Taylor & Francis
Taylor & Francis Group

Bioacoustics
The International Journal of Animal Sound and its Recording

ISSN: 0952-4622 (Print) 2165-0586 (Online) Journal homepage: http://www.tandfonline.com/loi/tbio20

Variability in the performance of the spectrogram correlation detector for North-east Pacific blue whale calls

Ana Širović

To cite this article: Ana Širović (2016) Variability in the performance of the spectrogram correlation detector for North-east Pacific blue whale calls, Bioacoustics, 25:2, 145-160, DOI: 10.1080/09524622.2015.1124248

To link to this article: http://dx.doi.org/10.1080/09524622.2015.1124248

Published online: 28 Dec 2015.

BIOACOUSTICS, 2016
VOL. 25, NO. 2, 145–160
thttp://dx.doi.org/10.1080/09524622.2015.1124248

Taylor & Francis
Taylor & Francis Group

Variability in the performance of the spectrogram correlation detector for North-east Pacific blue whale calls

Ana Širović

Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA

ABSTRACT

Spectrogram correlation has been used successfully for automatic detection of baleen whale calls. However, applying this method consistently to long time series can be challenging. To illustrate the potential challenges of the automatic detection process, recordings collected in the Southern California Bight between 2007 and 2012 were used for detection of North-east Pacific blue whale (*Balaenoptera musculus*) B calls. The effects of the following factors were investigated: blue whale B call frequency shift and appropriate kernel modification, seasonal variability in call abundance, analyst variability and noise. Due to intra- and inter-annual changes in the call frequency of blue whale B calls, seasonal and annual adjustments to the call detection kernel were needed. To account for seasonal variability in call production, evaluation of the detector against ground truth data was performed at multiple times during the year. Analyst variability did not affect overall long-term trends in detection, but it had an impact on the total number of detections, as well as call rate estimation. Noise, particularly from shipping, was negatively correlated with detections at hourly time scales. A detailed analysis of variability in the performance of spectrogram correlation detectors should be performed when applying this method to long-term acoustic data-sets.

ARTICLE HISTORY
Received 6 April 2015
Accepted 20 November 2015

KEYWORDS
Spectrogram correlation; blue whale calls; *Balaenoptera musculus*; call frequency shift

Introduction

Continued improvements in data acquisition raise the need for automatic detection methods to extract calls from exceedingly large acoustic data-sets. Manual picking of calls from these data is prohibitively slow and costly. Additionally, automatic detectors have an advantage over human analysts as they exhibit a relatively constant, quantifiable bias. In recent years, a variety of methods have been developed to facilitate the automatic detection of baleen whale calls, including artificial neural networks (Potter et al. 1994), spectrogram correlation (Mellinger and Clark 2000), energy detection (Širović et al. 2004), dynamic time warping (Mouy et al. 2009), generalized power law (Helble et al. 2012), etc. This variety of methods has been necessary because different detection algorithms lend themselves to the successful detection of different signal types.

CONTACT Ana Širović asirovic@ucsd.edu

A method that has been used successfully for automatic detection of stereotyped baleen whale calls is spectrogram correlation (Mellinger and Clark 2000; Munger et al. 2008). As described by Mellinger and Clark (1997), synthetic-kernel spectrogram correlation is a method that preforms two-dimensional correlation between a kernel's time-frequency image and a spectrogram. The kernel can be defined as multiple frequency modulated segments. The kernel is modified using a hat function, which creates a positive central lobe and negative flanks for the kernel (Mellinger and Clark 1997). This allows for good performance not just in white noise, but also under more realistic conditions encountered in the low-frequency range of the ambient sound spectrum.

Spectrogram correlation has been commonly applied to the detection of long-duration, stereotyped blue whale (*Balaenoptera musculus*) calls in long-term data-sets (Širović et al. 2004; Stafford et al. 2004, 2005; Wiggins et al. 2005; Oleson, Wiggins, et al. 2007; Samaran et al. 2013; Širović et al. 2015). However, when describing the performance of this detector in publications, the variability intrinsic to the automatic detection process has rarely been described or quantified. While some authors did not quantitatively describe detector performance (Stafford et al. 2004, 2005), others reported only false alarm rate (Samaran et al. 2013). Even when both false alarm and missed call rates were reported, details of how they were tested were not always clear (Širović et al. 2004; Wiggins et al. 2005). The most thorough evaluation was reported by Oleson, Wiggins, et al. (2007), which included testing detector performance with sampling over years and months, as well as a kernel modification to account for call frequency changes (McDonald et al. 2009).

North-east Pacific blue whales produce low-frequency, high-intensity calls, named A, B and D calls. A and B calls often form sequences called songs and are produced only by males (Oleson, Calambokidis, et al. 2007). B calls are slightly downswept, long (up to 18 s) tonal calls that generally exhibit multiple harmonics, with the third harmonic generally being the most energetic and thus most commonly recorded. North-east Pacific blue whale B calls are typically used to detect blue whale presence in passive acoustic data (Wiggins et al. 2005; Stafford et al. 2009; Širović et al. 2015) because their apparent stereotypy makes them amenable for automatic detection. However, the frequency of the B call, as well as other blue whale songs, decreases over time (McDonald et al. 2009). This decrease was shown not to be linear for Antarctic blue whale calls, and is greater within any one calling season than it is from one year to the next (Gavrilov et al. 2012). The overall abundance of blue whale B calls in the Southern California Bight (SCB) varies seasonally, with few calls detected in the spring, early summer and late fall, and large numbers detected during the late summer and early fall (Oleson, Wiggins, et al. 2007; Širović et al. 2015).

As long-term data-sets become more common, a standard process for evaluating automatic detector performance will be necessary to allow interpretation and inter-comparison across different analyses. Passive acoustic data can be used to analyse presence, seasonality and habitat use of species, as well as, more recently, density estimation. Based on the fixed-point density estimation framework described by Marques et al. (2011), data needed for density estimation include the number of calls per unit time, the false positive detection rate, the overall call rate (usually assessed from independent data sources) and the probability of detection. Ideally, a reliable automatic detector will have relatively constant precision and recall over time, allowing one to accurately assess detector performance across a large data-set using a small subset of the data in the valuation process. While this consistency is often implicitly assumed (e.g. Širović et al. 2004), detailed tests of this assumption have not been reported in the literature. In this paper, I test this assumption by investigating the impact of various factors on the detection process. These factors are variability in blue whale B call characteristics, their seasonal abundance and external factors such as human analyst variability and ambient noise. I also examine how these factors may affect data interpretation. Although this work describes the automatic detection of North-east Pacific blue whale B calls from a multi-year data-set, similar automatic detection processes could be applied to spectrogram correlation detection of other mysticete calls.

Methods

Data collection and processing

Data used in these analyses were collected using High-frequency Acoustic Recording Packages deployed in the SCB. Data were initially sampled at a 200 kHz sample rate, but were then low-pass filtered and decimated by a factor of 100. This produced a data-set with an effective bandwidth from 10 Hz to 1 kHz. Long-term spectral averages (LTSAs) with 5 s temporal and 1 Hz frequency resolution were created from the decimated data. These data were used for all analyses of blue whale B call presence and noise measures. Additionally, data used for the analysis of shipping presence (see 'Effects of noise' section below) were created by decimating the original data-set by a factor of 20, resulting in a data-set with an effective bandwidth from 10 Hz to 5 kHz. LTSAs created from these data had 5 s temporal and 10 Hz frequency resolution. Decimated data were stored as WAV and LTSA files, and both were used for subsequent analysis. Data from two different sites, M (33°30.89′ N and 119°14.88′ W) and H (32°50.55′ N and 119°10.27′ W), were used in the analysis. Data collection at sites M and H started in 2007 and 2009, respectively, and continued with minimal interruption until 2012.

Spectrogram correlation output, $a(t)$, is a dimensionless quantity given by:

$$a(t) = \sum_{t_0} \sum_{f} k\left(t_0, f\right) S(t - t_0, f)$$

where $S(t,f)$ is the spectrogram and $k(t,f)$ is the kernel; summation limits are specified by the size of the kernel (Mellinger and Clark 2000). Thus, to automatically detect calls using spectrogram correlation, call kernel (Mellinger and Clark 2000) needs to be defined (see 'Kernel measurements' section below). The kernel is first used for finding the desired threshold, *i.e.* the value of the spectrogram correlation output that will yield acceptable missed call and false alarm rates. For this analysis, that process consisted of developing a ground truth by manually logging all North-east Pacific blue whale B calls (referred to as blue whale B calls in the remainder of the paper) in a training data-set. The spectrogram correlation detector was then run with the appropriate kernel at a variety of thresholds over the same training set. Detector performance was measured by comparing detection times with the times of ground truth call picks. During testing, the automatic detector was generally run at thresholds ranging from 25 to 45. Precision and recall were used as detector performance measures *sensu* Roch et al. (2011). Precision is defined as the fraction of calls from the ground truth (*gc*) that are found by the detector and recall is defined as the fraction of all automatic detections (*d*) that are also in the ground truth. Mathematically, that can be stated as:

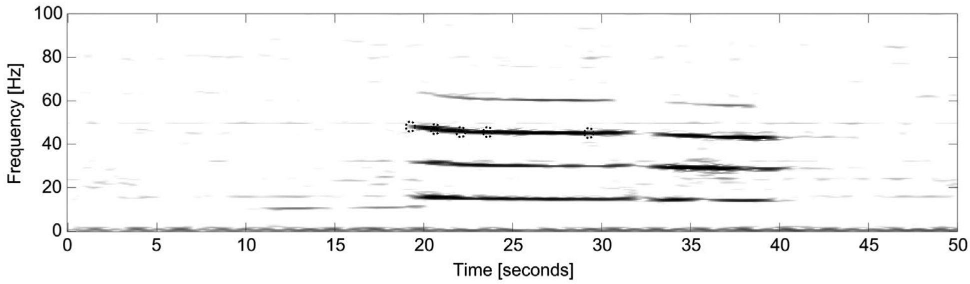

Figure 1. North-east Pacific blue whale B call (as recorded at site M in September 2010) with points at which data were collected for kernel creation, 0, 1.5, 3, 4.5 and 10 s from the start of the call, circled.

$$\text{recall} = \frac{\sum_{gc \in groundc} \text{match}(detections, gc)}{|groundc|}$$

$$\text{precision} = \frac{\sum_{d \in detections} \text{match}(d, groundc)}{|detections|}$$

where *groundc* is the set of all calls in the ground truth, *detections* is the set of all calls from the automatic detector and match(c1, c2) is an indicator function that returns one if call c2 has a matching detection in c1 and returns zero otherwise. The final threshold in this study is defined as the threshold at which the precision to recall ratio is closest to one.

Kernel measurements

To develop appropriate kernels, high-quality B calls first had to be manually identified in the data. Data scanning was done using a custom *MATLAB*-based software program called *Triton* (Wiggins and Hildebrand 2007). Presence of calls was first identified from the LTSAs and quality was assessed from the spectrogram representation of an individual call. When a suitable call was identified, a custom *MATLAB* code was used to automatically extract the call frequency at five points: 0, 1.5, 3, 4.5 and 10 s from the start of the call (Figure 1). Extracted frequencies were the frequencies at peak spectral values in a manually selected frequency band, at the pre-determined time periods from the start of the call. The frequency band was generally between 40 and 55 Hz, corresponding to the third harmonic of the blue whale B call. Spectra were calculated using 8000-point discrete Fourier transforms (DFT) with Hann window and 90% overlap, providing 0.25 Hz frequency resolution. These measurements were collected from at least 10 calls, separated by 24 h or more, for each month. This sampling regimen prevented oversampling of calls from a single animal. The kernel used for detection was finally calculated as a sequence of interpolated segments based on the frequency averages at the five measured points, with hat function width of 2 Hz (increase in hat function value did not impact detector performance).

To evaluate intra-annual variability in blue whale B call frequency in Southern California, monthly kernel measurements were obtained for each month with blue whale B calls over a period of 2 years. In addition, kernel measurements were taken each October during the study period to investigate inter-annual changes in call frequency. To test the impact of inter-annual changes in call frequency on detector performance, the spectrogram correlation detector was tested on data from 2009 to 2012 using the 2012 kernel. Ground truth was obtained by the same analyst for October of each year, and a detector with a kernel developed from 2012 data was applied to each ground truth data-set. The precision and recall curves were compared across years.

Analyses of detector performance variability

The performance of a spectrogram correlation detector may vary due to factors beyond the variability in call frequency. If there is a seasonal variation in call abundance, for example, then an automatic detector with a steady false detection rate will have poorer performance scores during periods of low calling. To test if this occurs with spectrogram correlation of blue whale B calls, ground truths were developed from one day of data collected in June, October and December and detector performance was tested for each month. Also, false detection rate was investigated over a two-month period with low call abundance (May and June) during 1 year of the study. All detections were evaluated for presence of blue whale B calls by a human analyst and those detections that did not contain a call (false detections) were summed to obtain a daily false detection rate. Additional detector performance tests conducted in this analysis were on the effect of analyst and ambient noise.

Effect of analyst

In addition to changes in whale calling, detector performance analysis can also be impacted by the human analyst. To evaluate individual analyst impact on detector evaluation scores and chosen threshold, three trained analysts performed ground truth analysis on data collected at site M from June 2010 through October 2012. Ground truth data for these months were obtained by manually picking approximately 200 calls in June and December, when call abundance is lower, and 400 calls in October, at peak call abundance. A variable number of calls was used to attempt to equalize the overall time period over which the detector was tested, while not overly reducing call sample size. Each analyst followed the above-described procedure for multiple annual evaluations to develop ground truth data-sets during different seasons (*i.e.* June, October and December). Each analyst ran the detector using the same (seasonally and annually adjusted) kernels across a range of thresholds, to estimate the final threshold for each time period. Recall and precision scores for each analyst were combined over this deployment period to estimate mean and standard deviations for each score for each analyst.

Additionally, the effect of the analyst on final data output was also evaluated. A subset of detections from October 2010 and 2012 was used to compare time series resulting from each analyst, and to test whether call rates calculated from each month's analysis were substantially different among analysts. For the latter comparison, a Kruskal–Wallis non-parametric one-way analysis of variance test with alpha level 0.05 was used.

Effect of noise

Spectrogram correlation, like most automatic detection methods, is susceptible to signal quality and its performance decreases with increased noise. For blue whale B calls, the most prominent potential source of masking is shipping noise, as it covers the same frequency range as the calls (McKenna et al. 2012). To evaluate the potential impact of ship noise on the blue whale B call detection rate, ship interference potential was measured in three different ways from continuous data collected at site M from September to November 2010. First, noise levels in the 50 Hz band were calculated as a proxy for ship presence from LTSAs with 5 s time and 1 Hz frequency resolution (Wiggins and Hildebrand 2007). These were further averaged into hourly and daily bins for use in the analyses. Second, ship passages were logged manually using the *MATLAB*-based program *Triton*. Start and end times of each nearby ship passage, as observed from LTSA plots calculated with 5 s temporal and 10 Hz frequency resolution (Wiggins and Hildebrand 2007) over the frequency range 0–5000 Hz, were logged into an *Excel* spreadsheet. Two metrics used for the analysis from this data-set were the total duration of all daily ship encounters and the total number of encounters per day. As none of these data were normally distributed, Spearman's rank correlation coefficient was used to evaluate the correlations between the total blue whale B call detections and four noise metrics: hourly noise level at 50 Hz, daily noise level at 50 Hz, daily number of ship encounters and number of minutes per day with ships passing. Hypotheses of no correlation between detections and the four noise metrics were tested using a Bonferroni adjusted α-level of 0.0125. For correlations with hourly data, only 60 random data points were chosen to have equivalent sample size across all correlation calculations.

Results

Variation in the North-east Pacific blue whale B calls

A frequency shift trend similar to Antarctic blue whale calls (Gavrilov et al. 2012) was found for blue whale B calls. The maximum call frequency in any given year occurred in July, at the start of the calling season in Southern California, and was higher than the minimum call frequency in the previous season, occurring in January or December in this region (Figure 2(A)). However, the change from year to year was not linear and in some years (2009 and 2012), call frequency in October was the same as it was the previous year (Figure 2(B)). Overall, the average decrease in blue whale B call start frequency between 2008 and 2012 was 0.22 Hz/year, about half of that reported previously (McDonald et al. 2009).

There is more variability in calls early in the calling season (Figure 2(A)) and the detector generally performs less well during that time (Figure 3). When a June-specific, rather than the average annual (i.e. October), kernel was used for testing the detector on June data, this kernel modification increased both detector precision and recall by 0.1 during a time of generally poor detector performance (Figure 3). On the other hand, a detector that performs at over 0.9 precision in October can have precision of only 0.5 when tested again on June ground truth data, when both are tested on one day of detections (Figure 3).

The spectrogram correlation detector did not have a constant false detection rate. During 77% of the time tested in this analysis, the false detection rate was very low (<8 calls/day). On about 14% of days it was intermediate (8–24 calls/day or 0.33–1 calls/h) and during the remaining 9% of days false detection rate was >1 call/h. The highest false detection

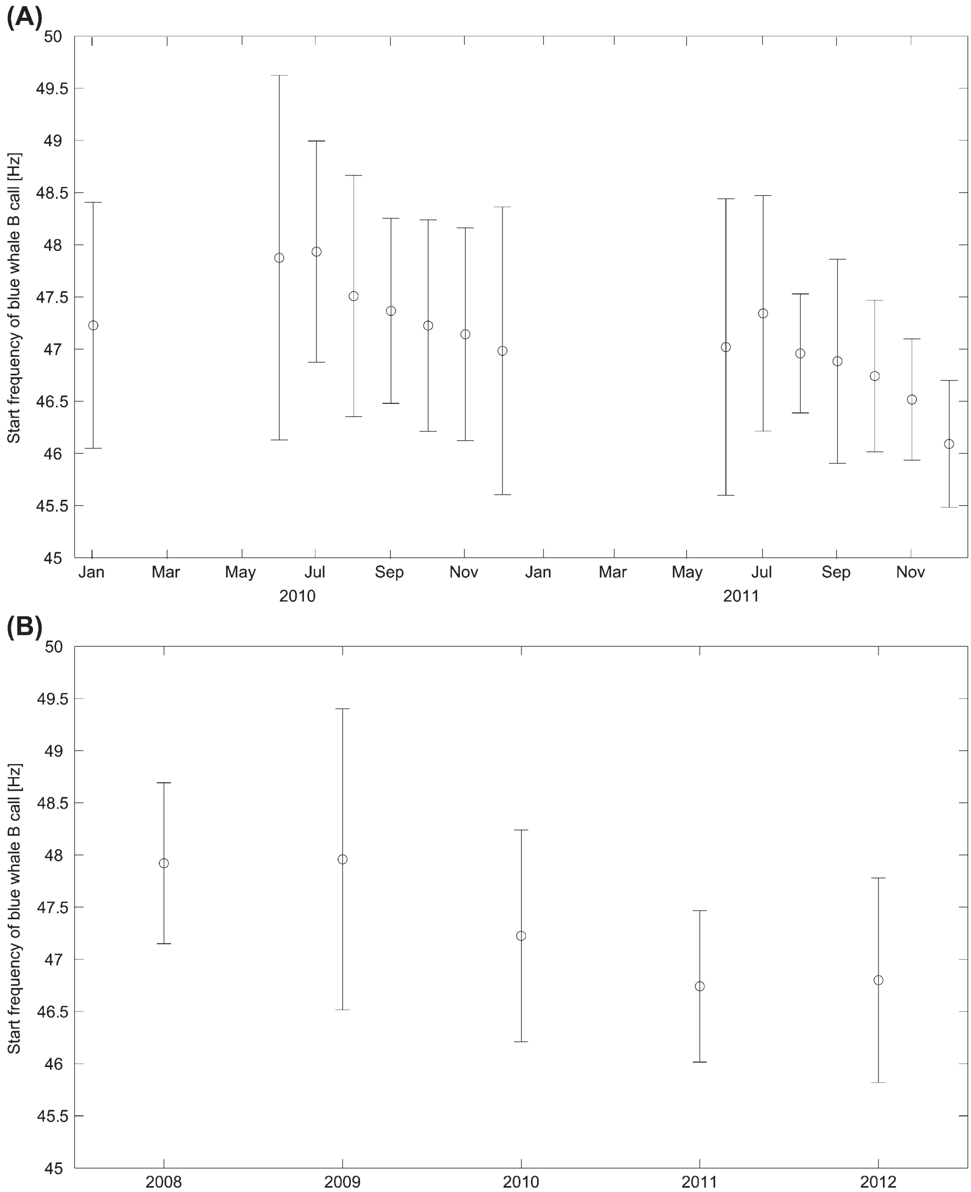

Figure 2. Start frequency (circles represent mean and bars are standard error) of the blue whale B calls recorded at site H (A) during every month with calls from January 2010 through December 2011 and (B) annual average (measured in October) from 2008 to 2012, showing an overall decreasing but non-linear trend.

rate in the tested data was just under 6.6 calls/h. The source of the false detections was most commonly ship noise, but other intermittent sounds from unidentified sources also contributed to false detections.

When a kernel from 2012 was applied to 2011 data, there was little change in detector performance (Figure 4). However, when the 2012 kernel was used for detector evaluation on data from earlier years, there was a decrease in recall by approximately 0.07 over the period

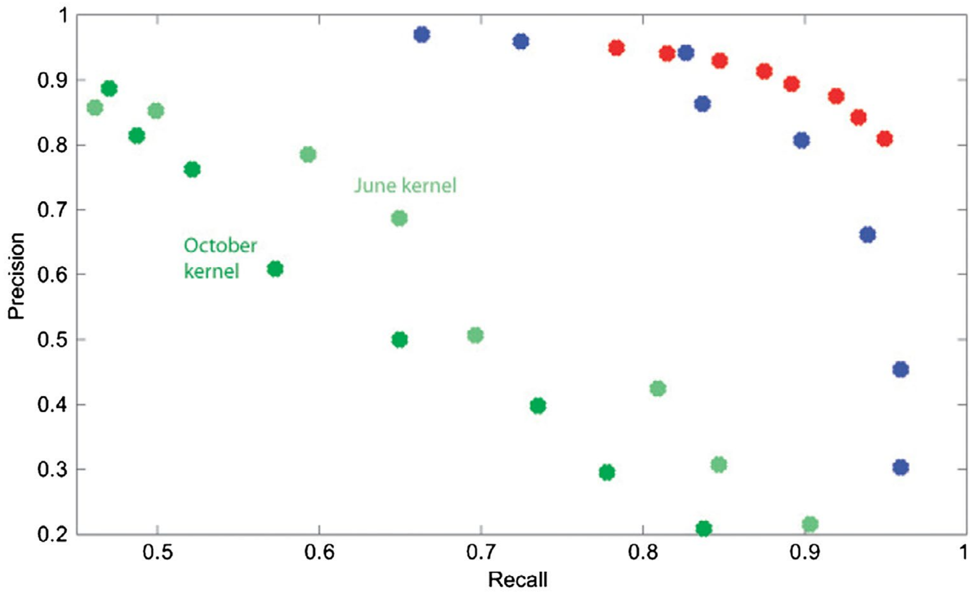

Figure 3. Variation in detector performance within a year when ground truth is based on set data duration, showing precision and recall values for ground truths consisting of 118 calls in June (green), 1600 calls in October (red) and 99 calls in December (blue).
Note: Lighter green also shows improved performance of the detector when a June-specific kernel was applied to the detector instead of the generic October one. All analyses were performed on data collected at site H in 2010.

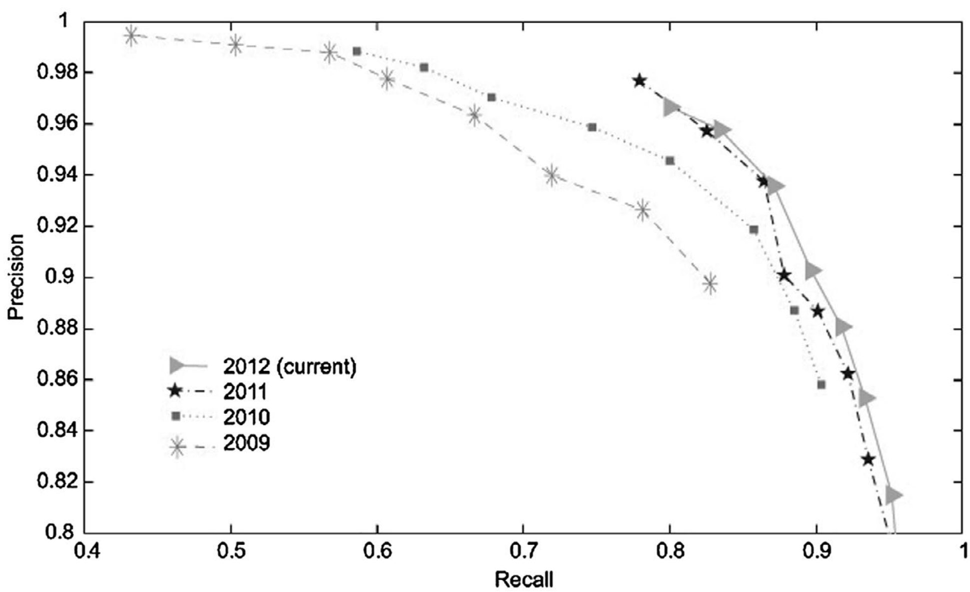

Figure 4. Decrease in detector performance when a kernel from one year is used for evaluation of detector in data from multiple years. Kernel from October 2012 was used for data from Site M from 2009 (asterisk), 2010 (square), 2011 (star) and 2012 (triangle).

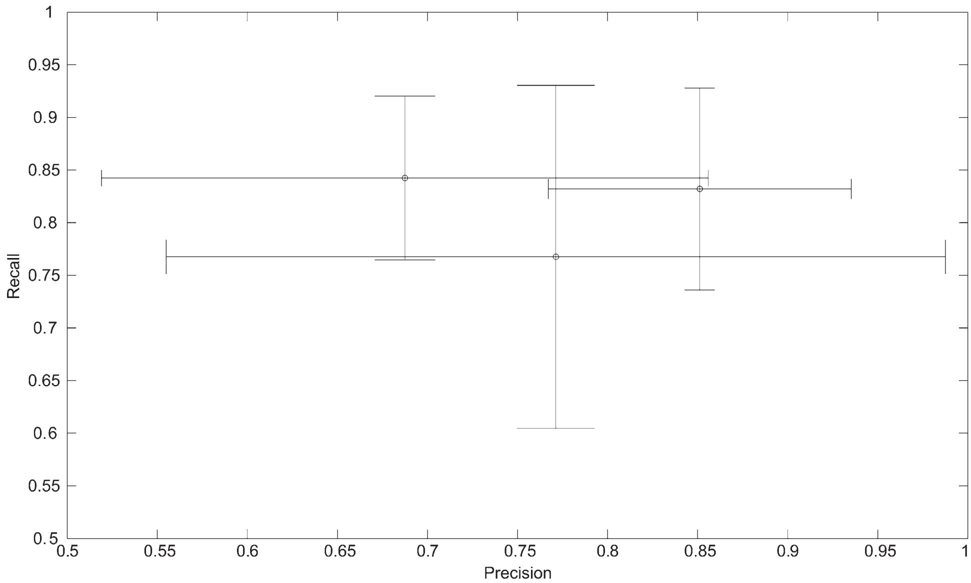

Figure 5. Comparison of precision and recall values (means and with standard deviation bars) produced by the detector evaluation process by three different analysts on the same data-sets collected at site M from June 2010 to October 2012, with ground truth data evaluation occurring during every June, October and December.

Table 1. Precision and recall, as well as the final chosen threshold, for the three analysts who developed ground truth for October 2010 and 2012 data from site M.

Analysis period	Analyst	Precision	Recall	Threshold
October 2010	1	0.66	0.85	33
October 2010	2	0.87	0.89	33
October 2010	3	0.82	0.92	29
October 2012	1	0.90	0.90	33
October 2012	2	0.87	0.89	43
October 2012	3	0.85	0.84	25

from 2009, while precision remained relatively unchanged (within 0.01; Figure 4). This means that in order to maintain a recall to precision ratio close to one, there was an overall decrease in the final threshold value in those years. The similarity in results between 2011 and 2012 is likely due to the fact that call frequency changed little between those two years (Figure 2(B)), supporting the importance of matching the kernel to that year's call frequency.

Effect of analyst

Generally, there was less variation between the analysts in recall, with averages differing by less than 0.08, but precision differed by more than 0.16 (Figure 5). Also, standard deviations of precision were much larger than they were for recall. Even though precision and recall (or false and missed rates) generally get reported for any automatic detection process, in spectrogram correlation these rates are used to set the detection threshold that ultimately determines the detections. In some cases, different precision and recall can produce identical

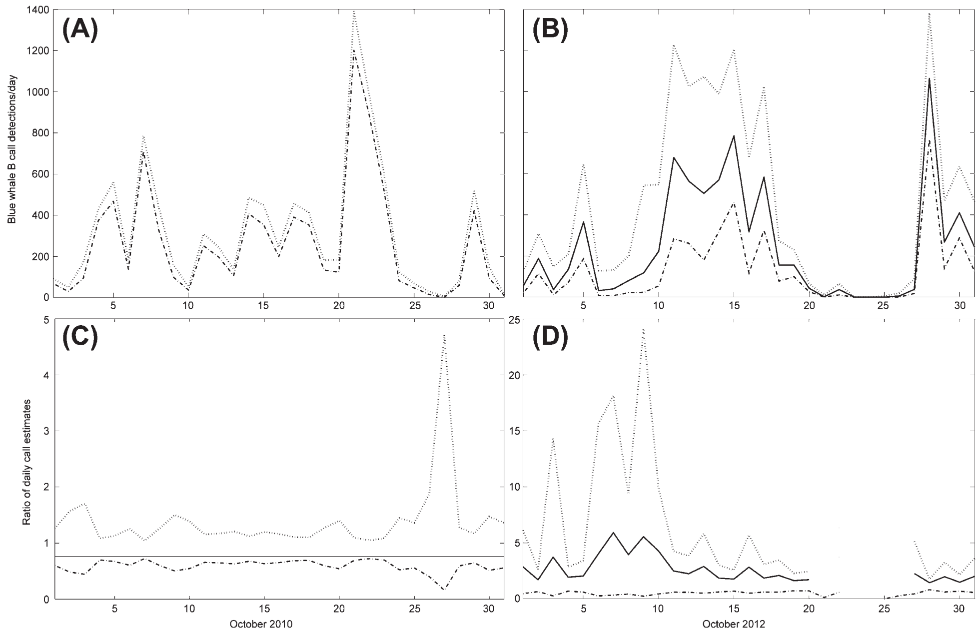

Figure 6. Daily blue whale B call detections from October (A) 2010 and (B) 2012 as analysed by different analysts. Dotted line in A represents results from analysts 1 and 2. Dotted line in B represents results from analyst 2 and analyst 1 results are represented by a solid line. Dash-dotted grey line is analyst 3 in both panels. Ratio between daily call estimates (*i.e.* factor by which number of detections*precision differ) for each pair of analysts is given for data from October (C) 2010 and (D) 2012.

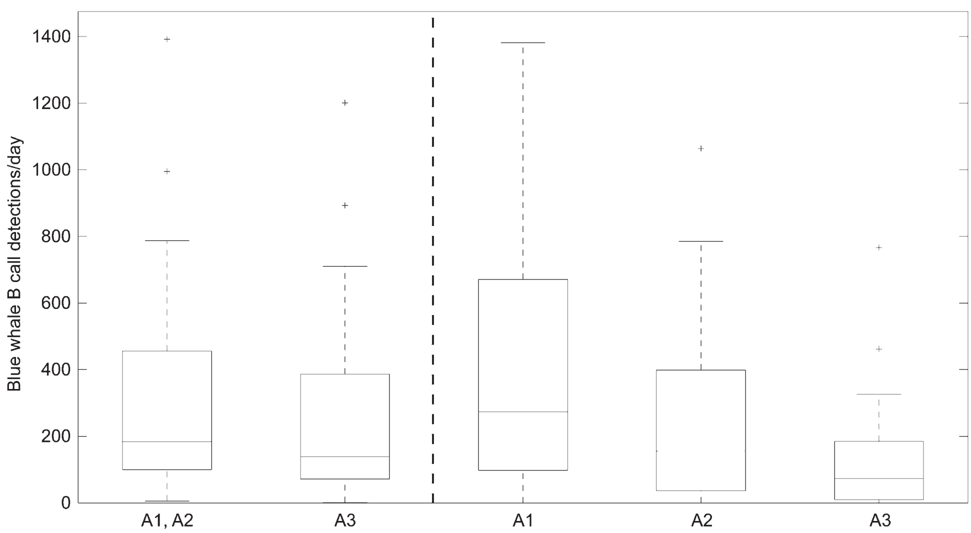

Figure 7. Box plots of median daily call detection rates obtained by different analysts (A1 to A3 represents analysts 1–3), with October 2010 data shown to the left of the broken line and October 2012 data to the right.

detector threshold settings (as for analysts 1 and 2 on October 2010 data; Table 1), while in others similar precision and recall can yield vastly different thresholds (October 2012 for same analysts). This difference in precision and recall from the same thresholds is due to the fact that human analysts do not always pick identical calls when developing a ground truth. The variability mostly results from some analysts picking calls with lower signal-to-noise ratio, shorter duration or otherwise modified appearance. Ultimately, the difference in thresholds impacts the final call detection counts, while precision and recall impact data interpretation.

When investigating overall presence or change in relative call abundance over time, different analysts' results show the same overall trends (Spearman's correlation coefficients $r > 0.93$ and $p<0.001$ for all cases) regardless of their actual threshold settings (Figure 6). Results for October 2012 had a larger amplitude difference, but the peaks and troughs matched very well across data from different analysts; the change in analyst did not change the general temporal trends in the data.

Other potential analyses that move beyond trends in calling animal presence require the use of call rate. In the example shown here, data from October 2010, as analysed by different analysts, produced median call rates of 184 and 138 blue whale B call detections/day (Figure 7). These results, however, were not significantly different ($\chi^2 = 1.04$, $p = 0.307$). Call rates from October 2012 ranged between 73 and 273 blue whale B call detections/day (Figure 7), which were significantly different ($\chi^2 = 11.29$, $p = 0.004$). At times with relatively low detection numbers, daily call estimates can vary by an order of magnitude due to analyst differences (Figure 6(D)), although during times with many detections the difference is likely to be by a much smaller fraction (3–20%). In a case where analysts had different precision rates but ended up using the same threshold, daily call estimate comparison resulted in a systemic offset (Figure 6(C), solid line).

Effect of noise

There were no significant correlations between the daily duration or the total daily number of ship passages and the overall daily blue whale B call detections ($r = 0.087$, $p = 0.508$ and $r = 0.036$, $p = 0.785$, respectively; Figure 8(A)). The relationship with the daily noise levels at 50 Hz was negative, although also not significant ($r = -0.270$, $p = 0.036$). There was, however, a strong negative relationship between the hourly noise levels at 50 Hz and blue whale B call detections ($r = -0.612$, $p<0.001$; Figure 8(B)). The latter results indicate that there is a close coupling between detector performance and noise levels over short time scales, but it gets obscured as data are pooled over longer time periods.

Long-term detection results

Blue whale B call presence was determined by an automatic spectrogram correlation detector from six years (2007–2012) of data in the SCB at site H. Over this deployment period, recall was 0.86 ± 0.11 and precision was 0.85 ± 0.07, with variable detector thresholds applied to each deployment (Table 2). Blue whale B calls were detected every year during this period, with peak detections during the fall. Blue whale B calls generally started being detected in June and July and tapered off in December and January, with more than 2000 call detections/day during peaks in September (Figure 9).

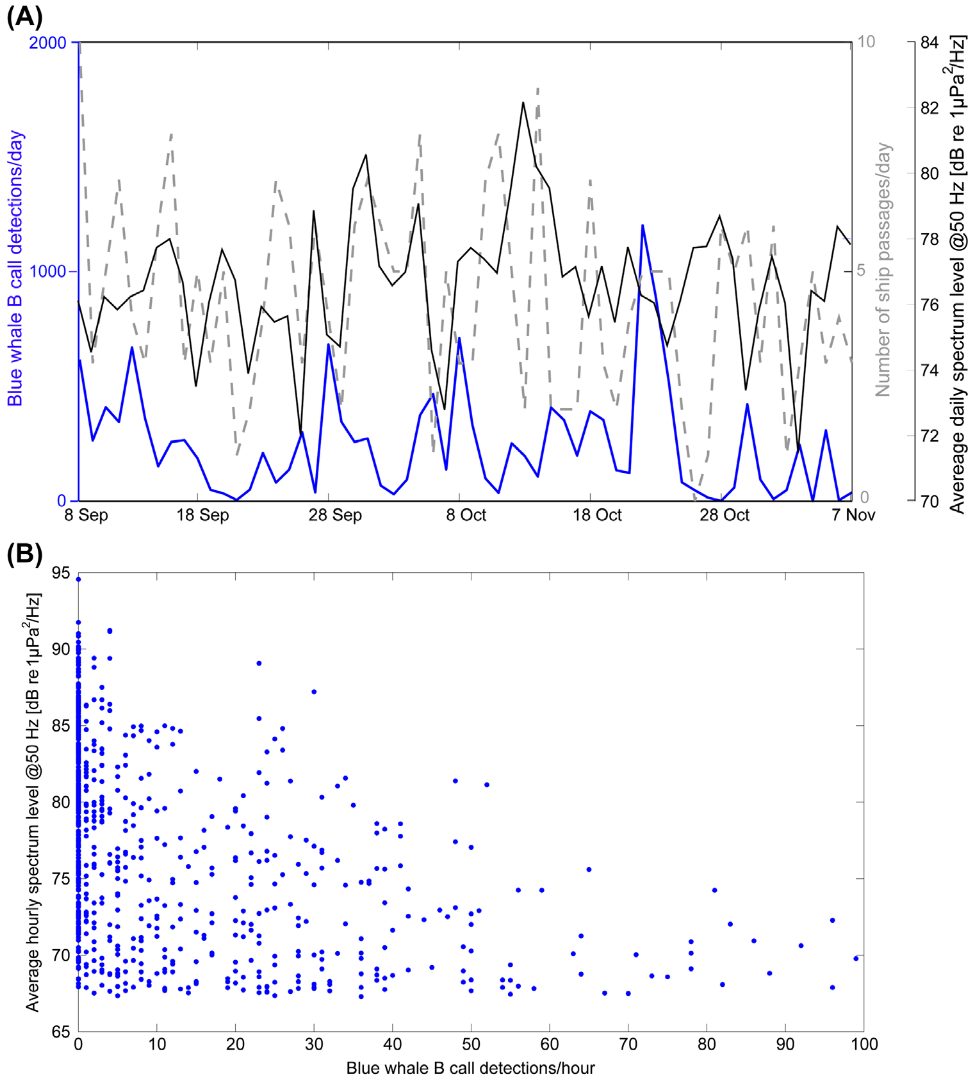

Figure 8. (A) Time series of daily blue whale B call detections (blue solid line) with number of daily ship passages (grey broken line) and average daily noise levels at 50 Hz (solid black line) recorded at site M from 8 September to 7 November 2010. (B) Scatterplot of hourly blue whale B call detections and average hourly noise levels at 50 Hz from the same location and over the same time period.

Discussion

The main goal of automatic detection on passive acoustic data is obtaining robust, long-term calls counts. As a result, the implementation and evaluation process often gets under-reported. However, a detailed accounting of detector performance variability is important for the correct use and interpretation of passive acoustic data and, therefore, deserves more attention. In the case of blue whales, multiple factors affect automatic spectrogram correlation performance, including changes in call frequency, seasonal changes in absolute call abundance, analyst performance and noise.

Table 2. Deployment and threshold details for periods when data collection occurred at site H between 2007 and 2012 (Figure 9).

Deployment start	Deployment end	Threshold
24 Jul 2007	16 Sep 2007	33
5 Jun 2008	29 Jul 2008	27
4 Aug 2008	27 Sep 2008	27
21 Oct 2008	14 Dec 2008	35
21 Dec 2008	8 Mar 2009	35
14 Mar 2009	7 May 2009	33
23 Jul 2009	15 Sep 2009	31
25 Sep 2009	18 Nov 2009	31
6 Dec 2009	29 Jan 2010	33
30 Jan 2010	22 Mar 2010	33
10 Apr 2010	22 Jul 2010	33
22 Jul 2010	8 Nov 2010	31
6 Dec 2010	17 Feb 2011	31
21 Feb 2011	17 Apr 2011	31
11 May 2011	12 Oct 2011	32
16 Oct 2011	5 Mar 2012	39
10 Aug 2012	20 Dec 2012	45

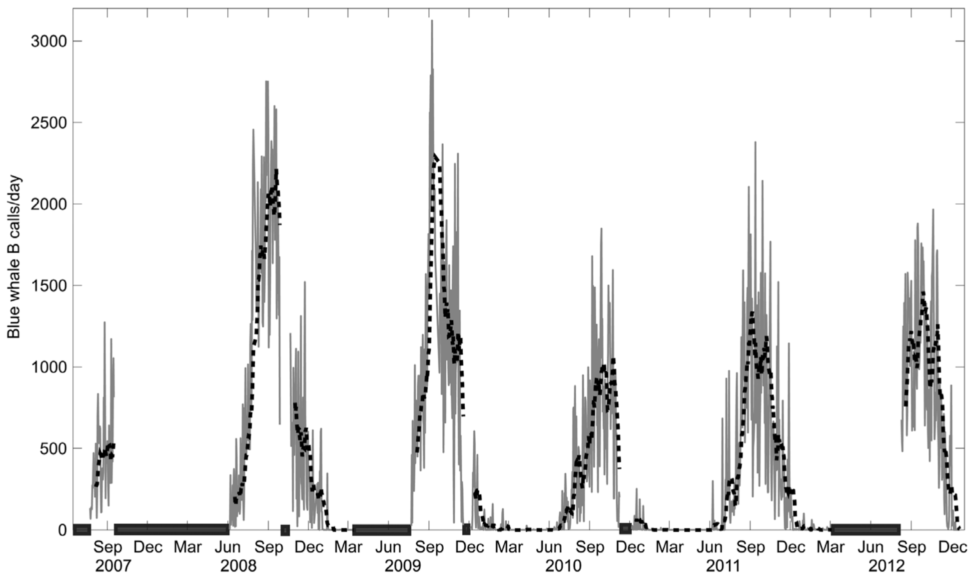

Figure 9. Time series of blue whale B calls detected each day (grey solid line) and a two-week running average (black broken line) at site H in the Southern California Bight from June 2007 to December 2012. Note: Dark boxes along the *x*-axis denote periods when no data were collected.Note: Dark boxes along the *x*-axis denote periods when no data were collected.

An optimal spectrogram correlation detector will take into account both the call's frequency characteristics and the variability in seasonal call abundance. Call frequency of the blue whale B calls in the SCB decreases over time in a non-linear fashion, therefore it is not sufficient to modify the kernel each year by subtracting the average rate of change from the previous year's kernel. Instead, a new kernel needs to be measured and applied each year. In fact, to appropriately account for changes in frequency and thus optimize detection, moving beyond a one-kernel-per-year model and developing multiple kernels within a year is necessary. When multi-year data are analysed, rather than reporting a single precision and recall (or false and missed rate) value, their averages and standard deviations may be the appropriate values to report. A similar process would be required for any other seasonally or annually changing call type.

Detector performance measures are affected by the sample size of ground truth and detection data-sets. Thus, to obtain a robust evaluation of detector performance, it may be necessary to develop ground truth data based on the number of calls wanted in the ground truth as well as the overall test data duration. In addition, when running a spectrogram correlation detector over a long-term time series, a number of false detections will occur during periods of low presence or anticipated absence. However, the rate of false detections seems unpredictable for blue whale B calls because many false detections result from irregular ship passages, and thus detection counts cannot be automatically corrected. Instead, in those cases, a good practice is to manually verify all detections and remove false ones to avoid reporting false calling whale presence (e.g. Širović et al. 2004; Širović et al. 2015).

Two additional factors that can impact detection assessment are analyst variability and noise. Any automatic detection process requires some input from a human analyst. There is anecdotal evidence that analyst abilities affect the final result, but published comparative accounts of human detection results generally include only comparisons of analysts to different algorithms (Deecke et al. 1999; Baumgartner and Mussoline 2011), not investigations of analyst impact on detector output. In the examples presented here, the analyst did not affect overall long-term detection trends, but had an impact on the total detection counts and call rate estimation. Some variability in precision among analysts is likely due to the fact that different analysts picked calls of different quality (variable signal-to-noise ratio, duration, etc.) in their ground truth. False detection rate estimation for density estimation purposes is generally conducted by subsampling detected data and evaluating this rate on detections, which is different from the process used here for estimating precision. However, analyses presented here point to the potential for different analysts to produce different false detection rates, which could result in a systematic error in the final density estimate. The variability in adjusted detection counts among analysts, if further applied to density estimation, could get offset by an analyst-specific detection probability function, but it is a step that should be addressed in any analysis of animal density from passive acoustic data. In general, the range of outputs impacted by an analyst will depend on the automatic detection process itself, but overall, it should be accounted in the results and deserves scrutiny in each such study.

Finally, noise will also impact the final detector output. There are two possible explanations for the negative relationship between call detections and hourly noise levels. One is masking (Southall et al. 2007); the animal may still be producing calls but the calls cannot be distinguished in noise. The other is cessation of calling due to noise (Southall et al. 2007). Since we do not know the intrinsic level of variability in blue whale hourly call rates, it is impossible to establish which one is occurring using only passive acoustic recordings. Practically, however, changes in noise result in a changing detection range so depending on the question being investigated, it may require a dynamic adjustment to the detection area. This could be achieved, for example, by removing high noise periods from overall effort. Accounting for hourly changes in noise can be computationally taxing and impractical, and most often, the interest in these long-term data is not in hourly time scales, but rather daily, weekly or monthly. A simplified solution, if comparing detections from multiple locations, would be to account for longer time scale variability in noise across sites by modelling average propagation characteristics given average noise conditions. Even though this process ignores the fine scale negative correlation between noise and detection rates, the prolonged noise periods will be reflected in the overall long-term noise levels. Thus, computational intensity is reduced allowing comparison over temporal scales that are more often of interest. As mentioned, shipping is the most common masking source for blue whale B calls, but the noise source that may affect the detection process will vary depending on the frequency of the studied call.

Overall, these factors point to complexities in the interpretation of results from an automatic detector, as well as a human analyst. A more detailed quantification of the variability in spectrogram correlation, as well as other automatic detector performance should be conducted when applying these methods to long-term acoustic data-sets. In addition, the interactions of external factors such as noise and variability in calls need to be considered when interpreting automatic detection results.

Acknowledgements

Data deployment, recovery and analysis were supported by Sophie Brandstatter, Tim Christianson, Emily Chou, Amanda Debich, Chris Garsha, John Hildebrand, Brent Hurley, Erin O'Neill, Mattie Parsell, Ally Rice, Bruce Thayer and Sean Wiggins. The manuscript was improved thanks to comments from three anonymous reviewers. This research was supported by the Office of Naval Research (ONR) under Award N00014-12-1-0904 (Michael Weise). Data used in the provided examples were collected with support from Chief of Naval Operations CNO-N45 (Frank Stone, Ernie Young, and Bob Gisiner) and US Pacific Fleet (Chip Johnson).

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by the Office of Naval Research (ONR) [grant number N00014-12-1-0904].

References

Baumgartner MF, Mussoline SE. 2011. A generalized baleen whale call detection and classification system. J Acoust Soc Am. 129:2889–2902.

Deecke VB, Ford JKB, Spong P. 1999. Quantifying complex patterns of bioacoustic variation: use of a neural network to compare killer whale (*Orcinus orca*) dialects. J Acoust Soc Am. 105:2499–2507.

Gavrilov AN, McCauley RD, Gedamke J. 2012. Steady inter and intra-annual decrease in the vocalization frequency of Antarctic blue whales. J Acoust Soc Am. 131:4476–4480.

Helble TA, Ierley GR, D'Spain GL, Roch MA, Hildebrand JA. 2012. A generalized power-law detection algorithm for humpback whale vocalizations. J Acoust Soc Am. 131:2682–2699.

Marques TA, Munger L, Thomas L, Wiggins S, Hildebrand JA. 2011. Estimating North Pacific right whale *Eubalaena japonica* density using passive acoustic cue counting. Endanger Species Res. 13:163–172.

McDonald MA, Hildebrand JA, Mesnick S. 2009. Worldwide decline in tonal frequencies of blue whale songs. Endanger Species Res. 9:13–21.

McKenna MF, Ross D, Wiggins SM, Hildebrand JA. 2012. Underwater radiated noise from modern commercial ships. J Acoust Soc Am. 131:92–103.

Mellinger DK, Clark CW. 1997. Methods for automatic detection of mysticete sounds. Mar Freshw Behav Phy. 29:163–181.

Mellinger DK, Clark CW. 2000. Recognizing transient low-frequency whale sounds by spectrogram correlation. J Acoust Soc Am. 107:3518–3529.

Mouy X, Bahoura M, Simard Y. 2009. Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence. J Acoust Soc Am. 126:2918–2928.

Munger LM, Wiggins SM, Moore SE, Hildebrand JA. 2008. North Pacific right whale (*Eubalaena japonica*) seasonal and diel calling patterns from long-term acoustic recordings in the southeastern Bering Sea, 2000–2006. Mar Mammal Sci. 24:795–814.

Oleson EM, Calambokidis J, Burgess WC, McDonald MA, LeDuc CA, Hildebrand JA. 2007. Behavioral context of call production by eastern North Pacific blue whales. Mar Ecol Prog Ser. 330:269–284.

Oleson EM, Wiggins S, Hildebrand JA. 2007. Temporal separation of blue whale call types on a southern California feeding ground. Anim Behav. 74:881–894.

Potter JR, Mellinger DK, Clark CW. 1994. Marine mammal call discrimination using artificial neural networks. J Acoust Soc Am. 96:1255–1262.

Roch MA, Brandes TS, Patel B, Barkley Y, Baumann-Pickering S, Soldevilla MS. 2011. Automated extraction of odontocete whistle contours. J Acoust Soc Am. 130:2212–2223.

Samaran F, Stafford KM, Branch TA, Gedamke J, Royer JY, Dziak RP, Guinet C. 2013. Seasonal and geographic variation of Southern Blue Whale subspecies in the Indian Ocean. PLoS One. 8(8):e71561.

Širović A, Hildebrand JA, Wiggins SM, McDonald MA, Moore SE, Thiele D. 2004. Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. Deep-Sea Res II. 51:2327–2344.

Širović A, Rice A, Chou E, Hildebrand JA, Wiggins SM, Roch MA. 2015. Seven years of blue and fin whale call abundance in the Southern California Bight. Endanger Species Res. 28:61–76.

Southall, BL, Bowles AE, Ellison WT, Finneran JJ, Gentry RL, Greene Jr. CR, Kastak D, Ketten DR, Miller JH, Nachtigall PE, et al. 2007. Marine mammal noise exposure criteria: Initial scientific recommendations. Aquat Mammal. 33:I–521.

Stafford KM, Bohnenstiehl DR, Tolstoy M, Chapp E, Mellinger DK, Moore SE. 2004. Antarctic-type blue whale calls recorded at low latitudes in the Indian and eastern Pacific Oceans. Deep-Sea Res I. 51:1337–1346.

Stafford KM, Moore SE, Fox CG. 2005. Diel variation in blue whale calls recorded in the eastern tropical Pacific. Anim Behav. 69:951–958.

Stafford KM, Citta JJ, Moore SE, Daher MA, George JE. 2009. Environmental correlates of blue and fin whale call detections in the North Pacific Ocean from 1997 to 2002. Mar Ecol Prog Ser. 395:37–53.

Wiggins SM, Hildebrand JA. 2007. High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring. Proc Int Symp on Underwater Technology 2007 and Int Workshop on Scientific Use of Submarine Cables and Related Technologies. IEEE, Tokyo, Japan.

Wiggins SM, Oleson EM, McDonald MA, Hildebrand JA. 2005. Blue whale (*Balaenoptera musculus*) diel calling patterns offshore of Southern California. Aquat Mammal. 31:161–168.