Fish recognition in underwater environments using deep learning and audio data

Laplante, Jean-François, Akhloufi, Moulay, Gervaise, Cédric

Jean-François Laplante, Moulay A. Akhloufi, Cédric Gervaise, "Fish recognition in underwater environments using deep learning and audio data," Proc. SPIE 11752, Ocean Sensing and Monitoring XIII, 117520G (12 April 2021); doi: 10.1117/12.2585991

SPIE.

Event: SPIE Defense + Commercial Sensing, 2021, Online Only

Fish Recognition in Underwater Environments using Deep Learning and Audio Data

Jean-François Laplante[a], Moulay A. Akhloufi[a], and Cédric Gervaise[b]

[a]Perception, Robotics, and Intelligent Machines Research Group (PRIME), Dept of Computer Science, Université de Moncton, Moncton, NB, Canada
[b]Institut CHORUS, Grenoble, France

ABSTRACT

Environmental conservation is a field where AI can provide significant help for many types of tasks. Oil, plastic, anthropogenic noise, overfishing and global warming are known to affect marine ecosystems (flora, fauna) inducing a drastic decrease of marine biodiversity and ecosystem services. The assessment of marine animals' distribution could benefit from automatic recognition of the presence of a species in a specific location. For this purpose, the passive acoustics monitoring can use underwater audio recordings and try to recognize the sound produced by the species. This work compares the performance of classical computer vision algorithms and modern deep learning methods for the task of identifying if a spectrogram contains the characteristic sound produced by the brown meagre. An accuracy of 95% was achieved using a deep convolutional neural network based on a recent architecture that was partially pretrained, outperforming classical computer vision algorithms.

Keywords: Fish recognition, Deep learning, Convolutional Neural Networks, Environmental conservation, Audio classification, Spectrograms

1. INTRODUCTION

Oceans cover more than 70% of the Earth's surface and its ecosystems, partly thanks to their biodiversity, serves the human race in many ways including climate regulation and as a food source. Since the industrial revolution, the extinction rate of species accelerates. Wilson[1]predicts an extinction of 50% of animal species by 2100. Marine biodiversity is in constant evolution following global changes (global warming), the increase of anthropic pressure (overfishing, pollution, maritime traffic), protection attempts (reserves, protected zones) and restoration attempts (artificial reefs). The evaluation of marine biodiversity is an issue of the 21st century. Solutions for remote sensing have been developed using optics and radars for the surface and acoustics for the depths of the water body.

For acoustics, two approaches coexist:[2] active acoustics, where researchers emit voluntarily a sound wave in order to listen to its echos to detect and localize submerged bodies, and passive acoustics, where researchers listen to naturally produced sounds by the sources forming what we call an acoustic landscape.[3] Recent developments of autonomous acoustics recorders allow us to accurately measure acoustic landscapes over long periods of time. Acoustics landscapes are made up of 3 components according to the nature of the source of the sounds. Anthropophony is emitted by human activities, geophony by natural phenomenons and biophony by animals. Marine biophony is mainly produced by cetacean, fish and benthic invertebrates.

Fish voluntarily emit sounds for the purpose of communication mainly during reproduction, for signaling their presence, to alert, and to defend a territory. The form of the signals are elaborated as they must carry information. However, the range of sounds stay limited as they are adapted to interactions with a distance going between from a couple meters to hundreds of meters. Fish also produce involuntary sounds when they feed. Amongst 28000 fish species, at least 700 have been identified as producers of sounds.[4] Fish produce sounds using 4 different mechanisms: by vibrating their swimming bladder, their tendons or other internal organs, by

Further author information: (Send correspondence to Moulay A. Akhloufi)
Moulay A. Akhloufi: E-mail: moulay.akhloufi@umoncton.ca, Telephone: +1 (506) 858-4120

Ocean Sensing and Monitoring XIII, edited by Weilin "Will" Hou, Proc. of SPIE Vol. 11752, 117520G · © 2021 SPIE · CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2585991

releasing gas from their swim bladder or by vibrating their teeth, bones or fins. The production by vibration of swim bladders and tendons are the most frequent and those that produce the highest sound intensity.

Depending on the emitting species, biophony covers a large frequency range (from 10Hz for mysticeti to more than 100 kHz for invertebrates and the clicks of odontoceti) and a large range (over 300 km for mysticeti to 300m for fish). Biophony is analyzed using 3 facets in acoustic ecology.[5] The first facet is for studying a particular species for which we know the acoustic repertoire. Recognizing and counting their signatures indicates presence periods, the number of individuals presidents and their activity using exchanged sounds (nutrition, reproduction, alert signals). The second facet is used for the quantification of marine biodiversity. As opposed to studying a particular species, we consider the sounds emitted by a global animal population. This ensemble forms an acoustic community. Evaluating the number of sounds, the number of sound families and the repartition of the sounds by family allows us to quantify the acoustic community. We have shown that a quantified acoustic community is a trustable index of marine biodiversity and that this index presents a better resolution than visual identification by divers for the purpose of evaluating the effects on marine biodiversity from the installation of marine reserves and protected marine areas. The third facet is used for the evaluation of the state of habitats through the acoustic community created by the habitat's fauna. We assume that a habitat in a good state holds a well-developed animal population that generates an acoustic community that presents a high acoustic score (from the second facet's standard). We have developed cartography tools that allow us to cartograph the biophony where each animal sound is localized and associated to a habitat,[6] and we have shown that different habitats emit different biophonies[7] and that the habitat's biophony is indicative of the state of marine prairies[8] and rocky habitats.[9] Other authors have shown that biophony is also indicative of the state of coral reefs.[10]

The assessment of marine animals' distribution (presence/absence, censusing, identification of vital functions such as feeding and breeding) could benefit from automatic recognition of the presence of a species in a specific location. For this purpose, the passive acoustics monitoring can use underwater audio recordings and try to recognize the sound produced by the species. The brown meagre is a specie of the sciaenid family, commonly called drums or croakers due to the drum-like croaking sound they produce by vibrating a muscle against their swim bladder. Due to a former intensive fishing and hunting, a drastic decrease of the population was observed in the 2 last decades and it is nowadays protected by a moratorium in French waters.

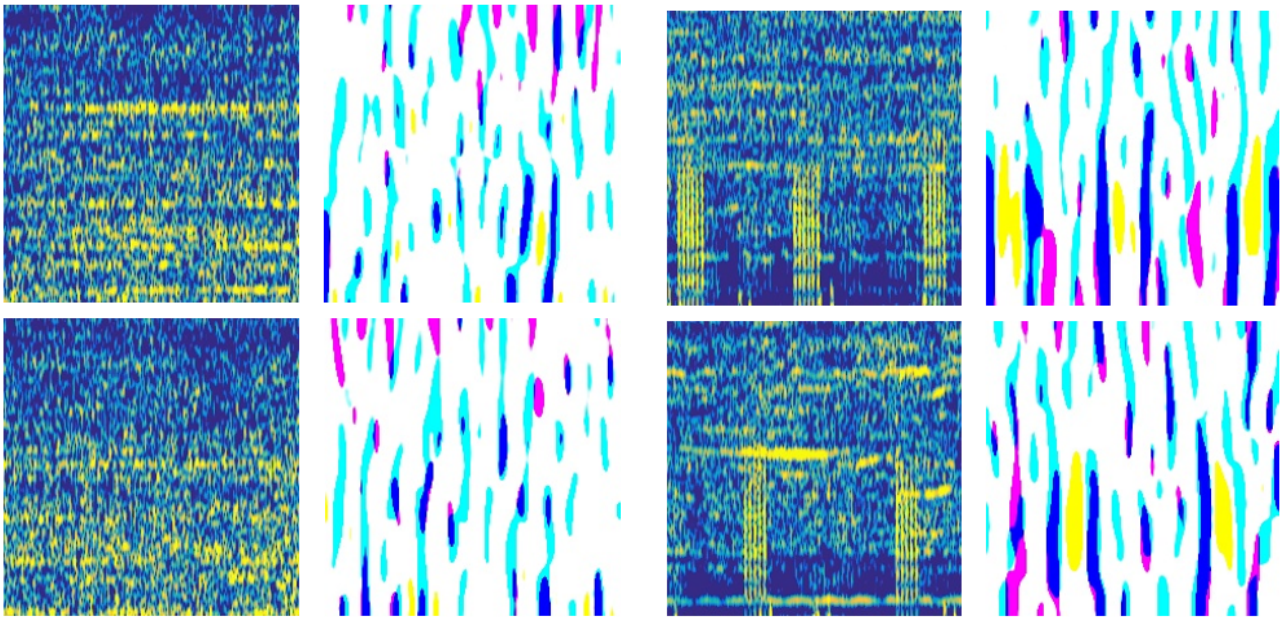
This work compares the performance of classical computer vision algorithms and modern deep learning methods for the task of identifying if a spectrogram contains the characteristic sound produced by the brown meagre thanks to a labeled data base of 9666 snapshots of 10 seconds containing noise only and 9926 snapshots of 10 seconds with brown meagre signatures.

2. PROPOSED APPROACH

5 methods for image processing, feature extraction and class inference were compared in this study:

- The average of the images' columns was taken to generate a histogram for column intensity on a per-image basis. These values were used to train a Neural Network.
- The images were transformed with a Gabor filter[11] to extract the characteristic columns. These processed images were used to train a Convolutional Neural Network. Examples are shown in figure 1.
- Features were extracted from the images with the NasNet Mobile network[12] and class inference was done with a neural network.
- Features were extracted from the images with the Efficientnet B3 network[13] and class inference was done with a neural network.
- Features were extracted from the images with the Efficientnet B7 network[13] and class inference was done with a neural network.

The NasNet Mobile, Efficientnet B3 and Efficientnet B7 networks are all deep learning models pretrained on the ImageNet database.[14] The ImageNet database contains more than 14 million labeled images. The upper layers, responsible for the feature extraction, are used as a base from which we can start our training of the inference layers and the fine-tuning of the feature extraction layers.

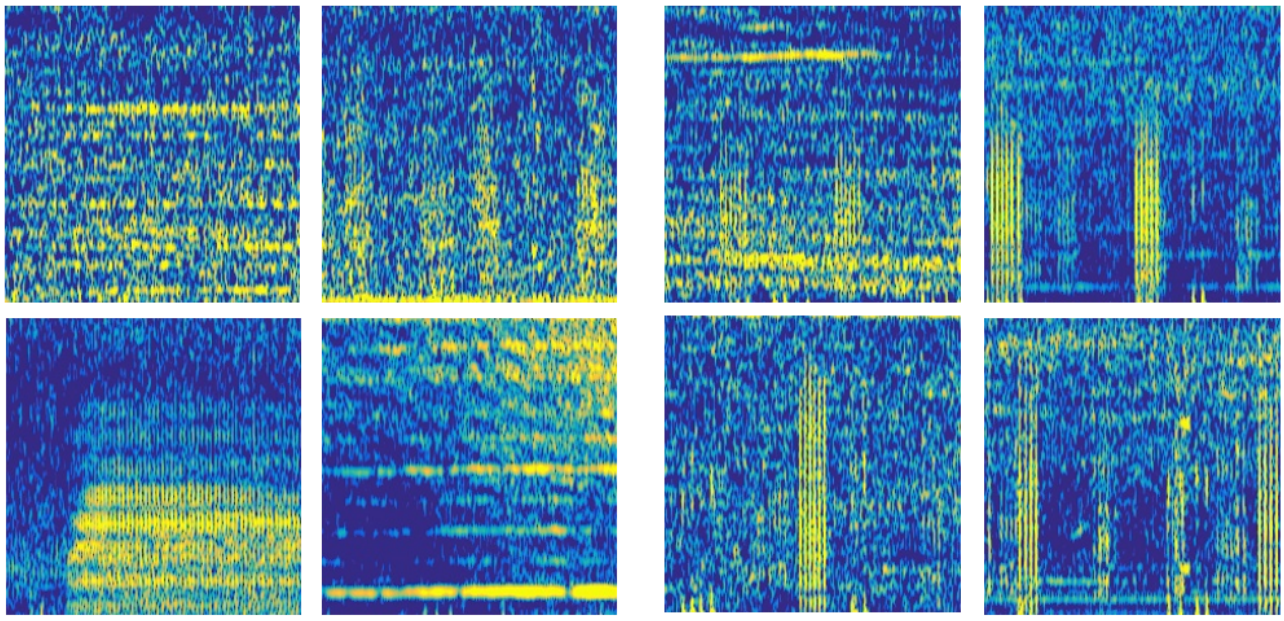


(a) Without corb (b) With corb

Figure 1: Example Gabor transformations of the spectrograms

3. DATASET

Our dataset contains 19592 spectrogram images (9926 of brown meagre sounds, 9666 of ambient sounds). The original images were 1678 by 1164 pixels, they were transformed to 224 by 224 pixel images containing only the spectrograms to serve as the deep network's input. Example spectrograms are given in figure 2.



(a) Without corb (b) With corb

Figure 2: Example spectrograms from our dataset

4. EXPERIMENTAL SETUP AND RESULTS

For training methods 1 and 2 we split the train and test sets into 0.7 : 0.3. For methods 3 through 5,we split the train, test and val sets into 0.7 : 0.15 : 0.15.

The performance of each method is shown in Table 1.

Table 1: Results of the spectrogram classification

	Method	Accuracy	Precision	Recall	F1 Score
1	Averaged columns + NN	0.77	0.88	0.64	0.74
2	Gabor + CNN	0.82	0.91	0.71	0.80
3	NasNet-M	0.78	0.96	0.59	0.73
4	EfficientNet-B3	0.91	**0.97**	0.85	0.91
5	EfficientNet-B7	**0.95**	0.96	**0.94**	**0.95**

Methods 1-4 noticeably have a high false negative rate (as seen with their relatively low recall). This is likely because these methods are not effective enough to accurately predict noisy images. The high precision of all methods is likely because these methods were all able to identify sound files with low noise. As the deeper architectures achieve a higher recall (methods 4-5), we suspect that the additional depth allows the networks to identify even the noisiest of images.

We note when visualizing misclassifications that many of them seem, to the human eye, like what the model incorrectly predicted. For example, the database's ambient noise spectrograms that were incorrectly predicted to be brown meagre sounds are shown in figure 3. A few of them show unusual spectrograms and a few more seem to bear the vertical lines we are looking for when trying to find brown meagre sounds.

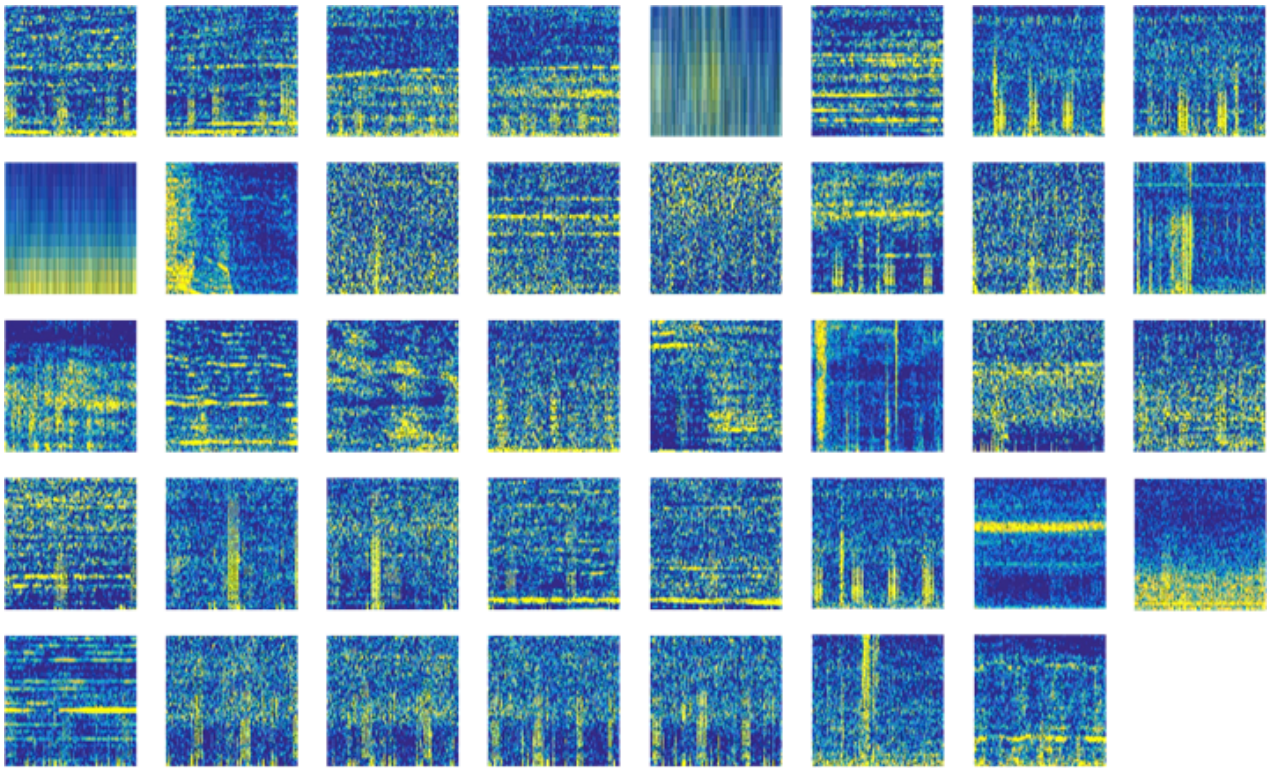


Figure 3: Examples labeled as brown meagre sounds incorrectly identified as ambient noise

As another example, the database's corb sound spectrograms that were incorrectly predicted to be noise are shown in figure 4. Many of these images show no sign of the characteristic vertical bars we are looking for. This seems to indicate a noise possibly caused by regular waves coming in contact with a boat, creating sounds which look like the brown meagre.

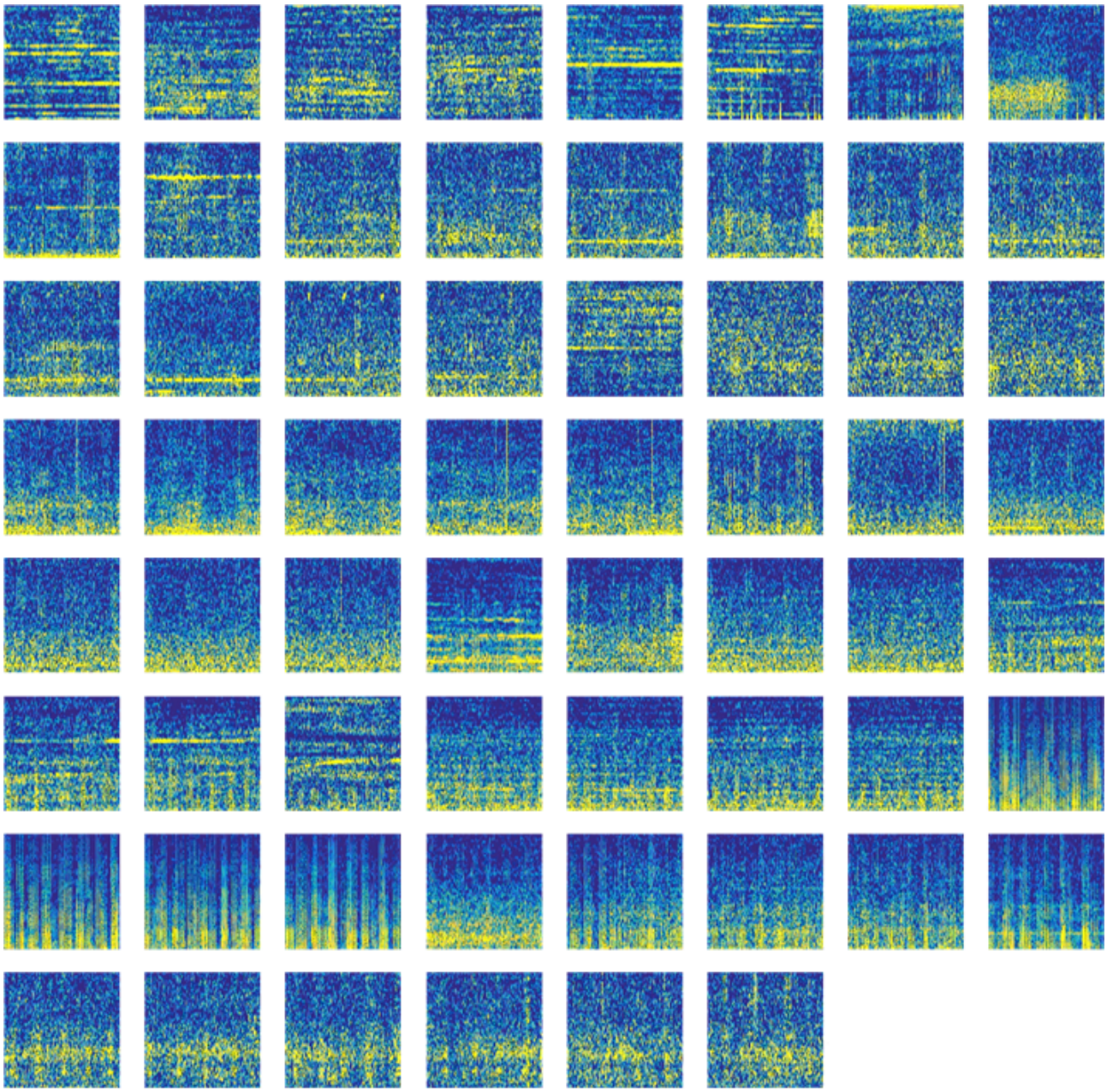


Figure 4: Examples labeled as ambient noise incorrectly identified as brown meagre sounds

5. CONCLUSION

This work proposes the use of the Efficientnet B7 network for the task of identifying whether or not a given spectrogram shows the presence of brown meagres and compares the performance of this network against other computer vision and deep learning methods. The proposed approach achieves very interesting results with an F1-score of 0.95.

Our results demonstrate the performance of deep learning methods on spectrograms containing the sound produced by brown meagres and their efficiency on classifying these types of sounds even in complex scenarios with the presence of noise.

Possibilities for future works includes the addition of new data and evaluating the effect of preprocessing the spectrogram images before feeding them to the network.

ACKNOWLEDGMENTS

This research was enabled in part by support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2018-06233.

REFERENCES

[1] Wilson, E., [*The future of life*], Alfred A. Knopf, New York (2002).

[2] Urick, R. J., [*Principles of Underwater Sound 3rd Edition*], Peninsula Pub (Aug. 1996).

[3] Pijanowski, B., Farina, A., Gage, S., Dumyahn, S., and Krause, B., "What is soundscape ecology? An introduction and overview of an emerging new science," *Landscape Ecology* **26**, 1213–1232 (Nov. 2011).

[4] Rountree, R. A., Gilmore, R. G., Goudey, C. A., Hawkins, A. D., Luczkovich, J. J., and Mann, D. A., "Listening to Fish," *Fisheries* **31**(9), 433–446 (2006).

[5] Sueur, J. and Farina, A., "Ecoacoustics: the Ecological Investigation and Interpretation of Environmental Sound," *Biosemiotics* **8**, 493–502 (Dec. 2015).

[6] Gervaise, C., Lossent, J., Valentini-Poirier, C. A., Boissery, P., Noel, C., and Di Iorio, L., "Three-dimensional mapping of the benthic invertebrates biophony with a compact four-hydrophones array," *Applied Acoustics* **148**, 175–193 (May 2019).

[7] Lossent, J., Iorio, L. D., Valentini-Poirier, C. A., Boissery, P., and Gervaise, C., "Mapping the diversity of spectral shapes discriminates between adjacent benthic biophonies," *Marine Ecology Progress Series* **585**, 31–48 (Dec. 2017).

[8] Iorio, L. D., Raick, X., Parmentier, E., Boissery, P., Valentini-Poirier, C.-A., and Gervaise, C., "'Posidonia meadows calling': a ubiquitous fish sound with monitoring potential," *Remote Sensing in Ecology and Conservation* **4**(3), 248–263 (2018).

[9] Desiderà, E., Guidetti, P., Panzalis, P., Navone, A., Valentini-Poirrier, C.-A., Boissery, P., Gervaise, C., and Iorio, L. D., "Acoustic fish communities: sound diversity of rocky habitats reflects fish species diversity," *Marine Ecology Progress Series* **608**, 183–197 (Jan. 2019).

[10] Freeman, L. and Freeman, S., "Rapidly obtained ecosystem indicators from coral reef soundscapes," *Marine Ecology Progress Series* **561** (Dec. 2016).

[11] Mehrotra, R., Namuduri, K. R., and Ranganathan, N., "Gabor filter-based edge detection," *Pattern Recognition* **25**, 1479–1494 (Dec. 1992).

[12] Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V., "Learning Transferable Architectures for Scalable Image Recognition," *arXiv:1707.07012 [cs, stat]* (Apr. 2018). arXiv: 1707.07012.

[13] Tan, M. and Le, Q. V., "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv:1905.11946 [cs, stat]* (Sept. 2020). arXiv: 1905.11946.

[14] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in [*2009 IEEE Conference on Computer Vision and Pattern Recognition*], 248–255 (June 2009). ISSN: 1063-6919.