JASA REVIEW-ARTICLE

Performance metrics for marine mammal signal detection and classification

John A. Hildebrand,[1,a)] Kaitlin E. Frasier,[1,b)] Tyler A. Helble,[2,c)] and Marie A. Roch[3,d)]
[1]*Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093, USA*
[2]*Naval Information Warfare Center Pacific, San Diego, California 92152, USA*
[3]*Department of Computer Science, San Diego State University, San Diego, California 92182, USA*

ABSTRACT:
Automatic algorithms for the detection and classification of sound are essential to the analysis of acoustic datasets with long duration. Metrics are needed to assess the performance characteristics of these algorithms. Four metrics for performance evaluation are discussed here: receiver-operating-characteristic (ROC) curves, detection-error-trade-off (DET) curves, precision-recall (PR) curves, and cost curves. These metrics were applied to the generalized power law detector for blue whale D calls [Helble, Ierley, D'Spain, Roch, and Hildebrand (2012). J. Acoust. Soc. Am. **131**(4), 2682–2699] and the click-clustering neural-net algorithm for Cuvier's beaked whale echolocation click detection [Frasier, Roch, Soldevilla, Wiggins, Garrison, and Hildebrand (2017). PLoS Comp. Biol. **13**(12), e1005823] using data prepared for the 2015 Detection, Classification, Localization and Density Estimation Workshop. Detection class imbalance, particularly the situation of rare occurrence, is common for long-term passive acoustic monitoring datasets and is a factor in the performance of ROC and DET curves with regard to the impact of false positive detections. PR curves overcome this shortcoming when calculated for individual detections and do not rely on the reporting of true negatives. Cost curves provide additional insight on the effective operating range for the detector based on the *a priori* probability of occurrence. Use of more than a single metric is helpful in understanding the performance of a detection algorithm.
https://doi.org/10.1121/10.0009270

(Received 16 August 2021; revised 13 November 2021; accepted 15 December 2021; published online 25 January 2022)
[Editor: James F. Lynch] Pages: 414–427

I. INTRODUCTION

The monitoring of marine mammals using their recorded underwater sounds is a rapidly advancing field, with large datasets now available from a diverse set of locations (Au and Lammers, 2016). Methods for detection and classification are needed to allow processing of these large acoustic datasets for use in ecological assessment and monitoring (Gibb *et al.*, 2019). Manual analysis of sound data is time consuming, subject to variation of individual analysts (Nguyen Hong Duc *et al.*, 2021), and impractical for long (months to years) time periods. Segments of data with calls are often sparse (Širović *et al.*, 2014) and when calls are present, they may be obscured by interfering sounds. Marine mammal calls can be complex and variable (Allen *et al.*, 2018) and evolve seasonally or annually (McDonald *et al.*, 2009). This complexity has resulted in a variety of approaches for detection and classification including spectrogram correlation, neural networks, hidden Markov models, and frequency contour tracking, among others (Mellinger and Clark, 2000; Roch *et al.*, 2011; Frasier *et al.*, 2017; Shiu *et al.*, 2020). Test sets of annotated sounds are a vital tool to assess the performance of these approaches (Mellinger and Clark, 2006), and large-data test sets have recently become available for a range of species and settings (DCLDE, 2015, 2018). In this paper, we evaluate the metrics used for assessment and comparison of detection and classification algorithms, their strengths and weaknesses when applied to detection of underwater marine mammal sounds, and areas for their future development.

The detection and classification of marine mammal sounds to species or call type (Bittle and Duncan, 2013), is typically the first step in their study. Detection is the process of deciding whether a signal, for example, an animal-made sound, is present or absent (Helstrom, 1968). Classification is the process of labeling detected sounds. Examples of labels that can be applied by classification include determination of the species that produced a sound, identifying a specific type of call, or determining information about the caller such as the gender or behavioral state (e.g., foraging). Detection can be considered a specialized type of classifier that merely denotes the presence or absence of sounds for a specific class. Detectors can be applied to broad classes of signals, such as for any sound that was produced by a baleen whale. It is very common to use a pair of classifiers, the first

[a)]Electronic mail: jhildebrand@ucsd.edu, ORCID: 0000-0002-5418-9799.
[b)]ORCID: 0000-0002-2401-8569.
[c)]ORCID: 0000-0003-4871-9615.
[d)]ORCID: 0000-0002-0687-2059.

one detecting sounds of interest (detector) and the second one assigning detected sounds to more refined categories (classifier). Both the sounds themselves and the background ocean noise against which they must be detected, are variable in space and time. In practice, the sounds that marine mammals produce are not well known in advance (Baumann-Pickering *et al.*, 2013), and neither is the noise encountered in the ocean strictly stochastic (Livina *et al.*, 2018).

The quality of an algorithm for automatic detection and classification of these sounds is typically evaluated by analyzing how well they perform on a labeled dataset. The predictions of the algorithm are compared to the labeled training and testing data, and metrics are calculated for the quality of the algorithm performance. Based on the outcome of these metrics, a new algorithm is created and the process is repeated until an acceptable performance level has been obtained. The final algorithm is assessed by its ability to classify the evaluation data, which have not been part of the development data.

The process can entail nominal class designation, where the predicted label is compared to the actual (true) label, or numerical scoring where a quantitative value is available to designate how well a particular sound belongs to a particular class. For example, discriminant function analysis provides only nominal predictions (Oswald *et al.*, 2003), whereas support vector machines (SVM) provide a numerical value for the prediction score (Tachibana *et al.*, 2014).

With the goal of minimizing the influence of error, the quality of the algorithm should be evaluated, including determining the limits imposed by statistical uncertainty (Lehmann, 1959). In practice, detection and classification algorithms are often assessed with a single metric, and which metric is selected may be related to the effort invested in the testing and training datasets, as well as lack of familiarity with the range of metrics available. The goal of this paper is to calculate multiple metrics on a set of algorithms to compare their characteristics. We restrict our analysis to the two class (detection) case, although the metrics presented are applicable to detections from multiclass problems.

II. METHODS

A. Ground truth data

Datasets that are labeled for known marine mammal sounds, present in realistic ocean noise, are essential for training and testing of detection and classification algorithms; we refer to these as ground truth. Thus far, human analysts are needed to create ground truth datasets, primarily due to the variability and complexity of both the marine mammal sounds and the background ocean noise, although unsupervised machine learning may change our ability to produce annotated datasets in the future. Some of the factors that lead to this variability are related to the animals that produce these sounds including: changes in the source level, changes due to behavioral state (e.g., foraging, traveling, breeding), geographic variations in sound production, animals' demographic differences (e.g., age and sex), and even orientation of the animal with respect to the sensor (particularly for high frequency echolocation clicks). Other environmental factors will change the received sound including propagation in the water column, interaction with the seabed, the presence of ice, seasonal changes in these, and changing conditions of ambient noise. In addition, the hardware that is used for sound recording is important since it may add variability such as: electronic self-noise, frequency response and dynamic range, as well as introducing artifacts in the sound data (e.g., sounds of moving components) that can interfere with detection and classification of desired sounds. These geographic, temporal and instrumental variabilities require labeled datasets with enough variety that they reflect what might be encountered in a novel dataset.

Manual examination of acoustic data is often the first step in detection and classification of acoustic signals. For large data sets it may be possible to only examine a small subset of the data. However, some level of manual examination may be critical to guide future steps in the analysis. Manual analysis allows identification of the range of acoustic signals present, and the potential interference from noise sources. By manually annotating a portion of the data, development and evaluation data are generated that allow for creation and assessment, respectively, of automated detection and classification methods.

The software package TRITON provides one approach to manual data analysis and annotation, especially suited for the analysis of large data sets (Wiggins and Hildebrand, 2007). The approach used by TRITON is to calculate a long-term spectral average (LTSA) for the entire dataset, prior to manual analysis. The LTSA provides a rapid means for scanning data and selection of data subsets for more detailed examination or annotation as spectrogram or time series. In this way, it is possible to conduct manual analysis of large datasets, with manual detections representing either individual calls or presence or absence of calls over a fixed time window, for instance, in hourly bins.

Although it is tempting to assume that the manually generated labels associated with a dataset are infallibly correct, human analysts produce classification errors as well as machine algorithms, and a system designed to learn from labeled data will be unlikely to perform well when provided manual labels are of poor quality. This points out the need for widely available datasets for performance benchmarks, and a process for curated annotations to be improved over time, as errors are discovered and corrected.

B. Training, testing, and evaluation

Standardized procedures are needed for training and testing of detection and classification algorithms. When developing detectors and classifiers that learn signal characteristics from the data, it is important to separately select those data that are used to train and test the classifier, from those data that are used to evaluate performance. As

mentioned earlier, both training/testing and evaluation data should span the expected diversity of the sounds to be detected and the variability of environments in which they occur. Another cardinal rule is that the set of training and test data must be disjoint from the evaluation data. This minimizes the possibility that the algorithm is customized for some peculiar feature of the training/testing data that is not present in the evaluation data, nor in a novel dataset that is presented for analysis.

Given a labeled dataset, the available data are partitioned into development data (with further train/test partitioning) and evaluation data (Fig. 1). Care must be taken to include a broad representation of the variability of signal conditions across the training, test, and evaluation data. As an example, data from multiple animal encounters under varying conditions (e.g., background noise, location, recording equipment) should appear in all three partitions. Care is needed, however, to avoid placing different calls from the same animal encounter in both the training and test sets. The rationale for this is that such sounds are likely to artificially improve test results due to matched conditions (e.g., the same animals in the same behavioral state, recorded under the same noise conditions and acoustic propagation effects) and that such a test set is unlikely to be reflective of field performance.

Classifier construction is typically an iterative process where the results of experiments with development data are used to refine the parameters of a given classification system, with the goal of improving performance. The algorithm is retrained with the new training parameters and reevaluated with the test data in a process sometimes referred to as tuning or plowing (Fig. 1). Labeled data are partitioned into development data, used to train and test the model and evaluation data, that are used for a final assessment of performance (Campbell and Reynolds, 1999). The development data are partitioned, perhaps multiple times, into train and test datasets. Training data are used to create models that then classify the test data, with application of metrics. Analysis of errors is used to adjust the model parameters, and a new model is trained, possibly using a different partition of the development data. This process is repeated until an acceptable final model has been trained. The final model is then used to classify the evaluation data, which have not been part of the development data. We note that in some branches of machine learning, the names for these data sets differ; validation is used in place of test, and the evaluation data is called the test set (Russell and Norvig, 2020).

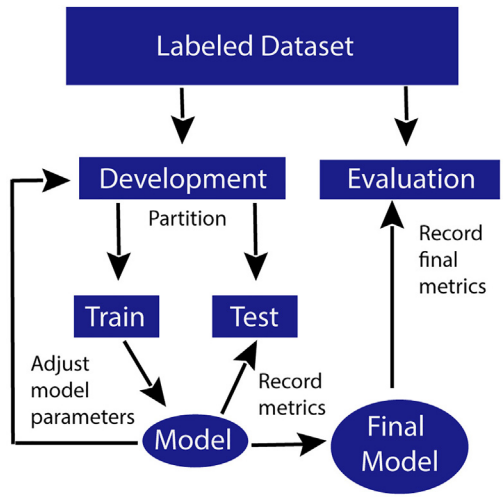

FIG. 1. (Color online) Labeled data are partitioned into development data (training and testing) and evaluation data. Data plowing is the iterative development of an algorithm by adjusting detector parameters, partitioning the development data into possibly varied nonoverlapping train and test sets training, and using metrics to evaluate performance. The evaluation data are not tested until a final algorithm has been obtained.

Examples of data plowing include adjusting model parameters such as a support vector machine's tuning parameter C (Shawe-Taylor and Zlicar, 2015) or modifying the set of features that are included in the signal description (e.g., spectra or cepstra, bandwidth, duration). This has the side-effect of making the classifier more adapted to the specific training and test datasets and can be seen as a weak form of training on the test data. As a consequence, system performance should be assessed with the use of a novel data set known as an evaluation data set. Once a classifier has been trained from the development data, it should be assessed against the evaluation data set to determine the classifier's ability to generalize to novel examples. This helps to evaluate if a classifier is too specific, or over trained, to the development data. In some circumstances, there may not be enough data for a separate evaluation dataset, but whenever possible an evaluation dataset should be used.

A key question is, how much data are needed to train, test, and evaluate a high-performance classifier? Unfortunately, there is no simple answer as it depends on factors such as the variability of signals and environments discussed above, the size of the features extracted from the sounds, and the complexity of the classifier itself. More complex classifiers typically require more data, and for complex classifiers in high-dimensional feature spaces, it is frequently difficult to have enough data.

Insufficient data can lead to overtraining. Sometimes, few data are available for algorithm development, and methods are needed to maximize use of the available data. One such method is cross-validation using either a bootstrap or an N-fold procedure. For the N-fold cross-validation approach, the development data are split into N different partitions. All but one partition is used for training, and the remaining partition is used for testing. This process is repeated, with each of the N partitions taking a turn as the test data and the other $N-1$ partitions comprising the training data for that experiment (Fig. 2). When feasible, it is recommended to have enough folds to characterize the bias and variability in performance. Kohavi (1995) notes that small N can lead to an underestimate of performance and recommends $N=10$ folds for most tasks. However, constructing $N=10$ folds should not be done at the expense of selecting appropriate partitions that prevent splitting similar data across the train/test boundary.

https://doi.org/10.1121/10.0009270

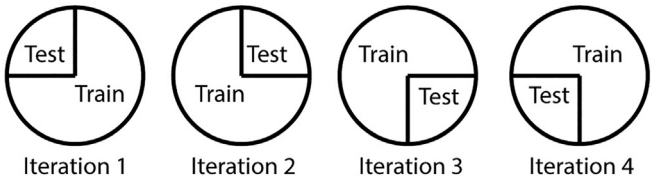

FIG. 2. Four-fold cross-validation scenario. Three different three-quarter selections of the development data are used for model training, with the remaining one-quarter selection used as test data to determine model performance.

The bootstrap cross-validation approach is similar to the *N*-fold approach except that samples are drawn for training and testing with replacement; which is to say, once selected the data are returned to the pool of available data and may be selected again for either the training or testing dataset. This approach is thought to have the advantage of providing information on the whole sampling distribution, rather than missing portions of the data when compared to the *N*-fold procedure. For example, if there are 100 000 exemplar sounds available, 70 000 of them are selected at random for training. As the random sample is with replacement, some sounds may be selected more than once. Data augmentation strategies are also increasingly popular, where variations of the same exemplar sound can be used, by moving the event around within the time window, stretching or shrinking it slightly in time or amplitude (Cui *et al.*, 2015). Repeating this process multiple times allows characterization of the expected variability in the development dataset (Gillespie *et al.*, 2013). A modification of the bootstrap procedure may increase the variability of the development data (Efron, 1982). Rather than taking a random sample from all data in the development set, data are first grouped into portions that should never be split, for instance, portions of data from the same animal encounter (sounds clustered in time and space, potentially from the same animal or a small number of animals). These un-splitable groups are then partitioned into training and test data, respectively, using a method such as cross-validation. Roch *et al.* (2015) illustrate how different splitting criteria can have non-trivial effects on overall performance.

C. Binary detection and classification

A binary classifier is one where a decision is made about the present or absence of a single type of signal (Chernoff and Moses, 1959; VanTrees, 1968). Is the signal from a particular class actually present in a specified time window, and does the algorithm predict it to be present? A binary contingency table (confusion matrix) is used to provide a tabulation between the actual and the predicted classes. The columns of the confusion matrix represent counts for the actual classes, while the rows represent counts for the predicted classes (Table I). The four resulting categories are true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Statistics are accumulated for each of the four categories into a contingency table and the detector performance is judged based on their combination

TABLE I. Confusion matrix for binary decisions. Actual classes are columns and predicted classes are rows. POS is the sum of actual positive cases detected correctly or incorrectly (TP+FN), NEG is the sum of actual negative cases (FP+TN), PPOS is the sum of predicted positive cases, PNEG is the sum of predicted negative cases, N is the total number of cases.

		Actual class		
		Positive	Negative	Total
Predicted class	Positive	TP	FP	PPOS
	Negative	FN	TN	PNEG
	Total	POS	NEG	N

into quantitative metrics. Note that in statistics, a FP is known as a type I error, and a FN is a type II error, whose probability are called alpha and beta, respectively (Lindgren, 1971).

An important aspect of the contingency table is whether or not it is constructed based on finite time windows of fixed length. If time windows are used the question becomes whether an actual and/or a predicted event are present within the specified time window. In this case all four potential outcomes of the confusion matrix are possible. In the absence of a finite time window for detection, only three of the four confusion matrix outcomes are possible (TP, FP, and FN). To determine that a TN has occurred, a finite time window is needed during which both the actual and the predicted detection were absent. The use or non-use of time windows allows the confusion matrix to be calculated on either basis, by event or by occurrence within a specified time window, but the lack of TNs in the event-based confusion matrix will limit the range of metrics that can be calculated from it, as discussed below.

Application of the confusion matrix categories to a time series that includes a click sequence from a sperm whale divided into one second segments is illustrated in Fig. 3. Five clicks are present in a patterned sequence, followed by a noise spike. Setting an amplitude threshold results in detection of four of the five sperm whale clicks as well as the noise spike. Dividing the time series into one-second time windows results in one TP (the first four clicks), one FN (the fifth click which is below the detection threshold), one FP (the noise spike), and one TN (the period between the clicks and the noise spike). Alternatively, using a detection framework without time windows results in four TP, one FN and one FP, with no assessment for TN. The difference in the balance between classes is noteworthy; in the assessment using time windows a more balanced set of outcomes is observed ($TP = FN = FP = TN$) relative to the detection-driven assessment that results in greater class imbalance ($TP > FN = FP$). In a long acoustic time series, we might expect only the occasional presence of sperm whales, and most of the time windows would not contain detections. As such, this would create a skew in the distribution of detection absent and detection present windows. As we will show, this can be problematic for many metrics. It is also important to recognize that different signal features

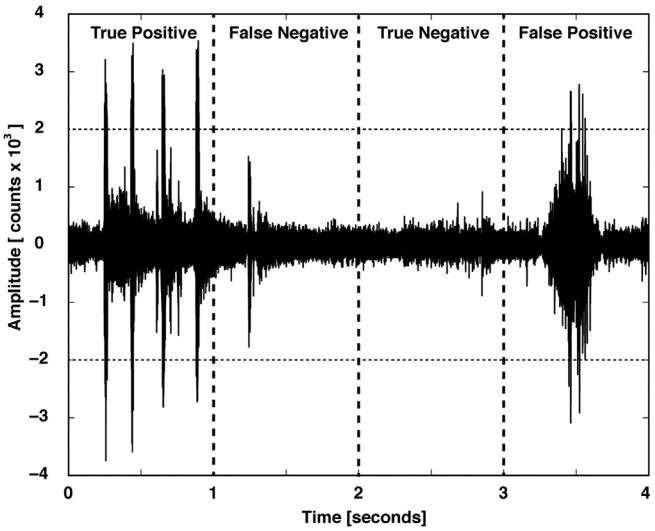

FIG. 3. Confusion matrix categories applied to a sperm whale patterned click train divided into four 1-s windows (dashed lines). The true signal consists of five rapid clicks seen between 0 and 1.3 s. The detection threshold (horizontal dotted lines) is at 50% of the maximum amplitude. In the first 1-s period (0–1 s), clicks are correctly detected (TP); between 1 and 2 s an actual click is present but below the level of the detection threshold (FN); between 2 and 3 s clicks are correctly judged to be absent (TN); and between 3 and 4 s noise is incorrectly detected as clicking (FP).

may be chosen to input to the detector. For instance, by filtering the data in Fig. 3 it may be possible to remove the noise spike and enhance the clicks, and by so doing improve the outcome of the detection problem.

Different performance metrics are calculated from the confusion matrix. A commonly used metric is the performance accuracy, which sums the diagonal elements of the confusion matrix (TP + TN) divided by the total number of cases (N),

$$Acc = \frac{TP + TN}{N}. \tag{1}$$

The drawback with accuracy is its poor performance for an imbalanced dataset, those with unequal counts in the columns of the confusion matrix. For instance, when positives (POS) are expected only 10% of the time, a classifier that designated all cases as negative would have a 90% accuracy. Likewise, the accuracy metric does not allow for non-uniform assessment of costs. The cost of missing (FN) the presence of a rare species (e.g., the North Atlantic right whale) may be much greater than the penalty for making a false detection (FP).

A classifier typically provides a numerical score for each case, and cases with higher scores are seen as being more likely to contain the target signal. Binary predictions are generated by applying a threshold to the scores. As the detector threshold is changed, the counts for each category change, and detector performance is plotted with respect to the threshold. By changing the detection threshold, the proportion of errors shifts, for instance, between FP and FN (Fig. 4).

The overlap between signal and noise distributions is set both by the difference of their means and by the shape of their distributions (Fig. 5). The final choice of operating threshold depends upon the goals of the detection task.

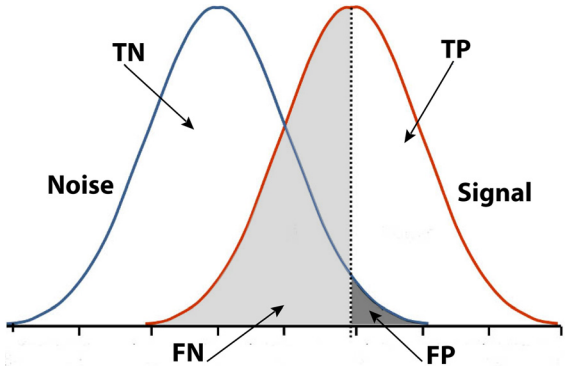

FIG. 4. (Color online) Gaussian distributions of signal (orange) and noise (blue). Threshold (dotted line) divides them into the four categories of the confusion matrix: TP (signal above threshold), FP (dark shaded area—noise above the threshold), FN (light shaded area—signal below the threshold), and TN (noise below the threshold).

Tasks that have little penalty for false positives select for a lower detection threshold, and those with little penalty for false negatives select for a higher threshold.

D. Receiver-operating-characteristic curves

The elements of the confusion matrix are used to calculate detection error curves for visualization and quantification of classifier performance. Among the most common is the receiver-operating-characteristic (ROC) curve (Peterson *et al.*, 1954; Fawcett, 2006), which is derived from each of the two columns of the confusion matrix (Table I). ROC curves plot the false positive rate (FPR) on the x axis and the true positive rate (TPR) on the y axis as follows:

$$FPR = \frac{FP}{FP + TN} \quad TPR = \frac{TP}{TP + FN}. \tag{2}$$

Values of the FPR and the TPR are plotted as the detection threshold is varied, creating a series of points, and a curve is drawn by linear interpolation between calculated points (Fig. 4). Detector performance is judged to be superior when it inhabits the upper-left-hand corner of the ROC plot, with a high probability of TP and low probability of FP. A detailed analysis of ROC curves for various statistics of signal and noise distributions is presented by Egan (1975).

E. Detection-error-tradeoff curve

An alternative representation of the confusion matrix is the detection-error-tradeoff (DET) curve (Martin *et al.*, 1997; Auckenthaler *et al.*, 2000), which plots false negative rate (FNR) on the y axis against the FPR on the x axis, expressing these values as a percentage,

$$FPR(\%) = \frac{FP}{FP + TN} \times 100,$$
$$FNR(\%) = \frac{FN}{FN + TP} \times 100. \tag{3}$$

The lower left-hand corner of the DET plot is the region of high performance. Unlike ROC plots, the DET scores are

https://doi.org/10.1121/10.0009270

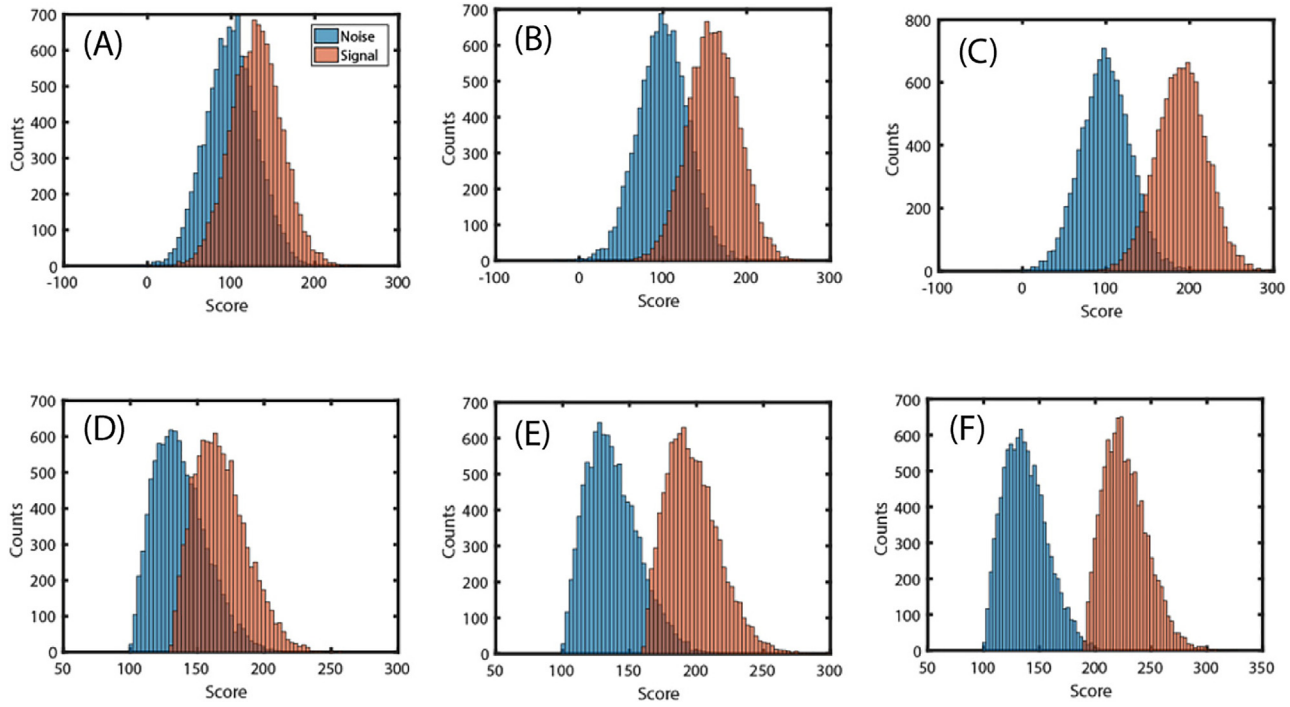

FIG. 5. (Color online) Signal (orange) and noise (blue) scores drawn from normal distributions (A, B, C) and Rayleigh distributions (D, E, F). All distributions have a mean of 100 and standard deviation of 30. Signal means are higher than noise means by one (A,D), two (B, E), or three (C, F) standard deviations.

transformed prior to display. They are scaled by their standard normal deviate (sFPR and sFNR) using the inverse error function,

$$sFPR = \sqrt{2}\, erfinv(FPR); \quad sFNR = \sqrt{2}\, erfinv(FNR), \tag{4}$$

which maps a uniform distribution [–1, 1] into a normal distribution [−inf, inf]. Using the inverse error function, normal distributions are estimated from the scores of each class and the distance in standard deviations from the mean are used as the axes of the DET curve. A consequence of this is that normally distributed score data appear as straight lines in the DET curve (Fig. 6).

Since DET curves are related to detection errors (FP and FN) on their axes, they have the advantage of being able to weight one type of error as being more important than the other. For instance, it may be important to not miss any calls when attempting to detect a rare species, so the detector threshold could be set for low FN with a corresponding increase in FP. Alternatively, it may be important to minimize FP if their cost is high. In a monitoring setting, excessive false alarms (FP) may result in lack of response to an individual alarm, and in this case one may wish to make false alarms costlier than missed detections.

F. Precision-recall curve

Precision-recall (PR) curves (Manning and Schütze, 1999) are constructed from the true positive rate, called the recall (same definition as the TPR of the ROC curve), on the x axis, and the rate at which positive predictions are correct, called the precision, on the y axis, as follows:

$$Recall = \frac{TP}{TP + FN}, \quad Precision = \frac{TP}{TP + FP}. \tag{5}$$

A superior detector in PR curve space inhabits the upper-right-hand corner of the plot, with high values for both the precision and the recall (Fig. 6). Unlike DET and ROC plots, the PR metric does not rely on the number of signal absent (TN) cases, which can lead to advantages in some situations. It has been suggested that PR curves may be superior to ROC (and by implication DET) curves, based on their ability to work with highly skewed (non-normally distributed) datasets (Davis and Goadrich, 2006), owing to their lack of dependence on signal absent (TN) cases. This type of skew is common in long-term monitoring situations where animals are only vocalizing within detection range over a relatively small portion of the analysis effort, and therefore the signal is absent most of the time.

PR curves can be calculated either using the individual detections as the trials, or they can be calculated based on fixed-length time windows. In the former, the times for the detections are all that is needed to determine the three elements of the confusion matrix (TP, FP, and FN). When there is overlap between the actual and predicted signal in time, the detection is judged to be a TP. Likewise, FP occurs when the predicted signal has no overlap with an actual signal, and a FN occurs when a true signal is not associated with a predicted signal. PR removes the need to construct trials with a time window, since it makes no use of the TN

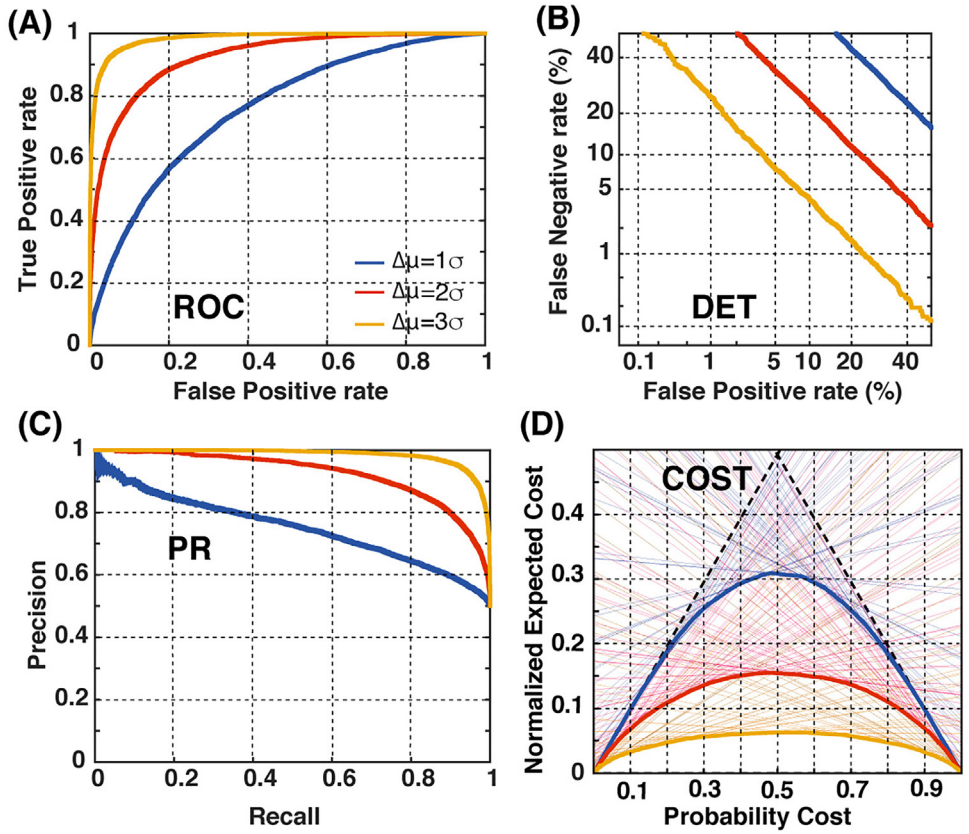

FIG. 6. (Color online) ROC (A), DET (B), PR (C), and cost curves (D) for normally distributed signal and noise from Figs. 5(A)–5(C). Mean difference ($\Delta\mu$) between signal and noise is one (blue), two (red), and three (orange) standard deviations.

parameter. Likewise, PR avoids the risk of uniform segmentation periods splitting TP signals into multiple trials. Alternatively, the PR can be assessed during defined periods for the trials, with the same procedure used for ROC or DET curves.

G. Cost curve

Cost curves provide a means for performance measurement that can be adapted to a specified cost function (Drummond and Holte, 2006). Cost curves are trial-based metrics for binary classifiers (+/− classes) that assume that there is a cost for misclassification, and that the cost for false positives C(+|−) and for false negatives C(−|+) may differ. Since it is rare for the cost of these two types of errors to be equivalent, a performance metric is needed which takes into account the differences. The probability cost function (PCF) has this characteristic,

$$PCF(+) = \frac{p(+)C(-|+)}{p(+)C(-|+) + p(-)C(+|-)} \quad (6)$$

and varies between [0,1] as a function of the potential misclassification costs [cost of negative given positive: $C(-|+)$ and vice versa $C(+|-)$] and the distribution of positive [$p(+)$] and negative [$p(-)$] samples. For equal misclassification costs, $C(+/-) = C(-/+)$, the PCF simplifies to the percentage of positive cases in the dataset. In cost space, PCF is plotted against the normalized expected cost (NEC), that is, the expected cost normalized by the cost of misclassifying every example,

$$NEC = \frac{(1 - TP)\,p(+)\,C(-|+) + FP\,p(-)\,C(+|-)}{p(+)\,C(-|+) + p(-)\,C(+|-)}. \quad (7)$$

In cost curves, the x axis represents the *a priori* probability that a signal is present (equal cost case) and hence varies between [0, 1]. The y axis is the expected cost, a linear combination of the false positive and false negative probabilities weighted by the cost of each type of error with the given *a priori* distribution. When misclassification costs are the same for FP and FN, the cost curve plots TP probability on the x axis and error rate on the y axis.

Each point in ROC space corresponds to a line in cost curve space. These lines show how the threshold (or other criteria used to select the operating point) varies in performance as the *a priori* probability of a signal being present varies between 0 and 1 [the probability cost (x) axis]. Consequently, the points of an ROC curve produce a set of lines in cost curve space, and the cost curve is defined as the lower envelope of these lines. The lower envelope has the lowest cost for any specific *a priori* probability of signal.

In the metrics discussed so far, it is assumed that data sets will have the same percentage of positive and negative cases. The cost curve method adapts to changes in the percentage of signal presence and absence, which often varies

https://doi.org/10.1121/10.0009270

between the data used to develop a detector and what is encountered in the field. For example, the development data may be relatively balanced, with a roughly equal number of signal present and signal absent cases. However, for species that are only seasonally present, multi-year field data would contain a greater number of signal absent cases. Cost curves account for these differences by enabling performance estimation over signal present/absent distributions that are best reflective of the circumstances under which the detector will be applied. It should be noted that the detector itself may have performance variability due to differing conditions (e.g., a noisier background in winter months) that are not accounted for by any of the metrics discussed.

As the *a priori* signal presence probability approaches 0 or 1, there is a point where the most efficient classifier is a trivial one that either labels everything as signal (probability→1) or as no signal (probability→0). The cost lines associated with these two trivial classifiers are plotted as dotted lines (Figs. 6 and 7), and when the costs of both types of error are equal, they intersect to form an isosceles triangle anchored at 0 cost on either end of the probability axis. Any place that the sides of this triangle pass under the cost curve, a trivial detector will perform better than the detector being measured, and this defines the range over which the detector is effective. The computation and conceptualization of cost curves is more complex than the other methods, but they provide additional insight on detector performance at a given probability of positive occurrence.

H. Comparison of performance metrics

To illustrate differences in the performance of these metrics, we present scores from signal and noise distributions, modeled as either normal or Rayleigh distributions, with varying degrees of overlap (Fig. 5) along with the ROC, DET, PR, and cost curves that result from these distributions (Figs. 6 and 7).

With greater separation between signal and noise, all the error curves show improved performance (yellow curves in Figs. 6 and 7). Normally distributed data result in curves that bend toward the ideal operating point for the ROC [upper left corner of Fig. 6(A)] and for the PR [upper right corner in Fig. 6(C)], and curves that are straight lines in the DET [Fig. 6(B)]. Cost curves for normally distributed data form symmetric convex arcs with greater error (expected cost) for equal numbers of positive and negative cases (probability = 0.5) than for other operating conditions.

Rayleigh distribution were selected because they reflect the statistics that occur when two Gaussian processes are combined. Rayleigh distributions are asymmetric with long-tails for high values. The performance metrics for Rayleigh distributed scores (Egan, 1975) are somewhat better than for normally distributed scores [compare Figs. 6(A) and 7(A), Figs. 6(B) and 7(B), and Figs. 6(C) and 7(C)]. For these data, it is somewhat easier to judge the performance from the DET plot given the log-scaled display, particularly for data points near the desired operating point [lower-left-hand corner of Figs. 6(B) and 7(B)]. For Rayleigh distributions, the cost curve reveals that the largest errors occur when the

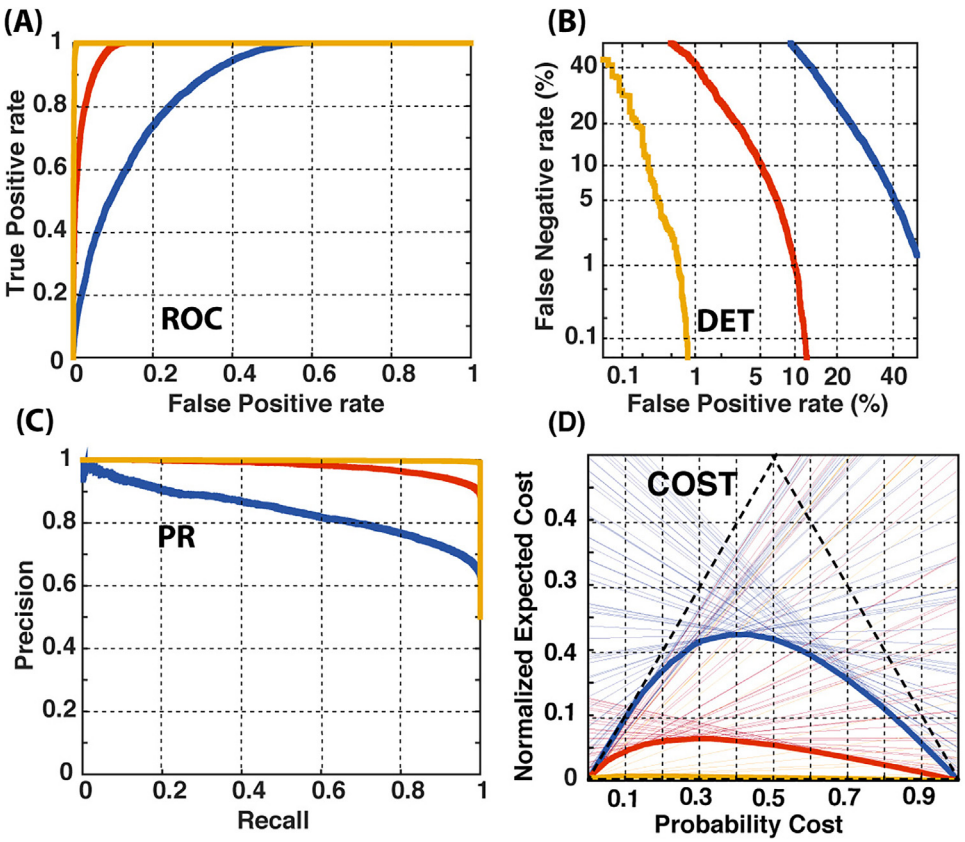

FIG. 7. (Color online) ROC (A), DET (B), PR (C), and cost curves (D) for Rayleigh distributed signal and noise from Figs. 5(D), 5(E). Mean difference between signal and noise is one (blue), two (red), and three (orange) standard deviations.

positive probability is in the range 0.3–0.4 (negative more likely than positive).

The above binary detection metrics are used for what is known as *monophonic sound detection*, estimating the presence or absence of a sequence of non-overlapping sound events. A more complex situation arises for *polyphonic sound event detection* (Mesaros *et al.*, 2016), where sounds from different sources occur in overlapping segments (Parascandolo *et al.*, 2016). This is the case when two species (e.g., blue and fin whale) are both present and producing calls that overlap in time. It may also be the case when different call types are produced by the same species and overlap in time (e.g., simultaneous blue whale B and D calls) and it is desired to detect both call types. Metrics are needed for polyphonic detection algorithms, and a recently described example is the *polyphonic sound detection score* (Bilen *et al.*, 2020). The key innovation of this metric is that a detection is taken as correct when the time overlap of a detection and of a corresponding ground truth label exceeds a threshold value [cf. Roch *et al.* (2011)]. Metrics for overlap metrics and multi-class situations, are then fused into a single score. In the current study, we have not considered polyphonic sound detection metrics, but they are a topic worthy of future study.

III. CASE STUDY RESULTS

Case studies for the application of these performance metrics were drawn from the Seventh International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals using Passive Acoustics (DCLDE, 2015). These acoustic data were collected off the southern and central coast of California at several locations, spanning all four seasons. Two different datasets were provided, one with low-frequency sampling (1 or 1.6 kHz), appropriate for study of mysticetes, and one with high-frequency sampling (200 or 320 kHz), appropriate for the study of odontocetes. The low-frequency dataset contains call-level markings for blue whale (*Balaenoptera musculus*) D calls (Thompson *et al.*, 1996) and fin whale (*B. physalus*) 40 Hz calls (Širović *et al.*, 2013). The high-frequency dataset consists of marked time periods for encounters with echolocation clicks of species commonly found along the U.S. West Coast, including Cuvier's (*Ziphius cavirostris*) and Baird's beaked (*Mesoplodon densirostris*) whales (Baumann-Pickering *et al.*, 2013), Risso's (*Grampus griseus*) and Pacific white-sided (*Lagenorhynchus obliquidens*) dolphins (Soldevilla *et al.*, 2008), sperm whales (*Physeter macrocephalus*), unidentified porpoises (*Phocoenidae spp.*) and unidentified odontocetes, as well as click-level markings for only Cuvier's beaked whales that supplemented the DCLDE (2015) dataset.

A. Blue whale D call dataset

The D calls of blue whales (Thompson *et al.*, 1996; Oleson *et al.*, 2007) are typically associated with multiple foraging animals and are less consistent in frequency and duration than the units of song (McDonald *et al.*, 2006) that are made as a sequence of calls by a single animal. D calls are typically 1–4 s in duration and have a down-swept frequency modulation, typically in the range of 90–25 Hz. The challenge for detection and classification of D calls is to account for the variability in these calls and similar confounding signals (Fig. 8). The evaluation dataset for these calls covers 908D calls produced over three week in January, late April/early May, and November at three different recording sites (CINMS-B, DCPP-A, DCPP-C) in 2011 and 2012. Using a 3-min window as the unit of analysis there were ∼104 time-bins in the dataset and 861 positive time windows, yielding an 8.5% *a priori* probability of positive occurrence.

The generalized power law (GPL) detection algorithm of Helble *et al.* (2012) is designed for signals that are narrowband, but may be variable within a broad range of monitored frequencies, well suited to the characteristics of blue whale D calls. Rather than a conventional energy detector (square of the Fourier amplitude) the GPL detector uses a higher power (6) to provide increased signal-to-noise ratio. The GPL detector also uses detection threshold parameters that are robust against highly varying ocean noise conditions. To evaluate the performance of the GPL algorithm for blue whale D calls, a parameter for D call slope was varied along with the detection threshold, producing a series of curves for each performance metric.

Performance curves were constructed for the GPL blue whale D call detector (Fig. 9). For methods that required reporting signal presence/absence over discrete time bins (DET, ROC, and COST curves), we segmented the recording effort into 3 min bins. If any portion of the analyst-reported (actual) calls occurred during a bin, a detection was expected for that bin. If the call crossed the bin boundary, detection was expected for both bins. The same process was repeated for the (predicted) GPL detections. This resulted in 3 min bins that could be marked as actual call present or absent and/or detection present or absent. From these, the count of TP, FP, TN, and FN counts were constructed and

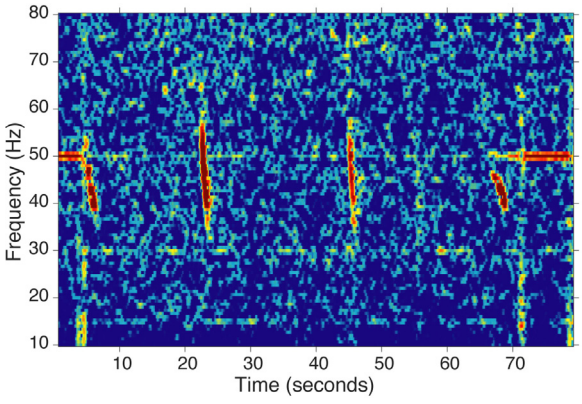

FIG. 8. (Color online) Four blue whale D calls from the DCLDE (2015) evaluation dataset, down swept from ∼55 to 38 Hz. Data were collected in the Santa Barbara Channel (34-17.126 N, 120-01.632 W, 580 m depth) on 23 June 2012 at 01:35 UTC.

https://doi.org/10.1121/10.0009270

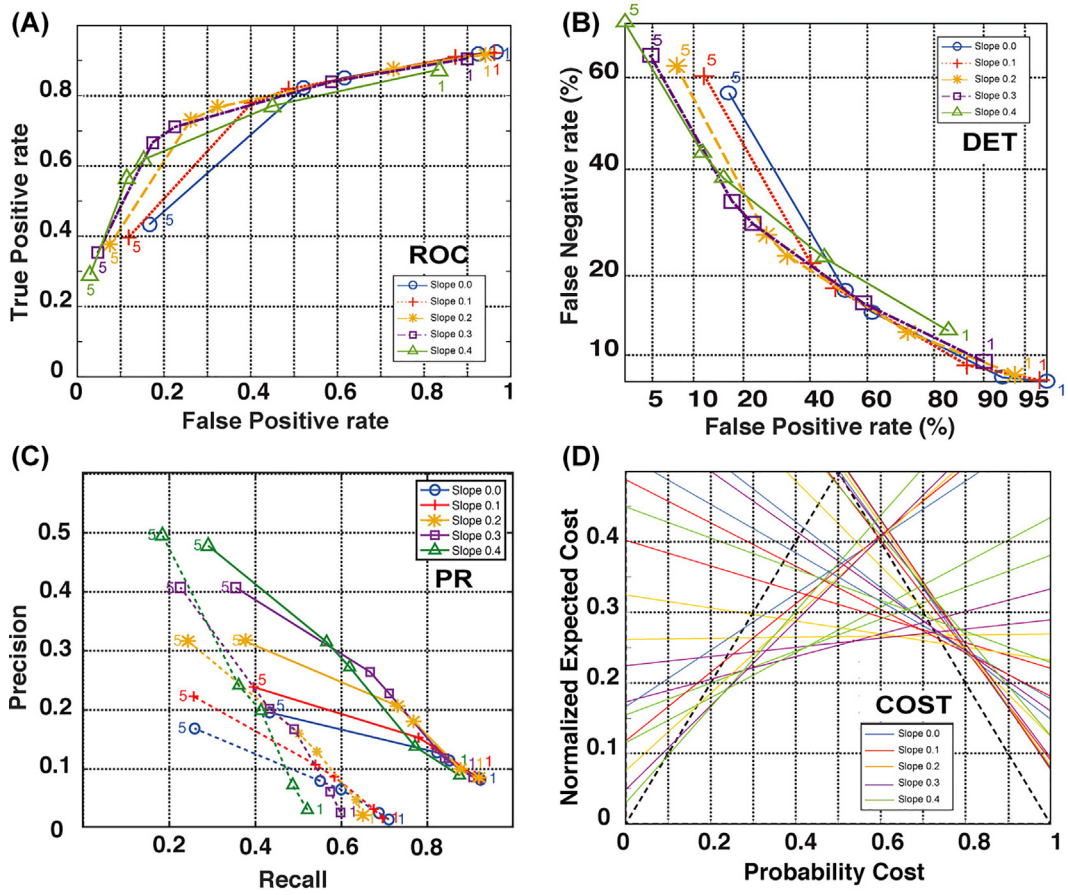

FIG. 9. (Color online) ROC (A), DET (B), PR (C), and cost curves (D) for the GPL algorithm for blue whale D call detections with varying parameters for call slope (line color) and detection threshold (1–5). PR curves are both for individual detections (dashed lines) and binned detection (solid lines).

the ROC, DET, and PR curves generated. PR curves were also computed by examining analyst and GPL detection time labels, where TP resulted from intersecting labels, FP were detection labels that did not overlap analyst labels, and FN were the analyst detection labels that did not overlap GPL detection labels. Cost curves were derived from the ROC curves as discussed above.

The ROC curve for the GPL algorithm [Fig. 9(A)] suggests moderate performance for D call detection. The point on the ROC curve closest to the upper-left hand corner of the plot (TP = 1, FP = 0) provided a 72% true positive performance with a 26% false positive rate. The detector had its best performance for call slope thresholds of ~0.3 Hz/s. Likewise, the DET curve [Fig. 9(B)] shows an equal error rate (FN = FP) at ~27% for call slope of 0.2 Hz/s and only slightly weaker performance for slope of 0.3 Hz/s. These performance rates are not greatly different than those of normally distributed synthetic data, with a 1-σ signal-to-noise ratio (Fig. 7). The PR curve [Fig. 9(C)] can be assessed both for individual call performance and for 3 min time window performance. Both have relatively low equal precision and recall operating points at 32% and 40%, respectively. The impact of the slope parameter is clearly shown in the spread of precision at a given threshold, with improved precision at higher slope (0.4 Hz/s) with little loss of recall. The cost curve [Fig. 9(D)] suggests that below 10% or above 75% *a priori* probability of signal presence (probability cost) the GPL algorithm performs more poorly than assignment of all calls to be false or true (respectively). Operating at equal probability cost (0.5), the expected cost of the algorithm (0.25) appears to be about half that of random guessing (0.5). It is worth noting that in Helble *et al.* (2012) the processing chain involves a human analyst to review the detections, which would have greatly reduced false positives had that been implemented here.

B. Cuvier's beaked whale echolocation click dataset

The DCLDE (2015) evaluation dataset was also used to study detection of Cuvier's beaked whale's echolocation clicks. The beaked whale clicks in the manually annotated dataset conform to previously reported characteristics for Cuvier's beaked echolocation clicks (Zimmer *et al.*, 2005; Hildebrand *et al.*, 2015); they are short (200 μs) frequency-upswept (35–45 kHz) signals, with a regular inter-click-interval (0.4–0.5 s; Fig. 10). The rapidly produced (<0.15 s inter-click-interval) buzz pulses that Cuvier's beaked whales are also known to produce while foraging (Zimmer *et al.*, 2005) were excluded from this analysis.

The DCLDE (2015) evaluation dataset for Cuvier's beaked whales contained three week of data recorded in January 2009, August 2010, and March 2013, from three

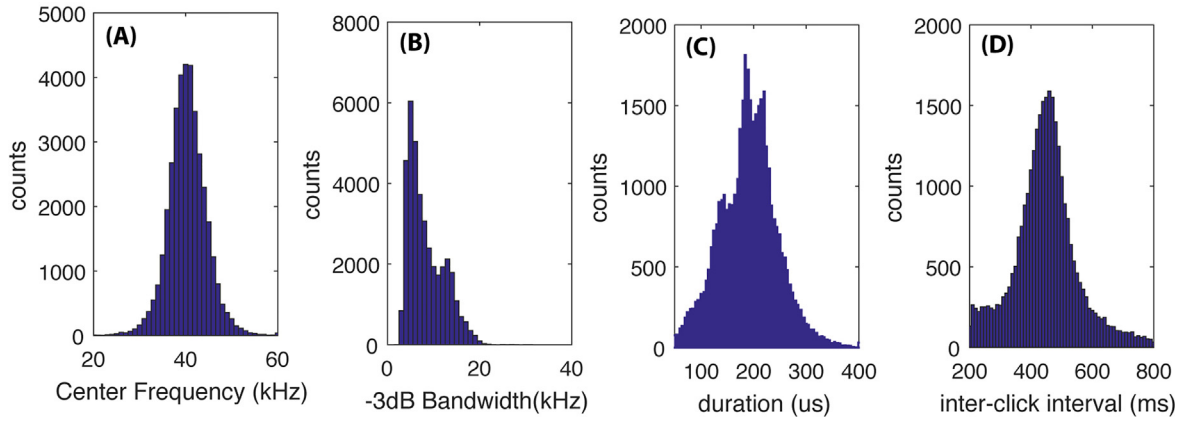

FIG. 10. (Color online) Cuvier's beaked whale click parameters for the manually annotated dataset. (A) Center-frequency, (B) −3 db bandwidth, (C) duration, and (D) inter-click-interval. Parameters for clicks with well-defined duration, N = 38 296 except inter-click-interval with N = 29 985. See Baumann-Pickering *et al.* (2013) for signal processing parameters.

different recording sites (SOCAL-E, SOCAL-R, DCPP-C, respectively) offshore from Southern California, at depths of 1000–1300 m (DCLDE, 2015). The time of individual Cuvier's beaked whale echolocation clicks was manually determined as follows. A filter was applied to all the data (fourth order elliptical bandpass filter with 0.1 dB bandpass ripple, 40 dB stop band attenuation, bandpass edges of 10 and 95 kHz). The filtered signal was then rectified, and whenever signal energy exceeding a threshold value (121 dB_{pp} re: 1 μPa) was observed, a 1 ms segment of data centered around each detection was saved. A 30 ms lockout period, where no further detections were allowed, followed each detection to prevent detection of multiple reflections or reverberant signals. This procedure resulted in 8×10^6 potential echolocation clicks. These detected data were further manually examined (Roch *et al.*, 2021) using custom software (Solsona-Berga *et al.*, 2020) that allows encounter-level visualization of data and elimination of incorrectly classified clicks to remove false (non-beaked whale) detections. The Cuvier's beaked whale echolocation detections remaining after this procedure (total of 43 034 clicks) were used as the manually annotated dataset. Using a one-hour window as the unit of analysis there were 504 time-bins in the dataset and 95 positive time windows, yielding a ∼19% *a priori* probability of positive occurrence.

We evaluated a Cuvier's beaked whale automated click detection method against the manual detection dataset described above. The automated method began with the full range of echolocation click detections output from the filtered raw acoustic data, as described above. A two stage unsupervised learning (clustering) algorithm (Frasier *et al.*, 2017) then identified recurrent signal types across all detections based on spectral features and modal inter-click intervals. In the first phase of the analysis, the algorithm reviewed echolocation clicks in five-minute time windows, computing mean spectra and ICI distributions for one or more detection types within each time window containing sufficient numbers of events (at least 50 detections). In the second phase, unsupervised clustering was used to identify a common set of consistent detection types across all sites based on the mean spectra and inter-click-interval distributions.

Seven signal categories were identified by the clustering algorithm, attributed by an analyst to Cuvier's beaked whale, Risso's dolphin, ships, echosounders, and three unidentified delphinid click types. A random subset of 442 000 detections were used to train a deep neural network to identify the seven signal types based on waveforms and spectra. The network, implemented in KERAS and TENSORFLOW, consisted of four dense hidden layers and a SOFTMAX output layer (Frasier, 2021). Once trained, the network was used to classify the 8×10^6 detections. For each event, the classifier returned a classification label (1–7) and a probability of belonging to the selected class (scale of 0–1). The majority (57%) of detections labeled as Cuvier's beaked whales were given a high probability (>95%) of belonging to that class. Probabilities as low as 30% were retained in the analysis, about twice the probability (14% = 1/7th) for random class selection. The data were analyzed into hourly time bins by segmenting the effort period by hour, and hourly bins with at least one click were set as predicted positive and those without clicks as predicted negative.

The beaked whale detector performance curves (Fig. 11) reveal an ROC curve with a ∼90% true positive rate and a 3%–5% false positive rate, with little variation due to the probability of class assignment. The DET curve shows similar results with a 9%–10% false negative, or detection miss, rate. The precise rates are somewhat easier to discern in the DET curve due to the non-linear scale. The PR curve is plotted for both hourly-bin performance (squares) and individual detection performance (circles) in Fig. 11. The two ways of assessing the PR performance show contrasting behaviors. The metric based on hourly bins shows declining precision at lower class probability thresholds, with little gain in recall, suggesting that lower probability detections add FP bins faster than additional correct detections (as might be expected). However, the individual detection PR curve (dotted line) shows that inclusion of lower probability detections

https://doi.org/10.1121/10.0009270

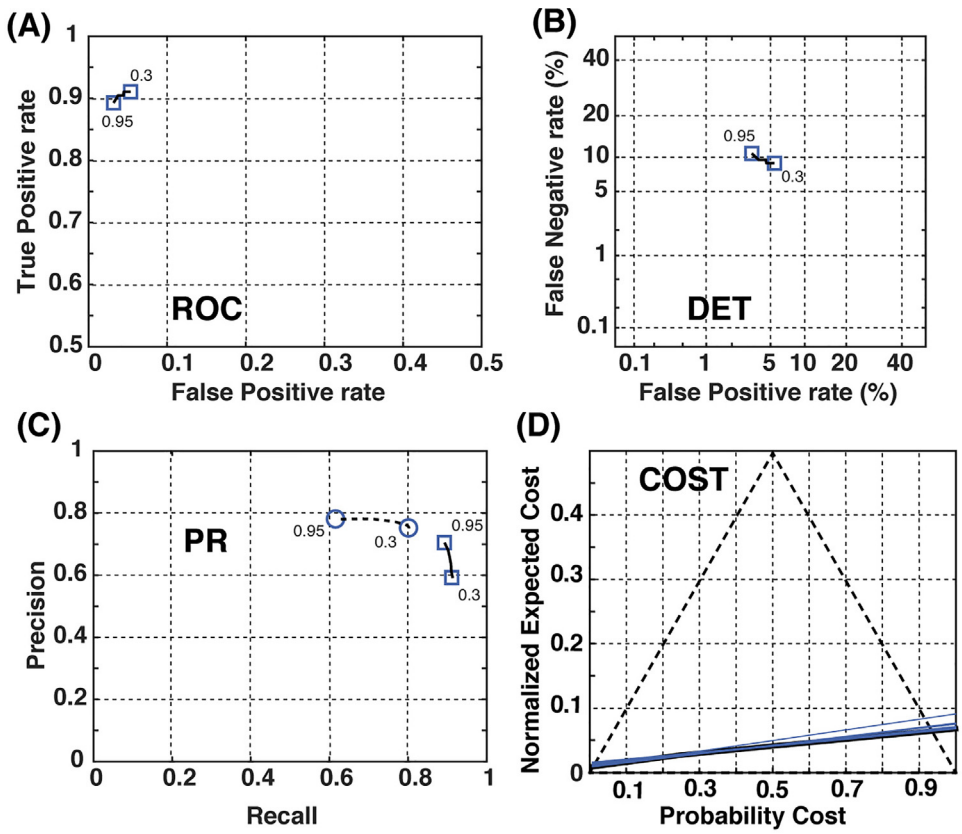

FIG. 11. (Color online) ROC (A), DET (B), PR (C), and cost curves (D) for the Cuvier's beaked whale detections. Data are plotted for all detections with >30% and >95% probability of belonging to the class. PR plot shows both for one-hour time bins (square) and individual detections (circle).

increases the recall performance with little impact on the precision. This is presumably due to many correctly detected beaked whale clicks being added which would have been counted as single detections in the binned metric. It can also be seen that the recall is much lower, as the binned detections have multiple opportunities to detect a single click. The cost curve reveals that the beaked whale detector performance for 50% probability cost, the case where true and false detections are equally likely, has an error of about 5% and that the detector is only inferior to guessing when the signal is known to be present in >93% of the observation windows.

IV. DISCUSSION

For the case of marine mammal calls, the presence of a call is often a relatively rare event, which can result in class-imbalance (large differences in the occupancy of cells within the confusion matrix), particularly for TN time bins. The selection of short periods (e.g., 3 min) for time bins will result in a large number of TN events, which will inherently reduce the false-positive-rate used in the ROC and DET metrics. Longer time bins may result in better detector performance but at the expense of temporal resolution. However, longer time periods can obscure problems with the detector by aggregating periods with good and poor performance. Although various approaches have been proposed to account for class imbalance it is still a significant problem for signal classification. For the ROC curve, when the data are highly imbalanced towards the signal absent case, changes in the TP counts are more easily reflected in the true positive rate than changes in the FP count are reflected in the false positive rate. The problem is somewhat addressed in the DET curve by using both FN and FP rates and a non-linear scaling of the axes, but for both ROC and DET curves the false positive rate may appear to suggest low numbers of FP detections, when in reality there is a distortion based on a large percentage of TN.

PR has a clear advantage in not requiring discrete time-binning of the data, and therefore may be less subject to the issue of class imbalance described above. For the GPL blue whale D call case study (Fig. 9), the PR curve which depends upon individual detections (dotted lines), rather than binned data (solid lines), suggests a weaker performance for the algorithm than what is observed for the ROC and DET curves. The GPL algorithm is designed for high recall but typically produces low precision, and to alleviate this it includes a pass by an analyst using a review tool to discard false positives, although this final step of manual editing was omitted for this test case. The PR curves based on a 3 min time window suggest that the GPL detector can achieve high recall (>90%) when the detector threshold is low, but with only 50%–70% recall for individual calls. The difference in PR based on individual call and time windows also reveals that the choice of detector slope has an impact, but only for high detector threshold with associated low recall rates, which is typically not the way that this algorithm is applied. The cost curve for the GPL algorithm provides additional insight, revealing that the algorithm is effective for a range of *a priori* detection probabilities (10%–75%). The selected 3 min time window provided an

8.5% detection probability, which is outside the range of effectiveness revealed by the cost curve, suggesting that a longer time window would be appropriate for this classifier without the manual review step. The GPL detector also illustrates the challenge of having a detector operate at a particular performance point while dealing with the multiple detector settings: detection threshold, call length, noise baseline, and power-law. All these settings are subject to adjustment—and knowing how to jointly adjust them is a general issue when operating with more than one detector parameter, one that we have not addressed here.

For the Cuvier's beaked whale case study (Fig. 11), even using a one-hour time window resulted in class imbalance due to the rarity of this species. The ROC and DET curves show exceptionally low false positive rates (3%–5%) but by making reference to the PR curve we learn that the precision is 60%–70% suggestive of high impact from the FP count. The PR based on individual calls gives confidence in the use of low score calls (>0.3) since they improve the recall significantly (80%) with little or no impact on precision (75%). The cost curve for the beaked whale study suggests low error across a broad range of operating conditions.

Recent studies of right whales calls (Kirsebom *et al.*, 2020; Shiu *et al.*, 2020) used both PR and a curve that showed FPR/h versus recall. This is hybrid precision-recall plot where the precision is translated into a count of false positives per unit time period. These plots do not provide additional information but do provide a human interpretable insight into the nuisance factor of false alarms. In many cases, a human analyst quality checks the detections, and high false positive rates make the use of an automated system cost prohibitive.

V. CONCLUSION

The use of multiple metrics can be helpful in understanding the performance of a detection algorithm. Visualization of two performance metrics, for instance, ROC or DET as well as PR, provides insight that cannot be obtained from a single metric alone. Likewise, calculation of PR both for individual call and for a time window allows assessment of how the choice of window length has impacted the assessment. Cost curves require some effort to understand, but they provide fresh insight into the range of operating conditions under which a detector is effective. Likewise, they provide the flexibility to specify the cost of different kinds of error, which is often an important consideration in an operational setting.

A drawback of all of the performance metrics discussed here is that they present curves representing the results of changing a single tunable parameter, typically a threshold. When the algorithm has a multidimensional parameter space, the interaction between these parameters becomes more difficult to represent and should be the topic of future work.

ACKNOWLEGMENTS

This work was funded by the U.S. Navy Living Marine Resources (LMR) Program and we thank Bob Gisiner, Anu Kumar, and Mandy Shoemaker for their support. The LMR Detection and Classification Committee—consisting of J.A.H, M.A.R, Ana Širović, Simone Baumann-Pickering, David Mellinger, Holger Klinck, Douglas Gillespie, Susan Jarvis, and T.A.H—spent many hours discussing the best approach for developing ground truth datasets and for implementing metrics for marine mammal call detection and classification, and we appreciate their contributions to this work. The comments of two anonymous reviewers greatly improved this paper.

Allen, J. A., Garland, E. C., Dunlop, R. A., and Noad, M. J. (**2018**). "Cultural revolutions reduce complexity in the songs of humpback whales," Proc. R. Soc. B: Biol. Sci. **285**(1891), 20182088.

Au, W. W. L., and Lammers, M. O. (**2016**). "Listening in the ocean: New discoveries and insights on marine life from autonomous passive acoustic recorders," in *Modern Acoustics and Signal Processing* (Springer, New York).

Auckenthaler, R., Carey, M., and Lloyd-Thomas, H. (**2000**). "Score normalization for text-independent speaker verification systems," Dig. Sign. Process. **10**(1), 42–54.

Baumann-Pickering, S., McDonald, M. A., Simonis, A. E., Berga, A. S., Merkens, K. P., Oleson, E. M., Roch, M. A., Wiggins, S. M., Rankin, S., Yack, T. M., and Hildebrand, J. (**2013**). "Species-specific beaked whale echolocation signals," J. Acoust. Soc. Am. **134**(3), 2293–2301.

Bilen, Ç., Ferroni, G., Tuveri, F., Azcarreta, J., and Krstulović, S. (**2020**). "A framework for the robust evaluation of sound event detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 61–65.

Bittle, M., and Duncan, A. (**2013**). "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring," in *Proceedings of Acoustics 2013*, Victor Harbor, Australia.

Campbell, J. J., and Reynolds, D. A. (**1999**). "Corpora for the evaluation of speaker recognition systems," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Phoenix, AZ.

Chernoff, H., and Moses, L. (**1959**). *Elementary Descision Theory* (Wiley, New York).

Cui, X. D., Goel, V., and Kingsbury, B. (**2015**). "Data augmentation for deep neural network acoustic modeling," IEEE-ACM Trans. Audio Speech Lang. Process. **23**(9), 1469–1477.

Davis, J., and Goadrich, M. (**2006**). "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ACM, Pittsburgh, PA, pp. 233–240.

DCLDE (**2015**). "Dataset documentation for the 2015 DCLDE workshop," http://www.cetus.ucsd.edu/dclde/datasetDocumentation.html (Last viewed 1/13/2022).

DCLDE (**2018**). "Dataset documentation for the 2018 DCLDE workshop," http://sabiod.univ-tln.fr/DCLDE/challenge.html#datasetDocumentation (Last viewed 1/13/2022).

Drummond, C., and Holte, R. C. (**2006**). "Cost curves: An improved method for visualizing classifier performance," Mach. Learn. **65**(1), 95–130.

Efron, B. (**1982**). "Transformation theory—How normal is a family of distributions," Ann. Stat. **10**(2), 323–339.

Egan, J. P. (**1975**). *Signal Detection Theory and ROC-Analysis* (Academic Press, New York).

Fawcett, T. (**2006**). "An introduction to ROC analysis," Pattern Recogn. Lett. **27**(8), 861–874.

Frasier, K. E. (**2021**). "A machine learning pipeline for classification of echolocation clicks in large underwater acoustic datasets," PLOS Comput. Biol. **17**, e1009613.

Frasier, K. E., Roch, M. A., Soldevilla, M. S., Wiggins, S. M., Garrison, L. P., and Hildebrand, J. A. (**2017**). "Automated classification of dolphin echolocation click types from the Gulf of Mexico," PLoS Comp. Biol. **13**(12), e1005823.

Gibb, R., Browning, E., Glover-Kapfer, P., and Jones, K. E. (**2019**). "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring," Meth. Ecol. Evol. **10**(2), 169–185.
Gillespie, D., Caillat, M., Gordon, J., and White, P. (**2013**). "Automatic detection and classification of odontocete whistles," J. Acoust. Soc. Am. **134**(3), 2427–2437.
Helble, T. A., Ierley, G. R., D'Spain, G. L., Roch, M. A., and Hildebrand, J. A. (**2012**). "A generalized power-law detection algorithm for humpback whale vocalizations," J. Acoust. Soc. Am. **131**(4), 2682–2699.
Helstrom, C. W. (**1968**). *Statistical Theory of Signal Detection*, 2nd ed. (Pergamon Press, Oxford, England).
Hildebrand, J. A., Baumann-Pickering, S., Frasier, K. E., Trickey, J. S., Merkens, K. P., Wiggins, S. M., McDonald, M. A., Garrison, L. P., Harris, D., Marques, T. A., and Thomas, L. (**2015**). "Passive acoustic monitoring of beaked whale densities in the Gulf of Mexico," Sci. Rep. **5**, 16343.
Kirsebom, O. S., Frazao, F., Simard, Y., Roy, N., Matwin, S., and Giard, S. (**2020**). "Performance of a deep neural network at detecting North Atlantic right whale upcalls," J. Acoust. Soc. Am. **147**(4), 2636–2646.
Kohavi, R. (**1995**). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Quebec, Canada, p. 7.
Lehmann, E. L. (**1959**). *Testing Statistical Hypotheses* (Wiley, New York), pp. 369.
Lindgren, B. W. (**1971**). *Elements of Decision Theory* (Macmillan, New York).
Livina, V. N., Brouwer, A., Harris, P., Wang, L., Sotirakopoulos, K., and Robinson, S. (**2018**). "Tipping point analysis of ocean acoustic noise," Nonlin. Proc. Geophys. **25**(1), 89–97.
Manning, C., and Schütze, H. (**1999**). *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).
Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (**1997**). "The DET curve in assessment of detection task performance," in *Proceedings of Eurospeech '97*, Rhodes, Greece (September, 1997), pp. 1895–1898.
McDonald, M. A., Hildebrand, J. A., and Mesnick, S. (**2009**). "Worldwide decline in tonal frequencies of blue whale songs," Endang. Spec. Res. **9**(1), 13–21.
McDonald, M. A., Messnick, S. L., and Hildebrand, J. A. (**2006**). "Biogeographic characterisation of blue whale song worldwide: Using song to identify populations," J. Cetacean Res. Manage. **8**(1), 55–65. https://escholarship.org/uc/item/5r16c2mz.
Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**(6), 3518–3529.
Mellinger, D. K., and Clark, C. W. (**2006**). "MobySound: A reference archive for studying automatic recognition of marine mammal sounds," Appl. Acoust. **67**(11-12), 1226–1242.
Mesaros, A., Heittola, T., and Virtanen, T. (**2016**). "Metrics for polyphonic sound event detection," Appl. Sci. **6**(6), 162.
Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P. R., Gérard, O., Adam, O., and Cazau, D. (**2021**). "Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics," Ecolog. Inf. **61**, 101185.
Oleson, E. M., Calambokidis, J., Burgess, W. C., McDonald, M. A., LeDuc, C. A., and Hildebrand, J. A. (**2007**). "Behavioral context of call production by eastern North Pacific blue whales," Mar. Ecology Prog. Ser. **330**, 269–284.
Oswald, J. N., Barlow, J., and Norris, T. F. (**2003**). "Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean," Mar. Mamm. Sci. **19**, 20–37.
Parascandolo, G., Huttunen, H., and Virtanen, T. (**2016**). "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6440–6444.
Peterson, W., Birdsall, T., and Fox, W. (**1954**). "The theory of signal detectability," Trans. IRE Profess. Group Inf. Theory **4**(4), 171–212.
Roch, M. A., Brandes, T. S., Patel, B., Barkley, Y., Baumann-Pickering, S., and Soldevilla, M. S. (**2011**). "Automated extraction of odontocete whistle contours," J. Acoust. Soc. Am. **130**, 2212–2223.
Roch, M. A., Lindeneau, S., Aurora, G. S., Frasier, K. E., Hildebrand, J. A., Glotin, H., and Baumann-Pickering, S. (**2021**). "Using context to train time-domain echolocation click detectors," J. Acoust. Soc. Am. **149**(5), 3301–3310.
Roch, M. A., Stinner-Sloan, J., Baumann-Pickering, S., and Wiggins, S. M. (**2015**). "Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks," J. Acoust. Soc. Am. **137**(1), 22–29.
Russell, S. J., and Norvig, P. (**2020**). *Artificial Intelligence: A Modern Approach*, 4th ed. (Prentice Hall, Upper Saddle River, NJ).
Shawe-Taylor, J., and Zlicar, B. (**2015**). "Novelty Detection with One-Class Support Vector Machines," in *Advances in Statistical Models for Data Analysis*, edited by I. Morlini, T. Minerva, and M. Vichi, pp. 231–257.
Shiu, Y., Palmer, K. J., Roch, M. A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (**2020**). "Deep neural networks for automated detection of marine mammal species," Sci. Rep. **10**(1), 607.
Širović, A., Johnson, S. C., Roche, L. K., Varga, L. M., Wiggins, S. M., and Hildebrand, J. A. (**2014**). "North Pacific right whales (*Eubalaena japonica*) recorded in the northeastern Pacific Ocean in 2013," Mar. Mammal Sci. **31**, 800–817.
Širović, A., Williams, L. N., Kerosky, S. M., Wiggins, S. M., and Hildebrand, J. A. (**2013**). "Temporal separation of two fin whale call types across the eastern North Pacific," Mar. Biol. **160**(1), 47–57.
Soldevilla, M. S., Henderson, E. E., Campbell, G. S., Wiggins, S. M., Hildebrand, J. A., and Roch, M. A. (**2008**). "Classification of Risso's and Pacific white-sided dolphins using spectral properties of echolocation clicks," J. Acoust. Soc. Am. **124**(1), 609–624.
Solsona-Berga, A., Frasier, K. E., Baumann-Pickering, S., Wiggins, S. M., and Hildebrand, J. A. (**2020**). "DetEdit: A graphical user interface for annotating and editing events detected in long-term acoustic monitoring data," PLoS Comp. Biol. **16**(1), e1007598.
Tachibana, R. O., Oosugi, N., and Okanoya, K. (**2014**). "Semi-Automatic Classification of Birdsong Elements Using a Linear Support Vector Machine," Plos One **9**, e92584.
Thompson, P. O., Findley, L. T., Vidal, O., and Cummings, W. C. (**1996**). "Underwater sounds of blue whales, *Balaenoptera musculus*, in the Gulf of California, Mexico," Mar. Mammal Sci. **12**(2), 288–293.
VanTrees, H. (**1968**). *Detection, Estimation, and Modulation Theory* (Wiley, New York), Vol. 1.
Wiggins, S. M., and Hildebrand, J. A. (**2007**). "High-frequency acoustic recording package (HARP) for broad-band, long-term marine mammal monitoring," in *International Symposium on Underwater Technology 2007 and International Workshop on Scientific Use of Submarine Cables & Related Technologies 2007*, Institute of Electrical and Electronics Engineers, Tokyo, Japan (April 17–20, 2007), pp. 551–557.
Zimmer, W., Johnson, M., Madsen, P., and Tyack, P. (**2005**). "Echolocation clicks of free-ranging Cuvier's beaked whales (*Ziphius cavirostris*)," J. Acoust. Soc. Am. **117**(6), 3919–3927.