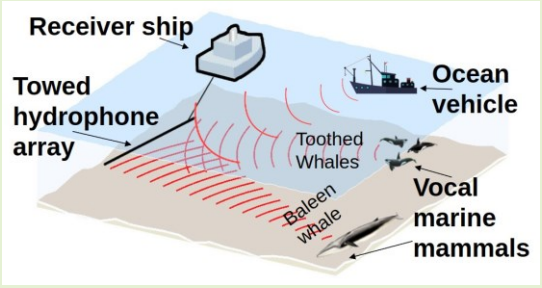
Real-Time Detection, Bearing Estimation, and Whale Species Vocalization Classification From Passive Underwater Acoustic Array Data

Hamed Mohebbi-Kalkhoran, Nicholas C. Makris, and Purnima Ratilal, *Member, IEEE*

Abstract **—Developing automatic algorithms for real-time monitoring of underwater acoustic events is essential in ocean acoustic applications. Most previous ocean acoustic ecosystem monitoring studies are non-real-time, focusing on data received on a single hydrophone or a specific analysis, such as bearing estimation or detection, without considering the full end-to-end analysis system. Here, we develop a unified framework for real-time ocean acoustic data analysis including beamforming, detection, bearing estimation, and classification of transient underwater acoustic events. To detect sound sources, thresholding on computed mel-scale per-channel energy normalization (PCEN) is applied, followed by morphological image opening to extract pixels with significant intensities. Next, connected component analysis is applied for grouping pixel detections. The bearing of signal detections is next estimated via nonmaximum suppression (NMS) of 3-D stacked beamformed spectrogram imageries. To classify a variety of whale species from their calls, time–frequency features are extracted from each detected signal's beamformed power spectrogram. These features are next applied to train three classifiers, including support vector machine (SVM), neural networks, and random forest (RF), to classify six whale vocalization categories: Fin, Sei, Unidentified Baleen, Minke, Humpback, and general Odontocetes. Best results are obtained with the RF classifier, which achieved 96.7% accuracy and 87.5% F1 score. A variety of accelerating approaches and fast algorithms are implemented to run on GPU. During an experiment in the U.S. Northeast coast in September 2021, the software and hardware advances developed here were used for near real-time analysis of underwater acoustic data received by Northeastern University's in-house fabricated 160-element coherent hydrophone array system.**



Index Terms **— Array processing, beamforming, coherent hydrophone array, data processing acceleration, detection, machine learning, passive acoustic, remote sensing.**

I. Introduction

UNDERWATER acoustic data usually contain signals from a myriad of sound sources including marine life such as marine mammals (MMs), fishes, and crustaceans; man-made machinery such as ships, wind farms, and seismic airguns; and natural geophysical processes such as earthquake and volcanic eruption [1], [2], [3], [4], [5], [6], [7]. MM vocalization classification is a challenging problem due to the transient nature of their broadband calls, high variation in the calls of a specie (intraclass variation), and high similarity between the calls of some species.

MMs are usually defined as mammals whose terrestrial predecessors have returned to life in the sea. The most important criterion is that they must get all or most of their food from the marine environment. MMs include at least 129 extant species divided into four groups [8], [9]. They are cetaceans (whales, dolphins, and porpoises), pinnipeds (seals, sea lions, and walruses), sirenians (sea cows that are now extinct, manatees, and dugongs), and fissipeds (one bear, the polar bear, the sea, and marine otters). Some of these groups such as fissipeds are much less completely adapted to living

Received 9 June 2024; revised 8 September 2024; accepted 11 September 2024. Date of publication 2 October 2024; date of current version 14 November 2024. This work was supported in part by the U.S. Office of Naval Research under Grant N00014-23-1-2327, Grant N00014-20-1-2026, and Grant N00014-20-1-2035; and in part by the U.S. National Science Foundation under Grant OCE-2219953 and Grant OCE-1736749. The associate editor coordinating the review of this article and approving it for publication was Prof. Dongsoo Har. *(Corresponding author: Hamed Mohebbi-Kalkhoran.)*

Hamed Mohebbi-Kalkhoran and Nicholas C. Makris are with the Department of Mechanical Engineering, MIT, Cambridge, MA 02139 USA (e-mail: hmohebbi@mit.edu; makris@mit.edu).

Purnima Ratilal is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: purnima@ece.neu.edu).

Digital Object Identifier 10.1109/JSEN.2024.3469112

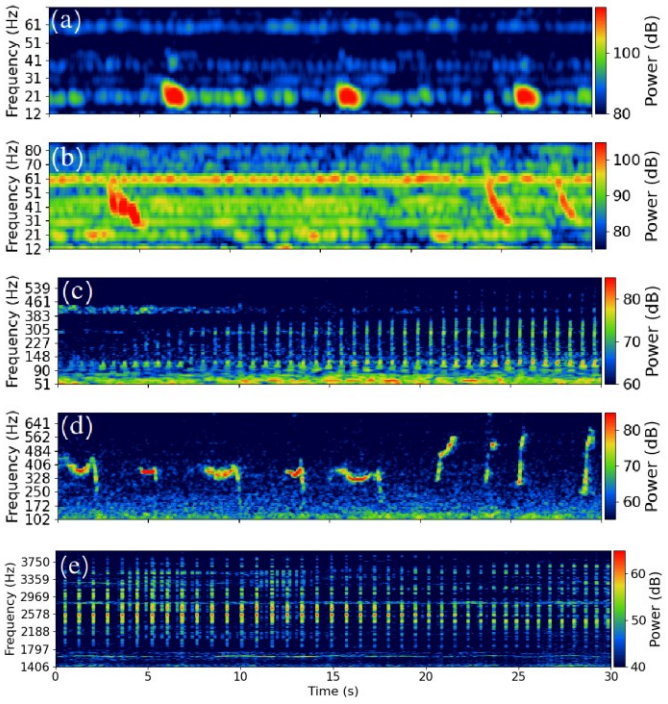


Fig. 1. Sample log-transformed beamformed power spectrograms of different whale vocalization categories from the GOM 2006 Experiment. (a) Fin whale 20-Hz pulses, (b) unidentified baleen whale downsweep chirp (UBDC on left) and pair of sei whale downsweep chirps (right), (c) minke whale buzz sequence, (d) humpback whale song sequence, and (e) odontocete whale click signals.

in the water compared with others such as cetaceans that are fully adapted to live their entire lives in the water [8].

There are currently 86 known living species of Cetacea, which can be divided into two suborders—baleen whales (Mysticeti) and toothed whales (Odontoceti). Baleen whales examples are Blue (*Balaenoptera musculus*), fin (*Balaenoptera physalus*), sei (*Balaenoptera borealis*), minke (*Balaenoptera acutorostrata*), and humpback (*Megaptera novaeangliae*). Toothed whales examples are sperm (*Physeter microcephalus*), killer (*Orcinus orca*), pilot (*Globicephala* spp.), and *Delphinid* species. Baleen whales are generally large, such as the blue whale (up to 33 m or more in length) and the smallest is the pygmy right whale (less than 7-m long). In contrast, the toothed whales are small to medium size, with the exception of the sperm whale which can get up to 18 m in length. Baleen whales do not have teeth, and instead in the mouth, the upper jaw is hung with baleen (stiff plates of keratin with fringes on the inside) [8].

MM vocalizations are associated with a variety of purposes such as echolocation, sexual display while mating, singing while migrating to breeding and feeding grounds, communication, and contact calls for coordinated movement during group feeding and other activities [1], [10]. In this work, we investigate whale species classification for six classes including: Fin, Sei, Unidentified Baleen Downsweep Chirp (UBDC), Minke, Humpback, and toothed whales or Odontocetes. Examples of calls in each of these categories are shown in Fig. 1, recorded during the Gulf of Maine (GOM) 2006 Experiment [1], [11], [12], [13], [14]. These sounds span a wide range of frequencies from low-frequency fin whale 20-Hz pulse calls, to odontocetes click signals recorded here up to 4 kHz due to sampling frequency limitation of 8 kHz. Odontocetes can make sounds with significant energies at much higher frequencies to about 150–170 kHz [15] from the smallest odontocetes.

Substantial volumes of underwater acoustic data are usually acquired in passive ocean acoustic waveguide remote sensing (POAWRS) experiments with a large-aperture densely populated coherent hydrophone array. The POAWRS technique has been widely used to detect, localize, track, and study underwater acoustic events such as whale behavior, population and distributions in time and space, and their interaction with prey species [1], [4], [7], [14], [16]. The task of processing and analyzing large volumes of underwater acoustic data is extremely laborious, especially when the data are drawn from measurements of a large-aperture densely populated coherent hydrophone array, since beamformed signals in multiple distinct bearing directions spanning 360° azimuths about the receiver array need to be analyzed concurrently. Developing automatic, accurate, and fast algorithms for the detection and classification of underwater acoustic events can help minimize the human effort needed for underwater acoustic monitoring, enabling real-time processing and analysis, and hence aiding rapid scientific discoveries at sea.

Prior works in the published literature on classification of MM sounds using machine learning approaches and developed in recent years include [7], [17], [18], [19], [20], [21], [22], [23], [24], and [25]. In [17], deep learning was used for detection and classification of Sperm whale clicks. They used convolutional neural network (CNN) to classify each 0.5 s of data into “click” and “no click” categories, and then used long short-term memory (LSTM) and gated recurrent unit (GRU) to perform classification tasks for coda type, vocal clan, and whale identity classification. In [18], wavelet denoising was used to reduce noise level in the audio signal, and then dual-threshold endpoint detection algorithm [26] was used to detect the beginning and ending positions of the whale clicks. Next, a set of extracted features emphasizing the duration and scale energy from wavelet coefficient matrix were extracted and used to train a set of classifiers including support vector machine (SVM) and neural networks to distinguish between sperm whale and pilot whale clicks. Zhang et al. [19] used transfer learning approach using pretrained CNN models on the ImageNet dataset [27], and combined with Mix-up data augmentation, they fine-tuned the models on the whale-call datasets to subclassify 16 whale family units known as “pods” from killer and pilot whales. In [21], CNN consisting of residual blocks is trained for killer whale sound classification. In [22], region-based CNN is used for detection of Fin whale 40 Hz, and Blue whale D calls while also providing number of calls and times they occur. To detect North Atlantic right whale up-calls, Ibrahim et al. [23] used spectrograms and scalograms as the input to a CNN and stacked autoencoder, respectively, to train the models and then fused the predictions from the two individual models for final prediction. In [24], CNNs are trained for killer and long-finned pilot whales’ whistle detection. In [25], five different types of Fin whale calls including 20-Hz single and double pulses,

18-Hz backbeat, 130-Hz upsweep, and 40–60-Hz downsweep chirps in beamformed data of a coherent hydrophone array were classified using a set of classifiers including SVM, decision tree (DT), and CNNs. In older works, such as [28], automatic detection, and classification system for baleen whale calls was developed using pitch-tracking and quadratic discriminant function analysis. In [29], subclassification of humpback whale downsweep moan calls into 13 subgroups was accomplished using K-means clustering after beamformed spectrogram analysis, pitch-tracking, and time–frequency feature extraction. Automatic classifiers were further developed in [7] to distinguish humpback whale song sequences from nonsong calls in the GOM by first applying bag of words to build feature vectors from beamformed time-series signals, calculating both power spectral density and mel frequency cepstral coefficients (MFCCs) features, and then using and comparing the performances of SVM, neural networks, and naive Bayes in the classification. Wavelet-transform-based approaches are also widely used in different applications. In [30], features from the discrete wavelet transform are fed into CNN architecture for fault detection in power systems. In [31], least-squares wavelet is applied to astronomical time-series analysis and shown to provide higher resolution time–frequency spectrogram since it considers correlated and systematic noises. Most existing approaches for whale call classification focus on just one to two species and are based on acoustic data acquired with a single hydrophone. Supervised deep-learning-based approaches such as CNN and recurrent neural network (RNN) usually require large volume of labeled data to reach high accuracy. This is especially challenging for coherent hydrophone array data since labeling is necessary for data in multiple beamformed directions and frequency ranges.

Here, we implement beamforming algorithms for high sample rate large-aperture coherent hydrophone array data and output the power spectrogram density in multiple bearings spanning 360° azimuths about the horizontal receiver array simultaneously. Next, approaches for rapid detection, bearing estimation, and time–frequency feature extraction of underwater acoustic events from high sample rate beamformed data are implemented, including adaptation of methods in computer vision, image processing, and perception science for underwater acoustic data analysis. The results are then used as inputs in classification of MM vocalizations for real-time applications. We use underwater acoustic data from a 160-element coherent hydrophone array and use the POAWRS technique to enable sensing and detections over instantaneous wide areas more than 100 km in diameter from the array. A variety of computational accelerating approaches, combining hardware and software, which make the methods desirable for real-time applications are also developed.

II. Materials and Methods

A. Datasets

In this section, we describe the datasets analyzed here which are from the GOM 2006 and the U.S. Northeast coast (USNE) 2021 experiments.

The GOM 2006 Experiment dataset [1], [12], [13], [14] was acquired in Fall 2006, from September 19 to October 6, in this important North Atlantic MM feeding ground containing large populations of spawning fish, the Atlantic herring [12], [13], [32]. Acoustic recordings of whale vocalizations were acquired using a large-aperture densely populated coherent hydrophone array with 160 elements, the ONR FORA array [33], [34] towed by a research vessel along designated tracks in Franklin Basin, north of Georges Bank. The acoustic data are sampled at 8000 Hz per element. Data from all 160 hydrophone elements nested into four subapertures are analyzed, where each subaperture contains 64 hydrophones for spatially and temporally unaliased sensing up to 4 kHz [1]. In some time periods of the experiment, several hydrophones were nonresponsive and were omitted from data processing, reducing the number of usable hydrophones from 160 to 132. The water depth ranged from 180 to 250 m at the array locations. The array tow depth was roughly 105 m, and tow speed was roughly 2 m/s.

Previous analysis in [1], [11], and [29] provided the labeled set of MM vocalization signals for this project. There the acoustic pressure–time series measured by sensors across the coherent hydrophone array which were converted into 2-D beam-time series by beamforming. A total of 64 beams were formed, separately for each subaperture, spanning 360° horizontal azimuth about the receiver array. Each beam-time series was converted into a beamformed spectrogram by short-time Fourier transform (STFT) (sampling frequency 8 kHz, frame 2048 samples, overlap 3/4, Hann window). Significant sounds present in the beamformed spectrograms were automatically detected by first applying a pixel intensity threshold detector [35] followed by pixel clustering and verified by visual inspection. Beamformed spectrogram pixels with local intensity values that stood 10 dB above the background were grouped using a clustering algorithm according to a nearest-neighbor criteria which determines whether the pixels can be grouped into one or more significant sound signals. MMs' vocalization signal detections were verified and labeled manually by visual inspection and listening to the sounds [1], [14], [29].

The USNE 2021 experiment dataset was acquired from September 3 to September 8, 2021 at the Great South Channel and the continental slope and deep water south of Rhode Island. The purpose of this experiment was to test the newly developed Northeastern University large-aperture coherent hydrophone array hardware, and data processing and analysis software systems for instantaneous wide-area passive acoustic monitoring of MMs, including detection and classification of MM sounds. The 160-element coherent hydrophone array hardware and software systems were designed, fabricated, and assembled in-house at Northeastern University (NU) [36], [37], [38], [39], [40]. The array was towed by a ship, and acoustic data were acquired at adjustable sampling frequencies ranging from 8 to 100 kHz per element. Whale vocalizations from a wide range of species, both mysticetes and odontocetes, were recorded in the frequency range spanning 10 Hz–50 kHz.

The experimental regions for both the GOM 2006 and USNE 2021 experiments are shown in Fig. 2. Detailed schematics of the coherent hydrophone array systems including hydrophone positioning are provided in [33] and [34] for

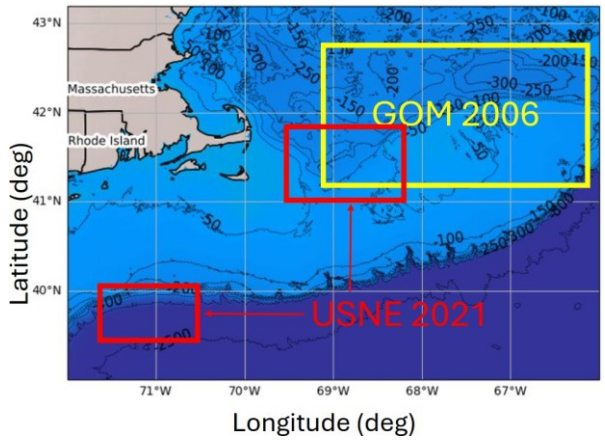


Fig. 2. Map of the region containing bathymetry contours for GOM 2006 and USNE experiments.

the ONR FORA hydrophone array used in GOM 2006 Experiment, and in [38] for the NU coherent hydrophone array used in USNE 2021 Experiment.

B. Beamforming

Beamforming is a spatial filtering technique using data from an array of sensors to enhance the signal-to-noise ratio (SNR) in the specified azimuthal direction or bearing and is widely used in acoustic and electromagnetic remote sensing and imaging systems [41], [42]. Relative bearing is the horizontal azimuth of a source with respect to the hydrophone array axis normal. Delay and sum array beamforming is used to amplify signals in specified relative bearing directions while simultaneously suppressing signals from other directions. The theoretical SNR gain from beamforming when using the n hydrophone array compared with using only one hydrophone is up to $10 \log_{10} n$ dB, so if $n = 64$, the SNR gain is up to 18 dB [1]. The actual array gain, which may differ from the 18-dB theoretical array gain, is dependent on noise coherence and signal wavelength relative to array aperture length [43]. A large densely populated linear array of hydrophones enables detection of acoustic events from greater distances than a single hydrophone. The time delay of signal arrival between two successive hydrophones in the array is computed using the following equation:

$$t = \frac{d}{c} \sin\theta \tag{1}$$

where c is the speed of sound propagation in water, d is the distance between two successive hydrophones in meters, and θ is the relative bearing. Data are beamformed in 147 directions spanning θ between -90° and $+90^\circ$ corresponding to back- and forward endfires, respectively, with 0° corresponding to array broad side. The relative bearing θ is converted into true bearing β in a clockwise direction from true north by correcting for the array heading α.

C. Detection of Transient Signals

Many sound sources in the ocean generate broadband transient signals, for example, MM vocalizations, fish-generated sounds, and human activities such as offshore piling and sonar transmissions. The detection of transient broadband signals can be significantly enhanced via data transformation in two key steps. They are nonuniform high-frequency compression via mel-scale transformation and removal of persistent tonal background signals via per-channel energy normalization (PCEN) [44], [45]. Morphological image processing operation is next applied to remove noisy background regions while finding the desired potential detections.

1) Mel Spectrogram: The mel-scale is a result of psychoacoustics [46], which is a scale of pitches judged by listeners to be equal in distance one from another and looks like a quasi-logarithmic function of acoustic frequency. It shows humans do not perceive sound frequency differences in a linear scale. For example, 500-Hz frequency differences at lower frequencies, such as between 500 and 1000 Hz, are more readily discernible than at higher frequencies, such as between 10 000 and 10 500 Hz. In [47], the empirically determined transformation from Hertz to mel-scale is defined as

$$m = C \log_{10}\left(1 + \frac{f}{f_0}\right) \tag{2}$$

where $f_0 = 700$ Hz is the corner frequency where human perception of frequency transitions from approximately linear to log dependence, and $C = 1000/\log_{10}(1 + (1000/700)) = 2595$ is a constant that ensures the mel-scale $m = 1000$ equals the physical frequency $f = 1000$ Hz.

In linear scale spectrogram, the frequency width of consecutive bins is a constant. However, in mel spectrogram, frequency resolution varies nonlinearly such that resolution is finer at low frequencies and broader at high frequencies. An advantage of mel-scale over linear scale spectrogram is a reduction in computational complexity for subsequent stages of the detection algorithm. For example, in STFT for spectrogram generation with desired temporal window length of 0.128 s, data sampled at 8000 Hz will lead to $8000 * 0.128 = 2^{10} = 1024$ number of FFT points for each temporal window. For data sampled at 100 kHz, there will be $2^{14} = 16\,384$ number of FFT points for the same temporal resolution. This increase in the number of FFT points will in consequence raise the computational complexity. Here, using mel-scale can help improve the processing time for the subsequent steps. In the ocean acoustic datasets analyzed here, the key frequency information contained in detected signals can usually be represented by just 256 mel bins since fine frequency resolution is often needed for acoustic signals centered at low frequencies, but the same resolution is not necessary for signals centered at high frequencies. For example, a frequency resolution of around 1 Hz is usually necessary for an acoustic event centered at 20 Hz, while a coarser frequency resolution of around 100–500 Hz would suffice for acoustic events at 10 kHz. This is because most natural underwater acoustic signals, such as whale calls, have increasing frequency bandwidths as their call central frequencies increase. Another advantage of using mel over linear frequency scale is the reduction in noise level in the mel-scale. That is because for computing the mel-scale from linear scale, some kind of averaging (usually triangular window filter bank) is often used to summarize the frequency information from

multiple frequency bands into a single mel-scale band. This in consequence increases the detection algorithm's accuracy. Nonlinearity in human perception also extends to acoustic intensities, where Weber's law applies in that the smallest resolvable change, a just-notable-difference, grows in direct proportion to the stimulus. In [48], Weber's law was found to be a consequence of attaining the theoretical minimum mean-square error possible, the Cramer–Rao lower bound, in resolving the intensity of naturally scintillating light and sound. Thus showing human intensity resolution is optimally adapted to the natural scintillation of light and sound.

2) Per-Channel Energy Normalization: PCEN [44] is an alternative to logarithmic transform, with the aim of providing better dynamic range compression (DRC), adaptive gain control, robustness to channel distortion, and learnable differentiable parameters that can be optimized using gradient-based optimizations [45]. The PCEN is applied to beamformed mel-scale transformed data prior to signal detection. Here, the equation used for PCEN, derived in the Appendix, is given by

$$\mathbf{PCEN}(t, f) = \left(\frac{\mathbf{E}(t, f)}{(\epsilon + \mathbf{M}(t, f))^{\alpha}} + \delta \right)^{r} - \delta^{r} \tag{3}$$

where $\mathbf{E}(t, f)$ is the instantaneous signal energy at time t and frequency f, $\mathbf{M}(t, f)$ is the mean signal energy obtained via time-averaging, ϵ prevents overflow error in quiet background regions, r is the exponential transform power, and $\delta > 1$ is a threshold parameter that enables the DRC to be adjusted. Instead of having a constant δ, we set it as a frequency-dependent parameter in a way that it fits the background frequency-dependent ocean noise. The formula for δ is described in the Appendix.

3) Thresholding and Morphological Operations: The goal for transient underwater acoustic event detection is to extract individual signals that stand above the background noise and to focus on capturing the dominant signal energy, without breaking a signal into parts. To do so, we apply a procedure consisting of pixel thresholding, morphological image opening, and connected components' labeling. A constant threshold equal to 2.5 is applied to PCEN spectrograms images to convert them into binary images. This threshold value is determined experimentally to provide a good tradeoff between false alarm rate and missed detections. Smaller threshold values will lead to more noisy detections while larger threshold values will cause some true detections to be missed. This constant thresholding is an advantage of working with PCEN since it scales out the frequency dependence of underwater stationary ambient noise. Otherwise a frequency-dependent and varying threshold would be needed, for instance, when working directly with linear or log-transformed acoustic energies.

Fig. 3(d) shows an example of the output of thresholding on PCEN spectrogram for frequency range below 100 Hz. These data include several Fin whale 20-Hz and Sei whale downsweep chirp calls. As can be seen, the PCEN image has frequency-independent background, enabling the transient events (whale calls) to be readily detected. To focus detection on dominant signal energies and to make the detection algorithm more robust against background noise regions, morphological image opening is applied. Morphological image opening consists of two steps, the first is morphological image erosion, and the second is morphological image dilation. In the erosion step, a kernel slides through the binary image, and in the output image a pixel value will be equal to 1 if all the values inside the kernel are 1, and else it will be 0. Erosion makes the foreground boundaries thinner and also removes small dot-like detections. Dilation is the opposite of erosion in that a pixel is 1 in the output image if at least one pixel in the kernel is 1, making the boundaries of the foreground thicker and also merging nearby detections, thus compensating for the effect of erosion on foreground boundaries. Fig. 3(e) shows the output of the opening operation on the binary image in Fig. 3(d) where the correct number signals are detected focusing on dominant signal energies.

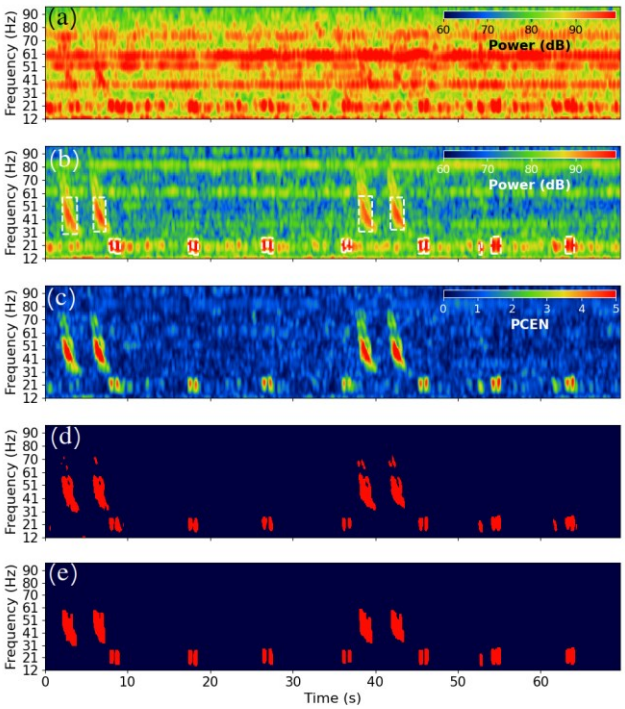


Fig. 3. Detection processing stages showing (a) log-transformed power spectrogram of single hydrophone, (b) log-transformed power spectrogram of beamformed data using 132 hydrophones, (c) PCEN transformed spectrogram, (d) thresholded binary image, and (e) binary image after image opening. The results of connected component labeling are shown as the white bounding boxes over the detected regions on the beamformed power spectrogram in (b).

Next, connected components' labeling [49] is applied to find the exact location of the detections and also separate them from each other. In a binary image, a connected component is a set of adjacent pixels whose values are 1. Fig. 3(b) shows the bounding boxes on the original beamformed power spectrogram after applying the connected component labeling on Fig. 3(e). The entire acoustic events' detection procedure is capable of detecting and separating individual transient signals.

D. Bearing Estimation

The PCEN and transient acoustic event detection algorithms are applied to each beamformed spectrogram imagery

to extract significant acoustic signals. Because of potential sidelobe effects and broad spatial beamwidths, especially at low frequencies, some detections may appear in multiple azimuthal directions. The relative bearing direction with the highest foreground energy is the correct direction for the detected signal. Furthermore, signals not overlapping in time and frequency should be considered as separate detections.

Here, we use nonmaximum suppression (NMS) to determine a detected signal's correct relative bearing. Note that NMS has been widely used for object detection in image processing [50], [51], [52]. In computer vision applications, many object detection methods produce multiple potential detections that are usually indicated by bounding boxes around the target objects. There NMS is used to select the bounding box with the highest detection score or probability among all the bounding boxes overlapping above some threshold. There are different criteria for measuring the percent of overlap between bounding boxes, among which the most commonly used is intersection over union (IoU), a similarity measure based on the Jaccard index. The IoU between two sets, A and B, is expressed as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{Area of Intersection}}{\text{Area of Union}}. \quad (4)$$

In NMS, among all the bounding boxes with IoU higher than some predefined threshold, the one with highest detection score is selected as the final detection.

For bearing estimation here, we use NMS and IoU overlapping measure to find the bounding box with the highest score among overlapping bounding boxes. To compute IoU, the signal energy in decibels is used as the detection score obtained from summing power of foreground time and frequency pixels inside detection bounding boxes in beamformed power spectrogram imagery.

Once the correct relative azimuthal angle is found, the potential left–right bearings with respect to true north are calculated using the following:

$$\beta_{\text{right}} = \alpha + (90 - \theta) \quad (5)$$

$$\beta_{\text{left}} = \alpha - (90 - \theta) \quad (6)$$

where θ is the relative bearing of the detected acoustic event with respect to the array heading. α is the array heading with respect to the north. β_{right} is the right bearing with respect to the north, and β_{left} is the left bearing with respect to the north, and only one of these is the true bearing, while the other is ambiguous. The inherent left–right ambiguity about the horizontal line-array's axis is resolved by varying ship heading as described in [12], [14], [53], and [54]. The bearing estimates for acoustic detections in the GOM 2006 and USNE 2021 Experiments are shown in Section III. Sequences of bearing estimates for acoustic detections from a particular source form a bearing-time trajectory. Bearing-time trajectories are essential for subsequent passive acoustic source localizations [53], [55].

E. Classification

For each detection, we extract a set of features from the time–frequency power spectrograms of beamformed signals. These features should be as discriminant as possible for different classes and at the same time be similar for samples of the same class. We extracted the following 13 features: 1) minimum frequency, $f_{\min}$ (Hz); 2) maximum frequency $f_{\max}$ (Hz); 3) average central frequency f^c (Hz); 4) bandwidth BW (Hz); 5) mean instantaneous bandwidth BW_{avg} (Hz); 6) power amplitude weighted average frequency $\bar{f}$ (Hz); 7) duration of the detected call τ (s); 8) SNR; 9) slope (Hz/s); 10) curvature (Hz/s^2) from the first- and second-order polynomial fit to the vocalization traces obtained via pitch-tracking; 11) area which is the number of pixels for the detected sound; 12) relative instantaneous bandwidth BW^r; and 13) instantaneous power-weighted average frequency $\bar{f}^w$ (Hz). Equations (7)–(20) provide the formulas to calculate these features

$$f_{\min} = \min_{i,j} f(i, j) \quad \forall i, j \quad (7)$$

$$f_{\max} = \max_{i,j} f(i, j) \quad \forall i, j \quad (8)$$

$$\text{BW} = f_{\max} - f_{\min} \quad (9)$$

$$\mathbf{f_t}(j) = f(i = t, j) \quad (10)$$

$$f_t^c = \left(\max_j(\mathbf{f_t}) + \min_j(\mathbf{f_t}) \right) / 2 \quad (11)$$

$$f^c = \sum_t f_t^c / N_T \quad (12)$$

$$\text{BW}_i = \max_j(\mathbf{f_t}) - \min_j(\mathbf{f_t}) \quad (13)$$

$$\text{BW}_{\text{avg}} = \sum_i \text{BW}_i / N_T \quad (14)$$

$$\bar{f}_i = \sum_j P(i, j) f(i, j) \Big/ \sum_j P(i, j) \quad (15)$$

$$\bar{f} = \sum_i \bar{f}_i / N_T \quad (16)$$

$$\text{BW}_i^r = \text{BW}_i / \bar{f}_i \quad (17)$$

$$\text{BW}^r = \sum_i \text{BW}_i^r / N_T \quad (18)$$

$$P_i = 10 \log_{10} \sum_j P(i, j) \quad (19)$$

$$\bar{f}^w = \sum_i \bar{f}_i P_i \Big/ \sum_i P_i \quad (20)$$

where N_T is the number of time steps, and $P(i, j)$ is the power for time step index i and frequency band index j. The final feature vector is defined as

$$\text{features} = [f_{\min}, f_{\max}, f^c, \text{BW}, \text{BW}_{\text{avg}}, \bar{f}, \tau, \text{SNR}, \text{slope}, \text{curvature}, \text{area}, \bar{f}^w, \text{BW}^r]. \quad (21)$$

We trained a set of classifiers including SVM, multilayer perceptron (MLP) neural network, DT, and random forest (RF). Data are first normalized so the features have zero mean and unit variance. For DT, we use Gini impurity criterion for splitting the features. For RF, we set the number of estimators to 100 trees, each trained on a subset of training data samples with replacement and the final prediction is the majority vote of DTs. For MLP, we set the number of

fully connected layers to 5, each containing 30 neurons and rectified linear unit (ReLU) activation functions. Stochastic gradient descent (SGD) with batch size 200 and Adam [56] optimizer with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e^{-8}$ is used to train the network by minimizing categorical cross-entropy loss. Weight decay of $1e^{-4}$ is used to reduce overfitting. For SVM, we set the parameter $C = 10$ and used RBF kernel with parameter $\gamma = 1/\text{numfeatures} = 1/13$ when 13 features are used or 1/7 when seven features are used.

III. Results

In this section, we apply the methods developed and discussed above to detect, analyze, and classify the significant sound sources including MM vocalization signals in the GOM 2006 and USNE 2021 experimental datasets. First, we provide a comprehensive time–frequency features' importance analysis, and then evaluate performances of various classifiers for distinguishing calls of different MM species present in our datasets. Next, we present the detection and bearing estimation results from the two experiments, including computational complexity analysis for real-time performance.

A. Features Importance for Classification

To analyze the separability of classes and features, we apply linear discriminant analysis (LDA) and principal component analysis (PCA) to reduce the feature dimensions from 13 to 3. Both LDA and PCA are methods for linear mapping where LDA is supervised while PCA is unsupervised. All data other than the test day data are used to find linear mapping using PCA and LDA which are then applied to the test data. Fig. 4 shows the result of this mapping on the test data day of October 2, 2006. It can be observed that the class Odontocestes is highly separable from the other classes because of a different frequency range for this class compared with other classes. The Sei and UBDC classes as well as the Humpback and Minke classes are not as well-separated because of overlapping frequency range and duration of calls in these class pairs, respectively.

To analyze features' importance, we calculate the mean decrease in impurity (MDI) for the RF classifier. The MDI measure counts the number of times a feature is used to split a node weighted by the number of samples it splits and is plotted in Fig. 5 for classification of detections into the six whale categories using the RF classifier. This calculation demonstrates that the frequency-based features have higher importance and play a more significant role in the classification process using the RF classifier.

Some of the features we selected are collinear or highly correlated, for instance, the mean instantaneous bandwidth BW_{avg} and total bandwidth BW. To quantify features correlation, we calculate Pearson's and Spearman's correlation coefficients for every pair of features. Pearson's correlation is a good indicator for Gaussian distribution, and there are no extreme outliers in the data. However, these assumptions may not hold for some datasets, such as the ordinal (categorical) variables and also non-Gaussian distributions. Spearman's correlation coefficient is a nonparametric measure of rank correlation

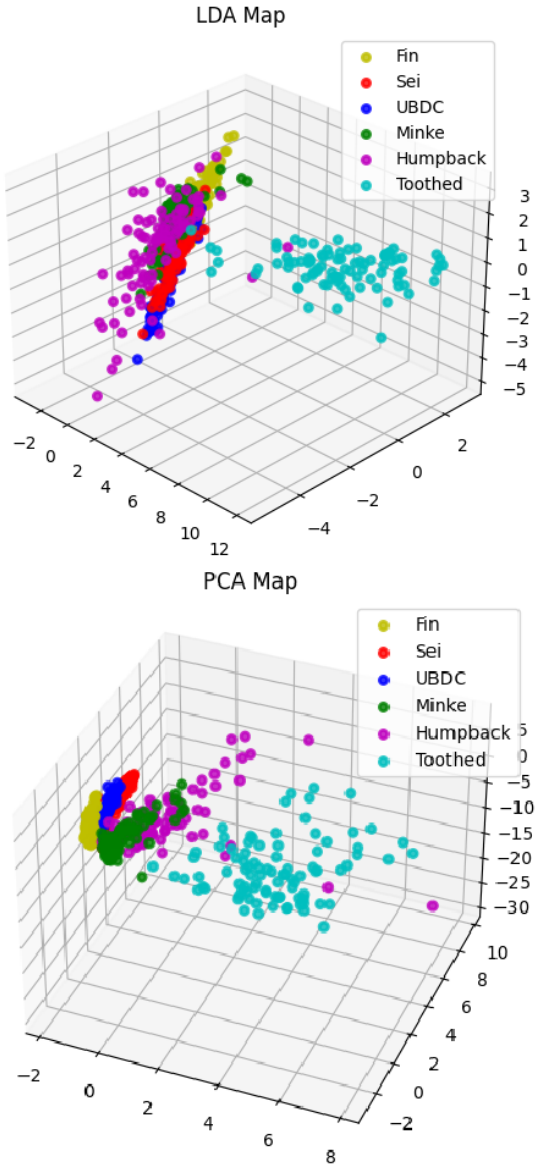


Fig. 4. Dimension reduction on the extracted features to visualize the separability of classes using LDA and PCA.

that can be used for both continuous and discrete ordinal variables and is relatively robust against outliers [57]. It is basically a Pearson correlation on the ranked variables, where the values are ordered and assigned as integers, instead of real numbers [58]. The absolute value of Spearman's correlation coefficients is plotted in Fig. 6(a).

Next, we apply Ward's minimum variance method for hierarchical agglomerative clustering of features based on Spearman's correlation, where a distance matrix is formed using the formula: $d(a, b) = 1 - |\text{corr}(a, b)|$, where $d(a, b)$ is the distance between two features a and b, and $|\text{corr}(a, b)|$ is the absolute value of Spearman's correlation between a and b. Fig. 6(b) shows the hierarchy linkage dendrogram for Ward's agglomerative clustering method. It can be seen that Ward's distances between $\tilde{f}^w$, $\tilde{f}^w$, f^c are low which imply that these features are highly correlated and similar. To obtain a more reliable estimate of feature importance, just one feature is

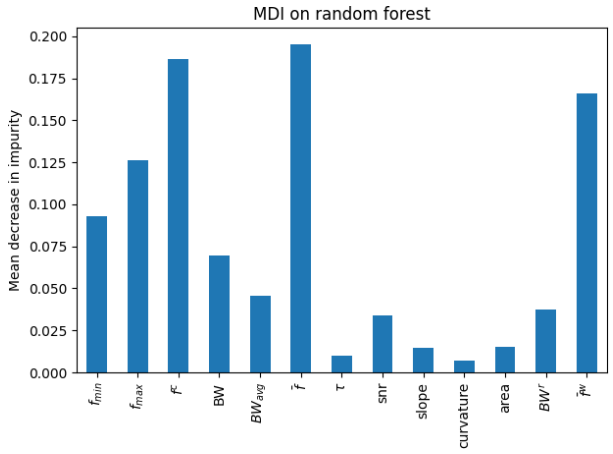


Fig. 5. Feature importance for the features used for six whale categories' classification using MDI on all 13 features set, and PFI measure on seven features selected using Ward's minimum variance method.

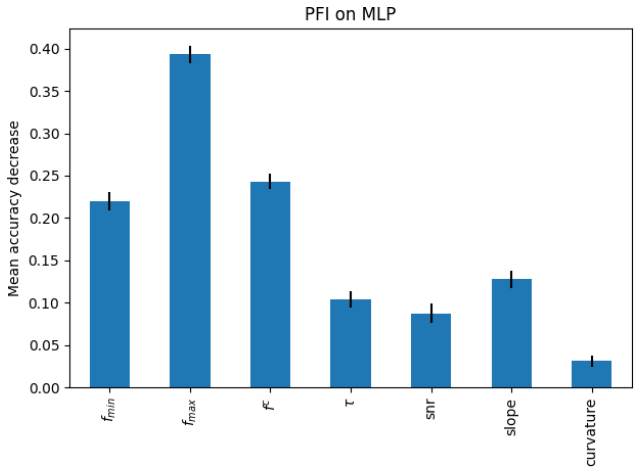


Fig. 7. Feature importance using PFI measure on seven features selected using Ward's minimum variance method.

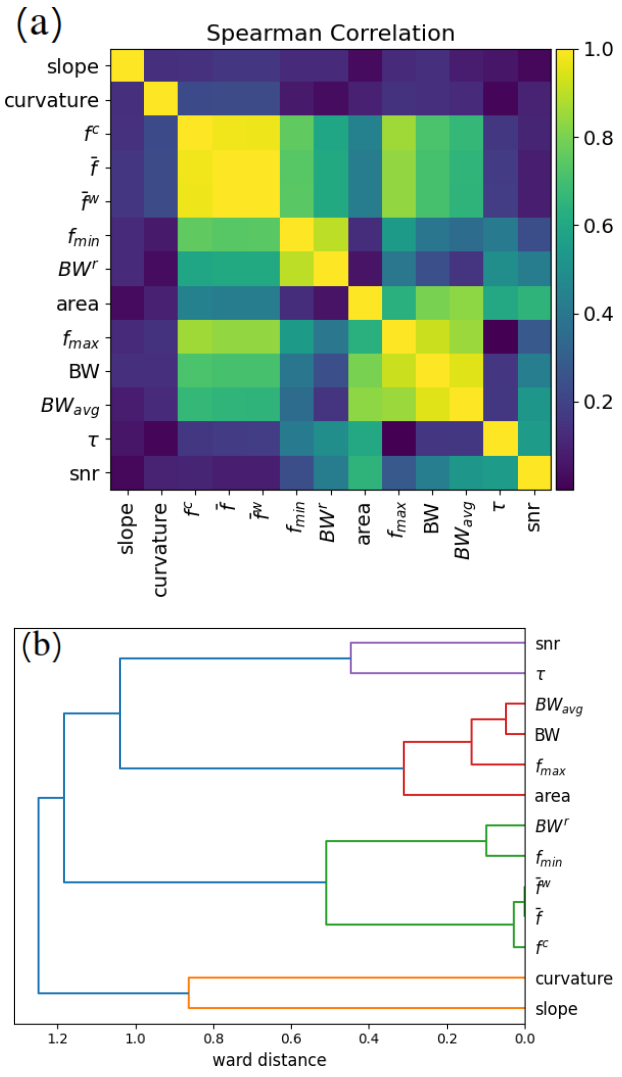


Fig. 6. (a) Spearman's correlation coefficients for the extracted features. (b) Hierarchy linkage dendrogram for Ward's agglomerative clustering method.

selected per highly correlated cluster, where the threshold for Ward's linkage clustering distance is set to 0.4. The resultant seven features used are: $f_{\min}$, $f_{\max}$, f^c, τ, SNR, slope, and curvature. The classifiers are retrained and evaluated on this subset of features, and the features' importance is recalculated.

The MDI measure has several limitations including: 1) being sensitive to overfitting because it is computed on just training data; 2) can only be used for tree-based models as it is quantified by splitting criterion; and 3) is biased to high cardinality such as numerical features compared with binary or categorical features. To overcome these limitations, another measure called permutation feature importance (PFI) is used. It can be used with any model and on both the training and test data. The PFI is defined as the decrease in a model score when a feature value is randomly shuffled [59]. If shuffling a feature results in a greater decrease in model score, it implies that the model is more dependent on that feature and in consequence that feature is more important for the classification using the specific model. Fig. 7 shows the PFI on MLP classifier on test data after seven feature selection using Ward's minimum variance method from October 2, 2006. For each feature, the permutation is repeated five times, and the final feature importance measure is the average of these five repeats. The standard deviations of PFI for these five repeats are shown as the black vertical lines around the average values, which is the top of bars in the plot.

B. Classification Results

We trained the classifiers on the GOM 2006 dataset. To evaluate the performance, we separate the training and test data by data collection day. Out of the 13 data collection days, we selected one day as the test day and the remaining 12 days for training. We repeated this procedure 13 times to get a complete set of predicted labels for all the days. For the first training set, 10% of training data are selected randomly to be used as the validation data to fine-tune the parameters for the classifiers, and these parameters are used for the rest of the experiment. Confusion matrices resulting from test data on October 3, 2006 are shown in Fig. 8. Each cell in the confusion matrix contains two numbers. The top quantity provides the

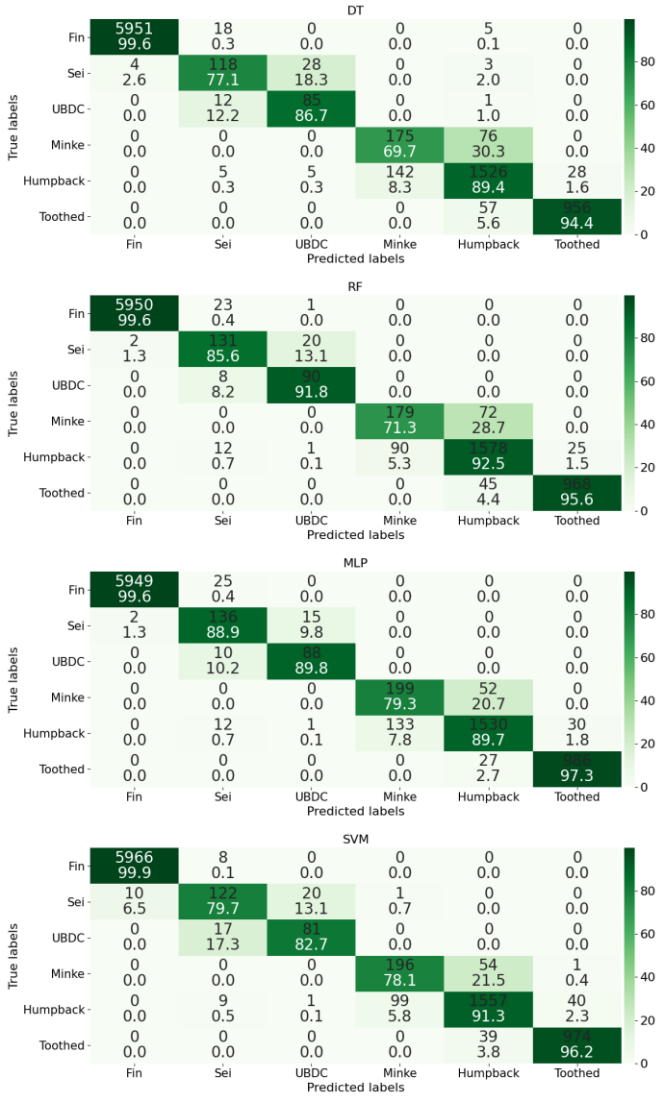


Fig. 8. Confusion matrices for whale species call classification using different classifiers on October 3, 2006 GOM dataset. Each cell in the confusion matrix contains two numbers. The top quantity provides the number of calls and the bottom is the percentage of calls automatically classified to the specific class.

number of calls and the bottom is the percentage of calls automatically classified to the specific class.

It can be seen from the confusion matrices that Fin and Toothed whales are highly discriminant from the rest of the classes, mainly because their frequency ranges do not overlap with those of the other classes. Misclassifications between Sei and UBDC, and between Minke and Humpback occur frequently since these classes have overlapping ranges for many features that make their classification more challenging.

Assuming the specific test data are drawn from a distribution, resampling the test data can change the performance measures such as accuracy or F1 score. This implies that for unseen future test data, the performance measures can vary, so these measures should be treated as random variables. To numerically express this uncertainty in the measures and better represent the performance estimation of models on unseen data, we calculate the F1 score confidence intervals

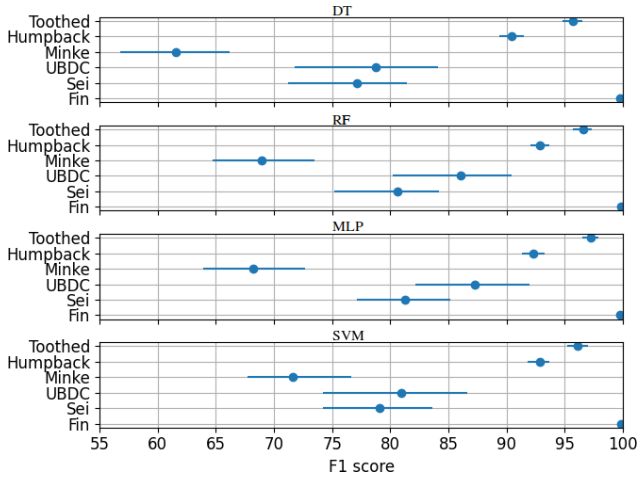


Fig. 9. F1 scores with confidence intervals for whale species vocalization classification using different classifiers on October 3, 2006 GOM dataset. The means are indicated by circles and the 95% confidence intervals by horizontal bars.

TABLE I
CLASSIFICATION PERFORMANCE MEASURES ON THE 13 EXTRACTED FEATURES AND SEVEN SUBSETS OF EXTRACTED FEATURES FOLLOWING WARD'S METHOD

	13 features			7 features		
	Acc	f1-m	f1-w	Acc	f1-m	f1-w
DT	95.82	83.90	95.96	95.80	84.26	95.95
RF	**96.75**	87.30	**96.80**	**96.75**	**87.55**	**96.83**
MLP	96.66	**87.60**	**96.80**	96.70	87.50	96.80
SVM	96.75	86.72	**96.80**	96.35	83.89	96.38

for each class using a bootstrapping [60] method which resamples the data with replacement [61]. Confidence interval is a method to compute the lower and upper bounds around the mean estimated value. After training the classifiers on training data, we resample the test data 200 times with replacement, and recalculate the F1 score each time. For the 95% confidence interval, we select the 2.5th and 97.5th percentiles of the 200 F1 scores as the lower and upper bounds, respectively. Fig. 9 shows the mean F1 score and 95% confidence interval for each class. It can be seen that uncertainties in F1 score for Sei, UBDC, and Minke whale vocalizations are larger than those for Fin, Humpback, and Toothed whales.

As can be seen from Table I, RF outperforms the other classifiers based on the seven subsets of extracted features. The performance measures including accuracy, macro average F1 score and weighted average F1 score for the test date of October 3, 2006 using seven subsets of extracted features are the highest with the RF classifier. In macro average, F1 score is first calculated for each class and then unweighted average values of these F1 scores are calculated. For the weighted average F1 score, after calculating the F1 score for each class, their average, weighted by the number of true instances for each class, is calculated.

The results and methods described above are for the classification of predetermined whale sound detections into the six specified categories. To enable whale sound classification into the six specified categories in the presence of other acoustic events, such as ship-generated noise, fish sound,

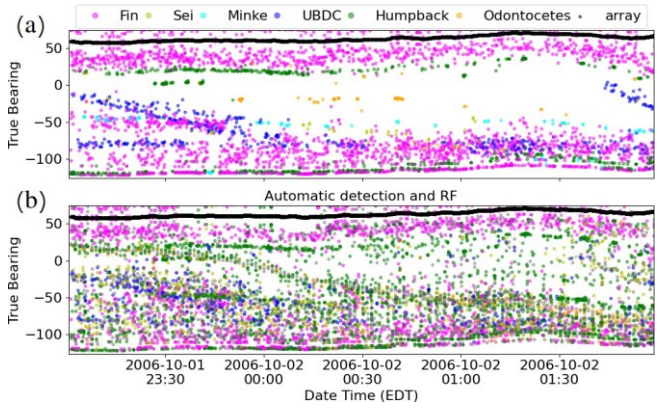


Fig. 10. Bearing-time trajectories for labeled and classified detections based on (a) manual inspection and (b) automatic processing for data below 4 kHz from the GOM 2006 Experiment dataset. Only the right-side bearings are shown.

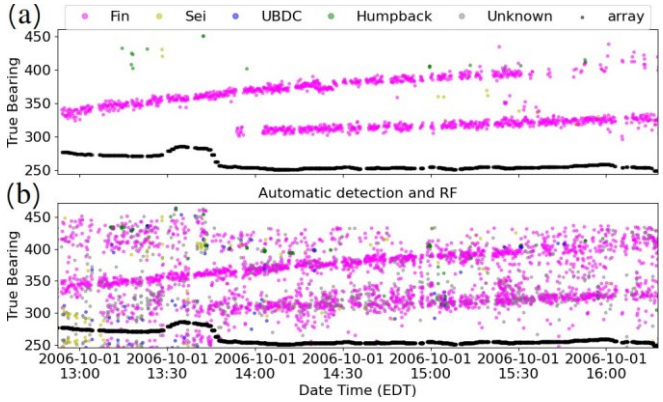


Fig. 11. Bearing-time trajectories for labeled and classified detections based on (a) manual inspection and (b) automatic processing for data below 110 Hz from the GOM 2006 Experiment dataset. Only the left-side bearings are shown.

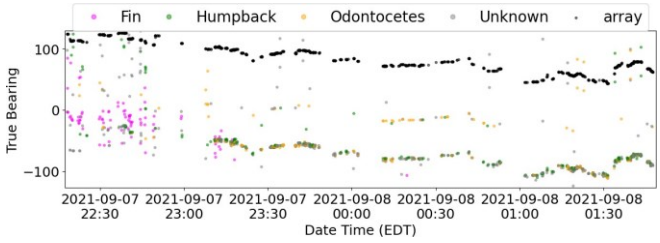


Fig. 12. Bearing-time trajectories, result of automatic detection, bearing estimation, and RF classification for the USNE 2021 experiment for detections up to 5 kHz only. The black dots show array heading direction with respect to true north direction.

or calls produced by other whale classes, we first compute the mean and standard deviation of the features for each of the six specified categories. Then for an arbitrary detected event, we apply one of the classifiers above to predict a label from any of the six predefined categories. Next we compute distance of the detected event's features from the mean value of the predicted class, and if this distance is more than three times the standard deviation of the predicted class' features, we label this event as an unknown class.

The bearing-time trajectories from automatic detection and labeling for signals in the six whale sound categories are plotted in Fig. 10 for three hours of recording from the GOM 2006 dataset. The corresponding results obtained from manual detection and classification are also shown for comparison. The detections are numerous since they span the full frequency range of the GOM 2006 dataset between 10 and 4000 Hz. We next focus on a subset of detections, below 110-Hz frequency range in Fig. 11 for closer comparison between the automatic and manual processing operations. It can be seen that the automatic approach provides bearing-time trajectories for fin whale calls that are highly similar to those obtained from manual processing. The detections forming well-defined bearing-time trajectories have high SNRs. We also note that there are more background detections in the automatic approach compared with the manually labeled calls. Part of these background detections are lower SNR fin whale calls which were ignored in the manual approach, and the remainder are random background noise.

We next apply the classifiers trained on the GOM 2006 dataset to a new dataset from the USNE 2021 experiment to find potential whale calls there. The USNE 2021 dataset is sampled at various frequencies up to 100 kHz per hydrophone element. Here, we focus on a subset of detections below 4 kHz in Fig. 12 for four hours of recording from 22:00 September 7, to 02:10 am September 8, 2021. The automatic detection and labeling using the trained RF classifier is effective in detecting many fin, humpback, and toothed whales call types, consistent with visual inspection. There are some false detections as well, especially in the frequency range of humpback calls due to other background sound sources. The sample spectrograms containing MM vocalizations from the USNE 2021 experiment are shown in [38, Figs. 8–11].

The computational complexities and run times for different stages of the POAWRS processing algorithms were analyzed in [36]. There, it was shown that real-time performance is achievable for processing 60-s-long acoustic data from 132 hydrophones, each sampled at 100 kHz and beamformed into 147 distinct relative bearing directions, followed by detection processing in full 360° horizontal azimuthal directions simultaneously. Speed comparison for runs on CPU and GPU for different algorithms was investigated, as well as FFT-based versus time domain delay and sum beamforming performance comparison. We showed that significantly faster processing time for most algorithms ran on GPU compared with CPU, for instance, about 338 times faster beamforming on GPU, enabling large-aperture coherent hydrophone array data to be analyzed in real-time.

IV. Conclusion

Instantaneous wide- area POAWRS technology implemented with a large-aperture densely populated coherent hydrophone array has been advanced here in several crucial ways. Taking advantage of combined hardware and software advances and optimizations, the analysis and processing of large-aperture high-sample rate hydrophone array data has achieved real-time performance combining multiple stages of processing. These include beamforming that enhances signal SNR, acoustic event detection, bearing estimation,

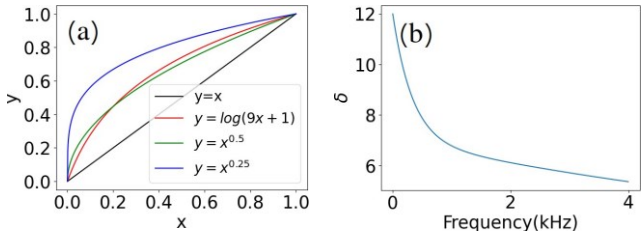


Fig. 13. (a) Nonlinearity behavior for log- and exponential transforms. (b) Frequency-dependent plot of parameter δ in PCEN showing exponentially decreasing function of frequency.

and classification for data sampled at 100 kHz per hydrophone element and beamformed in 147 directions. This huge volume of data requires special considerations to achieve real-time performance for underwater acoustics events' monitoring which we addressed in this study. Various methods including those widely used in other domains, such as computer vision and perception, have been adapted here to underwater acoustic data to improve signal detection. These methods include mel-scale frequency transformation, morphological image processing, PCEN, and NMS. A variety of machine learning methods have been incorporated to automatically classify MM vocalizations in several classes. We analyzed features' importance and class separability for six whale vocalization categories and achieved average F1 score of 87.5%.

Detection of calls with low SNR or large discontinuity in frequency or time is challenging for the image-processing-based approach implemented here. In addition, there are many other underwater acoustic events, such as ship-radiated sounds that can be detected or classified falsely. Applying other machine learning approaches such as CNNs can improve robustness in the classification. However, these approaches usually require large volume of labeled data. In the future, we plan to apply unsupervised learning approaches, which do not require manually labeled data, to cluster underwater acoustic events, and that can significantly reduce the burden and complexity of manual labeling.

APPENDIX
PCEN DERIVATION

To derive the formulation for PCEN [44], [45], a nonlinear monotonically increasing transformation for DRC is first selected. This transformation is important for applications where most values, such as the Fourier transform magnitude coefficients, are much smaller than a few large ones. This leads to poor contrast for small values and potential instability in numerical computations, for instance, when taking products of small values, falling below quantization accuracy thresholds, and numerical overflow or underflow may occur. For example, for input data in the range from 1 to 10^{12}, using a nonlinear transformation, such as the logarithm, will map the dynamic range to 0–12, which is much more compressed and so numerically more stable. An advantage of the transformation is that it mimics the hearing sensitivity of humans which resembles a logarithmic scale rather than a linear scale.

Fig. 13(a) shows the nonlinear behavior for three functions, $x^{1/2}$, $x^{1/4}$, and $\log x$, which increase the contrast for small values while compressing the dynamic range for larger values. Instead of log-transform, the function x^r, where $0 < r < 1$, can be used instead. From Fig. 13, it can be noted that the nonlinearity behavior of log-transform and $x^{0.5}$ is highly similar. Reducing the value of r results in more stretching (increasing the contrast) for smaller values and more compression (reducing contrast) for larger values. This is especially useful for adaptive DRC. For example, a smaller value of r may be used if the foreground audio source is relatively weak. Another advantage is that we can use different values of r for different regions of the signal and also optimize its value using any gradient-based optimization approach.

Let the log-transform of instantaneous signal energy at time t and frequency band f be $\log(\mathbf{E}(t, f))$. To enhance transient signal detection, for example, a whale call signal in the presence of random ambient noise and/or other persistent background signals, such as ship-radiated narrowband tonals, the log-transform of time-averaged signal energy $\mathbf{M}(t, f)$ can be subtracted from the log of instantaneous signal energy. The parameter $0 < \alpha < 1$ is introduced to control the amount of background noise cancellation

$$\log(\mathbf{E}(t, f)) - \alpha \log(\mathbf{M}(t, f)) = \log \left(\frac{\mathbf{E}(t, f)}{\mathbf{M}(t, f)^{\alpha}} \right). \quad (22)$$

Using the exponential transform for dynamic range control instead of log-transform, we have

$$\left(\frac{\mathbf{E}(t, f)}{(\epsilon + \mathbf{M}(t, f))^{\alpha}} \right)^{r} = (\mathbf{G}(t, f))^{r} \quad (23)$$

where ϵ is introduced to prevent numerical instability caused by dividing by too small numbers in the quiet background regions. A soft threshold parameter $\delta > 1$ is introduced to adjust the DRC

$$\mathbf{PCEN}(t, f) = \left(\frac{\mathbf{E}(t, f)}{(\epsilon + \mathbf{M}(t, f))^{\alpha}} + \delta \right)^{r} - \delta^{r}. \quad (24)$$

For quiet regions where $G \ll \delta$, PCEN will be close to 0, and for loud regions where $G \gg \delta$, PCEN will be higher and closer to G^r. For noisy regions where the value of G fluctuates more because of higher background noise, the larger value of δ is desirable. This parameter is useful, especially for underwater acoustic applications where the ambient noise level varies by time and frequency. On average, the sound pressure level of ocean acoustic ambient or background noise decays exponentially as a function of increasing frequency [62]. We propose the following exponential decay formula for δ as a function of increasing frequency:

$$\delta = 2 + 5e^{-af} + 5e^{-bf} \quad (25)$$

where a and b are constants, and f is the frequency. Based on our experimental data, $a = 0.003$ and $b = 0.0001$ are empirically determined to match the ambient noise in our dataset. The resultant plot for δ as a function of frequency is shown in Fig. 13(b).

The time-averaged signal energy $\mathbf{M}$ mainly carries the loudness profile, which is the stationary background noise. There are various approaches for estimating this background noise. In [44], a first-order IIR filter was used to estimate $\mathbf{M}$,

useful when the background noise mean is highly variable. Here, since the background noise within each data file of roughly 60 s duration is fairly consistent, a global average value for each frequency band within each recorded file can be used. Compared with the original PCEN [44], we apply another modification to compute **M** and use median value over time instead of average value. This is to reduce the effect of high SNR foreground event on background noise estimation, since the median value is less sensitive to extremely large or small values in the data compared with the mean value. Using median instead of mean can also be justified in the context of Bayes risk minimization, where if the mean square error or mean absolute error is used as the Bayes risk to be minimized, then we obtain the mean and median values as the solution, respectively. The absolute value of error is less sensitive to extreme values compared with the squared value of error [63].

REFERENCES

[1] D. Wang et al., "Vast assembly of vocal marine mammals from diverse species on fish spawning ground," *Nature*, vol. 531, no. 7594, pp. 366–370, Mar. 2016.

[2] R. P. Hodges, *Underwater Acoustics: Analysis, Design and Performance of Sonar*. Hoboken, NJ, USA: Wiley, 2011.

[3] K. W. Chung, A. Sutin, A. Sedunov, and M. Bruno, "DEMON acoustic ship signature measurements in an urban harbor," *Adv. Acoust. Vib.*, vol. 2011, pp. 1–13, 2011, Art. no. 952798.

[4] C. Zhu, S. G. Seri, H. Mohebbi-Kalkhoran, and P. Ratilal, "Long-range automatic detection, acoustic signature characterization and bearing-time estimation of multiple ships with coherent hydrophone array," *Remote Sens.*, vol. 12, no. 22, p. 3731, Nov. 2020.

[5] S. G. Seri, C. Zhu, M. Schinault, H. Garcia, N. O. Handegard, and P. Ratilal, "Long range passive ocean acoustic waveguide remote sensing (POAWRS) of seismo-acoustic airgun signals received on a coherent hydrophone array," in *Proc. OCEANS*, Oct. 2019, pp. 1–8.

[6] A. N. Popper et al., "Effects of exposure to seismic airgun use on hearing of three fish species," *J. Acoust. Soc. Amer.*, vol. 117, no. 6, pp. 3958–3971, Jun. 2005.

[7] H. Mohebbi-Kalkhoran, C. Zhu, M. Schinault, and P. Ratilal, "Classifying humpback whale calls to song and non-song vocalizations using bag of words descriptor on acoustic data," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 865–870.

[8] T. A. Jefferson, M. A. Webber, and R. Pitman, *Marine Mammals of the World: A Comprehensive Guide to Their Identification*. Amsterdam, The Netherlands: Elsevier, 2011.

[9] Y. Yuan et al., "Comparative genomics provides insights into the aquatic adaptations of mammals," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 37, 2021, Art. no. e2106080118.

[10] D. Cato, R. McCauley, T. Rogers, and M. Noad, "Passive acoustics for monitoring marine animals-progress and challenges," in *Proc. ACOUSTICS*, 2006, pp. 453–460.

[11] D. Wang, W. Huang, H. Garcia, and P. Ratilal, "Vocalization source level distributions and pulse compression gains of diverse baleen whale species in the Gulf of Maine," *Remote Sens.*, vol. 8, no. 11, p. 881, Oct. 2016.

[12] Z. Gong et al., "Low-frequency target strength and abundance of shoaling Atlantic herring (*Clupea harengus*) in the Gulf of Maine during the ocean acoustic waveguide remote sensing 2006 experiment," *J. Acoust. Soc. Amer.*, vol. 127, no. 1, pp. 104–123, Jan. 2010.

[13] N. C. Makris et al., "Critical population density triggers rapid formation of vast oceanic fish shoals," *Science*, vol. 323, no. 5922, pp. 1734–1737, Mar. 2009.

[14] Z. Gong et al., "Ecosystem scale acoustic sensing reveals humpback whale behavior synchronous with herring spawning processes and re-evaluation finds no effect of sonar on humpback song occurrence in the Gulf of Maine in fall 2006," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e104733.

[15] A. Galatius et al., "Raising your voice: Evolution of narrow-band high-frequency signals in toothed whales (Odontoceti)," *Biol. J. Linnean Soc.*, vol. 126, no. 2, pp. 213–224, Jan. 2019.

[16] D. D. Tran et al., "Using a coherent hydrophone array for observing sperm whale range, classification, and shallow-water dive profiles," *J. Acoust. Soc. Amer.*, vol. 136, no. 6, p. 2093, Oct. 2014.

[17] 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.*, vol. 9, no. 1, pp. 1–10, Aug. 2019.

[18] J.-J. Jiang et al., "Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network," *Appl. Acoust.*, vol. 141, pp. 26–34, Dec. 2018.

[19] L. Zhang, D. Wang, C. Bao, Y. Wang, and K. Xu, "Large-scale whale-call classification by transfer learning on multi-scale waveforms and time-frequency features," *Appl. Sci.*, vol. 9, no. 5, p. 1020, Mar. 2019.

[20] 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. Amer.*, vol. 147, no. 3, pp. 1834–1841, Mar. 2020.

[21] C. Bergler et al., "ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning," *Sci. Rep.*, vol. 9, no. 1, pp. 1–17, Jul. 2019.

[22] J. H. Rasmussen and A. Širovic´, "Automatic detection and classification of baleen whale social calls using convolutional neural networks," *J. Acoust. Soc. Amer.*, vol. 149, no. 5, pp. 3635–3644, May 2021.

[23] A. K. Ibrahim, H. Zhuang, L. M. Chérubin, N. Erdol, G. O'Corry-Crowe, and A. M. Ali, "A multimodel deep learning algorithm to detect North Atlantic right whale up-calls," *J. Acoust. Soc. Amer.*, vol. 150, no. 2, pp. 1264–1272, Aug. 2021.

[24] J.-J. Jiang et al., "Whistle detection and classification for whales based on convolutional neural networks," *Appl. Acoust.*, vol. 150, pp. 169–178, Jul. 2019.

[25] H. A. Garcia et al., "Comparing performances of five distinct automatic classifiers for fin whale vocalizations in beamformed spectrograms of coherent hydrophone array," *Remote Sens.*, vol. 12, no. 2, p. 326, Jan. 2020.

[26] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 201–212, Jun. 1976.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[28] M. F. Baumgartner and S. E. Mussoline, "A generalized baleen whale call detection and classification system," *J. Acoust. Soc. Amer.*, vol. 129, no. 5, pp. 2889–2902, May 2011.

[29] W. Huang, D. Wang, and P. Ratilal, "Diel and spatial dependence of humpback song and non-song vocalizations in fish spawning ground," *Remote Sens.*, vol. 8, no. 9, p. 712, Aug. 2016.

[30] D. Paraskevopoulos, C. Spandonidis, and F. Giannopoulos, "Hybrid wavelet–CNN fault diagnosis method for ships' power systems," *Signals*, vol. 4, no. 1, pp. 150–166, Feb. 2023.

[31] E. Ghaderpour and S. Ghaderpour, "Least-squares spectral and wavelet analyses of V455 andromedae time series: The life after the super-outburst," *Publications Astronomical Soc. Pacific*, vol. 132, no. 1017, Oct. 2020, Art. no. 114504.

[32] W. J. Overholtz, J. M. Jech, W. L. Michaels, L. D. Jacobson, and P. J. Sullivan, "Empirical comparisons of survey designs in acoustic surveys of Gulf of maine-georges bank Atlantic herring," *J. Northwest Atlantic Fishery Sci.*, vol. 36, pp. 127–144, Sep. 2006.

[33] K. M. Becker and J. R. Preston, "The ONR five octave research array (FORA) at Penn state," in *Proc. Oceans. Celebrating Past. Teaming Toward Future*, Jul. 2003, pp. 2607–2610.

[34] D. Wang and P. Ratilal, "Angular resolution enhancement provided by nonuniformly-spaced linear hydrophone arrays in ocean acoustic waveguide remote sensing," *Remote Sens.*, vol. 9, no. 10, p. 1036, Oct. 2017.

[35] M. I. Sezan, "A peak detection algorithm and its application to histogram-based image data reduction," *Comput. Vis., Graph., Image Process.*, vol. 47, no. 3, p. 396, Sep. 1989.

[36] H. Mohebbi-Kalkhoran, M. Schinault, N. C. Makris, and P. Ratilal, "Integrated computing system for real-time data processing, storage and communication with large aperture 160-element coherent hydrophone array," in *Proc. OCEANS*, Oct. 2022, pp. 1–9.

[37] P. Ratilal et al., "Continental shelf-scale passive ocean acoustic waveguide remote sensing of marine ecosystems, dynamics and directional soundscapes: Sensing whales, fish, ships and other sound producers in near real-time," in *Proc. OCEANS*, Oct. 2022, pp. 1–7.

[38] M. E. Schinault et al., "Development of a large-aperture 160-element coherent hydrophone array system for instantaneous wide area ocean acoustic sensing," in *Proc. OCEANS*, Oct. 2022, pp. 1–9.
[39] M. K. Radermacher, M. E. Schinault, S. G. Seri, and P. Ratilal, "Research and design for the power hierarchy of a 160-element linear towable ocean acoustic coherent hydrophone array," in *Proc. OCEANS*, Oct. 2022, pp. 1–7.
[40] H. Mohebbi-Kalkhoran, "Machine learning approaches for classification of myriad underwater acoustic events over continental-shelf scale regions with passive ocean acoustic waveguide remote sensing," Ph.D. dissertation, Dept. Elect. Comput. Eng., Northeastern Univ., Boston, MA, USA, 2022.
[41] J. V. DiFranco and W. L. Rubin, *Radar Detection*. Upper Saddle River, NJ, USA: Prentice-Hall, 1968.
[42] P. Chiariotti, M. Martarelli, and P. Castellini, "Acoustic beamforming for noise source localization—Reviews, methodology and applications," *Mech. Syst. Signal Process.*, vol. 120, pp. 422–448, Apr. 2019.
[43] H. A. Garcia et al., "Temporal–spatial, spectral, and source level distributions of fin whale vocalizations in the Norwegian sea observed with a coherent hydrophone array," *ICES J. Mar. Sci.*, vol. 76, no. 1, pp. 268–283, 2018.
[44] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5670–5674.
[45] V. Lostanlen et al., "Per-channel energy normalization: Why and how," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 39–43, Jan. 2019.
[46] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937.
[47] D. O'shaughnessy, *Speech Communications: Human and Machine*. Piscataway, NJ, USA: IEEE Press, 2000.
[48] S. Pednekar, A. Krishnadas, B. Cho, and N. C. Makris, "Weber's law of perception is a consequence of resolving the intensity of natural scintillating light and sound with the least possible error," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 479, no. 2271, Mar. 2023, Art. no. 20220626.
[49] A. Rosenfeld and J. L. Pfaltz, "Sequential operations in digital picture processing," *J. ACM*, vol. 13, no. 4, pp. 471–494, 1966.
[50] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[52] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[53] Z. Gong, D. D. Tran, and P. Ratilal, "Comparing passive source localization and tracking approaches with a towed horizontal receiver array in an ocean waveguide," *J. Acoust. Soc. Amer.*, vol. 134, no. 5, pp. 3705–3720, Nov. 2013.
[54] N. C. Makris, "Imaging ocean-basin reverberation via inversion," *J. Acoust. Soc. Amer.*, vol. 94, no. 2, pp. 983–993, Aug. 1993.
[55] Z. Gong, P. Ratilal, and N. C. Makris, "Simultaneous localization of multiple broadband non-impulsive acoustic sources in an ocean waveguide using the array invariant," *J. Acoust. Soc. Amer.*, vol. 138, no. 5, pp. 2649–2667, Nov. 2015.
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[57] J. C. Caruso and N. Cliff, "Empirical size, coverage, and power of confidence intervals for Spearman's Rho," *Educ. Psychol. Meas.*, vol. 57, no. 4, pp. 637–654, Aug. 1997.
[58] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.
[59] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.
[60] B. Efron and R. J. Tibshirani, *An Introduction to Bootstrap*. Boca Raton, FL, USA: CRC Press, 1994.
[61] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," 2018, *arXiv:1811.12808*.
[62] G. M. Wenz, "Acoustic ambient noise in the ocean: Spectra and sources," *J. Acoust. Soc. Amer.*, vol. 34, no. 12, pp. 1936–1956, Dec. 1962.
[63] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.

Hamed Mohebbi-Kalkhoran received the M.Sc. degree in biomedical engineering from the Sharif University of Technology, Tehran, Iran, in 2015, and the Ph.D. degree in electrical engineering from Northeastern University, Boston, MA, USA, in 2022.

After completing his doctoral studies, he conducted postdoctoral research at the Laboratory for Undersea Remote Sensing, Massachusetts Institute of Technology, Cambridge, MA, USA. His research interests include machine learning, image and signal processing, and acoustics.

Nicholas C. Makris is a Professor of Mechanical Engineering and the Director of Acoustics, Sensing and Undersea Remote Sensing at the Massachusetts Institute of Technology, Cambridge, MA, USA. He is a Secretary of the Navy/Chief of Naval Operations Scholar of Oceanographic Sciences and a William I. Koch Professor of Marine Technology. He has explored many diverse uses of sound and other waves in sensing, including applications to ocean sensing on Earth and planetary moons, evolution of sound power efficiency in stringed musical instruments, and fundamental aspects of human auditory and visual perception.

Purnima Ratilal (Member, IEEE) is a Professor of Electrical and Computer Engineering, as well as the Director of the Laboratory for Ocean Acoustics and Ecosystem Sensing, Northeastern University, Boston, MA, USA. With over 30 years of experience in the field of ocean acoustics and remote sensing, she works with fisheries and marine mammal ecologists to monitor, census, and study population dynamics and distributions of marine mammals and fish species.