Automatic detection and classification of bearded seal vocalizations in the northeastern Chukchi Sea using convolutional neural networks

Christian. D. Escobar-Amado, Mohsen. Badiey and Sean. Pecknold

Citation: The Journal of the Acoustical Society of America **151**, 299 (2022); doi: 10.1121/10.0009256
View online: https://doi.org/10.1121/10.0009256
View Table of Contents: https://asa.scitation.org/toc/jas/151/1
Published by the Acoustical Society of America

JASA ARTICLE

Automatic detection and classification of bearded seal vocalizations in the northeastern Chukchi Sea using convolutional neural networks[a)]

Christian. D. Escobar-Amado,[1,b)] Mohsen. Badiey,[1,c)] and Sean. Pecknold[2]
[1]*Department of Electrical Engineering, University of Delaware, Newark, Delaware 19716, USA*
[2]*Defence Research and Development Canada, Dartmouth, Nova Scotia, B3K 5X5, Nova Scotia, Canada*

ABSTRACT:
Bearded seals vocalizations are often analyzed manually or by using automatic detections that are manually validated. In this work, an automatic detection and classification system (DCS) based on convolutional neural networks (CNNs) is proposed. Bearded seal sounds were year-round recorded by four spatially separated receivers on the Chukchi Continental Slope in Alaska in 2016–2017. The DCS is divided in two sections. First, regions of interest (ROI) containing possible bearded seal vocalizations are found by using the two-dimensional normalized cross correlation of the measured spectrogram and a representative template of two main calls of interest. Second, CNNs are used to validate and classify the ROIs among several possible classes. The CNNs are trained on 80% of the ROIs manually labeled from one of the four spatially separated recorders. When validating on the remaining 20%, the CNNs show an accuracy above 95.5%. To assess the generalization performance of the networks, the CNNs are tested on the remaining recorders, located at different positions, with a precision above 89.2% for the main class of the two types of calls. The proposed technique reduces the laborious task of manual inspection prone to inconstant bias and possible errors in detections. https://doi.org/10.1121/10.0009256

(Received 28 July 2021; revised 14 November 2021; accepted 13 December 2021; published online 19 January 2022)
[Editor: James F. Lynch] Pages: 299–309

I. INTRODUCTION

Passive acoustic monitoring (PAM) has become a feasible method for investigating marine mammal activity over large spatial and temporal scales.[1] The large amount of acoustic data collected in PAM recorders allows researchers to analyze the seasonal and geographical variability of the vocal behavior of certain marine mammals, which can be used to measure their presence, density, and distribution across large areas. Bearded seals, in particular, are highly vocally active mammals and are one of the main contributors to the marine soundscape in the Arctic, especially during the mating season.[2] In the last decades, extensive studies of bearded seal vocalizations have been possible thanks to the large acoustic datasets collected *via* PAM.[1–11] Commonly, bearded seal calls are manually identified by trained analysts by visualizing the spectrogram representation of the audio recordings.[1,3–9] However, as the acoustic datasets grow larger, this manual inspection becomes a long and laborious task that might lead to an inconstant bias dependent on the degree of fatigue of the analyst.[12,13] Therefore, automatic detection systems become necessary.

Recently, some techniques for automatically detecting bearded seal vocalizations have been implemented. Hannay *et al.*[9] used time-frequency contours of normalized spectrograms for classifying the presence of a bearded seal call based on several extracted representative features. Halliday *et al.*[10,11] and Heimrich *et al.*[2] used the algorithm proposed in Ref. 14 for detecting acoustic events and classifying them among bearded seals and other mammals such as beluga whales, bowhead whales, and walruses using a random forest; then, all files with at least one automatic detection along with 5%–10% of files with no detections were manually analyzed.

Other detection and classification systems (DCSs) are commonly based on spectrogram correlation[13,15,16] and contour detectors,[9,12,15,17] and have been implemented for identifying other marine mammals such as ringed seals,[16] and a wide variety of whale species.[9,12,13,15,17] However, one common issue often present in detection systems is the need for setting an adequate threshold to balance the rates of false positives and true positives. When a system is too sensitive then it is more likely to not miss a signal of interest (true positives), however, undesired signals corresponding to noise generated by other sound sources are also detected (false positives).

To alleviate this trade-off and improve the performance of the detectors, recently, machine and deep learning approaches have been adopted for DCS, especially for identifying whale calls. Some efforts in the implementation of machine learning (ML) include the use of artificial neural networks,[13,18] logistic regression classifiers,[19] and a

[a)]This paper is part of a special issue on Ocean Acoustics in the Changing Arctic.
[b)]Electronic mail: escobarc@udel.edu, ORCID: 0000-0003-2907-7311.
[c)]ORCID: 0000-0002-5869-336X.

Boltzmann machine combined with a sparse auto-encoder.[20] Most recently, deep learning (DL) techniques have gained special interest for whale call detection and classification. Some DL approaches include convolutional,[21–25] residual,[26] recurrent,[24,27] and long short-term memory[27] neural networks.

In this work, we present an approach based on deep learning techniques for automatically detecting and classifying two main types of vocalizations of bearded seals year-round recorded at several positions in the northeastern edge of the Chukchi Continental Slope in 2016–2017. The center of the receivers was located at approximately 175 km northwest of point Barrow, Alaska, in between the 100 and 400 m isobath. The vocalizations are first detected using a two-dimensional (2D) spectrogram correlation method. To increase the number of true positive detections, a low threshold is selected; however, the false positives increase as well. To deal with this trade-off, these first detections are considered candidate regions containing potential bearded seal vocalizations. Then, convolutional neural networks (CNNs) are used for validating and classifying the signals detected by the 2D spectrogram correlation method. This study demonstrates the potential for convolutional neural networks to improve the performance of fast automated detectors applied to large PAM databases for analyzing the presence and abundance of bearded seal vocalizations.

This paper is structured as follows: Sec. II introduces the measured data used for this work. Section III presents the methodology of the DCS. Sections IV and V show the results and discussion, respectively, followed by the conclusions in Sec. VI.

II. MEASURED DATA

The data used in this paper were collected by several receivers deployed in the northeastern edge of the Chukchi Shelf, in the dynamic shelf break region during the Shallow Water Canada Basin Acoustic Propagation Experiment (SW-CANAPE) in 2016–2017.[28–30] Four autonomous multichannel acoustic recorder arrays were deployed by the Defense Research and Development Canada agency during the SW-CANAPE and are referred hereafter as A1, A2, A3, and A4. Figure 1 shows the location of the recorders overlaid on IBCAO v3.0 bathymetric data.[31]

Details about position and recording times for each receiver are listed in Table I. The receivers recorded from late October 2016 until mid October 2017. The dataset is composed of 32-min recordings starting at 00 h (for anthropogenic noise reception) and 3.5-min recordings starting at 02 h (for ambient noise measurement) every four hours for the duration of the experiment.

One challenge for the marine mammal detections in the SW CANAPE experiment is the fact that anthropogenic noise was present most of the time in the four recordings. As shown in Fig. 1(b), the Naval Research Laboratory deployed two sound sources, S1 and S2, close to A3 and A1, respectively. These sound sources transmitted Linear Frequency Modulated (LFM) signals as upsweep from S1 and down-sweep from S2 at the frequency bands of 700–1100 Hz and 1.5–4 KHz. Also, M-Sequences were transmitted at the same frequency bands of the LFM signals, as well as Continuous Wave (CW) signals at 890 Hz from S1 and 910 Hz from S2. Each source broadcast a 30-min sequence of transmissions approximately every four hours every day starting from October 24, 2016, for both sources and ending on June 26, 2017, for S1 and September 1, 2017, for S2.

III. METHODOLOGY

The detection and classification system (DCS) proposed in this work for detecting several bearded seals vocalizations is divided into two main sections, as shown in Fig. 2. The first step is the detection, where several candidate regions in the spectrograms of the audio recordings are selected based on a representative template of the vocalization. When the fast two-dimensional normalized cross correlation[32] (2D-NCC) between the template and the spectrogram is greater

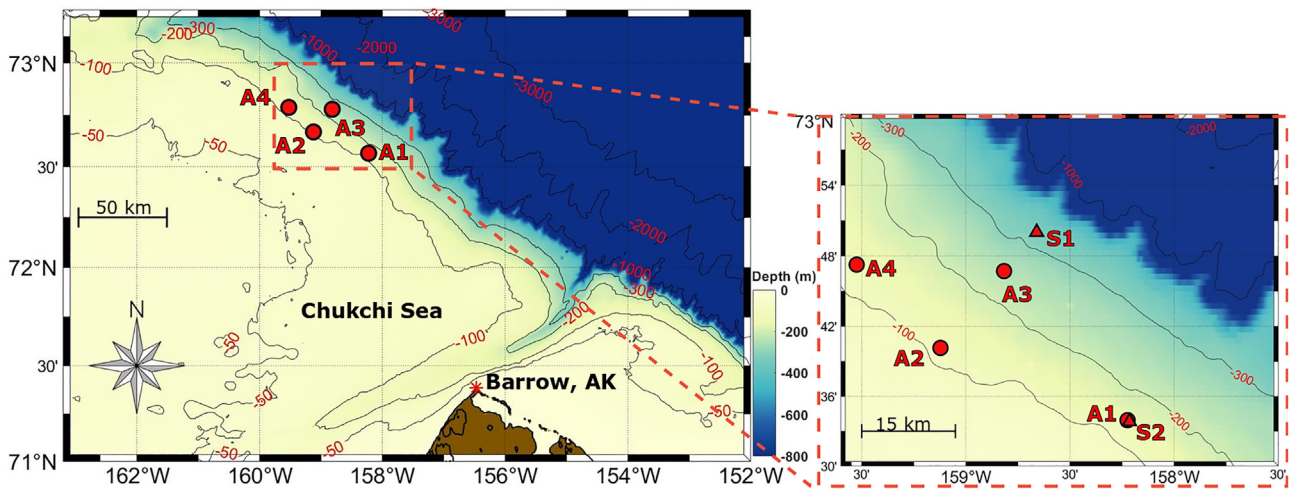

FIG. 1. (Color online) (a) Wide area map showing the locations of A1, A2, A3 and, A4 deployed in the Chukchi continental slope. (b) Close-up of the locations of the receivers relative to the sound sources. Maps are overlaid on IBCAO v3.0 bathymetric data.

https://doi.org/10.1121/10.0009256

TABLE I. Details of position, recording times, and water depths of the receivers.

Recorder	Lat(N)	Lon(W)	Water depth (m)	Recorded dates
A1	72.566	158.223	149	10/21/2016–10/12/2017
A2	72.669	159.122	106	10/21/2016–10/12/2017
A3	72.779	158.817	224	10/19/2016–10/10/2017
A4	72.788	159.524	123	10/24/2016–10/12/2017

or equal than the threshold, then that position represents a candidate region, also known as region of interest (ROI). Second, for the classification task, the portion of the spectrogram—selected in the first step—is input to a CNN that classifies the ROI as either noise or one type of bearded seal vocalization. These deep learning models are trained and validated using data collected and labeled from A4 and tested on the remaining three receivers (see Fig. 1).

A. Fast 2D normalized cross correlation

Studies of vocal repertoire of bearded seals from Alaska (AL) have categorized their calls into trills (T), moans (M), and ascents (A).[1,3,4] Some examples of these vocalizations are shown in Fig. 3, where the trills are labeled as AL1(T), AL1i(T), AL2(T), AL4(T), AL5(T), and AL6(T). Moans and ascents are labeled as AL3(M) and AL7(A), respectively. More details about the labeling and characteristics of the different types of calls can be found in Risch *et al.*[3] The scope of this work is to identify two representative vocalizations of the bearded seal repertoire that fall into several types of trills.

The first step of this DCS is to obtain the ROIs from the spectrograms that are extracted with a 40 ms time interval and a frequency step of 3.9 Hz. For this detection part of the system, two templates, also known as masks, are extracted from the recordings and represent the two vocalizations of interest. The first vocalization, denoted as Mask 1, is a trill with both ascending and descending components as shown in Fig. 4(a), and corresponds to AL1i(T) (see Fig. 3). Mask 1 is 12.28 s long with a bandwidth of 1840 Hz (328–2172 Hz). Mask 2 is the long downsweep call shown in Fig. 4(b), which is 26.44 s long and has a bandwidth of 1199 Hz (324–1523 Hz). These two templates in Figs. 4(a) and 4(b) are used for detecting the ROIs in the recordings using a spectrogram correlation technique based on the fast normalized 2D cross correlation method described in Ref. 32, where each template is slid in frequency and time across the spectrogram. The correlation value at each point between the template and the portion of the spectrogram under the template is computed as,

$$\gamma(u,v)=\frac{\sum_{t,f}\left[S(t,f)-\bar{S}_{t_T,f_T}\right]\left[T(t-t_T,f-f_T)-\bar{T}\right]}{\left\{\sum_{t,f}\left[S(t,f)-\bar{S}_{t_T,f_T}\right]^2\sum_{t,f}\left[T(t-t_T,f-f_T)-\bar{T}\right]^2\right\}^{0.5}}, \quad (1)$$

where S is the input spectrogram and the sums are over time t and frequency f under the window containing the template T positioned at t_T, f_T. And, $\bar{T}$ is the mean of the template while $\bar{S}_{t_T,f_T}$ is the mean of S in the region under the template.

The output of the 2D-NCC for Mask 1 and Mask 2 applied to the spectrogram in Fig. 4(e) is shown in Figs. 4(c) and 4(d) where Eq. (1) is computed by sliding a window across the spectrogram using a stride of one in both axes with no padding. The blank part of Figs. 4(c) and 4(d) represent areas where the mask overlap with the underline data does not have a value. To detect the regions, the spectrograms are divided into a grid where each portion has the same size of the template as shown by the white dashed lines in Figs. 4(c) and 4(d). When the maximum correlation value inside each region of the grid exceeds a threshold, then that point indicates the starting time and frequency (t_T, f_T) of a ROI as shown by the white circle markers.

Some examples of bearded seal vocalizations detected using this method are shown in Fig. 4(e). Detections using Masks 1 and 2 are shown in black and white rectangles, respectively. These rectangles encompass the ROIs where the white circle markers indicate the starting frequency and time, and are located at the same position as the white markers shown in Figs. 4(c) and 4(d).

The detections in Fig. 4(e) lead to several observations about the behavior of the 2D-NCC. First, when the threshold for Mask 1 is low, the system starts detecting AL1(T) calls, which are found at higher frequencies and have a behavior similar to AL1i(T) in the first 12 s but with a longer downsweep component. With a lower correlation value, Mask 1 also detects downsweep calls that have a slope similar to that of the template. Mask 2, on the other hand, is used for detecting long downsweep calls corresponding to signals such as AL2(T), AL2i(T), and the long descent portion of AL1(T). With a lower correlation threshold, Mask 2 starts detecting short downsweep calls such as AL5(T) and the descent portion of Mask 1, i.e., AL1i(T). However, one challenge in this dataset is that with a low threshold, the detection system is triggered by anthropogenic signals such as the M-sequences when using both masks as shown in Fig. 4(f), which increases the rate of false positives.

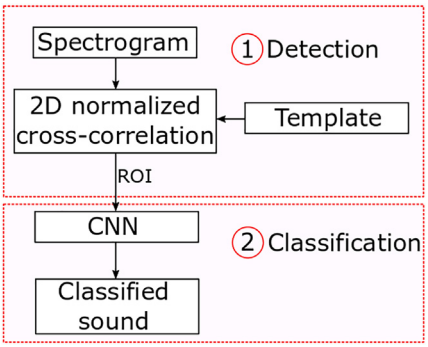

FIG. 2. (Color online) Detection and classification system diagram.

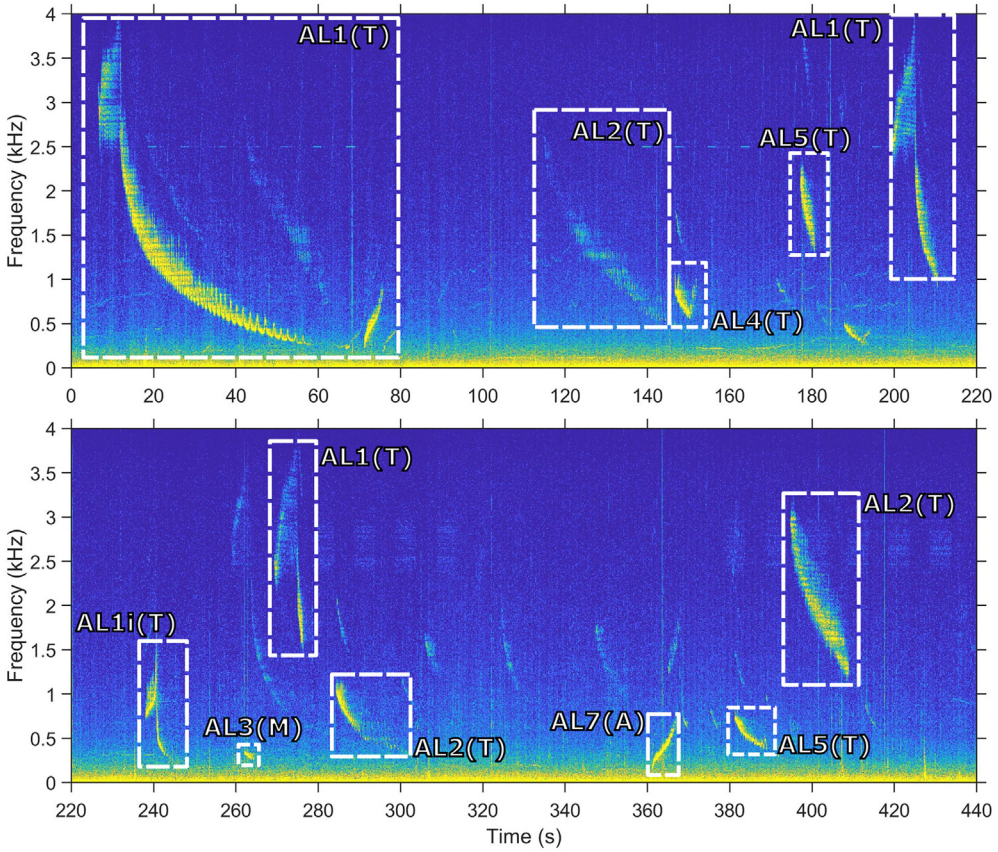

FIG. 3. (Color online) Representative examples of bearded seal vocalization repertoire. Data measured on A4 on April 25, 2017 from 20:08 until 20:15.

B. Data labeling

The selection of an adequate threshold to determine the system sensitivity and balance the rate of false positives and true positives is always a challenge in detection systems. To analyze the impact of the threshold selection, all the ROIs identified in one of the receivers have been manually labeled for both templates using a low detection threshold. Here, A4 is chosen for this task given that it is the most benign case since it is the recorder that is farthest from the sound sources S1 and S2.

For each mask, the data have been labeled among four classes as shown in Table II. *Class 0* corresponds to anthropogenic noise (M-Sequences), *Class 1* corresponds to the main sound the mask is intended for; Class 2 is background noise; and *Class 3* is a bearded seal sound similar to that of the main Class 1.

For Mask 1, Fig. 5(a) shows the detected ROIs as a function of geotime versus starting frequency where *Class 1*, i.e., AL1i(T) trills, are found at three distinguishable starting frequency bands as depicted by the green dots. On the other hand, AL1(T) trills are found at higher frequencies above 1.5 kHz. *Class 2* is spread over the entire frequency band and mostly happens in the same time frame as AL1i(T) and AL1(T) trills, mainly because the threshold is commonly triggered by other downsweep signals from the same species. In Fig. 5(b) it is observed that as the 2D-NCC threshold decreases, the rate of true positives increases exponentially; however, the rate of false positives increases with a steeper exponential behavior. The accuracy of the detection system [dashed magenta line in Fig. 5(b)] is above 80% with thresholds higher than 0.41 but decreases rapidly as the threshold is lowered.

Similarly, for Mask 2, the ROIs have been divided into the four classes shown in Table II. The long downsweep trills are commonly found at starting frequencies below 1.5 kHz as shown in Fig. 5(c). Furthermore, Mask 2 has a lower ratio of false positives than Mask 1 when decreasing the correlation threshold as observed in Fig. 5(d). This is due to the fact that Mask 2 has a longer duration and spans a broader number of vocalizations, i.e., any trill that has a long downsweep component falls into *class 1*. However, the false and true positive rates trade-off is still present with a detection accuracy of 50% with a 0.2 threshold.

When labeling the data, if two or more classes are present in the same ROI, then the priority order for both masks is *class 1*, *class 3*, *class 0*, and *class 2*. For instance, for Mask 1, if an M-Sequence and an AL1i(T) sound are present in the same detected region then that ROI is labeled as *class 1*.

C. CNN architecture

To deal with the trade-off between false and true positive rates due to the correlation threshold, convolutional neural networks are implemented for classifying the ROIs

https://doi.org/10.1121/10.0009256

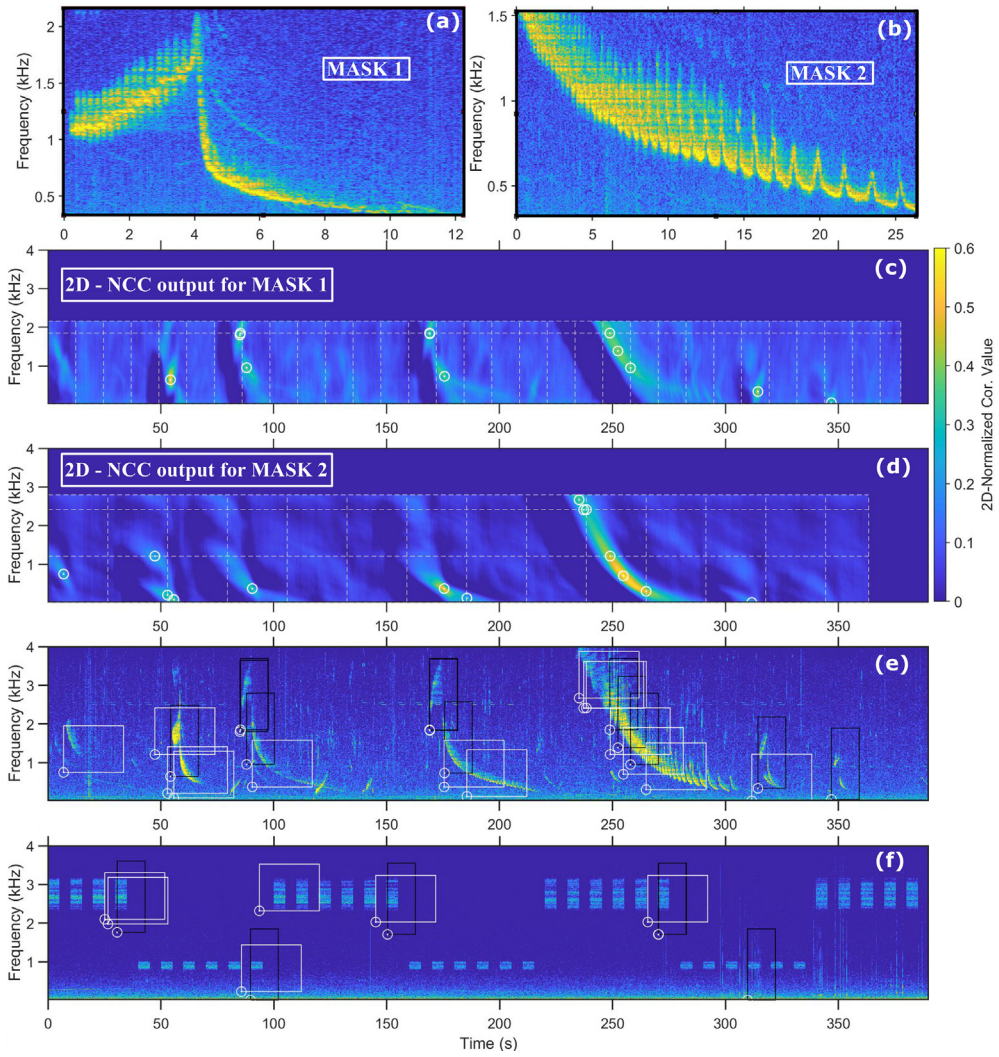

FIG. 4. (Color online) (a) Mask 1 for AL1i(T) sound. (b) Mask 2 for long down-sweep vocalization. (c), (d) Output from the 2D Normalized cross correlation of Mask 1 and Mask2. (e) ROI encompassing bearded seal vocalizations detected by the spectrogram correlation method. (f) ROI encompassing M-Sequences which correspond to false positive detections using the 2D-NCC.

detected during the spectrogram correlation step. CNNs are a deep learning tool used for learning representative features in the convolutional layers by sliding kernels across the data.[33]

The CNN topology used in this work is inspired by previous CNNs developed for classification tasks using acoustic signals.[34,35] The network architecture is shown in Table III where the parameters for each convolutional layer are presented as (kernel size), (stride), (padding), and (Number of channel outputs). The number of channels represents the number of feature maps learned at each layer when sliding the kernel or filter across the input. Using large kernels in the first layers but not in the deeper ones helped the networks to learn, similar to what was found in previous works.[34,35]

The deep learning algorithms used in this study were written in Python using the PyTorch framework.[36] The CNN is composed of five convolutional layers and one fully

TABLE II. Classes for Masks 1 and 2.

Mask 1		Mask 2	
Class	Description	Class	Description
0	M-Sequence	0	M-Sequence
1	AL1i(T)	1	Long downsweep vocalizations such as AL2(T), AL2i(T), and the long descent part of AL1(T)
2	Any sound different to classes 0, 1 or 3	2	Any sound different to classes 0, 1 or 3
3	AL1(T)	3	Short downsweep calls such as AL4(T), AL5(T), and the descent part of AL1i(T)

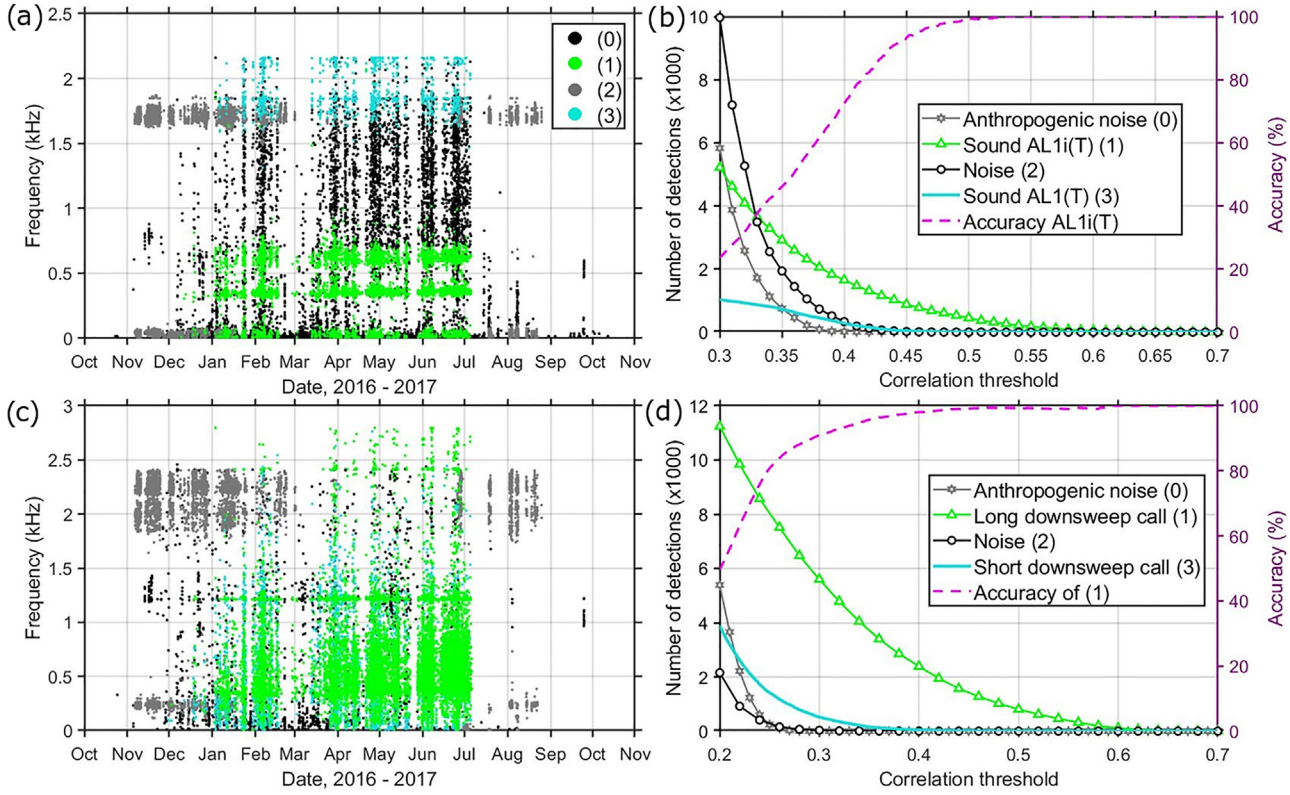

FIG. 5. (Color online) (a), (c) ROIs detected and labeled on A4 for (a) Mask 1 and (c) Mask 2 as a function of starting frequency vs geo-time. (b), (d) Number of detections as a function of the correlation threshold for (b) Mask 1 and (d) Mask 2.

connected (FC) layer with 2000 neurons followed by an output layer that assigns a score to each of the 4 classes—one associated with each class defined in Table II. Each layer is followed by a ReLu activation function[37] that is a computationally efficient way to include a sparse activation into the problem. Batch normalization is employed after ReLu for each convolutional layer to add a level of regularization and reduce the generalization error.[38] Max pooling is applied after the first two convolutional layers to decrease the dimensionality of the network. A 50% dropout is used before the FC layer for avoiding overfitting. The networks are trained on 60 epochs with a batch size of 32 using the Adam optimizer[39] with a learning rate scheduler.

TABLE III. CNN architecture. The convolutional layer parameters are presented as (kernel size), (stride), (padding), and (Number of channel outputs). The value next to FC corresponds to the number of neurons in that FC layer.

Layer	CNN Parameters
conv	$(11 \times 7)(3 \times 2)(5 \times 3)$ **(32)**
max pool	$(5 \times 5)(2 \times 2)(2 \times 2)$
conv	$(5 \times 5)(1 \times 1)(2 \times 2)$ **(64)**
max pool	$(5 \times 5)(1 \times 2)(2 \times 2)$
conv	$(5 \times 3)(2 \times 1)(2 \times 1)$ (128)
conv	$(3 \times 3)(1 \times 2)(1 \times 1)$ **(64)**
conv	$(5 \times 3)(2 \times 1)(1 \times 1)$ **(32)**
	Vectorization and dropout
	FC-2000
	Output (4 classes)

Using the data previously labeled for A4, two independent networks, named CNN1 and CNN2, are trained to reinforce the detections found with the 2D-NCC using Mask 1 and Mask 2, respectively. The two networks share the same CNN architecture presented in Table III for classifying the ROIs among the four possible classes described in Sec. III B for the two types of vocalizations of interest.

IV. RESULTS

Results of the DCS for the two types of bearded seal vocalizations of interest are presented in this section. The 2D-NCC technique described in Sec. III A has been applied to the four receivers detecting ROIs across the entire dataset. The networks have been trained and validated using the labeled data on A4. However, due to the close proximity of A1 with respect to the sound source S2, several downsweep LFM signals were being falsely classified (for A1 only) as long and short downsweep bearded seal vocalizations, i.e., as *class 1* and *class 3* for Mask 2 when using CNN2. Since A4 is far from S2, LFM signals were not detected as ROIs in the spectrogram correlation step and therefore they were not included originally in the training. To address this issue, a small portion of labeled ROIs from A1 containing LFM signals were included in the training of CNN2.

The metrics used for evaluating the performance of the networks during the training/validation stage are accuracy, precision, and recall. Accuracy is calculated by counting the number of times the CNN predicted the correct class.

https://doi.org/10.1121/10.0009256

TABLE IV. Number of ROIs detected on the six receivers using the 2D-NCC on the spectrograms.

Recorder	# ROIs MASK 1	# ROIs MASK 2	Daily recorded hours	Total Recorded hours
A1	79 920	1 27 441	3.58	1273.9
A2	35 064	29 495	3.58	1274.5
A3	56 637	61 076	3.58	1274.5
A4	22 056	22 683	3.58	1273.9
			Total	**5096.8**

Precision is the ratio between correctly predicted observations for a given class and the total of predicted observations for that class, and it is defined as

$$Precision = \frac{\#True\ positives}{\#True\ positives + \#False\ positives}. \quad (2)$$

Recall is the ratio between correctly predicted observations and the total number of observations of a given class and it is defined as

$$Recall = \frac{\#True\ positives}{\#True\ positives + \#False\ negatives}, \quad (3)$$

where a high precision index signifies a low false alarm rate and a high recall index indicates a high detection efficiency.[17]

Furthermore, for assessing the generalizability performance of the DCS, the trained networks are applied to the non-labeled ROIs detected on the three remaining receivers located at different positions.

A. Detection results

The detection system described in Sec. III A has been applied to the four receivers. The number of ROIs detected per receiver is shown in Table IV where A1 and A3 have the larger counts of candidate regions. The reason behind this is the proximity of A1 and A3 to the sound sources S2 and S1, respectively, where the anthropogenic signals are present with high SNR for most of the recording time, especially for A1.

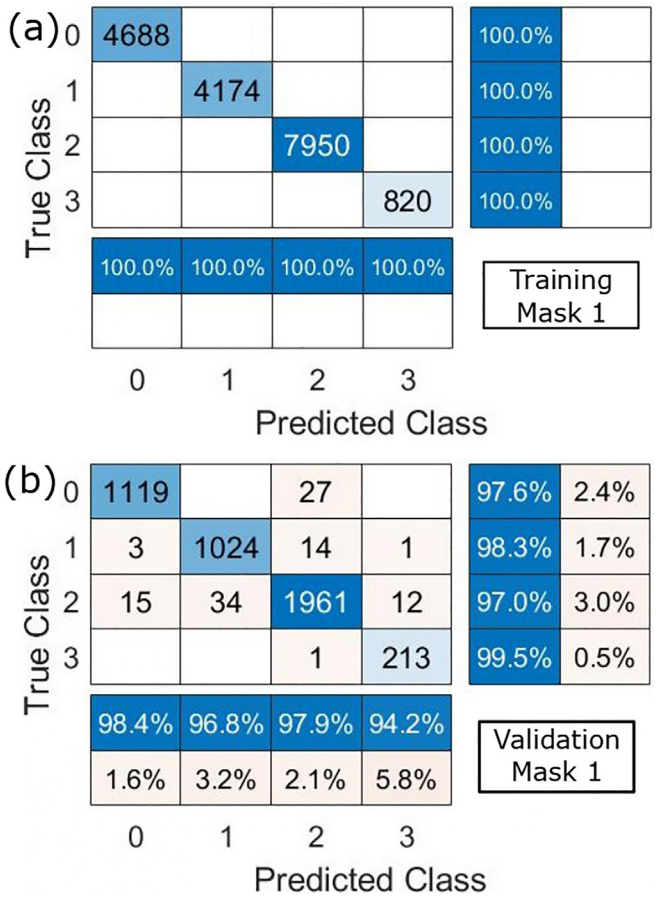

FIG. 6. (Color online) Confusion matrices for (a) training and (b) validation of ROIs detected with Mask 1. Precision for each class is presented by the two rows at the bottom. Recall is presented by the two columns at the right-hand side.

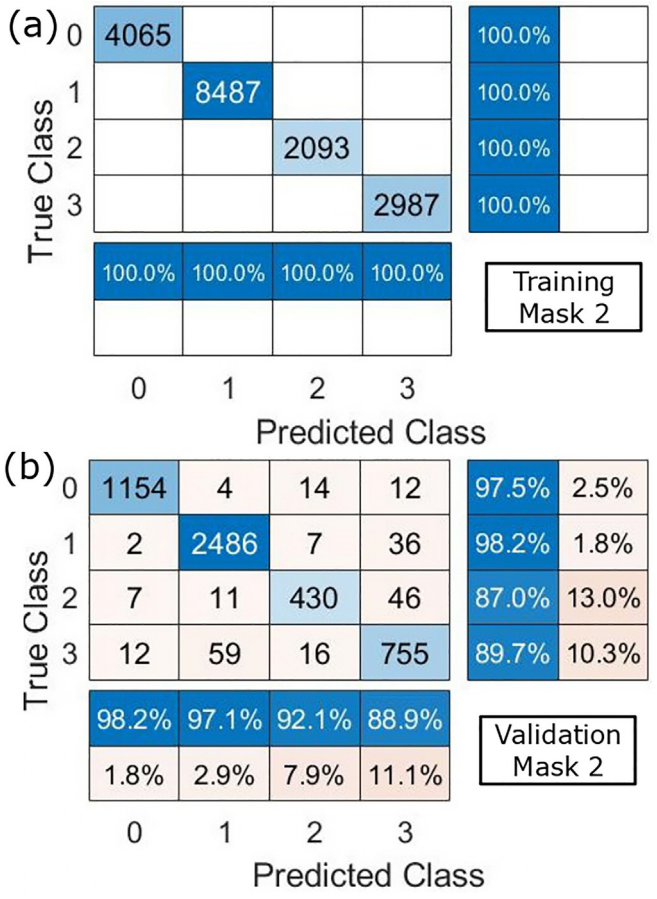

FIG. 7. (Color online) Confusion matrices for (a) training and (b) validation of ROIs detected with Mask 2. Precision for each class is presented by the two rows at the bottom. Recall is presented by the two columns at the right-hand side.

B. Training and validation results

CNN1 was trained and validated with an 80/20% random split using the 22 056 ROIs identified and labeled on A4 when correlating the spectrograms with Mask 1. The confusion matrix in Fig. 6(a) shows that the network had an accuracy of 100% in the training stage. For the validation stage, the overall accuracy was 97.6% with precision and recall values for *class 1* of 96.8% and 98.3%, respectively, as shown in the confusion matrix in Fig. 6(b).

CNN2, with an 80/20% random split, was trained and validated using 22 683 ROIs labeled on A4 plus 616 ROIs containing LFM signals from A1 labeled as *class 2*. Training and validation accuracies were 100% and 95.52%, respectively, as shown in Figs. 7(a) and 7(b). Precision and recall for *class 1* were 97.1% and 98.2% respectively. *Class 2* and *class 3* had lower recall and precision performance but still above 87%.

These validation results show that both CNNs have learned representative patterns from the labeled ROIs previously identified with the spectrogram correlation technique.

C. Generalization results

The ROIs classified by CNN1 for A1, A2, A3, and the validation portion of A4 are shown in Fig. 8 as a function of geotime vs starting frequency (for Mask 1) where the colors represent the predicted class. The behavior of the classifications is similar to the labeled signals shown in Fig. 5(a), where AL1i(T) sounds, i.e., *class 1*, are found at the same three starting frequency bands and AL1(T) sounds are found above 1.5 kHz. Furthermore, the frequency band of the classified M-Sequences (*class 0*), represented by the dark gray dot markers, match their actual transmitted frequency. This agreement in frequency and time of occurrence between the predicted classes and the labeled data provides evidence that the ROIs are correctly classified. To further prove this statement, the ROIs predicted as *class 1* and *class 3* are manually labeled to compute the precision performance for this generalizability assessment. Manual labeling of ROIs predicted as *class 0* and *class 2* is out of the scope of this work since it would take a considerable amount of time.

The precision results for the three recorders are presented in Table V. It is observed that A1 and A3 have the lowest precision for both, *class 1* and *class 3*, while A2, with lower anthropogenic noise activity, have precision above 99.6%.

Similarly, the predictions obtained using CNN2 for classifying the ROIs corresponding to the downsweep vocalizations are shown in Fig. 9. Results on the non-labeled data show that the starting frequency versus geotime behavior is similar to the labeled data shown in Fig. 5(c) where the long downsweep vocalizations are most commonly found below 1.5 kHz. Several occurrences happen above 1.5 kHz and correspond to AL2i(T) type of sounds which are less common to find in this particular area of the Chukchi shelf break.[4] ROIs predicted as *class 1* and *class 3* types of sounds have

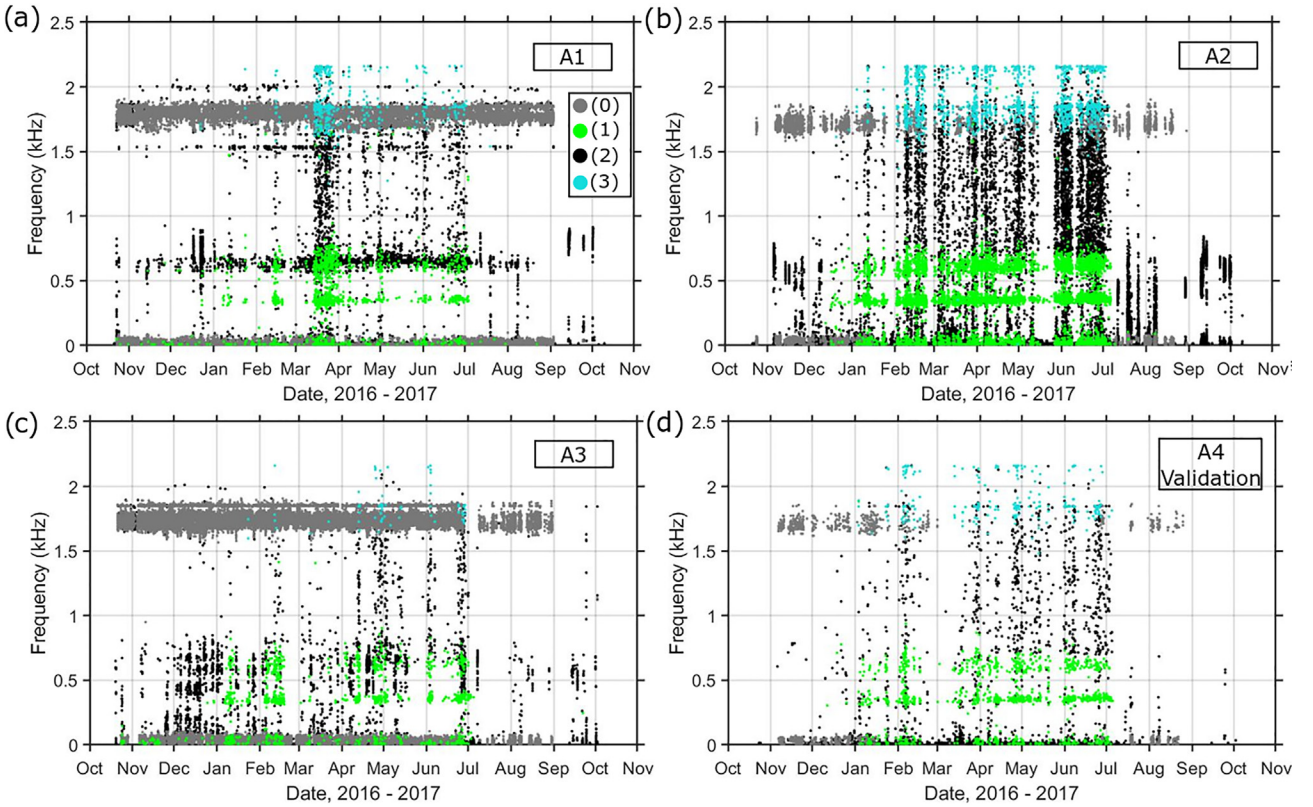

FIG. 8. (Color online) ROIs classified by CNN1 as a function of starting frequency and geo-time for (a) A1, (b) A2, (c) A3, and (d) the validation portion of A4 using Mask 1.

https://doi.org/10.1121/10.0009256

TABLE V. Precision for *class 1* and *class 3* using Mask 1.

True Class	Recorder	Predicted class 0	1	2	3	Precision (%)
	A1	20	1259	97	1	91.4
1	A2	5	6672	14	5	99.6
	A3	4	978	115	0	89.2
	A1	2	0	11	407	96.9
3	A2	0	0	0	1473	100.0
	A3	1	0	0	67	98.52

TABLE VI. Precision for *class 1* and *class 3* using Mask 2.

True Class	Recorder	Predicted class 0	1	2	3	Precision (%)
	A1	10	3692	33	0	98.8
1	A2	0	16047	3	0	99.9
	A3	1	2408	4	0	99.8
	A1	856	14	499	1399	50.5
3	A2	0	0	1	4484	99.9
	A3	87	0	11	769	88.7

been manually labeled to compute the precision performance of the network. The results are shown in Table VI, where the long downsweep vocalizations corresponding to *class 1* exhibit a precision above 98.8% for all the receivers. However, receivers with high anthropogenic noise activity (A1 and A3) had more difficulties for correctly classifying the short downsweep signals, i.e., *class 3*.

V. DISCUSSION

The 2D normalized cross correlation method implemented in Sec. III A provides a fast and functional detection system able to find the bearded seal vocalizations using the spectrogram representation of the acoustic signals of interest. However, an adequate threshold needs to be established to balance the false positive and true positive rates. While in detection systems it is common to heuristically set a threshold that balances this trade-off, in this paper, we have proposed an approach that allows the system to have a low threshold by reinforcing the detections using deep learning techniques.

A low threshold allows the detection system to identify several vocalizations that have a low SNR or that are masked by other noise events. However, this also makes the system identify other signals such as anthropogenic noise or other marine mammals vocalizations, which increases the false positive rate. In this work, these first detections are considered as ROIs and are 2D matrices found at a given frequency and geotime. To address the trade-off between false positive and true positive rates, CNNs are used for reinforcing the detections by classifying each ROI into several predetermined classes as described in Sec. III B.

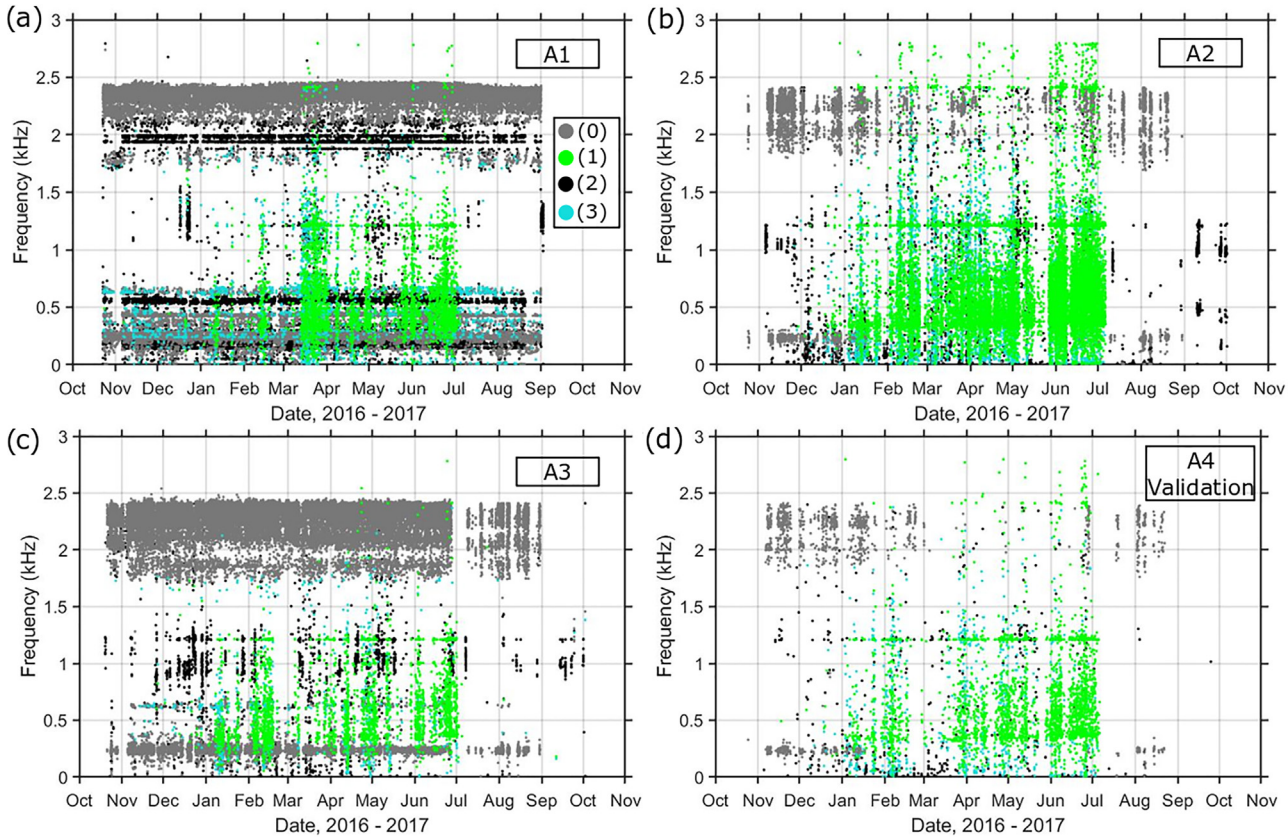

FIG. 9. (Color online) ROIs classified by CNN2 as a function of starting frequency and geo-time for (a) A1, (b) A2, (c) A3, and (d) validation portion of A4, using Mask 2.

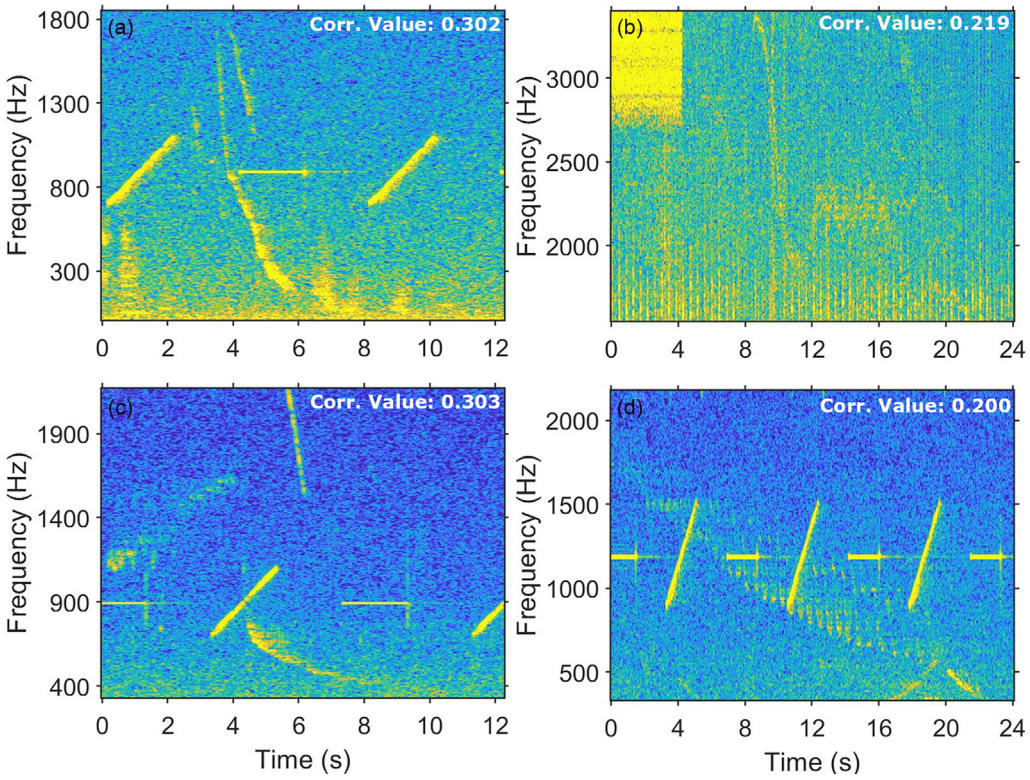

FIG. 10. (Color online) examples of the DCS for sounds measured on the receiver A3. (a), (b) ROIs falsely detected and classified as *Class 1* for (a) Mask 1 and (b) Mask 2. (c), (d) ROIs successfully detected and classified as *Class 1* for (a) Mask 1 and (b) Mask 2 under environmental noise.

To properly train a supervised machine learning model, a large labeled dataset is necessary. This database needs to have sufficient information that captures the variety found in the testing scenarios. For the ROIs found with Mask 1 [AL1i(T)], the CNN showed high performance when trained on data labeled for one single receiver and tested on the remaining three receivers at different locations. For Mask 2, corresponding to downsweep vocalizations, using data labeled for one single receiver (A4) was sufficient for most of the remaining recorders. However, since A1 was close to the sound source S2 which transmitted downsweep LFM signals, it was necessary to include a small portion of these ROIs found on A1 for the training of CNN2. This way, the network learns how to classify these type of signals that were not originally included when using only the data labeled for A4 which was the farthest from S2.

Figure 10 shows several examples of the DCS applied to noisy recordings in A3, where the correlation value is close to the threshold of both Masks. Figures 10(a) and 10(b) show the cases when the CNNs falsely classified background noise as *Class 1* for Mask 1 and Mask 2, respectively. These errors occurred when the noise had patterns similar to the vocalizations of interest. These few false positive classifications can be further alleviated by including more background noise signals in the training data set or by applying data augmentation techniques.[40] On the other hand, Figs. 10(c) and 10(d) show examples of the CNNs correctly classifying ROIs containing bearded seals vocalizations with low SNR and masked by anthropogenic signals.

VI. CONCLUSIONS

This study provides evidence that convolutional neural networks are suitable for classifying bearded seal vocalizations from candidate regions found by the spectrogram correlation technique. With this DCS, given that the ROIs will be classified by the CNNs, the sensitivity of the detection step can be increased to find more possible vocalizations even though many false positive events are triggered. For the CNNs, the training data must have enough information content for the networks to learn how to distinguish the signals of interest from other possible signals that might be present in the testing scenarios.

ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research Ocean Acoustics Program (ONR OA322) under Grant Nos. N00014-15-1-2110, N00014-18-1-2140, and N00014-21-1-2760.

[1]H. Frouin-Mouy, X. Mouy, B. Martin, and D. Hannay, "Underwater acoustic behavior of bearded seals (Erignathus barbatus) in the northeastern Chukchi Sea, 2007–2010," Mar. Mammal Sci. **32**(1), 141–160 (2016).
[2]A. F. Heimrich, W. D. Halliday, H. Frouin-Mouy, M. K. Pine, F. Juanes, and S. J. Insley, "Vocalizations of bearded seals (Erignathus barbatus) and their influence on the soundscape of the western Canadian Arctic," Mar. Mammal Sci. **37**(1), 173–192 (2021).
[3]D. Risch, C. W. Clark, P. J. Corkeron, A. Elepfandt, K. M. Kovacs, C. Lydersen, I. Stirling, and S. M. Van Parijs, "Vocalizations of male bearded seals, Erignathus barbatus: Classification and geographical variation," Animal Behav. **73**(5), 747–762 (2007).

[4]J. M. Jones, B. J. Thayre, E. H. Roth, M. Mahoney, I. Sia, K. Merculief, C. Jackson, C. Zeller, M. Clare, A. Bacon, S. Weaver, Z. Gentes, R. J. Small, I. Stirling, S. M. Wiggins, and J. A. Hildebrand, "Ringed, bearded, and ribbon seal vocalizations north of Barrow, Alaska: Seasonal presence and relationship with sea ice," Arctic **67**(2), 203–222 (2014).
[5]I. Parisi, G. de Vincenzi, M. Torri, E. Papale, S. Mazzola, A. Bonanno, and G. Buscaino, "Underwater vocal complexity of Arctic seal Erignathus barbatus in Kongsfjorden (Svalbard)," J. Acoust. Soc. Am. **142**(5), 3104–3115 (2017).
[6]H. J. Cleator, I. Stirling, and T. G. Smith, "Underwater vocalizations of the bearded seal (Erignathus barbatus)," Can. J. Zool. **67**(8), 1900–1910 (1989).
[7]K. Q. MacIntyre, K. M. Stafford, C. L. Berchok, and P. L. Boveng, "Year-round acoustic detection of bearded seals (Erignathus barbatus) in the Beaufort Sea relative to changing environmental conditions, 2008–2010," Polar Biol. **36**(8), 1161–1173 (2013).
[8]T. K. Boye, M. J. Simon, K. L. Laidre, F. Rigét, and K. M. Stafford, "Seasonal detections of bearded seal (Erignathus barbatus) vocalizations in Baffin Bay and Davis Strait in relation to sea ice concentration," Polar Biol. **43**(10), 1493–1502 (2020).
[9]D. E. Hannay, J. Delarue, X. Mouy, B. S. Martin, D. Leary, J. N. Oswald, and J. Vallarta, "Marine mammal acoustic detections in the northeastern Chukchi Sea, September 2007–July 2011," Continental Shelf Res. **67**, 127–146 (2013).
[10]W. D. Halliday, S. J. Insley, T. de Jong, and X. Mouy, "Seasonal patterns in acoustic detections of marine mammals near Sachs Harbour, Northwest Territories," Arct. Sci. **4**, 259–278 (2017).
[11]W. D. Halliday, M. K. Pine, S. J. Insley, R. N. Soares, P. Kortsalo, and X. Mouy, "Acoustic detections of arctic marine mammals near ulukhaktok, northwest territories, Canada," Can. J. Zool. **97**(1), 72–80 (2019).
[12]X. Mouy, M. Bahoura, and Y. Simard, "Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence," J. Acoust. Soc. Am. **126**(6), 2918–2928 (2009).
[13]D. K. Mellinger, "A comparison of methods for detecting right whale calls," Can. Acoust. **32**(2), 55–65 (2004).
[14]X. Mouy, J. Oswald, D. Leary, J. Delarue, J. Vallarta, b Rideout, D. Mellinger, C. Erbe, D. Hannay, and B. Martin, "Passive acoustic monitoring of marine mammals in the Arctic," in Detection, *Classification, Localization* of *Marine Mammals Using Passive Acoustics* (DIRAC NGO, Paris, France, 2013), pp. 185–224.
[15]B. Martin, K. Kowarski, X. Mouy, and H. Moors-Murphy, "Recording and identification of marine mammal vocalizations on the scotian shelf and slope," in *Proceedings of 2014 Oceans*, St. John's, Newfoundland, Canada (June 16–17, 2014).
[16]H. Frouin-Mouy, X. Mouy, C. L. Berchok, S. B. Blackwell, and K. M. Stafford, "Acoustic occurrence and behavior of ribbon seals (Histriophoca fasciata) in the Bering, Chukchi, and Beaufort seas," Polar Biol. **42**(4), 657–674 (2019).
[17]D. Gillespie, M. Caillat, J. Gordon, and P. White, "Automatic detection and classification of odontocete whistles," J. Acoust. Soc. Am. **134**(3), 2427–2437 (2013).
[18]J. R. Potter, D. K. Mellinger, and C. W. Clark, "Marine mammal call discrimination using artificial neural networks," J. Acoust. Soc. Am. **96**(3), 1255–1262 (1994).
[19]C. H. Ho, J. Joseph, H. Ming Jer, and T. Margolina, "Automated detection and identification of blue and fin whale foraging calls by combining pattern recognition and machine learning techniques," in *Proceedings of Oceans 2016 MTS/IEEE*, Monterey, CA (September 19–23, 2016). pp. 1–7.
[20]X. C. Halkias, S. Paris, and H. Glotin, "Classification of mysticete sounds using machine learning techniques," J. Acoust. Soc. Am. **134**(5), 3496–3505 (2013).
[21]S. Liu, M. Liu, M. Wang, T. Ma, and X. Qing, "Classification of Cetacean Whistles Based on Convolutional Neural Network," in *Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing*, Hangzhou, China (October 18–20, 2018).
[22]W. Luo, W. Yang, and Y. Zhang, "Convolutional neural network for detecting odontocete echolocation clicks," J. Acoust. Soc. Am. **145**(1), EL7–EL12 (2019).
[23]M. Zhong, M. Castellote, R. Dodhia, J. Lavista Ferres, M. Keogh, and A. Brewer, "Beluga whale acoustic signal classification using deep learning neural network models," J. Acoust. Soc. Am. **147**(3), 1834–1841 (2020).
[24]Y. Shiu, K. J. Palmer, M. A. Roch, E. Fleishman, X. Liu, E. M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," Sci. Rep. **10**(1), 1–12 (2020).
[25]M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," in *Machine Learning and Knowledge Discovery in Databases*, edited by U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet (Springer International Publishing, Cham, 2020), pp. 290–305.
[26]O. S. Kirsebom, F. Frazao, Y. Simard, N. Roy, S. Matwin, and S. Giard, "Performance of a deep neural network at detecting north atlantic right whale upcalls," arXiv:2636 (2020).
[27]P. C. Bermant, M. M. Bronstein, R. J. Wood, S. Gero, and D. F. Gruber, "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," Sci. Rep. **9**(1), 1–10 (2019).
[28]M. S. Ballard, M. Badiey, J. D. Sagers, J. A. Colosi, A. Turgut, S. Pecknold, Y.-T. Lin, A. Proshutinsky, R. Krishfield, P. F. Worcester, and M. A. Dzieciuch, "Temporal and spatial dependence of a yearlong record of sound propagation from the Canada Basin to the Chukchi Shelf," J. Acoust. Soc. Am. **148**(3), 1663–1680 (2020).
[29]M. D. Collins, A. Turgut, R. Menis, and J. A. Schindall, "Acoustic recordings and modeling under seasonally varying sea ice," Sci. Rep. **9**(1), 1–11 (2019).
[30]M. Badiey, L. Wan, S. Pecknold, and A. Turgut, "Azimuthal and temporal sound fluctuations on the Chukchi continental shelf during the Canada basin acoustic propagation experiment 2017," J. Acoust. Soc. Am. **146**(6), EL530–EL536 (2019).
[31]M. Jakobsson, L. Mayer, B. Coakley, J. A. Dowdeswell, S. Forbes, B. Fridman, H. Hodnesdal, R. Noormets, R. Pedersen, M. Rebesco, H. W. Schenke, Y. Zarayskaya, D. Accettella, A. Armstrong, R. M. Anderson, P. Bienhoff, A. Camerlenghi, I. Church, M. Edwards, J. V. Gardner, J. K. Hall, B. Hell, O. Hestvik, Y. Kristoffersen, C. Marcussen, R. Mohammad, D. Mosher, S. V. Nghiem, M. T. Pedrosa, P. G. Travaglini, and P. Weatherall, "The international bathymetric chart of the Arctic Ocean (IBCAO) version 3.0," Geophys. Res. Lett. **39**(12), 1–6, https://doi.org/10.1029/2012GL052219 (2012).
[32]J. P. Lewis, "Fast normalized cross-correlation," Industrial Light & Magic, http://scribblethink.org/Work/nvisionInterface/nip.html (Last viewed: January 10, 2022).
[33]I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[34]T. B. Neilsen, C. D. Escobar-Amado, M. C. Acree, W. S. Hodgkiss, D. F. Van Komen, D. P. Knobles, and M. Badiey, "Learning location and seabed type from a moving mid-frequency source," J. Acoust. Soc. Am. **149**, 692–705 (2021).
[35]C. D. Escobar-Amado, T. B. Neilsen, J. Castro-Correa, D. F. Van Komen, M. Badiey, D. P. Knobles, and W. S. Hodgkiss, "Seabed classification from merchant ship-radiated noise using a physics-based ensemble deep learning algorithms," J. Acoust. Soc. Am. **150**, 1434–1447 (2021).
[36]A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, and L. Antiga, *et al*. "PyTorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada* (2019), pp. 8024–8035
[37]V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Madison, WI (June 21–24, 2010), pp. 801–814.
[38]S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France (July 6–11, 2015), pp. 448–456.
[39]D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA (May 7–9, 2015).
[40]J. A. Castro-Correa, M. Badiey, T. B. Neilsen, D. P. Knobles, and W. S. Hodgkiss, "Impact of data augmentation on supervised learning for a moving mid-frequency source," J. Acoust. Soc. Am. **150**(5), 3914–3928 (2021).