NOAA Technical Memorandum NMFS

MARCH 2013

EVALUATION OF AN AUTOMATED ACOUSTIC BEAKED WHALE DETECTION ALGORITHM USING MULTIPLE VALIDATION AND ASSESSMENT METHODS

Eiren K. Jacobson[1], Tina M. Yack[1,2,3], Jay Barlow[1]

[1] NMFS/NOAA, Southwest Fisheries Science Center
8901 La Jolla Shores Dr., La Jolla, CA 92037 USA

[2] Bio-Waves Incorporated
3642 2nd St., Suite #3, Encinitas, CA 92024 USA

[3] Joint Doctoral Program in Ecology
San Diego State University
San Diego, CA 92182 USA

NOAA-TM-NMFS-SWFSC-509

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Southwest Fisheries Science Center

The National Oceanic and Atmospheric Administration (NOAA), organized in 1970, has evolved into an agency that establishes national policies and manages and conserves our oceanic, coastal, and atmospheric resources. An organizational element within NOAA, the Office of Fisheries, is responsible for fisheries policy and the direction of the National Marine Fisheries Service (NMFS).

In addition to its formal publications, the NMFS 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, however, reflect sound professional work and may be referenced in the formal scientific and technical literature.

NOAA Technical Memorandum NMFS

This TM series is used for documentation and timely communication of preliminary results, interim reports, or special purpose information. The TMs have not received complete formal review, editorial control, or detailed editing.

MARCH 2013

EVALUATION OF AN AUTOMATED ACOUSTIC BEAKED WHALE DETECTION ALGORITHM USING MULTIPLE VALIDATION AND ASSESSMENT METHODS

Eiren K. Jacobson[1], Tina M. Yack[1,2,3], Jay Barlow[1]

[1] NMFS/NOAA, Southwest Fisheries Science Center
8901 La Jolla Shores Dr., La Jolla, CA 92037 USA

[2] Bio-Waves Incorporated
3642 2nd St., Suite #3, Encinitas, CA 92024 USA

[3] Joint Doctoral Program in Ecology
San Diego State University
San Diego, CA 92182 USA

NOAA-TM-NMFS-SWFSC-509

U.S. DEPARTMENT OF COMMERCE
Rebecca M. Blank, Acting Secretary
National Oceanic and Atmospheric Administration
Dr. Kathryn D. Sullivan, Acting Administrator
National Marine Fisheries Service
Eric Schwaab, Assistant Administrator for Fisheries

Evaluation of an automated acoustic beaked whale detection algorithm using multiple validation and assessment methods

Eiren K. Jacobson[1], Tina M. Yack[1,2,3], Jay Barlow[1]

(1) Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla Shores Drive, La Jolla, CA 92037 USA
(2) Bio-Waves, Inc., 364 2nd St., Suite #3, Encinitas, CA 92024
(3) Joint Doctoral Program in Ecology, San Diego State University, San Diego, CA 92182 USA

Abstract

Currently, the acoustic detection of beaked whales during passive acoustic surveys requires trained acousticians to identify beaked whale signals with the aid of various software programs. The development of reliable automated detection and classification methods will enable passive acoustic approaches to better meet monitoring needs for real-time mitigation of industry and military impacts. During ongoing development of automated beaked whale detectors and classifiers it will be important for researchers at different institutions to utilize standardized metrics of performance. At the Southwest Fisheries Science Center (SWFSC), automated detection algorithms for Cuvier's beaked whale (*Ziphius cavirostris*) and Baird's beaked whale (*Berardius bairdii*) were developed using PAMGUARD software (Douglas Gillespie: www.pamguard.org). To evaluate the performance of these beaked whale detectors, 15 ten-minute recording segments were processed in PAMGUARD, and the resulting signal detections were compared to manual logs of beaked whale signals confirmed by an experienced acoustician. The comparison was conducted using three methods: precise timestamp matching between manual and automated detections, detection counts from one-minute time bins, and binary presence/absence detection classification of one-minute bins. The detections were scored as true positive, false positive, false negative or false classification. Detector efficacy was quantified using measures developed for information retrieval systems (precision, recall, and F-score) as well as the Receiver Operating Characteristic. Calculated performance scores were compared across evaluation methods. We found that the method used to evaluate detector functionality greatly influences the resulting performance scores and subsequently our perception of detector ability. Therefore, it will be important for researchers to clearly communicate methods and results of detector evaluation. To allow for greatest precision and applicability to different recording datasets, we recommend that beaked whale detectors be evaluated using timestamp matching between manual and automated detections in trial datasets and that F-scores be used to compare detectors. This approach avoids problems associated with binning datasets by eliminating the need for a measure of false negatives.

Introduction

Passive acoustic monitoring for beaked whales

Passive acoustic monitoring using automated acoustic detection algorithms is increasingly used to detect beaked whale species during cetacean ship surveys (Zimmer and Pavan 2008, Gillespie et al. 2009b, Yack et al. 2010). Beaked whales (family Ziphiidae) are a family of at least 21 visually elusive and deep diving species (Dalebout et al. 2004). Since these animals forage in deep waters, they are primarily distributed offshore, and global species distributions and population structures are poorly understood (Dalebout et al. 2004). Due to their long dive times and subtle surfacing behavior, beaked whales are not well suited for standard visual survey methods (Barlow 1999). The strong connection between beaked whale diving behavior and vocal activity, combined with their general elusiveness makes passive acoustic monitoring (PAM) an ideal method for surveying these whales (Marques et al. 2009). PAM is comprised of the detection, localization, and classification of a vocal signal and for beaked whales can be carried out either via an array of hydrophones towed behind a vessel (e.g., Gillespie et al. 2009b) or via an array of stationary hydrophones floating at the surface or mounted on the sea floor (e.g., Marques et al. 2009). While current PAM procedures for beaked whales (including implementation of automated acoustic detection algorithms) require manual support by an experienced acoustician, increased automation of PAM methods will enable non-experts to detect, localize, and classify beaked whales (Yack et al. 2009). Several researchers are currently working to improve beaked whale detection algorithms as evidenced by the proceedings of the Fifth International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals Using Passive Acoustics (Mt. Hood, Oregon, 2011). In order to compare detector performance across software platforms and datasets, it will be necessary for researchers to establish a standard metric of performance evaluation.

Beaked whales emit regular echolocation clicks while foraging, followed by rapid but weak buzz clicks during the final stages of prey capture (Johnson et al. 2004). Regular echolocation clicks are useful for PAM due to their strength and regularity (Zimmer et al. 2008). The frequency upsweep and long duration of beaked whale clicks make them distinctive from other cetacean clicks, even those that echolocate in similar frequency ranges (Johnson et al. 2006). These distinguishing characteristics are important in the development of effective automated detection methods. Beaked whale echolocation clicks have central frequencies of 20-40 kHz, inter-click intervals (ICIs) of 0.2-0.4 s, and durations near 200 ms (Johnson et al. 2004, Tyack et al. 2006, Zimmer et al. 2008). Frequency modulation of beaked whale clicks occurs at a rate of approximately 110 kHz per ms (Johnson et al. 2006). When beaked whales near their prey target (within ~3m, Johnson et al. 2004), clicks accelerate and the ICI decreases to the point that clicks are perceived as a buzz. Buzz clicks have been documented at rates of 250 clicks per second (Johnson et al. 2004), are higher in frequency than regular echolocation clicks, and are not frequency modulated (Johnson et al. 2006).

The two species of interest in this study were Cuvier's beaked whale (*Ziphius cavirostris*) and Baird's beaked whale (*Berardius bairdii*). These two species were chosen because they can be reliably identified both visually and acoustically in the study area. Cuvier's beaked whales produce regular echolocation clicks with a central frequency range of 38-42 kHz (Johnson et al. 2004, Zimmer et al. 2005) while Baird's beaked whales produce clicks with a central frequency range of 22-25 kHz. In addition to foraging clicks, Baird's beaked whales vocalize socially via tonal sounds and whistles at the surface (Dawson et al. 1998). In this study, target vocalizations for acoustic detection of both species were regular echolocation clicks.

Development and evaluation of automated acoustic detection algorithms

In the field of cetacean bioacoustics, automated detection and classification algorithms have been utilized for over a decade (Mellinger et al. 2007). Automated detection algorithms have advantages over manual signal analysis due to their ability to rapidly process large datasets with a quantifiable and consistent detection bias (Mellinger 2004, Mellinger et al. 2007). Approaches to automated detection and classification include neural networks, template matching, spectrogram correlation, and energy band comparisons, among others. Different vocalizations may be particularly suited to certain automated detection methods. For example, stereotyped vocalizations produced by mysticetes are amenable to template-matching methods, while variable tonal sounds produced by delphinids are better detected using band-limited energy summation methods (Mellinger et al. 2007).

In order for automated detector output data to be useful, the capabilities of the detector must be quantified. Four measures are commonly used to quantify detector performance. These include true positive, false positive, false negative, and true negative detection rates, as evaluated in direct comparison to manual signal detections (Mellinger and Clark 2000, Mellinger 2004, Munger et al. 2005, Zimmer et al. 2008, Marques et al. 2009, Yack et al. 2010). Detectors may be further evaluated in relation to variable detection thresholds, ICIs, and the rate of detection of cetacean individuals or groups. Ideally, a detection method would find only true signals, but in reality the sensitivity of the detector must be configured to optimize a trade-off between true positives, false positives and false negatives (Mellinger et al. 2007). The desired application of the detector will influence the optimization method, since for some contexts it is crucial that no true signals are missed (e.g., during real-time monitoring for impact mitigation), whereas for others it is important to eliminate false detections even if some true signals are missed (e.g., when using automated detection methods for density estimation). Ultimately, the characteristics of an automated detector are less important than accurately quantifying detector performance (Marques et al. 2009).

Mellinger and Clark (2000) addressed detector evaluation when they compared the performances of neural network and spectrogram correlation methods for detecting bowhead whale (*Balaena mysticetus*) vocalizations. Using plots of false negatives versus false positives over variable detection threshold levels, the authors selected the detector with the lowest combined error, or sum of false positive and false negative error rates, as optimal. For the detector using a neural network the combined error was 1.6% while for spectrogram correlation the combined error was 2.5%. Later work using these same two methods on a larger dataset of right whale (*Eubalaena japonica*) vocalizations found the neural network method to have a combined error of 6% while spectrogram correlation had a combined error of 26% (Mellinger 2004). The discrepancies in error rates between these two studies is likely due to the size and nature of the training datasets used. Below we will discuss potential problems with training datasets.

To conduct these detector comparisons, both Mellinger and Clark (2000) and Mellinger (2004) used a dataset of sound files a few seconds in length containing either target signals (bowhead whale or right whale calls) or noise (background noise, other marine mammal vocalizations, or equipment noise) sampled from larger recording datasets. Mellinger (2004) recommended selecting noise samples that are likely to trigger the detector rather than random recording samples. Training datasets in both studies contained large samples of signals and noise, 588 and 888 respectively in Mellinger and Clark (2000) and 1,857 and 6,359 respectively in Mellinger (2004). Short recording samples allowed the authors to score each sample as a true positive, false positive, true negative, or false negative and calculate error rates using these scores.

The use of short recording samples as a testing dataset may artificially deflate the error rates associated with an automated detector. Munger et al. (2005) report that a detector developed using short (one minute) and intermediate (one hour) recordings containing right whale calls became less effective when applied to datasets with durations increasing from one minute to one hour to one week. Over one-minute trial periods, the detector resulted in 19% missed detections and 25% false detections, while over a one-week trial period the same detector missed over 30% of target signals and produced over 90% false detections. Leary et al. (2011) found that when applying automated detectors to long-term recordings in the Arctic, detector performance did not stabilize until at least 400 random two-minute data samples were evaluated.

The increase in noise and decrease in occurrence of true signals over longer sampling periods can make detectors developed using short sample recordings ineffective when used on large datasets (Munger et al. 2005). These results suggest that short sample recordings are an unrealistic representation of whole datasets. For example, the ratios of noise samples to true signal samples of approximately 2:1 and 3:1 in the evaluation datasets used by Mellinger and Clark (2000) and Mellinger (2004) respectively are not comparable to the true prevalence of vocalizations in large recording datasets. Additionally, when a technician manually selects noise samples that they believe the detector will struggle to discern from true signals (as recommended by Mellinger 2004) a bias is introduced into the training dataset. A sample dataset constructed in this way contains only signals and loud noises in the frequency range of interest and is therefore not representative of the larger dataset. This discrepancy in dataset quality leads to inaccurate error quantification. Furthermore, binary classification of sample recordings containing target calls as either true positive or false negative will obscure multiple detections of a single true signal that can occur with some automated detection methods. A better approach to detector validation would be to use random representative recording subsamples containing target signals as well as a variety of typical noise conditions.

Evaluation of automated beaked whale acoustic detection algorithms

When developing automated detectors for beaked whale vocalizations, it is important to consider the context in which the detector will be applied. Here we consider issues relevant to the use of automated detectors for density estimation.

Zimmer et al. (2008) examined properties of beaked whale vocalizations in relation to dive behavior using a combination of bottom-mounted hydrophones and digital acoustic recording tags. The authors recommended using a two-part detection and classification scheme that would examine both the spectral upsweep feature of beaked whale clicks and the ICI of click series in order to determine beaked whale presence. This method requires that animals be on-axis to the receiver, so that the spectral upsweep of clicks is preserved, and close to the receiver, so that entire click trains are captured and accurate ICIs are calculated. The presence of multiple vocalizing animals may further complicate any classification scheme based on ICI.

Using multiple acoustic methods simultaneously can increase the precision of beaked whale abundance estimates. Ward et al. (2008), Marques et al. (2009), and Moretti et al. (2010) all used data collected from bottom-mounted hydrophones in conjunction with DTAGs and automated beaked whale detectors for abundance estimation. To evaluate results, Ward et al. (2008) compared the performances of matched filter and FFT-based signal detectors. The matched filter method detected 92% of clicks while the FFT detector detected 49% of clicks. Clicks emitted within 30 degrees of the hydrophone axis were detected at the greatest ranges. Marques et al. (2009) used the output of an FFT-based energy detector for beaked whale clicks to calculate the probability of detecting vocalizations, the vocalization rate, and the proportion of false positive detections produced. Output from the automated detector that was classified as a beaked whale signal was verified by manual examination. False positives comprised 55% of total detector yield. Moretti et al. (2010) used the same methods to verify detections using twenty uniformly spaced ten-minute samples and calculated a false positive rate of 26%. One caveat to detector evaluation using a combination of DTAG and bottom-mounted hydrophone data is that DTAGs are instrumented on beaked whales only on calm weather days, when animals can be easily seen at the surface and safely approached. Therefore, automated detection methods and resulting density estimation functions constructed and evaluated using these data may not accurately represent beaked whale signal detectability and density in stormy, noisy situations (Ward et al. 2010).

Beaked whale literature reviewed thus far contains analyses of recordings collected using bottom-mounted hydrophone arrays. At the Southwest Fisheries Science Center (SWFSC), shipboard acoustic line-transect surveys have been conducted in conjunction with visual surveys for over a decade (Rankin 2008). During these surveys, a towed hydrophone array is used to record marine mammal vocalizations during daylight hours. Advantages to a shipboard platform include the ability to cover a larger geographic area and the ability to obtain visual confirmation of species identification for acoustic encounters. However, towed hydrophone arrays are subject to higher levels of ambient noise than bottom-mounted hydrophones due to ship cavitation and relatively shallow hydrophone tow depth. Additionally, the speed of the recording platform, the rapid attenuation of beaked whale clicks, and the orientation-dependent detectability of the clicks pose challenges to surveys using towed array data to quantify beaked whale presence.

Yack et al. (2010) tested five beaked whale detection algorithms on towed-array data collected during a 2007 SWFSC shipboard survey. True and false positive detection rates were calculated using a subset of 60 minutes of data from two of the trial days. Sixty one-minute bins were evaluated for presence or absence of true beaked whale clicks. Test data also included five encounters when only Risso's dolphins (*Grampus griseus*) were present in the recordings in order to quantify false positive detection rates of beaked whale signals for this species. This species was chosen as a representative delphinid with echolocation clicks likely to be classified as beaked whale signals by the automated detector due to similar ICIs and peak frequencies of echolocation signals. The combined error associated with the PAMGUARD 1.0 (Gillespie et al. 2009a) click detector was 29%. A Gaussian mixture model (GMM) had a lower combined error (21%) but a higher rate of false detections (10% GMM versus 7% PAMGUARD). PAMGUARD was selected as a platform for further beaked whale detector development based on performance in this test, real-time detection capability, adaptability to different species and recording noise conditions, and ability for the user to refine and modify click classifiers in real-time as needed to optimize performance.

Receiver Operating Characteristic (ROC) and Precision-Recall (PR) Frameworks

The ROC framework is used to present the results of binary decision problems encountered in a range of disciplines, most notably in machine learning and medicine. The output of a binary decision algorithm (the hypothesized classes) can be projected onto the true classes (as determined by a human operator) and divided into the number of true positives, true negatives, false positives, and false negatives. These values make up a confusion matrix. From these values, the true positive rate (of all positive cases, how many did the algorithm identify as positive?) and false positive rate (of all negative cases, how many did the algorithm identify as positive?) can be calculated. An ideal algorithm would produce results with a true positive rate of one (100%) and a false positive rate of zero. Since classification algorithms are seldom perfect, the ROC framework is used to optimize the trade-off between true positives and false positives according to the research question at hand. The cost of increasing the true positive rate of an algorithm is usually an increase in the false positive rate. Plots of ROC values for different algorithms are often used to visually represent this tradeoff and to select the most appropriate algorithm (Figure 1). Swets et al. (2000) and Fawcett (2006) provide helpful reviews of ROC theory and applications.

The ROC framework is commonly used in a medical context and requires data points to be discrete units, like healthy and unhealthy patients. When applying this framework to the automated detection of cetacean vocalizations, and beaked whales in particular, the unit over which the algorithm is operating must be considered carefully. Since recordings are continuous, one approach is to break the recordings into discreet segments of a few seconds or a few minutes in length, as was done in Mellinger and Clark (2000) and Mellinger (2004). Since the mysticete vocalizations targeted in these studies were several seconds in length and non-overlapping, it was possible to extract discrete signal and noise samples from the continuous dataset. Due to the short durations of beaked whale clicks, it is difficult to define a time unit that can contain only one vocalization and therefore only one of the four possible evaluation classes. For example, a detector operating over a single second of true beaked whale recordings could produce true positives, false positives, and false negatives.

Setting aside the possibility of multiple scores within a single second time unit, beaked whales occur in such low densities that true signals are very rare within recording datasets. For example, within a ten-minute period of towed-array data, it would be extremely rare to detect even a hundred beaked whale vocalizations, and these occur within 0.4 seconds of one another. Therefore, an ideal detector would identify perhaps 50 beaked-whale-positive seconds in the 600-second period. The output of this ideal detector would be classed as 50 true positives, 0 false positives, 0 false negatives, and 550 true negatives. The problem with this scenario is that the number of seconds classed as true negatives is an order of magnitude higher than the number of seconds classed as true positives. If these 100 beaked whale vocalizations were the only detections within an entire hour or an entire day, the number of true positives would remain the same while the number of true negatives increases proportionate to the duration of the recording. The ROC framework requires a measure of true negatives in order to calculate a false positive rate, but in the case of beaked whale vocalizations, the false positive rate is artificially decreased with an increased sample size and therefore is not a very informative metric.

Due to the low densities of beaked whale vocalizations in towed array datasets, the output of automated detection algorithms has a skewed class distribution that can be better analyzed using a Precision-Recall (PR) framework (Davis and Goadrich 2006). The PR framework does not require discreet sampling units because it does not require a measure of true negatives. Events are scored as true positives, false positives, and false negatives and can be displayed in a contingency table, and the different evaluation scores are calculated as follows: the precision of a detector is the proportion of all automated detections that are true positives. The recall of a detector is the same as the true positive rate as defined in the ROC framework and is the proportion of all true events that are identified by the automated detector. Performance curves are used to compare different algorithms by plotting precision against recall. In this space both scores for an ideal detector would approach one (Figure X). The F-score is the geometric mean of precision and recall and can be weighted to emphasize the role of either precision or recall in optimizing detector performance.

The PR framework is not independent from the ROC framework but offers a more thorough approach to the evaluation of automated detectors. Examining algorithm performance with PR metrics can expose differences in algorithm performance that are not visible in the ROC framework. Davis and Goadrich (2006) prove that when comparing the performance of multiple algorithms, a particular algorithm can be optimal in ROC space only if it is optimal in PR space. In the case of odontocete vocalizations, we believe that the PR framework will provide an appropriate and thorough approach to the evaluation of automated detection algorithms.

Study Objectives

In the present study, we developed a beaked whale detector using an energy band comparison algorithm in PAMGUARD 1.9.01. The detector was applied to recordings from a towed hydrophone array survey, and the resulting detector output was analyzed with reference to methods used in previous studies of the characterization of beaked whale detectors. Through application of the ROC and PR evaluation frameworks to our particular dataset, we aim to better understand the limitations of our automated detector in relation to our research questions and to more effectively integrate automatic and manual detection methods.

Methods

Acoustic monitoring and recording during a line-transect survey

Data for this study were collected during the 2008 Oregon, California, and Washington Line-transect and Ecosystem (ORCAWALE) cruise and the 2009 Channel Islands Beaked Whale Acoustic Survey (BWAS). The ORCAWALE survey transited between the coast and 556 km offshore, covering a total of approximately 11,600 km of predetermined transect lines (Barlow 2010). Between 28 July and 10 December 2008, a five-element hydrophone array consisting of two mid-frequency hydrophones (frequency response 500Hz to 55kHz +/- 5dB re 1 V/µPa) and three high-frequency hydrophones (Reson TC4013 hydrophones with a frequency response of 1.5 to 150 kHz ±3 dB and a sensitivity of -170 dB re 1V/µPa after 40 dB pre-amplification) were towed 300 m behind the NOAA ship *MacArthur II* at a depth of four to eight meters. Data from two of the high frequency oil-filled array hydrophones were digitized at a rate of 480 k-samples/sec using a National Instruments USB-6251 interface and were continuously recorded to hard drives using Logger 2000 (Douglas Gillespie: www.ifaw.org/sotw) software. During daylight hours, a visual team of three observers surveyed for marine mammals from the flying bridge of the ship using 25x and hand-held binoculars. Five computers were operated by acousticians to detect cetaceans using a combination of manual and automated methods. Spectrographic displays were monitored by acousticians using ISHMAEL software (Mellinger 2001). Manually detected cetacean vocalizations were localized using cross-correlation algorithms in ISHMAEL and plotted in conjunction with GPS positions using Whaltrak 2.6 (Jay Barlow). Beaked whale vocalizations were detected using PAMGUARD 1.0 detection algorithms. Special protocols were conducted when either the visual or acoustic team detected beaked whales. When a beaked whale group was detected in good survey conditions, the vessel was maneuvered to obtain acoustic recordings and, if possible, to obtain sightings or re-sightings of the animals.

A second acoustic dataset was collected in 2009 in the Channel Islands. From 18 to 25 August 2009, a three-element hydrophone array was towed 100 m behind the sailing vessel *Nauti Buoys* over 950 km of trackline in the Southern California Bight. The hydrophone array was comprised of three high-frequency hydrophones (Reson TC4013 hydrophones with a frequency response of 1.5 to 150 kHz ±3 dB and a sensitivity of -165 dB re 1V/µPa after 40 dB pre-amplification). A Magrec was used to high-filter the analog signal at 2Hz. The signal was digitized at a sampling rate of 384 kHz using a National Instruments 6251 USB data acquisition board connected to a 12 V computer and recorded continuously using Logger 2000. Due to limited power on the research vessel, the entire acoustic system was run off 12 V batteries. In addition to aural and visual monitoring by a technician, Rainbow Click software was used for automatic detection and classification of beaked whale echolocation signals. A team of two visual observers used 7x50 handheld binoculars to search for marine mammals. More detailed survey methods can be found in Yack et al. 2011.

Processing survey recordings with PAMGUARD software

During the ORCAWALE cruise there were eighteen joint visual and acoustic encounters of beaked whales: seven of Baird's beaked whales, seven of Cuvier's beaked whales, one of *Mesoplodon* sp., and three of unidentified beaked whales (Barlow 2010). All visual detections were also detected acoustically in real-time. An additional 65 acoustic-only encounters were classified as unidentified beaked whale encounters and an additional 13 acoustic-only encounters were classified as possible Baird's beaked whale. During post-processing of ORCAWALE 2008 recordings, it was determined via trials on recordings of confirmed beaked whale encounters that improved classification algorithms in PAMGUARD 1.9.01 would yield more accurate and complete detection results. ORCAWALE 2008 high-frequency recordings were post-processed in the PAMGUARD 1.9.01 Mixed Mode. This PAMGUARD mode allows click detections to be linked to GPS data collected during the survey for localization purposes. All recordings from this survey were processed using a standardized energy band comparison click detector, which works by comparing the acoustic energy in test and control frequency bands. The energy in the test band must exceed that in the control band by a threshold of a user-defined number of decibels in order to trigger a detection. The detection parameters (Tables 1 and 2) were designed to classify signals into unidentified detections, general beaked whale detections (including Cuvier's beaked whales), Baird's beaked whale detections, and transducer noise detections. In real-time, the PAMGUARD detection parameters were adjusted throughout the day depending on weather and cavitation-related noise conditions. During post-processing, standardized parameters, including a fixed detection threshold, were applied to all survey recordings. Standardized parameters were used to enforce constant acoustic detection effort across the dataset.

Analyzing PAMGUARD detection and classification output

Initial review of the automated detector output was conducted with reference to real-time visual and acoustic encounters. Histograms of automated detections for each day of data were generated in one, five, and ten-minute bins. Histograms of intervals between automated detections (approximating ICIs) from periods with known beaked whale encounters were also generated.

Verification of beaked whale encounters

In addition to the methods described above, a secondary filtering method was developed that relies on the spectral characteristics of beaked whale clicks. All automated detection data were divided into ten-minute bins, and bins with fewer than five or more than 1,000 automated beaked whale detections were eliminated from consideration. Large numbers of automated beaked whale click detections indicate false positives due to ambient noise or non-relevant odontocete species. Rainbow Click files generated by PAMGUARD, which store the waveform characteristics of all auto-detected clicks, were reviewed to examine spectral properties of clicks for all of the qualifying bins in Legs 1-4 of the ORCAWALE survey. Waveforms, spectral plots, ICIs and Wigner plots were examined using Rainbow Click, and beaked whale clicks with peak frequency, ICI, and upsweep characteristics matching those of published descriptions were classified as true detections (Figure 2).

Evaluating automated detector performance

To quantify the performance of the automated detector over the entire survey, a series of 23 ten-minute test periods were selected for closer examination. These consisted of ten periods in which detections of Cuvier’s beaked whales or unidentified beaked whales occurred in real-time, five in which detections of Baird’s beaked whales or possible Baird’s beaked whales occurred in real-time, and eight with no real-time beaked whale detections. For the periods with Cuvier’s and Baird’s beaked whales present, only periods containing 15 or more clicks in the ten-minute test period were analyzed with multiple evaluation methods. This criterion allowed the acoustician to confidently classify the test period as containing beaked whales. All 230 minutes of evaluation data were pooled to determine total performance scores for Cuvier’s and Baird’s beaked whale detectors. Because this detector will be applied to data from line-transect surveys, test periods were selected from segments of standard acoustic and visual survey effort. Standard line-transect segments are straight-line segments of consistent effort; therefore, non-beaked whale vocalizations from other species that were visually detected and cavitation noise due to ship turns were minimal during the test periods. The test periods were selected with reference to only the effort and encounter databases, so that PAMGUARD detector results could not influence the selection of test periods. The ten-minute test periods were almost always chosen to be the ten minutes immediately following the time of the first acoustic detection in order to maximize the number of true clicks included before the ship moved out of range of the vocalizing animal. Exceptions were made if the encounter database noted particularly strong or frequent clicks at a different time during the encounter. None of the test periods contained vocalizations of both Cuvier’s and Baird’s beaked whales.

The sound files for the 23 test periods were copied into folders, ordered randomly, and labeled only by arbitrary period number. An experienced acoustician (TMY) browsed each ten-minute period and logged the start times of all beaked whale clicks. A technician (EKJ) extracted the start time and species codes for all automated detections within each of the test periods from databases generated for each day of data processed in PAMGUARD. Automated detections were aligned with manual detections and each individual true beaked whale click or automated detection was assigned one of the following scores:

True Positive (TP): Both time and species classification matched between manual and automated beaked whale detections.

False Positive (FP): Time of an automated detection did not match a manual detection. No true beaked whale signal existed.

False Negative (FN): Time of a manual detection did not match to an automated detection. Detector failed to recognize a true beaked whale signal.

False Classification (FC): Time of an automated detection matches a manual detection, but the species classification was incorrect.

Non-Relevant Classification (NC): Automated detections with species classifications not of interest.

Calculating Measures of Detector Performance

In order to evaluate the performance of the automated detector in the test periods, the following measures were calculated for each test period:

Precision (P): probability that an automated detection will be true

$$P = \frac{TP}{TP + FP}$$

Recall (R): probability that a true signal will be automatically detected

$$R = \frac{TP}{TP + FN}$$

F-score (F): relationship between precision and recall

$$F = \frac{2P * R}{P + R}$$

Methods for Evaluating Detector Performance

In addition to the measures listed above, three different methods of data analysis were used to evaluate detector performance.

Method A: Timestamp Matching

In this method, timestamps were compared from manual and automated detections. In order to count as a true positive, both the time and the classification of the detection had to match. Only true positives, false positives, and false negatives were used in calculations, where false classifications were counted as false negatives. No measure of true negatives was made. Precision, recall, and F-scores were calculated.

Method B: No Classification Scheme

To evaluate the classification scheme used by the detector, in this method no distinction was made between automated detection classification codes. Manual and automated detections had to match only in time to count as a true positive, so in this method TP = TP + FC as calculated previously. Similarly, false positives were counted for all automated detection classification codes, so FP = FP + NC. Precision, recall, and F-scores were calculated. This method is expected to result in higher recall but lower precision.

Method C: Binary Classification

In the previous evaluation methods described, no measure of true negatives was made. In order to calculate true negatives, test periods were binned by one minute, and each minute was evaluated for presence or absence of manual and automated detections. True positives were bins containing both manual and automated detections, and true negatives were bins containing neither manual nor automated detections. False positives and false negatives were also scored for each one-minute bin. Based on this analysis, precision, recall, and F-scores were calculated. Since the number of clicks is not evaluated when using this method, all scores are expected to increase.

For methods requiring a match between the timestamps of manual and automated detections, timestamps were matched to within one second. This degree of flexibility was allowed because recording times for sound files are accurate to one second, and automated and manual detection logs both record event times as seconds elapsed from the start of the file. Additionally, this one-second allowance accounts for any error associated with manual selection of the signal start time.

If more than one automated detection was recorded within one second of a true signal, the closest automated detection was scored as a true positive or false classification. If the closest detection was scored as a false classification and another automated detection within one second was the correct classification, the correctly classified automated detection was scored as a true positive and the incorrectly classified automated detection was scored as a non-relevant classification.

Measures of precision, recall, and F-score were calculated for each Cuvier's beaked whale and Baird's beaked whale test period and each method, resulting in nine scores for each test period. Measures of performance were summarized across periods through calculation of mean and total scores. Total scores were calculated by summing data across the 230-minute sample dataset, incorporating data not reflected in the species-specific mean scores. This total score represents a slightly more accurate representation of detector performance than mean scores over the entire dataset, including periods with noise and non-target species.

Comparisons of detector performance across datasets

A second filtering method was devised to compare the ICIs of automated detections against the expected range of 0.2 – 0.5 s (Madsen et al. 2005, Johnson et al. 2006). The ICIs of automated detections within test periods were calculated in a moving window as the distance between each automated detection and all automated detections that followed within one second. This method was applied to recordings of Cuvier's beaked whales from both ORCAWALE and the 2009 Channel Islands survey. We used a kernel-density plot to illustrate the probability density function of the ICIs. We expected that in a kernel-density plot of the values generated, periods with true beaked whale signals would show probability density peaks at ICIs of approximately 0, 0.4, and 0.8 seconds – surface bounce, first following click, and second following click for Cuvier's beaked whales.

To determine whether noise conditions were a primary determinant of detector performance in the ORCAWALE dataset, the same PAMGUARD automated detector (settings in Table 1 and Table 2) was applied to recordings of Cuvier's beaked whales collected during the ORCAWALE and 2009 Channel Islands surveys, but the threshold for detection was varied between 10 dB and 20 dB in a series of trials. Due to differences in survey platforms, the ORCAWALE data were considered to have a qualitatively low signal-to-noise ratio (SNR) while the Channel Islands data had a qualitatively high SNR, so it was expected that detector performance would differ across the datasets.

Results

Summary

Real-time acoustic monitoring effort was carried out for 762 hours over the course of the survey, and 976 hours of automated high-frequency recordings were collected. Of these recordings, approximately 624 hours were collected during standard line-transect effort. 230 minutes of standard effort recordings were selected as a sample dataset with which to evaluate detector performance. Approximately 490 hours of click file recordings were analyzed in Rainbow Click to verify the presence of beaked whales using spectral properties of the clicks and confirming the presence of frequency upsweeps.

Reviewing output from the automated beaked whale detection algorithm

Histograms and histogram count data of automated detections generated for each day of data were initially intended to determine beaked whale presence or absence throughout the dataset. Detector performance was not adequate to be used with a fixed click count criterion to unequivocally determine beaked whale presence in the midst of other species and variable noise conditions. It is likely that using standardized detection parameters across all recordings greatly increased false detections triggered by noise. A histogram from a single day of beaked whale detections, 9/6/2008, is included as Figure 3 to illustrate that true beaked whale clicks were difficult to distinguish from background noise and non-target species.

A method for determining beaked whale presence based on ICI was evaluated. This method used the intervals between each automated detection and all of the automated detections that followed within one second. This filtering mechanism worked well on quiet datasets (Figure 4, top panels) in which peak intervals were visible at multiples of the expected ICI, but failed on noisy datasets (Figure 4, bottom panels) in which not enough true clicks were detected and true ICIs were obscured by noise. This method was discarded as a possible indicator of beaked whale presence in the ORCAWALE 2008 dataset.

Automated detector performance evaluation

Of the 23 test periods selected for examination, only 11 were ultimately included in the comparison of detector evaluation methods. Three of the periods selected for Cuvier's beaked whales and one of the periods selected for Baird's beaked whales contained too few clicks to be positively classified as containing beaked whale vocalizations. Of the blank periods containing no real-time detections, four were found to have no cetacean vocalizations present, one contained sperm whale vocalizations, one contained dolphin vocalizations, and two were noted to be particularly noisy. All of these periods not containing beaked whales are included as representatives of the variable noise conditions present in this survey. Automated and manual detection counts for all 23 test periods are included in Table 3. A summary of performance evaluation results for the nine Cuvier's beaked whale and four Baird's beaked whale periods selected for further analysis is included in Tables 4 and 5, respectively.

Results of Cuvier's beaked whale detector evaluation

For evaluation of Cuvier's beaked whale detections using *Method A: Timestamp Matching*, the automated detector achieved a mean precision of 0.37, a mean recall of 0.14, and a mean F-score of 0.17. When the classification scheme was eliminated and these periods were evaluated using *Method B: No Classification Scheme*, the precision score decreased to a mean precision of 0.17, while the recall score increased to a mean recall of 0.41. The mean F-score for Method B was 0.18. For *Method C: Binary Classification*, the mean precision was 0.80, the mean recall was 0.64, and the mean F-score was 0.67. Depending on the evaluation method used, precision scores for individual periods containing Cuvier's beaked whales ranged from 0.04 to 1.00, recall scores ranged from 0.01 to 1.00, and F-scores ranged from 0.03 to 0.86. *Method A* produced the lowest F-scores while *Method C* produced the highest. Applying *Method A* over the entire 230-minute sample dataset, the Cuvier's beaked whale detector had a precision of 0.07, a recall of 0.07, and an F-score of 0.07.

Results of Baird's beaked whale detector evaluation

Using *Method A: Timestamp Matching* on Baird's beaked whale periods, the automated detector had a mean precision of 0.37, a mean recall of 0.07, and a mean F-score of 0.12. For *Method B: No Classification Scheme*, the mean precision was 0.27, the mean recall was 0.37, and the mean F-score was 0.31. *For Method C: Binary Classification*, the mean precision was 0.82, the mean recall was 0.76, and the mean F-score was 0.78. Precision scores for periods containing Baird's beaked whales ranged from 0.12 to 1.00, recall scores ranged from 0.13 to 1.00, and F-scores ranged from 0.17 to 1.00. *Method A* produced the lowest F-scores while *Method C* produced the highest. Applying *Method A* over the entire 230-minute sample dataset, the Baird's beaked whale detector had a precision of 0.01, a recall of 0.16, and an F-score of 0.02.

Detection of beaked whale groups

During post-processing, 90% of possible Baird's beaked whale acoustic encounters were manually verified and 85% of unidentified beaked whale acoustic encounters were manually verified and confirmed to be true beaked whale detections. All of the verified beaked whale encounters from the ORCAWALE 2008 survey were detected using PAMGUARD's automated classifier. A total of 84 beaked whale encounters were identified.

Discussion

The automated beaked whale detector configured using PAMGUARD software at the SWFSC is intended to detect and classify beaked whales both in real-time and in post-processing of recordings collected using a towed hydrophone array. The application of a standardized automated detector across all recordings from the ORCAWALE 2008 survey in conjunction with manual confirmation of detections will be used to build models of beaked whale distribution and habitat use in the California Current Ecosystem. While the automated detector configuration was optimized on sample datasets from the ORCAWALE 2008 cruise, it failed to clearly identify periods of beaked whale presence in the survey as a whole during post-processing. This was primarily due to the difficulty of assigning a standardized threshold parameter for application to the entire recording dataset. Extreme noise conditions in the recordings combined with variable sea states and the presence of non-beaked whale odontocete species made the results of the automated detector difficult to interpret. Histograms of automated detections over single days of data failed to clearly distinguish beaked whale encounters from encounters with other species (Figure 3). Therefore, it was necessary to manually verify the presence of beaked whale clicks. Future work will aim to optimize detectors to minimize the need for this intensive manual review. In our assessment of the performance of this automated beaked whale detector over recordings from the entire survey, we explored different methods for evaluating detector performance.

During initial configuration of this PAMGUARD beaked whale detector, various settings were tested on sample recordings from ORCAWALE. Detector performance was evaluated iteratively based on click count comparisons between manual and automated detections over one- and ten-minute periods. As shown in Table 3, matching click counts between manual and automated detections is not a reliable measure of performance in noisy datasets. Evaluation *Method A: Timestamp Matching* showed that absolute click counts can be deceiving. This method produced the lowest detector performance scores. Timestamp matching requires precision in aligning manual and automated click detections to avoid drift and thorough examination of any possible sources of alignment error. We assume that time is continuous between sound files; in reality there may be dropped samples that cause drift in timestamp alignment. In general, software programs record the manual or automated detection time from the beginning of the recording, so any timestamp error is introduced via the manual selection of signals. In spite of these challenges, using timestamp matching to evaluate detector performance gives an accurate picture of exactly what is triggering the automated detector and allows users to address any problems with false or missed detections.

Method B: No Classification Scheme was undertaken primarily to separate the performance of the detection algorithm from the classification algorithm. This method produced lower precision and higher recall scores than timestamp matching. For our application, an increase in recall is not worth the decrease in precision of the detector. The results of evaluation *Method C: Binary Classification* show that binning recording data by one minute likely produces inflated measures of performance. In this study, binning data by one minute obscured the high numbers of false positives and false negatives present in the data. Binning data may be appropriate for analyzing recordings with little extraneous noise, but it is not a precise method for comparing the performance of different automated detection methods across variable recording datasets. As was mentioned in the introduction, binning data by one second is another possible evaluation technique but would also be prone to the problem of multiple detections within a single second. For our purposes, timestamp matching in a precision-recall framework provides the most accurate and useful description of detector performance.

Previous authors have suggested using two criteria for determining beaked whale presence, the first based on spectral characteristics of the click, and the second based on the ICI of clicks detected (Zimmer et al. 2008). When this method was applied to recordings of Cuvier's beaked whale with a qualitatively high SNR from the 2009 Channel Islands survey, the probability density values for ICIs were indeed highest at the values expected for this species (Figure 4, top panels). However, when the same method was applied to one of the ORCAWALE test periods for this species, no such peaks in probability density at expected ICIs were present (Figure 4, bottom panels). The high proportion of missed detections (up to 97% for test periods) meant that the ICI algorithm did not have enough true data points to accurately construct an ICI histogram. A comparison of algorithm performance across several thresholds was conducted for the same high and low SNR test periods (Figure 5) and showed that the high SNR Channel Islands dataset was much more responsive to changes in the detector threshold. The low SNR dataset suffered from consistently high rates of missed detections across thresholds due to noise masking, while the false positive rate varied with the detector threshold. The extreme noise conditions in this low SNR ORCAWALE dataset presented challenges both to detector development and evaluation, highlighting the need to enhance recording quality through improvements to hardware and software design in future surveys.

Results from the manual verification of spectral upsweeps indicate that all real-time encounters of beaked whales were detected using a combination of manual and automated methods. This suggests that while the precision and recall scores of the detector are fairly low for individual clicks, groups of beaked whales were not missed using these automated post-processing techniques. In our study, the output of the automated detector served to identify regions of the dataset for closer examination. For population assessment and habitat modeling it is crucial to eliminate all false positive encounters of groups of beaked whales. Due to the low precision rate of the automated detector, it was necessary to manually confirm beaked whale presence in periods containing automated beaked whale detections. This method of filtering automated detection output is time consuming; however, requiring automated detections to be manually verified will prevent false detections from entering the dataset. With the integration of automated and manual methods used in this study, we have achieved a precision approaching one and are confident that beaked whale encounters that will be included in final assessments and future analyses represent true beaked whales.

Ideally, automated detectors for cetacean vocalizations should be evaluated using a subset of the recording dataset large enough that the values for precision and recall can converge on true values. The wide range of precision and recall values calculated from our evaluation dataset indicate that these scores were likely not stable and that a larger subset of data is required to evaluate detector performance. Using similar evaluation methods, Leary (2011) found that precision and recall values did not stabilize for a particular mysticete detector until 800 minutes of sample data had been evaluated. The amount of evaluation data required to arrive at stable precision and recall values will depend on the noise variability of the recording dataset. Recording datasets with low levels of extraneous noise and non-target species vocalizations may require very little evaluation data to arrive at stable precision and recall values compared to recording datasets with high noise levels. Our 230-minute sample dataset was likely not extensive enough to accurately quantify the performance of our automated detection algorithms. Because beaked whale vocalizations are so rare, we biased the sample dataset towards periods with known beaked whale presence, which likely inflated estimates of precision and recall. The contrast in F-scores across evaluation methods and between periods of target and non-target periods confirms that detector evaluation metrics must be carefully chosen to match the goals of a particular study. A more robust method for evaluating these automated detectors would use a large enough quantity of randomly subsampled data to capture true beaked whale vocalizations and also accurately represent the variety of noise conditions present in the dataset.

Conclusion

The comparison of performance evaluation methods for automated beaked whale detectors shows that binning data in order to generate the scores required in ROC methodology likely does not produce the most accurate measures of detector performance. By comparing noisy and quiet datasets, we show how detector performance scores can vary with recording quality. To allow for effective detector comparison we recommend that automated beaked whale detectors developed and used at different research institutions be evaluated with a standardized performance metric. From our comparison of performance measures, we recommend matching timestamps between manual and automated beaked whale detections and applying the precision-recall framework to summarize results. The F-score does not require a measure of true negatives and thus we recommend it as the most appropriate measure of performance for beaked whale detection applications where true signals are rare. Graphical representations (such as plots of precision versus recall or proportion missed versus proportion false) can be used to evaluate threshold-dependent detector performance. Future work in the automation of beaked whale detection should aim to minimize manual effort required to verify beaked whale detections. Guidelines for the amount of evaluation data required to produce stable precision and recall scores for automated detectors would be helpful to researchers when developing and implementing automated detectors. In the future, we will continue efforts to improve automated beaked whale detection techniques. In order to produce abundance estimates using this dataset, all manually verified beaked whale encounters will be re-localized. Our methods maximized the precision of our acoustic beaked whale identifications and will allow these data to be accurately used for abundance estimation and habitat modeling in subsequent analyses.

Acknowledgements

We would like to thank all of the participants in the ORCAWALE 2008 survey, including the survey coordinator, Annette Henry, marine mammal observers Jim Cotton, Chris Cutler, Mark Deakos, Jim Gilpatrick, Allan Ligon, Rich Pagen, Richard Rowlett, and Suzanne Yin, the acousticians Tina Yack and Shannon Rankin, the cruise leaders Jeremy Rusin, Jim Carretta, and Lisa Ballance, and the crew and officers of the NOAA R/V McArthur II. This manuscript benefited greatly from reviews by Karin Forney, Julie Oswald, and Bill Perrin. Shannon Rankin also provided valuable guidance on direction of this project. Thank you also to Doug Gillespie for developing the PAMGUARD software framework and detector modules used in this study.

References

Barlow, J. 1999. Trackline detection probability for long-diving whales. Garner et al. (eds). In: Marine Mammal Survey and Assessment Methods. Balkema, Rotterdam, Netherlands.

Barlow, J. 2010. Cetacean abundance in the California Current estimated from a 2008 ship-based line-transect survey. U.S. Department of Commerce, NOAA Technical Memorandum NMFS, NOAA-TM-NMFS-SWFSC-456, 19 pp.

Cox, T. M., T. J. Ragen, A. J. Read, E. Vos, R. W. Baird, K. Balcomb, J. Barlow, J. Caldwell, T. Cranford, L. Crum, A. D'Amico, G. D'Spain, A. Fernandez, J. Finneran, R. Gentry, W. Gerth, F. Gulland, J. Hildebrand, D. Houser, T. Hullar, P. D. Jepson, D. Ketten, C. D. MacLeod, P. Miller, S. Moore, D. C. Mountain, D. Palka, P. Ponganis, S. Rommel, T. Rowles, B. Taylor, P. Tyack, D. Wartzok, R. Gisiner, J. Mead, and L. Benner. 2006. Understanding the impacts of anthropogenic sound on beaked whales. Journal of Cetacean Resource Management **7**:177-187.

Dalebout, M. L., C. S. Baker, J. G. Mead, V. G. Cockcroft, and T. K. Yamada. 2004. A comprehensive and validated molecular taxonomy of beaked whales, family Ziphiidae. Journal of Heredity **95**:459-473.

Davis, J. and M. Goadrich. 2006. The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Dawson, S., J. Barlow, and D. Ljungblad. 1998. Sounds recorded from Baird's beaked whale, *Berardius bairdii*. Marine Mammal Science **14**:335-344.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters **27**:861-874.

Gillespie, D., D. K. Mellinger, J. Gordon, D. McLaren, P. Redmond, R. McHugh, P. Trinder, X.-Y. Deng, and A. Thode. 2009a. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans. The Journal of the Acoustical Society of America **125**:2547.

Gillespie, D., C. Dunn, J. Gordon, D. Claridge, C. Embling, and I. Boyd. 2009b. Field recordings of Gervais' beaked whales *Mesoplodon europaeus* from the Bahamas. Journal of the Acoustical Society of America **125**:3428-3433.

Johnson, M., P. T. Madsen, W. M. X. Zimmer, N. A. d. Soto, and P. L. Tyack. 2004. Beaked Whales Echolocate on Prey. Proceedings: Biological Sciences **271**:S383-S386.

Leary, D., X. Mouy, J. Oswald, B. Martin, and D. Hannay. 2011. Deriving call counts from automated classifiers in challenging noise conditions. Fifth International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals Using Passive Acoustics, Mt. Hood, OR.

Marques, T. A., L. Thomas, J. Ward, N. DiMarzio, and P. Tyack. 2009. Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales. Journal of the Acoustical Society of America **125**:1982-1994.

Mellinger, D. K. 2001. ISHMAEL 1.0 User's Guide. U.S. Department of Commerce, NOAA Technical Memorandum NMFS, OAR PMEL-120.

Mellinger, D. K. 2004. A comparison of methods for detecting right whale calls. Canadian Acoustics **32**:55-65.

Mellinger, D. K. and C. W. Clark. 2000. Recognizing transient low-frequency whale sounds by spectrogram correlation. Journal of the Acoustical Society of America **107**:3518-3529.

Mellinger, D. K., K. M. Stafford, and C. G. Fox. 2004. Seasonal occurrence of sperm whale (*Physeter macrocephalus*) sounds in the Gulf of Alaska, 1999-2001. Marine Mammal Science **20**:48-62.

Mellinger, D. K., K. M. Stafford, S. E. Moore, R. P. Dziak, and H. Matsumoto. 2007. An overview of fixed passive acoustic observation methods for cetaceans. Oceanography **20**:36-45.

Moretti, D., T. A. Marques, L. Thomas, N. DiMarzio, A. Dilley, R. Morrissey, E. McCarthy, J. Ward, and S. Jarvis. 2010. A dive counting density estimation method for Blainville's beaked whale (Mesoplodon densirostris) using a bottom-mounted hydrophone field as applied to a Mid-Frequency Active (MFA) sonar operation. Applied Acoustics **71**:1036-1042.

Munger, L. M., D. K. Mellinger, S. M. Wiggins, S. E. Moore, and J. A. Hildebrand. 2005. Performance of spectrogram cross-correlation in detecting right whale calls in long-term recordings from the Bering Sea. Canadian Acoustics **33**:25-34.

Rankin, S., J. Barlow, J. Oswald, and L. Ballance. 2008. Acoustic studies of marine mammals during seven years of combined visual and acoustic line-transect surveys for cetaceans in the Eastern and Central Pacific Ocean. U.S. Department of Commerce, NOAA Technical Memorandum NMFS, NOAA-TM-NMFS-SWFSC-429, 58pp.

Swets, J. A., Dawes, R. M., & Monahan, J. 2000. Better decisions through science. Scientific American **Oct. 2000**: 83-87.

Tyack, P., M. Johnson, N. Aguilar Soto, A. Sturlese, and P. T. Madsen. 2006. Extreme diving of beaked whales. The Journal of Experimental Biology **209**:4238-4253.

Ward, J., S. Jarvis, D. Moretti, R. Morrissey, N. DiMarzio, M. Johnson, P. Tyack, L. Thomas, and T. Marques. 2011. Beaked Whale (*Mesoplodon densirostris*) Passive acoustic detection in increasing ambient noise. Journal of the Acoustical Society of America **129**:662-669.

Ward, J., R. Morrissey, D. Moretti, N. DiMarzio, S. Jarvis, M. Johnson, P. Tyack, and C. White. 2008. Passive acoustic detection and localization of *Mesoplodon densirostris* (Blainville's beaked whale) vocalizations using distributed bottom-mounted hydrophones in conjunction with a digital tag (DTag) reocrding. Canadian Acoustics **36**:60-66.

Yack, T. M., J. Barlow, S. Rankin, and D. Gillespie. 2009. Testing and validation of automated whistle and click detectors using PAMGUARD 1.0. U.S. Department of Commerce, NOAA Technical Memorandum NMFS, NOAA-TM-NMFS-SWFSC-443, 55 pp.

Yack, T. M., J. Barlow, M. A. Roch, H. Klinck, S. Martin, D. K. Mellinger, and D. Gillespie. 2010. Comparison of beaked whale detection algorithms. Applied Acoustics **71**:1043-1049.

Yack, T. M., Barlow, J., Calambokidis, J., Ballance, L., Pitman, R., and McKenna, M. 2011. Passive Acoustic Beaked Whale Monitoring Survey of the Channel Islands, California. U.S. Department of Commerce, NOAA Technical Memorandum NMFS, NOAA-TM-NMFS-SWFSC-479, 26 pp.

Zimmer, W., & Pavan, G. 2008. Context dependent detection/classification of Cuvier's beaked whale (Ziphius cavirostris). In: New Trends for Environmental Monitoring Using Passive Systems.

Zimmer, W. M. X., J. Harwood, P. L. Tyack, M. P. Johnson, and P. T. Madsen. 2008. Passive acoustic detection of deep-diving beaked whales. The Journal of the Acoustical Society of America **124**:2823-2832.

Zimmer, W. M. X., M. P. Johnson, P. T. Madsen, and P. L. Tyack. 2005. Echolocation clicks of free-ranging Cuvier's beaked whales (*Ziphius cavirostris*). The Journal of the Acoustical Society of America **117**:3919-3927.

Tables

Table 1: Click detection settings used in PAMGUARD software to post-process all survey recordings

Click Detection Parameters

Menu Item	Field	Value
Source	Raw Data Source	Raw input data from Sound Acquisition
	Auto Grouping	One group
	Channel	Channel 0, Channel 1
Trigger	Threshold	18.0 dB
	Long Filter	0.00001000, Ch 0
	Long Filter 2	0.00000100, Ch 1
	Short Filter	0.10000000
Click Length	Min Click Separation	100 samples
	Max Click Length	1024 samples
	Pre Sample	40 samples
	Post Samples	0 samples
Noise	Create Sample Noise Measurements	Yes
	Interval	5.0 s

Table 2: Click classification settings used in PAMGUARD software to post-process all survey recordings

Classifier Parameters

	Menu Item	Field	Value
12K Transducer	**General**	Unique Code	3
		Symbol	Green Square
		Channel Options	Require positive identification on only one channel
		Restrict Parameter Extraction To	False
	Click Length	Enable	False
	Energy Bands	Enable	True
		Test Band	11000.0 to 12800.0 Hz
		Control Band	15000.0 to 45000.0 Hz, 3.0 dB Threshold
		Control Band	60000.0 to 80000.0 Hz, 3.0 dB Threshold
	Peak and Mean Frequency	Search and Integration Range	1000.0 to 90000.0 Hz, Smoothing 5 bins
		Peak Frequency Enable	True, 11000.0 to 12800.0 Hz
		Peak Width Enable	False
		Mean Frequency Enable	False
	Zero Crossings	Enable	False
Beaked Whale	**General**	Unique Code	1
		Symbol	Orange Diamond
		Channel Options	Require positive identification on only one channel
		Restrict Parameter Extraction To	False
	Click Length	Enable	False
	Energy Bands	Enable	True
		Test Band	32000.0 to 50000.0
		Control Band	12000.0 to 22000.0 Hz, 3.0 dB Threshold
		Control Band	140000.0 to 145000.0 Hz, 3.0 dB Threshold
	Peak and Mean Frequency	Search and Integration Range	20000.0 to 95000.0 Hz, Smoothing 5 bins
		Peak Frequency Enable	True, 32000.0 to 50000.0 Hz
		Peak Width Enable	False
		Mean Frequency Enable	False
	Zero Crossings	Enable	False
Berardius	**General**	Unique Code	2
		Symbol	Blue Diamond
		Channel Options	Require positive identification on only one channel
		Restrict Parameter Extraction To	False
	Click Length	Enable	False
	Energy Bands	Enable	True
		Test Band	11000.0 to 21000.0 Hz
		Control Band	70000.0 to 80000.0 Hz, 3.0 dB Threshold
		Control Band	140000.0 to 145000.0 Hz, 3.0 dB Threshold
	Peak and Mean Frequency	Search and Integration Range	10000.0 to 90000.0 Hz, Smoothing 5 bins
		Peak Frequency Enable	11000.0 to 21000.0 Hz
		Peak Width Enable	False
		Mean Frequency Enable	False
	Zero Crossings	Enable	False

Table 3: Manual and automated detection counts for all test periods. Cuvier's beaked whale and Baird's beaked whale periods containing fewer than 15 manual detections (Periods Zc_01, Zc_07, Zc_10, Bb_02) were discarded from further analysis.

Species	Period	Manual Sp. Detections	Automated Detection Counts				
			Unidentified	Beaked Whale	Berardius	12 kHz Transducer	Total
Cuvier's beaked whale	Zc_01	11	5	0	4	5	14
	Zc_02	108	32	8	59	4	103
	Zc_03	67	29	9	18	0	56
	Zc_04	78	235	21	179	22	457
	Zc_05	73	5	0	4	0	9
	Zc_06	17	63	21	99	10	193
	Zc_07	14	7	2	12	0	21
	Zc_08	33	14	2	14	2	32
	Zc_09	16	33	6	33	5	77
	Zc_10	0	24	5	31	4	64
Baird's beaked whale	Bb_01	19	9	1	12	0	22
	Bb_02	7	1	2	0	0	3
	Bb_03	68	55	12	76	7	150
	Bb_04	46	12	1	24	4	41
	Bb_05	23	12	1	16	2	31
Blank: no beaked whale vocalizations	Na_01	Noise	161	29	235	36	461
	Na_02	Blank	435	33	279	22	769
	Na_03	Sperm Whales	771	3	950	231	1955
	Na_04	Noise	77	6	124	12	219
	Na_05	Blank	23	6	62	21	112
	Na_06	Blank	3	1	1	0	5
	Na_07	Delphinid	426	154	46	0	626
	Na_08	Blank	438	74	663	109	1284

Table 4: Detection and classification performance scores calculated for periods containing Cuvier's beaked whale

	Method A: Timestamp Matching			Method B: No Classification Scheme			Method C: Binary Classification		
Period	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Zc_02	0.80	0.04	0.10	0.53	0.39	0.45	1.00	0.50	0.67
Zc_03	0.11	0.01	0.03	0.18	0.15	0.16	1.00	0.40	0.57
Zc_04	0.29	0.08	0.12	0.10	0.62	0.18	0.67	0.86	0.75
Zc_05	0.14	0.01	0.03	0.17	0.05	0.08	1.00	0.40	0.57
Zc_06	0.43	0.53	0.47	0.07	0.76	0.12	0.67	1.00	0.80
Zc_08	0.29	0.15	0.20	0.04	0.24	0.08	0.86	0.86	0.86
Zc_09	0.50	0.19	0.27	0.13	0.63	0.22	0.40	0.50	0.44
Mean	**0.37**	**0.14**	**0.17**	**0.17**	**0.41**	**0.18**	**0.80**	**0.64**	**0.67**

Table 5: Detection and classification performance scores calculated for periods containing Baird's beaked whale

	Method A: Timestamp Matching			Method B: No Classification Scheme			Method C: Binary Classification		
Period	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Bb_01	0.42	0.26	0.32	0.41	0.47	0.44	0.86	0.86	0.86
Bb_03	0.12	0.13	0.12	0.18	0.41	0.25	1.00	1.00	1.00
Bb_04	0.25	0.13	0.17	0.29	0.26	0.27	0.67	0.57	0.62
Bb_05	0.30	0.26	0.28	0.21	0.35	0.26	0.75	0.60	0.67
Mean	**0.27**	**0.20**	**0.22**	**0.27**	**0.37**	**0.31**	**0.82**	**0.76**	**0.78**

Figures

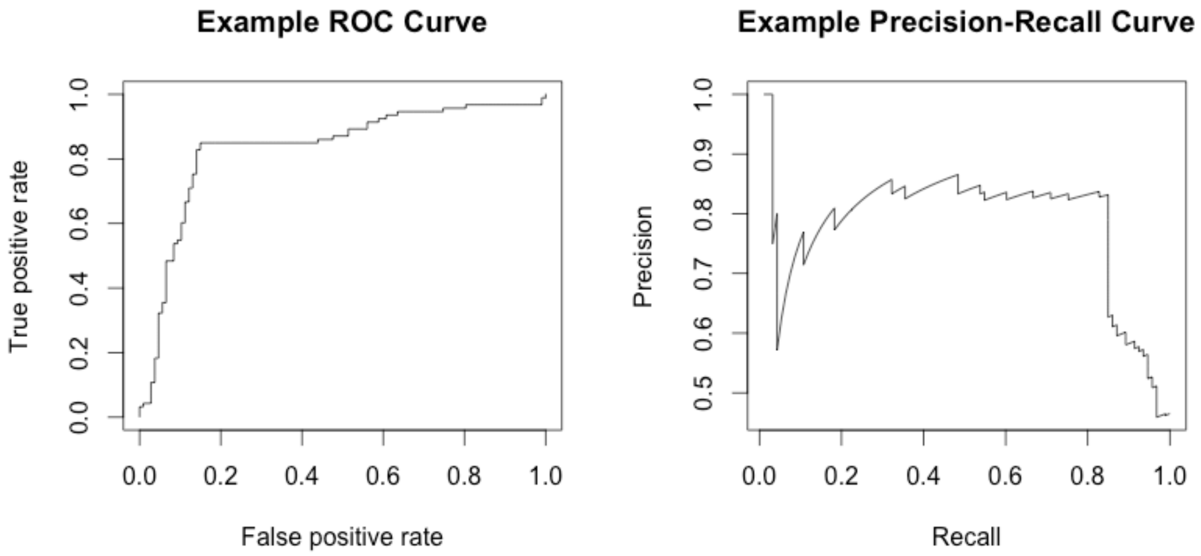

Figure 1: Examples of ROC and Precision-Recall curves generated the ROCR.simple dataset in R. These example plots demonstrate that a detector may appear to perform differently depending on which evaluation metrics are used.

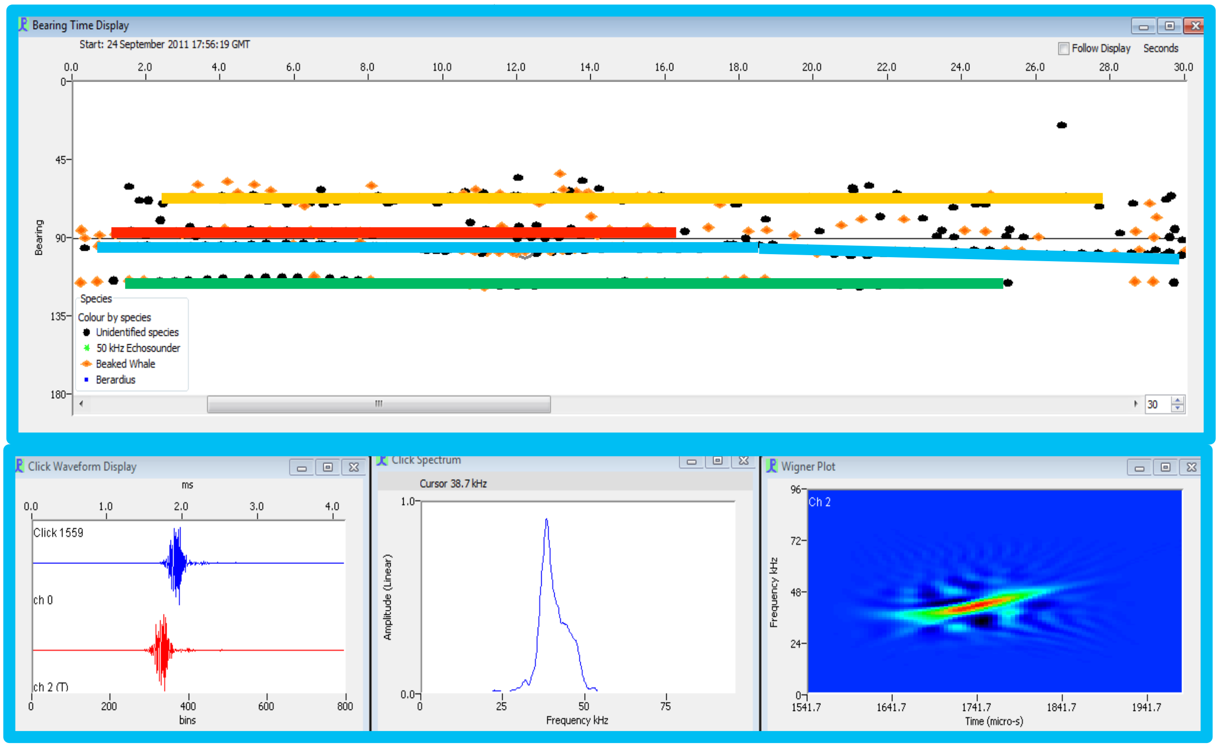

Figure 2: PAMGUARD click detection display showing an example Cuvier's beaked whale acoustic encounter. Colored lines in the top panel represent bearing angle trajectories for individual beaked whales. The lower left panel displays waveforms for signals received from each channel. The middle panel illustrates peak frequency, and the right panel is a Wigner plot showing a an upswept signal characteristic of beaked whale clicks.

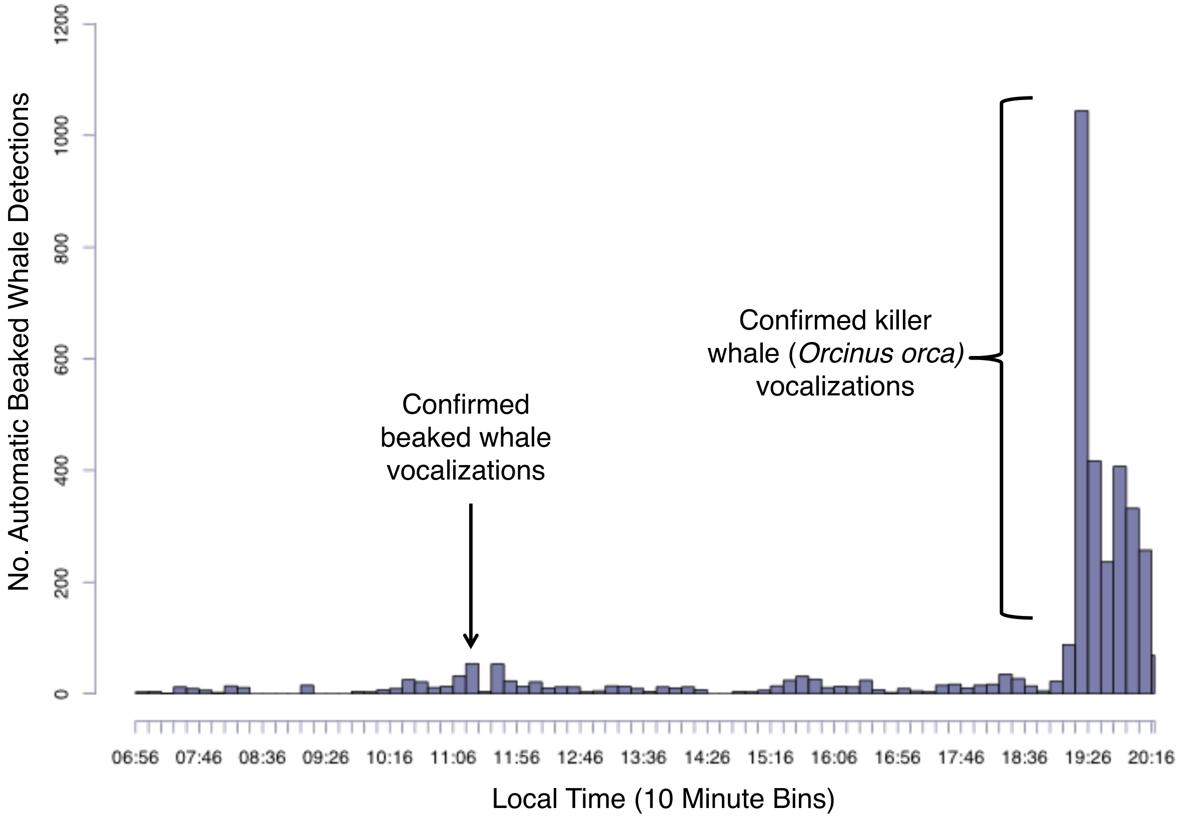

Figure 3: Histogrammed output of the PAMGUARD automated beaked whale detector over one day, 9/6/2008, illustrating the difficulty of extracting true beaked whale signals from background noise and the vocalizations of non-target species.

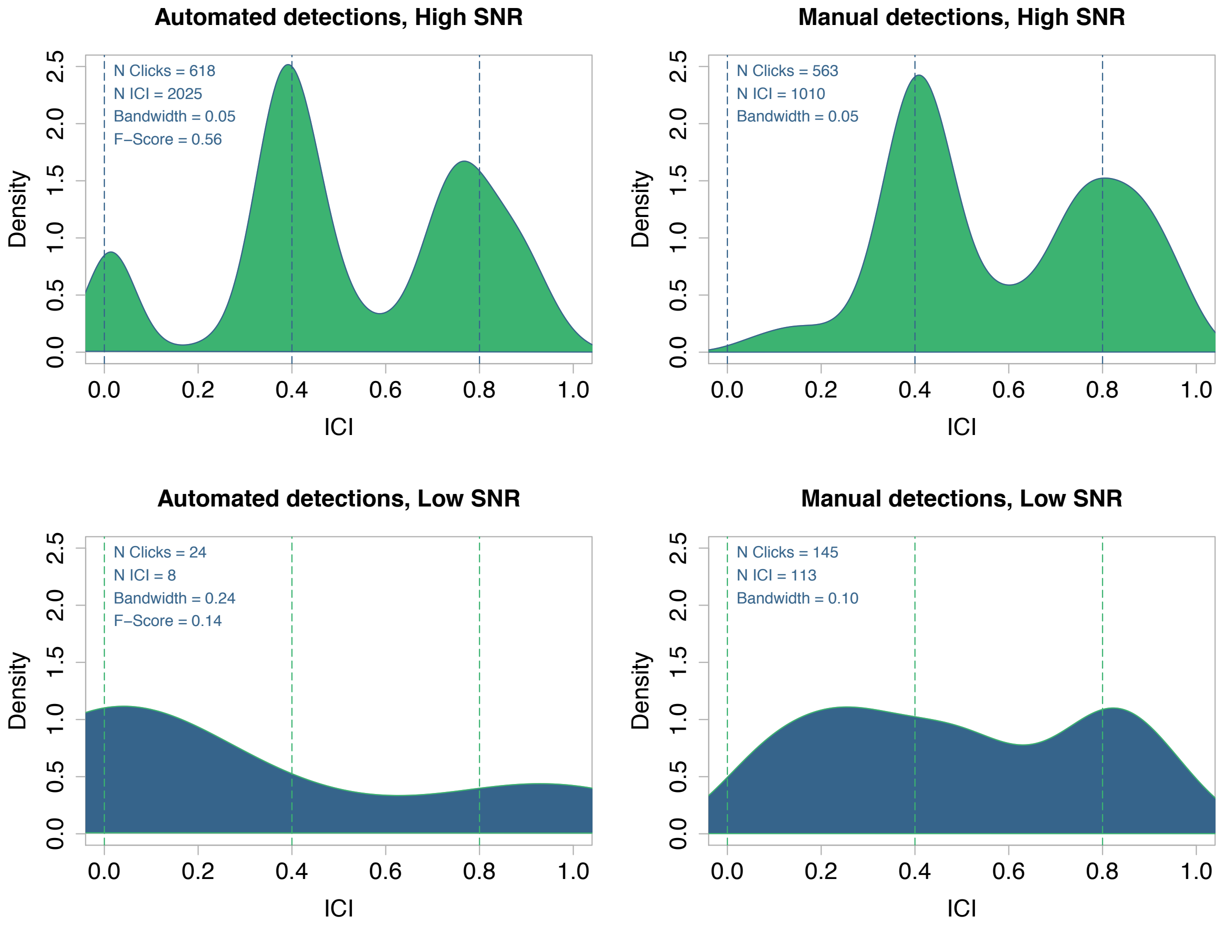

Figure 4: Kernel density plots of Cuvier's beaked whale click detection ICIs from a high SNR 2009 towed array recording (top two panels) and from a low SNR 2008 towed array recording (bottom two panels). The ICIs calculated from automated and manual click detections are displayed in the left and right columns, respectively. ICIs were calculated for a ten-minute period by determining the time difference between each automated detection and all automated detections that followed within one second. The dashed lines demonstrate the expected locations of click surface bounce, first following click, and second following click. There is a clear ICI pattern visible in the high SNR dataset and no clear ICI pattern visible in the low SNR dataset.

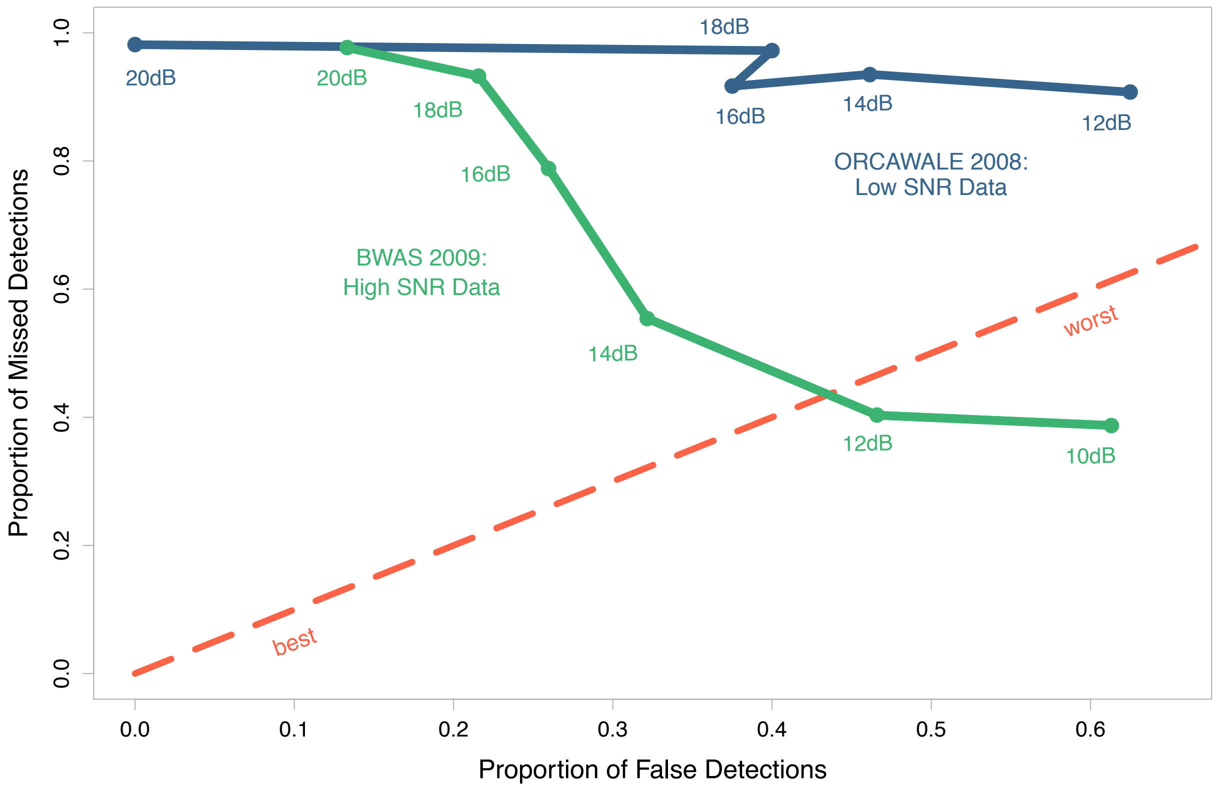

Figure 5: Relationship between missed and false automated detections of beaked whale clicks. The blue curve represents automated detector performance with a variable detection threshold on a 2008 towed array recording of a beaked whale with a low SNR. The green curve represents automated detector performance with a variable detection threshold on a 2009 towed array recording of a beaked whale with a high SNR. The optimal detector would score at the origin. Masking of true signals in the low SNR recordings results in high proportions of missed detections across the variable detection threshold.

RECENT TECHNICAL MEMORANDUMS

SWFSC Technical Memorandums are accessible online at the SWFSC web site (http://swfsc.noaa.gov). Copies are also available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161 (http://www.ntis.gov). Recent issues of NOAA Technical Memorandums from the NMFS Southwest Fisheries Science Center are listed below:

NOAA-TM-NMFS-SWFSC-499 Predictive modeling of cetacean densities in the California Current Ecosystem based on summer/fall ship surveys in 1991-2008.
E.A. BECKER, K.A. FORNEY, M.C. FERGUSON, J. BARLOW and J.V. REDFERN
(October 2012)

500 Marine mammal and seabird bycatch in California gillnet fisheries in 2011.
J.V. CARRETTA and L. ENRIQUEZ
(December 2012)

501 Assessment of the Pacific sardine resource in 2012 for U.S. management in 2013.
K.T. HILL, P.R. CRONE, N.C.H. LO, D.A. DEMER, J.P. ZWOLINSKI, and B.J. MACEWICZ
(December 2012)

502 Upper Klamath and Trinity River Chinook salmon Biological Review Team report.
T.H. WILLIAMS, J.C. GARZA, N.J. HETRICK, S.T. LINDLEY, M.S. MOHR, J.M. MYERS, M.R. O'FARRELL, R.M. QUINONES, and D.J. TEEL
(December 2012)

503 Proceedings of the National Marine Fisheries Service Productivity Workshop, Santa Cruz, California, June 11-12, 2012.
A.T. MAMULA and J.B. WALDEN
(December 2012)

504 U.S. Pacific Marine Mammal Stock Assessments: 2012.
J.V. CARRETTA, K.A. FORNEY, E. OLESON, K. MARTIEN, M.M. MUTO, M.S. LOWRY, J. BARLOW, J. BAKER, B. HANSON, D. LYNCH, L. CARSWELL, R.L. BROWNELL JR., J. ROBBINS, D.K. MATTILA, K. RALLS, and M.C. HILL
(January 2013)

505 Spawning biomass of Pacific Sardine (*Sardinops sagax*) off U.S. in 2012.
LO, N.C.H., B.J. MACEWICZ, AND D.A. GRIFFITH
(March 2013)

506 Probability of taking a western North Pacific gray whale during the postponed Makah hunt.
J. E. MOORE, and D. W. WELLER
(March 2013)

507 Report of the National Marine Fisheries Service gray whale stock identification workshop.
D. W. WELLER, S. BETTRIDGE, R. L. BROWNELL JR., J. L. LAAKE, J. E. MOORE, P. E. ROSEL, B. L. TAYLOR, and P. R. WADE
(March 2013)

508 Inferring trackline detection probabilities from differences in apparent densities of beaked whales and dwarf & pygmy sperm whales in different survey conditions.
J. BARLOW
(April 2013)