Silbido profundo: An open source package for the use of deep learning to detect odontocete whistles

Peter C. Conant, Pu Li, Xiaobai Liu, et al.

Citation: The Journal of the Acoustical Society of America **152**, 3800 (2022); doi: 10.1121/10.0016631

View online: https://doi.org/10.1121/10.0016631

View Table of Contents: https://asa.scitation.org/toc/jas/152/6

Published by the Acoustical Society of America

Silbido profundo: An open source package for the use of deep learning to detect odontocete whistles

Peter C. Conant,[1] Pu Li,[1] Xiaobai Liu,[1] Holger Klinck,[2] Erica Fleishman,[3] Douglas Gillespie,[4] Eva-Marie Nosal,[5] and Marie A. Roch[1,a)]

[1]*Department of Computer Science, San Diego State University, San Diego, California 92182, USA*

[2]*K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, New York, New York 14850, USA*

[3]*College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA*

[4]*Sea Mammal Research Unit, Scottish Oceans Institute, University of St. Andrews, St. Andrews, KY16 9AJ, United Kingdom*

[5]*Department of Ocean and Resources Engineering, University of Hawai'i at Mānoa, Honolulu, Hawaii 96822, USA*

ABSTRACT:

This work presents an open-source MATLAB software package for exploiting recent advances in extracting tonal signals from large acoustic data sets. A whistle extraction algorithm published by Li, Liu, Palmer, Fleishman, Gillespie, Nosal, Shiu, Klinck, Cholewiak, Helble, and Roch [(**2020**). *Proceedings of the International Joint Conference on Neural Networks*, July 19–24, Glasgow, Scotland, p. 10] is incorporated into *silbido*, an established software package for extraction of cetacean tonal calls. The precision and recall of the new system were over 96% and nearly 80%, respectively, when applied to a whistle extraction task on a challenging two-species subset of a conference-benchmark data set. A second data set was examined to assess whether the algorithm generalized to data that were collected across different recording devices and locations. These data included 487 h of weakly labeled, towed array data collected in the Pacific Ocean on two National Oceanographic and Atmospheric Administration (NOAA) cruises. Labels for these data consisted of regions of toothed whale presence for at least 15 species that were based on visual and acoustic observations and not limited to whistles. Although the lack of per whistle-level annotations prevented measurement of precision and recall, there was strong concurrence of automatic detections and the NOAA annotations, suggesting that the algorithm generalizes well to new data.

https://doi.org/10.1121/10.0016631

(Received 5 September 2022; revised 5 December 2022; accepted 8 December 2022; published online 27 December 2022)

[Editor: Shane Guan] Pages: 3800–3808

I. INTRODUCTION

Cetaceans, or members of the order that includes baleen and toothed whales, make a variety of sounds that are used extensively for navigation, foraging, and communication (Au and Hastings, 2008). The ability to reliably detect their calls enables passive acoustic monitoring (PAM) studies to address many research and management objectives (Van Parijs *et al.*, 2009). Examples of such objectives include localizing and tracking animals (e.g., Helble *et al.*, 2015), species identification (e.g., Gillespie *et al.*, 2013), potential identification of individuals (Gridley *et al.*, 2014; McCordic *et al.*, 2016), characterizing distributions and behavior (e.g., Baumann-Pickering *et al.*, 2014; Širović *et al.*, 2015), and estimating density (e.g., Marques *et al.*, 2011).

Many cetaceans produce tonal calls in which a narrowband signal varies in frequency over time (Au and Hastings, 2008). In toothed whales, these frequency-modulated calls, which are known as whistles, can carry information such as the identity of individuals (e.g., Caldwell and Caldwell, 1971; Janik and Sayigh, 2013) or populations (e.g., Bonato *et al.*, 2015; Van Cise *et al.*, 2017), and are believed to play a role in communication (e.g., King *et al.*, 2021). Many systems for detecting cetacean tonal calls report call presence in a certain frequency band but without extracting information about the fluctuating frequency itself (e.g., Thomas *et al.*, 2019; Kirsebom *et al.*, 2020; Shiu *et al.*, 2020). In some applications, such as identification of individuals or analysis of impacts of anthropogenic activities (e.g., Janik and Sayigh, 2013; Heiler *et al.*, 2016; Antichi *et al.*, 2022), analysis of detailed information about the time-varying frequency, such as conducted by Buck and Tyack (1993) or Deecke and Janik (2006), becomes critical. In addition, some forms of density estimation rely on call production rates, which can then be used to estimate density on the basis of the number of detected calls (Marques *et al.*, 2013). When systems only report presence, multiple calls within the same detection period may be undercounted, leading to a bias in the estimate. In this paper, we discuss an open-source, deep learning system for extracting time-frequency information on tonal calls in passive acoustic data. Although we focus on the example of toothed whales and hereafter

[a)]Electronic mail: marie.roch@sdsu.edu use the toothed whale-specific term "whistle," we expect the methods to be applicable to tonal calls produced by other taxa (e.g., baleen whale "moans").

Most methods for automating whistle extraction are performed over a time × frequency representation of a signal and attempt to extract contour ridges. Some methods connect peaks into fragments and then connect the fragments by searching along a predicted polynomial path to track the frequency of whistles over time (e.g., Mellinger *et al.*, 2011; Roch *et al.*, 2011; Gillespie *et al.*, 2013). Statistical filtering processes such as Kalman (Mallawaarachchi *et al.*, 2008), particle (White and Hadley, 2008; Roch *et al.*, 2011), and variants of particle hypothesis density filters (Gruden and White, 2016; Gruden and White, 2020) also are effective. Other methods include utilizing the instantaneous frequency of tonal signals to identify local maxima (Ioana *et al.*, 2010; Lin *et al.*, 2013), ridge detection (Serra *et al.*, 2020), and ridge regression (Kershenbaum and Roch, 2013).

Deep learning neural networks have been used to extract tonal information in human speech and musical tasks (Han and Wang, 2014; Bittner *et al.*, 2017), and have been used for detection of calls or identification of species in many other bioacousitc monitoring projects (see Stowell, 2022, for a recent review). These methods have demonstrated good performance on bioacoustic detection and classification tasks across a wide variety of taxonomic groups, including birds, mammals, and insects (Mac Aodha *et al.*, 2018; Bermant *et al.*, 2019; Oikarinen *et al.*, 2019; Stowell *et al.*, 2019; Frasier, 2021; Hoye *et al.*, 2021). Li *et al.* (2020) used deep learning-predicted peaks of odontocete whistles in conjunction with the graph search algorithm of Roch *et al.* (2011) to extract whistle annotations from the set of time-frequency predictions. Their method, *deep whistle*, is the focus of this paper.

Silbido is an open-source software package that uses a graph search algorithm to annotate cetacean tonal vocalizations (Roch *et al.*, 2011). The graphical user interface permits audition and spectrogram visualization of recordings and can invoke a graph search algorithm to generate annotations automatically. In addition, it supports analyst annotation of whistles by permitting users to specify control points (knots) that are joined via a cubic spline. Our goal was to improve the automated annotation performance of *silbido* by incorporating the Li *et al.* (2020) *deep whistle* model into software that can easily be used by the biology community. Throughout this paper, we refer to this implementation as *silbido profundo* (Spanish for deep whistle) to distinguish it from the original *deep whistle* implementation (Li *et al.*, 2020).

II. METHODS

A. Data sets

We used two data sets to examine the performance of *silbido profundo*; a smaller data set that provided detailed whistle annotations and a second with less-detailed labels over long-duration recordings. The first data set was a subset of time × frequency annotated dolphin whistles from the 2011 Detection, Classification, Localization and Density Estimation (DCLDE 2011) workshop data (DCLDE Organizing Committee, 2011). These data were recorded with ITC 1042 (International Trandsucer Corp., Santa Barabara, CA) and HS 150 (Sonar Research and Development Ltd., Beverly, UK) hydrophones that were sampled at 192 kHz with 16 or 24-bit quantization. Recordings were typically made from 10 to 30 m below the surface from a variety of platforms that were either towed or stationary. Each whistle was annotated by a trained analyst with time-varying frequency information. See Roch *et al.* (2011) for further details on data collection and analyst annotation protocols.

A subset of 7161 whistle contours produced by common (*Delphinus spp.*) and bottlenose (*Tursiops truncatus*) dolphins from the DCLDE 2011 data were used by Li *et al.* (2020) to train the *deep whistle* model. They sampled portions of spectrograms with and without whistle energy, creating 148 224 training patches that were used to train the model as summarized in the next section. To evaluate our implementation of their model, we used the same 911 bottlenose and common dolphin (*Tursiops truncatus* and *Delphinus capensis*) whistles[1] that Li *et al.* (2020) used for their test set. This is a challenging subset of the data that were reported by Roch *et al.* (2011).

A larger data set was used to gather evidence as to whether the *silbido profundo* methods are transferrable to signals that are from other species or regions or recorded with different equipment. The DCLDE 2022 data set from the National Oceanographic and Atmospheric Administration (NOAA) (NOAA Pacific Islands Fisheries Science Center, 2022) consisted of over 432 h of recordings from 47 days of effort during two towed-array expeditions aboard the R/V Lasker and R/V Sette. The cruises were conducted from July through November 2017 offshore of the Hawaiian Islands during the Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS). The cruises towed multichannel arrays of HTI-96-min hydrophones (High Tech Inc., Long Beach, MS) and custom preamplifiers. Data were sampled at 500 kHz with 16-bit quantization, and only the first channel of data was used in our experiments.

Toothed whale presence in these data were reported by NOAA visual and acoustic teams. The acoustic teams used PAMGuard (Gillespie *et al.*, 2008). The teams produced toothed whale annotations at the encounter level, noting the time, duration, and, when possible, species for each period of time during which a visual sighting occurred or acoustic cues (clicks or whistles) were present. Annotations for 276 encounters of at least 15 species of toothed whales were reported. Because the annotations are based on a combination of visual and acoustic cues and are not reported on a per-whistle basis, the cruise annotations are insufficient for computing the precision and recall of a whistle annotation task. However, the encounter-level annotations permit qualitative analysis of how well *silbido profundo* concurs with less-detailed analyst annotations of data that are substantially different than those used to train the system. Complete details on the HICEAS equipment, collection, and annotation protocols are in Yano *et al.* (2018).

B. Deep whistle

We used the *deep whistle* model proposed by Li *et al.* (2020). This neural network model used local convolutional kernels to learn contextual cues about whistle energy patterns from multiple fixed-duration spectrograms derived from a recording. The network produced confidence maps for whistle energy that could then be used by a traditional whistle extraction method to provide annotations for whistles (Fig. 1).

The network consists of ten convolutional layers that were trained on 100 ms by 6.25 kHz spectrogram patches with binary labels indicating the presence or absence of whistle energy. The first and last layers are standard convolutional layers that surround four residual blocks (He *et al.*, 2016) with two layers each. Hidden layers have 32 channels. Convolutional layers of the residual blocks are all followed by batch normalization (Ioffe and Szegedy, 2015), with a parametric rectified linear unit (He *et al.*, 2015) following the batch normalization of the first convolution in each residual block. The first and last convolutional layers had 5 × 5 convolutional kernels and the residual blocks had 3 × 3 kernels, resulting in a receptive field of 25 × 25, or 56 ms × 3.125 kHz. The network was initialized with Kaiming normalization (He *et al.*, 2015) and applied a Charbonnier loss (Charbonnier *et al.*, 1994) for the gradient calculation

$$\mathrm{Loss}(\hat{y} - y) = \sqrt{\|\hat{y} - y\|_2^2 + \varepsilon}, \qquad (1)$$

where y was a vector of ground truth labels for spectrogram nodes (0 → background, 1 → foreground whistle), $\hat{y}$ the network prediction, and ε a small constant. The learning rate was initially set to 0.001 and was decayed by a factor of 0.1 every 250 000 iterations. Additional details about the network architecture are in Li *et al.* (2020), and we use the network weights from the experiment of Li *et al.* (2020) that did not use synthetic data or labels (experiment "WGT").

C. Integrating the deep whistle method into silbido

The original *deep whistle* algorithm was not designed to be easily accessible to people outside of the machine learning community. The system relies on code written in Python, Java, and MATLAB (Mathworks, Natick, MA) with dependencies on PyTorch (Paszke *et al.*, 2019) and MATLAB toolboxes, each of which requires a separate download and installation. This system requires knowledge of multiple programming languages, and the manual steps necessary to move files and change code can slow processing and discourage the average user.

To make our software more usable by a diverse community, we sought to integrate *deep whistle* functionality within *silbido* and reduce dependencies to a few MATLAB toolboxes.[2] To achieve this goal, we reimplemented the original Python signal processing chain in MATLAB and migrated the PyTorch neural network into a form usable with the MATLAB deep learning toolbox. The processing chain uses discrete Fourier transforms with Hamming windowed frames of 8 ms (125 Hz resolution) advanced every 2 ms to create a log magnitude spectrogram. We restricted the dynamic range to 0 through 6 based on empirical results. As these are log magnitude values, one can multiply by 20 to recognize that the values correspond to the relative intensity range of 0 to

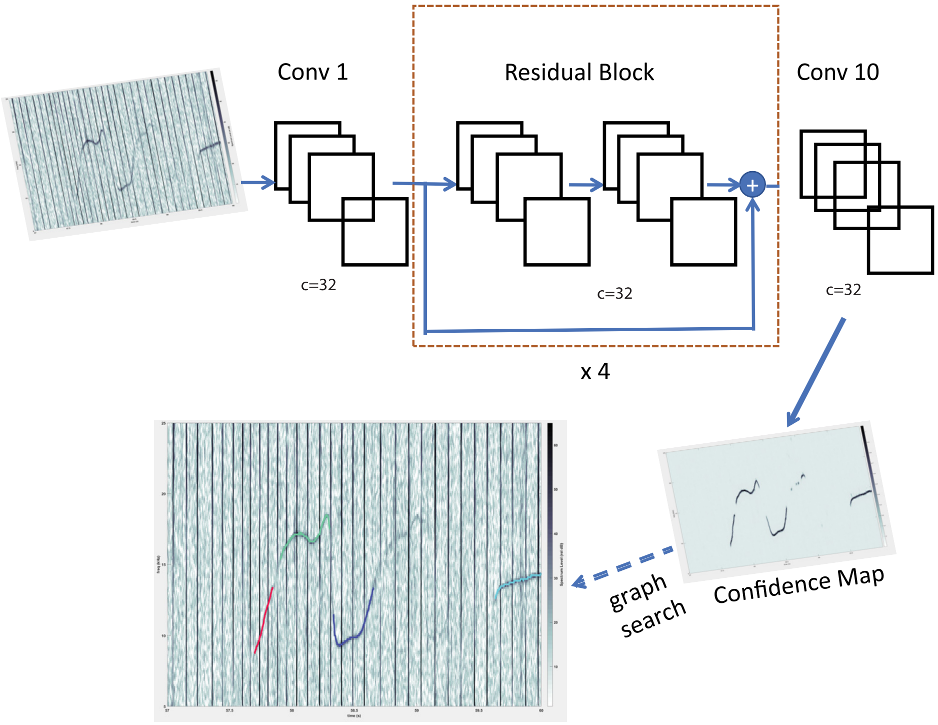

FIG. 1. (Color online) The spectrogram is processed with a deep convolutional neural network that has learned how to recognize whistle energy. This produces a confidence map that is processed with an existing graph search algorithm to extract whistles. State is maintained across consecutive confidence maps to reduce artificial breaks in whistles similar to Roch *et al.* (2011).

120 dB. The log magnitude values are normalized to the interval [0,1] by dividing by 6.

We converted the PyTorch model to an intermediate neural network description language, the Open Neural Network Exchange (ONNX) (Bai *et al.*, 2019). We used the MATLAB *Deep Learning Toolbox Converter for ONNX model Format* to convert the model to the directed acyclic graph (DAG) network format used by the MATLAB deep neural network toolbox.

In general, convolutional networks can operate on spectrograms of arbitrary sizes. The MATLAB DAG networks are restricted to a static input size of the network implementer's choice. This limitation would prevent users from changing spectrogram parameters such as the frequency analysis range. To enable more flexible analysis within the MATLAB DAG networks, we implemented code to dynamically generate DAG networks containing the *deep whistle* weights. We generated input and output layers sized appropriately to the audio sample rate and the current user-specified spectrogram analysis parameters. We then programmatically inserted the *deep whistle* network hidden layers. This process allows the analysis window to be changed to cover signals that occur in different frequency bands or that were sampled at different rates. Although the size of the input spectrogram changes, the time × frequency resolution remains constant as long as the spectrogram frame advance and length are of the same duration. Changing the temporal and frequency resolution is permitted, but more-than-minor deviations are likely to degrade performance of the *deep whistle* neural network unless the weights are adapted for the new resolution with additional training data. Figure 2 shows a sample input spectrogram, the intermediate stage of predicting a confidence map containing probabilities that time × frequency cells contain whistle energy, and the resulting whistles when the confidence map is incorporated into a traditional trajectory-tracking whistle algorithm. Incorporating this process into an established software package resulted in an automated whistle annotation system with the potential for broad use within the bioacoustics community.

III. TESTS AND RESULTS

A. Quantitative performance metrics

We measured the performance of *silbido profundo* on the basis of precision, recall, and F-1 score. Precision measures the percentage of detections that are correct and provides insight into the false positive rate. Recall, the fraction of expected detections that were retrieved provides insight into the rate of ground truth detections missed by the system. The F-1 score, the harmonic mean between precision and recall, can be used to summarize performance.

Ground truth data are required to measure precision and recall. The DLCDE 2011 data set provides analyst annotations that yield time × frequency data. We used the performance metrics outlined in Roch *et al.* (2011). We limited our measurements to whistles with durations ≥150 ms and a signal to noise ratio ≥10 dB relative over at least a third of their duration to remain consistent with the metrics of Li *et al.* (2020). When detections overlapped with ground truth

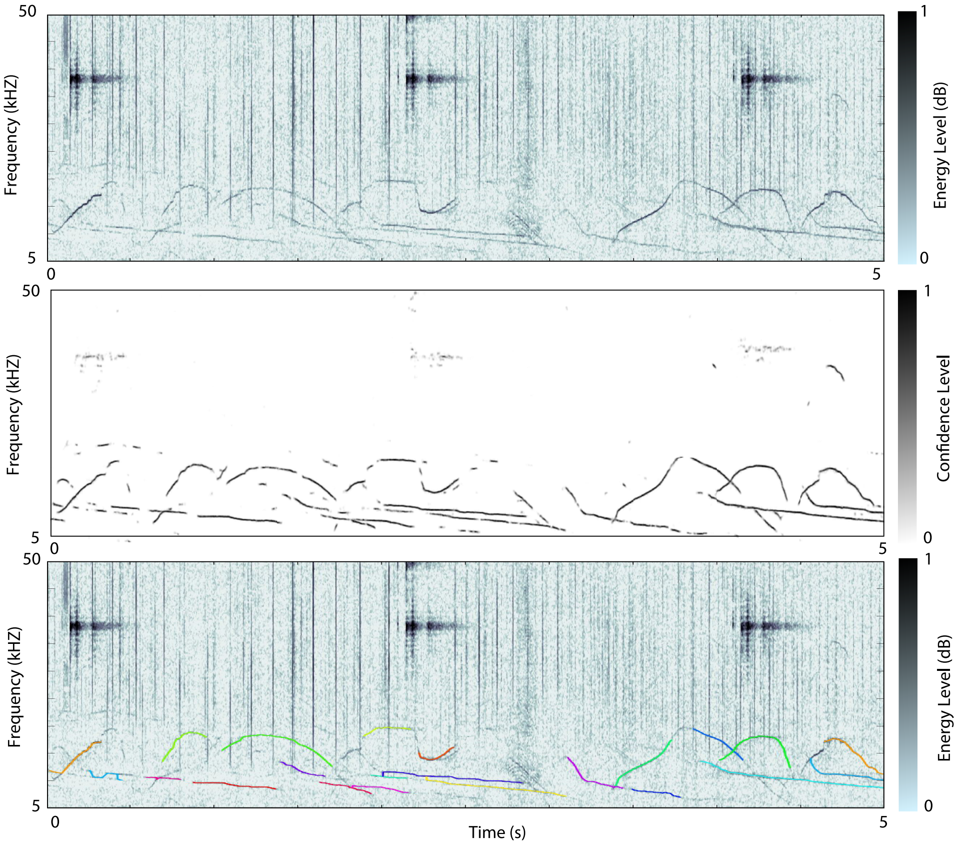

FIG. 2. (Color online) The *silbido profundo* whistle extraction system. Upper panel: Spectrogram illustrating whistles in the presence of echolocation clicks and a ship echosounder. Middle panel: Confidence map of neural network predictions of whistle energy. Lower panel: Whistles extracted (randomly colored) by the system.

annotations, we examined them to ensure that the detected tonal contour matched the whistle annotated by the analyst. This was accomplished by calculating the frequency deviation between each overlapping time × frequency bin of the ground truth and detected tonal. We computed a coverage metric (Roch *et al.*, 2011) that indicates the percentage of the matching overlap between the ground-truth whistle and detection. Detected tonal calls were marked as valid only if the mean deviation was ≤350 Hz.

Comparison results are reported with the default parameters of each algorithm, but variation in the threshold used in *silbido profundo*'s confidence map produces varying precision and recall (Fig. 3). On these data, the highest F-1 score of 87.2 results from a confidence threshold of 0.5 units, with reasonably stable performance in a region around the confidence metric, suggesting that the system is not overly sensitive to the confidence threshold. This performance is measured under the assumption that a maximum harmonic mean of the precision and recall is the goal. If either precision or recall is a higher priority, the confidence threshold can be adjusted.

We compared *silbido profundo* with three other detectors (Table I): *silbido* (the baseline method), the sequential Monte-Carlo probability hypothesis density (SMC-PHD) filter using the radial basis function motion model (Gruden and White, 2020), and the original implementation of *deep whistle* by Li *et al.* (2020). The *silbido* graph search results reflect that the test subset was more challenging than the full data set used in Roch *et al.* (2011). We used default parameters for all algorithms (Roch *et al.*, 2011; Gruden and White, 2020 Table I; Li *et al.*, 2020), with *silbido profundo* sharing the same defaults as *deep whistle*. All of the algorithms used DCLDE 2011 data in their development, but any of the algorithms might perform better if they were carefully tuned for this subset of DCLDE 2011 data.

By design, *silbido* discards detections that are shorter than 150 ms, which frequently are unreliable. To engender appropriate comparison, we reported SMC-PHD both with all detections and with those that are 150 ms or longer. Due to its stochastic nature, ten trials of SMC-PHD were conducted and the 80th decile F1 score was reported to provide a favorable but realistic expectation of the algorithm's performance. There are minor differences between the original *deep whistle* results and those of the MATLAB *silbido profundo* implementation, most likely attributable to the difference between neural network libraries and to numerical stability across different implementations of the underlying mathematics libraries.

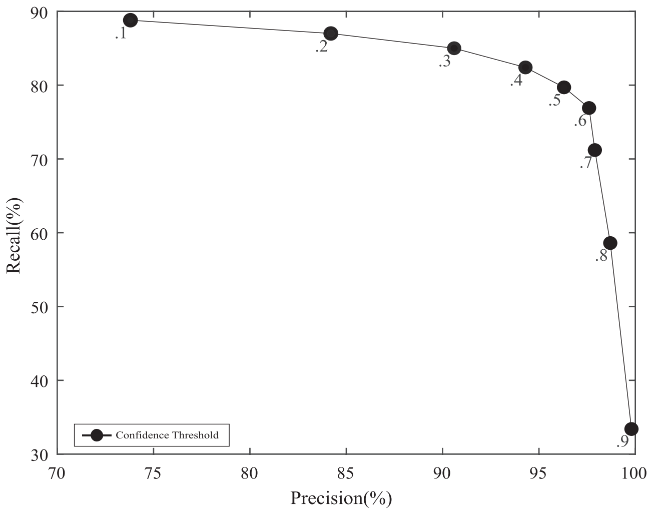

FIG. 3. Precision and recall based on *silbido profundo* confidence map thresholds of 0.1 through 0.9 for a subset of the DCLDE 2011 data. Comparison metrics are computed with the default threshold of 0.5, but the F1 score is optimized at 0.4.

TABLE I. A comparison of *silbido profundo* to other whistle extraction methods and the original *deep whistle* implementation using a subset of DCLDE 2011 data. All algorithms used default parameters and SMC-PHD used the radial basis function motion model. *Silbido profundo* maintained the level of performance of the original *deep whistle* implementation and outperformed other methods.

Method	Precision	Recall	F-1 Score	Coverage
***Silbido* Graph Search** (Roch *et al.*, 2011)	63.4	63.3	63.4	79.5 ± 22.5
SMC-PHD (Gruden and White, 2020)	70.5	92.6	80.1	70.5 ± 24.3
SMC-PHD **(det > 150 ms)**	96.1	61.4	74.9	73.6 ± 22.1
Deep Whistle (Li *et al.*, 2020)	95.6	82.2	88.4	86.6 ± 18.3
Silbido profundo	96.3	79.7	87.2	85.2 ± 19.1

B. Qualitative analysis of large data sets

To test *silbido profundo's* performance over a large data set, we used a Linux machine with an Intel i7–9700 processor (Intel Inc., Santa Clara, CA) and an NVIDIA RTX 2080 Ti (NVIDIA Inc., Santa Clara, CA) graphics processing unit (GPU). We processed the first channel of towed array data from the 439 h of the 500 kHz DCLDE 2022 data set (NOAA Pacific Islands Fisheries Science Center, 2022). On average, detections were processed 5.5 times faster than real time, although the rate varied depending on the complexity of the soundscape. We applied the same rules for duration and signal-to-noise ratio (SNR) described above to this test.

Next, 462 053 time-frequency contours were extracted from these audio data and stored in Tethys (Roch *et al.*, 2016), a database for organizing acoustic metadata. We arbitrarily grouped detections into blocks of 2 h and report the location and number of whistles within these periods (Fig. 4). It was not possible to compute per-call precision and recall with respect to these data because analyst annotations only reported the start and end times of groups of detections; annotating these data was beyond the scope of this work. We visually compared our results to the encounter periods reported by the teams of visual observers and acousticians onboard the R/V Lasker and R/V Sette cruises and observed good concurrence (subset of data shown in Fig. 5). We detected whistle signals in 186 of the 276 encounters.

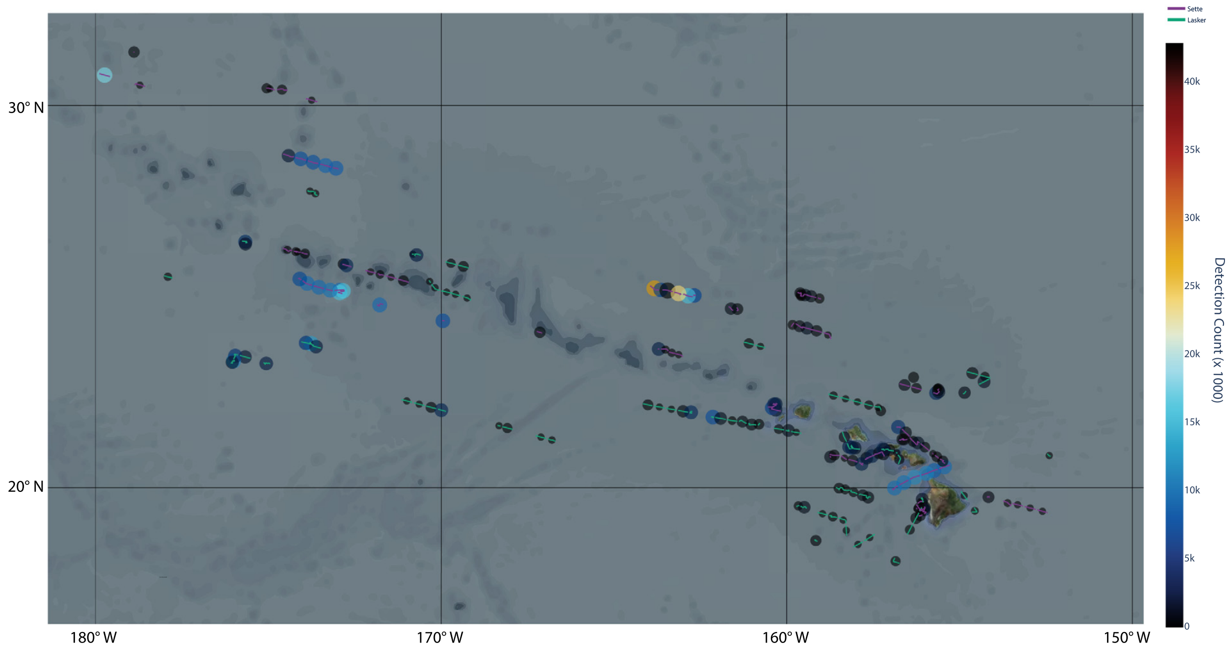

FIG. 4. (Color online) Regions of effort and reported whistle detections on the R/V Lasker and R/V Sette cruises around the Hawaiian Islands. Bubbles are overlayed on 2 h track segments with the number of whistles detected indicated by size (logarithmic scale) and color.

Analyst inspection of the remaining 90 encounters showed that they contained echolocation clicks without whistles, which *silbido profundo* is not designed to detect. Our system detected whistles outside of the NOAA labels, which we were able to verify as good detections. Due to the differences in annotation tasks, we did not compute a union over intersection statistic.

Most of the false positives observed in the DLCDE 2022 data set fell into three categories: spectral lines, rhythmic signals, and burst pulses. Spectral lines are recording artifacts where the energy level is stronger in a narrow band, and some are incorrectly detected as whistles. The graph search algorithm is designed to bridge small gaps in energy due to missed peaks; when there are series of short duration narrow-band energy, these signals can be incorrectly recognized as whistles. Burst pulses, rapid trains of echolocation clicks, can produce banding artifacts in a spectrogram (see Watkins, 1967, for a discussion of pulse trains and their impact on spectrograms), and it is not uncommon to detect portions of these as whistles.

IV. DISCUSSION

Deep learning outperforms other methods in a wide variety of contexts (LeCun *et al.*, 2015), and deep learning is being applied in many bioacoustics projects (see Stowell, 2022). Integration of *deep whistle* algorithm of Li *et al.* (2020) into *silbido profundo* provides access to an algorithm that substantially outperforms the graph search algorithm with heuristic peak identification (Roch *et al.*, 2011). *Silbido profundo* yields a stronger F-1 score than the SMC-PHD detector (Gruden and White, 2020), and detects greater portions of whistles (improved coverage metric). Application of the SMC-PHD filter on short whistles (<150 ms) that *silbido profundo* discards retrieves many more whistles, but with reduced precision.

The power of the neural network can likely be attributed to the convolutional kernel's ability to consider multiple peaks in context to their surroundings. Most of the methods discussed in the introduction detect whistle energy by searching for peaks within the spectrum of a single spectrogram frame. Consequently, these algorithms are sensitive to false positives created by transient signals or noise that may not be characteristic of narrow-band frequency modulated signals such as whistles. The *deep whistle* convolutional neural network has a receptive field of 56 ms × 3.125 kHz (Li *et al.*, 2020). As a result, the network can learn contexts that are relevant to predicting when an individual time × frequency node is attributable to whistle energy. Replacing the peak selection algorithm of the graph-search algorithm of Roch *et al.* (2011) with a deep neural network that provides more reliable peak selection offers large performance gains and is likely to provide benefits to other tonal extraction algorithms that replace their heuristic peak selection with *deep whistle* confidence maps.

Across the large DCLDE 2022 data set, *silbido profundo* and the NOAA annotations were, for the most part, in agreement. Detections by *silbido profundo* aligned closely with the temporal bounds of the analyst-specified encounter records and were processed in less than a fifth of the recording time. Differences between analyst annotations and periods of time in which *silbido profundo* detected whistles are attributable to multiple causes. In addition to whistle detections, the NOAA analysts reported visual and echolocation

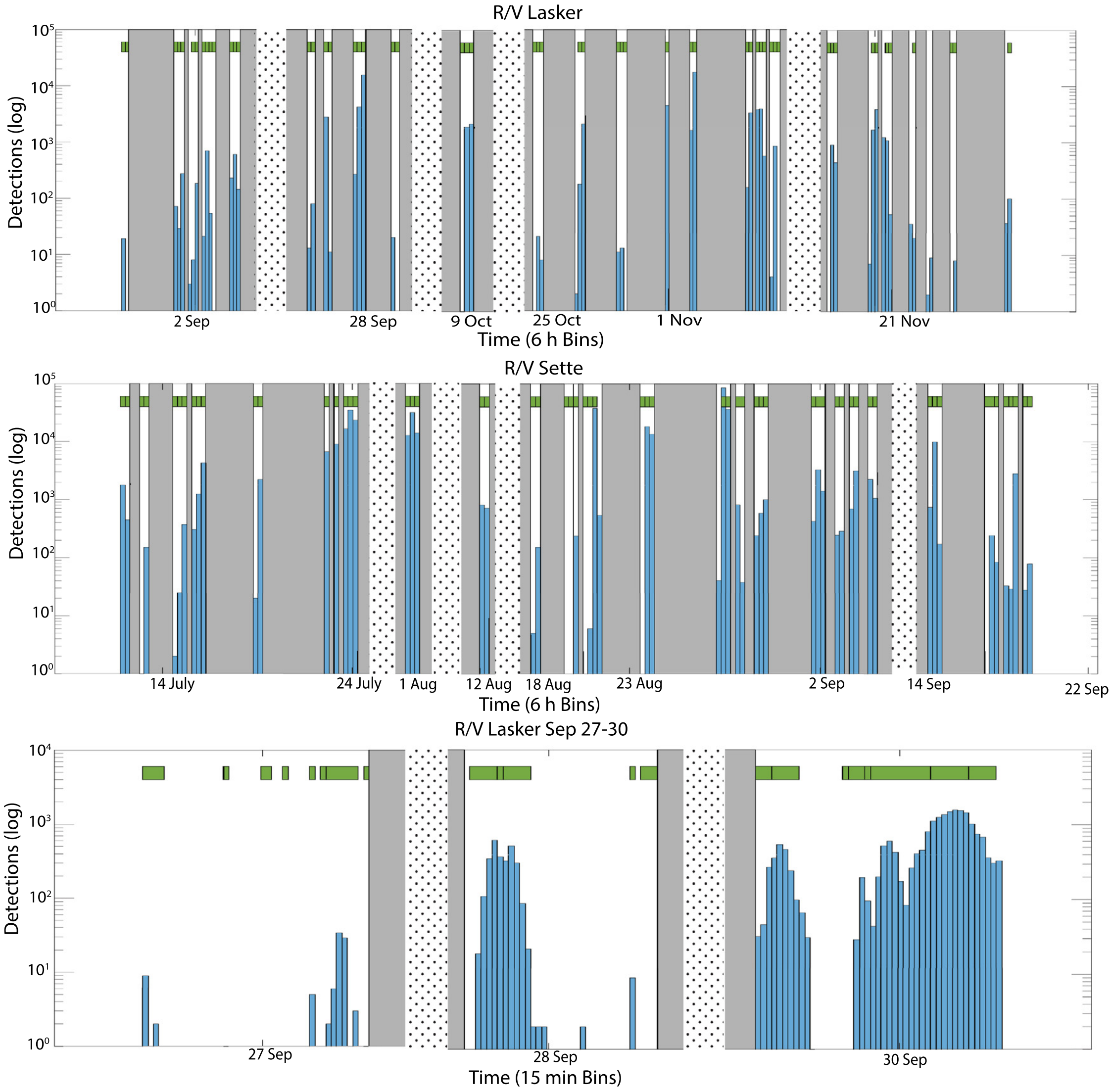

FIG. 5. (Color online) Representative counts of whistles detected over time by application of *silbido profundo* to large data sets. Rectangles at top of the figure show periods where analysts reported visual sightings or acoustic detections (whistles or echolocation clicks). Periods of no recording are represented by gray shading with a dotted pattern used to denote extended gaps between recordings. Upper and middle panels: Data from the DCLDE 2022 R/V Lasker and R/V Sette cruises with detection counts binned into 6 h periods. Bottom panel: A detailed subset of data from the R/V Lasker cruise (15 min bins).

detections, neither of which can be reported by tonal contour extractors. *Silbido profundo* had over 96% precision, but this means that about 4 in 100 detections were false positives, and the false positives occasionally occurred in regions without whistle activity. With the exception of burst pulses labeled as whistles, the false positives were rarely more frequent than 10 detections within 15 min. In some instances, *silbido profundo* made valid detections 5 to 10 min outside of the analyst-reported times (Fig. 5, bottom panel, September 30).

The application of *silbido profundo* has potential to detect a range of mammal vocalizations beyond the odontocete whistle. The neural network was trained to recognize frequency-modulated shapes at a specific temporal and spectral resolution. Although the neural network would likely need to be retrained to obtain optimal results, application to other narrow-band tonal signals, such as mysticete moans and tonal bird calls, shows potential. Figure 6 shows the results for a Speckled Warbler (*Pyrrholaemus sagittatus*) song. The neural network was not retrained to account for differences in frequency range or spectral resolution. Application to mysticete moans had similar results (not shown).

V. CONCLUSIONS

We have developed an open-source implementation of a recently proposed whistle extraction system that is easy for bioacousticians to use. This implementation increases access to deep learning technology for non-computer scientists. We tested our methods on a fully annotated data set of 1025 whistles and a larger, weakly annotated data set, and our results indicated substantial improvements to an existing

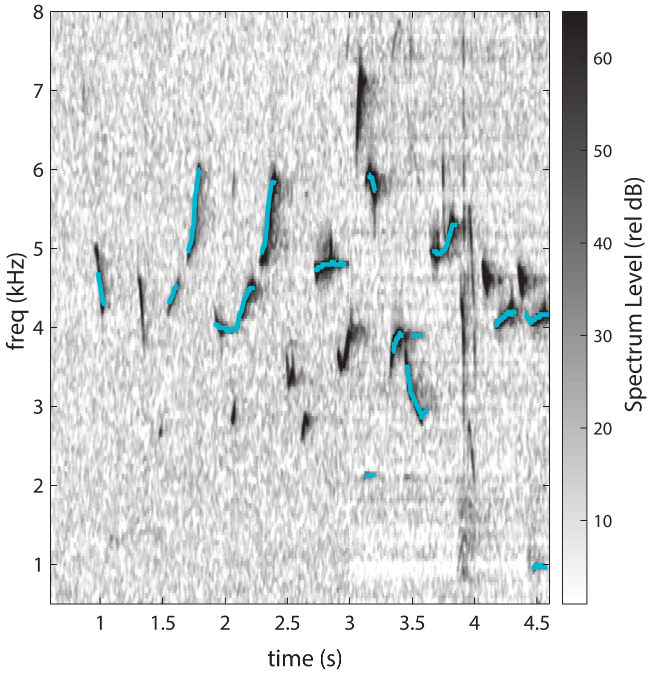

FIG. 6. (Color online) Application of *silbido profundo* to another species' tonal calls without retraining the neural network. A Speckled Warbler (*Pyrrholaemus sagittatus*) song (Lambert, 2014) was analyzed by *silbido profundo*. Spectrogram framing parameters were set to a 5 ms advance with a 30 ms length, analysis was restricted to between 0.5 and 8 kHz, and the following graph search parameters were changed from their default parameters defined in Roch *et al*. (2011); active set, 0.02 s, maximum gap between peaks 30 ms, and minimum duration 50 ms. Detections are shown in blue.

framework for extracting toothed whale whistles. The *silbido profundo* software is publicly available at https://github.com/MarineBioAcousticsRC/silbido.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Michael Weise of the Office of Naval Research (N00014-17-1-2867, N00014-17-1-2567) for supporting this project as well as the crews, engineers, technicians, observers, and analysts who are responsible for the publicly available data sets used in this project. We also thank Anu Kumar and Mandy Shoemaker of U.S. Navy Living Marine Resources for supporting development of the data management tools used in this work (N3943020C2202). Finally, we would like to thank the anonymous reviewers for their thoughtful suggestions that have been incorporated into this article.

[1] Test files from the DCLDE 2011 data: Qx-Tt-SCI0608-N1-060814-121518, palmyra092007FS192-070924-205305, palmyra092007FS192-070924-205730, Qx-Dc-CC0411-TAT11-CH2-041114-154040-s, Qx-Dc-SC03-TAT09-060516-171606, and QX-Dc-FLIP0610-VLA-061015-165000.

[2] The following subscription MATLAB toolboxes are used: Deep Learning, Statistics & Machine Learning, and Signal Processing. For manual annotations, the Image Processing toolbox is also required.

Antichi, S., Urban, R. J., Martinez-Aguilar, S., and Viloria-Gomora, L. (**2022**). "Changes in whistle parameters of two common bottlenose dolphin ecotypes as a result of the physical presence of the research vessel," PeerJ **10**, e14074.

Au, W. L., and Hastings, M. C. (**2008**). *Principles of Marine Bioacoustics* (Springer, New York), p. 679.

Bai, J., Lu, G. J., and Zhang, K. (**2019**). "ONNX: Open neural network exchange," https://github.com/onnx/onnx (Last viewed July 20, 2022).

Baumann-Pickering, S., Roch, M. A., Brownell, R. L., Jr., Simonis, A. E., McDonald, M. A., Solsona-Berga, A., Oleson, E. M., Wiggins, S. M., and Hildebrand, J. A. (**2014**). "Spatio-temporal patterns of beaked whale echolocation signals in the North Pacific," PLoS One **9**(1), e86072.

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.

Bittner, R. M., McFee, B., Salamon, J., Li, P., and Bello, J. P. (**2017**). "Deep salience representations for F0 estimation in polyphonic music," in *Proceedings of the International Society for Music Information and Retrieval Conference*, October 23–27, Suzhou, China, pp. 63–70.

Bonato, M., Papale, E., Pingitore, G., Ricca, S., Attoumane, A., Ouledi, A., and Giacoma, C. (**2015**). "Whistle characteristics of the spinner dolphin population in the Comoros Archipelago," J. Acoust. Soc. Am. **138**(5), 3262–3271.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *Tursiops truncatus* signature whistles," J. Acoust. Soc. Am. **94**(5), 2497–2506.

Caldwell, M. C., and Caldwell, D. K. (**1971**). "Statistical evidence for indvidual signature whistles in Pacific whitesided dolphins, *Lagenorhyncus obliquidens*," Cetology **3**(9), 1–9.

Charbonnier, P., Blanc-Feraud, L., Aubert, G., and Barlaud, M. (**1994**). "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proceedings of the International Conference on Image Processing*, November 13–16, Austin, TX, pp. 168–172.

DCLDE Organizing Committee (**2011**). "Detection, classification, localization, and density estimation (DCLDE) of marine mammals using passive acoustic monitoring workshop dataset," http://mobysound.org (Last viewed November 1, 2019).

Deecke, V. B., and Janik, V. M. (**2006**). "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," J. Acoust. Soc. Am. **119**(1), 645–653.

Frasier, K. E. (**2021**). "A machine learning pipeline for classification of cetacean echolocation clicks in large underwater acoustic datasets," PLoS Comput. Biol. **17**(12), e1009613.

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.

Gillespie, D., Gordon, J., McHugh, R., McLaren, D., Mellinger, D. K., Redmond, P., Thode, A., Trinder, P., and Deng, X.-Y. (**2008**). "PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans," in *Proceedings of the Institute on Acoustics*, October 14–15, Southampton, UK, pp. 54–62.

Gridley, T., Cockcroft, V. G., Hawkins, E. R., Blewitt, M. L., Morisaka, T., and Janik, V. M. (**2014**). "Signature whistles in free-ranging populations of Indo-Pacific bottlenose dolphins, *Tursiops aduncus*," Mar. Mam. Sci. **30**(2), 512–527.

Gruden, P., and White, P. R. (**2016**). "Automated tracking of dolphin whistles using Gaussian mixture probability hypothesis density filters," J. Acoust. Soc. Am. **140**(3), 1981–1991.

Gruden, P., and White, P. R. (**2020**). "Automated extraction of dolphin whistles—A sequential Monte Carlo probability hypothesis density approach," J. Acoust. Soc. Am. **148**(5), 3014–3026.

Han, K., and Wang, D. L. (**2014**). "Neural network based pitch tracking in very noisy speech," IEEE/ACM Trans. Audio. Speech. Lang. Process. **22**(12), 2158–2168.

He, K., Zhang, X., Ren, S., and Sun, J. (**2015**). "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, December 7–13, Santiago, Chile, pp. 1026–1034.

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *Proceedings of the IEEE CVPR Conference*, June 27–30, Las Vegas, NV, pp. 770–778.

Heiler, J., Elwen, S. H., Kriesell, H. J., and Gridley, T. (**2016**). "Changes in bottlenose dolphin whistle parameters related to vessel presence, surface behaviour and group composition," Animal Behav. **117**, 167–177.

Helble, T. A., Ierley, G. R., D'Spain, G. L., and Martin, S. W. (**2015**). "Automated acoustic localization and call association for vocalizing humpback whales on the Navy's Pacific Missile Range Facility," J. Acoust. Soc. Am. **137**(1), 11–21.

Hoye, T. T., Arje, J., Bjerge, K., Hansen, O. L. P., Iosifidis, A., Leese, F., Mann, H. M. R., Meissner, K., Melvad, C., and Raitoharju, J. (**2021**). "Deep learning and computer vision will transform entomology," Proc. Natl. Acad. Sci. U.S.A. **118**(2), e2002545117.

Ioana, C., Gervaise, C., Stéphan, Y., and Mars, J. I. (**2010**). "Analysis of underwater mammal vocalizations using time-frequency-phase tracker," Appl. Acous **71**(11), 1070–1080.

Ioffe, S., and Szegedy, C. (**2015**). "Batch normalization: Acclerating deep network training by reducing covariate shirt," in *Proceedings of the Machine Learning Research Conference*, July 7–9, Lille, France, pp. 448–456.

Janik, V. M., and Sayigh, L. S. (**2013**). "Communication in bottlenose dolphins: 50 years of signature whistle research," J. Comp. Physiol. A **199**(6), 479–489.

Kershenbaum, A., and Roch, M. A. (**2013**). "An image processing based paradigm for the extraction of tonal sounds in cetacean communications," J. Acoust. Soc. Am. **134**(6), 4435–4445.

King, S. L., Guarino, E., Donegan, K., McMullen, C., and Jaakkola, K. (**2021**). "Evidence that bottlenose dolphins can communicate with vocal signals to solve a cooperative task," R. Soc. Open Sci. **8**(3), 202073.

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.

Lambert, F. (**2014**). "Speckled warbler (*Pyrrholaemus sagittatus*) recording XC407950," https://xeno-canto.org/407950 (Last viewed November 18, 2022).

LeCun, Y., Bengio, Y., and Hinton, G. (**2015**). "Deep learning," Nature **521**(7553), 436–444.

Li, P., Liu, X., Palmer, K. J., Fleishman, E., Gillespie, D., Nosal, E.-M., Shiu, Y., Klinck, H., Cholewiak, D., Helble, T., and Roch, M. A. (**2020**). "Learning deep models from synthetic data for extracting dolphin whistle contour," in *Proceedings of the International Joint Conference on Neural Networks*, July 19–24, Glasgow, Scotland, p. 10.

Lin, T.-H., Chou, L.-S., Akamatsu, T., Chan, H.-C., and Chen, C.-F. (**2013**). "An automatic detection algorithm for extracting the representative frequency of cetacean tonal sounds," J. Acoust. Soc. Am. **134**(3), 2477–2485.

Mac Aodha, O., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., and Jones, K. E. (**2018**). "Bat detective-Deep learning tools for bat acoustic signal detection," PLoS Comput. Biol. **14**(3), e1005995.

Mallawaarachchi, A., Ong, S. H., Chitre, M., and Taylor, E. (**2008**). "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," J. Acoust. Soc. Am. **124**(2), 1159–1170.

Marques, T. A., Munger, L., Thomas, L., Wiggins, S., and Hildebrand, J. A. (**2011**). "Estimating North Pacific right whale *Eubalaena japonica* density using passive acoustic cue counting," Endang. Species Res. **13**(3), 163–172.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., and Tyack, P. L. (**2013**). "Estimating animal population density using passive acoustics," Biol. Rev. **88**(2), 287–309.

McCordic, J. A., Root-Gutteridge, H., Cusano, D. A., Denes, S. L., and Parks, S. E. (**2016**). "Calls of North Atlantic right whales *Eubalaena glacialis* contain information on individual identity and age class," Endang. Species Res. **30**, 157–169.

Mellinger, D. K., Martin, S. W., Morrissey, R. P., Thomas, L., and Yosco, J. J. (**2011**). "A method for detecting whistles, moans, and other frequency contour sounds," J. Acoust. Soc. Am. **129**(6), 4055–4061.

NOAA Pacific Islands Fisheries Science Center (**2022**). "Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS) towed array data. Edited and annotated for the 9th International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals Using Passive Acoustics (DCLDE 2022)" (NCEI, Washington, DC).

Oikarinen, T., Srinivasan, K., Meisner, O., Hyman, J. B., Parmar, S., Fanucci-Kiss, A., Desimone, R., Landman, R., and Feng, G. (**2019**). "Deep convolutional network for animal sound classification and source attribution using dual audio recordings," J. Acoust. Soc. Am. **145**(2), 654–662.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (**2019**). "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NIPS)*, December 8–14, Vancouver, BC, Canada, pp. 8026–8037.

Roch, M. A., Batchelor, H., Baumann-Pickering, S., Berchock, 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 Info **31**, 122–136.

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**(4), 2212–2223.

Serra, O. M., Martins, F. P. R., and Padovese, L. R. (**2020**). "Active contour-based detection of estuarine dolphin whistles in spectrogram images," Ecol. Info. **55**, 101036.

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., Rice, A., Chou, E., Hildebrand, J. A., and Roch, M. A. (**2015**). "Seven years of blue and fin whale call abundance in Southern California," Endang. Species Res. **28**, 61–75.

Stowell, D. (**2022**). "Computational bioacoustics with deep learning: A review and roadmap," PeerJ **10**, e13152.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., Glotin, H., and Orme, D. (**2019**). "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," Methods Ecol. Evol. **10**(3), 368–380.

Thomas, M., Martin, B., Kowarski, K., Gaudet, B., and Matwin, S. (**2019**). "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, September 13–17, Würzburg, Germany, p. 16.

Van Cise, A. M., Roch, M. A., Baird, R. W., Aran Mooney, T., and Barlow, J. (**2017**). "Acoustic differentiation of Shiho- and Naisa-type short-finned pilot whales in the Pacific Ocean," J. Acoust. Soc. Am. **141**(2), 737–748.

Van Parijs, S. M., Clark, C. W., Sousa-Lima, R. S., Parks, S. E., Rankin, S., Risch, D., and Van Opzeeland, I. C. (**2009**). "Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales," Mar. Ecol. Prog. Ser. **395**, 21–36.

Watkins, W. A. (**1967**). "The harmonic interval: Fact or artifact in spectral analysis of pulse trains," in *Symp. on Marine Bio-Acoustics*, edited by W. N. Tavolga (Pergamon Press, New York), pp. 15–43.

White, P. R., and Hadley, M. L. (**2008**). "Introduction to particle filters for tracking applications in the passive acoustic monitoring of cetaceans," Can. Acoust. **36**(1), 146–152.

Yano, K. M., Oleson, E. M., Keating, J. L., Ballance, L. T., Hill, M. C., Bradford, A. L., Allen, A. N., Joyce, T. W., Moore, J. E., and Henry, A. (**2018**). "Cetacean and seabird data collected during the Hawaiian islands cetacean and ecosystem assessment survey (HICEAS), July–December 2017," NMFS-PIFSC-72 (National Oceanic and Atmospheric Administration, Wahsington, DC), p. 100.