


MARCH 13 2024

Development of a machine learning detector for North Atlantic humpback whale song

Vincent Kather ; Fabian Seipel; Benoit Berges; Genevieve Davis; Catherine Gibson; Matt Harvey; Lea-Anne Henry; Andrew Stevenson; Denise Risch



J. Acoust. Soc. Am. 155, 2050–2064 (2024)

<https://doi.org/10.1121/10.0025275>



View Online

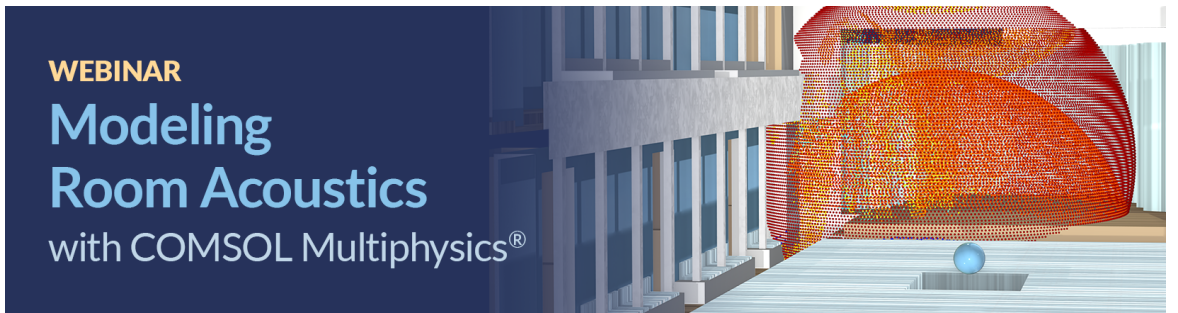


Export Citation

WEBINAR

Modeling Room Acoustics

with COMSOL Multiphysics®



Development of a machine learning detector for North Atlantic humpback whale song

Vincent Kather,^{1,a)}  Fabian Seipel,¹ Benoit Berges,² Genevieve Davis,³ Catherine Gibson,⁴ Matt Harvey,⁵ Lea-Anne Henry,⁶ Andrew Stevenson,⁷ and Denise Risch⁸

¹Audio Communication and Technology, Technical University Berlin, Einsteinufer 17c, 10587, Berlin, Germany

²Wageningen Marine Research, Wageningen University and Research, IJmuiden, Noord Holland, 1976 CP, Netherlands

³National Oceanic and Atmospheric Administration (NOAA) Northeast Fisheries Science Center, 166 Water Street, Woods Hole, Massachusetts 02543, USA

⁴School of Biological Sciences, Queens University Belfast, Belfast, BT9 5DL, Northern Ireland

⁵Google Inc., Mountain View, California 94043, USA

⁶School of GeoSciences, University of Edinburgh, James Hutton Road, EH9 3FE, Edinburgh, Scotland

⁷Whales Bermuda, 6 Overrock Hill, Pembroke, Bermuda

⁸Scottish Association for Marine Science, University of Highlands and Islands, Oban, PA37 1QJ, Scotland

ABSTRACT:

The study of humpback whale song using passive acoustic monitoring devices requires bioacousticians to manually review hours of audio recordings to annotate the signals. To vastly reduce the time of manual annotation through automation, a machine learning model was developed. Convolutional neural networks have made major advances in the previous decade, leading to a wide range of applications, including the detection of frequency modulated vocalizations by cetaceans. A large dataset of over 60 000 audio segments of 4 s length is collected from the North Atlantic and used to fine-tune an existing model for humpback whale song detection in the North Pacific (see Allen, Harvey, Harrell, Jansen, Merkens, Wall, Cattiau, and Oleson (2021). *Front. Mar. Sci.* **8**, 607321). Furthermore, different data augmentation techniques (time-shift, noise augmentation, and masking) are used to artificially increase the variability within the training set. Retraining and augmentation yield F-score values of 0.88 on context window basis and 0.89 on hourly basis with false positive rates of 0.05 on context window basis and 0.01 on hourly basis. If necessary, usage and retraining of the existing model is made convenient by a framework (*AcoDet*, acoustic detector) built during this project. Combining the tools provided by this framework could save researchers hours of manual annotation time and, thus, accelerate their research.

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1121/10.0025275>

(Received 5 July 2023; revised 22 February 2024; accepted 22 February 2024; published online 13 March 2024)

[Editor: Haiqiang Niu]

Pages: 2050–2064

I. INTRODUCTION

The rapid absorption of light in water requires organisms living in the marine environment to act on nonvisual cues. As a result, like many other species, cetaceans rely heavily on sounds for many vital life functions (Gordon and Tyack, 2001). Different species have developed advanced means of sensing and producing sound, enabling them to use sound in every aspect of their lives (Tyack, 1997; Mooney *et al.*, 2012). The different forms of sound production lead to very unique vocalizations, which can be used to differentiate species, populations, and, in certain cases, individual animals (Aide *et al.*, 2013). Scientists make use of this by collecting long-term acoustic data using passive acoustic monitoring (PAM) and classifying different cetacean species based on their vocalizations. Data collected in

this manner can be used to retrieve information on calling behavior, seasonal presence, migration patterns, as well as, under certain circumstances, determine relative or absolute abundance of populations (Parijs *et al.*, 2009; Thomas and Marques, 2012; Lin *et al.*, 2015).

Humpback whales (*Megaptera novaeangliae*) produce vocalizations that can be categorized into *social calls* and *songs*. *Social calls* can be described as “variable through time, interrupted by silent periods, apparently unpredictable, and not showing [...] rhythmic, consistent and continuous temporal pattern[s]” (Saloma *et al.*, 2022). Humpback whale song, on the other hand, is a deeply hierarchical pattern that can be subdivided into *themes*, which are made up of *phrases*, which, in turn, are made up of *units* (Payne and McVay, 1971; Payne, 1983; Cholewiak *et al.*, 2013). A unit is a single vocalization separated from its preceding and succeeding units by an audible pause. Songs are typically 9 to 25 min long and comprised of several hundred units.

^{a)}Email: vkather@gmail.com

Fundamental frequencies of humpback whale vocalizations typically lie under 1 kHz (Au *et al.*, 2006). When humpback whales are in their breeding grounds, *song sessions* (Cholewiak *et al.*, 2013) are common in which several songs are sung in repetition.

Humpback whale song, which is thought to be only sung by males (Clapham, 1996), has been the subject of in-depth studies for over half a century (Payne and McVay, 1971). Current research suggests that the song primarily functions in sexual selection (Cholewiak *et al.*, 2013; Garland and McGregor, 2020; Schulze *et al.*, 2022). A song is commonly sung by an entire population and reproduced by individuals within that population, showing forms of cultural transmission (Garland and McGregor, 2020). Although structurally similar, variations of songs within populations and between individuals do exist (Cholewiak *et al.*, 2013; Garland *et al.*, 2017). On some occasions members of one population copy songs of another (Garland *et al.*, 2011; Garland and McGregor, 2020), thereby bringing one unique song with a distinct pattern of themes, phrases, and units from another population to their own. This copying can spark a *song revolution*: a reoccurring (1.5 to 2 years) fundamental change of the structure of a song, sung within a population, yielding a new song (Allen *et al.*, 2018; Garland and McGregor, 2020). These song revolutions feature significant changes in the pattern of themes, phrases, and units enabling researchers to uniquely identify different songs (Allen *et al.*, 2018; Garland and McGregor, 2020). The complexity and spatiotemporal variation in male humpback whale song has been thoroughly described in several geographic areas (Hawaii, South Pacific, and Caribbean; Payne and McVay, 1971; Ryan *et al.*, 2014; Allen *et al.*, 2018, p. 18; Garland and McGregor, 2020; Wenzel *et al.*, 2020; Narganes Homfeldt *et al.*, 2022; Saloma *et al.*, 2022) but has yet to be researched in many others.

The west coast of Scotland is situated along the migratory route of North Atlantic humpback whales on their path from the warm breeding grounds off the Caribbean and west coast of Africa to their feeding grounds in subpolar and polar waters (Smith *et al.*, 1999; Wenzel *et al.*, 2009). Humpback whales and other baleen whale species, such as minke (*Balaenoptera acutorostrata*) and fin whales (*Balaenoptera physalus*), are sighted in Scottish waters particularly during spring and summer (Weir *et al.*, 2001). Humpback whales passing through Irish and Scottish waters are known to migrate (O'Neil *et al.*, 2019; Berrow *et al.*, 2021) from the two known breeding grounds which are located in the wider Caribbean region and near the Cape Verde Islands (Stevick *et al.*, 2003; Ryan *et al.*, 2014; Wenzel *et al.*, 2020). On their migratory route, they can be heard in many locations throughout the North Atlantic, like the Caribbean, Bermuda, the east coast of the United States (U.S.), Iceland, and Scotland. Understanding the spatiotemporal distribution of humpback whales throughout the year enables the development of more effective conservation measures. This is of particular importance for the Cape Verdian humpback whale population, which has been shown

to be low in numbers with only about 300 animals estimated for this population (Ryan *et al.*, 2014; Wenzel *et al.*, 2020).

PAM is a nonintrusive method to monitor species of vocally active marine mammals over periods long enough to detect meaningful changes in their seasonal distribution (Davis *et al.*, 2020; Todd *et al.*, 2022; White *et al.*, 2022). Long-term acoustic recordings have often been analyzed manually by human experts, which is an accurate method to detect species but is extremely time consuming. In recent years, with advances of computing power, methods to automate species detection and classification processes have, therefore, been developed for many different species (Gillespie *et al.*, 2009; Baumgartner and Mussoline, 2011; Bergler *et al.*, 2019; Bermant *et al.*, 2019; Shiu *et al.*, 2020; Thomas *et al.*, 2020; Zhong *et al.*, 2020; Allen *et al.*, 2021; Garibbo *et al.*, 2021; Kirsebom *et al.*, 2021; Hildebrand *et al.*, 2022). These methods include energy-based or generalized power-law detectors (Helble *et al.*, 2012; Frasier *et al.*, 2017) and are increasingly building on advancements in machine learning. Examples for this advancement are the recent successful applications of convolutional neural networks (CNNs) to detect marine mammals (Bergler *et al.*, 2019; Bermant *et al.*, 2019; Shiu *et al.*, 2020; Thomas *et al.*, 2020, p. 220; Zhong *et al.*, 2020; Allen *et al.*, 2021; Garibbo *et al.*, 2021; Kirsebom *et al.*, 2021). Whereas conventional detection algorithms rely on predefined parameters to detect vocalizations (for example, based on an energy threshold in a frequency band), deep neural networks models, a subset of machine learning models, have the capability to learn features (and map them to an output/classification) autonomously from raw data. A CNN is a deep neural network with millions of parameters that get *tuned* in the training process. This process requires large amounts of training data and computing power but if successful, yields a model which is capable of distinguishing specific vocalizations from other sounds in challenging noise environments.

In their study, Allen *et al.* (2021) developed a CNN model, which was trained on humpback whale song data from the North Pacific, reaching promising performance metrics of 97% average precision on the level of 3.9 s long spectrogram images over nine different deployment locations. When the National Oceanic and Atmospheric Administration (NOAA)/Google model was applied to humpback whale song recordings from the North Atlantic, average precision dropped to 79%. This loss in performance is most likely a result of different noise environments and distinct humpback whale song unit structures that vary across the two ocean basins.

To produce a well performing model for the North Atlantic humpback whale population, this study describes the fine tuning of the NOAA/Google model on a dataset from the North Atlantic. The dataset needed a large variation in location sites as well as recording dates to be able to encapsulate different noise environments and different humpback whale songs (due to song revolutions). Augmentation techniques (time-shift, noise augmentation, and masking) were employed during training to artificially

increase the variability of the training data. Finally, a framework for inference, training, and evaluation of machine learning models, *AcoDet* (acoustic detector) was built, which was used for all trainings and evaluations during this project. *AcoDet* is available for use online,¹ allowing researchers to use the model on their own datasets. The framework features a graphical user interface and outputs Raven annotation tables (Cornell Laboratory of Ornithology, 2014), allowing it to be integrated into existing workflows. *AcoDet* also allows researchers to train and evaluate models themselves. This enables them to apply it on a dataset of their own or apply the model to a different species. The main contributions of this study can be summarized in the two following aspects:

- (1) Provide a model for automated humpback whale song detection in the North Atlantic; and
- (2) present a (openly available) framework for inference, training, and evaluation of machine learning models for marine mammal vocalizations.

II. MATERIALS AND METHODS

To illustrate the unit level of a humpback whale song, Fig. 1 displays the hierarchical structure of a song. To train a CNN for humpback whale vocalizations [Fig. 1(D)], Allen *et al.* (2021) defined a fixed length to use for training. This length, the *context window length*, corresponds to a 3.9124-s long spectrogram image. A *context window* of 3.9124-s length is large enough to contain an entire unit but not large enough to contain two units (on average). In this study, a single manual annotation refers to the annotation of a single context window. When referring to predictions on context windows, what is implied is a generation of annotations for every context window. As can be observed in Fig. 1, the time scale changes for each of the hierarchical levels: starting with a 10-min excerpt of an entire song [Fig. 1(A)] and finishing with a single unit depicted in a context window of 3.9124 s length [Fig. 1(D)].

A. Data acquisition

The data used in this study were chosen to span a large geographic area encompassing many of the known humpback whale core habitats in the North Atlantic. In theory, humpback whales from Caribbean or Cape Verde populations would be recorded in different parts of the ocean basin, thereby having similar vocalizations but large variations in noise environments. The majority of the chosen data was selected from the month of March because in this month, North Atlantic humpback whales migrate to their northern feeding grounds (Vu *et al.*, 2012) and are likely to be heard in all sites included in the dataset. Furthermore, restrictions in storage capacity and transfer speeds of large acoustic datasets limited the size of the dataset in this study.

All data were recorded using stationary PAM devices. In marine environments, PAM devices typically feature underwater hydrophones that record for several months to

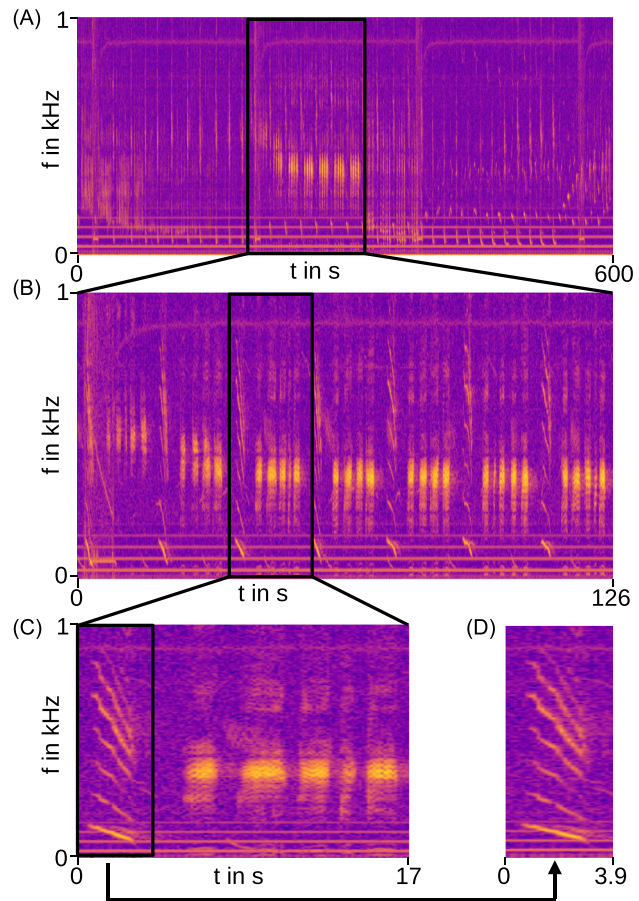


FIG. 1. (Color online) Visualization of humpback whale song structure in spectrograms. Each labeled spectrogram shows a different time scale, where (A) shows a 10 min excerpt from a humpback whale song. The highlighted portion of (A) contains a theme, a repeating subsection of a song, which is shown in spectrogram (B), where a phrase is highlighted, which is a repeating subsection of a theme. In (C), the highlighted portion shows a unit, which is the smallest structural element, and displayed in spectrogram (D).

several years (depending on recording site) either continuously or duty cycled (i.e., recording at defined intervals). All of the data used for training of the final model originate from 24 different deployments scattered over more than 15 locations, categorized into 4 regions: Scotland, Caribbean, the east coast of the United States (U.S.) and Bermuda, and Iceland. The deployments span a period of 15 years from 2005 to 2020. To ensure that the detector is not limited in its application to one recording setup, different recording devices were included in the dataset. The specifications of each setup can be found in the supplementary material, Table S1) and respective publications (Davis *et al.*, 2020; Narganes Homfeldt *et al.*, 2022; van Geel *et al.*, 2022; COMPASS, 2023).

Figure 2 shows the location of each respective deployment along with the amount of data linked to each deployment and the respective call to noise ratio. A call (positive) refers to a context window containing a humpback whale song unit (or part of one), whereas noise (negative) refers to a context window containing anything else other than a humpback whale song unit (can also include vocalizations from other cetaceans).

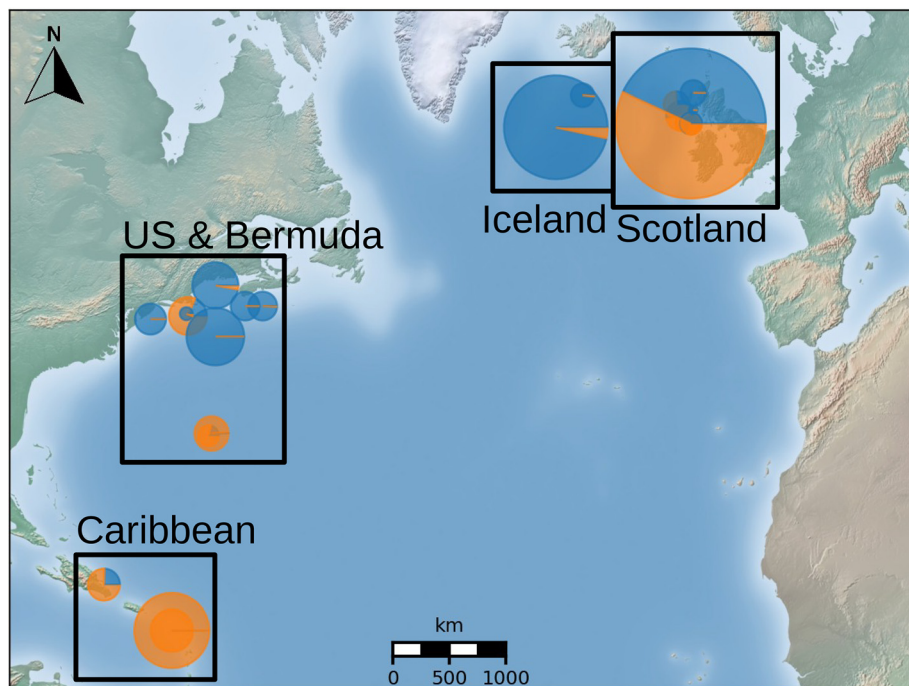


FIG. 2. (Color online) Map showing the North Atlantic Ocean alongside the location of annotated datasets. Sizes of circles correspond to dataset size. Blue corresponds to percentage of noise, and orange corresponds to percentage of calls. Refer to Secs. II C 1 and II C 2 for more information on the datasets.

B. Preprocessing

To create a machine learning model that is trained on data with varying recording conditions and settings, a standardized data preprocessing is necessary. Data preprocessing steps similar to those employed by Allen *et al.* (2021) were implemented, however, several parameters were adjusted. All final datasets were built with a sampling rate of 2 kHz, unlike the original 10 kHz (as used for the NOAA/Google model), as reducing the sampling rate from 10 to 2 kHz showed no significant loss in performance. Moreover, the lower sampling rate excluded high frequency noise and reduced the file sizes, allowing for faster processing. The reduced sampling rate results in a reduction of the context window length from the original 3.9124 s (39 124 samples at 10 kHz) to 3.8775 s (7755 samples at 2 kHz). Reducing the sampling rate leads to a change in context window length time as a result of rounding to the nearest number of time bins.

The model is run with nonoverlapping context windows to prevent detecting the same unit multiple times. However, if long units are split in a way in which there is enough in each of the context windows for the model to detect it as such, the unit would be counted twice. Table I shows all parameters relevant for data preprocessing.

The data preprocessing can be grouped into four steps:

- (1) Audio recording files are resampled to the model sampling rate. Unlike the original NOAA/Google model sampling rate of 10 kHz, a sampling rate of 2 kHz was used;
- (2) a standard short-time Fourier transform (STFT) with a Hann window of 1024 samples length and an output size fixed to 128×64 (time \times frequency) bins is applied. This step is identical to the preprocessing by Allen *et al.* (2021);

- (3) the matrices from step (2) are transformed into (Mel-) spectrogram images [STFT magnitudes are binned along the frequency axis using a triangular Mel filter bank (64 Mel filters) and then squared after binning]. Due to the change in sampling rate to 2 kHz, the 64 frequency bins are mapped onto the frequency range between 0 and 1 kHz; and
- (4) the spectrogram images are normalized using a per channel energy normalization (PCEN; Wang *et al.*, 2017). This, too, is analogous to the preprocessing in the NOAA/Google model.

For a more detailed description of preprocessing steps in the paper by Allen *et al.* (2021), please refer to the section therein entitled, “Acoustic Front End.”

C. Training progress

1. Training phase 1

Initially, a total of 22.7 h comprised from the COMPASS (2023) and the SAMOSAS (van Geel *et al.*, 2022) datasets (all sites located in Scottish waters) were

TABLE I. Settings used in preprocessing to generate training data and run model.

Name of setting	Value
Sample rate	2000 Hz
Context window length	3.8775 s (7755 samples)
Fast Fourier transform (FFT) size	1024 samples (512 ms)
FFT hop size	53 samples (26.5 ms)
Context window hop size	7755 samples (3.8775 s)
Spectrogram resolution	128×64 (time \times frequency)
Normalization	Per channel energy normalization (PCEN)

annotated (on context windows level) by an experienced bioacoustic analyst. The annotation protocol demanded only humpback whale song units to be manually annotated. After manual annotation was complete, positive context windows were generated by extracting 3.8775 s of audio starting at the onset of the manual annotation. A time-shift augmentation was later introduced to ensure variability of the unit placement within the context window. Negatives (implicit) were generated by using the audio between manual annotations. This data were used to fine-tune the pretrained NOAA/Google model (modified ResNet-50; Allen *et al.*, 2021), yielding a preliminary North Atlantic model (see training phase 1 in Fig. 3).

2. Training phase 2

The preliminary North Atlantic model was used to generate more annotations on the other available datasets across the North Atlantic and for hard negative mining (Sung and Poggio, 1995; Fig. 3, training phase 2). From the generated annotations, files with a high number of predictions (more than 100 annotations with a value > 0.5) were selected. From the generated annotations, missed positives were marked as positives, incorrectly annotated positives were

marked as explicit negatives, and the gaps between positives and explicit negatives were marked as implicit negatives. Using this strategy, another 37.2 h of annotated data (context windows) were generated. Following the generation of the dataset, augmentation techniques were introduced (detailed description in Sec. IID). Using the full dataset and the augmentations, the pretrained NOAA/Google model was fine-tuned to yield the final North Atlantic model.

3. Data splitting

Data were randomly divided into train, validation, and test sets. When creating train, validation, and test sets, context windows in the train and validation sets originated from the same files to ensure the same variability. For the test set, all of the context windows were unseen by the model, but the majority of the samples were extracted from files that had been used to previously extract train and validation data. To better test generalization capabilities of the model, a small portion of the context windows in the test set were generated from files that the model had not observed during training.

Table II shows the specific number of examples for train, validation, and test sets. As a result of the large availability data, Scotland is overrepresented in this study. Except for the unseen test files, data from Scotland are balanced, meaning that there are approximately equal numbers of positives and negatives in train, validation, and test sets. Data from the Caribbean, on the other hand, feature an imbalance toward positives due to the proximity of the recorders and the humpback whale breeding grounds. Data from the east coast of the U.S. and Bermuda, as well as Iceland, feature an imbalance toward the negative as the majority of the randomly selected files were predominantly noise. Because the test (unseen) files were also selected at random in Table II, the column “test (unseen) call” shows no positives for the Caribbean, U.S. and Bermuda, and Iceland. Scotland, on the other hand, yields only a single noise sample as the selected files were mainly filled with humpback whale song. The bottom two rows show the number of samples for the hourly and daily level datasets. No train and validation sets exist for hourly and daily based annotations. Aggregation metrics (see Sec. IIG) are used to produce hourly and daily level annotations; therefore, the model is evaluated but not trained on it.

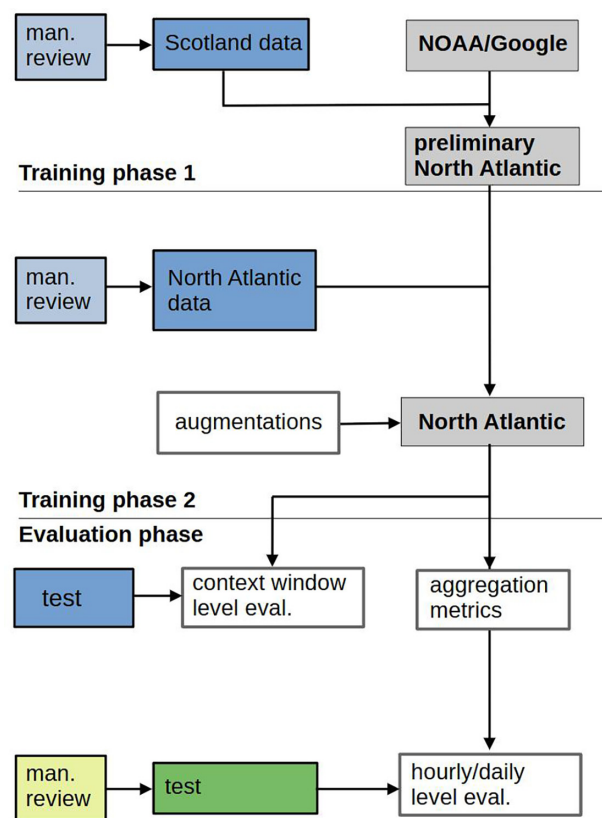


FIG. 3. (Color online) Flowchart showing usage of datasets and models throughout the project phases. Blue and green represent the context window-based and hourly/daily based datasets, respectively. Where applicable, the manual review effort (corresponding to the respective dataset) is displayed in a lighter shade of the same colour. Model versions are shown in gray. Refer to Secs. IIC1 and IIC2 for more information on the datasets.

4. Evaluation phase

After the training was completed in the evaluation phase (Fig. 3, bottom), the test set was used to evaluate the performance, compare different model versions and identify the strongest performing one (Fig. 3, bottom). To evaluate the model’s performance on the time scale of hours and days, test sets on hourly and daily basis (Fig. 3, green) were created. These datasets, which were (manually) annotated on an hourly and daily presence basis, were created for three sites off of the west coast of Scotland (SAMOSAS S1, N1, and EL1; see supplementary material, Table S1).

TABLE II. Number of samples [context windows (c), hours (h), or days (d)] in train, validation, and test sets with respective origin. Call refers to a sample containing humpback whale song units, whereas noise refers to a sample that does not contain humpback whale song units. The test set is split into samples that are extracted from files that have been observed during training (seen) and ones that have not (unseen). The total column shows the number of context windows summarized for all regions, as well as the number of hours.

Origin	Train call	Train noise	Validation call	Validation noise	Test (seen) call	Test (seen) noise	Test (unseen) call	Test (unseen) noise
Scotland (c)	7535	6588	2263	1884	3223	2658	208	1
Caribbean (c)	5141	188	1546	57	2194	77	0	42
US/Bermuda (c)	656	5994	200	1804	284	2553	0	1655
Iceland (c)	200	5789	56	1737	81	2480	0	0
Total (c)/in h	13 532/14.6	18 559/20	4065/4.4	5482/5.9	5782/6.2	7768/8.4	208/0.2	1698/1.8
Scotland (h)	—	—	—	—	—	—	1358	1954
Scotland (d)	—	—	—	—	—	—	30	108

D. Data augmentation

In machine learning, using complex models (like CNNs) to solve complex tasks always poses the risk of the model *memorizing* the training data instead of learning its patterns (*generalizing* from the training data). This issue is known as *overfitting*, “the phenomenon when a network learns a function with very high variance such as to perfectly model the training data” (Shorten and Khoshgoftaar, 2019). One way to reduce the risk of *overfitting* is the use of *data augmentation* techniques. Through data augmentation, copies of the training data are created, which are then artificially altered.

In marine environments, recorders in different conditions can be subject to a multitude of sounds, e.g., the sounds of boat motors, sonar, or sounds produced by the recording equipment, such as noise from the spin-up of magnetic hard drive platters. Distinguishing calls and noise in spectrograms is only successful if sufficient data with examples are provided for the model. Augmentation of training data increases the amount of data by creating copies containing alterations. In this study, three different types of data augmentation techniques were employed, all of which have been applied to machine learning in bioacoustics in the past: time-shift (Pandeya *et al.*, 2018), noise insertion (Pandeya *et al.*, 2018), and frequency and time masking (Anderson and Harte, 2021). Augmentations were implemented *on-the-fly*, meaning that they were computed continuously during the training process, thus, yielding copies with new alterations after every training iteration (epoch). In the following, the three data augmentation techniques that were employed in this study are presented.

1. Time-shift augmentation

Time-shift augmentation was used to reduce the risk of the model only detecting calls when the unit onset coincides with the beginning of the context window. The spectrogram training data were time shifted by selecting a random number of time bins within the first half of the spectrogram and reordering the image. The matrix in Eq. (1) shows a spectrogram image with dimensions $n \times m$ (frequency \times time):

$$x = \begin{bmatrix} x_{0,0} & x_{0,1} & \cdots & x_{0,m} \\ x_{1,0} & x_{1,1} & \cdots & x_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,m} \end{bmatrix}. \tag{1}$$

A value $\lambda \in [0, m/2]$ is randomly chosen, after which the spectrogram image is reordered such that values in the λ th and all succeeding columns are moved to the beginning, and all remaining columns are used to fill up the matrix. Equation (2) shows the reordered (time-shifted) matrix, $\tilde{x}_{\{ts\}}$:

$$\tilde{x}_{\{ts\}} = \begin{bmatrix} x_{\{0,\lambda\}} & x_{\{0,\lambda+1\}} & \cdots & x_{\{0,m\}} & x_{\{0,0\}} & \cdots & x_{\{0,\lambda-1\}} \\ x_{\{1,\lambda\}} & x_{\{1,\lambda+1\}} & \cdots & x_{\{1,m\}} & x_{\{1,0\}} & \cdots & x_{\{1,\lambda-1\}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{\{n,\lambda\}} & x_{\{n,\lambda+1\}} & \cdots & x_{\{n,m\}} & x_{\{n,0\}} & \cdots & x_{\{n,\lambda-1\}} \end{bmatrix}. \tag{2}$$

The first column is now the λ th column with all succeeding columns up to the m th column following it. The column after the m th column is now the first column of the original matrix, x , with all succeeding columns up to the $(\lambda - 1)$ th column following it. This reordering of the spectrogram image creates an artificially delayed onset of the humpback whale song unit.

The time-shift augmentation can be observed in the first row of Fig. 4. The visible down-sweep in the original spectrogram image on the left side is delayed in the augmented spectrogram image on the right.

2. Noise augmentation

As noise environments of different recording sites tend to vary a lot, noise augmentation provides a possibility to combine units recorded in one noise environment with those recorded in other noise environments. Noise is inserted by combining two spectrogram images into a new artificially created image:

$$\tilde{x}_{nu} = \alpha y + (1 - \alpha)x. \tag{3}$$

Equation (3) describes the formula by which two spectrogram images are combined, where y is a noise spectrogram,

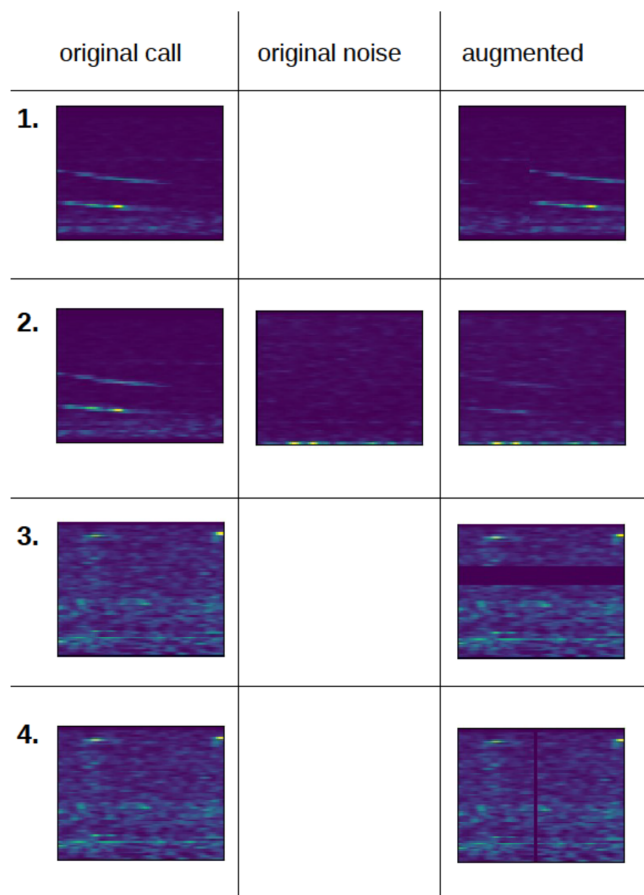


FIG. 4. (Color online) Different augmentation techniques employed in this study. Row 1 shows time-shift augmentation, row 2 shows noise augmentation, row 3 shows frequency masking (part of SpecAugment), and row 4 shows time masking (part of SpecAugment). For each row, the columns display the original spectrogram and their augmented counterpart.

x is a call spectrogram, \tilde{x}_{mu} is a resulting noise augmented spectrogram, and $\alpha \in [0,1]$ is the parameter determining the ratio between call and noise spectrogram amplitude.

As described in Sec. IID 2, the effectiveness of noise augmentation was increased by only using implicit noise for y [Eq. (3)] and combining them with explicit noise and calls. Noise augmentation was, thus, used to either combine humpback whale song units or explicit noise of one noise environment with implicit noise of another noise environment.

An example of noise augmentation can be observed in the second row in Fig. 4. The spectrogram image on the left (call) is combined with the spectrogram image in the middle (noise), yielding the augmented spectrogram image on the right, which exhibits the slightly weaker down-sweep unit overlaid with low frequency noise.

3. SpecAugment

The final augmentation technique originates from machine learning image recognition tasks and is based on the principle of masking. For this study, time and frequency masking were used to train recognition of units if parts of the units are masked and, thus, not visible. These masking algorithms, summarized under the name SpecAugment,

have proven to be effective in image recognition tasks (Park *et al.*, 2019). The time warping augmentation employed by Park *et al.* (2019) was not used in this study. Figure 4 shows examples of frequency masking in the third row and time masking in the fourth row. On the left side, the original training spectrogram image can be observed and, on the right, the masked spectrogram images show a dark blue bar of a randomly chosen width (for time masking, maximum of 10 bins) or height (for frequency masking, maximum of 10 bins) in a randomly chosen location.

Introducing augmentation techniques to the model reduces the risk of overfitting by increasing the amount and variability of the data. Further variability is added as all of the augmentation techniques listed above are constantly recomputed throughout the training session (on-the-fly). Each augmentation technique employed in the training was used on the entire dataset. In summary, the augmentations used in this study lead to a fivefold increase in the dataset size. Like the training data, all augmented spectrograms are subject to PCEN.

E. Model architecture and training parameters

The architecture of the CNN model used is based on the same modified ResNet-50 architecture as that for the NOAA/Google model’s architecture. The modified ResNet-50 diverges from the default ResNet-50 (He *et al.*, 2015) by having a smaller input spectrogram and, therefore, a stride of 1 instead of 2 in the first convolutional layer. The NOAA/Google model’s architecture amounts to 2.3×10^7 parameters.

The training of the CNN model was performed on a graphical processing unit computer using a NVIDIA RTX 3060 TI (Santa Clara, CA). During the training process, a binary cross-entropy loss function and an Adam optimizer were used. The final model converged after 43 epochs, each run for 1000 steps with a batch size of 32. The combination of batch size, number of epochs, and number of steps per epoch was determined empirically after running numerous training runs. The train set, including augmentations, consisted of approximately 150 000 samples. An exponentially decaying learning rate with an initial learning rate of 4×10^{-4} and a final learning rate of 3×10^{-6} were chosen. Several models were trained, however, for the results presented in this study, a single model was evaluated.

F. Comparison by region

To investigate differences in the model’s performance based on different noise environments and representation within the training set, the model’s performance was evaluated by region. To do so, the model was trained on the entire train set and subsequently evaluated with the test set corresponding to one region.

G. Aggregation metrics

Model performance was evaluated on different time scales: context window, hourly, and daily. Validation on

hourly basis was performed using the hourly based test set, which consists of manually annotated hourly presence annotations of each hour for 46 consecutive days and 3 different sites. This amounts to 3312 samples (24 h for 46 days and 3 sites; see Table II). Additionally, validation on daily basis was performed using the daily based test set, which consists of daily presence for 46 days and each of the sites, amounting to 138 samples. The hourly based test set contains binary values for humpback whale song presence in a given hour, and correspondingly, the same applies to the daily based dataset. To automatically produce the same format of binary presence-based annotations, two different aggregation metrics were compared: *simple limit* and *sequence limit*.

1. Simple limit

When used for inference, the North Atlantic model outputs a prediction value, $p \in [0,1]$, for each context window (3.8775 s). Once complete, a threshold is applied to the predictions, discarding all predictions with values under the threshold. If the number of remaining predictions in an hour of data does not exceed a predefined limit, the hour is marked as 0, which stands for no humpback whale song present. If the number of remaining predictions in an hour of data is exceeded, it is marked with 1, which stands for humpback whale song present.

To provide an example use case of the simple limit, consider a file with 120 context windows, each of which contains a prediction value. If only 12 of the 120 context windows contain values above 0.9, and the simple limit is applied with a threshold of 0.9 and a limit of 10, then the criterion would be met, and for the given hour, the metric would return the value 1, meaning humpback whale song is present in the given hour.

2. Sequence limit

Unlike the simple limit, the sequence limit takes into account that humpback whale songs last at least 9 min and, thus, contain at least about 130 units, which are sung in sequence. After applying a threshold to the predictions to discard lower value predictions, a predefined limit is used and needs to be exceeded in a string of N consecutive context windows in an hour. If this limit is exceeded at some point within the hour, the hour is marked with 1 for humpback whale song present. If not, it is marked with 0 for no humpback whale song present.

Considering the above example with 120 context windows, to compute the sequence limit, the order of the prediction values is relevant. If the 12 prediction values exceeding 0.9 are evenly distributed throughout the 120 context windows, we can assume to encounter one prediction value above 0.9 every 10 context windows. If N is 20 and we apply a threshold of 0.9 and a limit of 5, the criterion will not be met as we will never receive 5 predictions exceeding 0.9 in a consecutive string of 20 context windows.

Regardless of which aggregation metric is used for hourly presence, daily presence is acquired by checking and

compiling hourly presence. If for any of the 24 h, a value of 1 is returned, the daily presence also returns a value of 1. The aggregation metrics allow users to use the model to provide automatically generated annotations on hourly and daily levels.

III. RESULTS

A. Data augmentation

To compare augmentations, different instances of models [all modified ResNet-50 architecture (NOAA/Google)] were trained using different combinations of augmentations. All models were trained on the training set and evaluated using the *test set (unseen)*. By choosing the *test set (unseen)*, the ResNet-50 models were confronted with data from files that had not been included in the training. While using the entire test set would have provided a larger dataset for evaluation, the smaller *test set (unseen)* dataset allowed us to focus on the different model’s ability to generalize to new data. Table III shows model performance depending on the data augmentation techniques used. The first three columns indicate what augmentations were used in training. Average precision [area under the curve precision and recall (AUC-PR)] shows a threshold independent metric, indicating performance at different operating conditions. The final three columns show precision, recall, and F-score (harmonic mean of precision and recall) at a threshold of 0.5, thereby allowing for a comparison at a specific operating condition. The values of precision, recall, F-score, and false positive rate (FPR) are defined as specified in Eqs. (4)–(7):

$$precision = \frac{true\ positives}{true\ positives + false\ positives}, \tag{4}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}, \tag{5}$$

$$F\text{-score} = 2 \frac{precision \cdot recall}{precision + recall}, \tag{6}$$

$$FPR = \frac{false\ positives}{false\ positives + true\ negatives}. \tag{7}$$

TABLE III. Performance metrics of North Atlantic model by augmentation. The first three columns show values of 0 (augmentation not used) and 1 (augmentation used). The augmentations are time-shift (TS), noise augmentation (NA), and SpecAugment (SA), and applied on the train set. Values are displayed for area under the curve precision and recall (AUC-PR). $P_{0.5}$, $R_{0.5}$, and $F_{0.5}$ show precision, recall, and F-score values for a threshold of 0.5.

TS	NA	SA	AUC-PR	$P_{0.5}$	$R_{0.5}$	$F_{0.5}$
0	0	0	0.67	0.68	0.85	0.76
1	0	0	0.84	0.77	0.9	0.83
0	1	0	0.72	0.63	0.78	0.7
0	0	1	0.78	0.75	0.82	0.79
1	1	0	0.89	0.77	0.85	0.8
1	0	1	0.84	0.77	0.88	0.82
0	1	1	0.77	0.66	0.85	0.74
1	1	1	0.87	0.8	0.87	0.84

The first case, employing no augmentation, yields the worst performance for average precision, thereby justifying the usage of augmentations. Of all the different models, the two models yielding the best results achieve average precision values above 0.85 are the ones using all augmentations (last row) and all except for SpecAugment (row number five). When examining the performance by precision and recall, the model employing all augmentations is slightly better. This is more obvious when referring to the F-score, which is the highest for the model employing all augmentations. Interestingly, all trainings, including the time-shift augmentation, reach F-score values of 0.8 and above. The results highlight that the combination of all augmentation techniques forces the model to generalize from the training set rather than being able to memorize it (less prone to overfit). All subsequent trainings were, thus, executed with a pipeline containing all three augmentation techniques: time-shift, noise augmentation, and SpecAugment.

B. Comparison by region

Table IV shows the model performance for different regions. For this performance, the test sets of the four regions, Scotland, Caribbean, the east coast of the U.S. and Bermuda, and Iceland have been used as well as their combination for the overall North Atlantic performance.

The threshold independent metric average precision (AUC-PR) is the highest for the Caribbean (0.99); however, the North Atlantic (overall), Scotland, and the U.S. and Bermuda also reach high values exceeding 0.93. When comparing the precision and recall as well as the F-score at a threshold of 0.5, the Caribbean reaches the highest values with a precision of 0.98, a recall of 0.89, and an F-score of 0.93. When considering the precision metric for the Caribbean, Table II reveals that this region has an imbalanced test set in which the majority of the data are calls. Precision, recall, and F-score values are also displayed for a threshold of 0.9. At this operating condition, precision values of all regions increase reaching values of 0.93 and above while recall values decrease to 0.86 and below. The F-score values at a threshold of 0.9, ranging from 0.86 to 0.89, are lower than those at a threshold of 0.5, ranging from 0.86 to 0.93. The performance on the North Atlantic mirrors this, reaching an F-score of 0.91 at a threshold of 0.5 and a lower F-score value of 0.88 at a threshold of 0.9. Conversely, the values of the FPR reduce when increasing the threshold.

Although the values of the regions range from 0 to 0.31 at a threshold of 0.5, they reduce to a range of 0 to 0.12 at a threshold of 0.9. For the overall performance, the FPR at 0.5 reaches a value of 0.1, whereas it reduces to a value of 0.05 at a threshold of 0.9.

The precision and recall curve in Fig. 5 shows the model performance for the four different regions and overall. Ideally, a precision and recall curve goes through the point (1,1), signifying 100% precision and 100% recall. Recall values are depicted from 0.7 to 1 because the minimum recall for any threshold was 0.7. For low thresholds, only the Caribbean curve shows high precision values >0.9 while the other curves reach precision values of 0.6 and lower. At threshold values of 0.5, all curves reach precision values of 0.8 and higher (also shown in Table IV). The Iceland curve reaches a recall value of 0.8 while the other curves reach recall values of 0.88 and higher. When the threshold is further increased, precision values rise while recall quickly declines. At a threshold of 0.9, all curves reach precision values of 0.9 and higher with corresponding recall values of 0.77 and higher. The overall performance at a threshold of 0.9 reaches a recall value of 0.82 and a precision of 0.96 (see Table IV).

C. Aggregation metrics

Figure 6 shows F-score values of the sequence limit for hourly and daily presence as a function of the limit and threshold used. The ranges of F-score values differ between hourly and daily presence due to the improved results on daily presence. Performance on the three sites included in the *test set (unseen)* on hourly and daily levels were averaged for this visualization. The graphs show a very regular pattern for hourly presence [Fig. 6(B)]. For a threshold of 0.95, the maximum F-score is 0.93 (limit of 2). For lower thresholds, the F-score values decrease, reaching a maximum of 0.87 for a threshold of 0.7 (limit of 9). For daily presence [Fig. 6(A)], the highest F-score value of 0.965 is reached for a threshold of 0.85 and a limit of 6. For a threshold of 0.95, the curve reaches a maximum F-score value of 0.96 at a limit of 4. For a lower threshold value of 0.7, the maximum F-score value is 0.925 (limit of 9). When operating daily and hourly presence at the same conditions, the threshold of 0.9 with a limit of 4 yields the best combination of F-scores, 0.93 for hourly presence and 0.945 for daily presence (the same procedure was repeated for the simple

TABLE IV. Model performances for the four different regions represented in the validation datasets (Table I) and all regions combined [overall North Atlantic (NA)] are compared. AUC-PR values are used for comparison as well as precision, recall, F-score, and FPR values at fixed thresholds of 0.5 and 0.9.

Region	AUC-PR	$P_{0.5}$	$R_{0.5}$	$F_{0.5}$	$FPR_{0.5}$	$P_{0.9}$	$R_{0.9}$	$F_{0.9}$	$FPR_{0.9}$
Scotland	0.95	0.9	0.89	0.89	0.09	0.95	0.82	0.88	0.06
Caribbean	0.99	0.98	0.89	0.93	0.31	0.99	0.81	0.89	0.12
U.S. and Bermuda	0.93	0.81	0.92	0.86	0.01	0.93	0.86	0.89	< 0.01
Iceland	0.87	0.92	0.80	0.86	< 0.01	0.98	0.77	0.86	< 0.01
NA (overall)	0.96	0.92	0.89	0.91	0.1	0.96	0.82	0.88	0.05

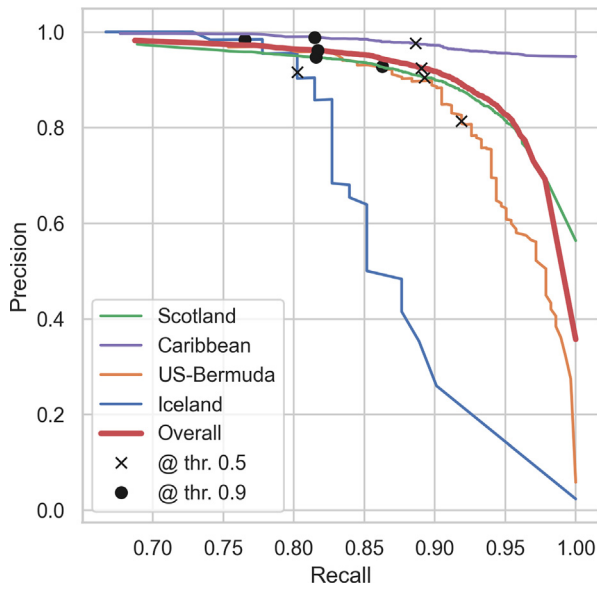


FIG. 5. (Color online) Precision and recall curve displaying the model performance (context window basis) by region and overall. The curves are generated by incrementally varying the threshold and calculating performance for every threshold value. Specific precision/recall values are highlighted by an “x” for performance at a fixed threshold of 0.5 and a circle at a fixed threshold of 0.9.

limit). Although this yields the best performance with regard to the F-score, taking into account the FPR, limits with slightly higher values were chosen to ensure more stability against false positives. For this reason, a limit of 6 was chosen for the sequence limit and a value of 18 was chosen for the simple limit.

Table V shows model performance by different time scales. Context window level performance values are compared with hourly and daily performances. Performance by F-score increases from context window to hourly to daily basis. Among the hourly presence results, the sequence limit (SQ) yields a slightly lower F-score with 0.89 compared to 0.9 by the simple limit (SL). Both reach a precision of 0.99. However, the FPR of the sequence limit is slightly lower with 0.01 vs 0.02 by the simple limit. The sequence limit is run with values of $N = 20$ and a limit of 6, whereas the simple limit is run with a limit of 18. These values were empirically determined (see below).

For daily presence, the difference between sequence limit and simple limit is very small. The most prominent difference is the FPR reaching a value of 0.14 for the simple limit on daily level and a value of 0.07 for the sequence limit. Recall values climb from context window level to hourly and daily starting at 0.82 and reaching 0.98. The precision values also increase from context window level with a value of 0.96 to hourly with a value of 0.99. However, for daily presence, the value drops again to 0.98 for sequence limit. F-score values climb as well, with all of the values reaching at least 0.88.

Figure 7 shows the output of the North Atlantic model using the framework *AcoDet* if hourly counts [Fig. 7(A)] and hourly presence [Fig. 7(B)] are generated. Date and time

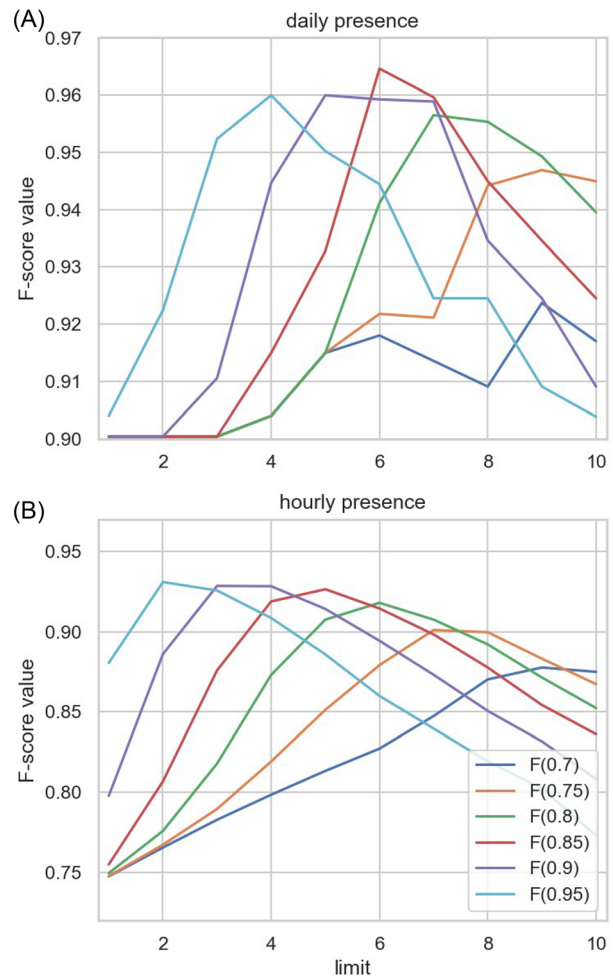


FIG. 6. (Color online) Changes in F-score as a result of changes in limit and threshold of sequence limit for hourly (A) and daily (B) presence. Variation in threshold is shown by colours over variation of limit values on x axis.

information are automatically extracted from the timestamps in the dataset and used to create visualizations that can be used to investigate diel activity patterns of whales. Both plots show a high number of vocalizations in the first three weeks of March 2021 at the recording site. Vocalizations are

TABLE V. Performance values for different time scales are compared. The time scales displayed are context window, hourly, and daily basis. For each time scale, precision (P), recall (R), F-score (F), and FPR values are shown along with the threshold that they are operated at. For hourly and daily presence, two different accumulation metrics, simple limit (SL) and sequence limit (SQ), are compared. For the aggregation metrics, the respective limits (number of predictions exceeding the threshold) are shown. For the sequence limit, the number of consecutive context windows evaluated (N) is displayed.

Basis of prediction	Metric	Limit	$P_{0.9}$	$R_{0.9}$	$F_{0.9}$	$FPR_{0.9}$
Context window basis	—	—	0.96	0.82	0.88	0.05
hourly presence	SL	18	0.99	0.82	0.9	0.02
Hourly presence	SQ ($N = 20$)	6	0.99	0.81	0.89	0.01
Daily presence	SL	18	0.97	0.95	0.96	0.14
Daily presence	SQ ($N = 20$)	6	0.98	0.94	0.96	0.07

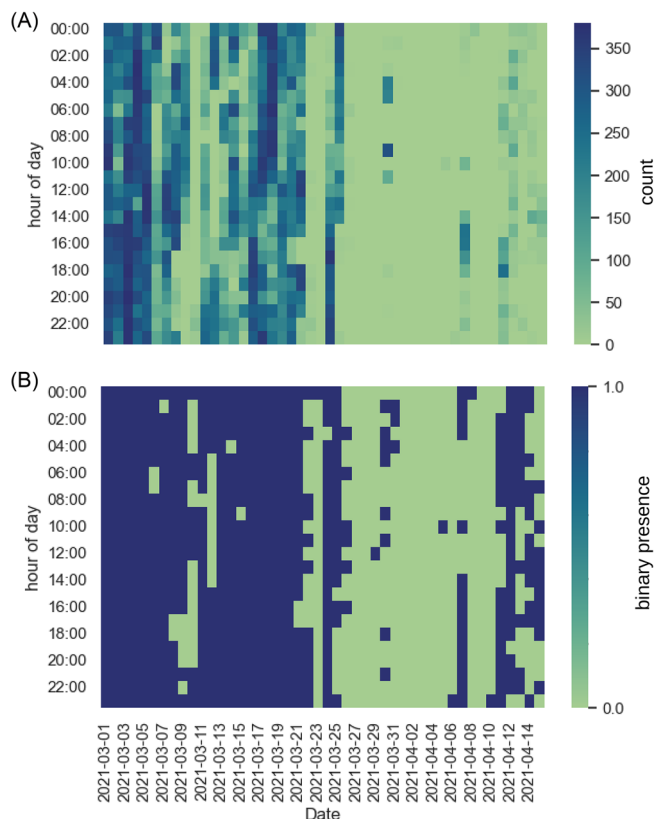


FIG. 7. (Color online) Example output of hourly counts and hourly presence for SAMOSAS site (EL1, located at 57.09847, -8.96888). The hours of day are depicted on the y axis and the recording dates appear on the x axis. For every hour and day, the framework (*AcoDet*) outputs the counted predictions (A) exceeding the threshold of 0.9 and limit of 6 (within sequence of $N=20$ context windows) in a given hour. Alongside the counted predictions, the framework also outputs the hourly presence (B), derived from the hourly counts (A) by applying the sequence limit and yielding a value of 1 if threshold and limit are exceeded.

irregular in the last week of March and decrease in April 2021. The hourly presence [Fig. 7(B)], which is a binary metric, shows that in the second week of April 2021, humpback whale song vocalizations increase again.

IV. DISCUSSION

In this study, an existing CNN model that was originally trained on the North Pacific for humpback whale song detection (Allen *et al.*, 2021) was successfully adapted for use in the North Atlantic. Using data provided by researchers from across the North Atlantic, a large representative dataset was created and used to fine-tune the existing NOAA/Google model, yielding the North Atlantic model. In the process of developing the new model, a user-friendly framework, *AcoDet*, was built, allowing researchers to apply the model to their own datasets on context window level as well as on hourly presence and daily presence level.

A. Preprocessing and augmentation

The reduced sampling rate of 2 kHz showed to be no hindrance in achieving satisfactory model results. Reducing

the sampling rate from 10 to 2 kHz has little effect on model performance, most likely because humpback whales tend to vocalize in lower frequency ranges (humpback whale song units rarely have fundamental frequencies that exceed 1 kHz; Au *et al.*, 2006). Down-sampling recordings of humpback whale song can preserve the crucial frequency components as long as the sampling rate is not below 2 kHz.

A key component of creating a detector capable of generalizing to new environments is a large variation in environment-specific data. Model performance will usually be improved if environment-specific data exist for the environment that a model is applied to. If only small amounts of such data are available, the usage of different data augmentation techniques can increase the variability within the dataset. Three established data augmentation techniques (time-shift, noise augmentation, and masking) were used and compared in this study. Because the majority of the data is produced in training phase 2 (see Sec. II C 2) through revision of automatically generated annotations, the context windows will feature units placed in different positions within the spectrogram as well as being cut off. This is likely the cause for the positive impact of the time-shift augmentation on the performance by F-score. Although noise augmentation and SpecAugment do not show such a clear pattern, when judging by precision and F-score, the combination of all three augmentation techniques during training proved to be the most effective in generalizing to unseen noise environments.

B. Model performance

The model performance by region emphasizes the difference in quality, balance, and amount of data for each of the regions. The strong imbalance between precision and recall for Iceland, Caribbean, and U.S. and Bermuda (Table IV) is likely caused by the imbalance between the number of calls and noise within the datasets (Table II) as well as their smaller dataset sizes. The comparatively high FPR of 0.31 for a threshold of 0.5 and 0.12 for a threshold of 0.9 reached within the Caribbean is a likely cause of the low amounts of noise for that region (only 119 noise samples; see Table II). Although a low number of noise samples does not directly influence the FPR [see Eq. (7)], it is likely to yield a smaller variation in noise. In this way, if the detector fails to correctly identify a specific type of noise within the dataset due to the low number of negatives, this could affect a substantial portion of the noise samples. Although the count of false positives might be low, the FPR would produce larger values. In the same way, the FPR for the U.S. and Bermuda region and Iceland aligns with their imbalance toward negative samples (very high number of negatives, i.e., likely to have a larger variation in noise). Scotland, which features a larger and more balanced dataset, yields a FPR in between the imbalanced datasets with 0.09 for a threshold of 0.5 and 0.06 for a threshold of 0.9. This is similar to the even larger and more balanced combined dataset, reaching a FPR of 0.05 for a threshold of 0.9. The same

similarity can be observed for precision, recall, and F-score between the combined dataset and Scotland. Both achieve high precision and recall values of 0.96 and 0.82 for the combined dataset and 0.95 and 0.82 for Scotland. With regard to the F-score, both the combined dataset and Scotland reach a performance of 0.88. These similarities highlight the importance of the size and balance of datasets when evaluating performance metrics. Moreover, the results reflect the large overlap between the Scotland data and the combined dataset.

To effectively reduce the amount of manual review necessary when detecting North Atlantic humpback whale song, the aim of this study was to develop a detector that produces low amounts of false positives. While reviewing the generated annotations in training phase 2 (Sec. II C 2), the main sources of false positives were boat noise and recording device related noise (especially noise from the spin-up of magnetic hard drive platters). After incorrect detections were marked as explicit negatives during training phase 2, using them in the noise augmentation (Sec. II D 2) made it possible to present the negatives in the context of different noise environments, thereby increasing their variability and occurrence in the dataset in an effort to train the model not to mistake them again. Furthermore, an operating threshold of 0.9 was set to further reduce the number of false positives. At this operating threshold, a FPR of 0.05 on context window level shows that in the areas included in the dataset, the model is able to yield only a small number of false positives. The use of aggregation metrics further improves this performance and allows using the model on hourly and daily levels. Figure 6 shows the effects that threshold and limit have on the F-score on hourly and daily levels when using the sequence limit. The hourly presence plot highlights this by showing a very uniform pattern for each of the F-score trajectories corresponding to unique thresholds. Thereby, threshold and limit of the sequence limit aggregation metric can be tuned to site-specific conditions. F-score values on hourly and daily bases (Table V) reach promising values with 0.93 on hourly basis and 0.94 on daily basis. Simultaneously, FPR values reach 0.01 on hourly and 0.07 on daily level when using the sequence limit. The drop of FPR values from context window (0.05) to hourly (0.01) highlights the aggregation metric's ability to filter noise while not increasing false negatives (recall only changes from 0.82 to 0.81 from context window level to hourly level). The increase in FPR from 0.01 to 0.07 when aggregating from hourly to daily level can be attributed to the simple aggregation (from hourly to daily as explained in Sec. II G) and its susceptibility to false positives if a single hour is a false positive. This effect is amplified because of the far lower (24 times less) number of samples for the daily vs hourly bases (Table II), increasing the negative effect of a single false positive. The increase in FPR is, furthermore, countered by an improvement in F-score from 0.9 to 0.96, showing that the overall performance does not decrease.

In the comparison between the two aggregation metrics, the sequence limit reached slightly better performance values than the simple limit when considering the FPR (0.01 vs 0.02 on hourly level and 0.07 vs 0.14 on daily level). An anthropogenic noise source, e.g., a vessel, might cause occasional prediction values to exceed the threshold. If this occurs to an extent that exceeds the limit set in the simple limit, the model will predict the presence of humpback whale song. The sequence limit, however, is more robust to infrequent high value predictions. This feature makes the sequence limit less prone to be triggered by infrequent anthropogenic sounds or vocalizations of other cetaceans.

Two known challenges for autonomous detection algorithms of humpback whale song are: In high latitude North Atlantic waters, bowhead whales are known to produce complex songs with units, which are likely to be mistaken for humpback whale song units (Erbs *et al.*, 2021). Similarly, on the east coast of the U.S., North Atlantic right whales are known to produce upsweeps similar to some of the humpback whale song unit upsweeps (Davis *et al.*, 2017). Whereas the sequence limit should prove helpful to filter infrequent vocalizations by other cetaceans, a retraining of the detector might be necessary. To provide a tool that can assist in addressing these challenges, *AcoDet* includes functionalities to retrain the North Atlantic model, thereby encouraging researchers to fine-tune the model to environmental conditions of a specific site or population if performance is unsatisfactory.

Whereas the framework *AcoDet* can be used for training and evaluation of new models, its main purpose is inference. Using the graphical user interface, a dataset is selected as input, parameters are specified, and once the computation is complete, Raven annotation tables (Cornell Laboratory of Ornithology, 2014) are generated, which can be directly imported into the Raven software for analysis of annotated spectrograms. Furthermore, the user can choose to generate hourly (or daily) counts as well as presence spreadsheets and visualizations to analyze diel activity patterns throughout the dataset. Figure 7 shows an example of hourly counts and hourly presence visualizations generated by *AcoDet* using the sequence limit. The same output is generated using the simple limit, allowing the user to choose the preferred aggregation metric.

Ease of use was the main driver in the development of *AcoDet*, requiring little to no prior coding experience to use the functionalities. In the past, novel machine learning models have produced promising results, however, the technical skillset required to apply and use the models on new datasets have left their potential unmet. Existing python frameworks, like ketos (Kirsebom *et al.*, 2021), koogu (Madhusudhana, 2022), or vak (Nicholson and Cohen, 2022), provide similar functionalities, yet at the time of writing the incorporation of custom models, such as the NOAA/Google model, into ketos, koogu, or vak proved more complicated than the creation of a custom framework. At the time of writing, researchers from Ireland, Scotland, and the U.S. have successfully used *AcoDet* on their datasets and vastly reduced

the time necessary to process large datasets. Furthermore, the North Atlantic model for humpback whale song detection can be used with the audio detection software PAMGuard (Gillespie *et al.*, 2009).

In summary, the framework *AcoDet* can be used to build training datasets from annotated files, train different model architectures on the training data, and evaluate the models. Furthermore, sample rate, context window length, and spectrogram resolution can be set by the user. By combining these functionalities, *AcoDet* is not limited to humpback whale song but can be applied to other species detection tasks. All necessary code and explanations for implementation can be found online.¹

V. CONCLUSION

In this study, an existing machine learning model for North Pacific humpback whale song is successfully adapted to the North Atlantic to provide researchers with a model for automated detection. A dataset spanning a large spatial and temporal variation was collated, comprised of approximately 60 000 samples from throughout the North Atlantic Ocean, which was used to fine-tune the existing North Pacific model. The resulting North Atlantic model was enhanced by the implementation of three different augmentation techniques [time-shift (Sec. IID 1), noise augmentation (Sec. IID 2), and SpecAugment (Sec. IID 3)], leading to promising performance results on data from the North Atlantic with F-score values of 0.88 on context window level and 0.96 on daily level and FPR values of 0.05 on context window level and 0.07 on daily level. Annotations are generated on context window level by default and can be generated on hourly and daily levels through the use of aggregation metrics. The newly developed North Atlantic model can be used for inference, training, and evaluation using the open-source framework *AcoDet*.

Countless hours of audio recordings containing humpback whale song have been recorded in the North Atlantic in previous decades. Automated detection algorithms that yield low FPRs have the potential of unfolding the information within archived datasets and datasets currently being built. This study aims to contribute to our understanding of North Atlantic humpback whale communication and behavior.

SUPPLEMENTARY MATERIAL

See the supplementary material for dataset metadata. Table S1 includes locations, recording equipment details, recording dates, and reference for the respective recording sites included in the dataset.

ACKNOWLEDGMENTS

The processing of large amounts of data requires large amounts of annotated data. For this the author would like to thank Tamara Narganes Homfeldt for providing hourly presence annotations for the Bermuda data. Furthermore, a

thank you to Steven Benjamins and Nienke Van Geel, who have annotated the hourly presence data used in Sec. III C. The authors would like to thank the anonymous reviewers for their comments that greatly improved this paper. The author would like to acknowledge, that PAM hydrophones for data from the Caribbean (Yarari Marine Mammal and Shark Reserve) were provided by Wageningen Marine Research while the Netherlands Ministry of Agriculture Nature and Food Quality funded their maintenance and data retrieval under project BO-43-117-003. The author would also like to thank Robert P. Dziak from NOAA Pacific Marine Environmental Laboratory (PMEL) and David K. Mellinger from both the Cooperative Institute for Marine Ecosystems and Resources Studies as well as the Marine Mammal Institute, both of which are part of the Oregon State University, as well as NOAA PMEL. Funding was provided by NOAA's Right Whale Grants Program and from the Office of Naval Research grant #N00014-03-1-0099. Data were provided by D.R., G.D., B.B., A.S., and L.H. Spectrograms were annotated and reviewed by C.G., D.R., and V.K. Code and advice for manipulation of the NOAA/Google model was provided by M.H. All code relevant for model training and *AcoDet* development was written by V.K. with advice by F.S. and M.H. The manuscript was written by V.K. All authors contributed to the concept and design of the conducted work, provided critical feedback, and helped shape the research, analysis, and manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare that they have no conflicts of interest associated with this work.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

¹See <https://github.com/vskode/acodet> (Last viewed 3 March 2024).

- Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., and Alvarez, R. (2013). "Real-time bioacoustics monitoring and automated species identification," *PeerJ* **1**, e103.
- Allen, A. N., Harvey, M., Harrell, L., Jansen, A., Merckens, K. P., Wall, C. C., Cattiau, J., and Oleson, E. M. (2021). "A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset," *Front. Mar. Sci.* **8**, 607321.
- Allen, J. A., Garland, E. C., Dunlop, R. A., and Noad, M. J. (2018). "Cultural revolutions reduce complexity in the songs of humpback whales," *Proc. R. Soc. B* **285**(1891), 20182088.
- Anderson, M., and Harte, N. (2021). "Bioacoustic event detection with prototypical networks and data augmentation," *arXiv:2112.09006*.
- Au, W. W. L., Pack, A. A., Lammers, M. O., Herman, L. M., Deakos, M. H., and Andrews, K. (2006). "Acoustic properties of humpback whale songs," *J. Acoust. Soc. Am.* **120**(2), 1103–1110.

- Baumgartner, M. F., and Mussoline, S. E. (2011). "A generalized baleen whale call detection and classification system," *J. Acoust. Soc. Am.* **129**(5), 2889–2902.
- Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., Hofer, H., and Maier, A. (2019). "ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning," *Sci. Rep.* **9**(1), 10997.
- Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. (2019). "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," *Sci. Rep.* **9**(1), 12588.
- Berrow, S. D., Massett, N., Whooley, P., Jann, B. V. M., Lopez-Suárez, P., Stevick, P. T., and Wenzel, F. W. (2021). "Resightings of humpback whales (*Megaptera novaeangliae*) from Ireland to a known breeding ground: Cabo Verde, West Africa," *Aquat. Mamm.* **47**(1), 63–70.
- Cholewiak, D. M., Risch, D., Valtierra, R., and Van Parijs, S. (2013). "Methods for passive acoustic tracking of marine mammals: Estimating calling rates, depths and detection probability for density estimation," in *Detection, Classification and Localization of Marine Mammals*, edited by O. Adam and F. Samaran (Dirac NGO, Paris, France), Chap. 6, pp. 107–145.
- Clapham, P. J. (1996). "The social and reproductive biology of humpback whales: An ecological perspective," *Mamm. Rev.* **26**(1), 27–49.
- COMPASS (2023). *The COMPASS Project—Collaborative Oceanography and Monitoring for Protected Areas and Species*, COMPASS, available at <https://compass-oceanscience.eu/> (Last viewed 12 November 2023).
- Cornell Laboratory of Ornithology (2014). "Raven Pro: Interactive Sound Analysis Software, version 1.5" (Cornell Laboratory of Ornithology, Ithaca, NY).
- Davis, G. E., Baumgartner, M. F., Bonnell, J. M., Bell, J., Berchok, C., Thornton, J. B., Brault, S., Buchanan, G., Charif, R. A., Cholewiak, D., Clark, C. W., Corkeron, P., Delarue, J., Dudzinski, K., Hatch, L., Hildebrand, J., Hodge, L., Klinck, H., Kraus, S., Martin, B., Mellinger, D. K., Moors-Murphy, H., Nieukirk, S., Nowacek, D. P., Parks, S., Read, A. J., Rice, A. N., Risch, D., Sirović, A., Soldevilla, M., Stafford, K., Stanistreet, J. E., Summers, E., Todd, S., Ward, A., and Van Parijs, S. M. (2017). "Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014," *Sci. Rep.* **7**(1), 13460.
- Davis, G. E., Baumgartner, M. F., Corkeron, P. J., Bell, J., Berchok, C., Bonnell, J. M., Thornton, J. B., Brault, S., Buchanan, G. A., Cholewiak, D. M., Clark, C. W., Delarue, J., Hatch, L. T., Klinck, H., Kraus, S. D., Martin, B., Mellinger, D. K., Moors-Murphy, H., Nieukirk, S., Nowacek, D. P., Parks, S. E., Parry, D., Pegg, N., Read, A. J., Rice, A. N., Risch, D., Scott, A., Soldevilla, M. S., Stafford, K. M., Stanistreet, J. E., Summers, E., Todd, S., and Van Parijs, S. M. (2020). "Exploring movement patterns and changing distributions of baleen whales in the western North Atlantic using a decade of passive acoustic data," *Global Change Biol.* **26**(9), 4812–4840.
- Erbs, F., van der Schaar, M., Weissenberger, J., Zaugg, S., and André, M. (2021). "Contribution to unravel variability in bowhead whale songs and better understand its ecological significance," *Sci. Rep.* **11**(1), 168.
- Frasier, K. E., Roch, M. A., Soldevilla, M. S., Wiggins, S. M., Garrison, L. P., and Hildebrand, J. A. (2017). "Automated classification of dolphin echolocation click types from the Gulf of Mexico," *PLoS Comput. Biol.* **13**(12), e1005823.
- Garibbo, S., Blondel, P., Heald, G., Heyburn, R., Hunter, A., and Williams, D. (2021). "Characterising and detecting fin whale calls using deep learning at the Lofoten-Vesterålen Observatory, Norway," *Proc. Mtgs. Acoust.* **44**(1), 070021.
- Garland, E. C., Goldizen, A. W., Rekdahl, M. L., Constantine, R., Garrigue, C., Hauser, N. D., Poole, M. M., Robbins, J., and Noad, M. J. (2011). "Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale," *Curr. Biol.* **21**(8), 687–691.
- Garland, E. C., and McGregor, P. K. (2020). "Cultural transmission, evolution, and revolution in vocal displays: Insights from bird and whale song," *Front. Psychol.* **11**, 544929.
- Garland, E. C., Rendell, L., Lilley, M. S., Poole, M. M., Allen, J., and Noad, M. J. (2017). "The devil is in the detail: Quantifying vocal variation in a complex, multi-levelled, and rapidly evolving display," *J. Acoust. Soc. Am.* **142**(1), 460–472.
- Gillespie, D., Mellinger, D. K., Gordon, J., McLaren, D., Redmond, P., McHugh, R., Trinder, P., Deng, X.-Y., and Thode, A. (2009). "PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans," *J. Acoust. Soc. Am.* **125**, 2547.
- Gordon, J., and Tyack, P. L. (2001). "Sound and cetaceans," in *Marine Mammals: Biology and Conservation*, edited by P. G. H. Evans and J. A. Raga (Springer US, Boston, MA), pp. 139–196.
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 27 June 2016 (Computer Vision Foundation, Washington, DC), pp. 770–778, available at https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html (Last viewed 12 November 2023).
- Helble, T. A., Ierley, G. R., D'Spain, G. L., Roch, M. A., and Hildebrand, J. A. (2012). "A generalized power-law detection algorithm for humpback whale vocalizations," *J. Acoust. Soc. Am.* **131**(4), 2682–2699.
- Hildebrand, J. A., Frasier, K. E., Helble, T. A., and Roch, M. A. (2022). "Performance metrics for marine mammal signal detection and classification," *J. Acoust. Soc. Am.* **151**(1), 414–427.
- Kirsebom, O. S., Frazao, F., Padovese, B., Sakib, S., and Matwin, S. (2021). "Ketos—A deep learning package for creating acoustic detectors and classifiers," *J. Acoust. Soc. Am.* **150**(4), A164.
- Lin, T.-H., Yu, H.-Y., Chen, C.-F., and Chou, L.-S. (2015). "Passive acoustic monitoring of the temporal variability of odontocete tonal sounds from a long-term marine observatory," *PLoS One* **10**(4), e0123943.
- Madhusudhana, S. (2022). "shyamblast/Koogu: version 0.7.1," Zenodo, available at <https://doi.org/10.5281/zenodo.7275319> (Last viewed 28 March 2023).
- Mooney, T. A., Yamato, M., and Branstetter, B. K. (2012). "Hearing in cetaceans: From natural history to experimental biology," in *Advances in Marine Biology*, edited by M. Lesser (Academic, New York), Chap. 4, pp. 197–246.
- Narganes Homfeldt, T., Risch, D., Stevenson, A., and Henry, L.-A. (2022). "Seasonal and diel patterns in singing activity of humpback whales migrating through Bermuda," *Front. Mar. Sci.* **9**, 941793.
- Nicholson, D., and Cohen, Y. (2022). "vak," Zenodo, available at <https://doi.org/10.5281/zenodo.6808839> (Last viewed 3 June 2023).
- O'Neil, K. E., Cunningham, E. G., and Moore, D. M. (2019). "Sudden seasonal occurrence of humpback whales *Megaptera novaeangliae* in the Firth of Forth, Scotland and first confirmed movement between high-latitude feeding grounds and United Kingdom waters," *Mar. Biodivers. Rec.* **12**(1), 12.
- Pandeya, Y. R., Kim, D., and Lee, J. (2018). "Domestic cat sound classification using learned features from deep neural nets," *Appl. Sci.* **8**(10), 1949.
- Parijs, S. M. V., Clark, C. W., Sousa-Lima, R. S., Parks, S. E., Rankin, S., Risch, D., and Van Opzeeland, I. C. (2009). "Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales," *Mar. Ecol. Prog. Ser.* **395**, 21–36.
- Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, Graz, Austria, 15 September 2019 (International Speech Communication Association, Baixas, France), pp. 2613–2617.
- Payne, K. (1983). "Progressive changes in the songs of humpback whales *Megaptera novaeangliae*: A detailed analysis of two seasons in Hawaii," in *Communication and Behavior of Whales* (Westview Press, Boulder, CO), pp. 9–57.
- Payne, R. S., and McVay, S. (1971). "Songs of humpback whales," *Science* **173**(3997), 585–597.
- Ryan, C., Wenzel, F. W., López-Suárez, P., and Berrow, S. (2014). "An abundance estimate for humpback whales *Megaptera novaeangliae* breeding around Boa Vista, Cape Verde Islands," available at <https://research.thea.ie/handle/20.500.12065/234> (Last viewed 20 February 2023).
- Saloma, A., Ratsimbazafindranahaka, M. N., Martin, M., Andrianarimisa, A., Huetz, C., Adam, O., and Charrier, I. (2022). "Social calls in humpback whale mother-calf groups off Sainte Marie breeding ground (Madagascar, Indian Ocean)," *PeerJ* **10**, e13785.
- Schulze, J. N., Denkinger, J., Oña, J., Poole, M. M., and Garland, E. C. (2022). "Humpback whale song revolutions continue to spread from the central into the eastern South Pacific," *R. Soc. Open Sci.* **9**(8), 220158.
- Shiu, Y., Palmer, K. J., Roch, M. A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (2020). "Deep

- neural networks for automated detection of marine mammal species,” *Sci. Rep.* **10**(1), 607.
- Shorten, C., and Khoshgoftaar, T. M. (2019). “A survey on image data augmentation for deep learning,” *J. Big Data* **6**(1), 60.
- Smith, T. D., Allen, J., Clapham, P. J., Hammond, P. S., Katona, S., Larsen, F., Lien, J., Mattila, D. K., Palsbøll, P. J., Sugurjónsson, J., Stevick, P. T., and Ølen, N. (1999). “An ocean-basin-wide mark-recapture study of the North Atlantic humpback whale (*Megaptera novaeangliae*),” *Mar. Mammal Sci.* **15**(1), 1–32.
- Stevick, P. T., Allen, J., Clapham, P. J., Friday, N., Katona, S. K., Larsen, F., Lien, J., Mattila, D. K., Palsbøll, P. J., Sugurjónsson, J., Smith, T. D., Øien, N., and Hammond, P. S. (2003). “North Atlantic humpback whale abundance and rate of increase four decades after protection from whaling,” *Mar. Ecol. Prog. Ser.* **258**, 263–273.
- Sung, K.-K., and Poggio, T. (1995). “Learning human face detection in cluttered scenes,” in *Computer Analysis of Images and Patterns*, Lecture Notes in Computer Science, edited by V. Hlaváč and R. Šára (Springer, Berlin), pp. 432–439.
- Thomas, L., and Marques, T. A. (2012). “Passive acoustic monitoring for estimating animal density,” *Acoust. Today* **8**(3), 35–44.
- Thomas, M., Martin, B., Kowarski, K., Gaudet, B., and Matwin, S. (2020). “Marine mammal species classification using convolutional neural networks and a novel acoustic representation,” in *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, edited by U. Brefeld (Springer International Publishing, Cham), pp. 290–305.
- Todd, N. R. E., Jessopp, M., Rogan, E., and Kavanagh, A. S. (2022). “Extracting foraging behavior from passive acoustic monitoring data to better understand harbor porpoise (*Phocoena phocoena*) foraging habitat use,” *Mar. Mammal Sci.* **38**, 1623–1642.
- Tyack, P. L. (1997). “Studying how cetaceans use sound to explore their environment,” in *Communication, Perspectives in Ethology*, edited by D. H. Owings, M. D. Beecher, and N. S. Thompson (Springer, Boston, MA), pp. 251–297.
- van Geel, N. C. F., Risch, D., Benjamins, S., Brook, T., Culloch, R. M., Edwards, W. J., Stevens, C., and Wilson, B. (2022). “Monitoring cetacean occurrence and variability in ambient sound in Scottish offshore waters,” *Front. Remote Sens.* **3**, 934681.
- Vu, E., Risch, D., Clark, C. W., Gaylord, L., Hatch, L. T., Thompson, M. A., Wiley, D. N., and Van Parijs, S. M. (2012). “Humpback whale song occurs extensively on feeding grounds in the western North Atlantic Ocean,” *Aquat. Biol.* **14**(2), 175–183.
- Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., and Saurous, R. A. (2017). “Trainable frontend for robust and far-field keyword spotting,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA (Institute of Electrical and Electronics Engineers Signal Processing Society, Piscataway, NJ), pp. 5670–5674.
- Weir, C. R., Pollock, C., Cronin, C., and Taylor, S. (2001). “Cetaceans of the Atlantic Frontier, north and west of Scotland,” *Cont. Shelf Res.* **21**(8), 1047–1071.
- Wenzel, F. W., Allen, J., Berrow, S., Hazevoet, C. J., Jann, B., Seton, R. E., Steiner, L., Stevick, P., Lopez Suárez, P., and Whooley, P. (2009). “Current knowledge on the distribution and relative abundance of humpback whales (*Megaptera novaeangliae*) off the Cape Verde Islands, Eastern North Atlantic,” *Aquat. Mamm.* **35**(4), 502–510.
- Wenzel, F. W., Broms, F., Lopez-Suárez, P., Lopes, K., Veiga, N., Yeoman, K., Rodrigues, M. S. D., Allen, J., Fernald, T. W., Stevick, P. T., Jones, L., Jann, B., Bouveret, L., Ryan, C., Berrow, S., and Corkeron, P. (2020). “Humpback whales (*Megaptera novaeangliae*) in the Cape Verde Islands: Migratory patterns, resightings, and abundance,” *Aquat. Mamm.* **46**(1), 21–31.
- White, E. L., White, P. R., Buli, J. M., Risch, D., Beck, S., and Edwards, E. W. J. (2022). “More than a whistle: Automated detection of marine sound sources with a convolutional neural network,” *Front. Mar. Sci.* **9**, 879145.
- Zhong, M., Castellote, M., Dodhia, R., Ferres, J. L., Keogh, M., and Brewer, A. (2020). “Beluga whale acoustic signal classification using deep learning neural network models,” *J. Acoust. Soc. Am.* **147**(3), 1834–1841.