Using context to train time-domain echolocation click detectors\textsuperscript{a)}

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

\textsuperscript{1}Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, California 92182-7720, USA
\textsuperscript{2}Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive #0205, La Jolla, California 92093, USA
\textsuperscript{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.

\textsuperscript{a)}This paper is part of a special issue on Machine Learning in Acoustics.
\textsuperscript{b)}Electronic mail: Marie.Roch@sdsu.edu

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 \textit{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 \textit{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 \textit{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 \textit{et al.}, 2008). Teager’s short-time energy operator (Kaiser, 1990) is an alternative energy...
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 ≥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 near the Dry Tortugas (25.539°N, 84.631°W) in fall and winter 2014–2015. The second deployment was recorded near the Dry Tortugas (25.539°N, 84.631°W) in fall and winter 2014–2015. The second deployment was recorded.

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 coeruleaodonta), 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 (Orcaius 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 (10–90kHz) 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 90kHz 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 2kHz 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 (RLpp) 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 × 106 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 RLpp, 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 200kHz) 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 $\geq 0.5$, 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 $\mu$s 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.

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 $\geq 0.5$ and received levels $\geq 112\, \text{dB re } 1\, \mu\text{Pa}$, 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 $\mu$Pa (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

### Table II. Dry Tortugas data used in threefold cross-validation experiment.

<table>
<thead>
<tr>
<th>Fold</th>
<th>Data interval (UTC)</th>
<th>Training clicks from other folds</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1. 2014/10/07 00:00–2014/10/12 23:59</td>
<td>27,129</td>
</tr>
<tr>
<td></td>
<td>2. 2015/02/17 00:00–2015/02/17 23:59</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3. 2015/02/19 00:00–2015/02/20 15:37</td>
<td>41,346</td>
</tr>
<tr>
<td>2</td>
<td>4. 2015/02/20 16:32–2015/02/23 23:59</td>
<td>44,523</td>
</tr>
</tbody>
</table>
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 \( \mu \text{s} \) 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 \( 7.2 \times 10^5 \) 500 \( \mu \text{s} \) 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 \( \mu \text{Pa} \) 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 \( \geq 0.5 \) 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...
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. 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.
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.
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; Mohl 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 ≥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

![Figure 8](https://doi.org/10.1121/10.0004992)
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 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 Klinkin for his helpful comments on a draft of this article as well as the anonymous reviewers and the editor for their constructive feedback.


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).