A generalized baleen whale call detection and classification system

Mark F. Baumgartner[a] and Sarah E. Mussoline
Biology Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

(Received 10 September 2010; revised 7 February 2011; accepted 14 February 2011)

Passive acoustic monitoring allows the assessment of marine mammal occurrence and distribution at greater temporal and spatial scales than is now possible with traditional visual surveys. However, the large volume of acoustic data and the lengthy and laborious task of manually analyzing these data have hindered broad application of this technique. To overcome these limitations, a generalized automated detection and classification system (DCS) was developed to efficiently and accurately identify low-frequency baleen whale calls. The DCS (1) accounts for persistent narrowband and transient broadband noise, (2) characterizes temporal variation of dominant call frequencies via pitch-tracking, and (3) classifies calls based on attributes of the resulting pitch tracks using quadratic discriminant function analysis (QDFA). Automated detections of sei whale (*Balaenoptera borealis*) downsweep calls and North Atlantic right whale (*Eubalaena glacialis*) upcalls were evaluated using recordings collected in the southwestern Gulf of Maine during the spring seasons of 2006 and 2007. The accuracy of the DCS was similar to that of a human analyst: variability in differences between the DCS and an analyst was similar to that between independent analysts, and temporal variability in call rates was similar among the DCS and several analysts.
 [DOI: 10.1121/1.3562166]

PACS number(s): 43.30.Sf [WWA] Pages: 2889–2902

I. INTRODUCTION

Marine mammal occurrence is currently assessed using visual surveys or passive acoustic monitoring. Both methods are challenged by detection uncertainty: visual surveys are often hindered by poor sighting conditions (e.g., nighttime, fog, and high sea states), uncertainty in species identification, and missed detections due to short surfacing intervals, whereas passive acoustic monitoring can be confounded by variable calling rates, noise, uncertainty in caller identity, and missed detections due to silent animals. Passive acoustic monitoring, however, has a distinct advantage over visual surveys: persistence. Whereas visual surveys are labor-intensive (i.e., expensive) and weather-dependent and are, therefore, limited to temporally sporadic sampling over short periods (days to weeks), acoustic recorders can sample continuously for periods ranging from hours to years (Moore *et al.*, 2006). The single greatest drawback of passive acoustic monitoring is the large volume of raw acoustic data returned that requires analysis to generate reliable species detections (Mellinger *et al.*, 2007; Van Parijs *et al.*, 2009). Manual analysis entails visually inspecting spectrograms of acoustic data, aurally reviewing putative calls, and classifying and logging confirmed calls. This method is extremely labor-intensive, inefficient, and unrealistic for most long-duration acoustic recordings. Not surprisingly, the rise in the use of passive acoustic monitoring applications over the past decade has spurred the development of automated methods to detect and classify calls. The overarching goal of this development effort is to significantly reduce the time required to derive detection information from acoustic recordings while maintaining a similar level of accuracy provided by a human analyst.

The advent of automated detection and classification algorithms for low-frequency baleen whale calls has been strongly motivated by conservation needs. In particular, the need for reliable occurrence data for the seriously endangered North Atlantic right whale (*Eubalaena glacialis*) to mitigate ship strikes and fishing gear entanglements has encouraged significant development over the past decade (Gillespie, 2004; Mellinger, 2004; Urazghildiiev and Clark, 2006, 2007a; Urazghildiiev *et al.*, 2009; Dugan *et al.*, 2010a,b). In addition to high accuracy requirements, there is an increasing emphasis on computational efficiency, as automated detection and classification systems (DCSs) are being incorporated into low-power instruments that can provide detections in real-time from a variety of autonomous platforms (Clark *et al.*, 2005; Johnson and Hurst, 2007). With such a capability, moored buoys and autonomous underwater vehicles can provide real-time marine mammal occurrence and distribution information over significantly longer temporal and spatial scales than is now possible with visual methods. Given the dearth of observations in remote areas (oceanic regions) and in seasons with rough seas (e.g., winter), real-time passive acoustic monitoring promises to contribute significantly to our understanding of global marine mammal distribution and ecology.

Most automated techniques for detecting low-frequency baleen whale calls operate primarily in the frequency domain by searching through a spectrogram for a call (although a few analyze the waveform directly; see Johansson and White, 2004; Urazghildiiev and Clark, 2006). Long-duration narrowband noise (e.g., ship noise) is minimized in most

[a]Author to whom correspondence should be addressed. Electronic mail: mbaumgartner@whoi.edu

applications by subtracting a running mean or median from each frequency band of the spectrogram (van Ijsselmuide and Beerens, 2004; Harland and Armstrong, 2004; Mellinger, 2004; Gillespie, 2004; Urazghildiiev and Clark, 2007a; Urazghildiiev *et al*., 2009). Even for detectors operating in the time domain, this narrowband noise is reduced by converting the data to the frequency domain with a Fourier transform, calculating the mean in each frequency band, using an inverse Fourier transform to convert the mean spectrum back to the time domain, and subtracting this waveform from the original digital acoustic data (Johansson and White, 2004; Urazghildiiev and Clark, 2006). This technique is referred to as pre-whitening or normalizing.

Detectors fall roughly into two categories. In the first, a spectrographic representation of a particular call (termed a kernel or template) is convolved with the spectrogram to produce a detection function; when the detection function exceeds a particular threshold, a detection of that call type is considered to have occurred. Examples of this type of detector include spectrogram cross-correlation (Mellinger and Clark, 2000; Mellinger, 2004; Baumgartner *et al*., 2008), neural networks (Mellinger, 2004), and banks of two-dimensional linear finite impulse response filters (Urazghildiiev and Clark, 2007a; Urazghildiiev *et al*., 2009). In some cases, normalization of the spectrogram is used prior to convolution (Urazghildiiev and Clark, 2007a; Urazghildiiev *et al*., 2009), but often the kernel is constructed specifically to account for potentially interfering noise (Mellinger and Clark, 2000). The second category of detectors seeks to identify any and all sounds in a spectrogram, extract attributes of those sounds (e.g., minimum frequency, maximum frequency, duration), and then classify the sound based on the similarity of these measured attributes to those of several call types comprised of tens to hundreds of exemplars. Detectors of this kind include an edge detector (Gillespie, 2004), connectivity algorithm (Harland and Armstrong, 2004), and a detector based on power-law and Page's test algorithms (van Ijsselmuide and Beerens, 2004). Urazghildiiev *et al*. (2009) described a hybrid of these two approaches, the feature vector testing (FVT) algorithm, which uses banks of two-dimensional finite impulse response filters to detect a particular call type (i.e., North Atlantic right whale upcalls), extracts attributes (features) from the resulting call, and compares the measured attributes to *a priori* call-specific limits for each of the features [Dugan *et al*. (2010a,b) extend the FVT classification procedure using neural networks, classification and regression trees, and multi-classifier combination methods].

The automated DCS presented here falls into the second category of detector/classifiers. Instead of extracting call attributes directly from the spectrogram, the time variation of the fundamental frequency is first estimated as a pitch track, and attributes of the call are extracted from this pitch track. This approach allows more efficient estimation of complex frequency modulation (e.g., calls with multiple inflection points) and potential incorporation of amplitude modulation. Pitch-tracking has been used to estimate high frequency contours of odontocete whistles (Buck and Tyack, 1993; Suzuki and Buck, 2000; Oswald *et al*., 2007; Shapiro and Wang, 2009) and minke whales (Mellinger *et al*., 2011), but it has yet to be applied to the detection and classification of low-frequency (<1 kHz) baleen whale calls. We refer to the DCS as generalized because the methods for pitch-tracking, extraction of attributes, and call classification are not specific to any particular call type (unlike, for example, kernel-based spectrogram cross-correlation methods), and are therefore applicable to any narrowband call. We present here a description of the algorithm and an evaluation of the DCS using recordings collected in the southwestern Gulf of Maine during the spring seasons of 2006 and 2007.

II. METHODS

The algorithm used for the DCS, in brief, is as follows (see Fig. 1 for an example). Spectrograms are smoothed using a Gaussian smoothing kernel [Fig. 1(a)], and tonal noise (such as that generated by ships) is reduced by subtracting a long-duration mean from each frequency band in the spectrogram [Fig. 1(b)]. Transient broadband noise is identified and removed from the spectrogram, putative calls are initially detected in the spectrogram using a simple amplitude threshold, and the time variation of the fundamental

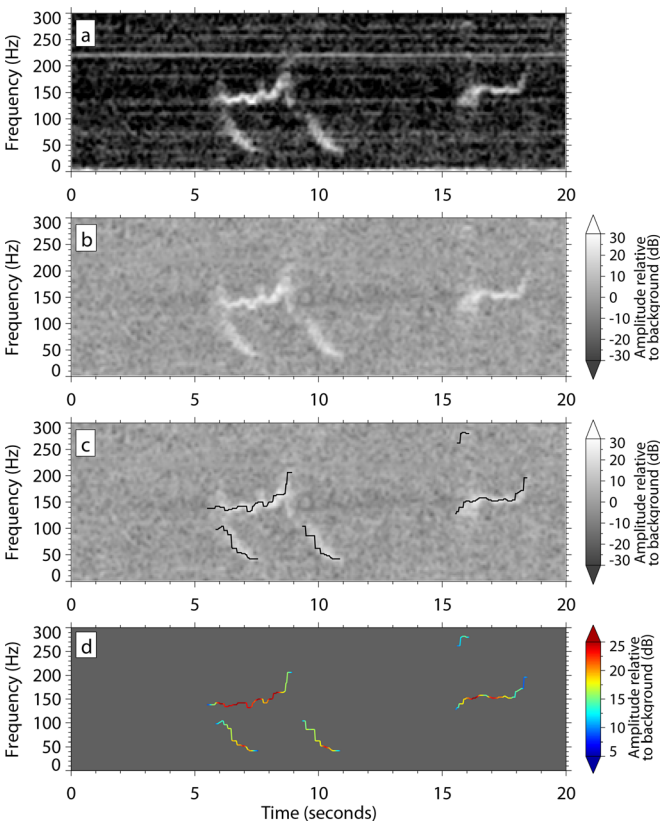

FIG. 1. A pitch-tracking example. (a) Spectrogram [S; Eq. (1)] created from short time Fourier transforms of the audio data (sampling frequency = 2048 Hz, frame = 640 samples, overlap = 80%, Hann window) and smoothed with the smoothing operator [M; Eq. (2)]. Four calls are present: two sei whale downsweeps (40–100 Hz, 1.4 s duration) and two right whale moans (120–170 Hz, 2.7 s duration). (b) Filtered spectrogram [A; Eq. (3)] created by subtracting a running mean from each discrete frequency band. Note removal of 220–225 Hz tonal noise. (c) Pitch tracks of calls with average amplitudes in excess of 12 dB relative to background. (d) Same pitch tracks in (c) with amplitude displayed in color.

frequency for each call is estimated as a pitch track [Figs. 1(c) and 1(d)]. Attributes or features of a call (e.g., duration, average frequency, and frequency sweep) are extracted from the pitch track, and quadratic discriminant function analysis (QDFA) is used for classification. Details of each step in the algorithm are described below.

A. Spectrogram smoothing

The short time Fourier transform was used to produce spectrograms from the digital audio data. The power spectrum for each frame of the spectrogram was produced using the fast Fourier transform with a Hann window and the resulting amplitudes were converted to units of decibels (i.e., $10 \log_{10}[P]$, where P is the power spectrum). Each spectrogram (R) was convolved with a 3×3 smoothing operator (M) to produce smoothed spectrograms (S) as

$$S_{i,j} = \frac{1}{\sum M} \sum_{p=i-1}^{i+1} \sum_{q=j-1}^{j+1} R_{p,q} M_{p-i+1,q-j+1}, \quad (1)$$

where

$$M = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \quad (2)$$

and i and j are the time and frequency indices in the spectrograms, respectively (after Gillespie, 2004). For the data presented below, spectrograms were produced from audio data sampled at 2048 Hz using a frame size of 640 samples and an overlap between frames of 80% (i.e., 80% of the digital audio data used to produce power spectra in successive frames of the spectrogram were identical); the resulting spectrograms had a temporal resolution of 0.0625 s and a frequency resolution of 3.2 Hz.

B. Tonal noise reduction

Continuous tonal noise produced by ships and slowly varying background noise (e.g., storms) was minimized in the spectrogram by subtracting an exponentially weighted running mean from each frequency band in the spectrogram (S) as

$$A_{i,j} = S_{i,j} - m_{i-1,j}, \quad (3)$$

where $m_{i\text{-}1,j}$ is the running mean for frequency band j and A is the resulting filtered spectrogram. The running mean is updated at each time step as

$$m_{i,j} = (1 - \varepsilon) m_{i-1,j} + \varepsilon S_{i,j}. \quad (4)$$

The coefficient ε is specified as

$$\varepsilon = 1 - e^{\log(0.15)\,\Delta t/T}, \quad (5)$$

where Δt is the time resolution of the spectrogram and T is the time constant for the exponentially weighted running mean. Since the weights used to calculate the running mean decrease exponentially over the time elapsed since the current time, the time constant, T, indicates the time at which this weight becomes 15% of the weight applied to the current value. In the filtered spectrogram (A), calls are identified as deviations in amplitude from the sound recorded in a window of time just before the call. The time constant of the running mean should be at least longer than the duration of the longest call expected in the acoustic record. For the data presented below, a time constant of 10 s was used.

C. Broadband noise reduction

Broadband noise is often present in passive acoustic recordings and can be caused by objects striking the hydrophone or the platform carrying the hydrophone, a loose recorder or platform component that creates noise when moved by waves or currents, the hydrophone or platform coming in contact with the bottom, platform-generated noise (e.g., buoyancy pumps on profiling floats or gliders and chain noise on moorings), environmental sources (e.g., ice), or biological sources (e.g., fish "thumping," right whale gunshot calls). Our initial goal was to design a system to detect and classify narrowband calls, so it was desirable to detect and eliminate broadband noises. Future development of this system will include classification of these broadband sounds based on amplitude- and frequency-modulated characteristics.

For each time step i in the filtered spectrogram (A), segments of broadband noise are detected as successive frequency bands, where $A_{i,j} > \alpha_{bb}$ and the frequency span of the segment exceeds a specified bandwidth (ϕ_{bbseg}). The frequency span of each detected segment is summed, and if this sum exceeds a second frequency bandwidth ($\phi_{\Sigma bbseg}$), the time step in the filtered spectrogram is considered to contain a broadband signal. A broadband sound is detected when broadband signals occur in successive time steps such that the sound exceeds a specified duration (τ_{bb}). The broadband sound is then "blanked" in the filtered spectrogram by setting all elements of the broadband sound to zero (i.e., $A_{i,j} = 0$ for all time steps i within the broadband sound, and all frequency elements j spanning the minimum and maximum frequencies of all the detected broadband signals). For the data presented below, the following parameters were used: $\alpha_{bb} = 9.6$ dB, $\phi_{bbseg} = 50$ Hz, $\phi_{\Sigma bbseg} = 150$ Hz, and $\tau_{bb} = 0.125$ s.

D. Pitch-tracking

Pitch-tracking estimates the frequency variation of the call over time using dynamic programming (Wang and Seneff, 2000; Shapiro and Wang, 2009). Candidate sounds for pitch-tracking are located in the spectrogram using a simple amplitude threshold: $A_{i,j} > \alpha_{pt}$ ($\alpha_{pt} = 10$ dB is used in the data presented below). Once an element in the spectrogram is found that satisfies this inequality, a pitch track of the sound is estimated. The algorithm begins with forward pitch-tracking to locate the end of the call, and then uses backward pitch-tracking to identify the entire call. Forward pitch-tracking is used first since the loudest part of a call

likely will occur later in time than when the call is first detected with the amplitude threshold.

Forward pitch-tracking starts at indices (i_0, j_0) in the filtered spectrogram (i.e., $A_{i0,j0} > \alpha_{pt}$). A cost $(C_{i,j})$ is computed for each successive spectrogram element $(i > i_0)$ as

$$C_{i,j} = \begin{cases} P(f_{j0}, f_j) - A_{i,j} & \text{for } i = i_0 + 1, \\ \min\limits_k \left[C_{i-1,*} + P(f_j, f_*) - A_{i,j} \right] & \text{for } i > i_0 + 1, \end{cases} \quad (6)$$

where

$$P(f_u, f_v) = \omega \left| \log_2 \left(\frac{f_u}{f_v} \right) \right|, \quad (7)$$

ω is a user-specified weight ($\omega = 20$ dB in the data presented below), the asterisk indexes all frequencies, and the function "$\min\limits_k[X_*]$" returns the minimum value of a vector X and the index (k) of that minimum value in X. The frequency indices from the "min" function above are retained as

$$J_{i,j} = \begin{cases} j_0 & for \quad i = i_0 + 1 \\ k & for \quad i > i_0 + 1. \end{cases} \quad (8)$$

The function $P(f_u f_v)$ represents a penalty incurred by a potential pitch track for jumping from frequency f_u to f_v in successive spectrogram frames; this penalty is minimized for small frequency changes [$P(f_u, f_v) = 0$ for $f_u = f_v$], is equal to ω for octave jumps, and grows with increasingly larger frequency changes. The value of $C_{i,j}$ for $i > i_0$ represents the cumulative cost of moving from (i_0, j_0) to (i, j), where the cost is defined as the penalty for frequency jumps between successive spectrogram frames minus the filtered spectrogram amplitude. The frequency indices (J) are used to reconstruct the best path (i.e., the one with least cost) from (i, j) back to (i_0, j_0) as $\{(i, j), (i-1, k_1), (i-2, k_2), (i-3, k_3), \ldots, (i_0, j_0)\}$, where $k_1 = J_{i,j}$, $k_2 = J_{i-1,k1}$, $k_3 = J_{i-2,k2}$. At each time step i, the best path is determined as $\min\limits_m [C_{i,*}]$, and the gradient in the cost over the last three points in this best path is computed as $G_i = C_{i-2,m2} - C_{i,m}$, where $m_2 = J_{i-1,m1}$ and $m_1 = J_{i,m}$. If the gradient drops below a threshold $(G_i < \gamma)$, pitch-tracking ceases and the path is ended at $(i_{end}, j_{end}) = (i-2, m_2)$. The gradient threshold used in the data presented below was $\gamma = 15$ dB.

Backward pitch-tracking proceeds from (i_{end}, j_{end}) backward in time $(i < i_{end})$ in exactly the same manner as described above for forward pitch-tracking. The best path determined during the backward pitch-tracking is used as the final call track. Amplitudes of the filtered spectrogram (A) are extracted along the call track, and the final output of the pitch tracking algorithm is a set of time, frequency, and amplitude triplets (t_p, f_p, A_p) for each spectrogram frame in the call.

After detection and pitch-tracking, calls are "blanked" from the filtered spectrogram (A) to ensure they are not included in the pitch tracks of subsequently detected calls. Blanking entails setting each element of the call and neighboring elements to zero: $A_{u,v} = 0$ where $i-3 \le u \le i+3$, $j-5 \le v \le j+5$ for each set of indices (i, j) of the call.

TABLE I. Attributes used to describe a pitch track for the QDFA. A pitch track consists of n sets of time (t_p), frequency (f_p), and amplitude (A_p) triplets. Each attribute is a weighted statistic where the weights are the call amplitudes (A_p).

Attribute	Formula
Average frequency ($\log_2$[Hz])	$\bar{f} = \frac{1}{\sum A} \sum_{p=1}^{n} A_p \log_2(f_p)$
Frequency variation ($\log_2$[Hz])	$\langle f \rangle = \sqrt{\frac{1}{\sum A} \sum_{p=1}^{n} A_p \left[\log_2(f_p) - \bar{f} \right]^2}$
Time variation (s)	$\langle t \rangle = \sqrt{\frac{1}{\sum A} \sum_{p=1}^{n} A_p \left[(t_p - t_0) - \bar{t} \right]^2}$
Slope (octaves per second)	$\beta = \frac{\left[\sum_{p=1}^{n} A_p (t_p - t_0) \sum_{p=1}^{n} A_p \log_2(f_p) \right] - \left[\sum_{p=1}^{n} A_p \sum_{p=1}^{n} A_p (t_p - t_0) \log_2(f_p) \right]}{\left[\sum_{p=1}^{n} A_p (t_p - t_0) \right]^2 - \left[\sum_{p=1}^{n} A_p \sum_{p=1}^{n} A_p (t_p - t_0)^2 \right]}$

Note: $\bar{t} = \frac{1}{\sum A} \sum_{p=1}^{n} A_p (t_p - t_0)$

E. Attribute extraction

For each call track, several attributes were calculated for use in the QDFA. Call attributes for classification typically include characteristics such as start frequency, end frequency, frequency range, duration, and slope of frequency variation (e.g., Gillespie, 2004; Urazghildiiev *et al.*, 2009); however, these attributes rely heavily on accurate estimates of the start and end of a call, which is often quite difficult to determine, particularly when calls are amplitude modulated (e.g., a call "ramps up" in sound level at the beginning or "ramps down" at the end). To minimize errors in the classification of calls caused by uncertainty in the start and end times and frequencies, we chose to make all attributes amplitude-weighted (AW) statistics (Table I). The four attributes used in the QDFA included the AW average frequency, frequency variation, time variation, and slope of the pitch track in time-frequency space; these attributes are AW proxies for the more traditionally used mid frequency, frequency range, duration, and slope, respectively. Amplitude weighting was applied within a pitch track so that louder parts of the call were weighted more heavily relative to softer parts of the call when computing the attributes (i.e., calls recorded with identical amplitude and frequency modulation but at different overall amplitudes would have identical attributes). We found that these AW attributes were less variable than their corresponding traditional attributes (Fig. 2) and therefore provided a more consistent representation of a call type in the QDFA. Frequencies were converted to base 2 logarithms when calculating the attributes since frequency is perceived on a logarithmic scale.

F. Discriminant function analysis

Exemplars of various call types from sei whales (*Balaenoptera borealis*) and right whales were extracted from passive acoustic recordings collected in the Northwestern Atlantic Ocean (Table II; see below). Pitch tracks were estimated and attributes calculated for all exemplars. For each call type (indexed by g), a vector of attribute means (μ_g), the

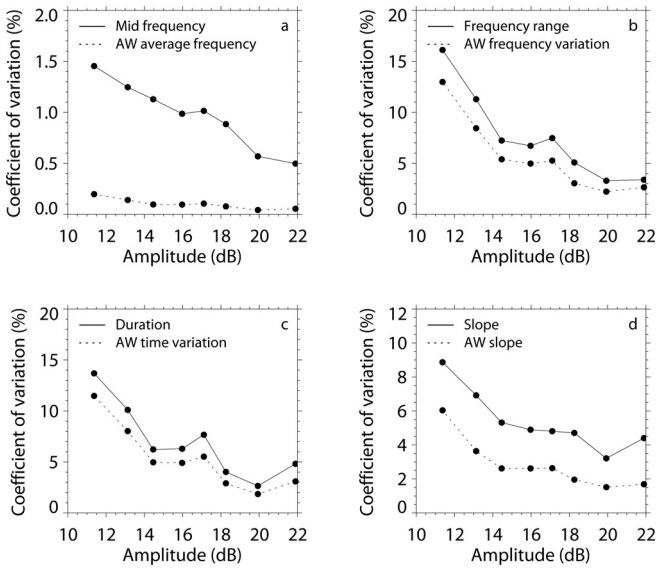

FIG. 2. Variation in traditional and AW attributes with increasing signal amplitude above background. Each filled circle represents the coefficient of variation for an attribute extracted from pitch tracks of 100 synthetic upsweeps (400–600 Hz over 0.75 s with amplitude "ramp up" and "ramp down" times of 0.25 s each and a background of randomly generated white noise).

attribute variance–covariance matrix (Σ_g), the inverse of the variance–covariance matrix (Σ_g^{-1}), and the determinant of the variance–covariance matrix ($|\Sigma_g|$) were computed and stored in a call library.

Attributes of calls were compared to those of each call type in the call library using QDFA (Johnson, 1998). Whereas linear discriminant function analysis assumes each call type has an identical attribute variance–covariance matrix, QDFA allows each call type to have a different attribute variance–covariance matrix. To begin, the Mahalanobis distance (d_g) between a new call and the mean attribute vector of each call type g was computed:

$$d_g(x) = \sqrt{\left(x - \mu_g\right)^T \Sigma_g^{-1} \left(x - \mu_g\right)}, \tag{9}$$

where x is a vector of attributes calculated from the pitch track of the unclassified call and $()^T$ denotes the vector transpose function (Johnson, 1998). The discriminant function (δ_g) was then computed as

$$\delta_g(x) = d_g(x)^2 + \log_e\left(|\Sigma_g|\right) - 2\log_e\left(p_g\right), \tag{10}$$

where p_g is the relative prior probability of the occurrence of call type g (Johnson, 1998). The relative posterior probability of the call belonging to call type g is

$$\pi_g(x) = \frac{e^{-1/2\delta_g(x)}}{\sum_{k=1}^{N} e^{-1/2\delta_k(x)}}, \tag{11}$$

where N is the total number of call types in the call library (Johnson, 1998). The call is classified simply by finding the call type with the maximum relative posterior probability (π_g). In cases where the new call is not represented in the call library, the QDFA can still classify the call with a high relative posterior probability, but the Mahalanobis distance will be quite large, indicating that the new call falls far outside the multivariate distribution of the call type to which it was classified (see example below). Therefore, accurate classification of baleen whale calls will make use of both the relative posterior probability and the Mahalanobis distance.

In most cases, no *a priori* knowledge of relative call occurrence is available, so equal prior probabilities can be used (i.e., $p_g = N^{-1}$ for each call type g). For the data presented below, equal prior probabilities were used. However, in certain situations, the relative prior probability for a particular call type can be increased if that call is known to occur more often than other calls. For example, humpback whales (*Megaptera novaeangliae*) often produce calls that are very similar to other species' calls, creating a challenge for any classification system (including humans). In cases where humpback whale song is detected, the prior probabilities for all humpback whale calls can be increased so that calls of questionable identity will be more likely assigned to humpback whales than if equal prior probabilities are used. Human analysts typically use this same approach: if humpback whales are known to be present, humpbacks are given the "benefit of the doubt" for questionable calls. QDFA provides a convenient means to incorporate this "benefit" in the relative prior probabilities.

TABLE II. Sources of exemplar calls comprising the call library for sei whale downsweeps and right whale upcalls.

Dates	Platform	Location	Number of exemplars	Source
Sei whale downsweep				
May 2005	Glider	Southwestern Gulf of Maine	43	Baumgartner and Fratantoni, 2008
Jul-Sep 2007	Mooring	Davis Strait	127	Stafford unpublished data
Sep 2007	Mooring	Mid-Atlantic Bight	47	Lynch unpublished data
Total			217	
Right whale upcall				
May 2005	Glider	Southwestern Gulf of Maine	63	Baumgartner and Fratantoni, 2008
Nov 2009	Glider	Central Gulf of Maine	191	Baumgartner and Fratantoni unpublished data
Total			254	

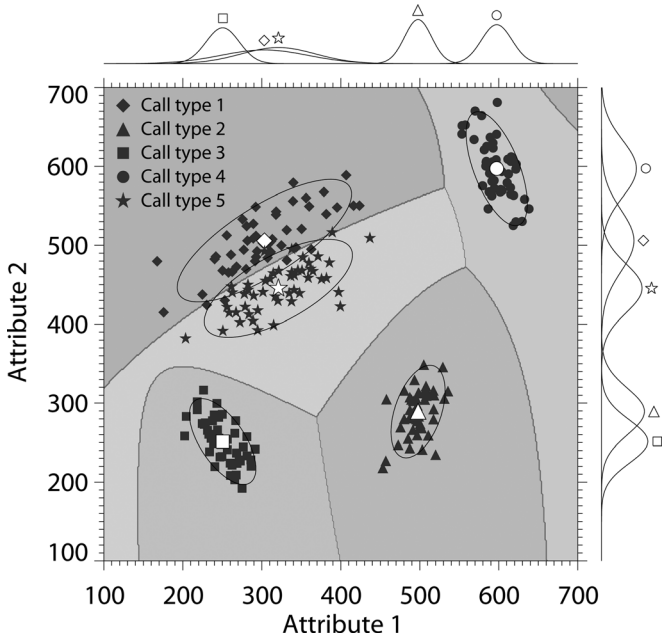

FIG. 3. An example of a call library of five call types with two attributes measured for each exemplar. Exemplars of each call type are shown as small filled symbols, and the mean attribute vector for each call type is indicated by the larger open symbol. The "territory" of each call type is shaded in gray and bounded by gray lines. The ellipses represent a Mahalanobis distance of 2 from each call type's mean attribute vector. The univariate distributions of the attributes for each call type are shown above and to the right of the axes for attribute 1 and attribute 2, respectively.

To illustrate how QDFA is used to classify calls, consider a call library consisting of five call types with two attributes measured for each exemplar. Exemplars of each call type should cluster together in a scatterplot of the attributes (Fig. 3). The mean attribute vector (μ_g) represents the centroid of the multivariate distribution for each call type and the variance–covariance matrix (Σ_g) describes the variability of the exemplars around this mean attribute vector (i.e., the spread of and correlation among the attributes). The Mahalanobis distance is a normalized distance from the mean attribute vector that accounts for the variance–covariance structure; in Fig. 3, any point on the ellipses around each mean vector has a Mahalanobis distance of 2 from that mean vector. To visualize how a new call would be classified given each call type's mean attribute vector and variance–covariance matrix, a "territory map" can be constructed (Fig. 3). For example, a new call with attribute $1 = 500$ and attribute $2 = 300$ would fall in the "territory" of call type 2, and would therefore be classified as call type 2. Similarly, a new call with attribute $1 = 600$ and attribute $2 = 200$ would also fall in the "territory" of call type 2 and would therefore also be classified as call type 2; however, the Mahalanobis distance of this new call from the mean attribute vector of call type 2 is very large (8.7), suggesting that this new call may not belong to any of the call types contained in the call library. Note that when the univariate distributions of call types overlap significantly in each attribute (e.g., call types 1 and 5 in Fig. 3), QDFA can often do a reasonable job of classifying new calls by taking advantage of the correlation structure among the attributes.

G. Call library

The DCS was evaluated for two call types: sei whale downsweep calls (Rankin and Barlow, 2007; Baumgartner *et al.*, 2008) and right whale upcalls (Schevill *et al.*, 1962; Clark, 1982, 1983). Exemplars of these two call types were manually extracted from several independent acoustic datasets collected in the Northwestern Atlantic Ocean from 2005 to 2009 (Table II). Pitch tracks were generated for all calls in these datasets, and sei whale downsweeps and right whale upcalls that were clearly audible and accurately pitch tracked were selected as exemplars. On average, sei whale downsweep exemplars were 16.8 dB above background ($n = 217$, standard deviation (SD) = 3.14 dB, and range = 11.1–27.7 dB) and right whale upcall exemplars were 16.2 dB above background ($n = 254$, SD = 2.68 dB, and range = 10.2–22.8 dB). AW attributes (Table I) were calculated for each exemplar so that the mean attribute vector (μ_g) and variance–covariance matrix (Σ_g) for all call types could be estimated. QDFA could then be conducted by measuring AW attributes for the pitch tracks of new calls (x) and applying Eqs. (9)–(11) with the mean attribute vector (μ_g) and variance–covariance matrix (Σ_g) for each call type (g) in the call library.

H. Evaluation

The accuracy of the DCS was evaluated using passive acoustic recordings collected in the Great South Channel of the southwestern Gulf of Maine during 2006 and 2007. On four separate occasions, arrays of four recoverable Cornell University Marine Autonomous Recording Units (MARUs) were deployed 2 miles to the north, south, east, and west of a central station for short periods of time (1–2 days) in the vicinity of right whales (Table III). The R/V *Albatross IV* occupied the central station for the duration of each study to collect collocated visual observations of whales as well as oceanographic and prey distribution measurements (Baumgartner *et al.*, 2008, 2011); therefore, significant ship noise was present in all recordings. Each MARU was moored with sand bags so that they floated 1.5–2 m above the sea floor. MARUs consisted of a digital audio recorder, hard drive, and batteries encased within an 18-in. glass sphere that was positively buoyant, vacuum sealed, and rated to a depth of 6700 m. Raw audio was captured with an HTI-94-SSQ series hydrophone (2 Hz–30 KHz frequency response) and internal preamplifier (combined maximum sensitivity of −165 dB re 1 V/μPa) mounted outside the plastic "hard hat" that protected the glass sphere. The MARUs were programmed to sample at a rate of 10 kHz, and the resulting digital audio data were low-pass filtered and decimated to 2048 Hz to allow efficient detection of low-frequency baleen whale calls. Recordings from only one MARU per station were used in the evaluation.

Sei whale downsweeps and right whale upcalls were identified in all recordings via manual review by a single analyst (analyst 1) using XBAT software (Figueroa, 2006). The original 10 kHz audio data were low-pass filtered and decimated to 2 kHz, and spectrograms were produced using a Hann window, fast Fourier transform frame size of 512 samples, and an overlap of 0.25 resulting in a frequency

TABLE III. Summary of each station in the southwestern Gulf of Maine where recordings were collected to evaluate the DCS.

Station	Start date/time	Location	Recorder deployments (h)	Water depth (m)
1	5/7/06 13:30	41 17.24 N, 69 08.89 W	25.5	103
2	5/23/06 15:30	41 15.06 N, 68 58.79 W	39.0	137
3	5/21/07 19:00	41 18.76 N, 69 03.28 W	41.5	160
4	6/6/07 20:00	41 56.52 N, 69 04.66 W	48.0	192

resolution of 3.91 Hz and a temporal resolution of 0.192 s. Sounds were classified based on both temporal and frequency characteristics observed in the spectrogram as well as during aural review. All detected calls, regardless of amplitude (signal-to-noise ratio), were used in the analyses described below. A typical right whale upcall lasts approximately 1 s and increases from 100 to 400 Hz (Clark, 1982, 1983). In contrast, sei whale downsweeps last approximately 1 s, range from 90 to 40 Hz, and occasionally occur in pairs (Baumgartner *et al.*, 2008). During manual review, the analyst "boxed" each call by selecting the start and end time and the minimum and maximum frequency. Manual review of right whale upcalls is sometimes confounded by the presence of humpback whale calls that are similar in structure. In cases where humpback whale calls were present, the viewing screen duration was increased from 30 to 240 s to determine whether the potential upcall occurred within a humpback whale song. There are no established criteria for distinguishing between right whale upcalls and similar humpback whale calls, but the repetitive quality of humpback whale song allows for a large number of questionable right whale upcalls to be rejected.

To facilitate an evaluation of between-analyst variability in detection rates, two independent analysts (analysts 2 and 3) also reviewed the acoustic data from station 1 to identify sei whale downsweeps and right whale upcalls. Analysts 1 and 3 conducted their manual review in a manner that is very typical for acoustic studies: viewing spectrograms to initially detect calls and then reviewing the call aurally to aid in classification. In contrast, analyst 2 reviewed and localized sei whale downsweep calls from station 1 as well as the other three stations using a kernel detector and spectrogram cross-correlation to aid in identifying potential calls (see Baumgartner *et al.*, 2008 for details of this manual analysis). The use of a kernel detector by analyst 2 allowed an assessment of how many calls analysts 1 and 3 missed because calls were too faint to be detected using traditional manual review methods. Detection rates from all three analysts and the DCS were compared for station 1, and hourly sei whale call rates were compared for all stations among analyst 1, analyst 3, and the DCS.

As in all DCSs, our DCS uses a threshold to decide when a call is correctly classified. In the evaluation below, only calls with a Mahalanobis distance of three or less were considered correctly classified. This threshold was determined based on the distribution of Mahalanobis distances in the call library. For sei whale downsweep calls, 93% ($n = 201$) of all 217 exemplars had Mahalanobis distances of three or less; similarly, 93% ($n = 235$) of all 254 right whale upcall exemplars had Mahalanobis distances of three or less. The choice of this threshold depends on the application; since our goal was to compare and evaluate the DCS, we set the threshold to correctly classify most calls (in contrast, if one's goal is to identify only high-quality calls, then the Mahalanobis distance threshold would be reduced).

In addition to a Mahalanobis distance threshold, only calls with average amplitudes of 12 dB or more above background were considered for classification (computed as the average of all amplitudes, A_p, of the pitch track; note that this 12 dB average amplitude threshold is different than α_{pt}, the amplitude threshold used to initiate pitch-tracking). Quieter calls are not only more difficult to detect (for the DCS and the human analyst; Urazghildiiev and Clark, 2007b) but also difficult to pitch track accurately. Simulations with synthetic upsweeps of varying amplitudes above background indicated that pitch tracks often became fragmented at amplitudes below 12 dB (Fig. 4); therefore, an amplitude threshold of 12 dB above background was used to reduce the false detection rate at the expense of increasing the missed call rate. For the evaluation presented below, only DCS detections below 12 dB were discarded; all manual detections were included regardless of amplitude.

III. RESULTS

Analyst 1 identified 1062 sei whale downsweep calls and 509 right whale upcalls in the acoustic recordings collected at stations 1–4, and the DCS generated pitch tracks within the time and frequency extents of 99.1% and 98.0% of these calls, respectively. Using a Mahalanobis distance of three and an average amplitude of 12 dB as thresholds, the DCS detected a total of 880 sei whale downsweeps, 570 (65%) of which were in agreement with analyst 1. Assuming the analyst detected and correctly identified all sei whale downsweeps in the recordings, the DCS apparently missed 46% of all downsweeps and incorrectly classified 35% of the downsweeps [Fig. 5(a)]. The DCS detected a total of 466 right whale upcalls, 244 (52%) of which were in agreement with analyst 1. Assuming the analyst detected and correctly identified all right whale upcalls in the recordings, the DCS apparently missed 52% of all upcalls and incorrectly classified 48% of the upcalls [Fig. 5(b)]. On average, DCS-detected sei whale downsweeps were 16.3 dB above background ($n = 880$, SD = 3.47 dB, and range = 12.0–33.0 dB) and right whale upcalls were 14.3 dB above background ($n = 466$, SD = 2.06 dB, and range = 12.0–24.8 dB).

To assess between-analyst variability in detection and classification and the effect this variability may have on the assessment of the DCS, detections from analysts 2 and 3 were compared to that of analyst 1 for station 1 only.

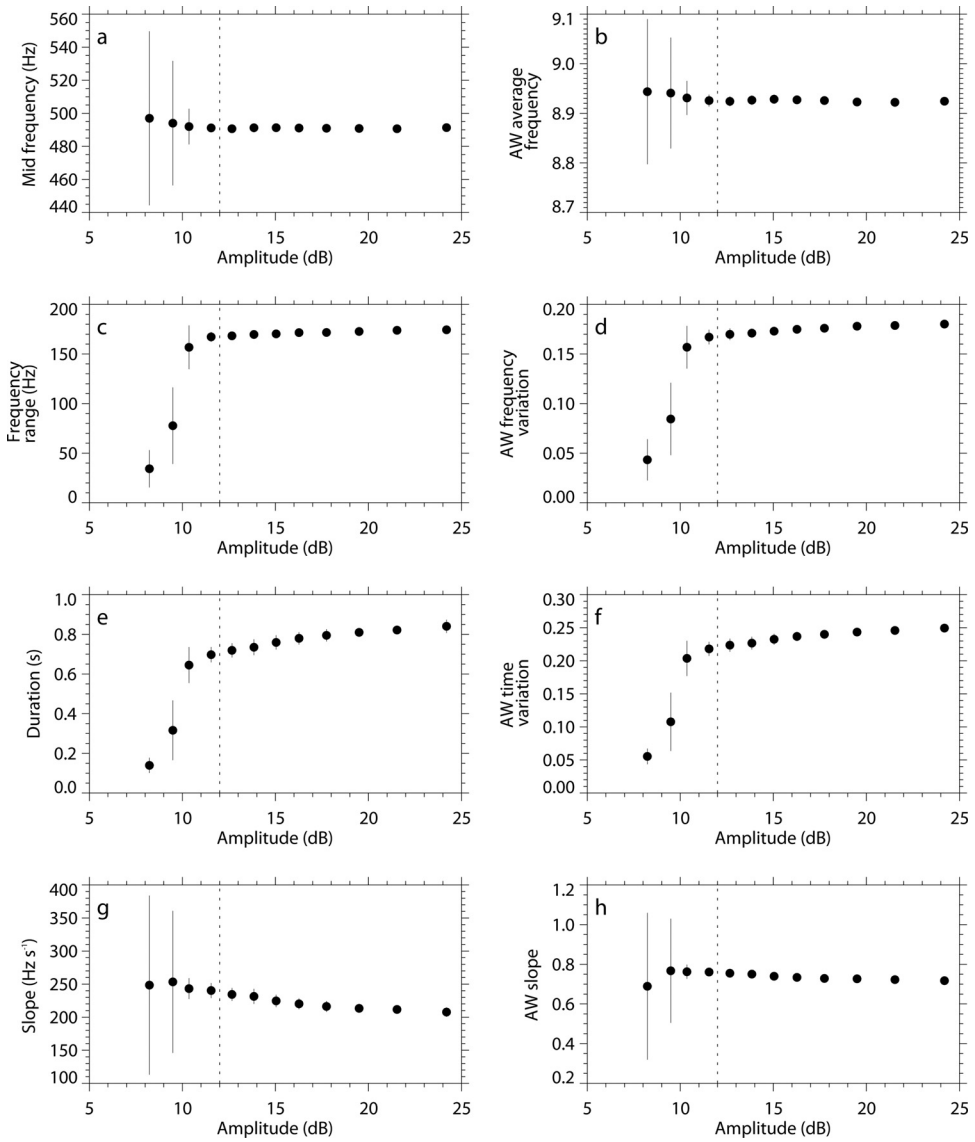

FIG. 4. Average (filled circle) and standard deviation (gray line) of attributes extracted from 12 sets of 100 synthetic upsweeps (400–600 Hz over 0.75 s with amplitude "ramp up" and "ramp down" times of 0.05 s each and a background of randomly generated white noise). Amplitude above random background noise was varied among sets of upsweeps. Traditional (a, c, e, and g) and AW (b, d, f, and h) attributes are shown. When fragmented pitch tracks were generated for an upsweep, attributes were extracted for the pitch track with the longest duration and largest frequency range. Vertical dotted line in each plot indicates 12 dB threshold (see text).

Analysts 1, 2, and 3 identified 570, 939, and 301 sei whale downsweeps during station 1, respectively, while the DCS identified 451 downsweeps. Using the same metrics of missed calls and false detections, analyst 2, analyst 3, and the DCS apparently missed 38, 51, and 33% of all downsweeps while incorrectly classifying 62, 7, and 15% of all downsweeps, respectively, when compared to analyst 1 [Fig. 5(a)]. Analysts 1, 2, and 3 identified 131, 224, and 165 right whale upcalls during station 1, respectively, while the DCS identified 115 upcalls. Again using the same metrics of missed calls and false detections, analyst 2, analyst 3, and the DCS apparently missed 54, 49, and 63% of all upcalls while incorrectly classifying 73, 59, and 58% of all upcalls, respectively, when compared to analyst 1 [Fig. 5(b)].

Despite a lack of perfect agreement between analyst 1 and the DCS for individual calls, agreement for hourly call rates was remarkably good, even during periods of prolonged humpback whale vocal activity during station 3 (Fig. 6; discrepancies during stations 3 and 4 for sei whale downsweeps are addressed in Sec. IV). Differences in call rates between analyst 1 and the DCS for both sei whale downsweeps and right whale upcalls were generally modest when calls were detected by analyst 1 [Figs. 7(a) and 7(c)] and were quite low when analyst 1 detected no calls [Figs. 7(b) and 7(d)]. There were some large disagreements for sei whale downsweeps when the analyst detected calls [Fig. 7(a)]; however, these underestimates by the DCS tended to occur when manually detected call rates were quite high ($\geq$ 30 calls per hour). The DCS tended to underestimate call rates for both sei whales and right whales relative to analyst 1; on average, DCS call rates were 0.66 and 0.79 times the analyst-detected rates for sei whale downsweeps and right whale upcalls, respectively. Overall, the DCS captured the variability in analyst-detected hourly call rates quite well. During 71 and 84% of the 153 hourly periods examined, the DCS and analyst 1 agreed to within three or fewer calls for sei whale downsweeps and right whale upcalls, respectively. Moreover, DCS hourly call rates were very low or zero when analyst-detected call rates were zero, indicating that the actual false detection rate is likely very low for the DCS.

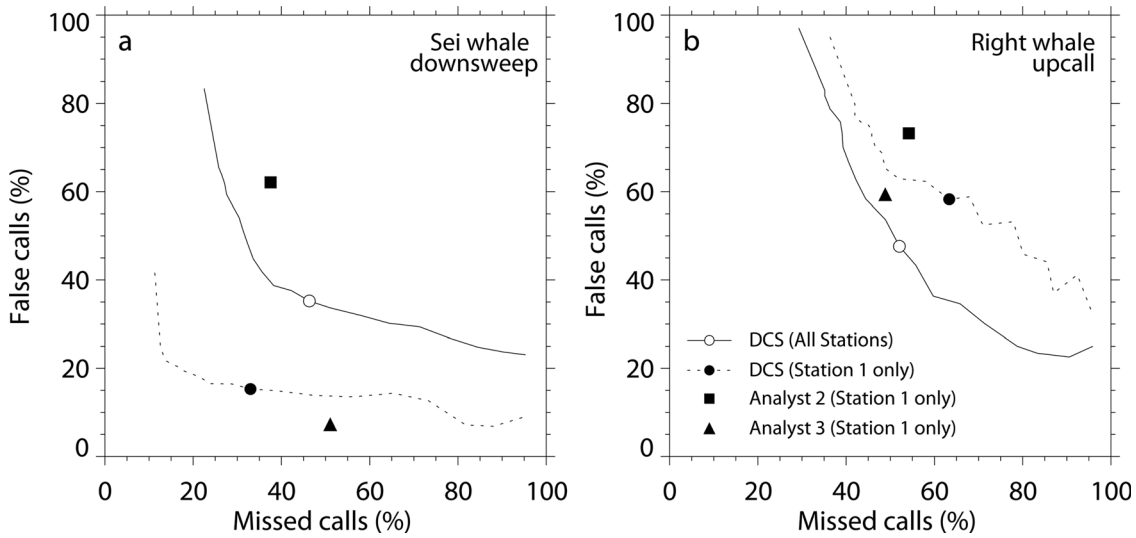

FIG. 5. (a) Performance of the DCS with respect to individual sei whale downsweep calls identified by analyst 1. The solid line indicates the performance of the DCS for varying Mahalanobis distance thresholds for all stations combined, whereas the dotted line indicates the same for station 1 only. The open circle indicates the performance of the DCS for a Mahalanobis distance threshold of three for all stations, and the filled circle indicates the same for station 1 only. The performance of analysts 2 (filled square) and 3 (filled triangle) are also shown with respect to analyst 1 for station 1 only. (b) Performance of the DCS, analyst 2, and analyst 3 with respect to individual right whale upcalls identified by analyst 1.

Although discrepancies in hourly call rates were observed between analyst 1 and the DCS, hourly call rates also varied among analysts when compared during station 1 (Fig. 8). When compared to analyst 1, sei whale downsweep call rates were, on average, higher for analyst 2, lower for analyst 3, and in close agreement for the DCS [Fig. 8(a)], whereas right whale upcall call rates were, on average, higher for analyst 2 and in close agreement for both analyst 3 and the DCS [Fig. 8(b)]. Recall that analyst 2 was aided by a kernel detector when identifying both right and sei whale calls, therefore it is not surprising that detection rates for this analyst were higher than the others. Despite discrepancies in the absolute call rate, the analysts and the DCS tended to agree very well on the relative call rate; that is, temporal

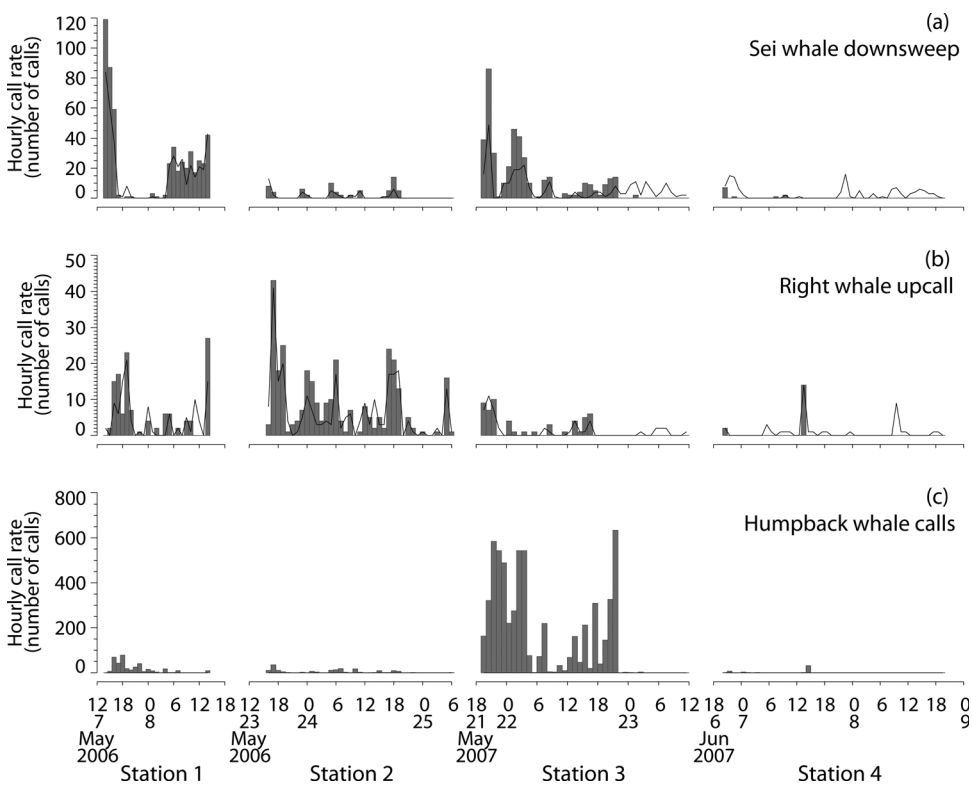

FIG. 6. Hourly call rates of (a) sei whale downsweeps, (b) right whale upcalls, and (c) various humpback whale calls observed by analyst 1 (filled gray bars) and the DCS (black line; a and b only). Humpback whale call rates are included here to indicate periods of potential interference with DCS detections of sei whale downsweeps and right whale upcalls. Discrepancies in (a) during stations 3 and 4 are addressed in Sec. IV and Fig 9.

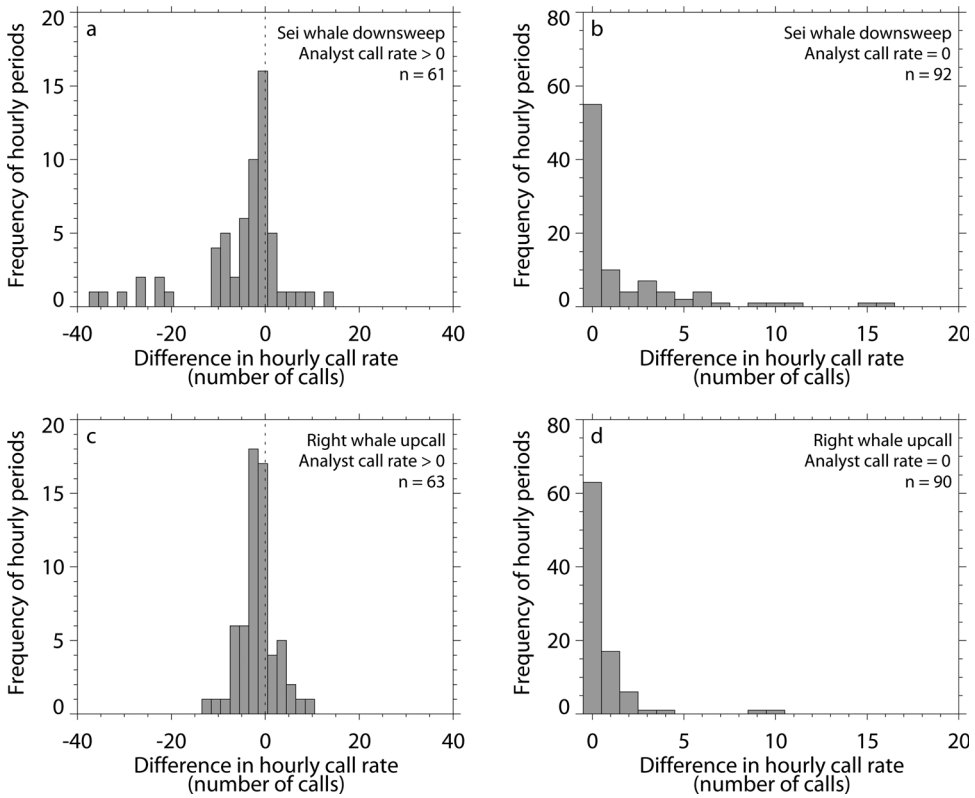

FIG. 7. Differences in hourly call rates observed by the DCS and analyst 1 for (a) sei whale downsweep calls when one or more downsweeps were detected by analyst 1, (b) sei whale downsweep calls when no downsweeps were detected by analyst 1, (c) right whale upcalls when one or more upcalls were detected by analyst 1, and (d) right whale upcalls when no upcalls were detected by analyst 1. Negative differences indicate that the DCS detected fewer calls than analyst 1.

variability in call rates was quite similar among the analysts and the DCS. All of the analysts and the DCS observed (1) high rates of sei whale downsweeps during the first 3 h of station 1, (2) a long period of relatively low call rates immediately afterward, and (3) higher call rates after 0400 on May 8. Likewise, the analysts and the DCS observed initially high rates of right whale upcalls followed by a long period of low but variable call rates until the last hour of the station when the call rate increased dramatically.

IV. DISCUSSION

The performance of automated DCSs for marine mammal sounds is always judged against detections from a

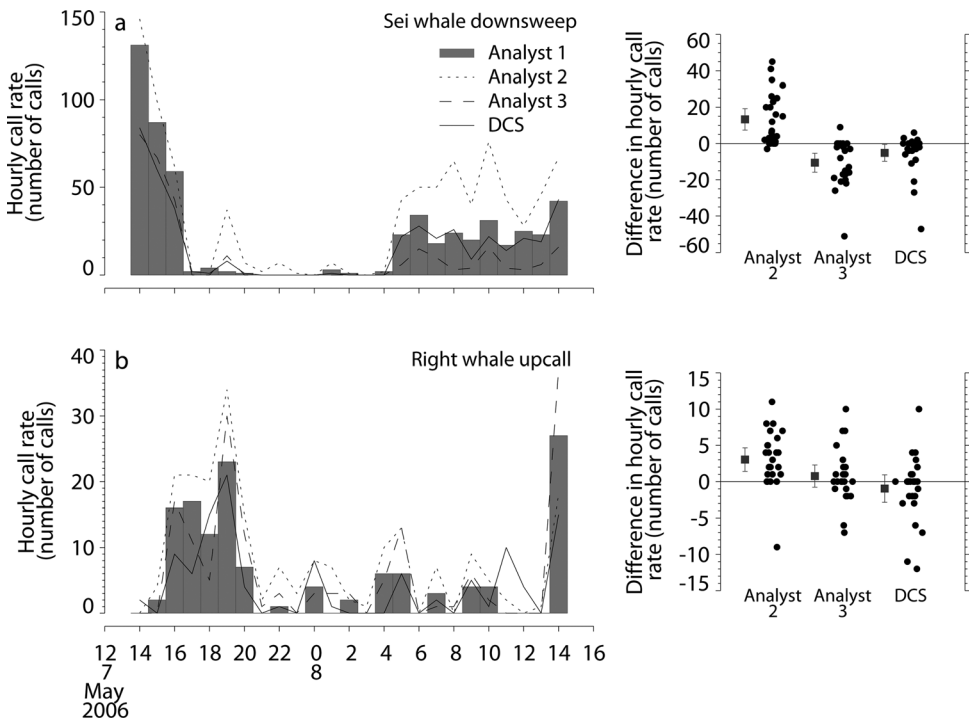

FIG. 8. Hourly call rates observed by the DCS and all analysts as well as the differences in call rates among the DCS, analyst 2, and analyst 3 relative to analyst 1 for (a) sei whale downsweep calls and (b) right whale upcalls during station 1. Negative differences indicate fewer calls were detected than analyst 1. Mean and 95% confidence interval of hourly call rate differences are indicated by a filled square with error bars.

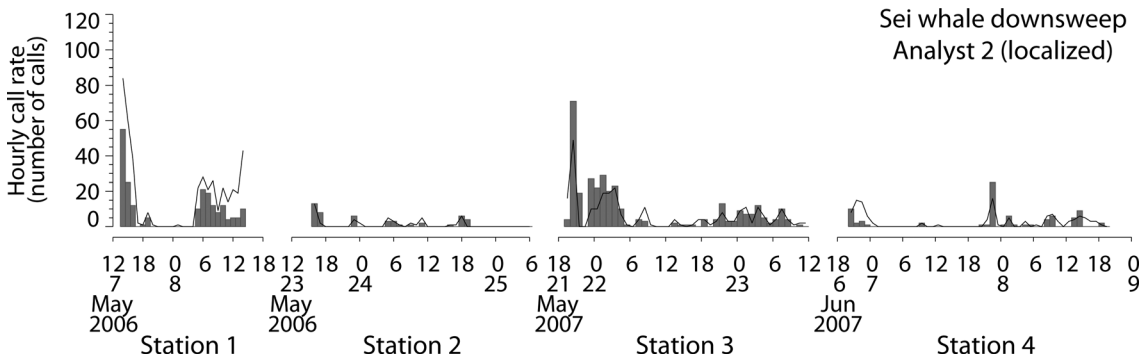

FIG. 9. Hourly call rates of sei whale downsweeps observed by analyst 2 (filled gray bars) and the DCS [black line; same as in Fig. 6(a)] to illustrate differences in analyst-detected call rates. Note differences in call rates between analyst 1 [Fig. 6(a)] and analyst 2 (this plot) toward the end of station 3 and throughout station 4. Analyst 2 was aided by a kernel detector to identify and classify calls (Baumgartner *et al.*, 2008); only localized calls are included here.

human analyst who has manually (visually and aurally) browsed the acoustic data. Differences between these two detection systems are often attributed to errors in the automated system, as the human analyst is considered perfect. There is little recognition that analysts can easily miss calls either because of incorrect spectrogram viewing parameters or fatigue (Urazghildiiev and Clark, 2007b). Moreover, many baleen whale calls are only nominally stereotypical; therefore, some degree of subjectivity is inherent in the classification of calls by an analyst. The three independent analysts in our study observed very different call rates during station 1: analyst 2 detected over three times as many sei whale downsweeps as analyst 3, and analyst 2 detected nearly twice as many right whale upcalls as analyst 1. While analysts 1 and 3 only manually browsed the audio data, analyst 2 had the benefit of using a kernel detector to help identify calls prior to manually browsing and could therefore detect much fainter calls than analysts 1 or 3. By comparing the call rates of analyst 2 to the other analysts, it is clear that analysts 1 and 3 missed a significant number of calls. Even when using the same manual methods to identify calls, analysts 1 and 3 observed different call rates owing to either differing detection rates or more likely, different rules for what constitutes a sei whale downsweep or a right whale upcall. This subjectivity in classification produces uncertainty in the analyst's call rates, which, in turn, causes discrepancies between the analyst and an automated system. For example, the DCS apparently had a high false detection rate for sei whale downsweeps toward the end of station 3 and throughout station 4 [Fig. 6(a)]; however, when compared to a time series of sei whale downsweep calls detected and localized by analyst 2 (Fig. 9; data from Baumgartner *et al.*, 2008), the DCS produced realistic call rates during these same time periods. It is unlikely that analyst 1 failed to detect the calls at the end of station 3 and throughout station 4; instead, analyst 1 subjectively decided that those calls were not sei whale downsweeps, and analyst 2 concluded that they were.

The three analysts in this study were trained acousticians with considerable experience identifying marine mammal sounds. Discrepancies among them are not attributable to inexperience. Instead, differences in detection rate and subjectivity in classification appear to be an unavoidable consequence of manual browsing. While an automated DCS cannot possibly be expected to perform better than a human analyst in classification, our DCS has one potential advantage over an analyst: the rules by which calls are classified are always fixed. Often an analyst's definition of a call type is not a concrete set of criteria against which calls are compared. Instead, the rules are somewhat fluid to accommodate new variants of the call or to allow consideration of neighboring calls when classifying. While this flexibility may lead to more accurate classifications in some cases, it also produces subjectivity, as no two analysts' rules can ever be identical. The DCS, in contrast, uses an objective set of criteria to classify calls, so that the rules are exactly the same each time a classification is made. This consistency is extremely useful when assessing relative call rates, since changes in call rates cannot be attributed to a change in classification rules over time [e.g., Figs. 6(a) and 9], but instead can be attributed to true changes in calling behavior.

Given significant between-analyst variability in detection and classification rates, it is important to assess the performance of an automated DCS against the relative performance of an analyst. Our results from station 1 (Figs. 5 and 8) suggest that the differences between the DCS and analyst 1 are very similar to the differences among the analysts. That is, discrepancies between the DCS and analyst 1 occur as often and are of similar magnitude as those between analyst 1 and the other two analysts. Moreover, the temporal patterns in hourly call rates are also quite similar among the analysts and the DCS. We therefore conclude that the DCS performs at least as well as a typical analyst.

The traditional manual review methods employed by analysts 1 and 3 resulted in missed calls when compared to the kernel detector aided detections of analyst 2. These differences were most likely attributable to the kernel detector's ability to identify times when calls are too faint to easily detect visually in the spectrogram during manual browsing without significant contrast enhancement in the displayed spectrogram (Urazghildiiev and Clark, 2007b). Whereas analysts 1 and 3 would not notice these calls, analyst 2 would aurally review each kernel-detection even when there was little evidence of the presence of a call in the spectrogram. In some cases, this review would result in the detection of a faint call. Intensive manual review to detect every call in an acoustic recording is

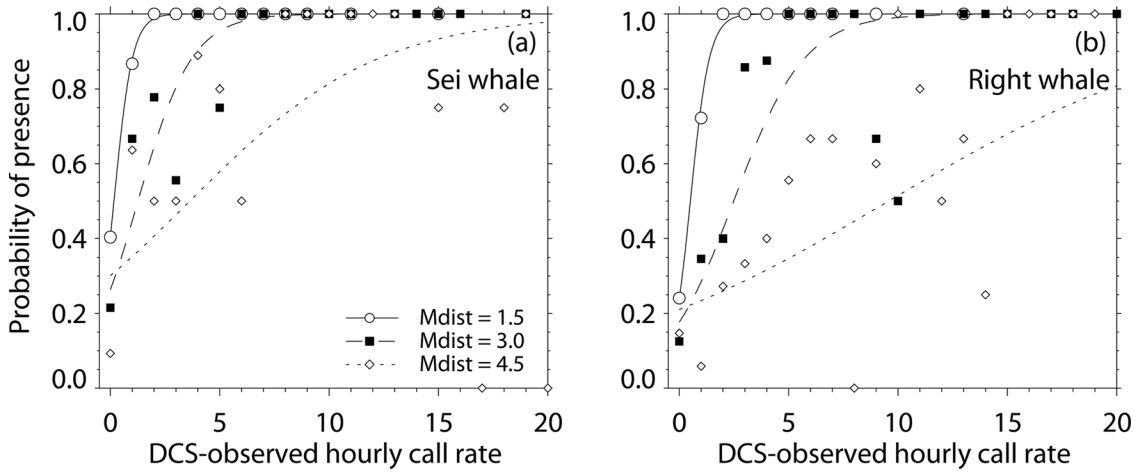

FIG. 10. Relationship between DCS-observed hourly call rates and analyst-observed probability of presence for (a) sei whales and (b) right whales based on detections of the sei whale downsweep and right whale upcall, respectively. Data shown for DCS-observed hourly call rates determined with Mahalanobis distance thresholds of 1.5 (open circles, solid line), 3.0 (filled squares, dashed line), and 4.5 (open diamonds, dotted line). Lines represent fitted logistic regression curves: $\text{logit}(\theta) = \beta_0 + \beta_1 \rho$, where θ is the analyst-observed probability of detection, ρ is the DCS-observed hourly call rate, and β_0 and β_1 are the model parameters. Analyst-observed hourly presence of sei whales was calculated from the combined downsweep detections of analysts 1 and 3, whereas hourly presence of right whales was observed by analyst 1 only.

very often unrealistic; for the 25.5 h of recordings during station 1 alone, analyst 2 reviewed over 11 000 kernel detections to identify 939 sei whale downsweeps and 224 right whale upcalls. The manual review of analysts 1 and 3 took roughly the same effort, and while these reviews detected fewer calls, the temporal pattern of call rates was identical to analyst 2 (Fig. 8). While the DCS also detected fewer calls than analyst 2, it produced similar temporal patterns in call rates to that observed by all of the analysts (Fig. 8). Moreover, the DCS took less than 10 min to detect and classify all calls, a tiny fraction of the time required by the analysts. If the goal of the analysis is to measure absolute call rates (i.e., to detect and classify *every* call), then using an analyst aided by an automated detector with a low threshold of detection is clearly the most accurate approach, albeit extremely laborious. However, if relative call rates are sufficient, then our study suggests that the DCS is far more efficient than an analyst and just as accurate. Assessment of temporal and spatial patterns in call rates rely far more on relative rates than absolute rates; that is, it is more important to know, for example, that call rates increase in one area versus another area or that call rates decrease during the day relative to night. Our results for sei whale downsweeps and right whale upcalls suggest that these relative rates can be efficiently and accurately assessed by the DCS.

The DCS can also be helpful in assessing occurrence. Acoustic data present a unique challenge when assessing occurrence; while the correct detection and classification of calls indicates the presence of one or more whales, the absence of detections does not necessarily imply an absence of whales (since whales may be silent). Similar to the assessment of absolute call rates, an analyst aided by an automated detector with a low threshold for detections is the only reliable way to detect *every* call and thereby correctly assess presence based on vocalizations. Although this process is laborious, it is necessary if a whale only calls on very rare occasions. However, if call rates are high and the false detection rate of a DCS is low, then missed calls do not pose significant challenges for detecting the presence of a species. With an automated DCS, changing the threshold with which calls are considered correctly classified can reduce the false detection rate significantly. In the case of our DCS, the Mahalanobis distance threshold for classification can be reduced so that only high-quality calls (i.e., calls that are well within the multivariate distribution of exemplar attributes) are considered. If call rates are reasonably high, then the detection of a small number of high-quality calls provides strong evidence of presence. For both sei whale downsweeps and right whale upcalls, the probability of analyst-observed presence was 1.0 when only two calls with Mahalanobis distances of 1.5 or less were detected per hour by the DCS (Fig. 10). In contrast, much higher call rates are required to indicate a high probability of analyst-observed presence when using greater Mahalanobis distance thresholds. For example, the probability of analyst-observed sei whale presence exceeded 90% on average when four and 14 downsweep calls with Mahalanobis distances of 3.0 and 4.5 or less were detected per hour by the DCS, respectively [estimated from logistic regressions shown in Fig. 10(a)]. Similarly, the probability of analyst-observed right whale presence exceeded 90% on average when seven and 26 upcalls with Mahalanobis distances of 3.0 and 4.5 or less were detected per hour by the DCS, respectively [Fig. 10(b)].

Recall that calls in this study were classified with a call library built from independent acoustic recordings (Table II). This permitted as rigorous an evaluation as possible. In many applications, the researcher's goal is to identify calls of interest as efficiently and accurately as possible, not to evaluate the DCS. In these cases, the call library can be built directly from the very acoustic recordings intended to be analyzed. This approach is particularly useful when calls vary over time (i.e., seasonally or annually), such as for humpback whale song. Evaluation of false detections can be easily accomplished by review of a

subset of auto-detected calls by a human analyst; however, missed calls can only be evaluated by comparing auto-detections with calls detected and classified during a lengthy independent manual review. When assessing relative call rates, missed calls pose a problem only if absolute call rates are very low. In such cases, the DCS (human or automated) must be able to detect and correctly classify nearly every call. In contrast, when absolute call rates are high, temporal trends or patterns in call rates can still be characterized despite missed calls. False detections are far more egregious when assessing relative call rates, but the rate at which the DCS produces false detections can be directly measured with manual review of a subset of the auto-detected calls.

Although evaluated here for sei whale downsweep calls and right whale upcalls, the DCS can be used for a wide variety of narrowband tonal and frequency-modulated calls, such as those produced by fin (*Balaenoptera physalus*), blue (*Balaenoptera musculus*), humpback, and bowhead (*Balaena mysticetus*) whales. The addition of new call types to a call library is trivial; only the mean attribute vector (μ_g), the inverse of the attribute variance–covariance matrix (Σ_g^{-1}), and the determinant of the attribute variance–covariance matrix ($|\Sigma|$) for each new call type are added to the call library based on pitch tracks from many tens to hundreds of exemplars. Broadband sounds (e.g., right whale gunshots) are not amenable to pitch-tracking, so the DCS described above does not classify these sounds. However, the DCS detects broadband sounds for the purposes of removing them from the spectrogram prior to pitch-tracking. In a manner exactly analogous to the DCS, attributes can be extracted and the broadband sounds can be classified using QDFA based on a separate broadband call library. Because detection and pitch-tracking are generalized and do not rely on call-specific templates (unlike, for example, kernel-based spectrogram cross-correlation methods), only a single pass through the spectrogram is required to detect and classify all call types, which significantly speeds processing for multi-species applications. By removing a long-duration running mean from each frequency band in the spectrogram prior to detection and pitch-tracking, the DCS makes all calculations relative to background noise levels. While this has the advantage of accounting for persistent tonal noise (e.g., noise generated by storms or ships), it also makes the DCS insensitive to changes in gain between different instruments; therefore, identical processing parameters can be used across instruments.

Spectrogram cross-correlation methods rely in great measure on the stereotypy of a call. The effectiveness of spectrogram cross-correlation methods to detect a call is reduced if there is significant variation in call characteristics because the kernel or template is fixed in frequency-time space. While multiple kernel detectors could be used to account for such variation, this approach can be quite computationally expensive. In contrast, the DCS inherently accounts for variability in call characteristics during classification with QDFA. For example, the right whale moan (Matthews *et al.*, 2001) is a low-frequency warble that is a poor candidate for kernel-based spectrogram cross-correlation because the duration varies from call to call. Because duration is simply another attribute used to classify pitch tracks, variation in this attribute poses no problem for the DCS.

In summary, the DCS provides an efficient means to detect and classify a wide variety of narrowband sounds produced by marine mammals. Although we have presented results here for right and sei whale calls, we have used the DCS on several other calls, including those of fin, blue, humpback, and bowhead whales as well as bearded (*Erignathus barbatus*) and ribbon (*Histriophoca fasciata*) seals. The exclusion of persistent tonal noise and transient broadband noise improves performance by reducing false detection rates. The system's capability to identify and exclude noise makes it particularly useful for deployment on autonomous platforms that may inadvertently produce sounds of their own. We have found that QDFA provides a convenient, parsimonious, and extendible framework with which calls can be accurately classified.

ACKNOWLEDGMENTS

We are grateful to Chris Clark and the Cornell Bioacoustics Research Program for the lease of the MARUs, Christopher Tremblay and Ingrid Biedron for their skillful deployment and recovery of the recorders, the officers and crew of the R/V *Albatross IV*, and chief scientist Fred Wenzel. We are also indebted to H. Carter Esch, Sofie Van Parijs, and Ann Warde for acoustic analyses. Kate Stafford, Arthur Newhall, Jim Lynch, and Dave Fratantoni contributed acoustic recordings from which independent exemplars were extracted for the call library. Kate Stafford and two anonymous reviewers provided helpful criticisms of this paper. Fieldwork was made possible by the contribution of vessel and personnel time by the NOAA Northeast Fisheries Science Center (NEFSC); we are particularly grateful for the support and encouragement of Sofie Van Parijs and Richard Merrick. Funding for the fieldwork was provided by the NOAA NEFSC, WHOI Ocean Life Institute, and the WHOI John E. and Anne W. Sawyer Endowed Fund. Development of the detection and classification system was supported by a grant from the Office of Naval Research.

Baumgartner, M. F., and Fratantoni, D. M. (**2008**). "Diel periodicity in both sei whale vocalization rates and the vertical migration of their copepod prey observed from ocean gliders," Limnol. Oceanogr. **53**, 2197–2209.

Baumgartner, M. F., Lysiak, N. S. J., Schuman, C., Urban-Rich, J., and Wenzel, F. W. (**2011**). "Diel vertical migration behavior of *Calanus finmarchicus* and its influence on right and sei whale occurrence," Mar. Ecol. Prog. Ser. **423**, 167–184.

Baumgartner, M. F., Van Parijs, S. M., Wenzel, F. W., Tremblay, C. J., Esch, H. C., and Warde, A. M. (**2008**). "Low frequency vocalizations attributed to sei whales (*Balaenoptera borealis*)," J. Acoust. Soc. Am. **124**, 1339–1349.

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

Clark, C. W. (**1982**). "The acoustic repertoire of the southern right whale, a quantitative analysis," Anim. Behav. **30**, 1060–1071.

Clark, C. W. (**1983**). "Acoustic communication and behavior of the southern right whale," in *Behavior and Communication of Whales*, edited by R. S. Payne (Westview Press, Boulder, CO), pp. 163–198.

Clark, C. W., Calupca, T., Gillespie, D., Von der Heydt, K., and Kemp, J. (**2005**). "A near-real-time acoustic detection and reporting system for endangered species in critical habitats," J. Acoust. Soc. Am. **117**, 2525.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., and Clark, C. W. (**2010a**). "North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms," in *IEEE Proceedings of the 2010 Long Island Systems, Applications and Technology Conference*, Farmingdale, NY, pp. 1–6.

Dugan, P. J., Rice, A. N., Urazghildiiev, I. R., and Clark, C. W. (**2010b**). "North Atlantic right whale acoustic signal processing: Part II. Improved decision architecture for auto-detection using multi-classifier combination methodology," in *IEEE Proceedings of the 2010 Long Island Systems, Applications and Technology Conference*, Farmingdale, NY, pp. 1–6.

Figueroa, H. (**2006**). "Extensible bioacoustic tool (XBAT)," http://xbat.org/home.html (Last viewed September 9, 2010).

Gillespie, D. (**2004**). "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," Can. Acoust. **32**, 39–47.

Harland, E. J., and Armstrong, M. S. (**2004**). "The real-time detection of the calls of cetacean species," Can. Acoust. **32**, 76–82.

Johansson, A. T., and White, P. R. (**2004**). "Detection and characterization of marine mammal calls by parametric modeling," Can. Acoust. **32**, 83–92.

Johnson, D. E. (**1998**). *Applied Multivariate Methods for Data Analysts* (Duxbury Press, Pacific Grove, CA), pp. 217–286.

Johnson, M., and Hurst, T. (**2007**). "The DMON: An open-hardware/open-software passive acoustic detector," in *Proceedings of the 3rd International Workshop on the Detection and Classification of Marine Mammals using Passive Acoustics*, Boston, MA, p. 12.

Matthews, J. N., Brown, S., Gillespie, D., Johnson, M., Mclanaghan, R., Moscrop, A., Nowacek, D., Leaper, R., Lewis, T., and Tyack, P. (**2001**). "Vocalisation rates of the North Atlantic right whale (*Eubalaena glacialis*)," J. Cetacean Res. Manage. **3**, 271–282.

Mellinger, D. K. (**2004**). "A comparison of methods for detecting right whale calls," Can. Acoust. **32**, 55–65.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529.

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

Mellinger, D. K., Stafford, K. M., Moore, S. E., Dziak, R. P., and Matsumoto, H. (**2007**). "An overview of fixed passive acoustic observation methods for cetaceans," Oceanogr. **20**, 36–45.

Moore, S. E., Stafford, K. M., Mellinger, D. K., and Hildebrand, J. A. (**2006**). "Listening for large whales in the offshore waters of Alaska," BioScience **56**, 49–55.

Oswald, J. N., Rankin, S., Barlow, J., and Lammers, M. O. (**2007**). "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **122**, 587–595.

Rankin, S., and Barlow, J. (**2007**). "Vocalizations of the sei whale *Balaenoptera borealis* off the Hawaiian Islands," Bioacoustics **16**, 137–145.

Schevill, W. E., Backus, R. H., and Hersey, J. B. (**1962**). "Sound production by marine animals," in *Bioacoustics*, edited by M. N. Hill (Wiley, New York), pp. 540–566.

Shapiro, A. D., and Wang, C. (**2009**). "A versatile pitch tracking algorithm: From human speech to killer whale vocalizations," J. Acoust. Soc. Am. **126**, 451–459.

Suzuki, R., and Buck, J. R. (**2000**). "Extraction and tracking of bottlenose dolphin whistle contours," J. Acoust. Soc. Am. **108**, 2635–2636.

Urazghildiiev, I. R., and Clark, C. W. (**2006**). "Acoustic detection of North Atlantic right whale contact calls using the generalized likelihood ratio test," J. Acoust. Soc. Am. **120**, 1956–1963.

Urazghildiiev, I. R., and Clark, C. W. (**2007a**). "Acoustic detection of North Atlantic right whale contact calls using spectrogram-based statistics," J. Acoust. Soc. Am. **122**, 769–776.

Urazghildiiev, I. R., and Clark, C. W. (**2007b**). "Detection performances of experienced human operators compared to a likelihood ratio based detector," J. Acoust. Soc. Am. **122**, 200–204.

Urazghildiiev, I. R., Clark, C. W., Krein, T. P., and Parks, S. E. (**2009**). "Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise," IEEE J. Ocean. Eng. **34**, 358–368.

van Ijsselmuide, S. P., and Beerens, S. P. (**2004**). "Detection and classification of marine mammals using an LFAS system," Can. Acoust. **32**, 93–106.

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

Wang, C., and Seneff, S. (**2000**). "Robust pitch tracking for prosodic modeling in telephone speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Piscataway, NJ), pp. 1143–1146.