NOAA Technical Memorandum NMFS-AFSC-412

INSTINCT: Infrastructure for Noise and Soundscape Tolerant Investigation of Nonspecific Call Types

D. Woodrich and C. Berchok

October 2020

U.S. DEPARTMENT OF COMMERCE

National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Alaska Fisheries Science Center

The National Marine Fisheries Service's Alaska Fisheries Science Center uses the NOAA Technical Memorandum series to issue informal scientific and technical publications when complete formal review and editorial processing are not appropriate or feasible. Documents within this series reflect sound professional work and may be referenced in the formal scientific and technical literature.

The NMFS-AFSC Technical Memorandum series of the Alaska Fisheries Science Center continues the NMFS-F/NWC series established in 1970 by the Northwest Fisheries Center. The NMFS-NWFSC series is currently used by the Northwest Fisheries Science Center.

This document should be cited as follows:

Woodrich, D., and C. Berchok. 2020. INSTINCT: Infrastructure for Noise and Soundscape Tolerant Investigation of Nonspecific Call Types. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-412, 58 p.

This document is available online at:

Document available: https://repository.library.noaa.gov/welcome

INSTINCT: Infrastructure for Noise and Soundscape Tolerant Investigation of Nonspecific Call Types

D. Woodrich[1] and C. Berchok[2]

[1]Cooperative Institute for Climate, Ocean and Ecosystem Studies (CICOES)
University of Washington
3737 Brooklyn Ave NE
Seattle WA 98195

[2]Marine Mammal Laboratory
Alaska Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
7600 Sand Point Way NE
Seattle, WA 98115

U.S. DEPARTMENT OF COMMERCE

National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Alaska Fisheries Science Center

NOAA Techncial Memorandum NOAA-TM-AFSC-412

October 2020

Executive Summary

The Cetacean Assessment and Ecology Program (CAEP) passive acoustics group of the Alaska Fisheries Science Center's (AFSC) Marine Mammal Laboratory (MML) analyzes passive acoustic data for the presence of cetaceans in Alaska waters. This analysis has historically been conducted manually, in large part due to the complexity of the calling repertoire of many cetacean species and tendency of high overlap of call types. Additionally, existing autodetection approaches have also performed poorly on simpler problems in our data due to high prevalence of recorder self-noise. Manual analysis protocol allows for the determination of binary species presence in several minute bins, allowing for efficient multi-species analysis at the expense of detail on call patterning and density. Extraction of individual calls requires additional fine-scale and labor-intensive processing and analysis. The custom software INSTINCT (Infrastructure for Noise and Soundscape Tolerant Investigation of Nonspecific Call Types) was developed in-house as a tool to extract individual calls of interest quickly and efficiently. It has been further developed to serve as a call type detector which can be applied to large datasets for a variety of applications.

INSTINCT is a generalized detection and classification system that has the novel property among mature software to train a classifier not only from positive exemplars of the signals of interest, but additionally from undefined negative exemplars specific to the system. Domain-specific differences in detectors generally restrict out of the box effectiveness: to address this, INSTINCT provides the tools to easily train a custom classifier using time-frequency boxes from real data containing exemplars and importantly, co-occurring noise events that resemble signals from the domain. As a result the domain is defined from the data provided, and wider generalization is possible with increased diversity of training data. This approach allows for transferability to different noise regimes and soundscapes to allow for successful deployment in dynamic environments.

INSTINCT detectors have introduced a new possibility for the MML to analyze cetacean acoustic behavior on call level data. This technical report demonstrates the use of INSTINCT in

developing detectors for five different calls. Two of these calls were deployed to assess fin whale occurrence and call density on a large scale. INSTINCT detectors were created for 1) North Pacific right whale (*Eubalaena japonica*) upcalls (code: RW); 2) North Pacific right whale gunshots (GS); 3) minke whale (*Balaenoptera acutorostrata*) boing calls (BN); 4) fin whale (*Balaenoptera physalus*) C calls (FN); and (5) backbeat calls (BB). The average precision (a statistic for total performance) for each INSTINCT detector presented are as follows: RW (0.82), GS (0.76), BN (0.88), FN (0.66), and BB (0.72). The FD workflow (FN & BB composite) demonstrated the ability to calibrate with results from manual analysis over eight test moorings ($R^2 = 0.98$, $p < 0.01$). More detailed performance metrics, design and tuning considerations, and strategies for implementation are reported for each INSTINCT detector, as well as future directions for continued development and application of the technology on U.S. Arctic data.

Contents

Introduction

The U.S. Arctic and subarctic are home to a wide variety of endangered marine mammal species (Moore et al. 2000, Moore et al. 2002, Clarke et al. 2013, Hanney et al. 2013, Vate Brattström et al. 2019). These migratory animals move with the yearly formation and retreat of sea ice, covering large distances across a remote, harsh environment that make monitoring via visual means difficult and cost-prohibitive. Passive acoustic monitoring (PAM) is a low-cost method used to monitor marine mammal species via long-term data collection of their calling activity (Van Parijs et al. 2009). The Alaska Fisheries Science Center's (AFSC) Marine Mammal Laboratory (MML) Cetacean Assessment and Ecology Program (CAEP) passive acoustics group has maintained a long-term PAM array from the southern Bering Sea through the high Arctic since the late 2000s, generating tens of terabytes of passive acoustic recordings. These data have historically been analyzed manually in three separate frequency bands: low (0-250 Hz), mid (0-800 Hz) and high (0-Nyquist), using the custom sound analysis program SoundChecker (Wright et al. 2018, Vate Brattström et al. 2019). These bands split up monitoring effort based on typical species calling range, which allows an analyst to review for many species at once. Therefore, this bundled effort represents a major speedup over manual analysis of individual species presence. The expected SoundChecker analysis rate for a given mooring deployment was approximately three months of analyst time per year of recordings for all three frequency bands combined on an average mooring deployment.

Manual analysis has been a reliable and successful method for analyzing complex and dynamic acoustic data for multiple species presence (Wright et al. 2018, Vate Brattström et al. 2019). To enable extensive coverage of the data, a binned analysis protocol was designed to establish presence in each frequency band per several minute bin. This low-resolution labeling of bins provides much needed year-round data on the presence and timing of multiple marine mammal species (as well as environmental and anthropogenic noise sources) at once, but is not well-suited for questions that rely on finer details such as call characteristics, patterning, and density.

Automated detection and classification methods were pursued to obtain higher resolution results and to provide a more efficient tool for analysis of individual call types. The primary objectives for this project were to develop workflows that could streamline detection of certain species and reduce the manual boxing effort needed for analysis of calling behavior. An existing detection and classification system, the Low Frequency Detection and Classification System (LFDCS) (Baumgartner and Mussoline 2011), was found to underperform when applied to these arctic datasets (Vate Brattström et al. 2019). Without a viable alternative to the LFDCS at the time, the software INSTINCT was developed in-house to better account for the high degree of noise present on our recorder moorings (including high self-noise from the trawl reinforced mooring construction) deployed in the Arctic. INSTINCT is designed to distinguish signal of interest (SOI) and these common noise events. Many approaches have demonstrated success using a binary classifier with a negative class to define signal from noise events similar to the signal (Mellinger 2004, Gillespie 2004, Esfahanian et al. 2017). Although these methods were promising, they were not made available as mature and generalized software systems. The LFDCS, a commonly used system, is not structured to classify SOI against noise events such as self-noise, instead classifying signal via distance to 'territories' of tonal call types (Baumgartner and Mussoline 2011). Noise events could be included as a separate positive class in the LFDCS and negatives could be selected and added manually as exemplars in rounds through hard negative training (iterative retraining with high confidence false positives). However, this would be time consuming and not well supported by the software. In addition, the development of the LFDCS has trended towards a low-power solution for real time detection of various call types instead of a system optimized to the needs of archival analysis (Baumgartner et al. 2013, Baumgartner et al. 2017). Software systems Raven, PAMGUARD, and ISHMAEL can be used for detection, but lack built in generalized classifiers to discriminate noise events from user defined SOI (Bioacoustics Research Program 2017, Gillespie et al. 2009, Mellinger 2001). PAMGuard has the module Real-time Odontocete Call Classification Algorithm (ROCCA), which uses the PAMGuard whistle and moan detector for automated detection and classification of delphinid whistles. However, ROCCA did not appear to have a retraining

capability for noise events in the classification stage, and classification design appears to be specific to delphinid whistles.

Although INSTINCT uses a generalized classification infrastructure to classify noise from SOI, it is still vulnerable to reduced performance when applied across time, location, and instrument differences. This problem is known as domain adaptation and is ubiquitous in algorithmic detection and classification (Beery et al. 2018). To address this issue, INSTINCT users characterize the domain of their system by providing segments of real data with hand-labeled SOI. INSTINCT then employs a binary random forest classifier to train probabilistic models based on these labels. Random forest classification performs well on these often polymorphic negative classes (Ross and Allen 2014), and also is considered insensitive to hyperparameter tuning (Liaw and Wiener 2002) allowing the classifier infrastructure to be transferable across various call types and their competing noise events. This classifier design has demonstrated good performance in various applications, including to classify polymorphic avian 'night calls' against false positives within single location-years (Ross and Allen 2014) and to classify minke whale pulse train positive detections against false positives in Stellwagen Bank (Popescu et al. 2013). INSTINCT has similarities to the two-stage technique and image analysis protocol of Esfahanian et al. (2017), which incorporates an initial computationally light detection round based on spectrogram features prior to the more computationally heavy features extraction.

Additionally, a modification on the traditional learning curve was designed to better illustrate the effect of classifier performance by the additional data segments for two INSTINCT detectors. As the data are not centrally located, pure random sampling was not feasible, so the common approach of increasing the size of the training set to track improvement in performance was not practical in our sampling design. A learning curve was designed to simulate the average effect on performance of adding an additional random data segment at each step. This approach used data segments as the smallest sampled unit, and due to high variance in classifier performance, the sampling was repeated over many trials to produce an average effect. Results were interpreted graphically by observing for stability in performance mean and variance.

Presented here is the successful implementation of INSTINCT, which has enabled detailed study of call behavior on small and large scales and unlocked massive new potential for analysis in the U.S. Arctic region, and potentially worldwide. Performance metrics, analysis considerations, and case studies from five cetacean call type detectors developed with INSTINCT are described. Future potential and considerations for the application of INSTINCT on new and challenging problems in cetacean detection and behavior analysis are discussed.

Methods

Detector Function

INSTINCT detectors work in an efficient two-stage process: an initial, computationally light detection stage limits the breadth of a second more computationally heavy classifier stage. Together, these stages allow the overall detector to be both computationally efficient while achieving high performance. The workflow to create a detector with INSTINCT can be grouped into three main phases: Detector design, training, and deployment. Detector design is a manual process (Fig. 1). First, ground truth data for later training is collected from sections of the data known to contain the call type, and hand annotated for positive SOI present in the ground truth. This sampling of data in chunks, as opposed more common random sampling, is necessary as the data are not already annotated on the call type resolution and data must be reviewed consecutively to ensure accuracy of annotations. These sections of ground truth data are here referred to as 'data segments'. Ground truth data should be selected by an analyst very familiar with the SOI and the system in which it occurs, and the ground truth dataset should be a minimal but comprehensive sampling of the expected variation in call characteristics and acoustic conditions. When available, unique data segments (data segments that are extracted from different mooring deployments) should be incorporated from a wide range of locations and years to account for spatiotemporal variability. For call types where the available data are sparse, non-unique (i.e., taken from different sections of the same mooring deployment) data segments can be chosen to bolster the training data size. These annotated ground truth data segments will be used in the training process, both to automatically assign labels to event detector outputs and produce performance statistics. Next, the event detection

protocol is created: this can be non-event specific if desired by using a single band limiting energy detector (BLED), or tailored specifically to the SOI using several BLEDs (such as the event detection design in Thode et al. 2012) and a custom algorithm to combine BLED 'minidetections' (Appendix section A). These BLEDs are implemented in Raven 1.5, enabled by the R package Rraven (Araya-Salas 2017). Also during this phase, a threshold is defined which indicates the percentage of true positive detections to be retained from the end of the event detection stage through the classifier stage.

The training phase is an automated process which outputs a trained model, detector outputs on the ground truth data, and produces performance statistics (Fig. 2). The event detector created in the design phase is applied to the ground truth data, and labeling of detections is performed based on comparison with the ground truth annotations. These labeled detections then undergo feature extraction, where various measurements are made in the time and frequency domains of the signal. Detections that have NA or infinite values for any measurement are removed from the dataset due to later incompatibility with the machine learning model. Binary random forest-based models are generated using R package 'randomForest' (Liaw and Weiner 2002). These models are generated with different sampling of the data, allowing for a cross-validated probability estimate for each detection (Appendix section B). These models are also retained for use in the deployment phase. The user-defined threshold is applied to the resulting probability estimates, and the remaining detections are retained as the final output. Various performance statistics are generated throughout this process, allowing for informed modification of detector design if performance is deemed inadequate for the deployment needs of the particular study.

The deployment phase is an automated process that can act on novel data. When acoustic data are presented, this phase applies and feeds the outputs of the event detection stage into the classifier stage to produce final detections (Fig. 3). Performance at this stage is best estimated by the statistics produced during training, although the common heterogeneity of acoustic data can make these estimates vary based on the contents of novel data presented.

Detector performance for a given probability threshold is assessed via automatically generated statistics such as precision, recall, multibox %, and overbox %. This type of statistic is here referred to as 'specification dependent' statistics (Table 1). Total detector performance can be evaluated with measurements of performance curves that compare tradeoffs based on a dynamic probability threshold (i.e., Precision-Recall curve (PR curve) and true positive rate-false positive rate curve (ROC curve)). Relevant statistics of performance curves include Area Under the Curve (AUC) score (ROC curve), average precision (PR curve), minimum distance recall, and minimum distance precision. These are 'specification independent' statistics which are not dependent on a probability threshold (Table 1). Recall is a measure of how effective the detector is at pulling out all instances of desired signal. Precision is a measure of the accuracy of returned detections. Recall and precision are calculated for the defined event detector and classifier stages, and for both stages combined (overall). Note that the binary classification operation functions as a part of the larger detection method, so performance statistics referring to the classifier stage refer to the effect of the binary classifier on the pool of detections returned from the defined event detection stage. Minimum distance recall and precision refer to the point along the PR curve lying closest to the perfect detector (recall = 1, precision = 1) and approximate an optimal threshold for maximum detector performance ignoring the application considerations of thresholding differences (Habibzadeh et al. 2016). Note that optimality here is used as a descriptor of model behavior and does not suggest appropriate tuning of the model which is ultimately determined by analysis goals and real world implications of misclassification (Perkins and Schisterman 2006). AUC is the value of the integral of the receiver operating characteristic (ROC) curve, and average precision is the value of the integral of the PR curve. These statistics indicate the total performance power of the detector, with better performing detectors approaching a perfect score of one. Multibox % and overbox % are custom statistics (defined in this report) designed to quantify the rate at which multiple returned detections are contained within a single ground truth detection (multibox %) and the rate at which multiple ground truth detections are contained within a single returned detection (overbox %). This allows for an assessment of the behavior of how true positive detections are lining up with the ground truth boxes. Statistics are calculated 'overall', where datasets with

more labels have heavier weights in any calculation, and also 'by-data segment', meaning that statistics are calculated individually for each data segment, and then averaged to equally weight the data segments regardless of size. The 'by-data segment' statistics help indicate variability of classifier performance across the individual data segments. Detector performance can also be evaluated qualitatively by analyzing returned detections, missed detections, and their component minidetections in Raven 1.5 on a per-detection level. Feature importance can also be compared within detectors using relative mean decrease in Gini coefficient (related to the ability of a variable to split mixed labels into pure categories) which allows for qualitative analysis of the classifier function (Ross and Allen 2014). As an aggregate, these statistics and assessment techniques are used to adjust BLED, defined event detector, and classifier parameters, which are compared over many iterations to optimize performance for the given analysis workflow. Given that the number of potential parameter combinations is computationally prohibitive to automated tuning, parameter adjustment is performed manually for a quantity of iterations needed to produce sufficient detector performance for the needs of the analysis (Mellinger 2004).

Compute

INSTINCT uses parallelization at different stages throughout the script to increase Central Processing Unit (CPU) utilization and decrease computation time. Parallelization is critical for time efficient analysis of large acoustic datasets, and when configured efficiently, can represent substantial reduction of processing time (Dugan et al. 2014). The degree of parallelization for INSTINCT is specified by the user, and is dependent on the number of CPU cores and memory availability for the machine. Parallelization can be utilized by INSTINCT in four separate stages: energy detection, defined event detector algorithm application, feature extraction, and model generation. Each of these stages has unique CPU and memory demands. Energy detection is characterized by low CPU usage and high memory demand, so can be efficiently parallelized with a high memory machine. Defined event detector algorithm application uses CPU efficiently, so this stage benefits from increased CPU capability, but also requires high memory due to the large datasets being processed. Feature extraction uses CPU

efficiently, and does not require high memory, so machines with high CPU capability can process this efficiently. Model generation is memory limited, so high memory machines will perform efficiently on this stage.

Given the differences in CPU or memory limitation for different parallel processes in INSTINCT (Fig. 4), computers which are low in CPU or memory will bottleneck at different stages. That said, the CPU limited stage of feature extraction is generally the bottleneck for computation time, meaning that higher CPU machines will have faster processing with INSTINCT. Also of importance is that large datasets can overwhelm the memory of smaller machines during defined event detection algorithm application, so chunking datasets into smaller components and running them individually may be necessary on low memory machines. Best performance will be achieved on high CPU and high memory capability machines.

Detectors

The flexibility of INSTINCT was demonstrated in analyses of several call types. As each analysis has different considerations depending on the purpose of the investigation (similar to Davis et al. 2017), call type frequency and behavior, species rarity, and prevalence of competing noise events against the call type, detector design and parameters were adjusted for the most appropriate optimization. For example, behavior analysis of a rare call type would necessitate extracting the maximum amount of calls at the expense of lower detector precision and more returned false positives. This would require the detector tuning for high recall and low precision and a stringent verification protocol to eliminate false positive detections. This is necessary, as the inclusion of false positives for a rare call type even in small numbers could misinform the findings of the analysis. This type of analysis is defined as an analysis tuned towards high recall. High recall analysis is particularly useful in situations where maximizing the number of calls returned is preferable, even at the expense of having to sift through a greater number of false positive detections. Alternatively, an investigation of a frequently produced call type may allow for lower recall and still capture species presence reliably at a daily time scale. In these situations, higher precision is generally required to reduce time spent on individual call

verification due to the very high quantity of returned detections. This is defined as an analysis tuned to high precision. Relationships between call detection rates and analyst determined presence in a high precision analysis can be modeled to predict presence based on count data (Baumgartner et al. 2019, Davis et al. 2020). Detectors can additionally be optimized somewhere between high recall and high precision to best balance the needs of the analysis.

We present the application of INSTINCT in the design of five call type detectors: North Pacific right whale (*Eubalaena japonica*; hereafter NPRW) upcalls (RW), NPRW gunshots (GS), minke whale (*Balaenoptera acutorostrata*) boing calls (BN), fin whale (*B. physalus*) C (FN) and backbeat (BB) calls combined into a larger detection workflow (FD). We demonstrate the adaptability of these INSTINCT detectors to analysis goals with a wide range of needs for design and specification. A case study for the FD data analysis workflow is presented to demonstrate the application of INSTINCT to large-scale detection challenges.

North Pacific right whale (RW and GS)

The eastern population of the NPRW is critically endangered with poorly understood migratory patterns (Wright et al. 2018). Due to logistical and funding constraints restricting vessel-based surveys, PAM has been effectively applied to collect data on their year-long presence in the Bering Sea. NPRW primarily make two stereotyped calls: a tonal frequency-modulated (FM) upsweep known as the 'upcall', and a high energy impulsive call known as the 'gunshot' (Crance et al. 2017, McDonald and Moore 2002). Although these call types are indicative of NPRW presence, neither are specific to NPRW as defined by measurable attributes of any individual call. Upcalls are similar to other common tonal upsweeps produced by humpback and bowhead whales. Gunshots are also produced by bowhead whales (Würsig and Clark 1993) and can sometimes resemble other impulsive signals (Crance et al. 2017). These ambiguities create a reliance on contextual information for correct classification of these calls to NPRW, which has been previously addressed by SoundChecker manual analysis protocol for these call types.

Although manual analysis enables the use of contextual information to classify call types in complex situations, the binned presence procedure does not allow for study of call rates and abundance which requires individual call counts. To improve data resolution for finer-scale behavioral analysis of NPRW, INSTINCT detectors for the NPRW upcall (coded RW) and NPRW gunshot (coded GS) were designed to extract calls of these types from manually analyzed data. Only subsetted data which contained the specific call type (as well as sounds unclassified by SoundChecker) were used to design and train each detector, eliminating interspecific interference. Due to the limited sampling of the training data, these detectors were designed with the intended use for call extraction in subsetted data (i.e., not optimized for full mooring analysis). The ground truth data for each of these detectors was composed of data segments from various years and seasons of Bering Sea mooring deployments known to contain exclusively upcalls and exclusively gunshots, respectively (see Results for details). For each call type detector, positive calls in each data segment were reviewed and hand labeled, defined event detectors were designed for initial detection based on the FM characteristics (Appendix section A), and labeled positives and negatives were assembled to train each classifier. The RW defined event detector was composed of 18 BLEDs with frequency bandwidths of 20Hz, overlapping by 10Hz, and covering a total frequency range of 60Hz to 250Hz. Each BLED shared identical parameters aside from frequency. The GS defined event detector was composed of 21 BLEDs with variable frequency bandwidth and overlap covering a total frequency range of 65Hz to 850Hz. BLED parameters were variable over the frequency range since propagation and noise effects were frequency dependent.

Minke boing calls (BN)

A minke whale boing INSTINCT detector (coded BN) was developed to automate the pulse repetition rate measurement to help determine the population identity of minke whales in the Bering Strait and Chukchi Sea (Delarue et al. 2012). Data that were manually labeled with SoundChecker to contain only boing calls and unclassified noise were concatenated into ground truth data segments (see Results for details). A defined event detector was designed to extract these calls based on the presence of harmonics creating ‘stacks’ of detections over a frequency

range (Appendix section A). The BN defined event detector was composed of 14 BLEDs with bandwidths of 100 Hz, overlapping by 50 Hz, and covering a total frequency range of 1,000 Hz to 1,750 Hz. Each BLED shared identical parameters aside from frequency. Given the original purpose of generating pulse repetition rate measurements, the detector was optimized towards high precision.

Fin whale (FD)

Two INSTINCT detectors (coded FN and BB) were developed to identify the two most common call types of Pacific fin whales: 'C' (also known as 40-20 Hz downsweeps, or 'classic') call (FN) and 'backbeats' (also known as 'B' calls) (BB) (Archer et al. 2018). These call types were chosen as the best proxies to match the fin whale class as identified by manual analysis, although other fin whale call types besides these two most common were included in the manual analysis. Data segments were chosen from data manually analyzed in SoundChecker to contain fin whale calls of any type. Positive calls from data segments across years and a broad spatial range were time-frequency boxed to create the ground truth data (see Results for details). Classifier negatives for each detector were composed of miscellaneous abiotic and possible biotic noise not considered to be fin positives. This unfiltered data segment configuration allowed for applicability of both detectors to full datasets. The FN detector featured a defined event detector composed of 14 BLEDs with frequency bandwidths of 4Hz and overlapping by 2Hz, covering a total frequency range of 16Hz to 46Hz. Each BLED shared identical parameters aside from frequency range. The BB detector was a single event detector due variable FM characteristics observed in the call type which is not compatible with defined event detection. Fin whale C calls and backbeats are often very prevalent in the data during periods of animal presence (Archer et al. 2018), so a high precision approach was implemented to take advantage of this repeatability and ensure good recall within larger time bins. This high precision approach was supplemented with a calibration stage to best tune the sensitivity of the detector to match fully manual approaches, with final manual verification of presence peaks reduce correlated false positive detections and confirm bouts of positive detections. The

ensemble of these detectors combined with a manual verification protocol was termed the FD detector workflow for fin whale detection.

The goal of the FD workflow protocol was to create a balanced detector for precision and recall (on the daily % presence of 5-minute bins level), which could be additionally manually verified to further improve confidence in results. A method was developed to best calibrate the performance of the three components of the protocol (FN detector, BB detector, and the manual verification) to manually analyzed fin whale data. The method began with rough optimization to enable a round of data cleaning, which helped to better support a more thorough optimization performed on the cleaned data (with manual verification built into the FD workflow, false positive peaks from noise and airguns are extracted by this verification, so the cleaned data are more representative of the final desired output). The FN detector and BB detector were independently correlated against seven and eight fully analyzed moorings, respectively, at sliding probability thresholds (one test mooring which had notably few C calls was removed from the FN comparison to improve relevance). The sliding threshold ranged from 0.45 to 0.99 with steps of 0.1. The threshold corresponding to the highest average correlation R score across each mooring was used to subset the data for the verification stage. Detections, determined by the probability threshold value, from the FN and BB detectors were grouped into hourly bins, and the local minima and maxima of these hourly bins were calculated to determine 'peaks' of presence. A manual verification stage was used to exclude peaks composed of false positives. Once the data were verified, hourly bins which were positively verified for both C and backbeat calls were extracted from the cleaned data to be used for a more thorough round of calibration against the manually analyzed test moorings. For this round, the parameterization of FN and BB was performed in a simultaneous comparison to best match the manual analysis with their combined binning in an FD class. The probability threshold combinations were compared using mean correlation R scores of the FD class to the manually analyzed moorings on the daily % presence level (5-minute bins/day with fin whale calls detected). The FN and BB thresholds representing the maximum mean correlation R score were used for a final comparison of the test moorings, and R^2 statistic, p-values, and over/under-

sensitivity were calculated for final determination of performance. The protocol was applied to the full extent of analyzed data archived by MML.

Learning Curve

To begin each trial, a randomly selected single ground truth data segment was used to generate a classifier that was applied to another randomly selected (without replacement) single holdout ground truth data segment to assess performance. The classifier performance differential (training performance - test performance) at this first step was registered and the holdout data segment was then incorporated into the ground truth. The model generated from the combined ground truth set was then tested against another random holdout ground truth data segment for the second step. This process was repeated for n (total data segments – 1) steps until all ground truth data segments were included in the training set, and trials were repeated to generate vectors of performance differential at each step. The mean and standard deviation of performance differential were calculated to indicate performance change for the overall process at each step. Due to high computational requirements and limited resources, learning curves were generated for only the RW and GS detectors. A total of 146 trials were conducted for the upcall learning curve and a total of 104 trials were conducted for gunshots.

Results

Detectors

Detector ground truth quantity ranged from 24.6 data hours to 91.4 data hours. Number of ground truth unique data segments ranged from 6 to 12 (Table 2). The five INSTINCT detectors demonstrated variable performance depending on the design of the detector and the analysis specification (Table 3). The scores for each detector are as follows: RW (ROC AUC: 0.96, average precision: 0.82), GS (0.95, 0.76), BN (0.97, 0.88), FN (0.97, 0.66), and BB (0.97, 0.92).

North Pacific right whale upcalls (RW)

The detector ground truth contained six locations throughout the Bering Sea from 5 years of effort (Figs. 5a and b). The RW defined event detector had a recall of 0.94 and a precision of 0.11 (Table 4). The upcall classifier stage had an ROC AUC score of 0.96, and the overall detector had an average precision of 0.82 (Figs. 5c and d). The relatively higher minimum distance precision of 0.81 compared to the minimum distance recall of 0.73 indicates that the detector is more optimized for higher recall analysis. The highest ranking classifier features, quantified by mean decrease in Gini coefficient, suggest that several measurements of slope as determined by application of the Hough lines algorithm (Duda and Hart 1972) (Appendix section C) were the most important in the model (Fig. 5e). Different statistics of the Θ (meanTheta.Hough and sdTheta.Hough) and the slope (MeanSlope.Hough and MedSlope.Hough) of the top scoring Hough lines were the most informative features, followed by the score of the single highest scoring Hough line (BestScore.Hough). With a classifier recall set at 0.95 for a high recall analysis, the overall detector recall was 0.89 at a precision of 0.35. At these levels, an analyst would expect to review ~2 false positives for every 1 true positive in the ground truth data. Multibox % and overbox % were low for the detector, each affecting less than 1 in 300 detections.

North Pacific right whale gunshots (GS)

The detector ground truth contained four locations in the southeastern Bering Sea over 5 years of effort (Figs. 6a and b). The GS defined event detector had a recall of 0.84 and a precision of 0.49 (Table 5). The gunshot classifier stage had an ROC AUC score of 0.95, and the overall detector had an average precision of 0.79 (Figs. 6c and d). The relatively higher minimum distance precision of 0.81 compared to the minimum distance recall of 0.75 indicates that the classifier is more optimized for higher precision analysis. Various types of measurements were informative to the classifier, but Hough features (Appendix section C) were particularly informative (Fig. 6e). The mean Θ (MeanTheta.Hough), median slope (MedSlope.Hough), and median y intercept (MedB.Hough) among top scoring Hough lines were the most informative measurements. Among the others, the standard deviation of summed

pixel values of x-axis quantiles in the binary spectrogram (AreaX.std) and the standard error of the short-term autocorrelation values for the waveform (autoc.se) were most informative. With a classifier recall set at 0.95 for a high recall analysis, the overall detector recall was 0.78 at a precision of 0.73. At these levels, an analyst would expect to review ~1 FP for every 3 TP in the ground truth data. Overbox occurred in ~7% of true detections, indicating these returned detections each overlapped at least one independent ground truth true positive. Only ~1% of detections (multibox) were redundant inside the boundaries of ground truth true positives. The skew towards overbox % indicates the detector is more sensitive to incorrectly grouping tightly associated signals into single detections as opposed to incorrectly splitting more dispersed single signals into multiple detections.

Minke boing calls (BN)

The detector ground truth contained five locations in the Bering and Chukchi seas over 7 years (Figs. 7a and b). The defined event detector had a recall of 0.91 and a precision of 0.42 (Table 6). The boing classifier stage had an ROC AUC score of 0.97 and the overall detector had an average precision of 0.88 (Figs. 7c and d). The relatively higher minimum distance precision of 0.90 compared to the minimum distance recall of 0.83 indicates that the classifier is more optimal for higher precision analysis. Spectrogram features related to shape object pixel area and dimensions were the most informative, corresponding to all five of the most informative features for the classifier (Fig. 7e). Specifically, these features refer to the average frequency range (sFreqrange), number (sCompared), and mean pixel area (sArea) of all considered shape objects, the pixel area of the largest shape object (AreaMax), and the ratio of the largest shape object pixel area to the sum of the other considered shape objects (AreaMax.Dom). With a classifier recall set at 0.95 for a high precision analysis, the overall detector recall was 0.88 at a precision of 0.83. At these levels, an analyst would expect to review ~1 false positive for every four true positives in the ground truth data. Multibox % and overbox % were somewhat high relative to the other detectors, indicating lower returned box quality for this detector.

Fin whale (FD)

Fin whale C calls (FN)

The detector ground truth contained five locations in the Bering and Chukchi seas over 5 years (Figs. 8a and b). The defined event detector had a recall of 0.93 and a precision of 0.13 (Table 7). The fin whale C call classifier stage had a ROC AUC score of 0.97 and the overall detector had an average precision of 0.66 (Figs. 8c and d). The similar minimum distance precision and minimum distance recall indicates that the detector is optimized for high recall or high precision application. The most informative features were various statistics on Hough lines and slope (Appendix section C), indicating the importance of the call frequency modulation in classification (Fig. 8e). The slope (BestSlope.Hough) and Θ (BestTheta.Hough) of the highest scoring Hough line, the median Θ (MeanTheta.Hough) and number (num.Goodlines) of top scoring Hough lines, and the mean slope for all considered shape objects (Meanslope) were the most informative measurements. With a classifier recall set at 0.62 for a high precision analysis, the overall detector recall was 0.50 at a precision of 0.95. Overall detector recall was lower than the expected recall (product of each stage recall) due to many detections being discarded in feature extraction for returning NA or infinite values. At this precision threshold an analyst would review 1 false positives for each 19 true positives present. Multibox % and overbox % were low at ~1.5% each, but the initial BLED detector had a tendency to split up calls into multiple detections at a multibox of 14%.

Fin whale backbeat calls (BB)

The detector ground truth contained six locations in the Bering Sea, Chukchi Sea, and Aleutian passes over 6 years (Figs. 9a and b). The event detector had a recall of 0.80 and a precision of 0.28 (Table 8). The fin whale backbeat classifier stage had a ROC AUC score of 0.97, and the overall detector had an average precision of 0.72 (Figs. 9c and d). The relatively higher minimum distance precision of 0.77 compared to the minimum distance recall of 0.68 indicates that the classifier is more optimized for higher precision analysis. The most informative features were statistics on the uniformity and strength of the backbeat call (Fig. 9e). The mean of pixel area among all considered shape objects (sArea), the mean of the number of shape objects along quantiles along the y-axis (SwitchesY.mean), the maximum pixel count for the

largest shape object (AreaMax), the score of the single highest scoring Hough line (Appendix section C) (BestScore.Hough), and the mean Hough line score among all considered shape objects (sScore) were the most informative measurements. With a classifier recall set at 0.59 for a high precision analysis, the overall detector recall was 0.49 at a precision of 0.97. At this threshold an analyst would review 1 false positive for each 32 true positives present. Multibox % and overbox % were low and balanced at < 1% each, suggesting boxes were encompassing call boundaries appropriately.

FD calibration

Eight test mooring deployments were used in the calibration protocol (2012 M2a, 2012 M2b, 2012 PH1, 2013 PH1, 2012 CL1, 2013 CL1, 2012 M8, and 2018 BS4). For the data cleaning stage, the maximum correlation R score for unreviewed detections on all the test moorings were at a FN threshold of 0.95 and a BB threshold of 0.92 (for this FN comparison, 2012 M2b was removed from the pool due to seasonal lack of this call type, reducing the utility of this comparison). The FN/BB parameter combination with the maximum correlation R score mean on the cleaned data was a FN threshold of 0.91, and a BB threshold of 0.90. Low standard deviation at this parameterization indicated consistent detector behavior among the test moorings, supporting the choice of thresholds (Fig. 10). After another round of peak verification at this calibration, the combined detector outputs had an R^2 of 0.98 ($p < 0.01$) to the test moorings in a combined regression (Table 9). Detector over-sensitivity (higher daily % presence of 5-minute bins value than the manual analysis) was more common at an overall ratio of ~4 over-sensitive days to 3 under-sensitive days, but 2018 BS4 had the highest single mooring sensitivity ratio, at ~71 under-sensitive days to 1 oversensitive day. The 2012 M2a mooring had the highest R^2 value of all test moorings at an R^2 of 1.00 (Fig. 11), the 2012 CL1 mooring had the median R^2 value of 0.97 (Fig. 12), and the 2018 BS4 mooring had the lowest R^2 value of 0.89 (Fig. 13) (Table 9). Daily % presence peaks at PH1 exhibited notable over-sensitivity annually (Fig. 14). The FD workflow was deployed on 216 mooring years of data, resulting in over 8 million BB detections and 7 million FN detections. Over 122,000 detections representing call peaks were manually verified over the corpus.

Learning Curve

The upcall learning curve showed consistent improvement in precision difference with the addition of data segments to the ground truth dataset. No apparent plateau was reached, but the curve showed diminishing returns on the third quartile (lower bound) around n = 6. The gunshot learning curve showed more dramatic improvement in precision difference at n = 1 followed by gradual improvement in third quartile precision difference that appears to plateau around n = 5 (Fig. 15).

Compute

Complete retraining for each call type detector took between one to four hours on a 16 CPU core 128 GB RAM virtual machine. INSTINCT is currently designed to run on windows OS, and requires a licensed copy of Raven 1.5. The FN detector had an average of 40 hours of run time on a full mooring deployment, and the BB detector had an average of 45 hours of run time.

Discussion

Detectors

INSTINCT was successful in enabling the development of a variety of call type detectors, which have allowed the AFSC MML to streamline certain analysis workflows. This technology emerged as the first viable solution to enable call type analyses at large scale in challenging U.S. Arctic datasets. The PR curve skew of the INSTINCT detectors tended to favor high precision analysis, due to a certain number of signals being lost during event detection due to time frequency masking or a low signal to noise ratio. Because of this, INSTINCT particularly excelled at large-scale deployment of detectors at high precision, which allowed for streamlined detection of call types in large datasets. However, the customizable nature of INSTINCT allows for its application to a wide range of analysis specifications. The ability for customization and the streamlined process of data labeling and detector creation has positioned INSTINCT as a powerful and flexible tool for high performance detection of call types in archival datasets.

The individual call outputs of INSTINCT will be valuable for further species or population classification, as the spacing of calls can indicate call patterns. For instance, while a single gunshot could be produced by a bowhead whale or NPRW, the occurrence of this signal in a stereotyped pattern indicates an origin from a NPRW (Crance et al. 2019). Similarly, fin whale notes are ubiquitous between populations, but differences in note interval and patterning can indicate the stock of the calling animal (Širović et al. 2017, Delarue et al. 2009). Developing pattern recognition techniques into mature workflows represents a promising area of future application for INSTINCT.

North Pacific right whale upcalls (RW)

The upcall detector demonstrated a capability for high recall to extract calls on a subset of data where positive presence was already established. This configuration is appropriate for the intended application, as NPRW calling is infrequent and the species is highly endangered. This allowed the detector to be useful for analysis of upcall calling rates and patterning on all previously analyzed data. The precision statistic at the currently specified recall was tolerable for this application considering there are few data hours after subsampling. However when applied to full mooring-year datasets for analysis the high false positive rate would become unmanageable (Table 2). Based on the skew of the precision-recall curve the detector at present would demonstrate better performance for full mooring application at higher precision analysis. Given the rarity of the species and the low calling rate, testing would need to be performed to ensure that bouts are not being missed at lower recall for a full mooring-year deployment.

To further develop the upcall detector for high recall application, it will be necessary to expand the capability of the defined event detector to enable a higher recall ceiling (recall ceiling refers to the number of true positives that are initially identified by the defined event detector before the classifier stage). This would have the trade-off of increasing the amount of false positive detections to process during feature extraction and decreasing computational efficiency. Once a recall ceiling is reached, it will be worth reexamining the recall statistic itself to make sure that it properly qualifies the success of the analysis. One such strategy to better analyze performance may be to adjust definitions of true calls to exclude lower quality calls from the ground truth data which are less important to include than high quality calls, but at

present contribute equally to the recall statistic. This will have the effect of increasing the recall statistic, as INSTINCT will preferentially return high quality calls. Alternatively, instead of throwing out low quality calls, a weighted recall statistic can be used to favor inclusion of higher quality upcalls for a similar effect. Further, it will be beneficial to define a high recall ceiling as a ceiling that is expected between two human analysts instead of an arbitrary (and typically unachievable in any case) target of 1.0. An experiment could be designed to compare call type boxing between two experienced observers, which would give a more realistic target for recall (as in Baumgartner et al. 2011). Recall could then be reported as a % of this attainable target statistic, allowing for more room to improve analysis precision while keeping recall high. While these changes would largely be semantic, they would help direct the analysis to its target goals while keeping the detector function closer to optimal.

Application of this detector to full mooring deployment data would also require expansion of the ground truth data to include biotic noise factors. This version of the detector would need to identify all upsweeps in the NPRW frequency range, not just NPRW upcalls, which is a category that would be expanded to include humpback whale and bowhead whale calls, which are considered to be indistinguishable without context. Context can be valuable to improve the performance of detection and classification tasks (Roch et al. 2018). Future work for NPRW upcall detection could be a multiple stage process, which would involve extracting upsweeps with INSTINCT, and using a convolutional neural network (CNN) or other deep learning method for species classification using contextual features sourced from outside of the call boundaries. This would be an innovative approach that combines INSTINCT for fast detection of call types, and leverages the ability of CNNs to learn complicated patterns in image data.

North Pacific right whale gunshots (GS)

Despite good classifier performance at high precision, the recall ceiling for the gunshot detector was lower as many low quality true positive calls were screened out during the defined event detection step. Defined event detection requires a set of rules to be applied to BLED minidetections to produce a detection, and these rules must encompass the variation in a call type to achieve accurate detection. Gunshot defined event detection is a difficult problem

over the range of variation, as lower quality/more dispersed received gunshots resemble tonal sweeps with a longer duration, while higher quality/less dispersed received gunshots are broadband with a very short duration. To achieve sufficient detector precision for gunshots, the detector recall for very low quality (high dispersion) calls was reduced, favoring the higher quality gunshots. This had the overall effect of lowering the ceiling for a high recall gunshot analysis. However, this approach had the benefit of tailoring the detector to the more important high quality signals and increasing performance on this more relevant class of gunshots.

Additionally, minimizing multibox % was favored over balancing overbox % and multibox %. Gunshots can be tightly coupled in 'doublet gunshot' patterns (Crance et al. 2019) or can have a long received duration due to modal dispersion. These tendencies made it difficult to define gunshot duration, resulting in a large overbox % if the detector was configured to appropriately bound highly dispersed gunshots, or a large multibox % if the detector was configured to appropriately bound doublet gunshot. Although identifying doublet gunshots can be helpful for future behavior analysis, it was determined that the more relevant need in the short term was a detector that could reliably bound single gunshots across a high degree of received duration variability. It is also possible to address high overbox % using another round of classification to separate doublet gunshots from single gunshots should the need to make this distinction arise in the future.

Minke boing calls (BN)

The minke boing detector was unrefined compared to the other detectors as it was designed to enable a calculation of pulse repetition rate from previously extracted minke boing calls. The BN detector had excellent precision/recall performance metrics compared with the other detectors. This success is likely due to the relatively low rate of competing noise present in the ground truth data at the higher frequency band of these calls. The high importance of pixel count features reflects the defining attribute of the harmonic nature of these calls. Instead of occurring as a single pitch modulated feature, the pulsive nature of these calls leads to their appearance on spectrograms as frequency modulated harmonics (Watkins 1968), and so their

identification relies on the association of the shape and area of these components (Appendix section A). The ability of the INSTINCT classifier to identify these characteristics reflects its versatility when applied to call types that cannot be defined by a single frequency modulated shape object. This is in contrast with existing pitch tracking methods (e.g., LFDCS, ROCCA) which would be unable to account for these associations. High variability seen in the by-data-segment performance metrics hints at an imbalance in the ground truth data, which is largely dominated by 2 years at mooring CL1 with particularly high calling activity. This is by necessity, given that the selected data segments were a comprehensive sampling of minke boing calls in the manually analyzed data. However, because of this data imbalance, it is difficult to predict the success of this detector when applied to other regions, years, and seasons. Any future analyzed data that are found to contain minke whale boing calls (via SoundChecker, data exploration with INSTINCT, or data from outside the lab) can be included in training to increase the balance and generalization of the detector ground truth to whatever domain is desired for its application.

Fin whale (FD)

The fin whale detector achieved generalization to call types sufficient to describe seasonal fin whale presence in the Bering and Chukchi seas. Although there was notable undersensitivity in Unimak Pass (BS4), the calibration protocol allowed for a compromise in sensitivity such that seasonal trends are best preserved between all considered locations. This protocol will allow for a rapid semi-automated analysis of the entire MML archival data in the Alaska region, allowing us to build a robust dataset of annual fin whale presence throughout a wide spatial range. These data will inform our knowledge of fin whale spatiotemporal distribution in a region where they have received limited study.

In addition to insights on fin whale acoustic presence in this region, the FD workflow has provided a blueprint for future model creation, calibration, and deployment with INSTINCT. One of the biggest challenges in training and assessing performance in the FN and BB detectors was the lack of robust training data, given that the low frequency band containing fins was among the least analyzed. In future model creation for thoroughly analyzed call types, more

spatial and temporal coverage will be available to better assess these differences while creating the model, which should diminish the need for calibration and manual verification to alleviate sensitivity differences in the model. Time spent training and assessing model performance in different environments is valuable to prevent computationally expensive redeployment down the road if performance is later found to be inadequate.

Learning Curve

Similar to the recall/precision tradeoff inherent to detection methods, there is a human effort versus detector generalization tradeoff. Research into the appropriate amount of data to generalize a detector to the relevant system is essential to navigate this tradeoff (Beery et al. 2018). It is difficult to know in the moment the appropriate amount of data segments that should be included for a new detector given the sometimes high degree of variability in performance between individual data segments. Therefore, exploring the average effect of adding data segments to training data are valuable to better inform this decision in practice.

Both the upcall and gunshot learning curves demonstrated a reduction in the precision differential with the addition of more data segments. The more gradual trajectory of the RW learning curve suggests those data segments generate steadily improving but similar models, while the more dramatic GS learning curve suggests models generated from fewer data segments do not necessarily transfer well when applied to other data segments. However, once a few data segments are included, the model produces consistent results (evidenced by the much reduced lower bound by n = 3), indicating commonalities are found between data segments with apparent high variability in call characteristics. The upcall results indicates that it may be possible to further increase the upcall detector performance by adding data segments until a more consistent plateau is reached, whereas the gunshot detector seems to have reached stable performance, indicating that more data may have been provided than optimal for efficiency in labeling effort. This learning curve may be useful to repeat as a tool in detector construction to determine whether performance has stabilized with ground truth effort or if more data may be beneficial to achieve generalization.

Compute

Processing speed varies with the amount of putative detections to process, making the FD workflow extremely computationally intensive due to the amount of true positives in the data. Additionally, detectors which demonstrate high false positive/hour metrics should also be expected to be computationally intensive. Sound file size was not observed to be a bottleneck on run time based on more limited experiments with higher frequency call types, although the most extensive profiling comes from data that were decimated down to very low frequency range (maximum 64Hz). Attention should be paid to other bottlenecks that may surface as INSTINCT is applied to larger file sizes at an equivalent scale.

Conclusions

INSTINCT has introduced a new possibility for automated detection and classification of marine mammal call types in U.S. Arctic passive acoustic data. In addition to demonstrating excellent performance on challenging datasets, INSTINCT provides a platform to develop and assess new artificial intelligence/machine learning methods for the advancement of detection capability in an increasingly automated scientific landscape. The transferability of INSTINCT provides a powerful out of the box approach for new analyses, and the scaling potential allows for application of analyses to large spatiotemporal ranges. In addition to the ability for INSTINCT to inform animal presence on new or unanalyzed data, it can be applied for behavior analysis of call types on new data or data previously labeled with SoundChecker. INSTINCT has enabled new analysis potential for MML and could become a go-to tool for call type analysis in the animal bioacoustics community.

Acknowledgments

The data used for this work were collected over many years and for numerous research projects. The authors give thanks to those responsible for the collection of the data, namely the captains, ship crews, and science crews of the FV *Aquila*, FV *Mystery Bay*, FV *Alaskan Enterprise*, RV *Ocean Starr*, USCGC *Healy*, and the NOAA ship *Oscar Dyson*. We also thank the following people for their contributions as follows: Phyllis Stabeno (NOAA/OAR/Pacific Marine Environmental Laboratory) for inviting us to attach our passive acoustic recorders to her oceanographic moorings in the Bering Sea. The analysts of the MML; Arial Brewer, Eric Braen, Holly Calahan, Jessica Crance, Stephanie Grassia, Jenna Harlacher, Eliza Ives, Brynn Kimber, Nick Tucker, Megan Wood, and Dana Wright for the essential data labeling for training and calibration of INSTINCT. Dana Wright for her guidance and leadership on NPRW detection. AFSC's OFIS Division (Ben Hou, Mike Brown, Ajith Abraham, other contributors) for the software and hardware support that has enabled scaling of INSTINCT through VM. Dan Morris for advice and perspective on best practice in AI. Internal reviewers Jessica Crance, Dana Wright, and Yuval Boss for helping to refine and improve this document. Funding for data collection and manual SoundChecker analysis was provided by BOEM through the following projects (Interagency Numbers and Program Managers listed in parentheses): North Pacific Right Whale Study (M07RG13267; Cathy Coon/Chuck Monnett), Chukchi Acoustics, Oceanography, and Zooplankton Study and Extension (M09PG00016; Heather Crowley/Chuck Monnett), Arctic Whale Ecology Study (M12PG00021; Jeff Denton/Carol Fairfield/Chuck Monnett) projects. Funding was also provided to Dana Wright for her NPRW analysis work in the Bering Sea by the Marine Mammal Commission, the National Fish and Wildlife Foundation, and the International Fund for Animal Welfare. Funding for more recent (i.e., 2016 on) manual SoundChecker analysis was provided by grants from the Office of Naval Research (Award # N000141812792; Michael Weise) and the NOAA Office of Science and Technology (Mridula Srinivasan/Jason Gedamke).

Citations

Araya-Salas, M. 2017. Rraven: connecting R and Raven bioacoustic software. R package version 1.0.2.

Archer, F. I., S. Rankin, K. Stafford, M. Castellote, and J. Delarue. 2019. Quantifying spatial and temporal variation of North Pacific fin whale (*Balaenoptera physalus*) acoustic behavior. Mar. Mammal Sci. 36(1): 224-245.

Beery, S., G. Van Horn, and P. Perona. 2018. Recognition in terra incognita. 2018 Proceedings of the European Conference on Computer Vision. Munich, Germany. Cornell University, Ithaca, NY 14850. https://arxiv.org/abs/1807.04975v2.

Bioacoustics Research Program. 2017. Raven Pro: Interactive Sound Analysis Software (Version 1.5) [Computer software]. Cornell University, Ithaca, NY 14850: The Cornell Lab of Ornithology. http://www.birds.cornell.edu/raven.

Baumgartner, M., and S. Mussoline. 2011. A generalized baleen whale call detection and classification system. J. Acoust. Soc. Am. 129(5): 2889-2902.

Baumgartner, M., D. Fratantoni, T. Hurst, M. Brown, T. Cole, S. Van Parijs, and M. Johnson. 2013. Real-time reporting of baleen whale passive acoustic detections from ocean gliders. J. Acoust. Soc. Am. 134(3): 1814-1823.

Baumgartner, M., J. Bonnell, S. Van Parijs, P. Corkeron, C. Hotchkin, K. Ball, L. Pelletier, J. Partan, D. Peters, J. Kemp, J. Pietro, K. Newhall, A. Stokes, T. Cole, E. Quintana, and S. Kraus. 2019. Persistent near real-time passive acoustic monitoring for baleen whales from a moored buoy: System description and evaluation. Methods Ecol. Evol. 10(9): 1476-1489.

Clarke, J., K. Stafford, S. Moore, B. Rone, L. Aerts, and J. Crance. 2013. Subarctic cetaceans in the southern Chukchi Sea: Evidence of recovery or response to a changing ecosystem. Oceanography 26(4): 136-149.

Collado, J., C. Hilario, A. de La Escalera, and J. Armingol. 2006. Self-calibration of an on-board stereo-vision system for driver assistance systems. 2006 Institute of Electrical and Electronics Engineers Intelligent Vehicles Symposium 156-162. Tokyo, Japan. doi:10.1109/IVS.2006.1689621.

Crance, J., C. Berchok, and J. Keating. 2017. Gunshot call production by the North Pacific right whale (*Eubalaena japonica*) in the southeastern Bering Sea. Endang. Species Res. 34: 251-267.

Crance, J., C. Berchok, D. Wright, A. Brewer, and D. Woodrich. 2019. Song production by the North Pacific right whale, *Eubalaena japonica*. J. Acoust. Soc. Am. 145(6): 3467-3479.

Davis, G., M. Baumgartner, J. Bonnell, J. Bell, C. Berchok, J. Thornton, S. Brault, G. Buchanan, R. Charif, D. Cholewiak, C. Clark, P. Corkeron, J. Delarue, K. Dudzinski, L. Hatch, J. Hildebrand, L. Hodge, H. Klinck, S. Kraus, B. Martin, D. Mellinger, H. Moors-Murphy, S. Nieukirk, D. Nowacek, S. Parks, A. Read, A. Rice, D. Risch, A. Širović, M. Soldevilla, K. Stafford, J. Stanistreet, E. Summers, S. Todd, A. Warde, and S. Van Parijs. 2017. Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014. Sci. Rep. *7*(1), No. 13460.

Davis, G., M. Baumgartner, P. Corkeron, J. Bell, C. Berchok, J. Bonnell, J. Bort Thornton, S. Brault, G. Buchanan, D. Cholewiak, C. Clark, J. Delarue, L. Hatch, H. Klinck, S. Kraus, B. Martin, D. Mellinger, H. Moors-Murphy, S. Nieukirk, D. Nowacek, S. Parks, D. Parry, N. Pegg, A. Read, A. Rice, D. Risch, A. Scott, M. Soldevilla, K. Stafford, J. Stanistreet, E. Summers, S. Todd, and S. Van Parijs. 2020. Exploring movement patterns and changing distributions of baleen whales in the western North Atlantic using a decade of passive acoustic data. Global Change Biol. 26(9): 4812-4840. https://doi.org/10.1111/gcb.15191.

Delarue, J., S. Todd, S. Van Parijs, and L. Di Iorio. 2009. Geographic variation in Northwest Atlantic fin whale (*Balaenoptera physalus*) song: Implications for stock structure assessment. J. Acoust. Soc. Am. 125(3): 1774-1782.

Delarue, J., B. Martin, and D. Hannay. 2012. Minke whale boing sound detections in the northeastern Chukchi Sea. Mar. Mammal Sci. 29(3): E333-E341.

Duda, R., and P. Hart. 1972. Use of the Hough transformation to detect lines and curves in pictures. Comm. ACM 15(1): 11-15.

Dugan, P., J. Zollweg, M. Popescu, D. Risch, H. Glotin, Y. LeCun, and C. Clark. 2014. High performance computer acoustic data accelerator: a New system for exploring marine mammal acoustics for big data applications. 2014 International Conference on Machine Learning. Beijing, China. Cornell University, Ithaca, NY 14850. https://arxiv.org/abs/1509.03591.

Esfahanian, M., N. Erdol, E. Gerstein, and H. Zhuang. 2017. Two-stage detection of north Atlantic right whale upcalls using local binary patterns and machine learning algorithms. Appl. Acoust. 120: 158-166.

Erbe, C., and A. King. 2008. Automatic detection of marine mammals using information entropy. J. Acoust. Soc. Am. 124(5): 2833-2840.

Gillespie, D. 2004. Detection and classification of right whale calls using an “edge” detector operating on a smoothed spectrogram. Can. Acoust. 32(2): 39-47.

Gillespie, D., D. Mellinger, J. Gordon, D. McLaren, P. Redmond, R. McHugh, P. Trinder, X-Y. Deng, and A. Thode. 2009. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans. J. Acoust. Soc. Am. 125(4): 2547.

Habibzadeh, F., P. Habibzadeh, and M. Yadollahie. 2016. On determining the most appropriate test cut-off value: the case of tests with continuous results. Biochemia Medica 26(3): 297-307.

Hannay, D., J. Delarue, X. Mouy, B. Martin, D. Leary, J. Oswald, and J. Vallarta. 2013. Marine mammal acoustic detections in the northeastern Chukchi Sea, September 2007–July 2011. Cont. Shelf Res. 67: 127-146.

Jahan, R., P. Suman, and D. Singh. 2018. Lane detection using Canny edge detection and Hough transform on Raspberry Pi. Int. J. Adv. Res. Comp. Sci. 9(Special Issue 2): 85-89.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R News 2/3: 18–22.

Mcdonald, M., and S. Moore. 2002. Calls recorded from North Pacific right whales (*Eubalaena japonica*) in the eastern Bering Sea. J. Cetacean Res. Manage. 4(3):261-266.

Mellinger, D. 2001. Ishmael 1.0 user's guide. ISHMAEL: Integrated System for Holistic Multi-channel Acoustic Exploration and Localization. U.S. Dep. Commer., NOAA Tech. Memo. OAR PMEL-120, 26 p.

Mellinger, D. 2004. A Comparison of methods for detecting right whale calls. Can. Acoust. 32(2): 55-65.

Moore, S., D. DeMaster, and P. Dayton. 2000. Cetacean habitat selection in the Alaskan Arctic during summer and autumn. Arctic 53(4): 432-447.

Moore, S., J. Waite, N. Friday, and T. Honkalehto. 2002. Cetacean distribution and relative abundance on the central–eastern and the southeastern Bering Sea shelf with reference to oceanographic domains. Progr. Oceanogr. 55(1-2): 249-261.

Perkins, N., and E. Schisterman. 2006. The Inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am. J. Epidemiol. 163(7): 670-675.

Popescu, M., P. Dugan, M. Pourhomayoun, D. Risch, H. Lewis III, and C. Clark. 2013. Bioacoustical periodic pulse train signal detection and classification using spectrogram intensity binarization and energy projection. 2013 International Conference on Machine Learning workshop on Machine Learning for Bioacoustics. Atlanta, GA. Cornell University, Ithaca, NY 14850. https://arxiv.org/abs/1305.3250.

Roch, M., S. Baumann-Pickering, D. Cholewiak, E. Fleishman, K. Frasier, H. Glotin, T. Helble, J. Hildebrand, H. Klinck, S. Lindeneau, X. Liu, E. M. Nosal, K. Palmer, Y. Shiu, and G. Singh. 2018. The use of context in machine learning for bioacoustics. J Acoust. Soc. Am. 144(3): 1728.

Ross, J., and P. Allen. 2014. Random Forest for improved analysis efficiency in passive acoustic monitoring. Ecol. Inform. 21: 34-39.

Širović, A., E. Oleson, J. Buccowich, A. Rice, and A. Bayless. 2017. Fin whale song variability in southern California and the Gulf of California. Sci. Rep. 7(1): 1-11.

Thode, A., K. Kim, S. Blackwell, C. Greene, C. Nations, T. McDonald, and A. Macrander. 2012. Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys. J. Acoust. Soc. Am. 131(5): 3726-3747.

Van Parijs, S., C. Clark, R. Sousa-Lima, S. Parks, S. Rankin, D. Risch, and I. Van Opzeeland. 2009. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. Mar. Ecol. Progr. Ser. 395: 21-36.

Van der Walt, S., J. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors. 2014. scikit-image: Image processing in Python. https://doi.org/10.7717/peerj.453.

Vate Brattström, L., J. Mocklin, J. Crance, and N. Friday (editors). 2019. Arctic Whale Ecology Study (ARCWEST): Use of the Chukchi Sea by Endangered Baleen and Other Whales (Westward Extension of the BOWFEST). Final Report of the Arctic Whale Ecology Study (ARCWEST), OCS Study BOEM 2018-022. Marine Mammal Laboratory, Alaska Fisheries Science Center, NMFS, NOAA, 7600 Sand Point Way NE, Seattle, WA 98115-6349.

Watkins, W. 1968. The harmonic interval: fact or artifact in spectral analysis of pulse trains. Woods Hole Oceanographic Institution, Reference no. 68-13, p. 15-43.

Wright, D., M. Castellote, C. Berchok, D. Ponirakis, J. Crance, and P. Clapham. 2018. Acoustic detection of North Pacific right whales in a high-traffic Aleutian Pass, 2009-2015. Endang. Species Res. 37: 77-90.

Würsig B., and C. Clark. 1993. The bowhead whale. Soc. Marine Mammal. Special Publication (2): 157−199.

Table 1. -- Definitions for relevant performance statistics. TP = True Positive. FN = False Negative. FP = False Positive. AUC = Area Under Curve. ROC = Receiver Operating Characteristic. PR = Precision-Recall.

Metric	Definition	Specification-dependent
Recall	=TP/(TP+FN) ; hit rate of the detector for all true calls.	Y
Precision	=TP/(TP+FP) ; Accuracy of returned detections	Y
Multibox %	Percent of duplicate returned detections within a single ground truth time frequency box	Y
Overbox %	Percent of duplicate ground truth frequency boxes within a single returned detection	Y
AUC score	Integral of ROC; total detector performance	N
Average precision	Integral of PR curve; total detector performance	N
Minimum distance recall	Recall at the point along the ROC or PR curve corresponding to the minimum distance to a perfect detector	N
Minimum distance precision	Precision at the point along the ROC or PR curve corresponding to the minimum distance to a perfect detector	N

Table 2. -- Ground truth effort for each detector. See text for column definitions.

Detector	Data segments	Unique data segments	Mean data segment effort (hr±SD)	Total effort (hr)	Total calls	Calls/hr
RW	12	12	7.6 ± 4.5	91.4	5793	63.4
GS	11	9	4.6 ± 3.1	50.1	9289	185.3
BN	12	9	2.1 ± 2.8	24.6	7347	299.0
FN	6	6	15.2 ± 5.0	90.9	9962	109.6
BB	9	8	7.1 ± 3.7	63.8	13539	212.3

Table 3. -- Training data characteristics for each detector. Tuning refers to whether the detector was designed for high recall or high precision. Per hour refers to data hour.

Detector	Tuning	Training data type	True detections/hr	Missed detections/hr	False detections/hr
RW	Recall	Subset	56.5	6.8	103.9
GS	Recall	Subset	161.41	23.93	105.25
BN	Precision	Subset	198.7	100.4	8.1
FN	Precision	Raw	55.4	55.2	2.7
BB	Precision	Raw	103.1	109.1	3.0

Table 4. -- Performance statistics of defined event detector, classifier, and combined detector for North Pacific right whale upcalls (RW). Variability indicated by standard deviation. Column definitions in text.

Stage	Recall	Precision	Multibox %	Overbox %	Min. dist. recall	Min. dist. precision
Event Detector	0.94	0.11	0.42	0.07		
Classifier	0.95	0.35			0.90	0.83
Overall	0.89	0.35	0.31	0.07	0.73	0.81
Overall, by data segment	0.89 ± 0.04	0.35 ± 0.08	0.28 ± 0.24	0.06 ± 0.09		

Table 5. -- Performance statistics of defined event detector, classifier, and combined detector for North Pacific right whale gunshots (GS). Variability indicated by standard deviation. Column definitions in text.

Stage	Recall	Precision	Multibox %	Overbox %	Min. dist. recall	Min. dist. precision
Event Detector	0.82	0.45	1.4	7.0		
Classifier	0.95	0.73			0.88	0.89
Overall	0.78	0.73	1.0	7.4	0.75	0.82
Overall, by data segment	0.76 ± 0.16	0.72 ± 0.09	1.0 ± 1.0	5.7 ± 5.4		

Table 6. -- Performance statistics of defined event detector, classifier, and combined detector for minke whale boing calls (BN). Variability indicated by standard deviation. Column definitions in text.

Stage	Recall	Precision	Multibox %	Overbox %	Min. dist. recall	Min. dist. precision
Event Detector	0.91	0.42	5.80	7.05		
Classifier	0.95	0.83			0.92	0.94
Overall	0.88	0.83	4.00	7.05	0.83	0.90
Overall, by data segment	0.92 ± 0.06	0.57 ± 0.27	2.10 ± 2.32	7.25 ± 9.34		

Table 7. -- Performance statistics of defined event detector, classifier, and combined detector for fin whale C calls (FN). Variability indicated by standard deviation. Column definitions in text.

Stage	Recall	Precision	Multibox %	Overbox %	Min. dist. recall	Min. dist. precision
Event Detector	0.93	0.13	14.05	1.62		
Classifier	0.62	0.95			0.91	0.93
Overall	0.50	0.95	1.70	1.41	0.79	0.83
Overall, by data segment	0.46 ± 0.11	0.96 ± 0.03	2.52 ± 3.52	0.65 ± 1.20		

Table 8. -- Performance statistics of defined event detector, classifier, and combined detector for fin whale backbeat calls (BB). Variability indicated by standard deviation. Column definitions in text.

Stage	Recall	Precision	Multibox %	Overbox %	Min. dist. recall	Min. dist. precision
Event Detector	0.80	0.28	4.32	0.88		
Classifier	0.59	0.97			0.91	0.90
Overall	0.49	0.97	0.55	0.67	0.68	0.77
Overall, by data segment	0.44 ± 0.24	0.97 ± 0.02	0.22 ± 0.32	0.53 ± 0.57		

Table 9. -- Calibration statistics for each FD test mooring. All regressions returned p-values of < 0.01.

Deployment	R^2	# days over-sensitive	# days under-sensitive	% days sampled in training data
2012 M2a	1.00	44	54	4.2
2012 M2b	0.95	89	92	<1
2012 PH1	0.96	61	9	0
2013 PH1	0.93	83	21	2.1
2012 CL1	0.97	43	4	3.5
2013 CL1	0.99	25	30	0
2012 M8	0.99	58	92	<1
2018 BS4	0.89	2	143	0
Total	0.98	405	302	1.3

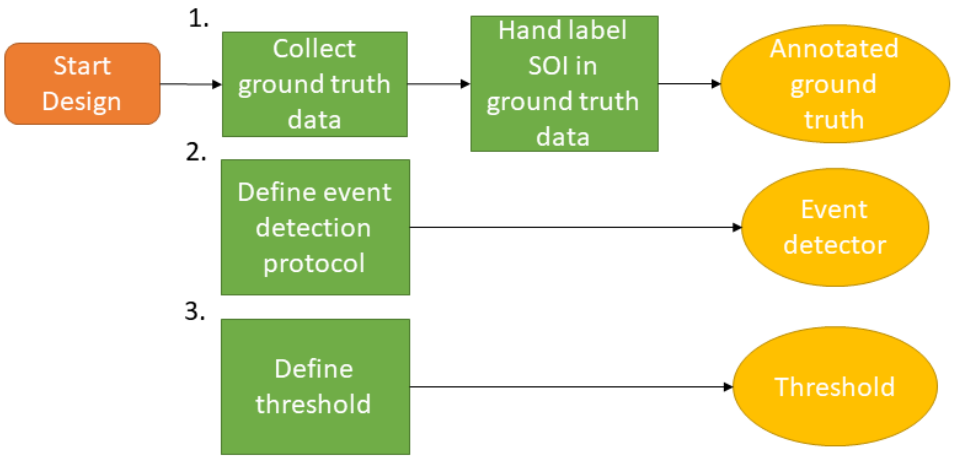


Figure 1. -- Flow chart of INSTINCT design phase. Square boxes indicate a process, and circles indicate a result. Green and yellow colors indicate manual processes and manually obtained modules, respectively. SOI = signal of interest.

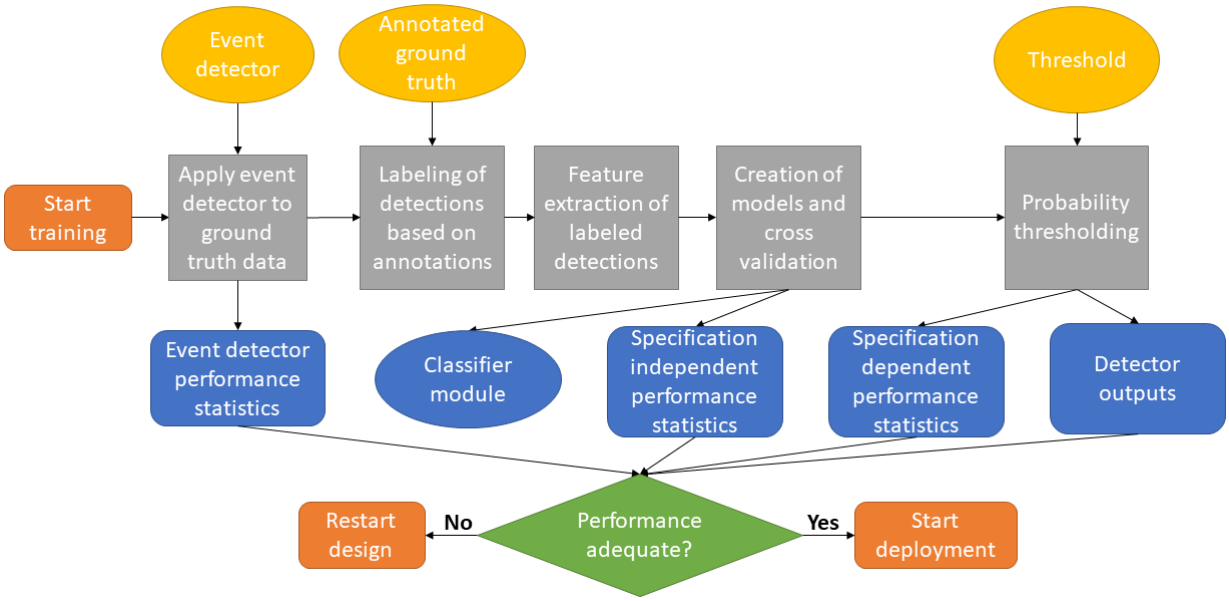


Figure 2. -- Flow chart of INSTINCT training phase. Square boxes indicate a process, circles indicate a module, diamonds represent a decision, and rounded boxes represent an output or initialization. Orange indicates an initialization, gray indicates an automatic process, yellow indicates a manually obtained module, blue indicates an automated result or module, and green indicates a manual decision.

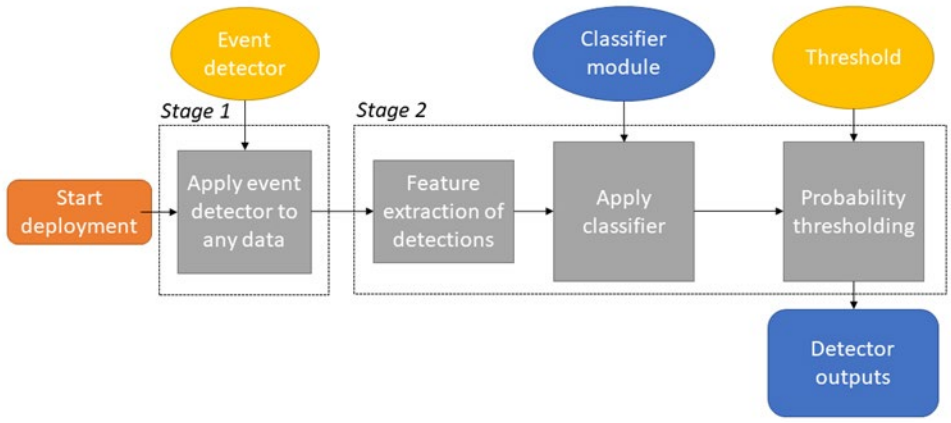


Figure 3. -- Flow chart of INSTINCT deployment phase. Square boxes indicate a process, circles indicate a module, and rounded boxes represent an output or initialization. Orange indicates an initialization, gray indicates an automatic process, yellow indicates a manually obtained module, and blue indicates an automated result or module.

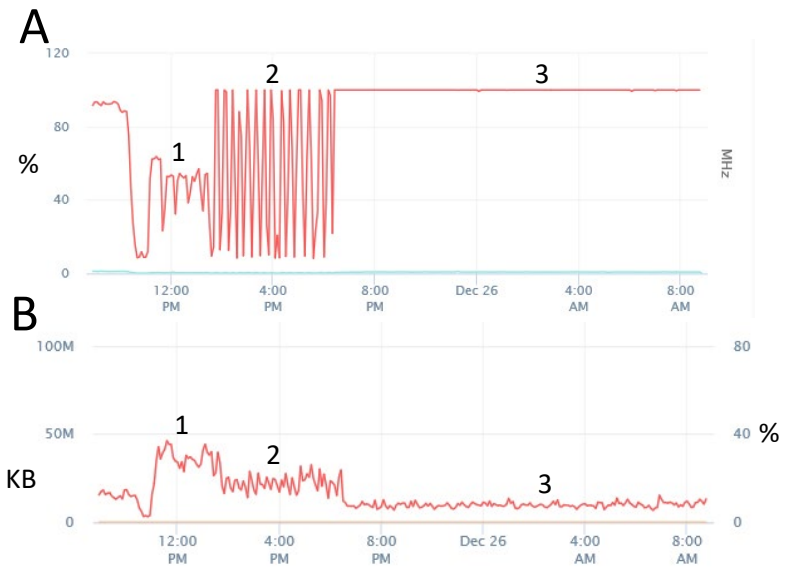


Figure 4. -- (A) Partial Central Processing Unit (CPU) and (B) memory profiling of INSTINCT application to a full year dataset on an Alaska Fisheries Science Center (AFSC) virtual machine (16 cores CPU, 128GB memory). Stage numbers correspond to the first through third parallelized stages of INSTINCT: Energy detection (1), defined event detection algorithm application (2), and feature extraction (3). Not shown is model generation (4). Note the high CPU utilization and low memory utilization, indicating the process would benefit from additional CPU cores (M. Brown, AFSC-OFIS, pers. comm. 26 December 2019).

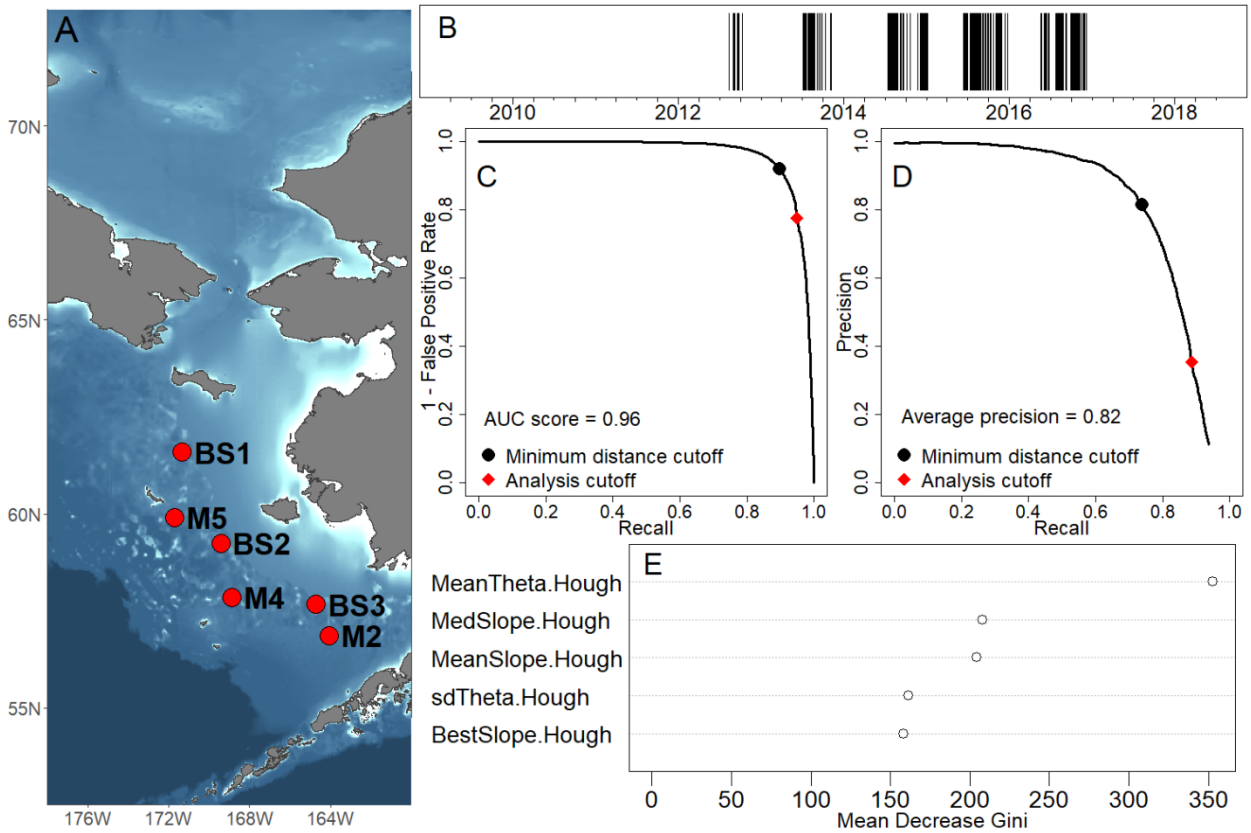


Figure 5. -- Results for North Pacific right whale upcall (RW) detector. (A) Locations where ground truth data segments were sampled. (B) Days in which ground truth data segments were sampled. (C) Receiver operating characteristic curve for upcall classifier. AUC = Area Under the Curve. (D) Precision-recall curve for entire upcall detector. (E) Top five important features for the classifier as determined by mean decrease in Gini coefficient (see text for definitions).

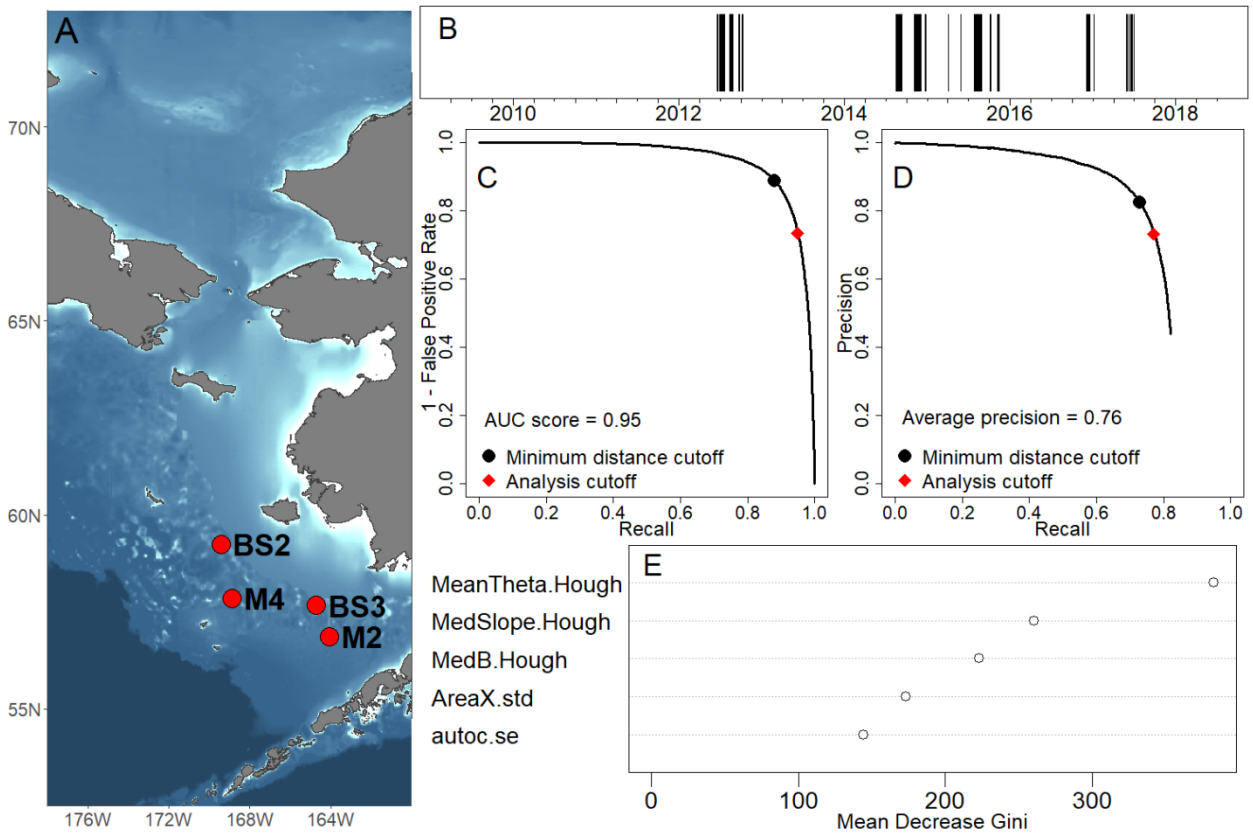


Figure 6. -- Results for North Pacific right whale gunshot (GS) detector. (A) Locations where ground truth data segments were sampled. (B) Days in which ground truth data segments were sampled. (C) Receiver operating characteristic curve for gunshot classifier. AUC = Area Under the Curve. (D) Precision-recall curve for entire gunshot detector. (E) Top five important features for the classifier as determined by mean decrease in Gini coefficient (see text for definitions).

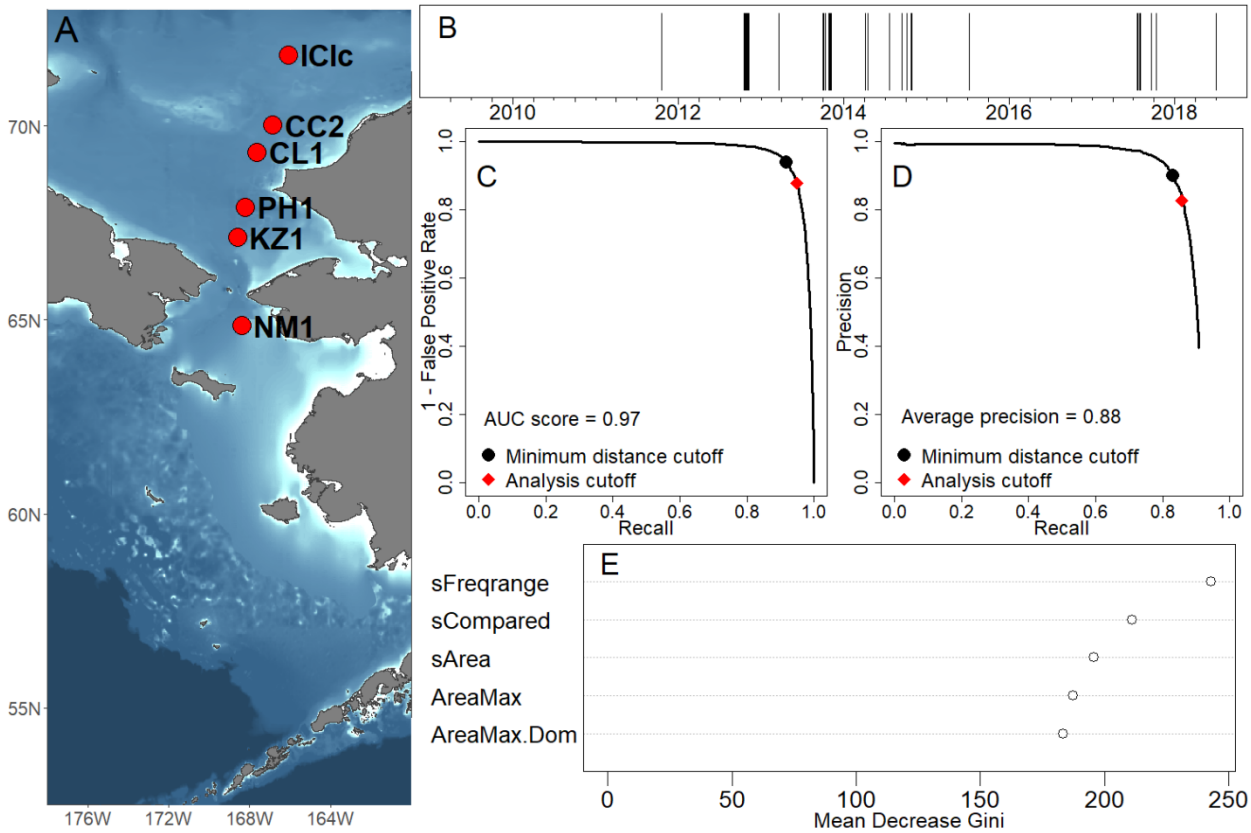


Figure 7. -- Results for minke whale boing call (BN) detector. (A) Locations where ground truth data segments were sampled. (B) Days in which ground truth data segments were sampled. (C) Receiver operating characteristic curve for boing classifier. AUC = Area Under the Curve. (D) Precision-recall curve for entire boing detector. (E) Top five important features in the model, as determined by mean decrease in Gini coefficient (see text for definitions).

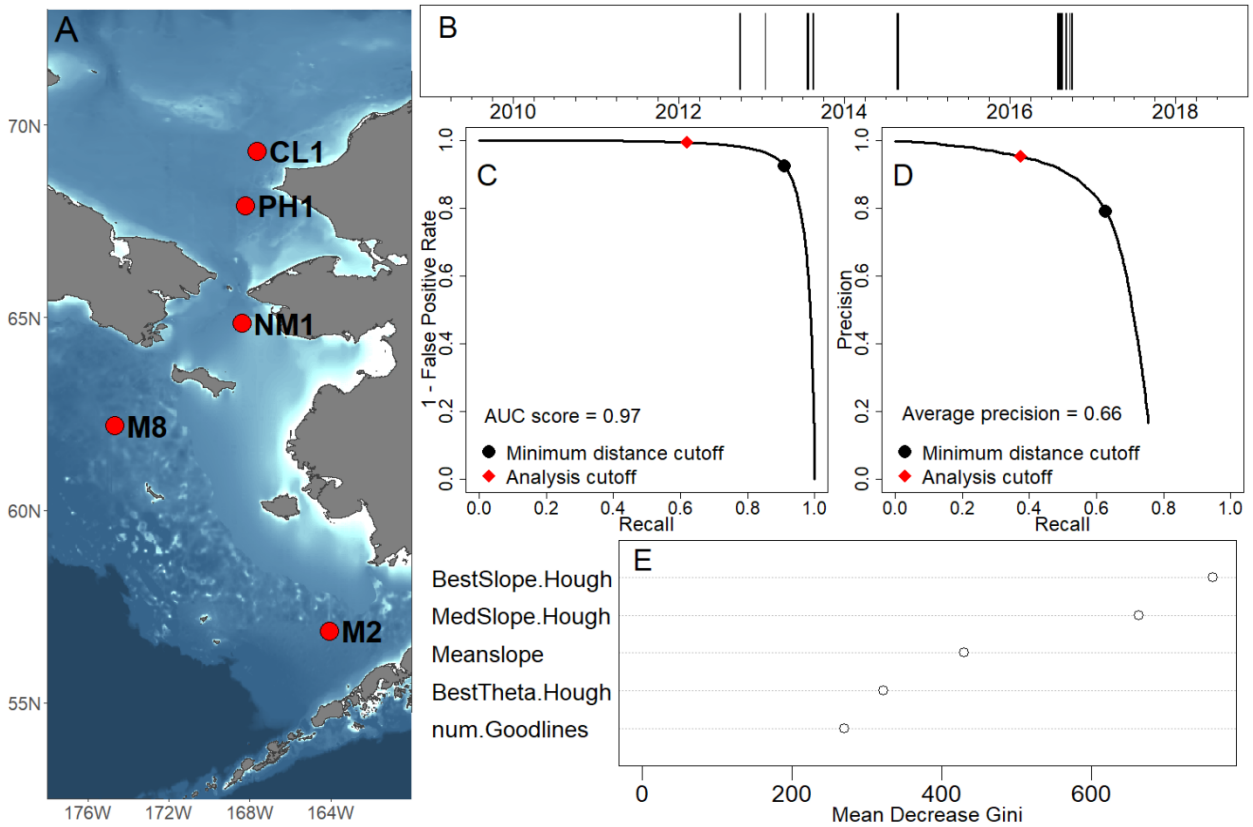


Figure 8. -- Results for fin whale C call (FN) detector. (A) Locations where ground truth data segments were sampled. (B) Days in which ground truth data segments were sampled. (C) Receiver operating characteristic curve for C call classifier. AUC = Area Under the Curve. (D) Precision-recall curve for entire C call detector. (E) Top five important features in the model, as determined by mean decrease in gini coefficient (see text for definitions).

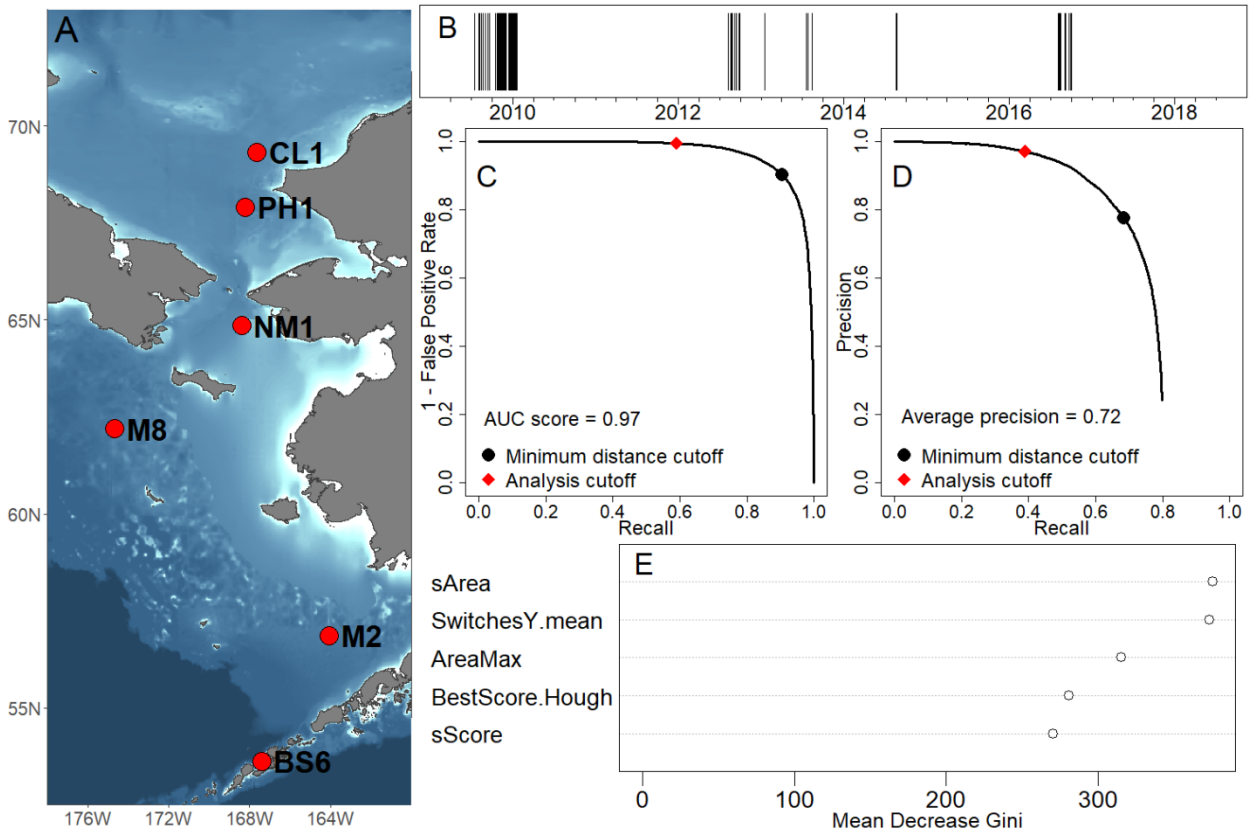


Figure 9. -- Results for fin whale backbeat call (BB) detector. (A) Locations where ground truth data segments were sampled. (B) Days in which ground truth data segments were sampled. (C) Receiver operating characteristic curve for C call classifier. AUC = Area Under the Curve. (D) Precision-recall curve for entire C call detector. (E) Top five important features in the model, as determined by mean decrease in Gini coefficient (see text for definitions).

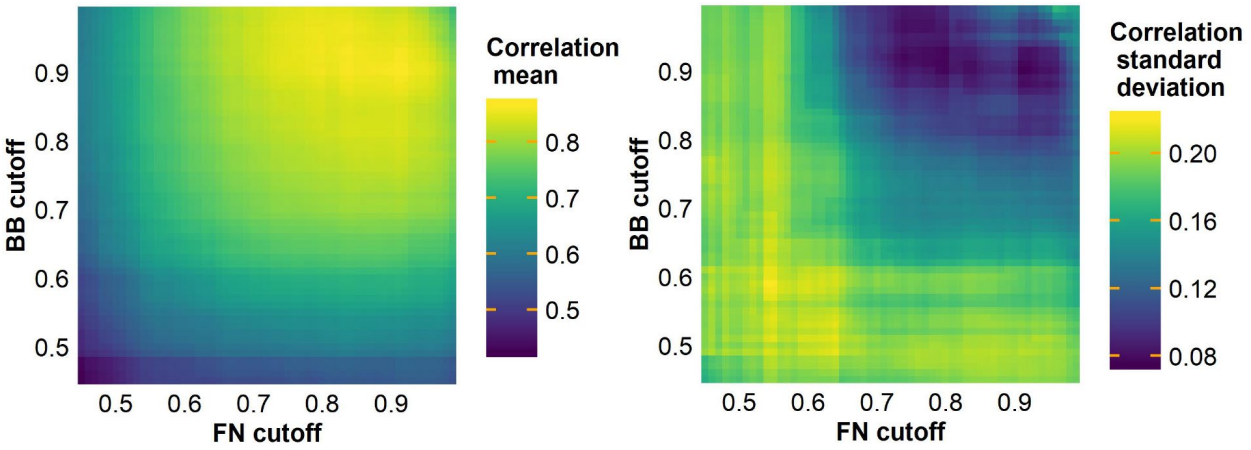


Figure 10. -- Correlation mean and standard deviation of correlation R scores for the FD calibration.

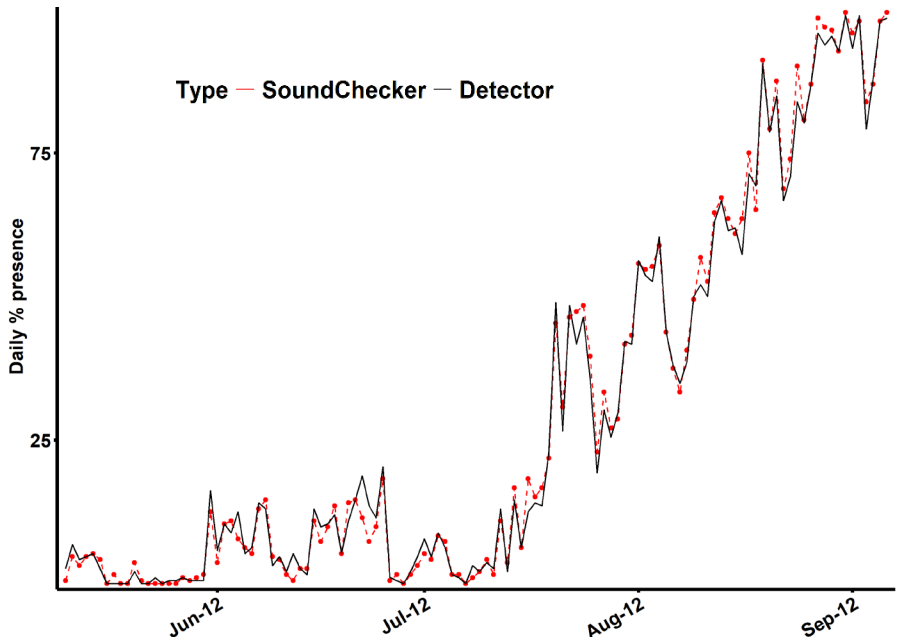


Figure 11. -- Daily % presence (% 5-minute bins/day with fin whale calls detected) for INSTINCT (black line) and SoundChecker (dashed red line) over the 2012 M2a mooring deployment. x-axis in month-years.

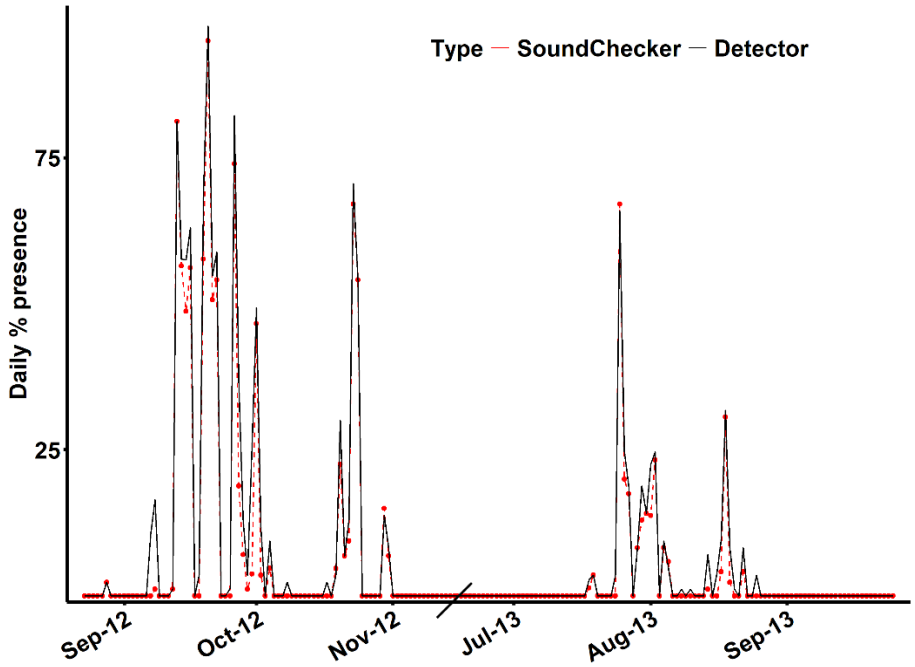


Figure 12. -- Daily % presence (% 5-minute bins/day with fin whale calls detected) for INSTINCT (black line) and SoundChecker (dashed red line) over the 2012 CL1 mooring deployment. x-axis in month-years.

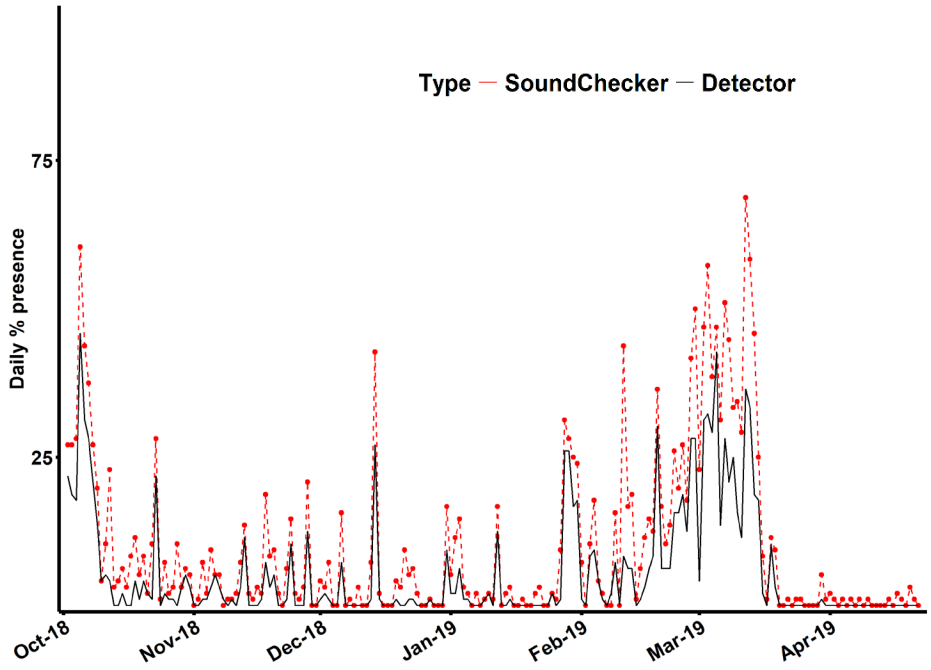


Figure 13. -- Daily % presence (% 5-minute bins/day with fin whale calls detected) for INSTINCT (black line) and SoundChecker (dashed red line) over the 2018 BS4 mooring deployment. x-axis in month-years.

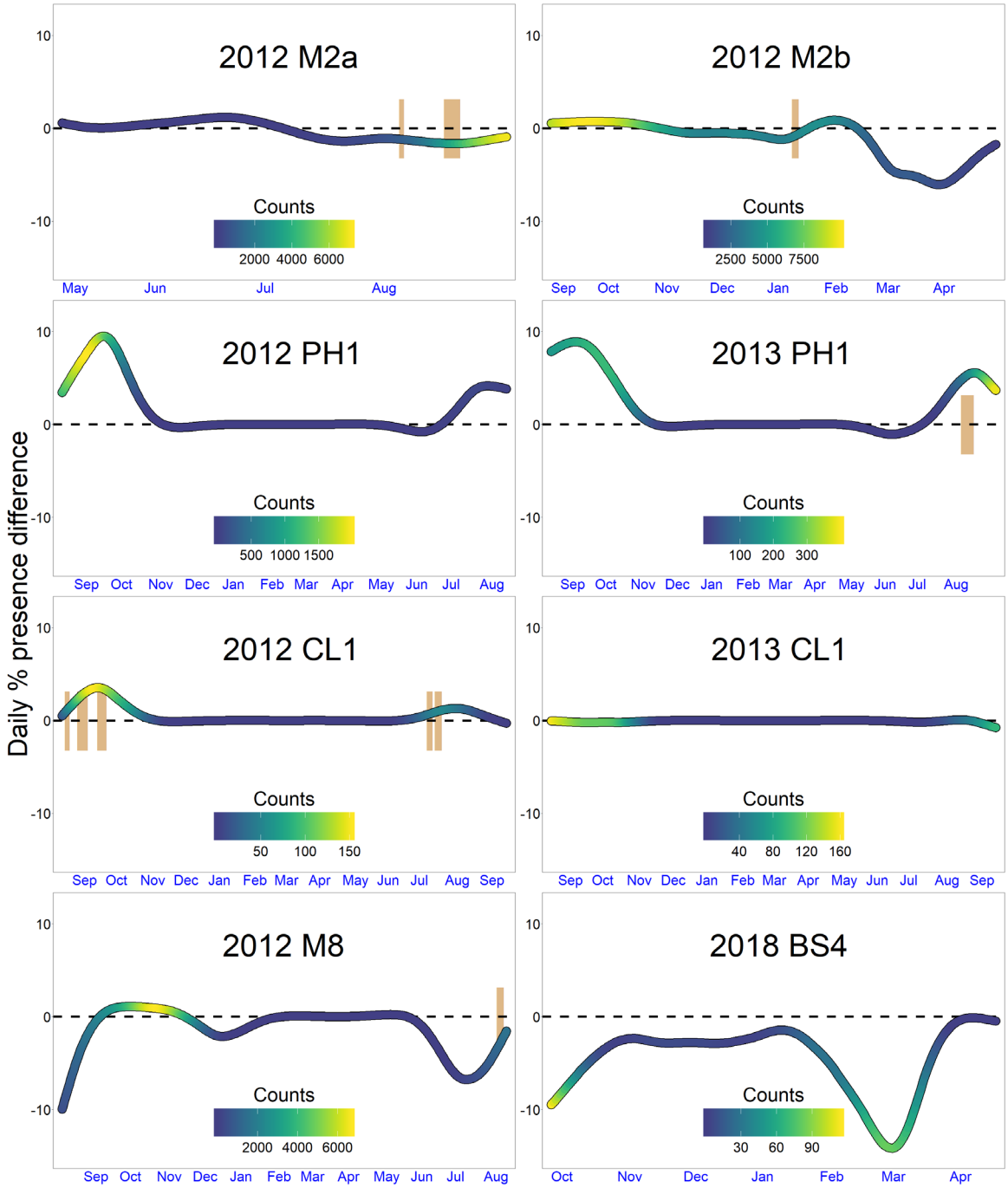


Figure 14. -- Difference in daily % presence (% 5-minute bins/day with fin whale calls detected) for eight test moorings for FD calibration. Negative values indicates detector under-sensitivity, positive values indicate detector over-sensitivity. Days used for ground truth for either FN or BB detectors are represented by brown rectangles. Counts represents the total call counts per day. Both daily % presence difference and counts are processed with a cubic smoothing spline with a smoothing parameter of 0.75. Note that color scale for counts varies among plots.

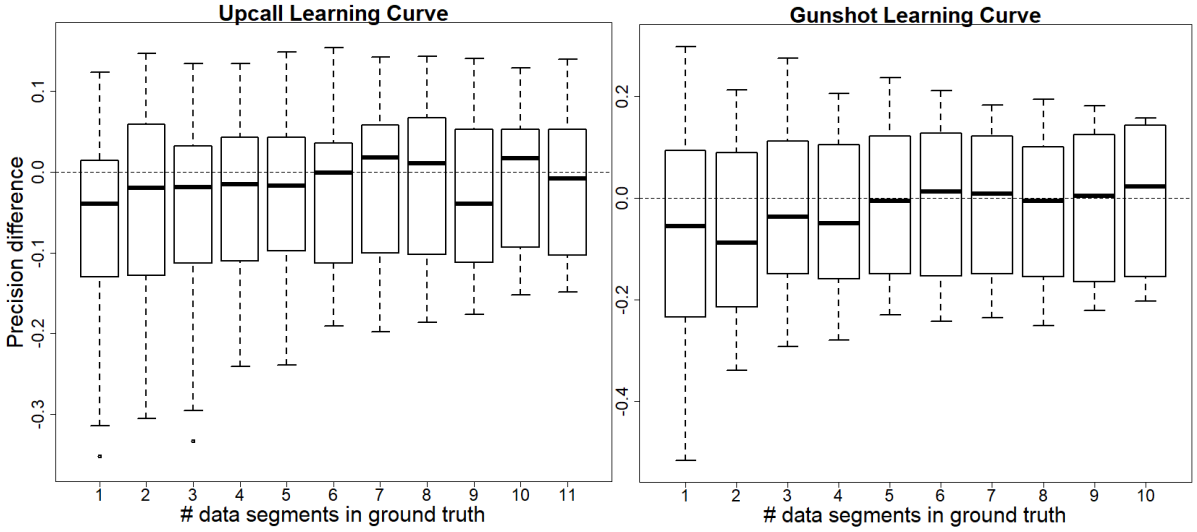


Figure 15. -- Learning curve of upcalls (left) and gunshots (right). The x-axis indicates the number of ground truth data segments used to generate the classifier for the given step, and the y-axis indicates the precision differential between the current classifier on its own ground truth and the next data segment.

Appendix

This appendix contains supporting details on the INSTINCT defined event detectors (A), classifier (B), and the Hough transform during feature extraction (C). It also contains information on the mooring deployments used in this study (D; Appendix Table 1; Appendix Fig. 9).

A: Defined event detector

Event detectors attempt to identify events based on received values that exceed a calculation of noise by a defined threshold (Erbe and King 2008). Spectrogram based event detectors (BLEDs) use amplitude values within a frequency band to distinguish background noise from events based on user parameters (Bioacoustics Research Program 2017). Each BLED is a nonspecific detector, which will be triggered by eligible events of specified signal-to noise ratio. However, making multiple BLEDs in series allows for rule based detection of events by comparing detections from different bands as an aggregate. This permits for discrimination of unlikely signals while retaining the computational efficiency of the event detector. This process is here called 'defined event detection'. Defined event detectors are specific to the attributes of each call type. They are constructed manually in an iterative process that attempts to maximize recall while keeping precision at manageable levels (low precision is tolerable during this stage, under the expectation that most false positives will be filtered out via classification). Defined event detection allows for a computationally light detection round using any definable time-frequency tendencies of the general call. Four defined event detectors were designed, corresponding to the RW, GS, BN, and FN INSTINCT detectors. Of these, two defined event detectors used algorithms that relied only on FM properties of the general signal (RW and FN), one used an algorithm that relied only on temporal association (BN), and one used an algorithm which dynamically incorporated both methods (GS). Of the five INSTINCT detectors presented, only the BB detector lacked a defined event detector due to inconsistent frequency modulation of the call type. Because of this, a single event detector was used for BB. In this appendix qualitative descriptions for defined event detection methods are presented.

The upcall (RW) defined event detector uses a custom algorithm to search for ascending sequences of 'minidetections' (individual detections within a BLED) within groups of minidetections designated by time similarity (Appendix Fig. 1). Time similarity refers to grouping by a fixed time interval, where a consecutive minidetection that falls outside of this interval is assigned to a new group. At current parameters, 'runs' of at least three minidetections are required for these groups to be considered a valid detection. All qualified runs are compared by length to determine the single best run, the minidetections of which are combined to define the putative detection. If lengths of multiple runs are the same, the following tie-breakers are used: first, the smallest maximum frequency jump in frequency range in between any two minidetections in the run, second, the fewest interruptions in the run (number of times in the run that consecutive streak is broken). By rejecting smaller runs, harmonics are designed to be disqualified as detections, and the fundamental frequency of the call is retained as the putative detection (Appendix Fig. 2).

The gunshot defined event detector custom algorithm searches for runs of stacked minidetections higher in the expected frequency range, and stacked or descending runs of minidetections lower in the expected frequency range (Appendix Fig. 3). The algorithm changes tolerance for accepting descending minidetections further with reduced frequency to account for the tendency of propagation to induce longer duration multipath arrival in the lower frequencies. These parameters were designed to account for the range of distortion present in signals with high or low dispersion, and validated with iterative performance testing and qualitative evaluation of performance on application to data segments. Stacked minidetections optionally transitioning into downsweeps are queried in these groups, and runs which fit the minimum four minidetection distance criteria are combined to be retained as the putative detection.

The boing defined event detector custom algorithm searches for stacked minidetections within groups (temporal similarity). At current parameters, two or more stacked minidetections are necessary to constitute a detection (Appendix Fig. 4). The fin whale C call defined event detector custom algorithm searches for downsweeps within groups. At current parameters, it

returns any downsweeps within a group provided they have at least four minidetections in the run (Appendix Fig. 5).

B: Classifier

INSTINCT generates bootstrapped binary random forest models to perform probabilistic classification of positive and negative signals based on ground truth data. The ground truth data are subset n-fold on a randomly selected three quarter partition of the data which then are rebalanced to equal positive/negative ratio by randomly excluding excess items from the larger class (typically the negative class) to produce n iterations of training datasets. Random forest models are generated for each training dataset with R package 'randomForest' (Liaw and Weiner 2002) with hyperparameter mtry = 11 and ntree = 500. Hyperparameter values are fixed as they have a small observed effect on detector performance and perform well between applications at these parameters. These models are applied differently depending on whether INSTINCT is being applied for performance assessment or deployed on novel data. When applied for performance assessment, these models are applied to the one quarter corresponding test data held out from the training partition for each iteration. Probabilities from each instance in the test data are recorded in a matrix of n columns, and rows equal to total items in the ground truth, resulting in each item having a probability vector of variable length depending on the number of times it was randomly selected in test data over n trials. These vectors are averaged, resulting in a cross validated probability for each instance in the training data. When applied to novel data, these models are generated in the same fashion, but instead are applied to each item of novel data resulting in probability vectors of length n that are averaged to produce a final probability. The training data partition was set to 75% for all classifiers after limited experimentation, but an optimal partition size for each detector would likely vary depending on species, energy detector configuration, soundscape characteristics, and training data available.

One notable aspect of INSTINCT is that due to the random sampling step during training, and the potential for any given instance to be underrepresented in the test data, detector performance is not deterministic and output probabilities may vary slightly in each application. Despite this, total performance statistics are very consistent between applications. If

repeatability is desired, the random sampling can be kept consistent between applications using the 'set.seed' function in R.

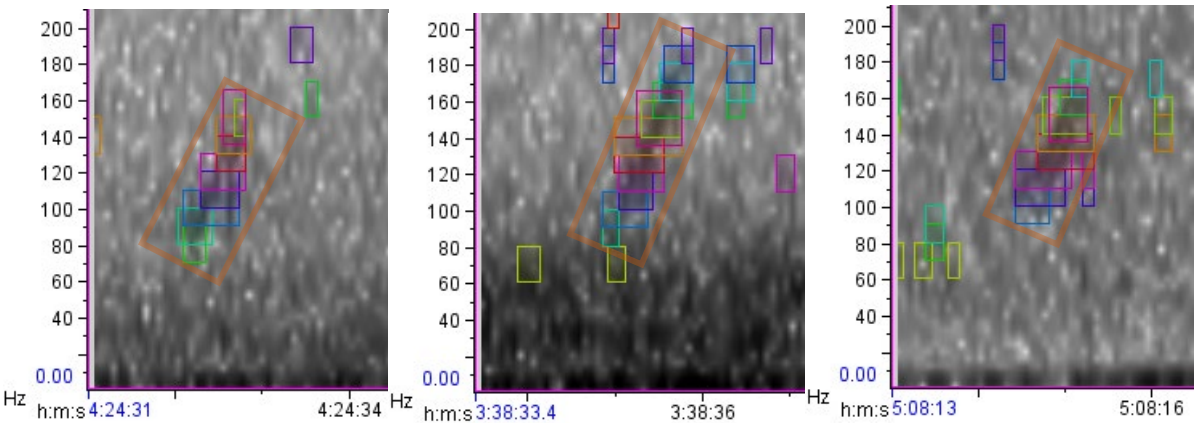
C: Hough line features

At present, there are 135 features extracted for use in the classifier. Of these, Hough line features are notably informative, especially in the FM signals. The Hough transform is a popular method in computer vision, allowing for detection of simple shapes (straight lines, circles, ellipses, etc.) within noisy or imperfect images, often in combination with edge extraction to produce the boundaries of more complex shapes (Collado et al. 2006, Jahan and Singh 2018). The Hough transform uses an accumulator matrix that accumulates votes for potential lines present in the image for each eligible ρ and Θ value for the image (Duda and Hart 1972). Votes for each ρ/Θ combination represent the line 'score' (Appendix Fig. 6).

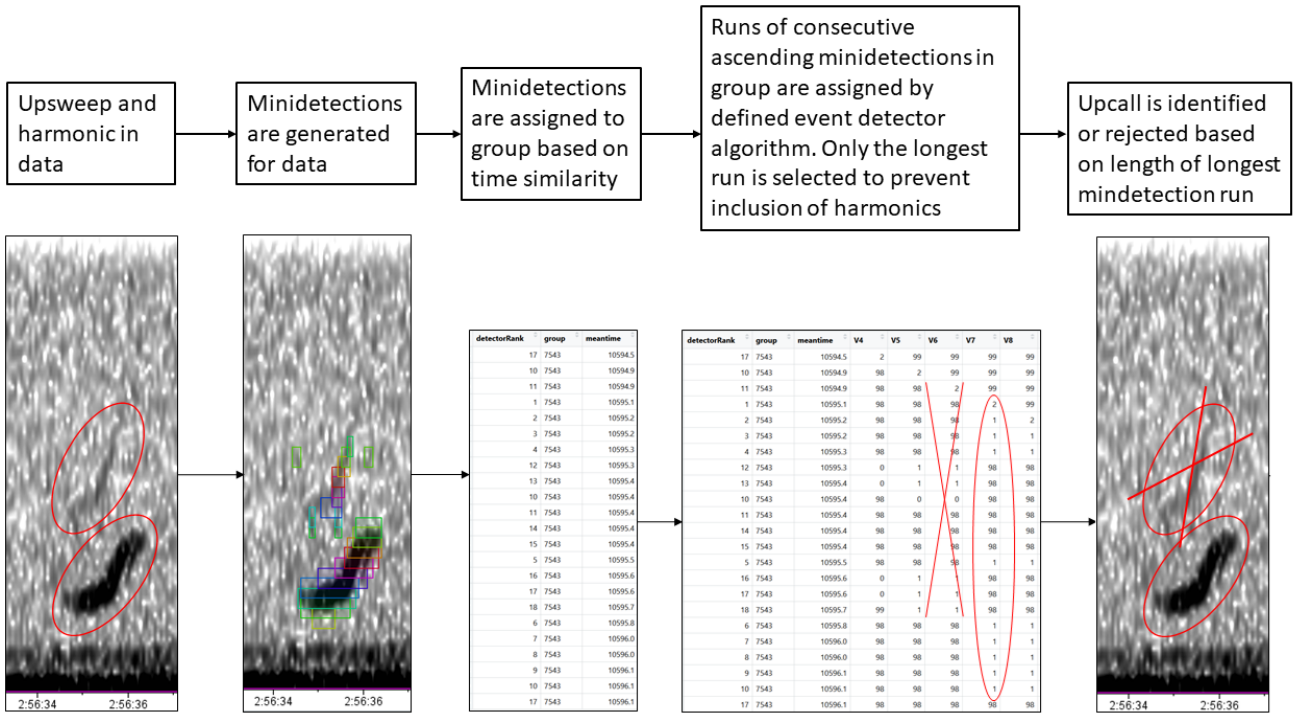
In the present application, Hough transformation was used to identify straight lines from each putative detection during feature extraction (Appendix Fig. 7). The attributes from the single best line from the Hough transform were extracted as features. These features corresponded to the ρ, Θ, slope (when finite), y-intercept (when finite) and score of the single best Hough line (Appendix Fig. 8). Features were also extracted from a subset of the top 30% of high scoring lines. These features corresponded to the mean, median, and standard deviation of ρ, Θ, slope (when finite), y-intercept (when finite) and score from lines in this pool. Additionally, these same Hough transform values were extracted for the highest scoring line from each independent shape object in the binary spectrogram, and the mean, median, and standard deviation were computed as additional fragmentation resistant features. Fragmentation refers to the splitting of calls into multiple shape objects that can occur during spectrogram thresholding in low signal-to-noise ratio conditions. Aside from artifacts from the thresholding process, fragmentation can occur from environmental effects on received signals. The Hough features are included with the other handcrafted extracted features, and used to inform the classifier. As a proxy for slope in noisy images, the Hough line features were particularly informative in the very consistently frequency modulated RW and FN call types. For gunshots, which are less consistently FM, Hough features related to slope nevertheless were

the top three most informative features. Even BB, which are not thought to have consistent FM, had one Hough feature in the top five informative features. Following our observations of BB characteristics, this feature represented the strength of the line, not a particular slope measurement. Only BN did not have Hough features within the top five for feature importance.

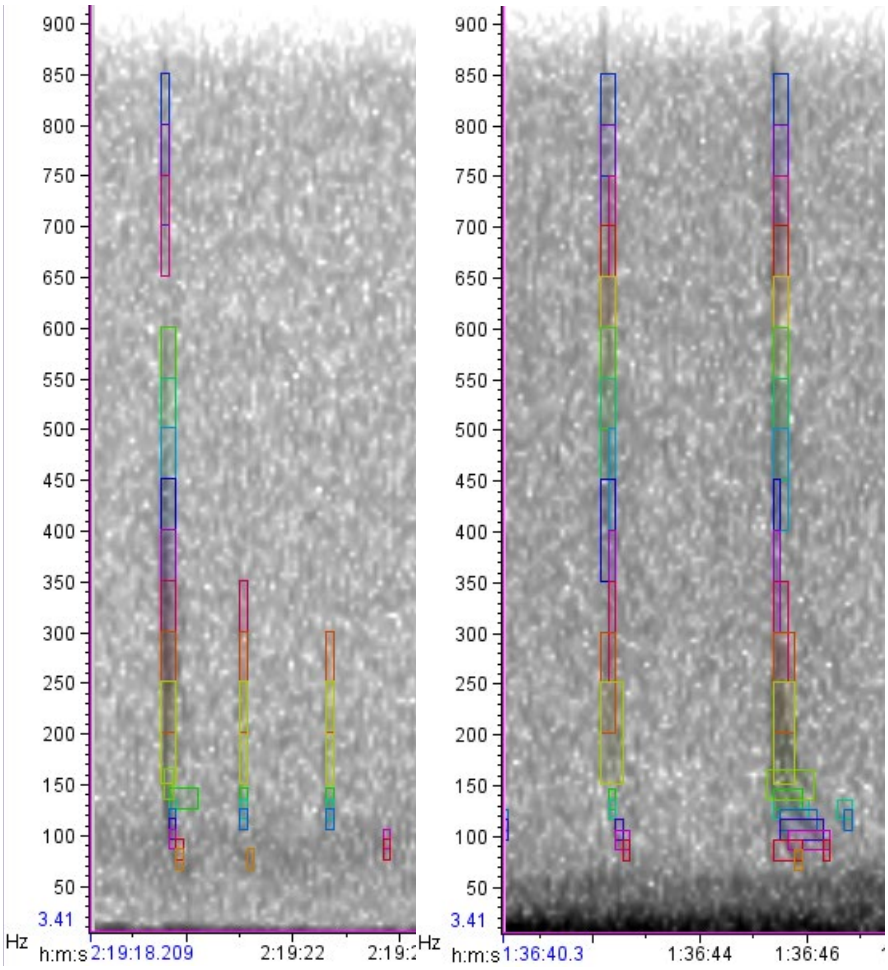
The consistent power of the Hough transform to inform the classifier, particularly for FM signals, indicates its strength for classification of call types. The current implementation is only used to find straight lines in images, but this algorithm can be applied to find other consistent shapes which may be meaningful for some stereotyped signals. To our knowledge, there is no literature that uses Hough lines as noise tolerant features in marine mammal acoustics, and the success in this application may be transferable to improve the performance of other model designs.



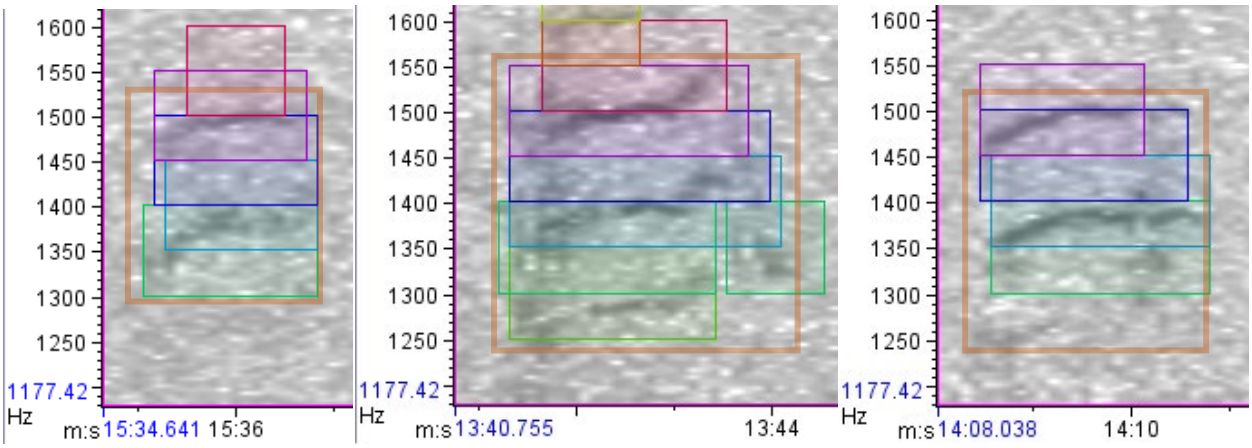
Appendix Figure 1. -- Three examples of the upcall defined event detector on true positive calls. True positive calls shaded with transparent orange box.



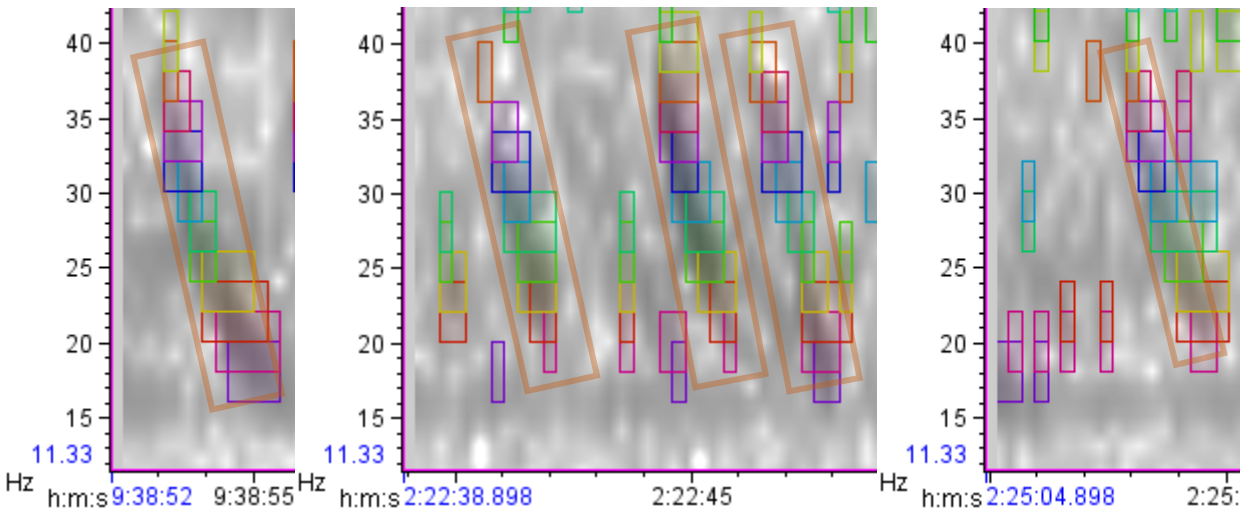
Appendix Figure 2. -- RW defined event detector algorithm.



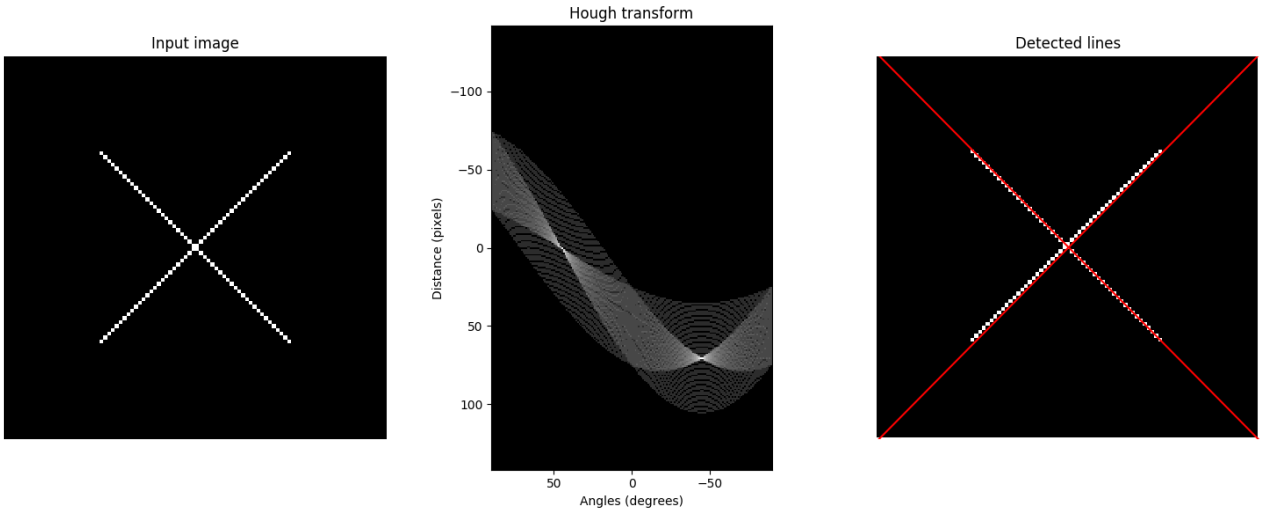
Appendix Figure 3. -- Three examples of the gunshot defined event detector on true positive calls. True positive calls shaded with transparent orange box.



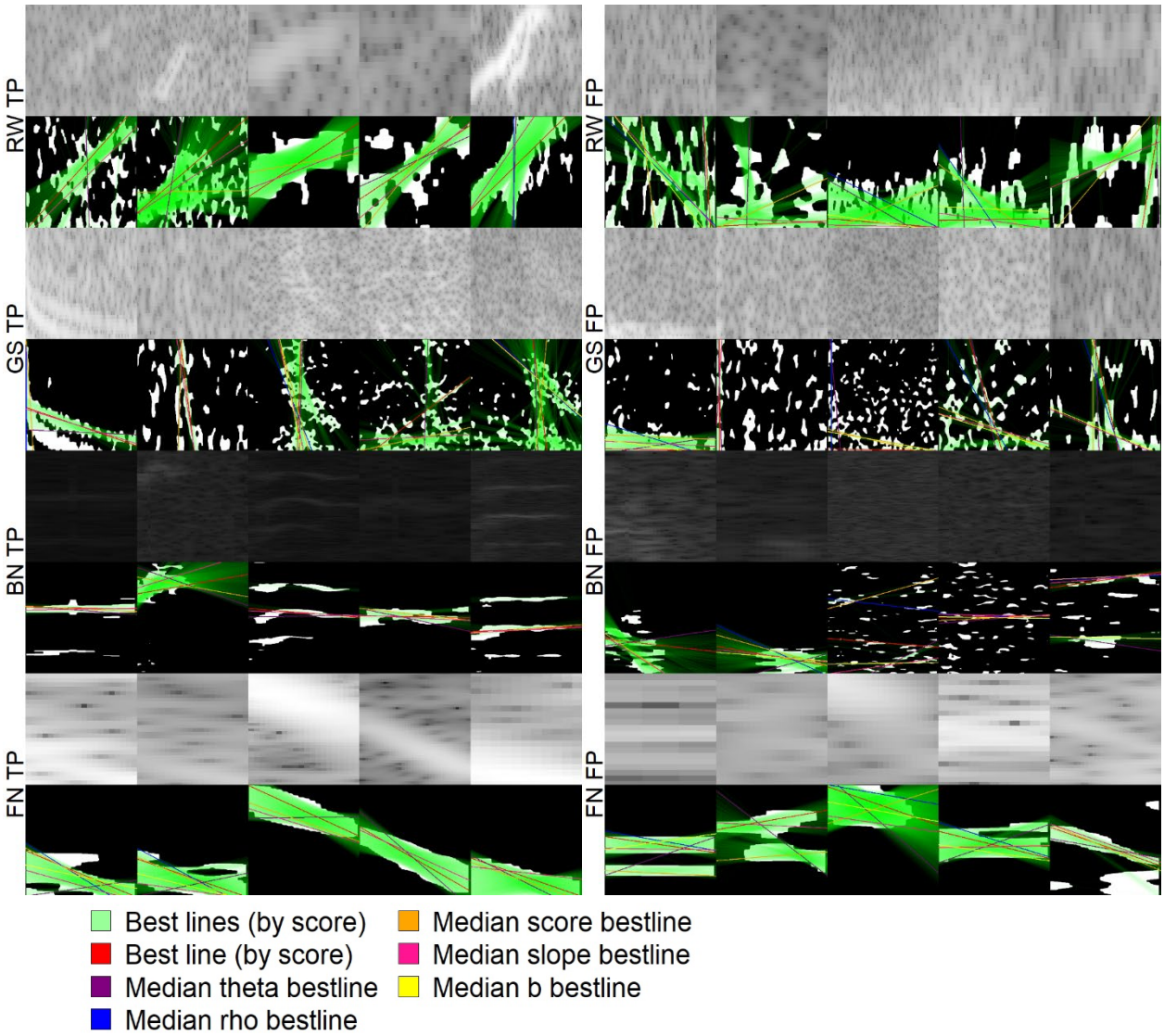
Appendix Figure 4. -- Three examples of the boing defined event detector on true positive calls.



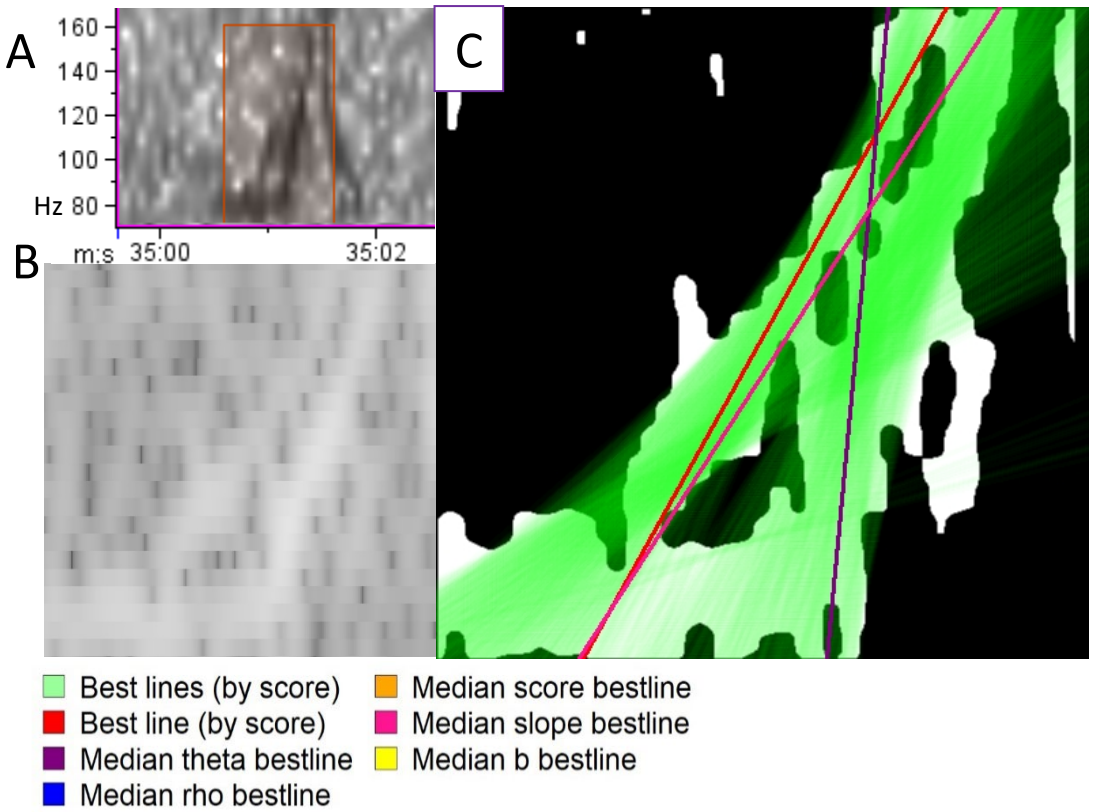
Appendix Figure 5. -- Three examples of the fin whale C call defined event detector on true positive calls. True positive calls shaded with transparent orange box.



Appendix Figure 6. -- Process of Hough line detection (from Van der Walt et al. 2014). In the Hough transform step, y-axis is equivalent to ρ, and x-axis is equivalent to Θ.



Appendix Figure 7. -- True positive and false positive spectrograms (top) and binarized spectrograms with superimposed top scoring Hough lines (bottom) for all INSTINCT detectors. Examples randomly selected from each class.



Appendix Figure 8. -- (A) Spectrogram of a true positive right whale upcall from training data with superimposed detection, viewed in Raven 1.5. (B) Spectrogram tile of the same call rendered during feature extraction. (C) Binarized spectrogram with top scoring Hough lines (green) superimposed. Colored lines indicate the best line (red), and lines which equal the values for the median of each parameter.

D: Organized list of mooring deployments used

Appendix Table 1. --Mooring deployment metadata.

Mooring Name	Mooring Site	Date Data Start	Date Data End	Duty Cycle (min) Record On	Duty Cycle (min) Period	Latitude (N)	Longitude (W)	Water Depth (m)	Sampling Rate
CZ11_AU_IC1c	IC1	9/3/2011	8/21/2012	85	300	70.817	163.136	43	16384
AL16_AU_CC2	CC2	9/20/2016	8/7/2017	80	300	70.016	166.86	47	16384
AL17_AU_CC2	CC2	8/9/2017	8/11/2018	80	300	70.016	166.86	47	16384
AW12_AU_CL1	CL1	8/23/2012	8/25/2013	85	300	69.307	167.648	48	16384
AW13_AU_CL1	CL1	8/25/2013	9/25/2014	80	300	69.316	167.632	48	16384
AW15_AU_CL1	CL1	9/21/2015	4/2/2017	80	300	69.317	167.623	49	16384
AW12_AU_PH1	PH1	8/22/2012	8/22/2013	85	300	67.909	168.195	58	16384
AW13_AU_PH1	PH1	8/24/2013	9/29/2014	80	300	67.907	168.203	55	16384
AW14_AU_PH1	PH1	9/17/2014	9/20/2015	80	300	67.908	168.202	68	16384
AW12_AU_KZ1	KZ1	8/21/2012	8/22/2013	85	300	67.125	168.602	43	16384
AW14_AU_NM1	NM1	9/22/2014	8/20/2015	80	300	64.849	168.39	48	16384
AW15_AU_NM1	NM1	9/10/2015	9/23/2016	80	300	64.848	168.39	44	16384
BS12_AU_08	M8	8/14/2012	8/20/2013	85	300	62.195	174.661	71	16384
AL16_AU_BS1	BS1	9/26/2016	9/28/2017	80	300	61.585	171.319	52	16384
BS16_AU_05	M5	9/28/2016	9/27/2017	80	300	59.911	171.731	68	16384
AW12_AU_BS2	BS2	8/12/2012	8/17/2013	85	300	59.244	169.413	53	16384
AW14_AU_BS2	BS2	10/18/2014	9/26/2015	80	300	59.243	169.414	52	16384
AW15_AU_BS2	BS2	9/27/2015	9/27/2016	80	300	59.243	169.413	53	16384
BS13_AU_04	M4	9/18/2013	10/17/2014	80	300	57.867	168.873	75	16384
BS14_AU_04	M4	10/19/2014	9/26/2015	80	300	57.882	168.879	70	16384
BS15_AU_04	M4	9/27/2015	9/27/2016	80	300	57.895	168.878	70	16384
AW12_AU_BS3	BS3	8/11/2012	9/13/2013	85	300	57.67	164.725	52	16384
AW14_AU_BS3	BS3	10/20/2014	9/27/2015	80	300	57.671	164.719	59	16384
AW15_AU_BS3	BS3	9/28/2015	9/28/2016	80	300	57.675	164.718	53	16384
AL16_AU_BS3	BS3	9/29/2016	10/1/2017	80	300	57.676	164.716	52	16384
BS12_AU_02a	M2	5/27/2012	11/8/2012	40	60	56.865	164.059	73	8192
BS12_AU_02b	M2	9/6/2012	5/5/2013	135	300	56.866	164.057	73	16384
BS15_AU_02a	M2	5/2/2015	9/27/2015	180	300	56.867	164.067	73	16384
BS15_AU_02b	M2	9/29/2015	5/4/2016	165	300	56.878	164.069	70	16384
BS16_AU_02a	M2	5/14/2016	9/29/2016	180	300	56.873	164.053	72	16384
AL18_AU_BS4	BS4	10/2/2018	4/22/2019	8	30	54.428	165.269	166	16384
RW09_EA_01	BS6	7/16/2009	1/20/2010	6.7	60	53.632	167.393	91	4000

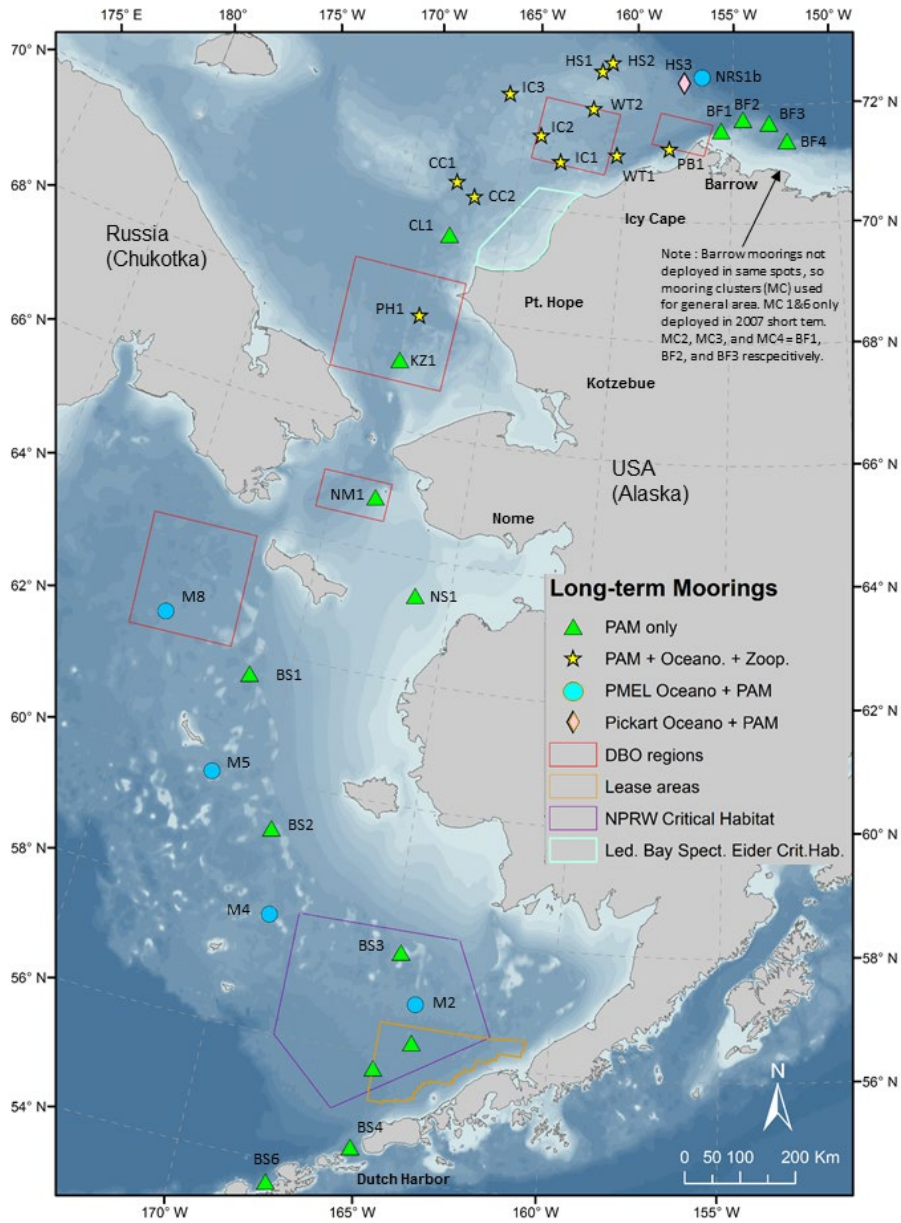
RW ground truth: AW12_AU_BS2, AW14_AU_BS2, AW14_AU_BS3, AW15_AU_BS2, AL16_AU_BS1, BS13_AU_04, BS14_AU_04, BS15_AU_02a, BS15_AU_02b, BS15_AU_04, BS16_AU_02a, BS16_AU_05

GS ground truth: AW12_AU_BS3, AW14_AU_BS3, AW15_AU_BS2, AW15_AU_BS3, AL16_AU_BS3, BS12_AU_02a, BS12_AU_02b, BS13_AU_04, BS14_AU_04

BN ground truth: CZ11_AU_IC1c, AW12_AU_KZ1, AW12_AU_CL1, AW14_AU_NM1, AW13_AU_PH1, AW14_AU_NM1, AW14_AU_PH1, AL16_AU_CC2, AL17_AU_CC2

FN ground truth: AW13_AU_PH1, AW15_AU_NM1, AW15_AU_CL1, BS12_AU_02b, BS12_AU_08a

BB ground truth: AW12_AU_CL1, AW13_AU_PH1, AW15_AU_CL1, AW15_AU_NM1, BS12_AU_02a, BS12_AU_02b, BS12_AU_08a, RW09_EA_01



Appendix Figure 9. -- Locations of mooring deployments.

U.S. Secretary of Commerce
Wilbur L. Ross, Jr.

Acting Under Secretary of Commerce for Oceans and Atmosphere
Dr. Neil Jacobs

Assistant Administrator for Fisheries
Chris Oliver

October 2020

www.nmfs.noaa.gov

OFFICIAL BUSINESS

National Marine Fisheries Service
Alaska Fisheries Science Center
7600 Sand Point Way N.E.
Seattle, WA 98115-6349