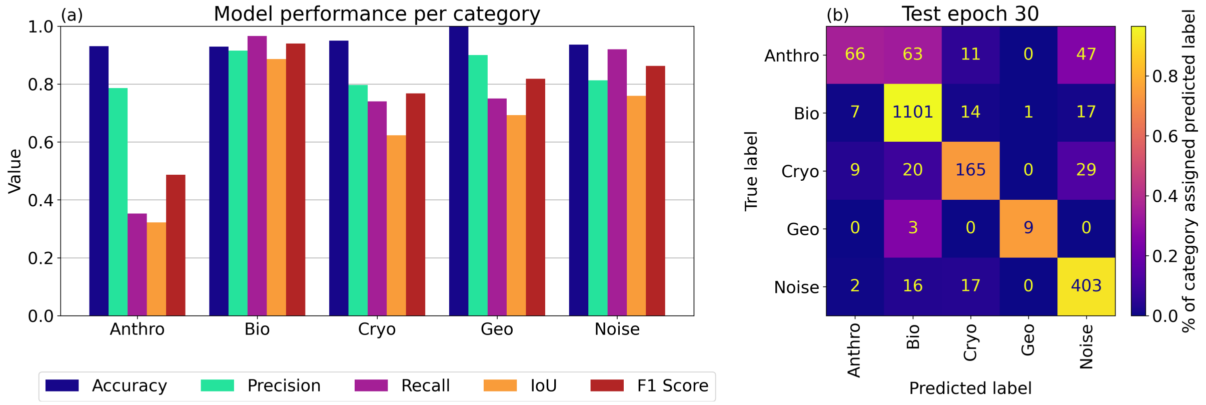

Figure 2: Model performance after training/testing for 30 epochs with broad category labels. (a) Performance metrics. IoU ('Intersection over Union') is also known as the Jaccard Index. (b) Confusion matrix of true vs. predicted labels.

The model struggled at correctly labelling snippets featuring anthropophonic signals, but the number of false positives was low. False negatives are as a result of anthropophony being largely faint against ambient noise except for the start and end of deployment (close proximity of the icebreaker deploying and retrieving the moorings), amongst other occurrences. Mislabelling of the noise category may have occurred due to the presence of single acoustic signatures of unknown origin, be they transients or short tones.

3.2. TRAINING AND TESTING *AVES* WITH SUB-CATEGORY LABELS

The performance of *AVES-all* after 30 train/test epochs with sub-category labels is presented in Figure 3. As expected for a model built for encoding animal vocalisations, *AVES-all* was better at correctly labelling sub-categories concerning temporally short events rather than continuous processes. The model was able to label most anthropophonic transients ('Anthro1'), but performance was poor for anthropophonic tones ('Anthro0'), though it should be noted that only a few isolated instances of anthropophony were included in the annotated dataset.

For biophonic activity, *AVES-all* was most effective at labelling files containing bearded seal mating calls ('Bio0'), perhaps due to these call types being the most consistent in time-frequency signatures. The majority of snippets featuring trains of downsweep calls ('Bio1') were labelled correctly, with lower precision and a significant number of mislabellings into other biophonic sub-categories. Low frequency calls ('Bio2') were labelled with good precision, but with low recall, in part due to calls being warped during propagation, causing some confusion with anthropophonic transients and tones. For mixed biophonic sub-categories, the model was more successful in labelling snippets with non-overlapping calls ('Bio3'), but with a notable amount of false positives and negatives. Snippets containing overlapping calls ('Bio4') were poorly labelled, with more false positives than true positives. Notably, 'Bio1', 'Bio3' and 'Bio4' shared many false negatives and positives, so if these sub-categories were merged together as a broader '$>$ 100 Hz calls' sub-category, the model may have been able to perform better.

Performance decreased for cryophonic events. While a decent precision was achieved, instances of cryophony were missed for both sub-categories, resulting in a lower recall. For cryophonic transients ('Cryo0'), the majority of false negatives were labelled as 'Noise' because transients were faint, and other false negatives arose along other biophonic sub-categories,

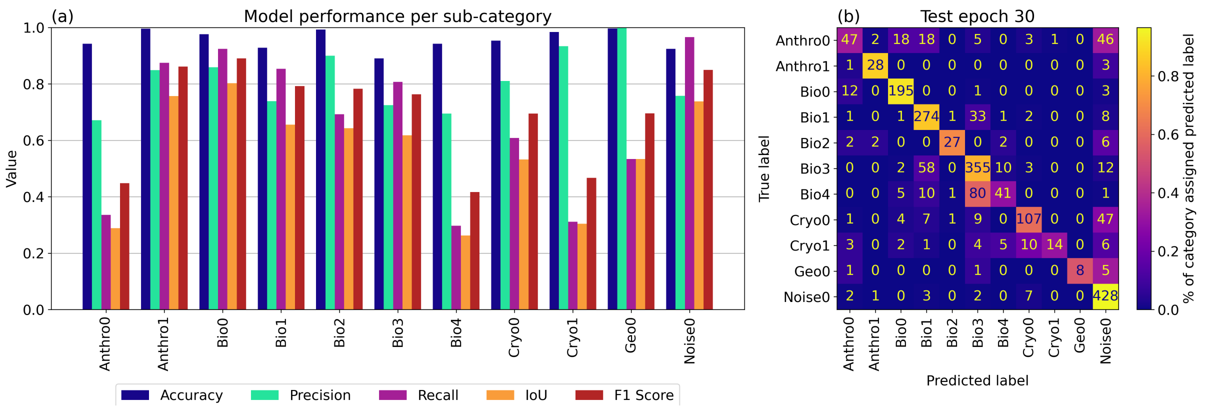

Figure 3: Model performance after training/testing for 30 epochs with sub-category labels. (a) Performance metrics. (b) Confusion matrix of true vs. predicted labels.

due to co-occurrence. For cryophonic tones (‘Cryo1’), many false negatives were labelled as ‘Cryo0’, likely because ice tones were rarely separate from transient events. Geophony snippets showed lower recalls, despite remaining a single category, with most false negatives labelled as noise. Snippets labelled as ‘Noise’ were recalled well, although there were a significant number of false positives across all other sub-categories, likely due to the sounds being faint.

3.3. PROCESSING THE ENTIRE DATASET WITH *AVES*

Temporal patterns of soundscape categorisation across the NERSC2 dataset are given in Figure 4 for both broad category and sub-category labels. The ‘summed weighted power spectral density (PSD) index’ ($swPSD$), an adaptation of the ‘bioacoustic index’ used in ecoacoustic studies, is computed using *scikit-maad* [12]. It is plotted here for the frequency bands of 10 - 100 Hz and 100 - 488 Hz (the highest frequency achievable here), highlighting the variance of acoustic energy above the ambient noise floor [13].

Between 100 Hz and 488 Hz, $swPSD$ shows a near consistent level in acoustic activity, regardless of what sources are present, with slight drops from December 2019 to mid-January 2020 and from mid-May to mid-June 2020. This coincides with time frames along the dataset labelled as predominantly featuring biophony. In winter months, ‘Bio1’, ‘Bio2’ and ‘Bio4’

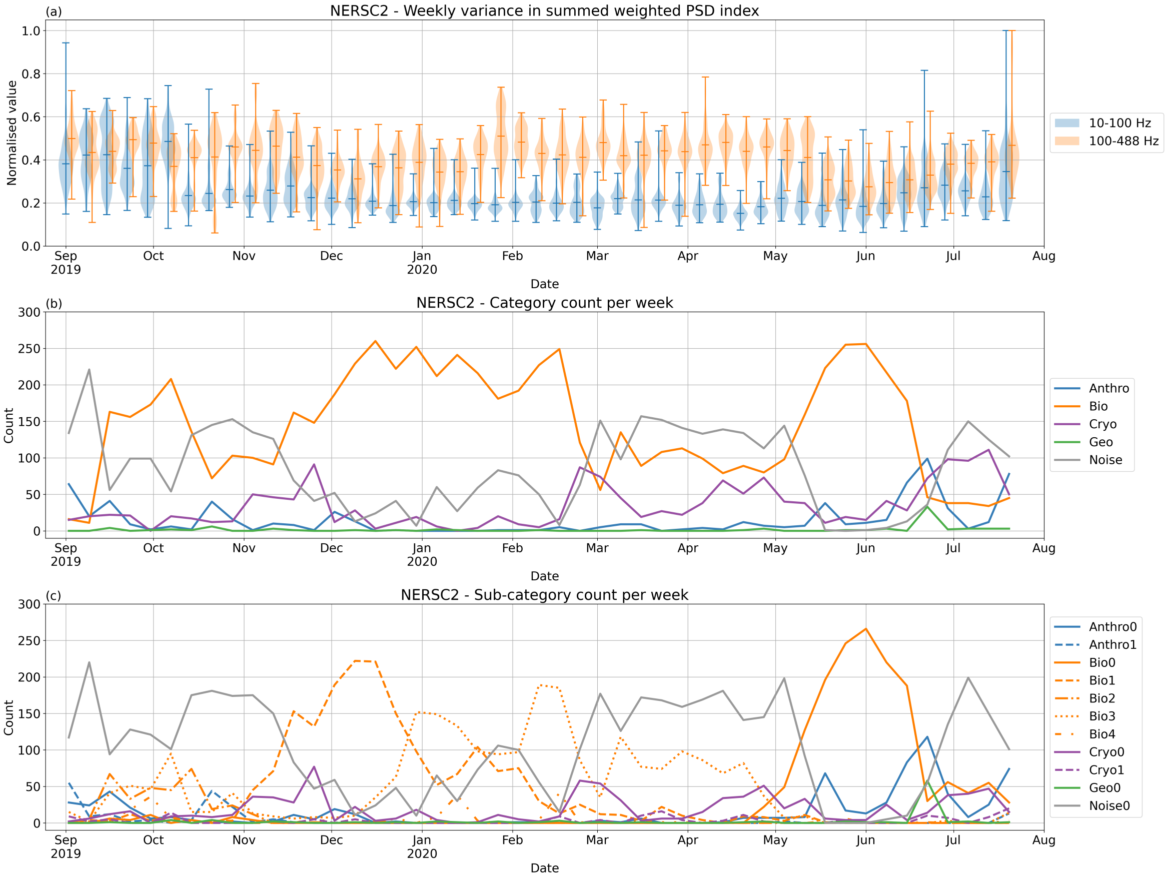

Figure 4: (a) Weekly variance in summed weighted PSD spectrum, across the frequency bands 10-100 Hz and 100-488 Hz. Weekly count of (b) broad category and (c) sub-category labels across entire NERSC2 dataset.

were the prevailing biophonic categories and they are attributed to bowhead whales. This is comparable to the vocal activity of bowhead whales observed in the Fram Strait during winter months [2, 14]. Bearded seal downsweep calls were observed only from mid-May to July 2020, which is comparable, but perhaps more restricted than other regions of the Arctic [15]. Sub-100 Hz calls were only observed between mid-September to November 2019, potentially from rorqual whale species visiting the region.

Anthropophony was largely restricted to the summer and autumn months, as shown not only by the increase in label count across these months, but from an increase in $swPSD$ when computed across 10-100 Hz. Some anthropophonic tones were noted up to December 2019, with sparse anthropophony next identified in March 2020. This is consistent with AIS records.

Along the dataset, increases in cryophony were often in line with increases in ambient noise. As presented in the confusion matrices in Figures 2 and 3, many cryophonic events may have been mislabelled as just noise. Around April 2020, ice cracking sounds may have been so frequent that they just accumulated into a raised ambient noise level. The number of snippets labelled as geophonic remained low across most of the dataset. The increase in geophonic snippets in June 2020 was partly from the presence of seismic events, but also from mislabelling of loud, modulating anthropophonic tones.

Comparing the two approaches of labelling the dataset, temporal patterns of sub-category counts are largely in agreement with their combined categories. If trained on more samples of isolated anthropophony, cryophony and geophony, the performance of *AVES* for identifying long-duration events, something it was not initially designed for, would improve.

4. CONCLUSIONS

AVES is a ML model originally built to identify animal vocalisations using audio waveforms. We show here that it can be used to broadly and automatically annotate large datasets from Arctic, ice-covered waters. Running the model across the entire deployment period identifies the temporal patterns of prevalent anthropophony, biophony and cryophony, along with isolated seismic events. The model allows for a quick first pass of annotation of acoustic datasets and it will next be applied to future datasets where temporal soundscape contributions are unknown. To further improve the applications of this approach, a set of models, e.g. *AVES* or other ML models, can be set to run in synchronicity to identify individual categories, assigning multiple and increasingly complex labels to recordings.

5. ACKNOWLEDGEMENTS

The CAATEX data used in this paper are used with permission from Principal Investigators of the CAATEX projects: Dr. Hanne Sagen, NERSC, (Project No. 280531) funded by the Research Council of Norway, and Dr. Matthew Dzieciuch, Scripps Institution of Oceanography, University of California (Project No. N00014-18-1-2698) funded by the Office of Naval Research. The HiAOOS project is funded by the European Union Horizon Europe Programme (Grant Agreement No. 101094621, UK participants supported by UKRI Grant No. 10071903). JC would like to acknowledge Gagan Narula (Earth Species Project) for his prompt correspondence and great advice in getting set up with *AVES*. This work has been completed as part of JC's EPSRC-funded PhD studentship (EP/W524712/1) supervised by PB.

REFERENCES

[1] Duarte, C. M., *et al.* : "The soundscape of the Anthropocene ocean", *Science* , **371**, eaba4658 **(2021)**

[2] Cleverly, J., Blondel, Ph., Sagen, H., Dzieciuch, M. and Storheim, E. : "Automatic identification of biophonics and sea ice processes in large datasets from the High Arctic Ocean", *International Conference on Underwater Acoustics, ICUA 2024, Bath, UK*, **46 (2024)**

[3] Worcester, P., Dzieciuch, M. and Sagen, H. : "Ocean Acoustics in the Rapidly Changing Arctic", *Acoustics Today*, **16**(1), 55-64 **(2020)**

[4] Hagiwara, M. : "AVES: Animal Vocalization Encoder based on Self-Supervision", *arXiv pre-print*, arXiv:2210.14493v1 **(2022)**

[5] NERSC : "CAATEX - Coordinated Arctic Acoustic Thermometry Experiment" https://caatex.nersc.no/ (last accessed: May 2025)

[6] NERSC : "HiAOOS - High Arctic Ocean Observation System" https://hiaoos.eu/ (last accessed: May 2025)

[7] Copernicus Marine Service : "Arctic Ocean Sea Ice Analysis and Forecast", https://doi.org/10.48670/moi-00004 (last accessed: May 2025)

[8] Williams, T., Korosov, A., Rampal, P. and Ólason, E. : "Presentation and evaluation of the Arctic sea ice forecasting system neXtSIM-F", *The Cryosphere*, **15**, 3207-3227 **(2021)**

[9] Met Office : "Cartopy: a cartographic Python library with a Matplotlib interface", https://scitools.org.uk/cartopy/docs/latest/ (last accessed: May 2025)

[10] Hsu, W.-N., *et al.*: "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", *arXiv pre-print*, arXiv:2106.07447 **(2021)**

[11] Narula, G. : "esp-aves 1.0.0", Earth Species Project **(2025)**, https://pypi.org/project/esp-aves/ (last accessed: May 2025)

[12] Ulloa, J. S., Haupert, S., Latorre, J. F., Aubin, T. and Sueur, J. : "scikit-maad: An open-source and modular toolbox for quantitative soundscape analysis in Python", *Methods in Ecology and Evolution*, **12**(12), 2273-2500 **(2021)**

[13] Bradfer-Lawrence, T., *et al.* : "The Acoustic Index User's Guide: A practical manual for defining, generating and understanding current and future acoustic indices", *Methods in Ecology and Evolution*, **00**, 1 - 11 **(2024)**

[14] Ahonen, H., Stafford, K. M., de Steur, L., Lydersen, C. and Wiig, Ø. : "The underwater soundscape in western Fram Strait: Breeding ground of Spitsbergen's endangered bowhead whales", *Marine Pollution Bulletin*, **123**, 97-112 **(2017)**

[15] Escobar-Amado, C. D., Badiey, M. and Pecknold, S. : "Automatic detection and classification of bearded seal vocalizations in the northeastern Chukchi Sea using convolutional neural networks", *The Journal of the Acoustical Society of America*, **151**(1), 299-309 **(2022)**