



HAL
open science

Detecting, classifying, and counting blue whale calls with Siamese neural networks

Ming Zhong, Maelle Torterotot, Trevor Branch, Kathleen Stafford, Jean-Yves Royer, Rahul Dodhia, Juan Lavista Ferres

► **To cite this version:**

Ming Zhong, Maelle Torterotot, Trevor Branch, Kathleen Stafford, Jean-Yves Royer, et al.. Detecting, classifying, and counting blue whale calls with Siamese neural networks. *Journal of the Acoustical Society of America*, 2021, 149 (5), pp.3086-3094. 10.1121/10.0004828 . hal-03436284

HAL Id: hal-03436284

<https://hal.science/hal-03436284>

Submitted on 23 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Detecting, Classifying, and Counting Blue Whale Calls**

2 **with Siamese Neural Networks**

3
4 Ming Zhong,¹ Maelle Torterotot,² Trevor A. Branch,³ Kathleen M. Stafford,⁴ Jean-Yves Royer,²

5 Rahul Dodhia,¹ Juan Lavista Ferres¹

6 ¹AI for Good Research Lab, Microsoft, Redmond, WA 98052, USA

7 ²Laboratory Geosciences Ocean, University of Brest & CNRS, Brest, France

8 ³School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98105, USA

9 ⁴Applied Physics Laboratory, University of Washington, Seattle, WA 98105, USA

10

11

12

13

14

15

16

17

18

19 This paper is part of a special issue on Machine Learning in Acoustics.

20 **Abstract:** Blue whales are endangered worldwide, and there are widely recognized to be at
21 least four clearly distinct populations of blue whales in the Indian Ocean, largely based on
22 different song types associated with each population. The goal of this project is to use acoustic
23 signatures to detect, classify and count the calls of each acoustic population so that, ultimately,
24 the conservation status of each population can be better assessed. We used manual
25 annotations from 350 hours of audio recordings from the underwater hydrophones in the
26 Indian Ocean to build a deep learning model to detect, classify, and count the calls from four
27 acoustic song types. The method we used was Siamese Networks, a class of neural network
28 architectures that are used to find the similarity of the inputs by comparing its feature vectors,
29 finding that they outperformed the more widely used convolutional neural networks (CNN).
30 Specifically, the Siamese Networks outperform a CNN with 2% accuracy improvement in
31 population classification and 1.7% - 6.4% accuracy improvement in call count estimation for
32 each blue whale population. In addition, even though we treat the call count estimation
33 problem as a classification task and encode the number of calls in each spectrogram as
34 categorical variable, SNN surprisingly learned the ordinal relationship among them. Siamese
35 Networks are robust and shown here to be an effective way to automatically mine large
36 acoustic data sets for blue whale calls.

37

38 Keywords: Machine learning, bioacoustics, Convolutional Neural networks

39

40

41

42

43 **I. Introduction**

44 *A. Background*

45 The blue whale *Balaenoptera musculus* is the largest of the mysticete (baleen) whales, with
46 lengths exceeding 30 meters (McClain *et al.* 2015). They are endangered worldwide, although
47 their population status differs from one location to another. The Indian Ocean, particularly its
48 southern extent, is one of the oceans with the greatest blue whale acoustic diversity (Stafford
49 *et al.* 2011). Blue whale subspecies present in the Indian Ocean include the Antarctic blue whale
50 (*Balaenoptera musculus intermedia*) and the pygmy blue whale (*B. m. breviceuda*); and pygmy
51 blue whales are further separated into multiple acoustic populations and possibly additional
52 subspecies (e.g., *B. m. indica*). In the absence of extensive genetic data from Indian Ocean blue
53 whales to determine speciation, the different song types of Indian Ocean blue whales, which
54 are acoustically somewhat geographically distinct, are used to broadly define populations of
55 blue whales. Prior to intensive commercial whaling beginning in the early 1900s, blue whales
56 were once abundant in the Southern Hemisphere. This was particularly true in the Southern
57 Ocean, where as many as 239,000 Antarctic blue whales congregated in summer to feed
58 (Branch, Matsuoka & Miyashita, 2004), primarily on Antarctic krill *Euphausia superba*.

59 Despite being the largest animal ever to exist on Earth, there is relatively little known about the
60 distribution and migration of blue whales in the Indian Ocean. The Antarctic blue whale has

61 been declared as “Critically Endangered” and pygmy blue whales are listed as “Data Deficient”
62 by the International Union for the Conservation of Nature (Cooke, 2019) due to lack of
63 sufficient data to assess their conservation status. Monitoring blue whales remains a challenge
64 because of the relative scarcity of individuals as well as their pelagic distribution which largely
65 encompasses remote and inaccessible regions of the ocean. Moreover, identifying pygmy from
66 Antarctic blue whales by visual observation is difficult, as they look almost identical at sea,
67 despite the smaller length of pygmy blue whales (Ichihara, 1966). Thus, most of the knowledge
68 about blue whales in the Indian Ocean comes from whaling data (Branch *et al.*, 2007, 2009),
69 and from passive acoustic monitoring (Samaran *et al.*, 2010a, 2013; Stafford *et al.*, 2011; Leroy
70 *et al.*, 2016; Dréo *et al.*, 2018; Torterotot *et al.* 2020). Such monitoring efforts are widespread in
71 the world’s oceans and often result in many terabytes of digital data, which requires big data
72 analysis efforts to analyze efficiently and robustly. Blue whale signals are particularly good
73 candidates for this type of observation, because of their repetitive, long (more than 15 s), loud
74 (more than 180 dB ref 1 μ Pa at 1 m) and low frequency (20–100 Hz) highly stereotyped calls
75 (Cummings and Thompson, 1971). Blue whale song calls (hereafter calls) vary from one region
76 to another and have been used to define acoustic populations which are geographically distinct
77 (McDonald *et al.* 2006; Stafford *et al.* 2011). Taking advantage of the temporal and frequency
78 differences among song units, we used machine learning methods to automatically detect,
79 classify and count blue whale calls from a subset of acoustic recordings from the southern
80 Indian Ocean. By developing a robust machine learning methodology to identify when and
81 where each population occurs, this opens up a pathway to allocate historical catches and recent
82 abundance estimates among the various populations, allowing us to assess the current status of

83 each identified acoustic population. Such status assessments form the basis for appropriate
84 management efforts to conserve these populations for the future.

85 *B. Motivation for the work*

86 Technological advances in the past two decades have allowed researchers to record and archive
87 passive acoustic data from remote underwater ocean moorings. The mooring deployments can
88 be from months to years with acoustic data archived on digital media in the instrument either
89 continuously or on a duty cycle. The acoustic data is retrieved periodically resulting in up to
90 many terabytes of data collected for each site. It is impractical to analyze all of the data
91 manually or in real time. The way to efficiently process such a large volume of acoustic
92 recordings has been the subject of many efforts in the past twenty years and has resulted in a
93 rich body of literature on automated detection methods, particularly for blue whales (e.g.,
94 Stafford *et al.* 2004, 2011, Mouy *et al.* 2009, Širović *et al.* 2009, Gavrilov and McCauley 2013).
95 Detection methods based on bespoke detectors and conventional machine learning classifiers
96 are the most prominent methods used during the last two decades (Kowarski and Moors-
97 Murphy 2020). For example, a non-parametric classification tree analysis (CART) and a Random
98 Forest analysis were implemented to provide robust results to classify 34 identifiable call types
99 of beluga whale vocalizations from the eastern Beaufort Sea population (Garland *et al.* 2015).
100 To investigate the vocal repertoire of Southeast Alaskan humpback whales, three classification
101 systems were used, including aural spectrogram analysis, statistical cluster analysis, and
102 discriminant function analysis, to describe and classify vocalizations; and a hierarchical acoustic
103 structure was identified to classify vocalizations into 16 individual call types nested within four

104 vocal classes (Fournet *et al.* 2015). For blue whale signals in particular, most detection methods
105 have been based on detection either in the time domain (e.g., matched filtering, Stafford *et al.*
106 1998) or in the frequency domain (spectrogram correlation, e.g., Širović *et al.* 2009) although
107 more recent efforts have involved more novel methods, including sparse representation of
108 signals (e.g., Socheleau *et al.* 2015, Torterotot *et al.* 2019).

109 More recently, the rapid development of artificial intelligence and deep learning algorithms
110 provide another approach for intelligent classification and prediction. In classifying animal
111 sounds, deep neural networks (DNN) methods have progressed tremendously with accessibility
112 to large training data and increasing computational power. Using spectrograms generated from
113 raw audio recordings as input, researchers have applied Convolutional Neural Networks (CNN),
114 either by training the model from scratch, or using transfer learning with pre-trained model
115 weights, to classify calls from different species (Bergler *et al.* 2019, Yang *et al.* 2020, Zhong *et al.*
116 2020, Kirsebom *et al.* 2020). Another approach is Recurrent Neural Networks (RNN), which
117 utilizes temporal information of animal calls for classification tasks (Ibrahim *et al.* 2018, Shiu *et*
118 *al.* 2020).

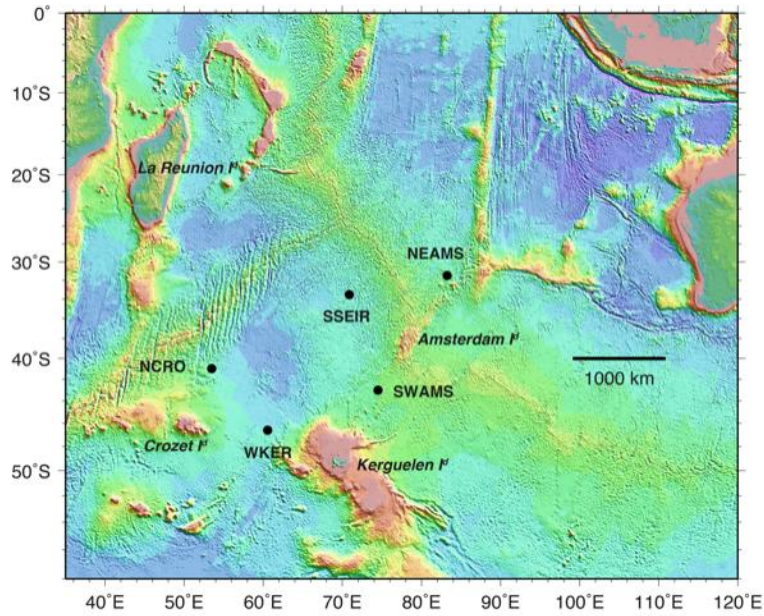
119 While the deep neural network models CNN and RNN have achieved great success in many
120 classification tasks, they have limitations that typically these models rely on large size of
121 datasets to train millions of parameters. For classification purposes of audio recording
122 classifications, all we need from these models is good embedding representations for
123 spectrograms. For same classes, we would expect the learned embeddings to be close to each
124 other in the latent space; for different classes, the learned embeddings are far apart. In this
125 paper, we proposed using Siamese Neural Networks (SNN) (Koch *et al.* 2015) as an alternative

126 of widely used CNN to conduct classifications, especially when the size of training data is
127 limited. Siamese Networks focuses on learning embeddings in the deeper layer that place the
128 same classes close together. Hence, it can learn semantic similarity effectively.

129 **II. Data**

130 **A. Data Sources and Data Annotation**

131 The acoustic data used in this study was recorded by the OHASISBIO (Observatoire Hydro-
132 Acoustique de la SISmicité et de la Biodiversité) hydrophone network (Royer 2009), located in
133 the Southwest Indian Ocean (see Fig. 1). The network was deployed in December 2009 and was
134 still recording as of the date of this publication. To provide a testing and training dataset, we
135 manually annotated signals from four populations of blue whales (Antarctic blue whale and
136 three pygmy blue whale populations) using data from 5 of 11 available mooring sites (see Table
137 I, Figure 2). Originally, song types were named based on the first location where calls were
138 recorded. More recently, with the realization that the extent of each population is greater than
139 originally understood, this naming convention has been updated (IWC 2020) to refer to broad
140 geographical regions as follows (with abbreviation and first location): central Indian Ocean (CIO,
141 Sri Lanka), southwest Indian Ocean (SWIO, Madagascar), southeast Indian Ocean (SEIO,
142 Australia/Indonesia), and Antarctic blue whales. In addition, there are two additional song types
143 of pygmy-type blue whales not yet reported on the OHASISBIO network: southwest Pacific
144 Ocean (SWPO, New Zealand), and northwest Indian Ocean (NWIO, Oman, Cerchio et al. 2020).
145 We follow this regional naming convention throughout the present study (Antarctic, SEIO,
146 SWIO, CIO).



147

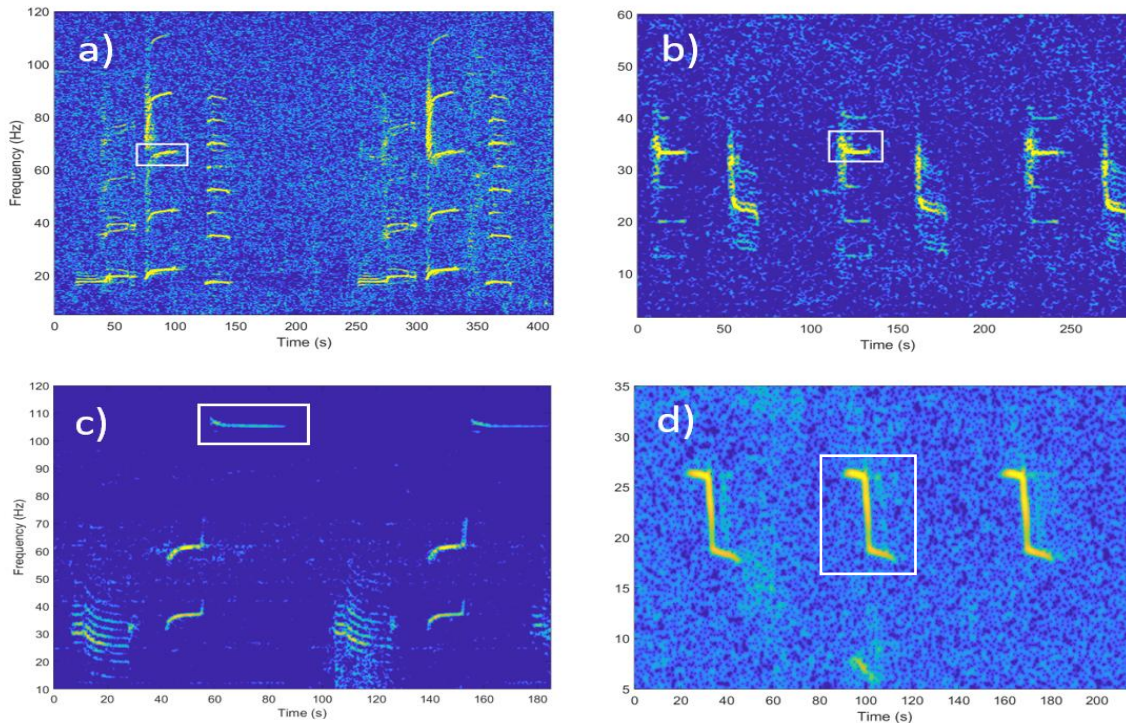
148 **FIG. 1.** Map of the southern Indian Ocean. Black dots represent moorings of the OHASISIBIO
 149 hydrophone network from which data were used in this paper: north of Crozet archipelago
 150 (NCRO); west of Kerguelen Island (WKER); southwest and northeast of St Paul and Amsterdam
 151 islands (SWAMS and NEAMS); south of the southeast Indian Ridge (SSEIR); south of Kerguelen
 152 plateau (ELAN).

153 **TABLE I:** Manually annotated acoustic data from 5 mooring sites for four populations of blue
 154 whales by hours and number of annotations per site.

Mooring Site	Antarctic	SEIO	SWIO	CIO
SSEIR	–	–	–	19.5 h, 138 calls
NCRO	–	–	71.5 h, 1503 calls	–

WKER	32.5 h, 801 calls	13 h, 109 calls	19.5 h, 334 calls	–
SWAMS	26 h, 698 calls	26 h, 572 calls	–	78 h, 537 calls
NEAMS	–	52 h, 769 calls	19.5 h, 841 calls	–

155 Manual annotation was performed with Raven Pro 1.5 (Cornell Lab of Ornithology software) by
156 a single bioacoustics expert. Given the distinct geographical distribution of the four blue whale
157 acoustic populations, four datasets were annotated, one for each call type. The audio files
158 composing each dataset were chosen among the OHASISBIO 2015 recordings, to cover a broad
159 range of acoustic scenarios, from high to low SNR calls. Ten-minute spectrograms with fixed
160 parameters (Hanning windows with 50% overlap and 512-point FFT) were screened for blue
161 whale calls. For pygmy blue whales (CIO, SWIO, SEIO), only the strongest unit was annotated
162 (see white boxes on Fig. 2) whereas for Antarctic blue whales, the whole call was annotated.



163

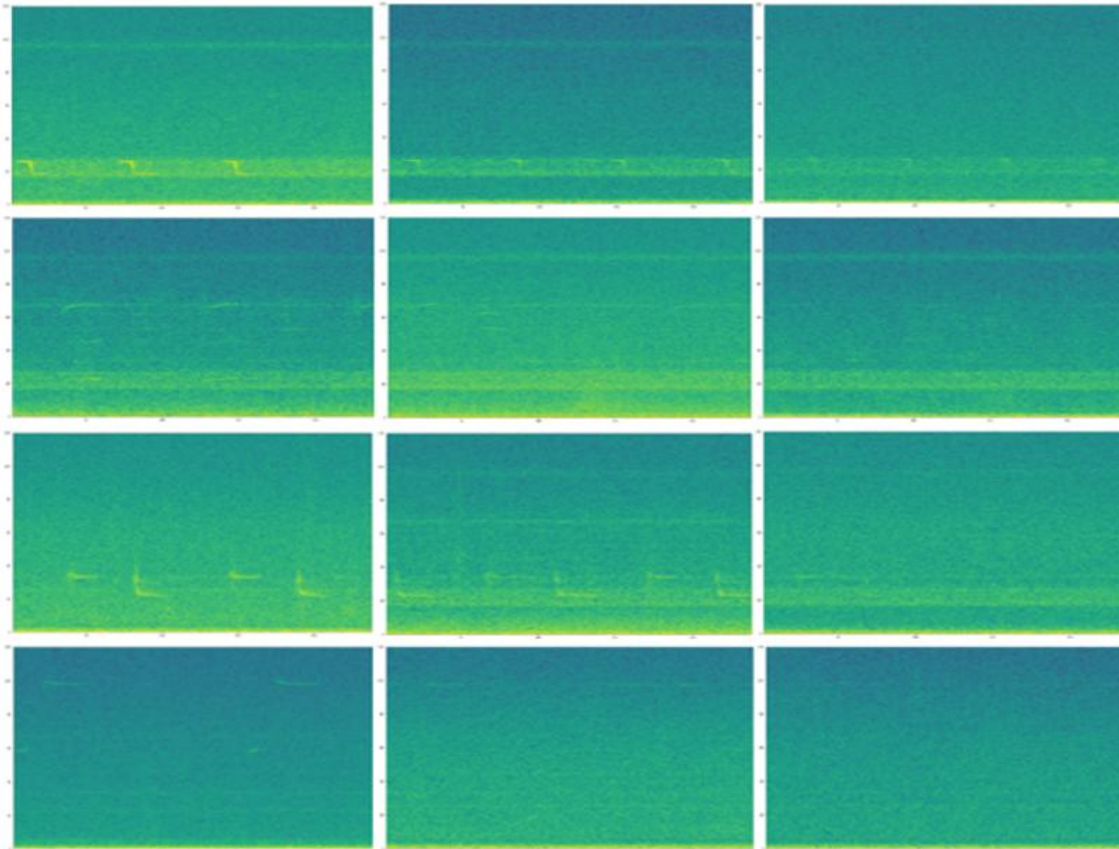
164 **FIG. 2.** Examples of annotated blue whale call with Raven Pro 1.5. a) SEIO pygmy blue whales, b)
 165 SWIO pygmy blue whales, c) CIO blue whales and d) Antarctic blue whales.

166 *B. Data for modeling*

167 For all four acoustic populations of blue whales, calls range from 6 to 40 seconds duration.

168 Using custom written scripts in Python 3.6, spectrograms were produced from audio files (with
 169 NFFT = 1024 and 75% overlap, Hanning window). Each spectrogram was generated from a 240-s
 170 audio segment that contained either one or multiple annotated blue whale calls and was
 171 resized as 224 pixels by 224 pixels with RGB channels (Fig.3). During the annotation process, we
 172 only focused on the presence of one blue whale population in each acoustic file. However, as
 173 part of the temporal and geographical distributions overlap among these blue whale
 174 populations, their acoustic co-occurrence is common. As a result, for each extracted
 175 spectrogram, its corresponding label (the name of blue whale population, and the number of

176 calls associated with the spectrogram) only represented the presence of that particular
177 population but did not indicate absence of the other three populations.



178
179 **FIG. 3.** Example spectrograms from different acoustic populations of blue whale in the Indian
180 Ocean that illustrate the range of signal-to-noise ratios in the data from loudest to faintest from
181 left to right. Row 1: Antarctic; row 2: SEIO; row 3: SWIO; row 4: CIO.

182 While the spectrograms extracted from annotated audio segments corresponded to positive
183 labels (i.e., presence of a blue whale population with at least one call), we also extracted
184 spectrograms that associated with negative labels (i.e., absence of a blue whale population with
185 no call). For each of the four populations, we randomly selected audio clips that did not contain
186 any annotated calls.

187 In total, we extracted 12,155 spectrograms (see Table II for breakdown by population), each
188 representing a 240-second-long audio clip. These spectrograms, along with their associated
189 labels, were used as input for building classification models.

190 **TABLE II:** Number of labeled data for each population of blue whale. The number of true signals
191 is shown in the left-hand column and the number of spectrograms with no calls used as
192 negative training data is shown in the right-hand column.

Population Name	Annotated calls used for training	Null data used for training
Antarctic	1,491	1,099
SEIO	1,459	1,988
SWIO	2,670	1,187
CIO	659	1,602

193

194 **III. Approaches**

195 We assessed the performance of Convolutional Neural Networks (CNN) and a newer technique,
196 Siamese Neural Networks (SNN), to determine which best identified and classified blue whale
197 calls.

198 *A. Classification Models using Convolutional Neural Network (CNN)*

199 Convolutional Neural Networks (CNN) have been widely used for image classification tasks, and
200 their success has also been proven in bioacoustic classification applications (Bianco *et al.* 2019).

201 Here we used the DenseNet-201 architecture (Huang *et al.* 2016) as a baseline to classify calls
202 of the four blue whale populations, and to count the number of calls in each 240-s spectrogram.

203 DenseNet was developed specifically to improve the declined accuracy caused by the vanishing
204 gradient in high-level neural networks and has the advantage of improving feature propagation

205 both in forward as well as backward fashion. In a DenseNet architecture, each layer is
206 connected to every other layer and obtains additional inputs from all preceding layers, and then
207 passes its own feature-maps to all subsequent layers.

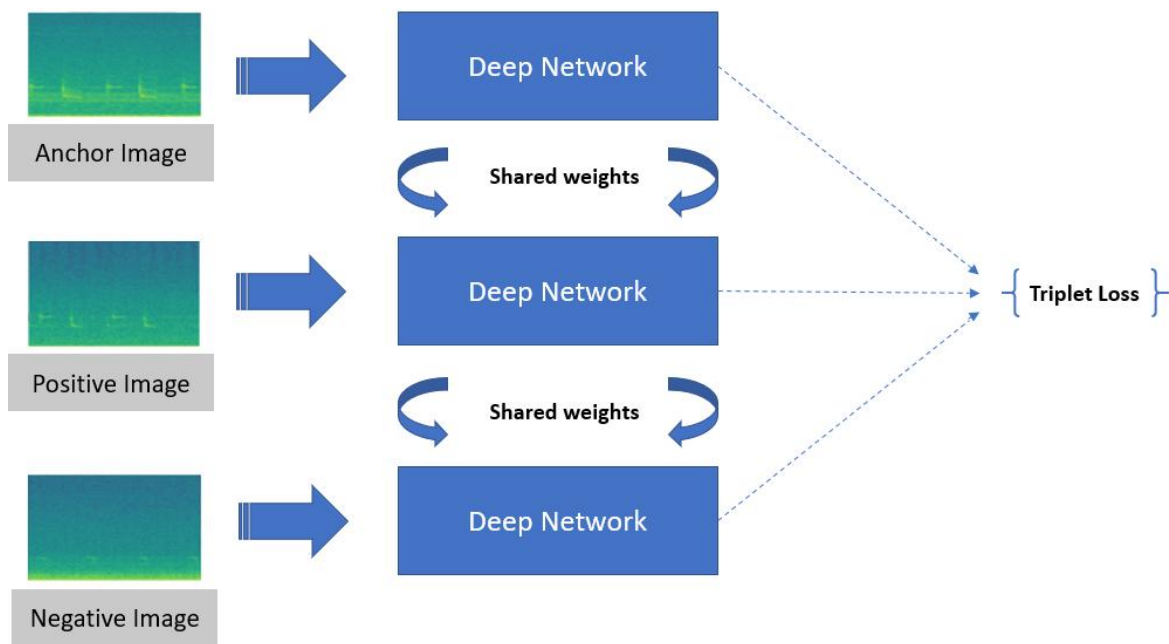
208 *B. Classification Models using Siamese Neural Network (SNN)*

209 Siamese Neural Networks (SNN) are a class of neural network architectures that contain two or
210 more identical subnetworks. “Identical” here means that they have the same configuration with
211 the same parameters and weights. Parameter updating is mirrored across both sub-networks.
212 SNN focuses on learning image embeddings in the deeper layers that place the same classes
213 close together. Hence, it can be used to measure the similarity of the inputs by comparing their
214 feature vectors and make decisions on whether the two images belong to the same category or
215 different categories.

216 Since training of Siamese networks involves pairwise learning, cross entropy loss cannot be
217 used in this case. Instead, we used another loss function called triplet loss (Hoffer and Ailon,
218 2015). This is a loss function where an anchor (baseline) image is compared to a positive image
219 (i.e., an image that is in the same category as the anchor image) and a negative image (i.e., an
220 image that is in a different category as the anchor image). The distance (here we used squared
221 Euclidean distance) from the anchor image to the positive image is minimized, and the distance
222 from the anchor image to the negative image is maximized. As shown in formula (1), $D(x, y)$
223 represents the distance between the learned vector representation of spectrograms x and y ,
224 and α is a margin term used to stretch the distance differences between similar and dissimilar
225 pairs in the triplet, and the remaining parameters represent the feature embeddings for the
226 anchor (a), positive (p), and negative (n) images.

227
$$L(a, p, n) = \max(0, D(a, p) - D(a, n) + \alpha) \quad (1)$$

228 During the training process, an image triplet (anchor image, positive image, negative image) is
 229 fed into the model as a single sample (see Fig. 4). The distance between the anchor and
 230 positive images should be smaller than that between the anchor and negative images. For many
 231 deep learning models, a large training data set is needed to achieve good performance. While
 232 this may not be practical in many real applications, the architecture of Siamese Networks
 233 enables these networks to learn from very little data.



234
 235 **FIG. 4.** Architecture of Siamese Networks with triplet loss.

236 When triplets are generated for model training, as the training continues, some of the
 237 additional triplets are easy to deal with (their loss value is very small or even 0), preventing the
 238 network from further improvement. A good training strategy would be to constantly “mine” out
 239 those difficult cases in each epoch, based on the current performance of model’s snapshot, so
 240 that the model will always have certain percentage of hard cases in the training loop from

241 which it still struggles to tell a difference. This is similar to the triplet mining in FaceNet (Schroff
242 *et al.* 2015). In our training process, we choose batch size = 5. Within each batch, we first
243 generated 5 triplets randomly and kept the 2 hardest examples, and then generated another 3
244 triplets randomly.

245 *C. Implementation*

246 For Convolutional Neural Networks (CNN), since our training data were weakly labeled (that is,
247 for each spectrogram, the corresponding label only indicated the presence or absence of one
248 blue whale call type, without labeling whether there were calls from the remaining three
249 acoustic populations), during the model training, we used a custom binary cross-entropy loss
250 function that only penalized the population category with known labels. For each spectrogram
251 in the training data, therefore, the loss function calculated the loss for the one blue whale call
252 type with a known (either positive or negative) label and did not assess the remaining three
253 populations.

254 For Siamese Neural Networks (SNN), the model outputs an n -dimensional embedding for each
255 spectrogram, where n corresponds to the dimension of the vector before the last (output)
256 layer. For DenseNet-201 that we used, the corresponding $n = 1920$. For each spectrogram in the
257 testing set, we compared its embedding vector with all the embedding vectors of the
258 spectrograms in the training set by calculating distance, and then assigned the label to the
259 population that has the smallest distance (here we used closest 10 training spectrograms from
260 each population).

261 When counting the number of blue whale calls, we only classified the spectrograms that had at
262 least one annotated call, and the model was fit separately to each of the four blue whale
263 acoustic populations, as the call densities varied from one population to another. Only 5% of
264 the training dataset spectrograms had 5 or more annotated calls, and 1% had 6 or more, so we
265 created categorical labels of “1”, “2”, “3”, “4”, and “5+” to correspond the number of calls in
266 each spectrogram.

267 **IV. Results**

268 We have two classification tasks: the first is to detect and classify the presence or absence of
269 calls from each of the four blue whale populations; and the second is to estimate the number of
270 calls from each of these populations in the training dataset and eventually, novel acoustic
271 datasets. For the two tasks, we compared the performance of the CNN and SNN methods. The
272 annotated data was randomly split into training, validation, and testing sets (which account for
273 49%, 21% and 30% of the annotated data, respectively), and the model results were reported
274 on the testing set.

275 *A. Model performance for classifying the presence of blue whale calls*

276 For Convolutional Neural Networks (CNN), the multi-class classification model outputs the
277 predicted probability of blue whale call presence for each population, and can be assessed with
278 commonly used metrics, including accuracy, sensitivity, specificity, and Area Under the Curve
279 (AUC). For Siamese Neural Networks (SNN), the output is not probability based and there is no
280 “threshold score”, and thus no AUC which is measured at various threshold settings.

281 To have a fair comparison of the outputs of the two models, we will then use three metrics:
 282 accuracy, sensitivity, and specificity. To determine these, we denote annotated calls that were
 283 correctly identified as true positives (TP), spectrograms with no calls that were correctly
 284 classified as true negatives (TN), calls that were identified as blue whales but were not
 285 annotated as false positives (FP), and annotated calls that were not correctly identified as false
 286 negatives (FN). Accuracy is the fraction of predictions that model got right (i.e., $(TP + TN)/(TP +$
 287 $FP + TN + FN)$); sensitivity, or true positive rate, measures the percentage of presence that was
 288 correctly predicted (i.e., $TP/(TP + FN)$); and specificity, or true negative rate, measures the
 289 percentage of absence that was correctly predicted (i.e., $TN/(TN + FP)$). Since sensitivity and
 290 specificity in CNN model are dependent on the choice of threshold score, we used a default
 291 neutral threshold score of 0.5. For all three metrics, the Siamese Networks model outperforms
 292 CNN in overall metrics and almost for each individual population, although CNN is slightly
 293 better in Sensitivity for SEIO and Specificity for SWIO (Table III).

294 **TABLE III:** Model results for classifying the presence of blue whale calls for the CNN and SNN
 295 models. Highest performance for each measure and acoustic population is in bold type.

296

Population	CNN			SNN		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
All 4 populations	0.901	0.893	0.909	0.922	0.921	0.922
Antarctic	0.911	0.900	0.922	0.943	0.949	0.936
SEIO	0.908	0.917	0.899	0.909	0.895	0.919

SWIO	0.907	0.905	0.910	0.928	0.957	0.863
CIO	0.838	0.779	0.899	0.908	0.787	0.963

297

298 *B. Model performance for counting the number of blue whale calls*

299 Although treated as a classification task, using standard metrics alone (such as accuracy) that
300 are commonly used to evaluate multi-class classification models may not be appropriate or
301 comprehensive here, as the classes here actually have ordinal implications. Therefore, we used
302 the prediction percentage error as the evaluation metric (see Table IV). The Siamese Networks
303 provided a higher prediction accuracy (lower prediction error) than CNN.

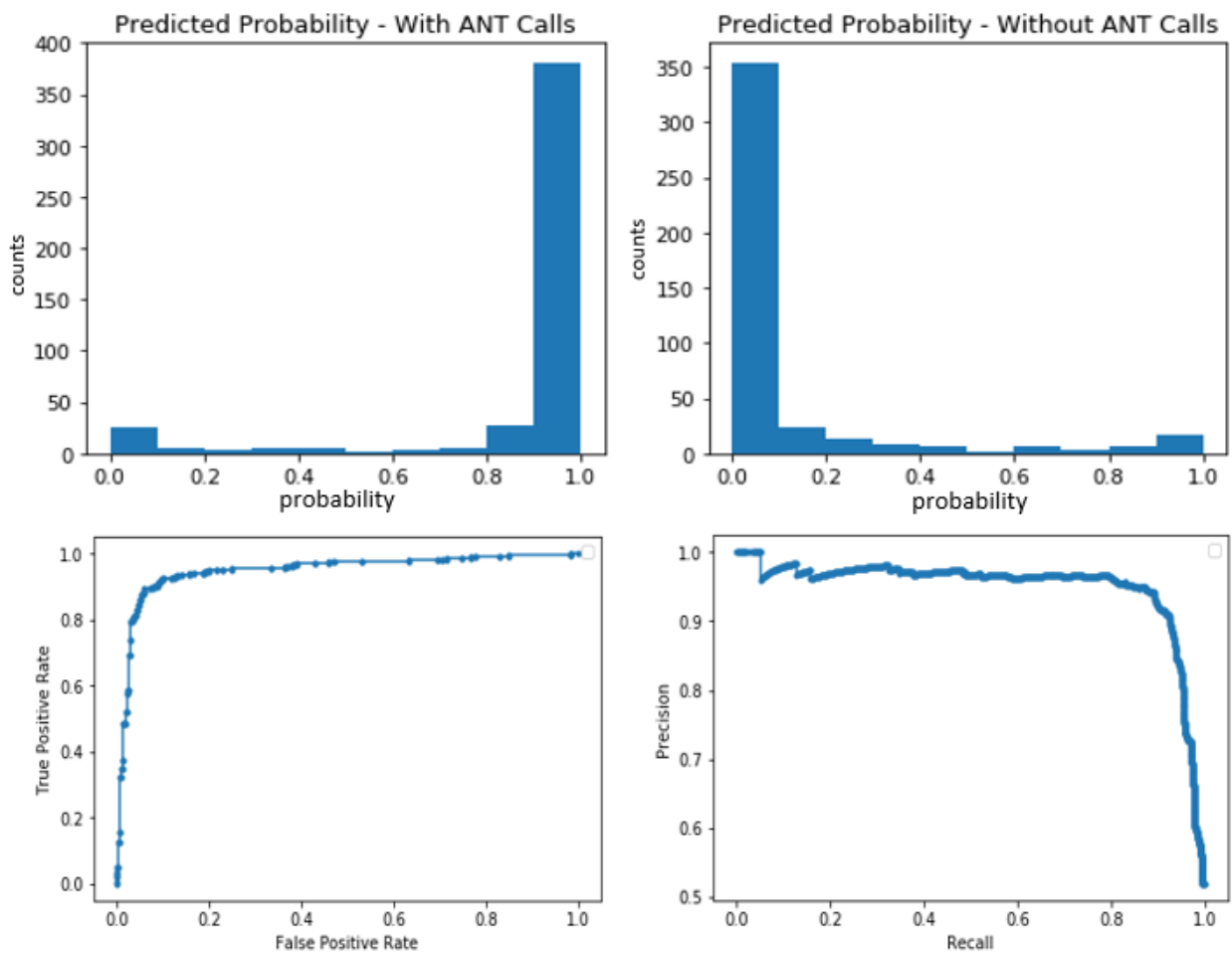
304 **TABLE IV:** Model results for predicting the number of calls by CNN and SNN.

Population	Annotated number of calls	Predicted number of calls by CNN	Predicted number of calls by SNN	Prediction percentage error by CNN	Prediction percentage error by SNN
Antarctic	1478	1552	1504	5%	1.76%
SEIO	889	957	878	7.65%	1.24%
SWIO	2187	2087	2124	4.57%	2.88%
CIO	316	305	311	3.48%	1.58%

305

306 *C. Further comparisons of two models*

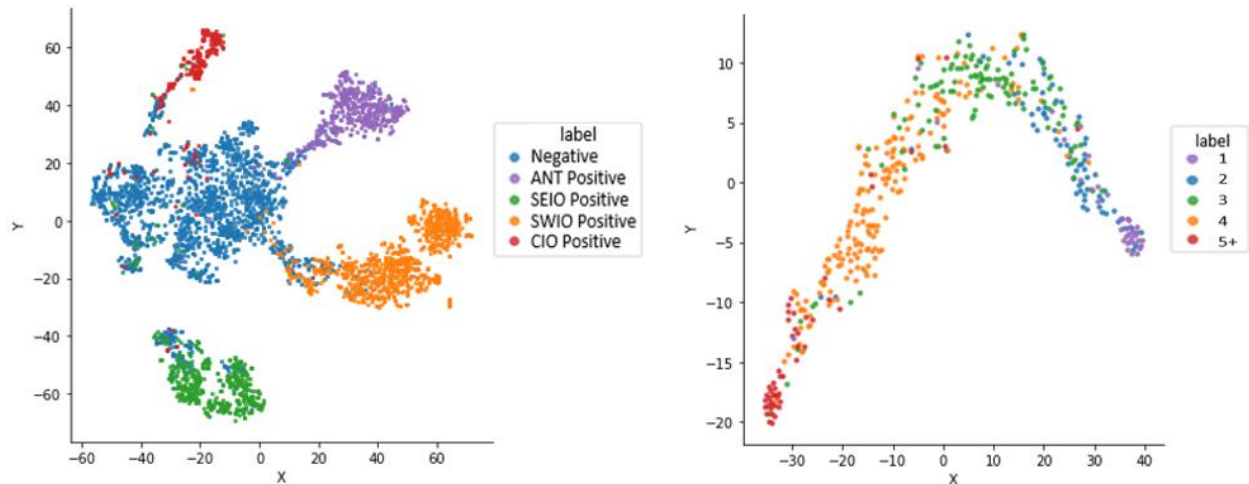
307 Even though Convolutional Neural Networks (CNN) did not perform as well as Siamese Neural
308 Networks (SNN) in this dataset, CNN has its advantages of making predictions with probability
309 score. This makes it convenient for the users to have better understanding of how confident the
310 model is when making classifications and under which circumstances the model may make
311 mistakes. In practical implementations, it also allows users to choose appropriate threshold
312 scores to have either less false positives or less false negatives depending on their specific
313 needs (Fig. 5).



314

315 **FIG. 5.** Illustration of the results of call classification task by CNN. Top left and top right:
316 Histograms for predicted probabilities of positive and negative samples in the testing set.
317 Bottom left: receiver operating characteristic (ROC) curve. Bottom right: precision-recall curve.

318 In contrast, Siamese Networks, at the end of the common network in its architecture, output a
319 vectored representation for each input image, thus providing an easy way to visualize in a 2-
320 dimension t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) plot.
321 t-SNE is a nonlinear dimensionality reduction technique well-suited for embedding high-
322 dimensional data for visualization in a low-dimensional space of two or three dimensions.
323 Specifically, a Siamese Network models each high-dimensional object by a two- or three-
324 dimensional point in such a way that similar objects are modeled by nearby points and
325 dissimilar objects are modeled by distant points with high probability. Fig. 6 shows the t-SNE
326 plots of the testing set for the two classification tasks. From the plot, we can see that the
327 classifications for each of the four blue whale calls are distinct from each other. The “Negative”
328 class, which included “no call” samples for each population, sits in the middle of the four
329 “Positive” classes and overlaps very little with any of them. In the second classification model
330 the number of blue whale calls present in a spectrogram is estimated for each population (Fig
331 4b). Although we encoded the number of calls as categorical variables which ignored their
332 ordinal implications (that is, category “1” should be closer to category “2” than category “3”,
333 and category “2” should be closer to category “3” than category “4” or “5+”, etc.), the Siamese
334 Networks clearly learned such ordinal relationships.



335
336

337 **FIG. 6.** (a) t-SNE plot for the model that classifies the presence or absence of blue whale calls
338 from each of the four populations. (b) t-SNE plot for the model that estimates the number of
339 Antarctic blue whale calls (for the other three populations, the plots show similar patterns).

340 **V. Discussion**

341 We built classification models to detect, classify and count the number of calls by each of four
342 blue whale acoustic populations in the Indian Ocean. In comparison to Convolutional Neural
343 Networks (CNN) which have shown success in several prior research in classifying bioacoustics
344 for multiple species (Bianco *et al.* 2019), Siamese Networks achieved better performance in this
345 study.

346 While Siamese Networks are particularly suitable for scenarios where there are only a few
347 samples in each class (i.e., few-shot learning), they can also be applied to larger datasets, like
348 the one we used in this study. However, since Siamese Networks learns from quadratic pairs (to
349 make use of all information available), the training is much slower than pointwise learning
350 models such as CNN. Additionally, instead of outputting probabilities of the prediction, they

351 output the distance from closest training samples in each class instead. In practice, CNN and
352 SNN can be used together to complement each other. Given that the learning mechanism of
353 SNN is somewhat different from CNN, their ensembled results are likely to perform even better.

354 While both models performed well in general on classifying calls from 4 populations of blue
355 whales, their performance differed among different populations. Classification of Antarctic blue
356 whale calls had the highest accuracy among 4 populations, while CIO had the lowest accuracy.

357 One possible reason is that Antarctic, SEIO and SWIO have larger sizes of training samples
358 compared to CIO, but more likely is that Antarctic blue whale calls (Z-calls) have more
359 frequency modulation on the spectrograms, compared to that of the CIO blue whale calls
360 (which looks like a flat line). Another factor is the call loudness in the audio recordings. In
361 general, CIO blue whale calls have lower signal-to-noise ratios in the annotated data, which
362 increases the difficulty for the model to classify correctly with high confidence. The lower
363 signal-to-noise ratios for CIO blue whale calls could be due to a number of factors among which
364 we cannot currently distinguish. These include the CIO call having a lower source level than
365 other calls; there are only a few source levels reported for blue whale signals globally, and none
366 for CIO blue whale calls. It is also likely that the animals producing these signals are further
367 from the hydrophones than the other populations, given what is known about their
368 distributions, although since the hydrophones are omni-directional we cannot ascertain this for
369 certain. This signal is the highest frequency signal we detected and as such would be subject to
370 greater transmission loss than the other signals.

371 Compared to traditional methods which rely heavily on manual verification by a human user or
372 template matching by software, the method presented here uses deep learning models and has

373 the advantage of flexibility with regards to temporal and frequency variations in a dataset.
374 Notably for blue whale calls, the call frequency has been getting lower in all populations over
375 time (McDonald et al. 2009, Leroy et al. 2018), and one major advantage of this approach is
376 that it looks for the shape of the call independent of the frequency of the call. Siamese
377 Networks can easily classify and count multiple types of calls from several populations at the
378 same time and have the ability to classify novel datasets that were collected from different
379 mooring sites or different years. Even at the sites that have somewhat different underwater
380 environments, the model still detected and classified the signals. An additional, and future
381 advantage is that the model can easily scale up to include other species or call types with the
382 addition of annotated data.

383 Although we treated the call count estimation problem as a classification task and encoded the
384 number of calls in each spectrogram as categorical variable, SNN surprisingly learned the
385 ordinal relationship among them. Call counts, or cue rates (how often a signal occurs over a
386 fixed time period, or number of individuals), are critical elements of density estimation
387 methods for marine mammals. Density estimation is one of the key ways to determine trends in
388 marine mammal populations using single instrument passive acoustic data and estimates of call
389 counts (Küsel *et al.* 2011, Marques *et al.* 2013). In this way, Siamese Networks are robust and
390 shown here to be an effective way to automatically mine large acoustic data sets for the
391 presence and number of blue whale calls.

392

393 **Acknowledgements**

394 This work was supported by AI for Earth grants at Microsoft. Our appreciation to Dan Morris
395 for connecting different parties for fruitful discussions and helpful online materials. Passive
396 acoustic data collection was funded by the French Polar Institute and the French Oceanographic
397 Fleet, with additional support from INSU-CNRS. M.T. acknowledges support from a PhD
398 Fellowship of the University of Brest and a travel grant from the Isblue project (ANR-17-EURE-
399 0015) to visit APL at the University of Washington.

400 **References**

401 Bergler, C., Schröter, H., Cheng, R. X., Barth, V., Weber, M., Nöth, E., Hofer, H., and Maier, A.
402 (2019) ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning. Sci.
403 Rep. 9(1), 10997.

404 Bianco, M.J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M.A., Gannot, S., Deledalle, C.A., and Li,
405 W. (2019) Machine learning in acoustics: Theory and applications,” The Journal of the
406 Acoustical Society of America, 146, 3590–3628.

407 Branch, T.A., Matsuoka, K. & Miyashita, T. (2004) Evidence for increases in Antarctic blue
408 whales based on Bayesian modelling. Marine Mammal Science, 20, 726–754.

409 Cerchio, S., Willson, A., Leroy, E. C., Muirhead, C., Al Harthi, S., Baldwin, R., Cholewiak, D.,
410 Collins, T., Minton, G., Rasoloarijao, T., Rogers, T. L., and Willson, M. S. (2020). A new blue
411 whale song-type described for the Arabian Sea and Western Indian Ocean, Endangered Species
412 Research, 43, 495-515.

413 Cooke, J., 2019. *Balaenoptera Musculus*, (Errata Version Published in 2019). Technical Report.
414 The IUCN Red List of Threatened Species, 2018.

415 Fournet, M.E., Szabo, A., Mellinger, D.K. (2015) Repertoire and classification of non-song calls in
416 Southeast Alaskan humpback whales (*Megaptera novaeangliae*). The Journal of the Acoustical
417 Society of America, 137, 1–10.

418 Garland, E.C., Castellote, M., Berchok, C.L. (2015) Beluga whale (*Delphinapterus leucas*)
419 vocalizations and call classification from the eastern Beaufort Sea population. The Journal of the
420 Acoustical Society of America, 137, 3054–3067.

421 Gavrilov, A.N., McCauley, R. (2013) Acoustic detection and long-term monitoring of pygmy blue
422 whales over the continental slope in southwest Australia. Journal of the Acoustical Society of
423 America, 134, 2505–2513.

424 Hoffer, E., and Ailon, N. (2015) Deep metric learning using triplet network. In International
425 Workshop on Similarity-Based Pattern Recognition.

426 Huang, G., Liu, Z., and Weinberger, K. Q. (2016) Densely connected convolutional networks.
427 arXiv preprint arXiv:1608.06993.

428 Ibrahim, A.K., Zhuang, H., Cherubin, L. M., Schärer-Umpierre, M.T., and Erdol, N. (2018)
429 Automatic classification of grouper species by their sounds using deep neural networks. Journal
430 of the Acoustical Society of America, 144, 196–202.

431 IWC (2020) Report of the Scientific Committee, Virtual Meetings, 11-24 May 2020. Section
432 8.2.1. International Whaling Commission, Cambridge, UK.

433 Kirsebom, O.S., Frazao, F., Simard, Y., Roy, N., Matwin, S., and Giard, S. (2020) Performance of a
434 deep neural network at detecting North Atlantic right whale upcalls. The Journal of the
435 Acoustical Society of America, 147, 2636–2646.

436 Koch, G., Zemel, R., and Salakhutdinov, R. (2015) Siamese neural networks for one-shot image
437 recognition. In ICML Deep Learning workshop.

438 Kowarski, K.A., Moors-Murphy, H. (2020) A review of big data analysis methods for baleen
439 whale passive acoustic monitoring. *Marine Mammal Science* 60:1-22.

440 Küsel E.T., Mellinger D.K., Thomas L., Marques T.A., Moretti D., Ward J. (2011) Cetacean
441 population density estimation from single fixed sensors using passive acoustics. *Journal of the*
442 *Acoustical Society of America*; 129(6):3610–22.

443 Leroy E.C., Samaran F., Stafford K.M., Bonnel J., Royer J.-Y. (2018) Broad-scale study of the
444 seasonal and geographic occurrence of blue and fin whales in the Southern Indian Ocean.
445 *Endangered Species Research* 37:289–300.

446 Marques T.A., Thomas L., Martin S.W., Mellinger D.K., Ward J.A., Moretti D.J., Harris D., Tyack
447 P.L. (2013) Estimating animal population density using passive acoustics. *Biological Reviews of*
448 *the Cambridge Philosophical Society* 88(2):287–309.

449 McClain, C. R., Balk, M. A., Benfield, M. C., Branch, T. A., Chen, C., Cosgrove, J., Dove, A. D. M.,
450 Helm, R. R., Hochberg, F. G., Gaskins, L. C., Lee, F. B., Marshall, A., McMurray, S. E., Schanche,
451 C., Stone, S. N., and Thaler, A. D. (2015) Sizing ocean giants: patterns of intraspecific size
452 variation in marine megafauna, *PeerJ*, 2, e715.

453 McDonald, M. A., Hildebrand, J. A., and Mesnick, S. L. (2006) Biogeographic characterization of
454 blue whale song worldwide: using song to identify populations, *J Cetacean Res Manage*, 8, 55-
455 65.

456 McDonald, M. A., Hildebrand, J. A., and Mesnick, S. (2009) Worldwide decline in tonal
457 frequencies of blue whale songs. *Endangered Species Research*, 9, 13-21.

458 Mouy, X., Bahoura, M., Simard, Y. (2009) Automatic recognition of fin and blue whale calls for
459 real-time monitoring in the St. Lawrence. *The Journal of the Acoustical Society of America*, 126,
460 2918–2928.

461 Royer, J.-Y. (2009) OHASISBIO - Hydroacoustic Observatory for the Seismicity and Biodiversity in
462 the Indian Ocean. Technical report. University of Brest (<https://doi.org/10.18142/229>).

463 Schroff, F., Kalenichenko, D., and Philbin, J. (2015) A unified embedding for face recognition and
464 clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*,
465 2015, 815–823.

466 Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E-M., Helble, T., Cholewiak, D.,
467 Gillespie, D., and Klinck, H. (2020) Deep neural networks for automated detection of marine
468 mammal species. *Sci. Rep.* 10, 607.

469 Širović, A., Hildebrand, J.A., Wiggins, S.M., Thiele, D. (2009) Blue and fin whale acoustic
470 presence around Antarctica during 2003 and 2004. *Marine Mammal Science* 25:125–136.

471 Socheleau, F.-X., Leroy, E., Carvallo Pecci A., Samaran, F., Bonnel, J., Royer, J-Y. (2015)
472 Automated detection of Antarctic blue whale calls. *The Journal of the Acoustical Society of*
473 *America*, 138, 3105–3117.

474 Stafford, K.M., Bohnenstiehl, D.R., Tolstoy, M., Chapp, E., Mellinger, D.K., Moore, S.E. (2004)
475 Antarctic-type blue whale calls recorded at low latitudes in the Indian and eastern Pacific
476 Oceans. *Deep Sea Research Part I: Oceanographic Research Papers* 51:1337–1346.

477 Stafford, K.M., Chapp, E., Bohnenstielh, D.R., Tolstoy, M. (2011) Seasonal detection of three
478 types of “pygmy” blue whale calls in the Indian Ocean. *Marine Mammal Science* 27:828–840.

479 Stafford, K.M., Fox, C.G., Clark, D.S. (1998) Long-range acoustic detection and localization of
480 blue whale calls in the northeast Pacific Ocean. *The Journal of the Acoustical Society of*
481 *America*, 104, 3616–3625.

482 Torterotot, M., Royer, J.-Y., Samaran, F. (2019) Detection strategy for long-term acoustic
483 monitoring of blue whale stereotyped and non-stereotyped calls in the Southern Indian Ocean.
484 *OCEANS 2019 - Marseille*:1–10 (doi: 10.1109/OCEANSE.2019.8867271).

485 Torterotot, M., Samaran, F., Stafford, K. M., and Royer, J.-Y. (2020) Distribution of blue whale
486 populations in the Southern Indian Ocean based on a decade of acoustic monitoring, *Deep-Sea*
487 *Research II*, 179, 104874 (doi: j.dsr2.2020.104874).

488 van der Maaten, L. and Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine*
489 *Learning Research*, 9(Nov):2579–2605.

490 Yang, W., Luo, W. and Zhang, Y. (2020) Classification of odontocete echolocation clicks using
491 convolutional neural network. *The Journal of the Acoustical Society of America*, 147(1), 49–55.

492 Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., and Brewer, A. (2020) Beluga
493 whale acoustic signal classification using deep learning neural network models. *The Journal of*
494 *the Acoustical Society of America*, 147(3), 1834–1841.