JASA ARTICLE

Using context to train time-domain echolocation click detectors[a)]

Marie A. Roch,[1,b)] Scott Lindeneau,[1] Gurisht Singh Aurora,[1] Kaitlin E. Frasier,[2] John A. Hildebrand,[2] Hervé Glotin,[3] and Simone Baumann-Pickering[2]

[1]*Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-7720, USA*

[2]*Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive #0205, La Jolla, California 92093, USA*

[3]*Université de Toulon, BP 20132, 83957 La Garde Cedex, France*

ABSTRACT:

This work demonstrates the effectiveness of using humans in the loop processes for constructing large training sets for machine learning tasks. A corpus of over 57 000 toothed whale echolocation clicks was developed by using a permissive energy-based echolocation detector followed by a machine-assisted quality control process that exploits contextual cues. Subsets of these data were used to train feed forward neural networks that detected over 850 000 echolocation clicks that were validated using the same quality control process. It is shown that this network architecture performs well in a variety of contexts and is evaluated against a withheld data set that was collected nearly five years apart from the development data at a location over 600 km distant. The system was capable of finding echolocation bouts that were missed by human analysts, and the patterns of error in the classifier consist primarily of anthropogenic sources that were not included as counter-training examples. In the absence of such events, typical false positive rates are under ten events per hour even at low thresholds.

 https://doi.org/10.1121/10.0004992

(Received 2 February 2021; revised 23 April 2021; accepted 26 April 2021; published online 18 May 2021)

[Editor: James F. Lynch] Pages: 3301–3310

I. INTRODUCTION

Toothed whales (suborder Odontoceti) use echolocation for a variety of purposes, including navigation and foraging, on a regular basis, making these signals common candidates for passive acoustic monitoring. Detecting and characterizing echolocation clicks can be challenging, as toothed whales have evolved to produce highly directional signals that are focused into a narrow beam in front of the animal (Au, 1993; Cranford, 2000). This results in signals that have varying temporal and spectral characteristics, depending upon the angle between the toothed whale's longitudinal axis and the receiver (Au *et al.*, 2012a,b).

The directional variability of echolocation clicks makes their detection more difficult. Clean on-axis echolocation signals are easily recognized with highly characteristic shapes. Off-axis clicks may contain species identity signals (Soldevilla, 2008) for some species, but as the angle increases, these clicks become more difficult to detect. At closer ranges, off-axis clicks that have less energy frequently dominate the number of echolocation clicks received. As a consequence, the performance of most automated click detectors has been described either anecdotally (e.g., Roch *et al.*, 2011) by calibration to an existing click detector, or by evaluation on small data sets (e.g., Kandia and Stylianou, 2006). In this work, we present results of an approximately 30 times faster-than-real-time neural network-based click detector of over 850 000 analyst-verified echolocation clicks without regard to the species that produced them from different times and locations. Both development and evaluation data have been labeled through a semiautomated process that enables analysts to exploit context to efficiently quality control potential echolocation clicks.

Click detection has traditionally been treated as a signal processing problem, such as Gillespie's Rainbow Click (Gillespie, 1997), that triggered on the amplitude of a signal that had been detrended by a low-passed version of itself. Most echolocation detectors are energy threshold detectors (e.g., Houser *et al.*, 1999). These detectors typically compute peak to peak or root mean square received levels (RL_{pp} and RL_{RMS}, respectively) and trigger detections when these values exceed an energy threshold. Some variants trigger on signal-to-noise ratio (SNR) instead of the absolute energy measurement or compare energy in different bands (e.g., Klinck and Mellinger, 2011). It is common to bandpass or high-pass filter the time-domain signal prior to computing the energy (e.g., Gillespie and Caillat, 2008), as the frequency band below the echolocation signal tends to be much noisier. Decisions are commonly made based on a threshold (e.g., Mellinger, 2001) or less frequently a hypothesis test (e.g., Zimmer *et al.*, 2008). Teager's short-time energy operator (Kaiser, 1990) is an alternative energy

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

[b)]Electronic mail: Marie.Roch@sdsu.edu

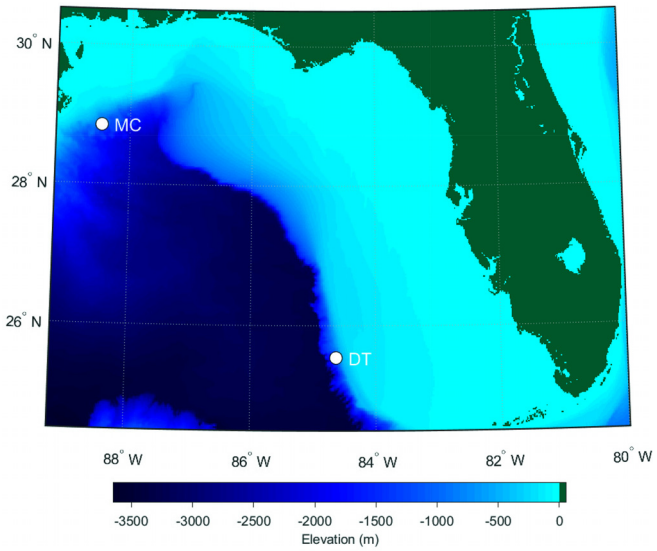

FIG. 1. (Color online) Dry Tortugas (DT) and Mississippi Canyon (MC) deployment sites in the Gulf of Mexico. Bathymetry courtesy of the United States National Center for Environmental Information (Amante and Eakins, 2009).

measurement that estimates energy over very short windows of a few samples. It was first proposed for echolocation detection by Kandia and Stylianou (2006) and can be smoothed when working on high-frequency data for effective click detection (Soldevilla *et al.*, 2008). Characteristics of the energy detections are frequently examined to determine if the detected signal should be considered a click, frequently looking at features such as duration, peak frequency, bandwidth, envelope shape, etc. (e.g., Soldevilla, 2008; Frasier, 2015; Madhusudhana *et al.*, 2015). Additional methods examine spectral characteristics (Bermant *et al.*, 2019), spectral and temporal characteristics (Zimmer *et al.*, 2005a), and phase changes in the group delay of the time-domain signal (Kandia and Stylianou, 2008). Machine learning approaches are rarely used, although there are some examples, such as Bermant *et al.* (2019), who used convolutional neural networks on spectrogram representations of audio, and Luo *et al.* (2019), who used a deep one-dimensional (1D) convolutional net on a modest data set from the publicly available Mobysound archive. Machine learning approaches are used for determining the species that produced an echolocation click after the click has been detected but usually rely on an energy-based method for the detection step [e.g., Ferrari *et al.* (2020), which uses time series of Teager energy detected clicks for classification to species].

For a machine learning approach to detection, a large corpus of training data is needed. Frasier *et al.* (2017) have developed methods to accomplish this; a very permissive peak to peak energy-based detector was created to detect candidate echolocation clicks for the purpose of unsupervised learning of click archetypes. All candidates with $\geq$115 dB_{pp} re 1 μPa in bandpassed data (10–90 kHz) were admitted by this detector and edited with a tool designed to provide contextual assistance for quality controlling echolocation detections. DETEDIT (Solsona-Berga *et al.*, 2020) provides interactive displays that show detections in relation to one another, providing information about the inter-detection interval, which tends to cluster about the inter-click interval, long-term spectrograms to place detections in a broader context, averaged waveforms and spectra of groups of detections as well as the ability to compare and contrast with easily selectable subgroups, and comparisons of different energy and click duration measures. DETEDIT assisted editing can quality control large numbers of detections much more efficiently than by simply examining time series and spectrograms. This opens the door to developing high-quality training sets that can be exploited by machine learning.

II. METHODS

A. Data

Data were collected at two sites in the Gulf of Mexico (Fig. 1) using calibrated high-frequency acoustic recording package data loggers (Wiggins and Hildebrand, 2007). These consisted of ITC 1042 hydrophones (International Transducer Corp., Santa Barbara, CA) and custom preamplifier boards, sampled continuously at 200 kHz with 16 bit sampling. We used subsets of these data (Table I) from two deep-water locations, chosen to have a variety of quiet times, ships, and echolocation activity. One deployment was near the Dry Tortugas (25.539° N, 84.631° W) in fall and winter 2014–2015. The second deployment was recorded

TABLE I. Data from two instruments deployed in the Gulf of Mexico. All times are reported in universal coordinated time (UTC). Dry Tortugas data are used for cross-validation experiments in the development of the classifier and Mississippi Canyon data for evaluation. Number of training clicks indicates the quantity of verified echolocation clicks contributing to cross-validation training data. As there were many other clicks in the data, "validated clicks" details the number of correctly detected echolocation clicks that had a peak to peak received level $\geq$115 dB re 1 μPa. NA, not available. Further details on these clicks can be found in Secs. II B and II D.

Site	Depth (m)	Data interval (UTC)	Training clicks	Validated clicks
Dry Tortugas	1189	1. 2014/10/07 00:00–2014/10/12 23:59	28 688	230 024
		2. 2015/02/17 00:00–2015/02/17 23:59	1682	27 516
		3. 2015/02/19 00:00–2015/02/20 15:37	14 153	147 063
		4. 2015/02/20 16:32–2015/02/23 23:59	12 976	107 083
Mississippi Canyon	980	1. 2010/12/21 03:00–2010/12/24 03:00	NA	449 184
Total			57 499	853 787

https://doi.org/10.1121/10.0004992

near Mississippi Canyon (28.846° N, 88.465° W) in early winter 2010.

A wide variety of pelagic toothed whales are present in the Gulf of Mexico (Frasier, 2015; Hildebrand *et al.*, 2015; Frasier *et al.*, 2017; Hildebrand *et al.*, 2019). These include the delphinids: pantropical spotted (*Stenella attenuata*), spinner (*Stenella longirostris*), Risso's (*Grampus griseus*), striped (*Stenella coeruleoalba*), rough-toothed (*Steno bredanensis*), Clymene (*Stenella clymene*), Fraser's (*Lagenodelphis hosei*) dolphins and short-finned pilot (*Globicephala macrorhynchus*), melon-headed (*Peponocephala electra*), false killer (*Pseudorca crassidens*), and killer whales (*Orcinus orca*); beaked whales: Blainville's, Gervais', and Cuvier's (*Mesoplodon densirostris, Mesoplodon europaeus*, and *Ziphius cavirostris*); and sperm whales (*Physeter macrocephalus*) and pygmy (*Kogia breviceps*) and dwarf (*Kogia sima*) sperm whales. Classification to species of the echolocation clicks is beyond the scope of this paper, but clicks in these data are most commonly produced by delphinids and beaked whales.

Data were stratified (Table I) into a development dataset, used in cross-validation tests while designing the system, and an evaluation dataset, only used once development was complete and the model parameters and signal processing chain were no longer modified. Approximately 12 days of acoustic data were used for development and 3 days for evaluation.

B. Signal processing

High-quality echolocation click candidates were identified using a permissive energy detector (Frasier *et al.*, 2017) on bandpass filtered (10–90 kHz) data that triggered when the peak to peak received level exceeded 115 dB re 1 μPa and the peak frequency was between 15 and 85 kHz. The threshold was set low to identify a large number of clicks with a potentially high false positive rate. These were quality controlled using the aforementioned detection editing software (DETEDIT; Solsona-Berga *et al.*, 2020) to reject false positives with a focus on retaining high-quality echolocation clicks. DETEDIT provides the ability to amplify analysis effort by large margins, making the analysis of sizable numbers of clicks feasible. The echolocation clicks identified by this toolchain provided a training corpus for the detectors in this work. As the focus was on identifying examples of unambiguous clicks, these data were only used in training and not as a ground truth corpus other than to verify that these clicks were a subset of the detected clicks in the test folds.

For the detection system, we similarly bandpass filtered data between 10 and 90 kHz using a finite impulse response equiripple filter designed to provide 1 dB of ripple in the passband and 80 dB of attenuation in the stop bands. Transition bands of 2 kHz occurred on either side of the passband. In our experiments, these data were prefiltered, and reported experiment time does not account for filtering, which is typically quite fast. Echolocation clicks for most odontocetes are shorter than a couple hundred μs [Table VII.2 of Au (1993)] with off-axis clicks having varying and longer duration. The family of beaked whales have longer echolocation clicks. Most but not all beaked whale clicks are shorter than 500 μs (Baumann-Pickering *et al.*, 2013). Data were partitioned into 500 μs trial bins to select frames that would capture most echolocation clicks and typically provide contextual ambient noise.

In early experiments, we found that providing either the peak to peak received level or the SNR of a bin was useful (see Fig. 8 in Sec. IV), and in the experiments reported here, we estimated the SNR of each trial bin by dividing it into thirds. Under the hypothesis that most clicks are significantly shorter than 500 μs, we split each 500 μs trial bin into three 166.67 μs segments. Based on the duration of most clicks and their random position within the 500 μs, it is expected that clicks will not cover the full 500 μs of the analysis bin, and it is almost certain that the high energy portion of the click will not. When a click is present, it is expected that at least one segment will be dominated by signal energy and at least one by ambient noise, which in some cases may be mixed in with the low intensity tail of a very long click. When clicks are absent, we expect all three segments to have less variability in intensity. We elected not to use more than three segments, as this would increase the chances that the intensity measurement was weakened due to the main portion of the click being split across segments.

We estimated the SNR from the ratio (log difference) of the strongest and weakest peak to peak received level (RL_{pp}) amongst the three segments. We considered using order statistics as an alternative method of estimating the ambient noise level, but the sort time required for 7.2×10^6 analysis windows per hour would have been prohibitive. Although our data were calibrated, we elected to use the SNR as a feature as opposed to RL_{pp}, as SNR can be used when calibration information is unavailable.

C. Modeling

Training data were selected by framing the training dataset into 500 μs non-overlapping segments starting at the first sample of each contiguous block of data. The start of a click provided in the ground truth data was used to select frames that were labeled as clicks. This resulted in echolocation clicks that could fall anywhere within a 500 μs segment and at times resulted in truncated clicks. Examples of training data without clicks were selected by using the next segment after a click except when (1) the start of the click was more than 30% into the frame or (2) the next frame was marked as containing a click. In both cases, a click absent example was selected from the next non-click frame.

Detection was accomplished via a small neural network that is publicly available on bitbucket (Roch, 2019). Custom code was developed in PYTHON 3 using the Keras 2.2.4 and Tensorflow 1.13.1 neural network libraries (Abadi *et al.*, 2015; Chollet *et al.*, 2015). Network input consisted of bandpassed samples associated with nonoverlapped 500 μs trial frames (100 samples at 200 kHz) and the appended

SNR. *Ad hoc* experiments showed that normalization of the input resulted in overtraining with poor generalization. In these *ad hoc* experiments, the time series was normalized by the maximum absolute value of each trial frame and the SNR by a high source level to restrict the SNR range to approximately [0, 1]. Batch normalization would have been inappropriate due to the rarity of the click present case during classification. We concluded that intensity of the time series was an important feature and presented unaltered waveforms and SNR to the network.

These data were passed through two densely connected layers of 101 units with rectified linear unit activation functions (Nair and Hinton, 2010). Each dense layer was followed by a dropout layer (Srivastava *et al.*, 2014), which provided regularization by dropping dense layer nodes with a 20% probability during training. This was followed by a two-node output layer that used a softmax activation function to provide a smooth estimate of the arg max function.

Some versions of the network not reported here used a 1D convolutional layer [in the same vein as the work of Luo *et al.* (2019)], but our decision to append the SNR estimate to the time series input prevented our including that in the final network design without resorting to a parallel network path for the SNR measurement. Parameter estimation used the Adam optimizer (Kingma and Ba, 2015) with Keras defaults and a categorical cross-entropy loss metric. Optimization was stopped at 75 epochs, and there was no attempt to use early stopping or further explore the parameter space, as the loss function demonstrated reasonable behavior and cross-validation performance was acceptable.

For each candidate detection frame where the network estimated the probability that a click was present to be ≥ 0.5, we identified the timestamp of the maximum magnitude sample within the click. For additional processing described in Sec. II D, we retained a time series consisting of 200 μs before and after the peak amplitude, even if this resulted in samples being retained from a prior or subsequent trial bin.

D. Experiment design

All experiments in which parameters were selected were performed on the Dry Tortugas development data. We used a threefold cross-validation, ensuring that all examples within a contiguous set of data (Table I) were placed either in training or test datasets. This resulted in uneven experiment folds (Table II) but allowed us to examine whether there were large differences in performance between data that were temporally close to one another vs months apart. As there were four time-contiguous regions in the development data, some folds occasionally had three training regions, while others had two. Fold 0 primarily tested data from October 2014 with 1 day of February 2015 data and was trained entirely on 2015 data. Folds 1 and 2 contained both October and February training data and were tested on February data, resulting in better matched conditions. We do not report variations in experimental parameters, as we found that many networks were capable of learning to predict echolocation clicks well and that small variations in parameters did not tend to result in large variances in performance. Also, the machine-assisted quality control process described below is time intensive, and it would not be feasible to repeat it for a large number of network architectures.

TABLE II. Dry Tortugas data used in threefold cross-validation experiment. Each fold uses the data listed for testing, with the remaining folds being used for training data.

Fold	Data interval (UTC)	Training clicks from other folds
0	1. 2014/10/07 00:00–2014/10/12 23:59 2. 2015/02/17 00:00–2015/02/17 23:59	27 129
1	3. 2015/02/19 00:00–2015/02/20 15:37	41 346
2	4. 2015/02/20 16:32–2015/02/23 23:59	44 523

As the training data were designed to provide good quality exemplars, their timing information was an incomplete ground truth label set for the detection task. Manual analysis on a large scale was infeasible, and we adopted a strategy of manually verifying that the detector did not miss acoustic encounters and using the DETEDIT toolchain as described below to validate detections.

We used calibration data managed by the Tethys metadata system that allowed us to automatically retrieve calibration information for recordings (Roch *et al.*, 2016) and limited detections to those that had prediction probabilities ≥ 0.5 and received levels ≥ 112 dB re 1 μPa, 3 dB less than the original threshold. During the DETEDIT analysis, we were conservative in what we accepted as echolocation clicks, using the displays in DETEDIT to verify temporal and spectral properties as well as frequently returning to the time series data to verify fine-scale temporal context. We do not believe that our labels are error free, but that they are a reasonable approximation of the truth with respect to the clicks that were reported by our detector. They do not represent all echolocation clicks, as there are many more apparent clicks that were not detected by the energy- or neural network-based detectors. While we audited all detections, we tried to be especially careful with small groups of echolocation clicks detected at times outside of regions suggested by our high-quality training clicks. In general, we only accepted detections if we could see characteristic temporal or spectral patterns that matched well-described species, looking for regular inter-click intervals, appropriate spectral shape, and time series that roughly matched descriptions in the literature (e.g., Zimmer *et al.*, 2005a; Zimmer *et al.*, 2005b; Johnson *et al.*, 2006; Soldevilla *et al.*, 2008; Baumann-Pickering *et al.*, 2013; Fais *et al.*, 2015; Frasier *et al.*, 2017). Isolated clicks were rejected unless we could see clear toothed whale activity within several minutes of the isolated detection.

To measure performance as a function of precision and recall, we set a received level threshold of 115 dB re 1 μPa (see Sec. IV for rationale). Precision is the rate at which detections are correct, and recall is the rate at which expected ground truth clicks are retrieved. Precision and recall are a better metric for highly skewed classes (Davis

https://doi.org/10.1121/10.0004992

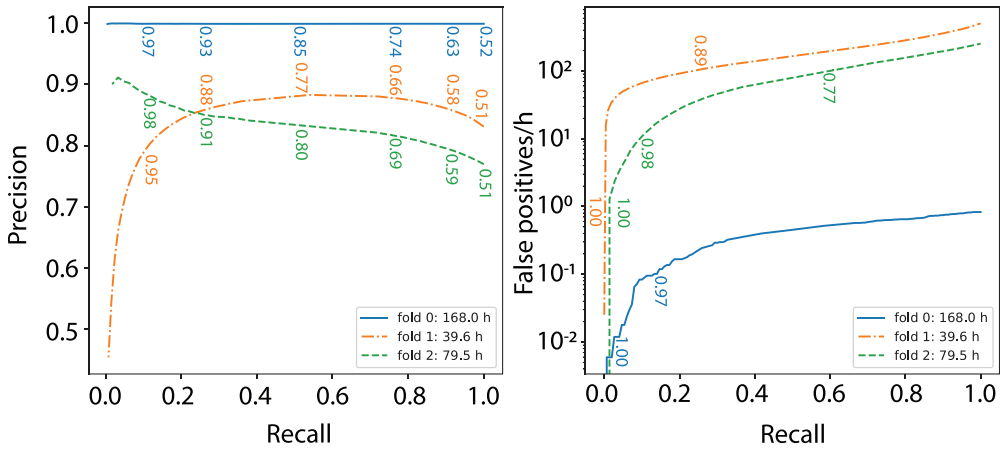

FIG. 2. (Color online) Precision and false positive per hour (log scale) over recall curves on a threefold experiment on the Dry Tortugas development data. Only detections with peak to peak received levels ≥115 dB re 1 μPa were evaluated. Numbers show operating thresholds at various recall points.

and Goadrich, 2006) than metrics such as receiver operating characteristic (Fawcett, 2006) or detection error trade-off curves (Martin *et al.*, 1997). Detections were considered to be valid if the timestamp associated with the peak amplitude was within 500 μs of the ground truth timestamp.

In addition, we computed the number of false positives per hour as a function of recall. For any given threshold resulting in a specific recall, we divided the number of false positives at that threshold by the number of trial bins of test data. As there are 7.2×10^6 500 μs trials per hour, even a high number of false positives can lead to very low false positive rates, e.g., 10 000 false positives per hour leads to a false positive rate of 0.0014. This metric provides better insight into the "nuisance" factor of false alarms; see Shiu *et al.* (2020) for a discussion.

III. RESULTS

In general, the detector was able to detect most of the labeled echolocation clicks at or above 115 dB re 1 μPa even at moderately conservative thresholds (Fig. 2). Except at very high thresholds, precision was always above 0.8. Overall false positive per hour rates never exceeded more than a few hundred false positives per hour when averaged across the data set, although specific regions of data had high false positive rates.

Folds 1 and 2 had lower precision than fold 0 even though most of the fold 0 test data were recorded 4 months earlier than the training data used in the model. In contrast, folds 1 and 2 had more training data that were recorded in close temporal proximity to their test data, resulting in better matched conditions across the train and test barrier. We note that changing the threshold has a large impact on recall but a weaker impact on precision, suggesting that one should not interpret prediction scores as a confidence metric. Neural network prediction scores are well known to not be well calibrated, a subject of recent investigation by Thulasidasan *et al.* (2019). The lower precision in folds 1 and 2 can be attributed to three sources of error: ship noise, echosounders, and instrument noise. These patterns are easily observed when unmatched detections with probability ≥0.5 are observed in a histogram with 15 min bins (Fig. 3). The histogram also shows the number of echolocation clicks in the ground truth data, demonstrating that the number of false positives is not a function of the echolocation click density. On February 20, 2015, a ship passed near the instrument, producing cavitation noise that had energy across the 80 kHz of analysis bandwidth. This produced high

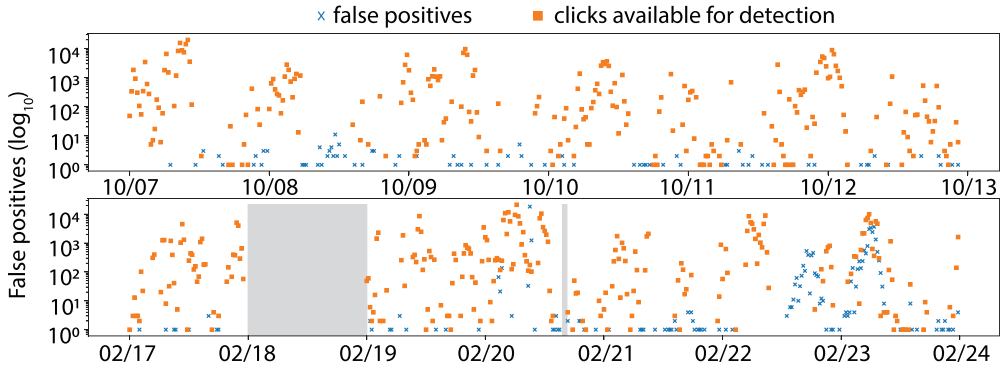

FIG. 3. (Color online) Fifteen min log-scale counts of clicks available for detection (squares) and false positive predictions (x's, probability ≥0.5) over time across the Dry Tortugas data. Top row, October 2014 data; bottom row, February 2015 data; gray shading, periods of no analysis effort. Strong peaks in the 2015 false positive counts are attributable to anthropogenic sources; see text for details.

confidence predictions of clicks and is responsible for the drop in precision in the fold 1 curve. As the threshold is raised, more and more of the predictions are dominated by the cavitation noise, lowering the precision. What is presumed to be a shipboard echosounder was present during parts of February 22 and 23. Both types of events produced large numbers of false positives. The presumed echosounder consisted of a short impulse followed by a frequency-modulated chirp that repeated approximately every 0.7 s. We assume that this was vessel mounted as the intensity varied, and there were other times that the echosounder was present in the data without causing significant interference. Both of these types of events caused spikes in the false positive rates. The final trend in error in folds 1 and 2 is from instrument self-noise. Toward the end of the disk write that occurred every 75 s, a click-like noise was observed, and this was sometimes mistakenly detected as a click. This was not as evident in fold 0, where most of the data were recorded on a different magnetic disk within the same instrument. Prediction probabilities for instrument noise events tended to be very low, with predicted click probabilities generally less than 0.55, suggesting that self-noise false positives were only influential when operating at low thresholds.

Apart from noisy conditions, which did not occur in the 6 days of fold 0, all models behaved reasonably similarly. We therefore selected the model associated with fold 0 to apply to evaluation data. This created an unmatched condition experiment from two sites that were separated by 621 km (335 nm) and nearly 5 years between recording dates.

Performance on these data (Fig. 4) was reasonably consistent in behavior with what was observed in the development data. Overall, the detector provides good performance at retrieving clicks with high precision except in areas with ship passages (one on December 21 and two on December 22). The ship passages contributed to higher false positive per hour rates at low thresholds and precisions that were between those of fold 0 and the other folds in the development data.

In general, the network predictor was able to recover the echolocation bouts that would characteristically be found by analysts as well as additional bouts that were frequently impossible to see in common visualization tools, such as long-term spectral averages (Fig. 5).

Examples of detected clicks both within bouts and in relative isolation (Fig. 6) illustrate the high precision of the system in the absence of infrequently occurring anthropogenic signals that caused the spikes in false positive rate. Missed clicks in these examples fall under the 115 dB re 1 μPa threshold.

IV. DISCUSSION

Application of machine learning-based click detectors resulted in the recovery of large numbers of validated echolocation clicks with low false positive per hour rates. Use of the machine-assisted analysis of Solsona-Berga *et al.* (2020) allowed the construction of large labeled datasets. The system showed high precision and recall except in areas with confounding anthropogenic signals. The training data did not provide any examples of the confounding signals, and it may be possible to reduce this using hard negative mining (Sung and Poggio, 1995), where difficult counterexamples are added to the training data. We also see ample opportunity for using contextual cues and are pursuing this direction in other experiments.

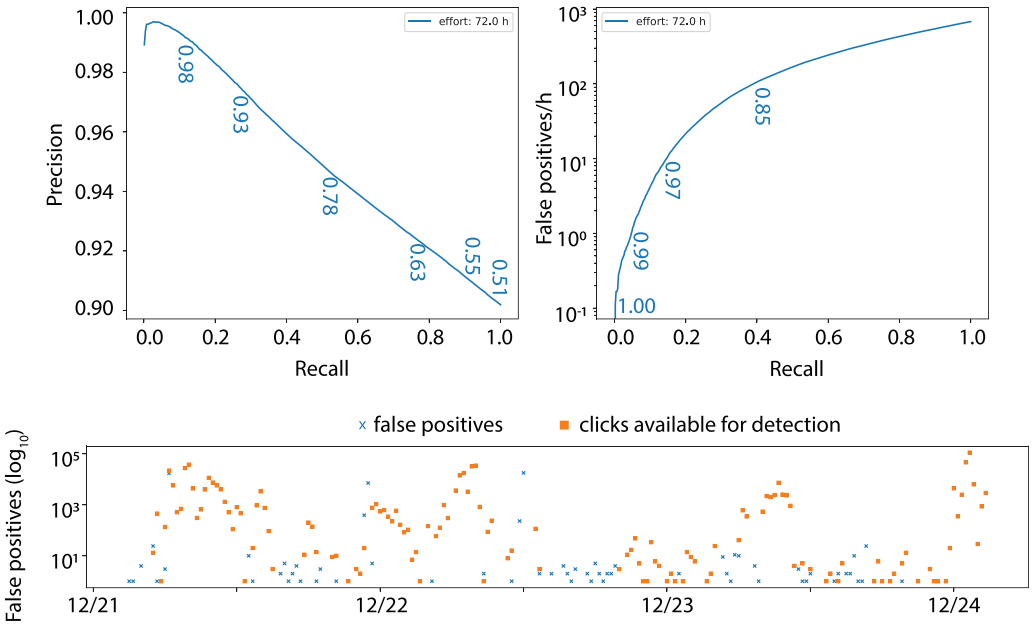

FIG. 4. (Color online) Evaluation of classifier developed from Dry Tortugas data on Mississippi Canyon data recorded 621 km away and 5 years earlier. Precision and false positive per hour over recall curves are presented in the top panel, and 15-min log-scale counts of clicks available for detection (squares) and false positives (x's) are presented in the lower panel.

https://doi.org/10.1121/10.0004992

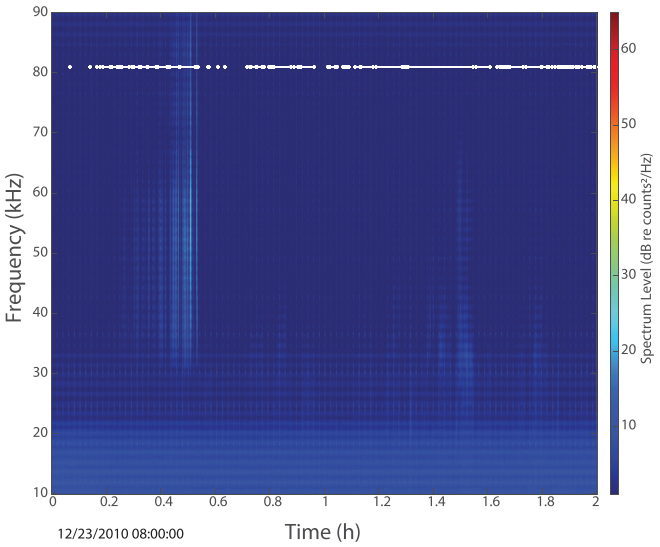

FIG. 5. (Color online) Long-term spectral average (Wiggins and Hildebrand, 2007) of 2 h of Mississippi Canyon evaluation data showing regions with predicted clicks of >0.5 probability. Spectral average was over 5 s windows of spectra, and detected clicks are denoted by white points plotted at 80 kHz. Lines are drawn across regions of clicks that are separated by <1 min. There were no false positives in this region, and many clicks are not visible in this long-term spectral average, a common tool used by analysts to identify periods of activity in long-term recordings.

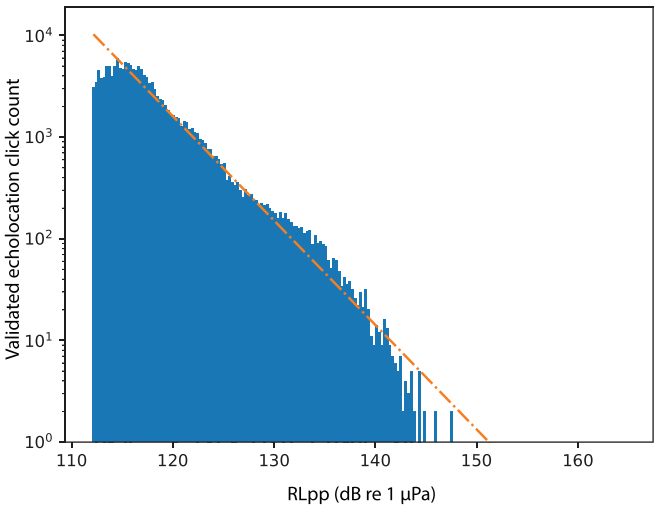

FIG. 7. (Color online) Histogram of received level intensity vs count of validated echolocation clicks in fold 1 of the development data. The regression line shows an ordinary least squares linear fit of $\log_{10}$ counts ($R = -0.978$). Other folds showed similar patterns with correlation coefficients of −0.973 and −0.983.

The detector had the ability to detect echolocation clicks with lower received levels than the echolocation clicks that were used to train the network. However, we observed that below 115 dB re 1 μPa, the number of detections was not increasing as quickly (Fig. 7). Acoustic modeling informs us that the number of detections should rise exponentially with increasing distance from the sensor (Frasier *et al.*, 2016). In the case of a fixed hydrophone, each bin in Fig. 7 corresponds to an annulus whose area grows the farther it is from the hydrophone, with lower intensity annuli tending to contain detections from animals

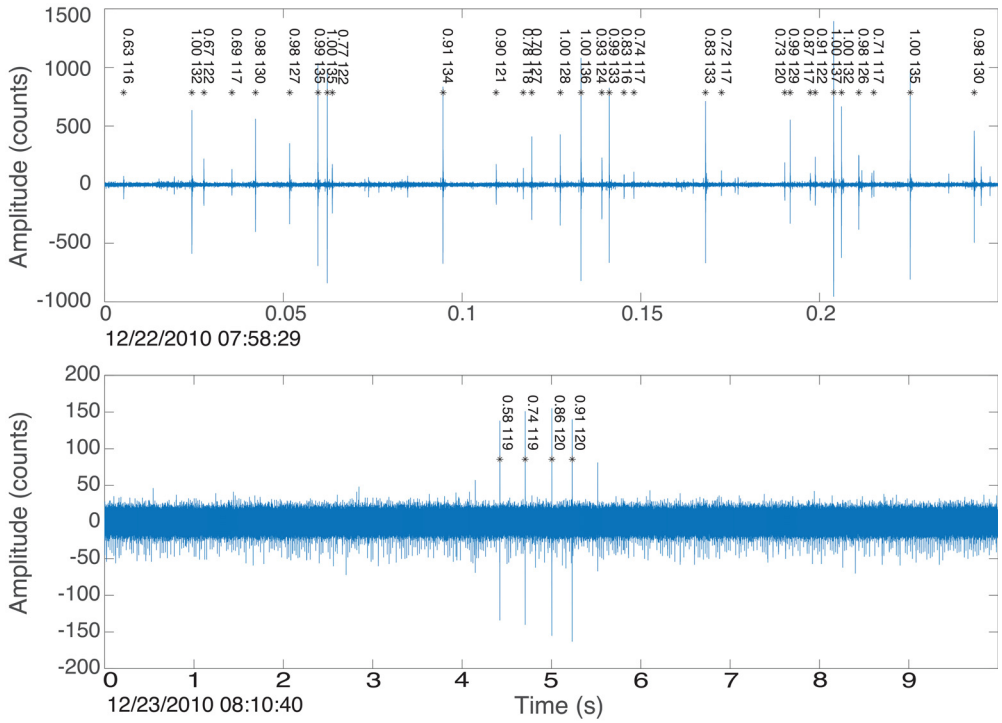

FIG. 6. (Color online) Examples of time-domain detections in the evaluation data. Upper panel shows 0.25 s of data taken from an encounter with dense echolocation clicks. Lower panel shows 10 s of data corresponding to an isolated group of detections in the long-term spectrogram of Fig. 6. Detections are marked with asterisks and labeled with the probability of echolocation click followed by the received level in dB re 1 μPa. Unlabeled echolocation clicks in these data fall beneath the 115 dB re 1 μPa threshold and were not expected to be detected.

that are farther away. While we may expect some variations of this due to the wide body of evidence that source levels vary both by species and context (e.g., Au *et al.*, 1985; Møhl *et al.*, 2003; Wahlberg *et al.*, 2011), propagation conditions (Helble *et al.*, 2013), and directionality of echolocation clicks (Au, 1993, p. 44), detection counts generally showed an inverse exponential relationship with intensity. Consequently, we assumed that the detector was beginning to miss many echolocation clicks below 115 dB re 1 μPa and only considered detections that met this received level criterion.

The addition of SNR to the time series feature vector provided valuable information for separating clicks from non-clicks, but the SNR by itself is insufficient to describe the results obtained, and the network attended to the features of the time series as well. The SNR distribution was time varying and had heavy tails (Fig. 8).

The recall in these experiments merits discussion. The recall is defined with respect to varying thresholds of a set of detections that have probabilities $\geq$0.5, meeting a received level criterion. In a traditional ground truth label set, we would have found every echolocation click meeting selection criteria in the 2 weeks of development and evaluation data. As described in Sec. II, we verified that the detector did not miss any major encounters and that detectors could retrieve all of the echolocation clicks marked as training from the initial label set, and we verified each detection produced by the detector. This leaves the opportunity for missed detections that would lower the recall if they had been recorded. As the DETEDIT process focuses on verifying detections, there is the possibility of unaccounted for clicks that meet our selection criteria. However, the focus on manual inspection of time series data along with the DETEDIT process suggests that such echolocation clicks would not play a significant role in lowering the recall. The vast majority of echolocation clicks that were not detected are due to not meeting the received level threshold (e.g., upper panel, Fig. 6). Characterizing all possible echolocation clicks is difficult; there are many times that analysts are unable to determine whether weak or isolated signals are echolocation clicks. Even strong clicks can be difficult to identify with certainty when they are in isolation.

These issues are particularly important for density estimation studies that rely on being able to adequately characterize the detection function. Strategies used for these studies are to limit the detection range by thresholding received level (e.g., Hildebrand *et al.*, 2015) or to use a trigger based on detections of groups of clicks that provide a cue, such as indicating the start of a dive for beaked whales (e.g., Marques *et al.*, 2009). For this type of work, improved click detectors may not be necessary, although they will increase the monitoring area, which is advantageous, especially for studies with poor spatial coverage. The detection of less obvious echolocation clicks with lower intensity or higher ambient noise may have stronger applications in tracking studies where the strong directivity index of the echolocation click (Au *et al.*, 1986) contributes to the difficulty of solving the association problem between clicks received on different hydrophones. It also increases the number of echolocation examples available for other tasks, such as classification to species, group, etc. In general, the classifier presented here not only improved the number of clicks found, but also discovered entire groups of echolocation clicks that were not apparent in long-term spectrograms or reported by the energy-based detector. Finally, being able to reliably detect calls is useful for understanding long-term trends and behavior, such as diel and seasonal patterns. An important component of this is detecting presence, and high false positive rates make this difficult. When operating on data outside of the potentially addressable problematic conditions outlined above (e.g., high ship noise, echosounders), our detector typically has very low false positive rates that can be described in the tens of clicks per hour, simplifying the presence/absence task.

While we have shown that this method can generalize to other sites and times, it should be noted that calibration data are not used in the data processed by this time-domain detector (apart from pruning low intensity trial bins during

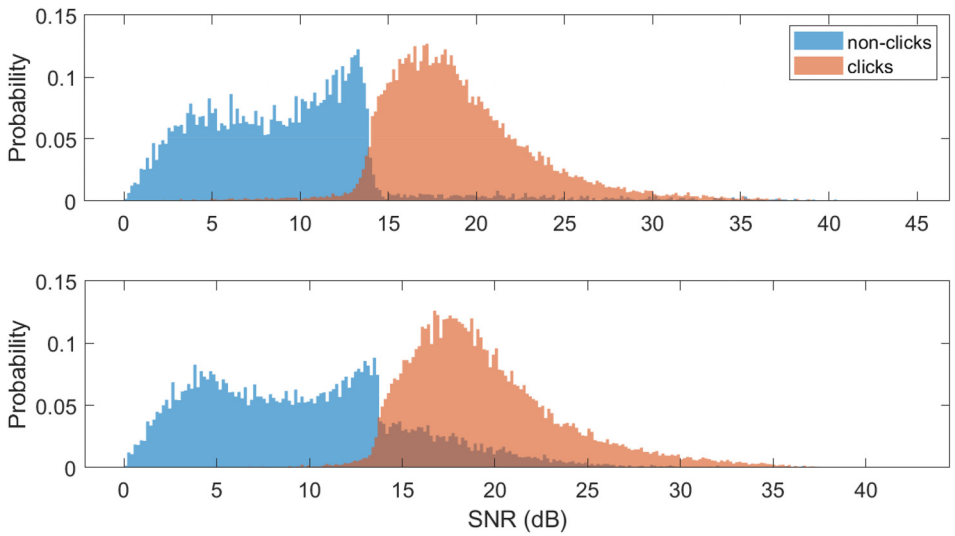

FIG. 8. (Color online) Distributions of SNRs of click present and click absent 500 μs trial bins that exceed 115 dB re 1 μPa RL_{pp}. The upper panel contains data on a day without apparent ship passages (02/19/2015). The lower panel contains data from a day with ship passages (02/22/2015). Classification based only on SNR would result in an equal error rate of 6.5% and 15.1%, respectively.

post-processing) and that instruments with significantly different characteristics may require retraining. We have shown that the system can be trained using data collected from semiautomated toolchains and that varying the amount of training data between 27 000 and 40 000 clicks does not appear to have a significant impact on classification accuracy. When applying this system to other locations, we suspect that the largest problem will be difficult classification situations for specific environments, such as snapping shrimp (genus *Alpheus* and *Synalpheus*) in shallow water deployments. Whether these can be addressed with temporal context and hard negative mining remains to be seen. Additional issues are the presence of novel species as well as differences in sample rate, the latter of which could potentially be addressed by resampling.

The network used in this study is relatively small and has approximately 20 000 parameters, making it a very modest sized network by today's standards. Execution is fast. On data that were previously bandpass filtered, the system operated at approximately 30 times real time on an Intel Core i7–9700K processor with an NVIDIA Geforce RTX 2080Ti graphics processing unit. Consequently, we believe that this algorithm could be used in real-time studies and is suitable for processing large archival data sets.

V. CONCLUSION

We have developed a time-domain based click detector that performs analysis without the use of a Fourier transform. It provides higher resolution than most spectral based methods, finding the maximal amplitude part of the click to the nearest sample. It operates on 500 μs samples at nearly 30 times real time and produces false positive rates of 1–10 false positives per hour in areas without nearby ships or echosounders, these latter situations being a focus of future research. The use of machine learning as opposed to more commonly used thresholding mechanisms allows the system to attend to the shape of the signal, and the development of recent annotation tools has enabled the generation of large training sets that provide many examples of the variations in echolocation clicks that vary both within and across species. This system has been shown to be robust across both space and time, demonstrating the ability to recognize echolocation clicks that were recorded nearly five years prior to and over 600 km distant from the site that provided training data.

ACKNOWLEDGMENTS

We thank the teams of the Scripps Institution of Oceanography–Marine Bioacoustics Research Collaborative for their dedicated work building, deploying, and processing instrumentation and data used in this paper. Funding for acoustic data collection and analysis was provided by the Natural Resource Damage Assessment partners (Grant No. 20105138). Our work was supported by the United States Office of Naval Research, Dr. Michael Weise (Grant Nos. N00014-15-1-2299 and N00014-17-1-2867). We also wish to thank Dr. Holger Klinck for his helpful comments on a draft of this article as well as the anonymous reviewers and the editor for their constructive feedback.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (**2015**). "TensorFlow: Large-scale machine learning on heterogeneous systems," http://download.tensorflow.org/paper/whitepaper2015.pdf (Last viewed May 10, 2021).

Amante, C., and Eakins, B. W. (**2009**). "ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis," NOAA Technical Memorandum NESDIS NGDC-24 (National Oceanic and Atmospheric Administration, Boulder, CO).

Au, W. W. L. (**1993**). *The Sonar of Dolphins* (Springer-Verlag, New York).

Au, W. W. L., Branstetter, B., Moore, P. W., and Finneran, J. J. (**2012a**). "The biosonar field around an Atlantic bottlenose dolphin (*Tursiops truncatus*)," J. Acoust. Soc. Am. **131**(1), 569–576.

Au, W. W. L., Branstetter, B., Moore, P. W., and Finneran, J. J. (**2012b**). "Dolphin biosonar signals measured at extreme off-axis angles: Insights to sound propagation in the head," J. Acoust. Soc. Am. **132**(2), 1199–1206.

Au, W. W. L., Carder, D. A., Penner, R. H., and Scronce, B. L. (**1985**). "Demonstration of adaptation in beluga whale echolocation signals," J. Acoust. Soc. Am. **77**(2), 726–730.

Au, W. W. L., Moore, P. W. B., and Pawloski, D. (**1986**). "Echolocation transmitting beam of the atlantic bottle-nosed-dolphin," J. Acoust. Soc. Am. **80**(2), 688–691.

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

Bermant, P. C., Bronstein, M. M., Wood, R. J., Gero, S., and Gruber, D. F. (**2019**). "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," Sci. Rep. **9**(1), 12588.

Chollet, F., O'Malley, T., Tan, Z., Bileschi, S., Gibson, A., and Allaire, J. J. (**2015**). "Keras," https://keras.io (Last viewed May 10, 2021).

Cranford, T. W. (**2000**). "In search of impulse sound sources in odontocetes," in *Hearing by Whales and Dolphins*, edited by W. W. L. Au, A. N. Popper, and R. R. Fay (Springer-Verlag, New York), pp. 109–155.

Davis, J., and Goadrich, M. (**2006**). "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd International Conference on Machine Learning, June 25–29, Pittsburgh, PA, pp. 233-240.

Fais, A., Aguilar Soto, N., Johnson, M., Pérez-González, C., Miller, P. J. O., and Madsen, P. T. (**2015**). "Sperm whale echolocation behaviour reveals a directed, prior-based search strategy informed by prey distribution," Behav. Ecol. Sociobiol. **69**, 663–674.

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

Ferrari, M., Glotin, H., Marxer, R., and Asch, M. (**2020**). "DOCC10: Open access dataset of marine mammal transient studies and end-to-end CNN classification," in Proceedings of the International Joint Conference on Neural Networks (IJCNN), July 19–24, Glasgow, UK, p. 8.

Frasier, K. E. (**2015**). "Density estimation of delphinids using passive acoustics: A case study in the Gulf of Mexico," Ph.D. thesis, University of California, San Diego, La Jolla, CA.

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.

Frasier, K. E., Wiggins, S. M., Harris, D., Marques, T. A., Thomas, L., and Hildebrand, J. A. (**2016**). "Delphinid echolocation click detection probability on near-seafloor sensors," J. Acoust. Soc. Am. **140**(3), 1918–1930.

Gillespie, D. M. (**1997**). "An acoustic survey for sperm whales in the Southern Ocean sanctuary conducted from the R/V Aurora Australis,"

Report 47 (International Whaling Commission, Cambridge, UK), pp. 897–908.
Gillespie, D., and Caillat, M. (**2008**). "Statistical classification of odontocete clicks," Can. Acoust. **36**(1), 20–26.
Helble, T. A., D'Spain, G. L., Hildebrand, J. A., Campbell, G. S., Campbell, R. L., and Heaney, K. D. (**2013**). "Site specific probability of passive acoustic detection of humpback whale calls from single fixed hydrophones," J. Acoust. Soc. Am. **134**(3), 2556–2570.
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.
Hildebrand, J. A., Frasier, K. E., Baumann-Pickering, S., Wiggins, S. M., Merkens, K. P., Garrison, L. P., Soldevilla, M. S., and McDonald, M. A. (**2019**). "Assessing seasonality and density from passive acoustic monitoring of signals presumed to be from pygmy and dwarf sperm whales in the Gulf of Mexico," Front. Mar. Sci. **6**, 66.
Houser, D. S., Helweg, D. A., and Moore, P. W. (**1999**). "Classification of dolphin echolocation clicks by energy and frequency distributions," J. Acoust. Soc. Am. **106**(4), 1579–1585.
Johnson, M., Madsen, P. T., Zimmer, W. M. X., de Soto, N. A., and Tyack, P. L. (**2006**). "Foraging Blainville's beaked whales (*Mesoplodon densirostris*) produce distinct click types matched to different phases of echolocation," J. Exp. Biol. **209**(24), 5038–5050.
Kaiser, J. F. (**1990**). "On a simple algorithm to calculate the 'energy' of a signal," in Proceedings of the 23rd International Conference on Acoustics, Speech, and Signal Processing (ICASSP), April 3–6, Albuquerque, NM, pp. 381-384.
Kandia, V., and Stylianou, Y. (**2006**). "Detection of sperm whale clicks based on the Teager-Kaiser energy operator," Appl. Acoust. **67**(11), 1144–1163.
Kandia, V., and Stylianou, Y. (**2008**). "A phase based detector of whale clicks," in Proceedings of 2008 New Trends for Environmental Monitoring Using Passive Systems, October 14–17, Hyeres, France, p. 6.
Kingma, D. P., and Ba, J. (**2015**). "Adam: A method for stochastic optimization," in Proceedings of the 3rd International Conference for Learning Representations, May 7–9, San Diego, CA, p. 15.
Klinck, H., and Mellinger, D. K. (**2011**). "The energy ratio mapping algorithm: A tool to improve the energy-based detection of odontocete echolocation clicks," J. Acoust. Soc. Am. **129**(4), 1807–1812.
Luo, W., Yang, W., and Zhang, Y. (**2019**). "Convolutional neural network for detecting odontocete echolocation clicks," J. Acoust. Soc. Am. **145**(1), EL7–12.
Madhusudhana, S., Gavrilov, A., and Erbe, C. (**2015**). "Automatic detection of echolocation clicks based on a Gabor model of their waveform," J. Acoust. Soc. Am. **137**(6), 3077–3086.
Marques, T. A., Thomas, L., Ward, J., DiMarzio, N., and Tyack, P. L. (**2009**). "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales," J. Acoust. Soc. Am. **125**(4), 1982–1994.
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, September 22–25, Rhodes, Greece, pp. 1895-1898.
Mellinger, D. K. (**2001**). "Ishmael 1.0 User's Guide," NOAA Technical Memorandum OAR-PMEL-120 (NOAA/Pacific Marine Environmental Laboratory, Seattle, WA).
Møhl, B., Wahlberg, M., Madsen, P. T., Heerfordt, A., and Lund, A. (**2003**). "The monopulsed nature of sperm whale clicks," J. Acoust. Soc. Am. **114**(2), 1143–1154.
Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted Boltzmann machines," in Proceedings of the International Conference on Machine Learning (ICML), June 21–24, Haifa, Israel.
Roch, M. (**2019**). "clicknet," https://bitbucket.org/marie_r/clicknet (Last viewed May 10, 2021).
Roch, M. A., Batchelor, H., Baumann-Pickering, S., Berchok, C. L., Cholewiak, D., Fujioka, E., Garland, E. C., Herbert, S., Hildebrand, J. A., Oleson, E. M., Van Parijs, S. M., Risch, D., and Širović, A. (**2016**). "Management of acoustic metadata for bioacoustics," Ecol. Inform. **31**, 122–136.
Roch, M. A., Klinck, H., Baumann-Pickering, S., Mellinger, D. K., Qui, S., Soldevilla, M. S., and Hildebrand, J. A. (**2011**). "Classification of echolocation clicks from odontocetes in the Southern California Bight," J. Acoust. Soc. Am. **129**(1), 467–475.
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.
Soldevilla, M. S. (**2008**). "Risso's and pacific white-sided dolphins in the Southern California bight using echolocation clicks to study dolphin ecology," in *Oceanography* (University of California at San Diego, La Jolla, CA).
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.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (**2014**). "Dropout: A simple way to prevent neural networks from overfitting," J. Mach. Learn. Res. **15**, 1929–1958.
Sung, K. K., and Poggio, T. (**1995**). "Learning human face detection in cluttered scenes," in *Comp. Anal. Images Patterns*, edited by V. Hlaváč and R. Šára (Springer, Berlin), pp. 432-439.
Thulasidasan, S., Chennupati, G., Bilmes, J., Bahattacharya, T., and Michalak, S. (**2019**). "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), December 10–12, Vancouver, Canada, p. 15.
Wahlberg, M., Jensen, F. H., Soto, N. A., Beedholm, K., Bejder, L., Oliveira, C., Rasmussen, M., Simon, M., Villadsgaard, A., and Madsen, P. T. (**2011**). "Source parameters of echolocation clicks from wild bottlenose dolphins (*Tursiops aduncus* and *Tursiops truncatus*)," J. Acoust. Soc. Am. **130**(4), 2263–2274.
Wiggins, S. M., and Hildebrand, J. A. (**2007**). "High-frequency acoustic recording package (HARP) for broad-band, long-term marine mammal monitoring," in Proceedings of the International Symposium on Underwater Technology, April 17–20, Tokyo, Japan, pp. 551-557.
Zimmer, W. M. X., Harwood, J., Tyack, P. L., Johnson, M. P., and Madsen, P. T. (**2008**). "Passive acoustic detection of deep-diving beaked whales," J. Acoust. Soc. Am. **124**(5), 2823–2832.
Zimmer, W. M. X., Johnson, M. P., Madsen, P. T., and Tyack, P. L. (**2005a**). "Echolocation clicks of free-ranging Cuvier's beaked whales (*Ziphius cavirostris*)," J. Acoust. Soc. Am. **117**(6), 3919–3927.
Zimmer, W. M. X., Madsen, P. T., Teloni, V., Johnson, M. P., and Tyack, P. L. (**2005b**). "Off-axis effects on the multipulse structure of sperm whale usual clicks with implications for sound production," J. Acoust. Soc. Am. **118**(5), 3337–3345.