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

Detecting, Classifying, and Counting Blue Whale Calls with Siamese Neural Networks

Ming Zhong,[1] Maelle Torterotot,[2] Trevor A. Branch,[3] Kathleen M. Stafford,[4] Jean-Yves Royer,[2] Rahul Dodhia,[1] Juan Lavista Ferres[1]

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

[2] Laboratory Geosciences Ocean, University of Brest & CNRS, Brest, France

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

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

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

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

Keywords: Machine learning, bioacoustics, Convolutional Neural networks

I. Introduction

A. Background

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

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

B. Motivation for the work

Technological advances in the past two decades have allowed researchers to record and archive passive acoustic data from remote underwater ocean moorings. The mooring deployments can be from months to years with acoustic data archived on digital media in the instrument either continuously or on a duty cycle. The acoustic data is retrieved periodically resulting in up to many terabytes of data collected for each site. It is impractical to analyze all of the data manually or in real time. The way to efficiently process such a large volume of acoustic recordings has been the subject of many efforts in the past twenty years and has resulted in a rich body of literature on automated detection methods, particularly for blue whales (e.g., Stafford *et al*. 2004, 2011, Mouy *et al*. 2009, Širović *et al*. 2009, Gavrilov and McCauley 2013).

Detection methods based on bespoke detectors and conventional machine learning classifiers are the most prominent methods used during the last two decades (Kowarski and Moors-Murphy 2020). For example, a non-parametric classification tree analysis (CART) and a Random Forest analysis were implemented to provide robust results to classify 34 identifiable call types of beluga whale vocalizations from the eastern Beaufort Sea population (Garland *et al*. 2015). To investigate the vocal repertoire of Southeast Alaskan humpback whales, three classification systems were used, including aural spectrogram analysis, statistical cluster analysis, and discriminant function analysis, to describe and classify vocalizations; and a hierarchical acoustic structure was identified to classify vocalizations into 16 individual call types nested within four vocal classes (Fournet *et al*. 2015). For blue whale signals in particular, most detection methods have been based on detection either in the time domain (e.g., matched filtering, Stafford *et al*. 1998) or in the frequency domain (spectrogram correlation, e.g., Širović *et al*. 2009) although more recent efforts have involved more novel methods, including sparse representation of signals (e.g., Socheleau *et al*. 2015, Torterotot *et al*. 2019).

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

While the deep neural network models CNN and RNN have achieved great success in many classification tasks, they have limitations that typically these models rely on large size of datasets to train millions of parameters. For classification purposes of audio recording classifications, all we need from these models is good embedding representations for spectrograms. For same classes, we would expect the learned embeddings to be close to each other in the latent space; for different classes, the learned embeddings are far apart. In this paper, we proposed using Siamese Neural Networks (SNN) (Koch *et al.* 2015) as an alternative of widely used CNN to conduct classifications, especially when the size of training data is limited. Siamese Networks focuses on learning embeddings in the deeper layer that place the same classes close together. Hence, it can learn semantic similarity effectively.

II. Data

A. Data Sources and Data Annotation

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

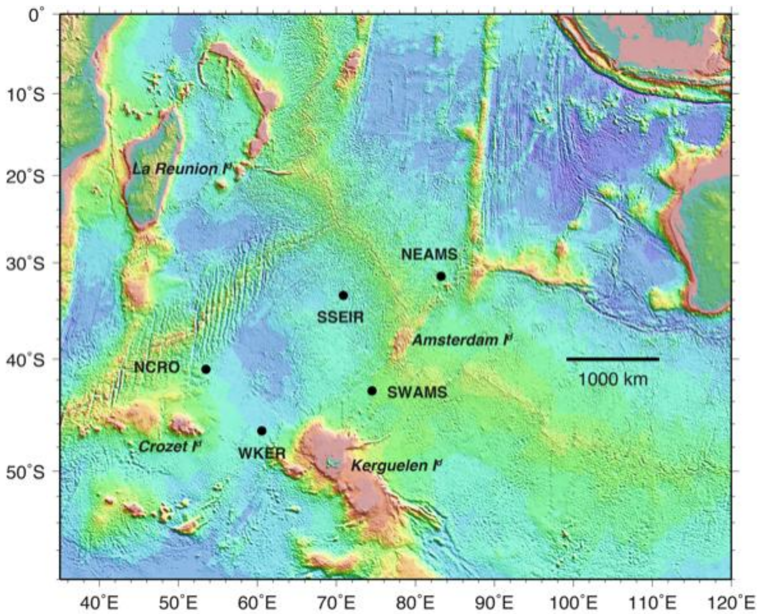

FIG. 1. Map of the southern Indian Ocean. Black dots represent moorings of the OHASISIBIO hydrophone network from which data were used in this paper: north of Crozet archipelago (NCRO); west of Kerguelen Island (WKER); southwest and northeast of St Paul and Amsterdam islands (SWAMS and NEAMS); south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

TABLE I: Manually annotated acoustic data from 5 mooring sites for four populations of blue 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	_

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

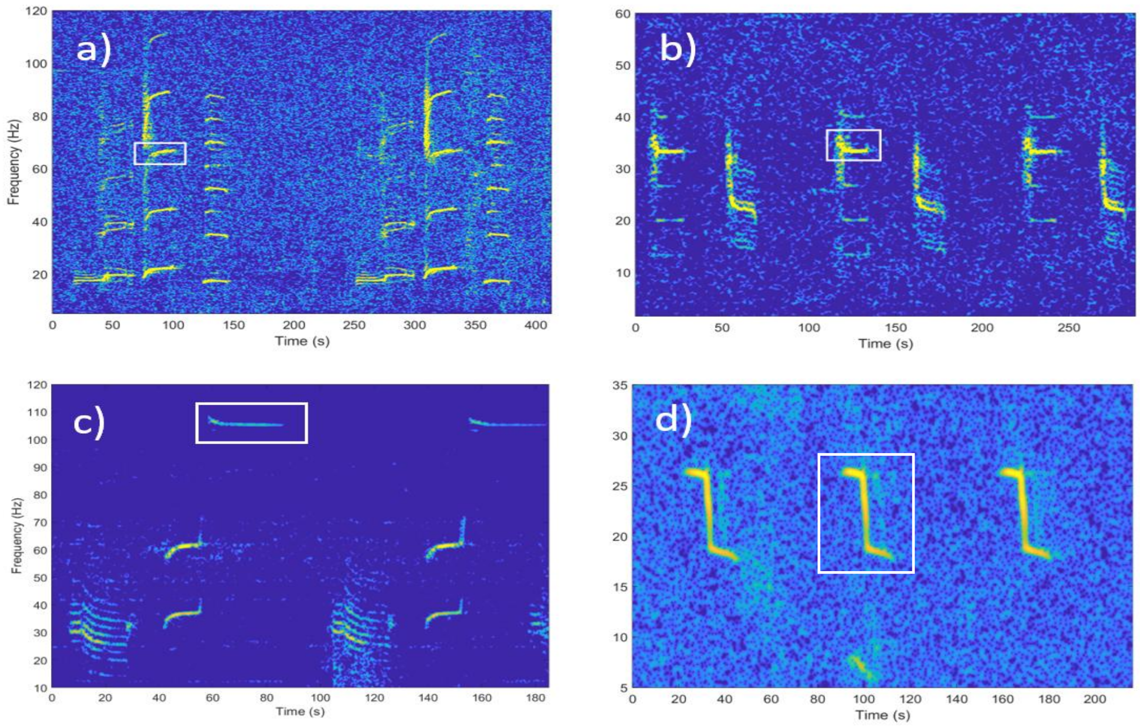

FIG. 2. Examples of annotated blue whale call with Raven Pro 1.5. a) SEIO pygmy blue whales, b) SWIO pygmy blue whales, c) CIO blue whales and d) Antarctic blue whales.

B. *Data for modeling*

For all four acoustic populations of blue whales, calls range from 6 to 40 seconds duration. Using custom written scripts in Python 3.6, spectrograms were produced from audio files (with NFFT = 1024 and 75% overlap, Hanning window). Each spectrogram was generated from a 240-s audio segment that contained either one or multiple annotated blue whale calls and was resized as 224 pixels by 224 pixels with RGB channels (Fig.3). During the annotation process, we only focused on the presence of one blue whale population in each acoustic file. However, as part of the temporal and geographical distributions overlap among these blue whale populations, their acoustic co-occurrence is common. As a result, for each extracted spectrogram, its corresponding label (the name of blue whale population, and the number of calls associated with the spectrogram) only represented the presence of that particular population but did not indicate absence of the other three populations.

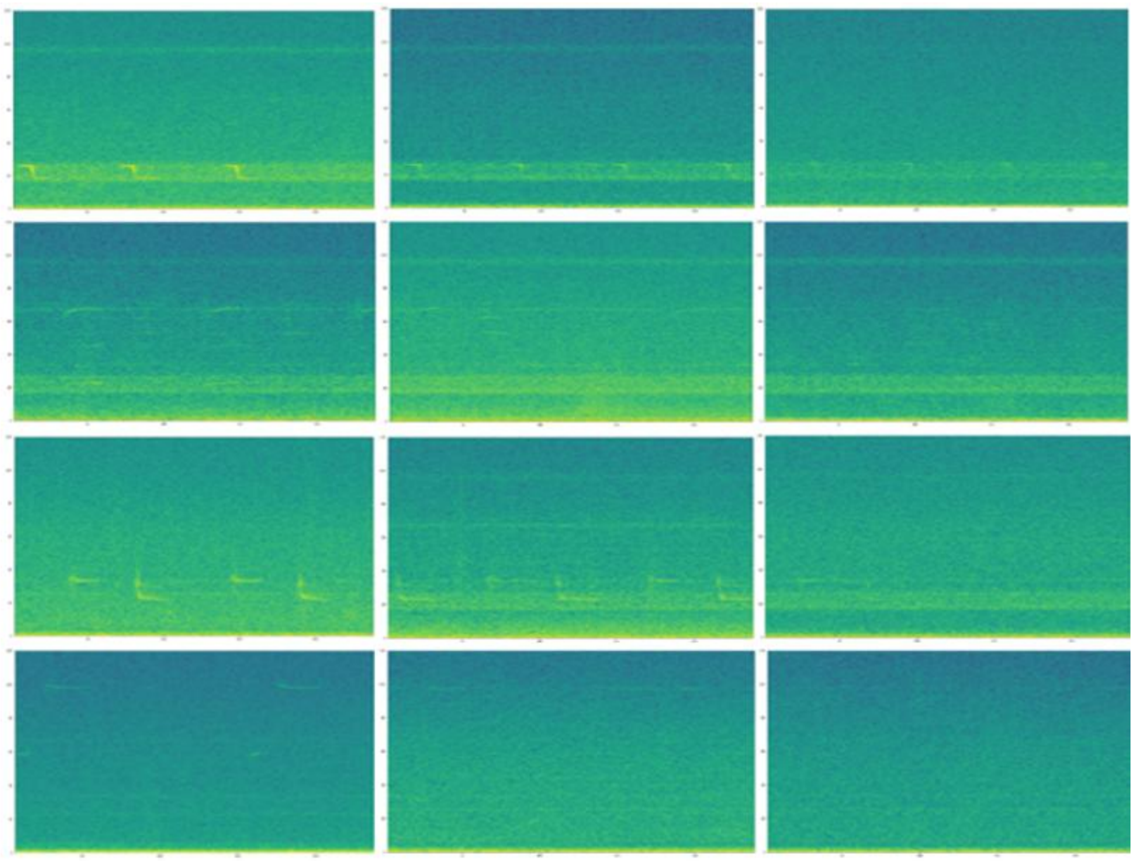

FIG. 3. Example spectrograms from different acoustic populations of blue whale in the Indian Ocean that illustrate the range of signal-to-noise ratios in the data from loudest to faintest from left to right. Row 1: Antarctic; row 2: SEIO; row 3: SWIO; row 4: CIO.

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

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

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

III. Approaches

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

A. Classification Models using Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNN) have been widely used for image classification tasks, and their success has also been proven in bioacoustic classification applications (Bianco *et al*. 2019). Here we used the DenseNet-201 architecture (Huang *et al.* 2016) as a baseline to classify calls of the four blue whale populations, and to count the number of calls in each 240-s spectrogram. DenseNet was developed specifically to improve the declined accuracy caused by the vanishing gradient in high-level neural networks and has the advantage of improving feature propagation both in forward as well as backward fashion. In a DenseNet architecture, each layer is connected to every other layer and obtains additional inputs from all preceding layers, and then passes its own feature-maps to all subsequent layers.

B. Classification Models using Siamese Neural Network (SNN)

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

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

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

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

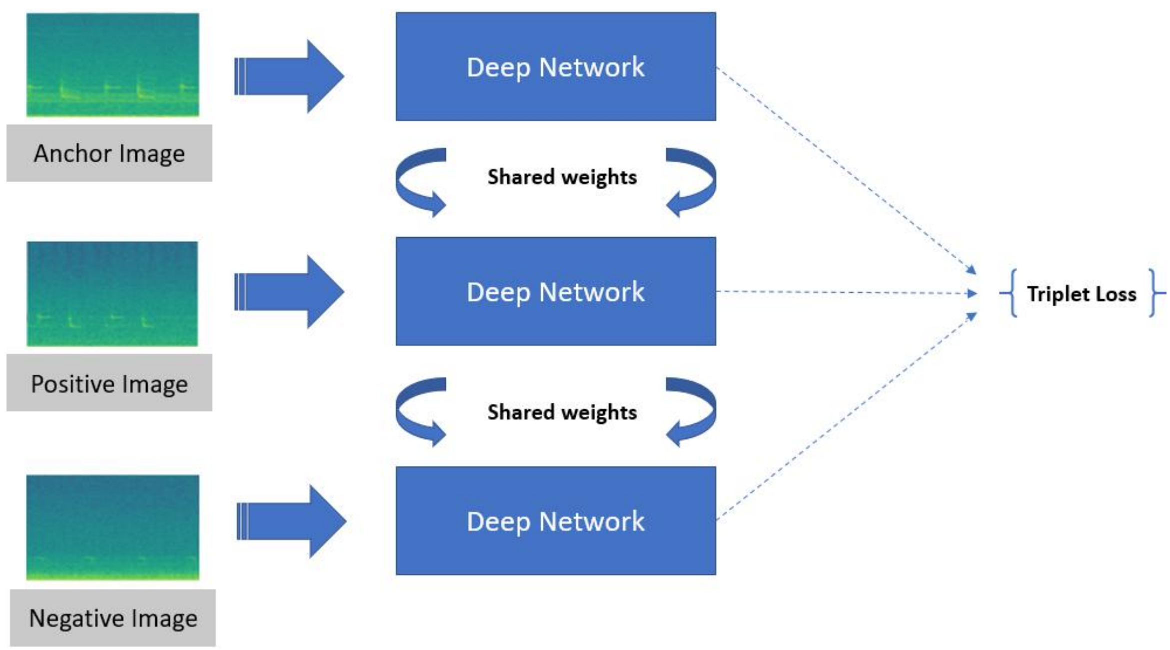

FIG. 4. Architecture of Siamese Networks with triplet loss.

When triplets are generated for model training, as the training continues, some of the additional triplets are easy to deal with (their loss value is very small or even 0), preventing the network from further improvement. A good training strategy would be to constantly “mine” out those difficult cases in each epoch, based on the current performance of model’s snapshot, so that the model will always have certain percentage of hard cases in the training loop from which it still struggles to tell a difference. This is similar to the triplet mining in FaceNet (Schroff *et al.* 2015). In our training process, we choose batch size = 5. Within each batch, we first generated 5 triplets randomly and kept the 2 hardest examples, and then generated another 3 triplets randomly.

C. Implementation

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

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

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

IV. Results

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

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

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

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

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

	CNN			SNN		
Population	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**

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

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

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%

C. Further comparisons of two models

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

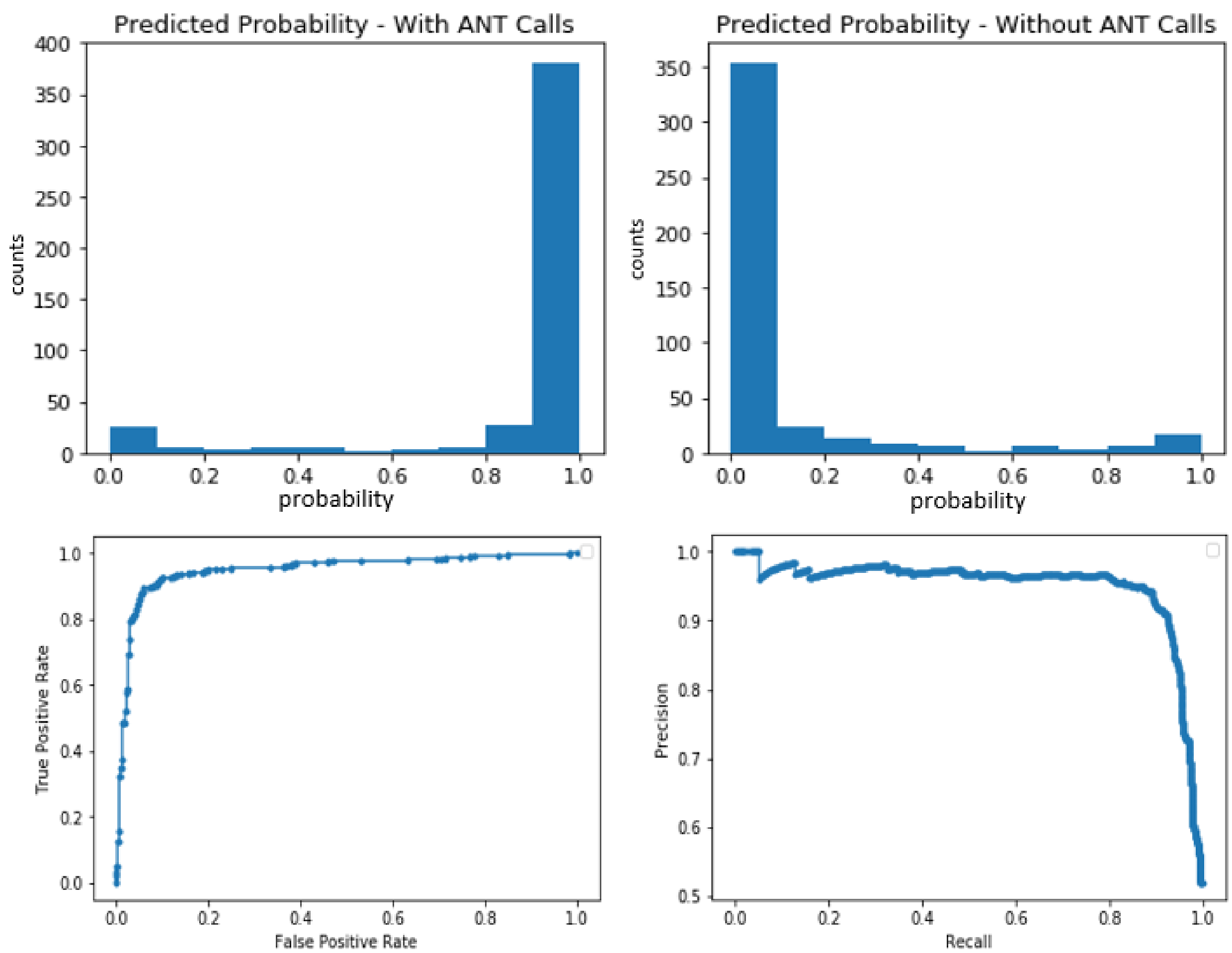

FIG. 5. Illustraiont of the results of call classification task by CNN. Top left and top right: Histograms for predicted probabilities of positive and negative samples in the testing set. Bottom left: receiver operating characteristic (ROC) curve. Bottom right: precision-recall curve.

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

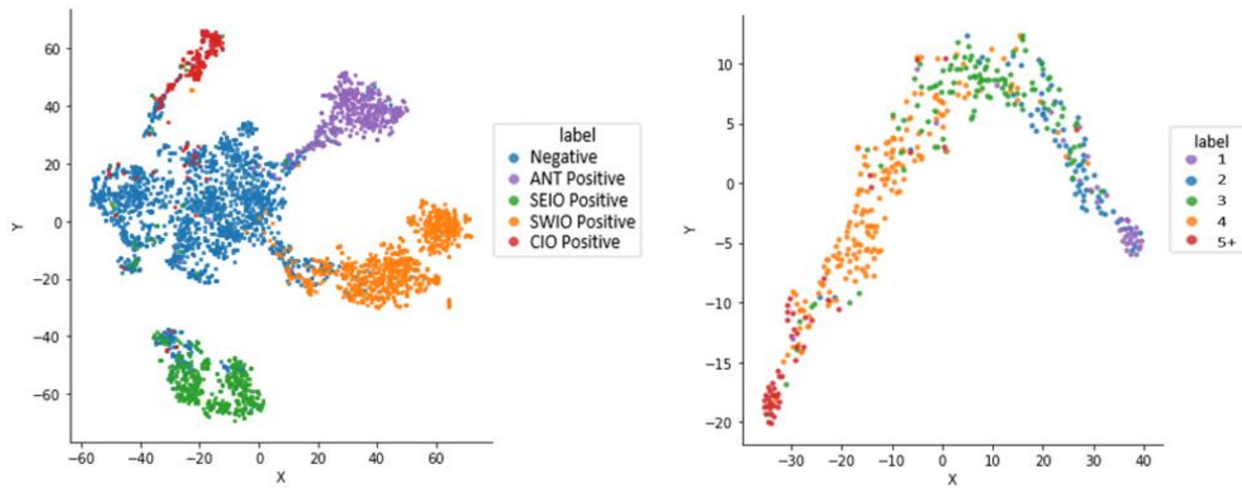

FIG. 6. (a) t-SNE plot for the model that classifies the presence or absence of blue whale calls from each of the four populations. (b) t-SNE plot for the model that estimates the number of Antarctic blue whale calls (for the other three populations, the plots show similar patterns).

V. Discussion

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

While Siamese Networks are particularly suitable for scenarios where there are only a few samples in each class (i.e., few-shot learning), they can also be applied to larger datasets, like the one we used in this study. However, since Siamese Networks learns from quadratic pairs (to make use of all information available), the training is much slower than pointwise learning models such as CNN. Additionally, instead of outputting probabilities of the prediction, they output the distance from closest training samples in each class instead. In practice, CNN and SNN can be used together to complement each other. Given that the learning mechanism of SNN is somewhat different from CNN, their ensembled results are likely to perform even better.

While both models performed well in general on classifying calls from 4 populations of blue whales, their performance differed among different populations. Classification of Antarctic blue whale calls had the highest accuracy among 4 populations, while CIO had the lowest accuracy. One possible reason is that Antarctic, SEIO and SWIO have larger sizes of training samples compared to CIO, but more likely is that Antarctic blue whale calls (Z-calls) have more frequency modulation on the spectrograms, compared to that of the CIO blue whale calls (which looks like a flat line). Another factor is the call loudness in the audio recordings. In general, CIO blue whale calls have lower signal-to-noise ratios in the annotated data, which increases the difficulty for the model to classify correctly with high confidence. The lower signal-to-noise ratios for CIO blue whale calls could be due to a number of factors among which we cannot currently distinguish. These include the CIO call having a lower source level than other calls; there are only a few source levels reported for blue whale signals globally, and none for CIO blue whale calls. It is also likely that the animals producing these signals are further from the hydrophones than the other populations, given what is known about their distributions, although since the hydrophones are omni-directional we cannot ascertain this for certain. This signal is the highest frequency signal we detected and as such would be subject to greater transmission loss than the other signals.

Compared to traditional methods which rely heavily on manual verification by a human user or template matching by software, the method presented here uses deep learning models and has the advantage of flexibility with regards to temporal and frequency variations in a dataset. Notably for blue whale calls, the call frequency has been getting lower in all populations over time (McDonald et al. 2009, Leroy et al. 2018), and one major advantage of this approach is that it looks for the shape of the call independent of the frequency of the call. Siamese Networks can easily classify and count multiple types of calls from several populations at the same time and have the ability to classify novel datasets that were collected from different mooring sites or different years. Even at the sites that have somewhat different underwater environments, the model still detected and classified the signals. An additional, and future advantage is that the model can easily scale up to include other species or call types with the addition of annotated data.

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

Acknowledgements

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

References

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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, 607.

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

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

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

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

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

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

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

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

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

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