Title

Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys

Journal

The Journal of the Acoustical Society of America, 131(5)

ISSN

0001-4966

Authors

Thode, Aaron M
Kim, Katherine H
Blackwell, Susanna B
et al.

Publication Date

2012-05-01

DOI

10.1121/1.3699247

Peer reviewed

Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys

Aaron M. Thode[a)]
Marine Physical Laboratory, Scripps Institution of Oceanography, San Diego, California 92093-0205

Katherine H. Kim, Susanna B. Blackwell, and Charles R. Greene, Jr.
Greeneridge Sciences, Inc., 6160-C Wallace Becknell Road, Santa Barbara, California 93117

Christopher S. Nations and Trent L. McDonald
Western EcoSystems Technology, Inc., 2003 Central Avenue, Cheyenne, Wyoming 82001

A. Michael Macrander
Shell Exploration and Production Co., 3601 C St. Suite 1000, Anchorage, Alaska 99503

(Received 29 April 2011; revised 31 January 2012; accepted 10 March 2012)

An automated procedure has been developed for detecting and localizing frequency-modulated bowhead whale sounds in the presence of seismic airgun surveys. The procedure was applied to four years of data, collected from over 30 directional autonomous recording packages deployed over a 280 km span of continental shelf in the Alaskan Beaufort Sea. The procedure has six sequential stages that begin by extracting 25-element feature vectors from spectrograms of potential call candidates. Two cascaded neural networks then classify some feature vectors as bowhead calls, and the procedure then matches calls between recorders to triangulate locations. To train the networks, manual analysts flagged 219 471 bowhead call examples from 2008 and 2009. Manual analyses were also used to identify 1.17 million transient signals that were not whale calls. The network output thresholds were adjusted to reject 20% of whale calls in the training data. Validation runs using 2007 and 2010 data found that the procedure missed 30%–40% of manually detected calls. Furthermore, 20%–40% of the sounds flagged as calls are not present in the manual analyses; however, these extra detections incorporate legitimate whale calls overlooked by human analysts. Both manual and automated methods produce similar spatial and temporal call distributions.
 [http://dx.doi.org/10.1121/1.3699247]

PACS number(s): 43.30.Sf, 43.60.Np, 43.30.Wi, 43.80.Nd [KGF] Pages: 3726–3747

I. INTRODUCTION

In 2007 and 2008, the Shell Exploration and Production Company (SEPCO) conducted a series of seismic exploration surveys in the Beaufort Sea during the late summer and early fall in relatively shallow arctic waters off Alaska. In order to evaluate the potential impact of airgun sounds on bowhead whales (*Balaena mysticetus*) during their westward fall migration, SEPCO commissioned Greeneridge Sciences, Inc. to deploy at least 35 "Directional Autonomous Seafloor Acoustic Recorders" (DASARs) (Greene *et al.*, 2004), divided among five sites over a 280 km swath in the coastal Beaufort Sea. Whale sounds were recorded over at least 35 days, covering times before, during, and after local survey activities for both years. The study complements past and present acoustic studies of bowheads in the Arctic (Clark and Ellison, 2000; Moore *et al.*, 2006; Blackwell *et al.*, 2007; Delarue *et al.*, 2009). The same deployments were also conducted in 2009 and 2010, when no large-scale airgun surveys occurred nearby; however, signals from more distant airgun surveys were present, along with signals generated by shallow hazard surveys.

Every year the acoustic data contained hundreds of thousands of whale calls. The scale of the dataset, combined with a need for timely analysis, motivated the development of methods for automatically detecting, classifying, and localizing bowhead whale sounds, even during active seismic surveys. This paper describes the multi-stage automated algorithm that has been developed to process these multi-year data sets.

After reviewing relevant details of the DASARs and their deployment in Sec. II, Sec. III reviews previous work on automated bowhead whale call detection and describes the characteristics of bowhead whale sounds and interfering acoustic sources, including seismic airguns and other marine mammals. Section IV presents the six-stage algorithm in detail, and Sec. V presents the results of processing four seasons of data with the algorithm. Section VI discusses the observed false alarm rates of the automated procedure and considers whether a portion of these "excess calls" are actually legitimate whale calls overlooked by the manual analyses. A detailed comparison of the performance of the automated vs the manual results is reserved for a companion paper, as the methods for evaluating the statistical similarity

[a)]Author to whom correspondence should be addressed. Electronic mail: athode@ucsd.edu

 0001-4966/2012/131(5)/3726/22/$30.00

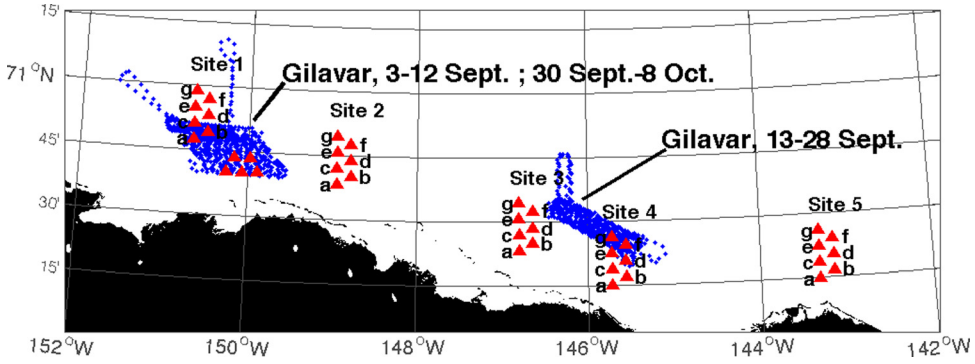

FIG. 1. (Color online) Stereographic projection of Directional Autonomous Seafloor Acoustic Recorders (DASARs) deployment geometries in the Beaufort Sea, 2008. The DASARs at each site are roughly 7 km apart. The deployments in other years are similar, except that the five unlabeled DASARS arranged as a trapezoid near site 1 are absent. GPS tracks of the seismic vessel Gilavar are shown whenever airguns are firing between September 8 and October 8, 2008.

between two temporal-spatial distributions require considerable exposition.

II. EQUIPMENT

A. DASAR description

The acoustic data were recorded on DASARs (Greene *et al.*, 2004), autonomous acoustic recording packages equipped with an omnidirectional acoustic pressure sensor (−149 dB re V/1 μPa) and two horizontal directional sensors capable of measuring the north-south and east-west components of acoustic particle velocity, an arrangement that permits the azimuth of bowhead whale sounds to be measured from each DASAR. All time series are sampled at 1 kHz and have a maximum usable acoustic frequency of 450 Hz. The GPS-synchronized time is noted to within a second whenever each DASAR is started and stopped. In addition, shortly after the deployment and before the recovery of every DASAR, a calibrated frequency-modulated signal is sequentially broadcast at roughly 4 km range from three to six positions around each DASAR. These playback start times are also time-stamped with GPS data, which are used to correct each DASAR's clock drift during post-processing. Coincident bearings to calls are combined to localize them using the methods discussed in Sec. IV G. The bearing uncertainty of the DASARs, derived by comparing the acoustic intensity measured along orthogonal directions, is around 15°–20° for signals with signal-to-noise ratios (SNR) of 5 dB or less, and 1°–2° for signals with SNR greater than 10 dB.

B. Deployment geometry and timelines

During the 2007 through 2010 field seasons, between 35 and 40 DASARs were deployed across a 280 km swath off the Alaskan North Slope, on the continental shelf in water depths between 20 and 53 m. The deployments were grouped into "sites," labeled 1–5 traveling from west to east (Fig. 1). Most sites contained seven DASARs deployed in a triangular grid with 7 km separating the DASARs. The southernmost DASAR was labeled "a," and the northernmost labeled "g," and the label "DASAR 5g" refers to DASAR data collected at site 5, location g. In general, the mean depth at each site increased from south to north, and among sites increased from west to east. For example, the deepest DASAR at sites 1–5 sat at 23, 32, 41, 41, and 53 m depth, respectively.

The data analyzed here were recorded during the dates reported in Table I. There was significant local airgun survey activity in 2007 and 2008, during dates also reported in Table I. In 2009 and 2010, there were fewer or no local airgun surveys, but up to three distant seismic surveys were present, similar to the number of distant surveys also detected in 2008.

III. BACKGROUND

A. Bowhead whale call diversity and previous automation research

The primary challenge in automating the passive acoustic detection and localization of bowhead whale calls is coping with the diversity of their calls. The complete repertoire of bowhead whales is highly variable and difficult to

TABLE I. Significant dates discussed in paper.

Year	Deployment dates [total days]	Dates of significant airgun survey activity [sites]	Dates used to train neural networks [total days, time of day, network trained]	Dates used for manual validation in Fig. 10 [total days, time of day]
2007	Aug. 21–Oct. 11 [51]	Sept. 17–Oct. 3 [3 & 4][a]	None	Aug. 8; Sept. 4, 5, 12, 17, 25, 29; Oct. 2 [8 days, midnight–noon]
2008	Aug. 19–Oct. 3 [45]	Sept. 3–12 [1] Sept. 13–28 [3 & 4] Sept. 30–Oct. 8 [1][a]	Aug. 21, 28; Sept. 6, 13, 21, 29 [6 days, midnight–noon, first network]	Same as training data
2009	Aug. 19–Oct. 6 [48]	Distant only	Aug. 27; Sept. 3, 8, 12,[b] 14, 18, 25, 30 [8 days, midnight–noon, second network]	Same as training data[c]
2010	Aug. 6–Oct. 6 [61]	Distant only	None	Aug. 15, 21, 29; Sept. 5, 13, 27 [6 days, midnight–noon]

[a]Up to three distant surveys also logged during all dates.
[b]Particular focus on pinniped sounds logged between 0500 and 0700 on site 2.
[c]Figure 11 uses data from noon to midnight.

organize into discrete classes. A large portion of their calls consists of frequency-modulated (FM) sounds between 25 and 500 Hz, which vary substantially in modulation pattern and frequency range (Clark and Johnson, 1984) and can contain concurrent higher-frequency sidebands. The first four spectrograms in Fig. 2 display examples of these bowhead whale calls.

Other types of signals include "pulsed-tone" and "amplitude/frequency-modulated (AM/FM)" calls, both of which appear as multiple frequency-modulated contours in spectrograms where the FFT length of the spectrogram is greater than the pulse interval or modulation rate (Clark and Johnson, 1984). Under such conditions, pulsed-tonal signals appear as closely spaced "combs" of contours 50 Hz or less apart, and AM/FM signals display one or two sidebands with similar frequency separations. Sometimes an individual animal will generate long sequences of similarly-modulated sounds, which in fall seasons have been labeled "call sequences" (Clark and Johnson, 1984; Blackwell *et al.*, 2007). Bowhead "songs" (e.g., Stafford *et al.*, 2008; Delarue *et al.*, 2009) have not been detected during the fall Beaufort migration discussed here.

In subsequent discussion it becomes convenient to define a "local," "global," and "total" bandwidth of an FM or AM call with sidebands or harmonics. The "local" bandwidth is the frequency range of the most intense FM component of a received call measured over a short duration of time, typically the FFT length used to create a spectrogram. The "global" bandwidth is the frequency range traversed by the midpoint of a given FM band throughout a call. For example, in Fig. 2(b), the lowest frequency component of the harmonic call descends from 162 Hz down to 75 Hz over 1 s, and so this component has a global bandwidth of 87 Hz. The "total" bandwidth is the difference between the maximum and minimum frequency attained by any harmonic or sideband throughout the entire duration of a call, and thus defines the vertical dimension of a "bounding box" that could be placed around a complete call on a spectrogram.

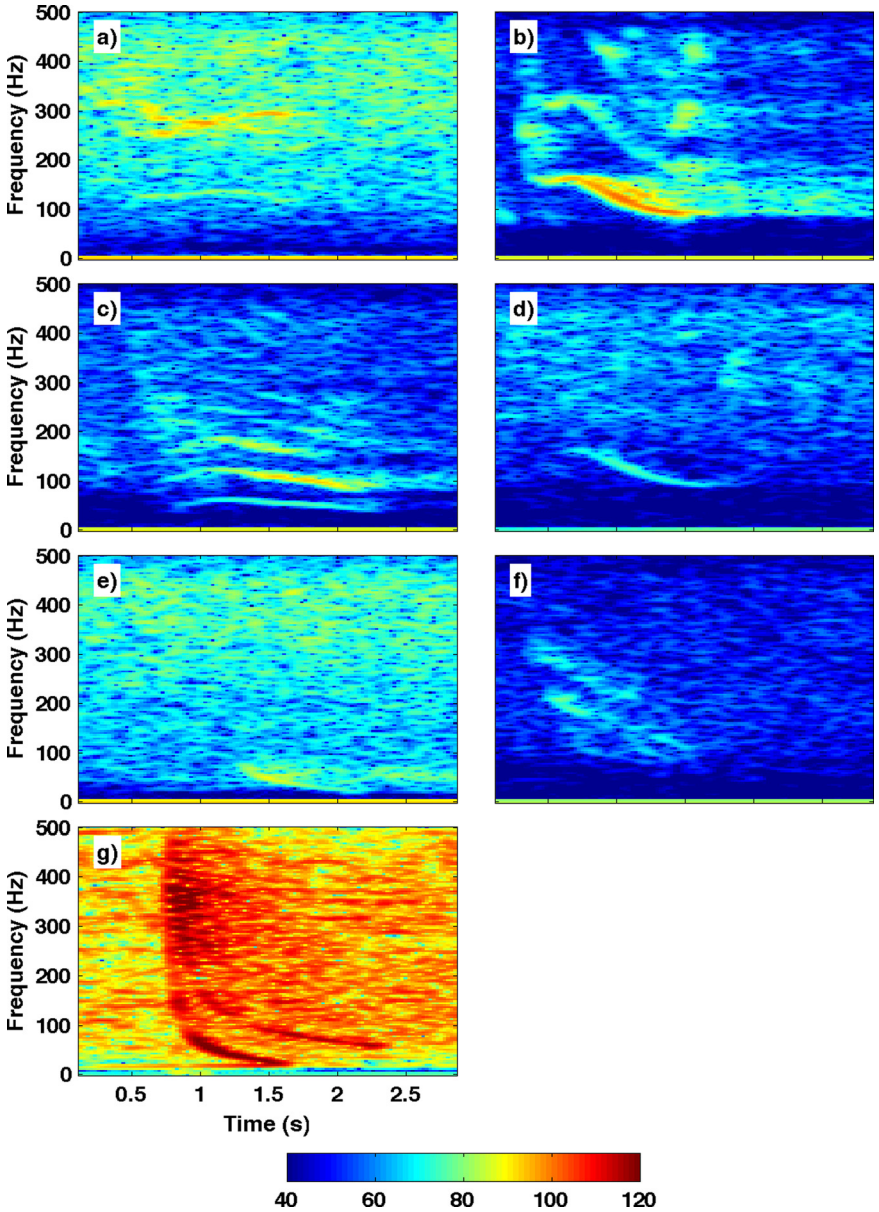

FIG. 2. (Color online) Spectrograms of bowhead and airgun sounds from the Beaufort Sea (256 pt FFT, 90% overlap): (a) two bowhead whale calls covering different frequency ranges during relatively high ambient noise conditions, DASAR 2g, 2008; (b) downswept whale call with harmonics, multipath, and reverberation, DASAR 2f, 2008; (c) "n-shaped" bowhead whale call with harmonics, DASAR 2f, 2008; (d) same call as (b), recorded on DASAR 2a, 2008;(e) distant airgun signal, DASAR 5g, 2008; (f) distant airgun signal, DASAR 2e, 2009; (g) strong airgun signal, DASAR 3g, 2008, generated at 21 km range. Intensity levels are in units of dB re 1 μPa^2/Hz.

Thus, in Fig. 2(b), the total bandwidth of the fundamental and first harmonic is slightly greater than 200 Hz.

Some of the earliest attempts to automatically detect marine mammal sounds used bowhead whale calls, but relatively little public peer-reviewed literature exists. In 1993, Weisburn *et al.* used a matched filter and an early version of a hidden Markov model (HMM) (Rabiner, 1989) to distinguish 114 bowhead sound samples from other transient sounds and white Gaussian noise. Several years later, Mellinger *et al.* compared the performance of spectrogram correlation methods, HMM, neural networks, and matched filtering when applied to a subclass of sounds called "endnotes," culled from bowhead call sequences recorded off Barrow, AK (Mellinger and Clark, 2000). After determining that spectrogram correlation methods had a higher performance than an HMM, the authors conducted a more extensive comparison between the spectrogram correlation method, a neural network, and a matched filter. Their neural network used 192 inputs that contained the time-frequency bins of an 11×21 spectrogram grid centered around the call in question. Mellinger *et al.* concluded that the neural network had the highest overall performance in terms of combining the error rate between "false alarms" and "missed calls," but that the spectrogram correlation approach required less training data and was more intuitive to understand than the neural network.

Neural networks were again applied to bowhead endnotes by Potter *et al.* (1994); here, the entire spectrogram was also used as an input into a three-layer network, and the neural network demonstrated a better overall performance than spectrogram correlation methods. More recently, Mouy *et al.* (2008) presented results that used Gaussian mixture models (GMM) to distinguish bowhead calls from other biological signals in the Chukchi Sea, using band-averaged cepstral coefficients and Daubechies wavelets to extract relevant features to input into the GMM. When testing the classifier on the original training data set, they found that classification based on cepstral coefficients correctly identified 80% of a sample size of 275 calls, with a 2% false alarm rate. Heimlich *et al.* (2009) has also applied a variety of classifiers to data collected in the Beaufort Sea in 2007 and 2008.

B. Seismic survey signals and other interfering transients

Another fundamental challenge provided by this data set is the presence of concurrent signals that are not bowhead whale sounds but are also variable in terms of duration, bandwidth, and frequency modulation. The most common type of interference arises from airgun surveys, which include not just the local SEPCO surveys but also distant seismic survey activity in the Arctic basin. At times, up to three seismic surveys were detected simultaneously [Fig. 3(a)]. The airgun pulses produced during seismic surveys occur at regular intervals (typically between 10 and 15 s) and contain energy distributed over the entire DASAR analysis band of 10–450 Hz.

While designed to be repetitive and reproducible, the far-field signature of an airgun array is azimuthally dependent, and the time-frequency structure of its signal alters substantially as it propagates long distances through shallow water, morphing from pulses into extended FM downsweeps.

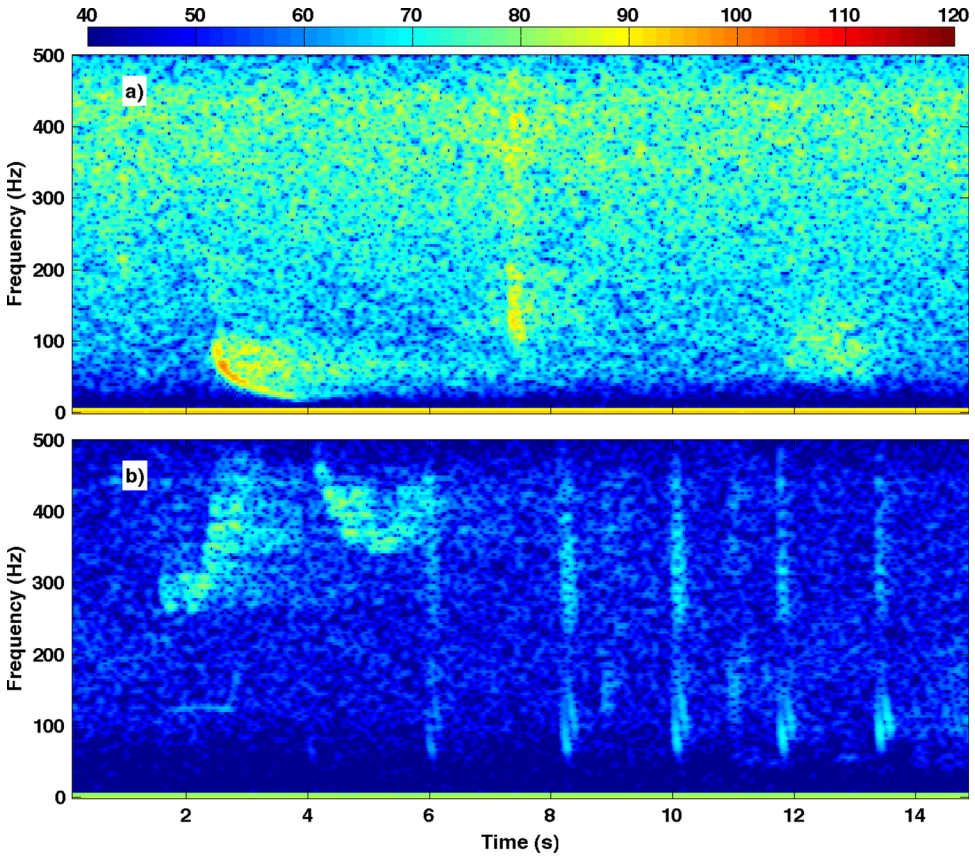

FIG. 3. (Color online) (a) Airgun examples from three simultaneous seismic surveys in the Arctic, DASAR 5g, 2008, visible at 3, 7.5, and 13 s. (b) Examples of bearded seal and other biological signals, DASAR 2e, 2009. Intensity levels are in units of dB re 1 μPa^2/Hz.

Indeed, a single airgun shot can split into three or four separate downsweeps due to geometric dispersion effects in the waveguide, with each propagating mode yielding a different FM downswept structure. While this structure can be exploited to estimate source range (D'Spain *et al.*, 1995; Thode, 1999; Wiggins *et al.*, 2004), these multipath effects are a nuisance for automated detection because the structure of the received signal from a nearby survey can vary considerably over time as the range to the survey varies. Figures 2(e) and 2(f) illustrate examples of distant airgun signals, while Fig. 2(g) illustrates a relatively close airgun pulse generated 21 km from a DASAR. Note how the two modal arrivals in Fig. 2(g) persist substantially longer than the higher-frequency pulsed arrival and how their global bandwidths and slopes are similar to those of bowhead whale calls.

Another source of interference is provided by other biological sounds [Fig. 3(b)], including sounds produced by bearded seals (*Erignathus barbatus*), walrus (*Odobenus rosmarus divergens*), and possibly ringed seals (*Phoca hispida*) and gray whales (*Eschrichtius robustus*). Whereas bearded seal calls are usually higher in frequency than most bowhead calls and readily distinguishable [at least by ear (Risch *et al.*, 2007)], walruses and bowhead whales can be difficult to distinguish even by trained listeners.

Ship noise is a third, but rare, form of interference. Unlike bowhead calls, ship noise is generally more continuous; however, this noise is also typically non-stationary and can generate acoustic fluctuations that can appear as transient signals.

IV. AUTOMATED PROCEDURE

A. Overview

The complete post-processing automated procedure is subdivided into six stages, plus a data preprocessing stage that converts the raw acoustic data into a form more amenable for automated analysis (Fig. 4). The first four stages are applied independently to data from each DASAR, and consist of (1) applying an "event detector" to flag any potential transient event of interest, (2) applying an "interval filter" that removes from further consideration a significant fraction of airgun pulses from distant and close range airgun surveys, (3) running an image processor that extracts 25 descriptive features from an equalized spectrogram of candidate detections, and (4) exploiting two cascaded feed-forward neural networks to winnow candidate detections based on their feature values. The remaining stages combine the neural network outputs from all DASARs at a site by (5) matching detected calls between DASARs and (6) computing the position of the whale by triangulating the geographic bearings computed from the matched call sets.

The fourth stage, which uses neural networks, required training data provided by manual analyses in order to adjust the network weights and output thresholds. The training data were obtained by running the first three stages on subsets of acoustic data from 2008 and 2009 that had been reviewed manually for bowhead whale calls. A comparison of the automated results with the manual data divided the automated results into appropriate "whale" and "non-whale" classes, producing the training sets. Section IV E 1 details the construction of these large-scale training sets, while also discussing some subtleties that arose. Once the networks had been trained and their weights fixed, all stages were then applied to the complete four-year acoustic data sets.

Each of the following sections describes the six stages in more detail. Table II summarizes the parameters used across all stages. These parameters have not been systematically optimized; the parameters for the first two stages (event detection and interval filtering) were estimated using local (direct-search) optimizations on selected days of 2008 data,

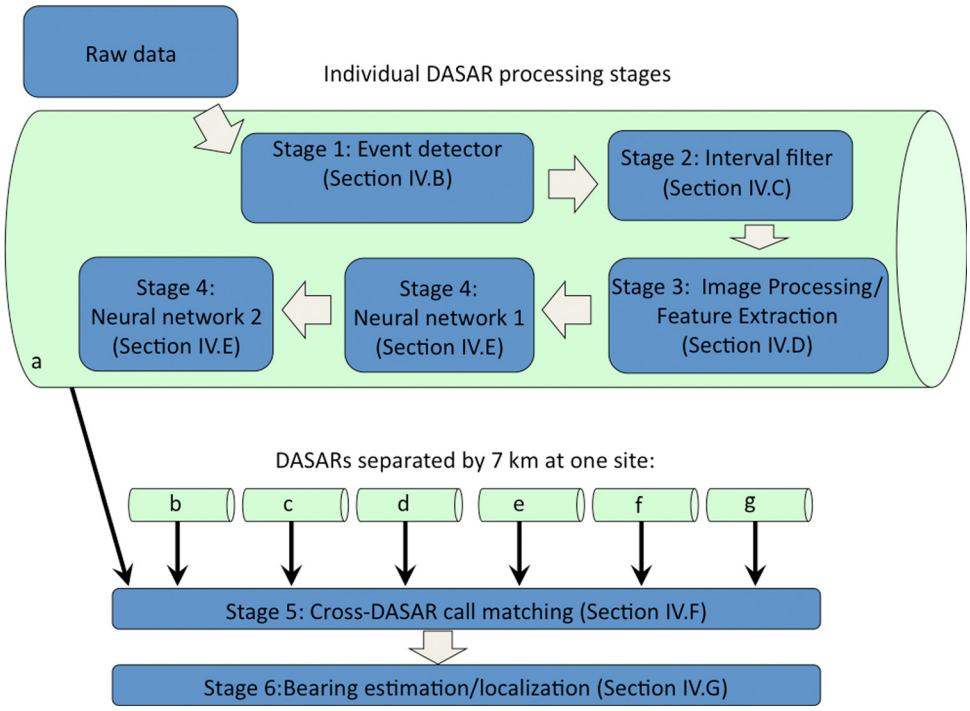

FIG. 4. (Color online) Schematic of automated detection, classification, and localization algorithm for one site.

TABLE II. Parameters and values used in the algorithm.

Event detection (Sec. IV B)	
Parameter	Value [range]
FFT length	256 samples
Percent overlap	50%
Minimum frequency	25 Hz
Maximum frequency	500 Hz
Equalization time	24 s
Detector bandwidth	37 Hz
Detector bandwidth overlap	50%
SNR threshold	6 dB
Minimum time between events	50 ms
Event duration	[0.1–6] s
Interval detection (Sec. IV C)	
Interval range	[5–42] s
Number of interval tests per candidate interval	20
Bearing tolerance	15°
Minimum match fraction	8/20 = 0.40
Timing tolerance of kth interval	$\sqrt{k}$ (s)
Maximum deviation from adjacent intervals permitted	0.4 s
Fraction of adjacent intervals that must lie within maximum deviation	7/20 = 0.35
Image processing (Sec. IV D)	
Thresholding	
Min SNR of ridge peak	13 dB
Maximum bandwidth	250 Hz
dB difference between ridge peak and edge	3 dB
Contour threshold	10 dB
Morphological processing	
Minimum time-bandwidth product	1 s Hz (8 pixels)
Ridge closing element bandwidth	11 Hz
Ridge closing element duration	40 ms
Contour closing element bandwidth	20 Hz
Contour closing element duration	100 ms
Crude feature filtering	*Permitted range*
Ridge duration	[0.15–6.0] s
Global contour bandwidth	[0–300] Hz
Maximum contour frequency	[20–400] Hz
Minimum contour frequency	[5–500] Hz
Segment splicing	
Maximum time gap	100 ms
Maximum frequency gap	20 Hz
Overtone matching	
Minimum overlap	25%
Maximum band separation	50 Hz
Neural network Processing (Sec. IV E)	
Number of networks	2
Input features	25
Hidden units	10 each
Default network output thresholds for both networks	[−0.8; 0.8]
Cross-DASAR call linking (Sec. IV F)	
Image correlation threshold	0.42
Physically permissible time window	±7 s
Bearing estimation and localization (Sec. IV G)	
Bootstrap sample size for bearing uncertainty estimate	100
Huber localization tuning constant	1.5

the first data set available when the procedures were first being developed. The parameters for the image processing and cross-DASAR call matching stages were estimated via trial-and-error work on selected 2008 data sets as well. The parameters for the localization stage were adopted from parameters reported in (Greene *et al.*, 2004).

B. Event detection

The first automated stage simply seeks to flag any transient "event" that occurs in the acoustic data, using a version of the cell averaging/clutter map constant false alarm rate detector (CFAR) (Nitzberg, 1986; Levanon, 1988), also known as an "energy detector" in the bioacoustics literature (Mellinger, 2002). First, a running spectrogram is created, using fast Fourier transforms (FFTs) of 256 samples (0.256 s) overlapped 50%. (Note that the spectrograms in Figs. 2 and 3, 5, and 6 are computed using 90% overlap to provide greater visual clarity.)

Next, a set of "detection functions" $D(f_n,t)$ is created by integrating the FFT output over a set of 37 Hz frequency bands f_n between 10 and 450 Hz, with 50% overlap between the detection bands. Whenever a new FFT sample is acquired at time t_i, the set of detection functions is updated. The new value of each detection function, divided by the current value of a "background noise equalization" function $B(f_n,t)$, is compared to a threshold of 6 dB (or 6.3 in terms of linear signal-to-noise ratio). If the new value exceeds the threshold, then the detector is "triggered," a new event is flagged, and the equalization function is not updated. If the new value does not exceed the threshold, then the equalization function is updated via the following expression:

$$B(f_n, t_{i+1}) = (1 - \alpha)D(f_n, t_{i+1}) + \alpha B(f_n, t_i). \quad (1)$$

The value of α is defined such that the contribution of a new sample to B will decay to 5% of its original contribution in 24 s, yielding $\alpha = 0.982$. Thus, the equalization function B becomes a running exponentially weighted average of the "smoothed" background noise levels, with an adaptive response of about half a minute. Once a single detection function has been triggered, the triggering of additional detection functions at other bandwidths will not initiate new events. Once all detection functions triggered by an event have fallen below the threshold, the duration of the event is derived. If the event's duration is longer than 100 ms, it is logged for further analysis. A new event is prohibited from being flagged until 50 ms have elapsed since the end of the previous event.

The reason a set of narrowband detection functions is used, instead of a single broadband detection function, is that the former approach should permit a larger SNR threshold to be defined, emphasizing the detection of locally narrowband signals vs signals that have energy distributed across a wide bandwidth at the same time (i.e., locally broadband pulse). Note also that a single event may actually harbor one or more distinctive signals from multiple sources, if the received signals overlap in time at that DASAR, even if they do not overlap in frequency.

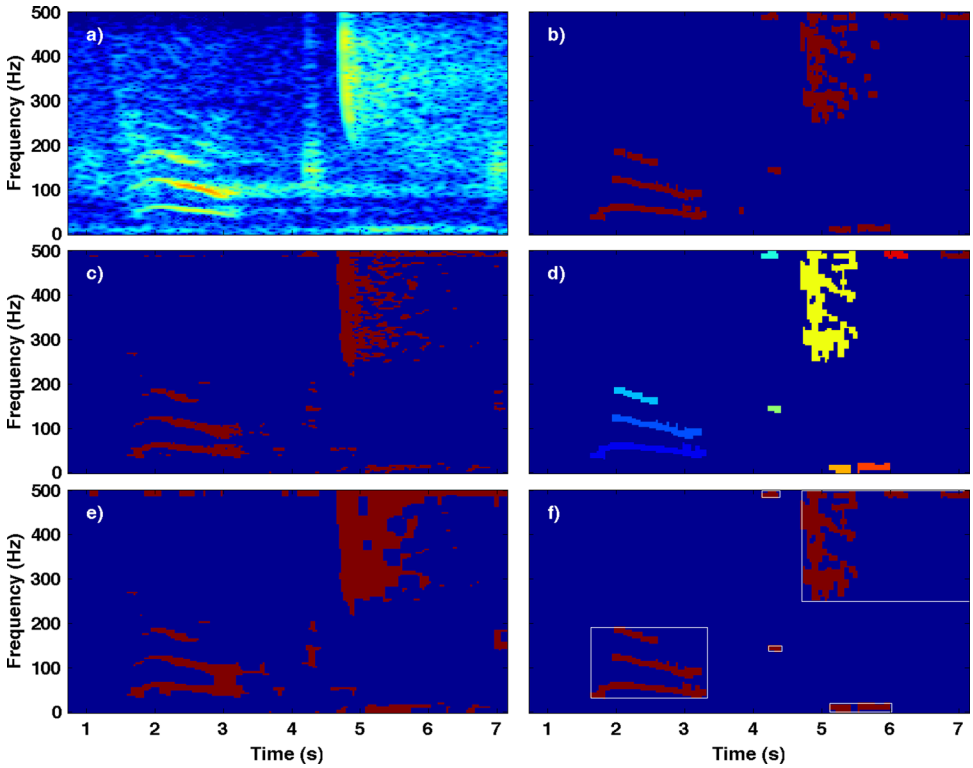

FIG. 5. (Color online) Key steps in the image processing stage, demonstrated on the harmonic whale call in Fig. 2(c), along with an airgun pulse after 4 s: (a) Original spectrogram (256 pt FFT, 90% overlap); (b) "ridge image" of equalized spectrogram using 13 dB ridge threshold; (c) "contour image" of equalized spectrogram using 10 dB SNR contour threshold; (d) labeled ridge image after morphological opening, closing, and connected-component labeling; (e) contour image after morphological opening and closing; (f) final "transients" indicated by the large white boxes, comprised of ridge segments linked using the methods described in Secs. IV D 4 and IV D 5. Note how the harmonics of the bowhead whale call at 2 s have been successfully "linked."

C. Interval filtering

The second automated stage seeks to determine whether a set of events occurs at regular intervals from a consistent direction and thereby can be flagged as a sequence of airgun pulses. The goal of this stage is to remove the numerous, low-intensity, short-duration, airgun pulses from distant airgun surveys that are ubiquitous among the acoustic records during the Beaufort Sea summer [e.g., Fig. 3(a)]. This stage begins by marching through every first-stage event and computes its geographic bearing, using the methods to be discussed in Sec. IV G. For a particular "current" event, the program searches 5 to 40 s into both the past and future for the presence of any other events that arrived within 15° of the azimuth logged for the current event. (This 30° swath is needed to guarantee detection of most pulses from distant airgun surveys; bearing estimates from these surveys have

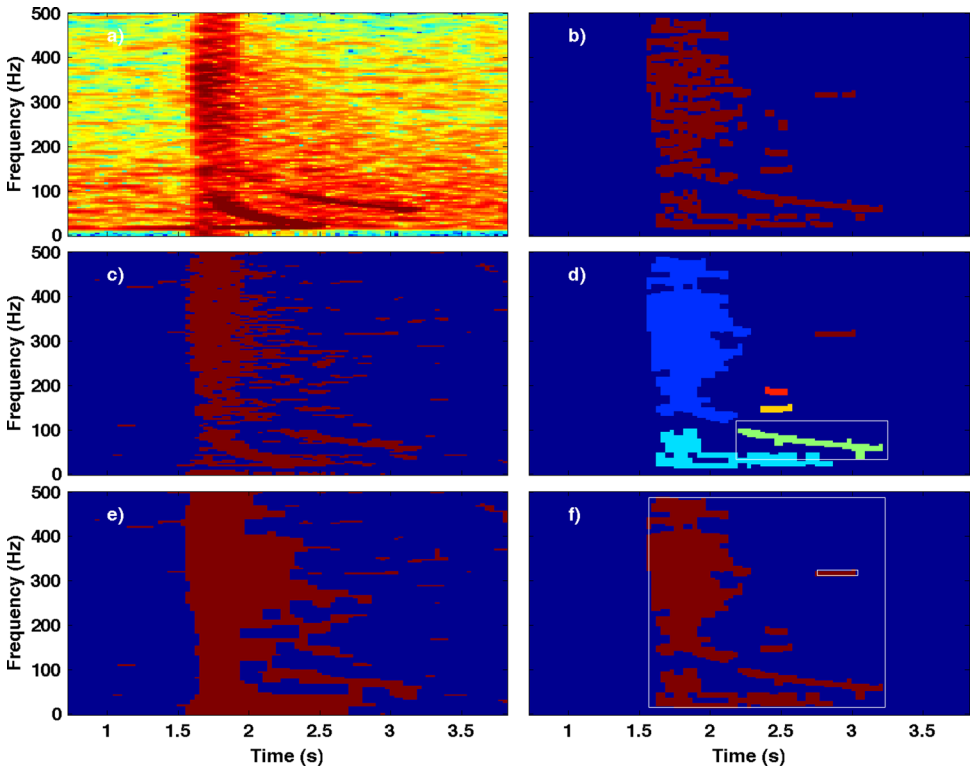

FIG. 6. (Color online) Same as Fig. 5, but with steps performed on the strong airgun signal shown in Fig. 2(g). The small white box in (d) highlights a ridge component that would have been flagged as a bowhead call had it not been linked to the main pulse arrival through the shared contour segment in (e). The small white box in (f) indicates a ridge segment that was not linked to the airgun pulse via a common shared contour segment in (e).

standard deviations of around 20°). These other events, if they exist, provide a set of "candidate" pulse intervals to test. Each candidate interval is individually tested by searching for the presence of other events at ten intervals into the future and past, relative to the current event. A candidate interval is awarded a "hit" if an additional event with an appropriate bearing is present within $\sqrt{K}$ s of a given candidate interval prediction, where K is the number of pulse intervals being projected either forward or backward in time from the current event. If a "hit" is logged, the candidate pulse interval estimate is recalculated to incorporate the measured detection time of the new event, thereby reducing the variance of the pulse interval estimate. If eight or more of the 20 candidate interval test times yield "hits," then the current event is flagged as an airgun pulse and assigned the candidate pulse interval. The program then advances to the next event, relabeling it as the current event, and repeats the entire procedure.

During times of heavy bowhead whale call activity, such as during bowhead whale call sequences, or during times when several bowhead whales are calling from the same azimuth, whale calls can be assigned a pulse interval; however, these assigned intervals usually have values that differ substantially from nearby interval values, because even call sequences do not display the degree of timing regularity exhibited by airgun pulses. Therefore, after a first pass through all events, the procedure checks that an interval assigned to an event lies within 0.4 s of at least seven out of 20 intervals assigned to nearby events. Events that pass this consistency check are flagged as airgun pulses and removed from further consideration.

While the parameters of this stage have been set with the goal of minimizing the rejection of true whale calls rather than eliminating all airgun pulses, Sec. V A will illustrate how this relatively simple stage can still remove between 25% and 75% of airgun signals before executing the more computationally expensive stages.

D. Image processing and feature extraction

The third automated stage, image processing and feature extraction, is the workhorse of the algorithm, requiring nearly 80% of the total processing time. There is a growing literature on using image processing and other techniques to extract features from frequency-modulated signals (Sturtivant and Datta, 1995; Datta and Sturtivant, 2002; Lammers *et al.*, 2003; Oswald *et al.*, 2007; Roch *et al.*, 2007; Asitha *et al.*, 2008; Madhusudhana *et al.*, 2008; Top, 2009), but this area is still an active research topic (Lampert and O'Keefe, 2010a,b), and methods for handling sidebands remain underdeveloped (Heller and Pinezich, 2008).

During this stage a time series stored from the output file generated by the event detector and interval filter is read into memory and converted into a spectrogram, using a 256-point FFT and 90% overlap between samples. In the following discussion the spectrogram will often be described as an "image," stored in a matrix with the rows designating frequency and the columns designating time, with a "pixel" being a specific time-frequency element in the spectrogram. This complex stage has multiple steps: (1) spectrogram creation and equalization; (2) generation of two binary images via "ridge" and "contour" thresholding; (3) morphological processing; (4) connected component labeling; (5) binary and SNR-weighted feature extraction; (6) ridge segment splicing and harmonic/sideband linking; and (7) contour/ridge association and final feature vector assembly. Figures 5 and 6 illustrate some of the key steps using the whale call and the close-range airgun pulse illustrated in Figs. 2(c) and 2(g), respectively. An excellent introduction into many of the techniques used in this stage, including thresholding techniques and morphological processing, can be found in Gonzalez and Woods (2002).

1. Spectrogram equalization and double-thresholding

After the spectrogram of each event is generated, it is equalized using the stored values of the equalization function B [Eq. (1)], producing an equalized image B_{eq} with pixel values in units of SNR. The application of Gaussian and other low-pass spatial filters to B_{eq} was initially explored (Gillespie, 2004), but spatial filtering was found to yield little or no improvement in performance, as subsequent morphological processing methods achieved the same goal.

This equalized continuous-scale image is converted into two binary images using two different thresholding methods. Due to the equalization process the thresholds can be expressed in terms of SNR. The ridge threshold approach cycles through each B_{eq} column, seeking local maxima with respect to frequency whose values are at least 13 dB SNR. The threshold then selects all vertical (frequency) pixels contiguous to each local maximum that lie within 3 dB of that maximum. The result is a "ridge image" [Figs. 5(b) and 6(b)] that traces high-resolution ridges of frequency-modulated narrowband signals. The contour thresholding method simply selects all pixels with values greater than 10 dB SNR [Fig. 5(c) and 6(c)]. The Niblack thresholding method described in the work of Asitha *et al.* (2008) is not used in this work, because no reliable *a priori* knowledge exists of what percentage of a typical spectrogram would consist of noise, and this percentage is strongly time-dependent.

There are two motivations for using the parallel thresholding methods. The first is to produce a larger set of potential features to extract from a spectrogram, since contrasts between features extracted from the two binary images can be useful for discriminating bowhead calls from other signals. For example, a tonal FM bowhead call should yield similar local bandwidths when subjected to the different methods [e.g., Figs. 6(b) and 6(c)], while broadband impulsive sounds, such as airgun signals, can yield different measurements. The ridge image of an impulsive sound will contain many narrowband fragments that lie over local maxima in the pulse spectrum, while the "contour image" will consist of a single broadband segment; thus, the ridge and contour bandwidths will be substantially different. The second motivation arises from the use of the contour image to link components of the ridge images together, providing a simple form of contextual-based processing. This last motivation is explained in more detail in Sec. IV D 5.

2. Morphological processing and connected component labeling

A morphological opening procedure (Gonzalez and Woods, 2002; Asitha *et al.*, 2008) is first applied to each image to remove components comprised of eight pixels or less, which corresponds to a time-bandwidth product (or spectrogram "area") of 1, given the chosen spectrogram parameters. Because calls often vary in intensity over their duration, the binary ridge image of the call is often split into a set of disconnected components. Thus, a morphological closing operation is next applied to both images, filling small gaps between components by using a rectangular structural element that spans 11 Hz bandwidth and 40 ms of time (or 3 frequency pixels and 2 time pixels in the spectrogram). Figures 5(d) and 6(d) and 5(e) and 6(e) display the results of the morphological operations on both types of binary images.

Finally, "eight-connected" components in both images (i.e., collections of pixels that adjoin each other horizontally, vertically, and/or diagonally) are labeled using a run-length encoding procedure, followed by the construction of an adjacency matrix to compute the "equivalence classes" (Haralick and Shapiro, 1992). The net result is the transformation of the binary image into a "labeled" image [Figs. 5(d) and 6(d)], where each set of connected components, or "segment," has been assigned a unique integer. The following text will refer to a "ridge segment" as a segment derived from the ridge image, etc., with the white box in Fig. 6(d) illustrating an example of a ridge segment that superficially resembles a FM bowhead whale call. Ridge segments that are less than 0.15 s or greater than 6 s are discarded at this point. The labeled image becomes fundamentally important in Sec. IV F, where whale calls are matched between DASARs.

3. Binary and SNR-weighted ridge feature extraction

Various features can now be extracted from each ridge image segment, and Table III lists these features, subdivided into two categories. Binary features are derived from the binary images alone, while SNR-weighted features are obtained by using the binary image to mask the equalized spectrogram B_{eq}, and then computing features using the SNR values of B_{eq} as weights.

Binary features include each segment's area (related to the time-bandwidth product), duration, and global bandwidth. Other segment features include (i) the median local bandwidth, which is the median of the set of local bandwidths measured across each column of the segment, (ii) the eccentricity, which is defined as the eccentricity of an ellipse that has the same normalized second central moments as the segment, (iii) the orientation, which is the angle between the ellipse's major axis and the horizontal, and (iv) the solidity, a quantitative measure of compactness, defined as the ratio between the area of the segment (i.e., the number of pixels) and its convex hull. A heuristic definition of the last term is the convex area (in pixels) that would lie within a rubber band placed around the segment. Thus an image of a ring has low solidity, while a filled circle has high solidity. In principle, a moderately frequency-modulated bowhead call would be expected to have a high solidity and eccentricity, and relatively low orientation and median local bandwidth. A highly frequency-modulated ("n" or "u"-shaped) call segment will have lower solidity and eccentricity.

TABLE III. Feature inputs into neural networks.

Binary features, (primary) ridge segment with largest time-bandwidth product	Binary features, linked ridge segments
Orientation (deg.)	Total minimum frequency (Hz)
Eccentricity	Total maximum frequency (Hz)
Solidity	Total duration (s)
Duration (s)	Total SNR (dB)
Median local bandwidth (Hz)	Total SEL (dB re 1 μPa^2-s)
Time-band product (s-Hz)	
Number of overtones (sidebands, harmonics)	
Band spacing (Hz)	
SNR-weighted features, (primary) ridge segment with largest time-band product	**Binary features: contour segment associated with primary ridge segment**
Median weighted local kurtosis	Time-bandwidth product (s-Hz)
Median weighted local bandwidth (Hz)	Global bandwidth (Hz)
Weighted minimum frequency (Hz)	Duration (s)
Weighted maximum frequency (Hz)	Median local bandwidth (Hz)
SNR (dB)	Minimum frequency (Hz)
SEL (dB re 1 μPa^2-Hz)	Maximum frequency (Hz)

The SNR-weighted features use weights w defined from the masked equalized spectrogram B_{eq} as follows:

$$w(f,t) \equiv B_{eq}(t,f) - \min\left[B_{eq}(f,t)\right], \tag{2}$$

where the global minimization is performed with respect to all time/frequency values of B_{eq}. The local mean frequency is then

$$\bar{f}(t) = \int_{f_1}^{f_2} w(f,t) f \, df \Big/ \int_{f_1}^{f_2} w(f,t) \, df, \tag{3}$$

where f_1 and f_2 are the lower and upper frequencies of a binary segment at time t, with $t=0$ defined as the start of the segment. The weighted minimum and maximum frequencies of a segment are simply the minimum and maximum values obtained by Eq. (3) over the duration of a given segment, and the weighted start and end frequencies are the values of Eq. (3) at the start and end of the segment. While the slope and curvature of a segment can also be inferred from the local mean frequency, they are not used here, due to the high variability in bowhead whale call modulation structure. The median weighted local bandwidth (MWLB) of a segment is defined as

$$\text{MWLB} = \text{median}\left(2 \sqrt{\frac{\int_{f_1(t)}^{f_2(t)} w(f,t)\left[f - \bar{f}(t)\right]^2 df}{\int_{f_1(t)}^{f_2(t)} w(f,t) \, df}} \right), \tag{4}$$

where the median is taken with respect to all time samples of the segment. The MWLB is a more robust estimate of the local bandwidth than the binary local bandwidth. Finally, the median weighted kurtosis (MWK) is defined as

$$\mathrm{MWK} = \mathrm{median}\left(\frac{\int_{f_1(t)}^{f_2(t)} w(f,t)[f-\bar{f}(t)]^4 df \Big/ \int_{f_1(t)}^{f_2(t)} w(f,t)\, df}{(MWLB/2)^4} \right), \tag{5}$$

where the median is taken with respect to all time samples of the segment.

4. Ridge segment splicing and linking

While the morphological closing operation in Sec. IV D 2 merges small ridge segments together, additional steps are necessary to "splice" together short breaks in calls. If the weighted end frequency of one ridge segment lies within 0.1 s and 20 Hz of the start of a second segment, both segments are assigned the same integer in the labeled image, and their features are merged and/or recomputed as necessary.

Ridge segments can also be related via harmonics, pulsed tone frequency combs, and AM sidebands. These "parallel" segments are "linked" whenever segments overlap in time by more than 25% and the median of the separation of their local mean frequencies is less than 50 Hz. These criteria are representative of high-SNR harmonic bowhead signals. Unlike spliced segments, the features of parallel segments are not merged; they retain unique integers in the labeled image, but are now identified as part of the same "transient" event. Figure 5(f) demonstrates how the "parallel" ridge segments of a harmonic bowhead whale call in Fig. 5(d) have been linked together (indicated with a white box) using parallel linking criteria.

Seven new features are extracted from the complete linked transient: the number of linked segments, the "total" frequency minimum and maximum of the transient, the "total" transient duration, the mean frequency separation between the segments, and the total transient signal-to-noise ratio (SNR) and sound exposure level (SEL) (Madsen, 2005), also known as "energy flux density." The inclusion of the last two features is important; for example, the duration and bandwidth of a given whale call typically shrink as SNR decreases, so the measured bandwidth and duration of the same whale call will change depending on receiver distance [e.g., Figs. 2(b) and 2(d)]. A classifier needs to incorporate measured received levels and SNR in order to account for this effect.

5. Contour/ridge linking and feature vector assembly

The previous two steps (feature extraction and segment splicing/linking) are only performed on the ridge image in order to reduce processing time. In this step, linked ridge segments are associated with their corresponding contour segments, and previously unlinked ridges that share a common contour segment are now linked into a single transient, thereby changing the "total" minimum and maximum frequencies, "total" duration, and SNR and SEL estimates. Further features are extracted from the contour segment encompassing the primary ridge segment (Table III). Subplots (d), (e), and (f) in Figs. 5 and 6 demonstrate how the contour image of airgun pulses can be used to link ridge segments arising from the same pulse. For example, five distinct ridge segments in Fig. 6(d) (including the bowhead-like contour highlighted by the white box) cannot be linked via the approaches presented in the previous subsection. However, Fig. 6(e) shows that the five segments share a common contour segment, and so they become linked together (assigned the same color) within the white bounding box in Fig. 6(f). However, the second small white box in Fig. 6(f) indicates a ridge segment that was not associated with the large airgun contour segment, and thus will pass on to subsequent stages as a transient separate from the other ridge segments. In general, contour segments not associated with ridge segments [Figs. 5(e) and 6(e)] have no impact on the final image [Figs. 5(f) and 6(f)]. Thus the (e) subplots are only used to consolidate the ridge segments in the (d) subplot, and are not used to add additional segments to the transient.

Having successfully linked appropriate segments into a common transient, the program then conducts some simple "feature filtering" to remove ridge segments that are clearly associated with close-range airgun signals. As illustrated in Fig. 6(e), these signals tend to produce contour segments with large global bandwidths, low contour segment minimum frequencies, and high contour segment maximum frequencies. Thus, if a contour segment exceeds some simple thresholds set for these features, then all ridge segments linked to that contour [e.g., segments in the large white bounding box in Fig. 6(f)] are eliminated. The program thereby eliminates dispersed airgun multipath arrivals that otherwise resemble bowhead FM sweeps. These simple steps, along with restrictions on the minimum and maximum detection durations, can eliminate anywhere between 10 and 30% of the total events passing through the image processing stage. The final output of the image processing stage is a set of transient detections, with each transient comprised of a set of spliced and linked ridge segments and one contour segment, the features associated with every segment, the 'total' features associated with the transient as a whole, and a labeled image of all linked ridge segments. The net effect is that although a single event detection from the first stage is often decomposed into numerous ridge segments, four distinct methods are used during the image processing stage to attempt to consolidate these segments into smaller groups of transient detections: morphological closing, segment splicing, harmonic/sideband linking, and linking ridge segments that share a common contour segment.

Finally, the image processing stage assembles a feature vector from the 25 features listed in Table III. The selected features include those extracted from the ridge segment with the greatest time-bandwidth product, features of the contour segment associated with the transient set, and "total" features of the complete transient such as the total duration, total bandwidth, the SNR, and the SEL.

E. Feature filtering: neural network processing

The feature vectors emerging from the image-processing step have already been subjected to some simple thresholding. However, many of the features are coupled, and a more sophisticated nonlinear classifier is needed. In the fourth stage of the algorithm, two cascaded multilayer feed-forward neural networks (Rumelhart, 1986; Rumelhart *et al.*, 1986; Bishop, 1995; LeCun *et al.*, 1998; Duda *et al.*, 2001; Bishop, 2006), also commonly called "multilayered perceptrons," process feature vectors from each individual DASAR. The first network makes a binary decision whether a vector represents a "biologic" vs "non-biologic" signal, while the second network makes a binary decision as to whether a vector is a "bowhead" call or "other biologic" sound. The reasoning behind using two networks instead of one is provided in the next subsection, which describes the development of the training sets.

The motivation for using multilayer perceptrons is that, in principle, these networks can reproduce any arbitrary classification scheme (Hecht-Nielsen, 1989; Kùrková, 1992). These networks have been applied to bowhead whale calls (Potter *et al.*, 1994; Mellinger and Clark, 2000), as well as primate sounds (Pozzi *et al.*, 2010), certain killer whale sounds (Deecke *et al.*, 1999), insects (Ganchev and Potamitis, 2007), and blue whales (Bahoura and Simard, 2010).

With the exception of Bahoura and Simard (2010), the approach used in this effort differs from previous applications of neural networks on whale calls (Potter *et al.*, 1994) in that here the networks are applied to extracted features of the calls, instead of the raw spectrogram time-frequency pixel values used in most previous efforts. Thus, instead of hundreds of raw input variables provided by a spectrogram, the inputs to the network are reduced to just 25 descriptive variables. The application of principal component analysis (PCA) to further reduce the dimensionality of the feature vector yielded no performance improvement.

The application of the neural network stage can be divided into three parts: (1) the assembling of an appropriate data set to "train" the networks, (2) the selection of an appropriate architecture and training protocol for the two networks, and (3) the application of the trained networks to novel feature vectors.

1. Creation of training data sets

A neural network requires a set of "training" examples from all desired output classes. Greeneridge Sciences' analysts reviewed acoustic data and recorded the times of occurrence of bowhead whale sounds, as well as the frequency band, duration, bearing, and general "shape" of the signal. The analysts used a program that displayed 60 s spectrograms for all DASARs simultaneously at a given site, and permitted analysts to listen to sound selections, as well as to estimate the quality of bearings obtained from the selection. An analyst could then use all this information to draw "bounding boxes" around the same call, as detected on different DASARs, and then localize the call. The availability of such manual analyses was an important resource for the automated processing of the SEPCO data sets, but three practical issues had to be resolved: which subsets of data to manually analyze, how to generate training examples of non-whale detection, and how to account for inconsistencies in the treatment of biological signals other than bowhead calls.

A typical DASAR deployment lasts approximately 45–60 days, but time and budget constraints limit manual analyses to 10%–20% of the data record. Six to eight non-contiguous days, roughly evenly spaced over time, were selected that displayed different levels of whale, local airgun, and distant airgun activity. The manual analysts would then review DASAR data from all sites for each day. The first 12 h of each day were used to train the algorithm, while the final 12 h of each day were reserved for evaluating algorithm performance.

The next issue was deciding how to generate training samples of transient events not associated with bowhead whale calls. Manually logging every transient event would have been prohibitively expensive. Instead, the first three stages of the automated procedure were used to derive this training data set. The first 12 h from each analyzed day were processed through the event, interval, and feature-extraction stages to produce a large collection of feature vectors. Each automated transient a was then compared with a manual analysis record m to determine the "time overlap":

$$t_{\text{ovlap}}(a,m) = \frac{\min\left(t_{d,a}, t_{d,m}\right) - \max\left(t_{0,a}, t_{0,m}\right)}{\max\left(t_{d,a}, t_{d,m}\right) - \min\left(t_{0,a}, t_{0,m}\right)}, \tag{6}$$

with $t_{0,x}$ and $t_{d,x}$ being the start and end time of transient x. A similar "frequency overlap" was defined using the minimum and maximum frequencies obtained by the automated and manual results. If an automated detection had at least 50% overlap in time and frequency with a manually logged bowhead detection, its corresponding feature vector was assigned as belonging to the "true" class; otherwise, the feature vector was assigned to the "false" class. If multiple automated results matched the overlap criteria with the same manual detection, the automated result with the greatest time overlap was selected. This approach assumes that manual analysts flag every bowhead whale call, even relatively weak and brief ones. Later, in Sec. VI, this assumption is revealed to be incorrect.

A related issue concerned the treatment of other biological sounds by the manual analysts. The original training data were collected in 2008, and originally a single neural network was used to analyze the data. In 2009 it became clear that single network was not effectively discriminating between bowhead, bearded seal, and walrus calls. The primary reason for this difficulty was the relatively small proportion of non-whale biological sounds in the 2008 data. The solution was to ask highly experienced analysts to generate a second training dataset with a high proportion of pinniped calls relative to bowhead calls, using the relatively pinniped-rich 2009 data set. A second network was then trained to distinguish between pinnipeds and whale calls, which could then be applied to feature vectors surviving the first neural network. Thus manually analyzed data from both 2008 and 2009 were used to train the

neural network stage. Specific dates used for the training data are provided in Table I.

2. Architecture and training protocol

Both neural networks are standard feed-forward networks consisting of (i) one input layer, (ii) one hidden layer of ten units, each using a hyperbolic tangent sigmoid activation function with biasing, and (iii) a two-unit output layer using linear transfer functions. Convergence tests indicated no improved performance using more than ten units per hidden layer. Each input variable in the training set was normalized such that the mean and standard deviation of the input distribution was 0 and 1, respectively (LeCun *et al.*, 1998).

The appropriate network weights and biases were derived using a scaled conjugate gradient backpropagation stochastic training method (Moller, 1993) with a mean-squared error performance function. A "cross-entropy" performance function was also examined (Bishop, 1995), but no discernable performance improvement was noted. Out of all the samples available from the training data, 60% were used to adjust the network weights in batch training mode, 20% were held in reserve to evaluate when to stop training the network, and 20% were held to validate the final network on completely novel inputs, to confirm that the network had not been over-optimized on the data set. There is no general consensus on how the training data should be divided between training, evaluation, and validation, but the 60/20/20 split used here is the default value recommended by commercial software packages such as MATLAB.

For every feature vector each network produces a scalar output that ranges between -1 and $+1$, with more positive values indicating a greater likelihood of the feature vector arising from a whale call. Each network was assigned a threshold such that feature vectors generating outputs more negative than the threshold were rejected.

A transient that survives passage through both neural networks, along with its associated feature vector, is now defined as a "call" in subsequent discussion. A set of calls detected on different DASARs that likely arise from the same source event is defined as a "call set." The estimated position derived from a call set is defined as a "call localization" or simply a "localization."

F. Cross-DASAR call matching

Up to this stage each DASAR has been processed independently, but at the fifth stage, calls between DASARs are matched to produce "call sets" as a precursor to final localization. The challenge of this stage is identifying and "matching" the same call detected on DASARs separated by several kilometers in water depths of less than 50 m. As discussed in Sec. III, at these water depths, acoustic signals with frequency content less than 500 Hz experience substantial geometric dispersion and bottom absorption, which alters the phase structure of the signal and often eliminates weaker sidebands and harmonics at larger ranges [e.g., Fig. 2(b) vs 2(d)]. Thus, phase-based techniques, such as matched filtering and/or signal cross-correlation, which work well for matching signals detected between deep water hydrophones in places like U.S. Navy test ranges (Ward *et al.*, 2000; Morrissey *et al.*, 2006; Ward *et al.*, 2008), did not perform well in this environment for most bowhead whale calls, as the peak normalized correlation coefficient was often less than 0.3, even for signals arising from the same source. More complex strategies are required for recognizing a common origin between calls detected at two different DASARs. The problem is exacerbated by the fact that bowhead calls can occur so frequently that a given detection on one DASAR can potentially be matched with several other detections on the second DASAR, within the physically permissible time window that encompasses the travel time of a sound between the two locations.

The eventual strategy employed for cross-DASAR matching used spectrograms, which only retained the magnitude of the signal spectra, and converted the matching problem into one of comparing images. By rejecting phase information, much precision is lost in estimating the relative arrival time of a signal on different sensors, but a compensating advantage of DASARs is that their localization method does not require relative arrival times, as is the case for conventional omnidirectional hydrophones. Thus, cross-DASAR matching only requires the ability to recognize the same signal on different instruments with a relative timing precision on the order of a second.

A call set is constructed by first defining an "anchor:" a particular DASAR that contains a call that is to be matched against other "target" DASARs. The stage begins by designating the southernmost DASAR (a) as the anchor, and designating a call from this DASAR as the "anchor call." The maximum travel time that is physically possible between the anchor and the two closest target DASARs is computed, and calls on a target DASAR that lie within the permissible time window relative to the anchor call are flagged as "candidates." Due to the triangular arrangements of the DASARs at a site, the two closest DASARs are always about 7 km apart, yielding a time window of 4.6 s. This window is expanded to 7 s in Table II to account for time misalignments between DASARs. During times of frequent calling or seismic activity, multiple candidates on a target DASAR are often available for each anchor call, and some criteria for evaluating the similarity between two calls are required.

Calls detected on different DASARs could not be matched simply by comparing their total frequency minima, maxima, or duration; some form of template matching was required. In principle the forward Hausdorff distance (Rucklidge, 1996) would be an ideal tool for matching a partial image to a complete image, but attempts to implement the method were slow and inaccurate. Eventually a variant of the spectrogram correlation method (Mellinger and Clark, 2000) was used, by exploiting the labeled images generated by the image processing stage.

Specifically, the ridge segment with the largest time-bandwidth product (pixel count in the spectrogram) on the anchor call is flagged, and a binary image $A(i,j)$ consisting of that one segment is created, with i being a row (frequency) index, and j being a column (time) index. A similar target binary image $T(i,j)$ is created for each candidate call on the target DASAR. The "min-norm distance coefficient" $C_{A,T}$ is defined such that

$$C_{A,T} = 1 - \frac{\max_k \left\{ \sum_{i,j} A(i,j+k)T(i,j) \right\}}{\min \left\{ \sum_{i,j} A(i,j), \sum_{i,j} T(i,j) \right\}}, \quad (7)$$

where k is the discrete time offset between the images, and the second term in Eq. (7) can be defined as a "min-norm correlation coefficient." In other words, after discovering the time offset that creates the greatest overlapping area (rough time-bandwidth product) between the images, the min-norm correlation coefficient is that maximum overlapping area divided by the area of the smaller segment. The min-norm distance coefficient is then simply the min-norm correlation coefficient subtracted from one. Note that the standard definition of a correlation coefficient would use the geometric mean of the two areas in the denominator; using that definition here would penalize a typical situation where substantial differences in area exist between the two segments.

Equation (7) is computed for every candidate target call, and values greater than 0.42 are rejected. Surviving candidate calls are then subjected to a reciprocal check by computing Eq. (7) between each candidate call and all physically permissible anchor calls, and confirming that the original anchor call provides the lowest value of Eq. (7) for each candidate call. If not, that candidate is rejected. Thus a "two-way" match is established between the anchor and target call, preventing a call from participating in multiple matched sets. If multiple candidates pass these tests, then the candidate call with the lowest value from Eq. (7) is finally accepted. The program then designates a new anchor call and repeats the entire process. With a little bookkeeping, Eq. (7) needs to be only executed once for a given anchor-target call pair.

Once all calls on the anchor DASAR have been matched to calls on the two closest DASARs, then the role of anchor is assigned to the next southernmost DASAR (B), and two additional DASARs that lie north of B (C, D) are the new targets. Since all calls on the B DASAR have previously been matched to the A DASAR, a match between a B call and C call will automatically result in a match between C and A, if the B call has been previously matched to an A call. Since the role of anchor is shouldered by all DASARs, calls can eventually be matched between distant DASARs, even if the structure of the signal changes substantially over the aperture of the system, since the actual comparisons are only conducted between adjacent DASARs.

As shown in Sec. V B, the cross-DASAR matching stage can play a prominent role in eliminating false detections.

G. Bearing estimation and localization

Once the call sets have been established, the sixth stage estimates a bearing for every call in that set, using methods nearly identical to previously-published methods used to locate bowhead whales from DASARs (Greene *et al.*, 2004). Along with the bearing estimate, the uncertainty of the measurement is also quantified by modeling the bearing estimation likelihood function as a Von Mises distribution and estimating the concentration parameter κ (which is analogous to the standard deviation of a Gaussian distribution), by conducting 100 bootstrap samples of the time series.

Finally, the localization stage uses the bearings from each call set to estimate a robust maximum-likelihood position of the animal, along with the 90% confidence ellipse (Lenth, 1981a; Greene *et al.*, 2004). A robust method is needed to reduce vulnerability to directional outliers, which are bearing estimates that deviate substantially from the bearing to a tentative location estimate. These outliers can arise from interference with other discrete signals or from incorrect matches within the call sets, and often arise during manual analysis as well. The maximum-likelihood method used here uses an "M-weighting" approach, where the contributions of each bearing to the localization estimate are initially given the same weight. After an initial solution is obtained, the weights are recomputed by inserting the estimated value of κ and the mismatch between the measured bearing and current maximum-likelihood estimate into a "Huber" weighting function (Huber, 1964; Lenth, 1981b) with a tuning parameter of 1.5. The new weights are used to compute a new maximum-likelihood position. After a few iterations the weights of directional outliers are reduced substantially, provided that 2 to 3 other DASAR bearings roughly converge. The final output is a location, bounded by a 90% confidence ellipse.

V. RESULTS

A. Neural network training datasets, bulk processing runs, and threshold test runs

As discussed in Sec. IV E 1, two separate neural networks were trained. In 2008, all five sites were manually analyzed on the days listed in Table I, with the first 12 h of each day providing the training sets for the first network. There were 141 796 automated transients (as defined in Sec. IV D 5) that satisfied the time and overlap criteria of Eq. (6) for a manually logged whale call, and will be designated as "whale transients." In addition, using the same criteria, 1 153 506 automated transients were not associated with manually logged whale calls, and will be designated as "excess transients." Figures 7 and 8 show two-dimensional histograms of the distributions of the minimum frequency and duration of whale and excess transients that comprise the first training set. Clearly, some separation between the classes can be obtained from these two parameters alone: substantial numbers of non-whale transients are less than 0.5 s long and have minimum frequencies below 25 Hz.

The manual analysts also generated 50 934 call sets from the 2008 data, where a "call set" has been defined in Sec. IV E as a set of calls detected on different DASARs which likely share a common origin. Call sets are the final outputs of the automated procedure before localization. Out of all the manually obtained call sets, 42 637 had two or more calls and 35 643 were localized successfully.

The training set for the second network was generated from the 2009 data (Table I), yielding 77 675 whale transients and 20 467 pinniped transients.

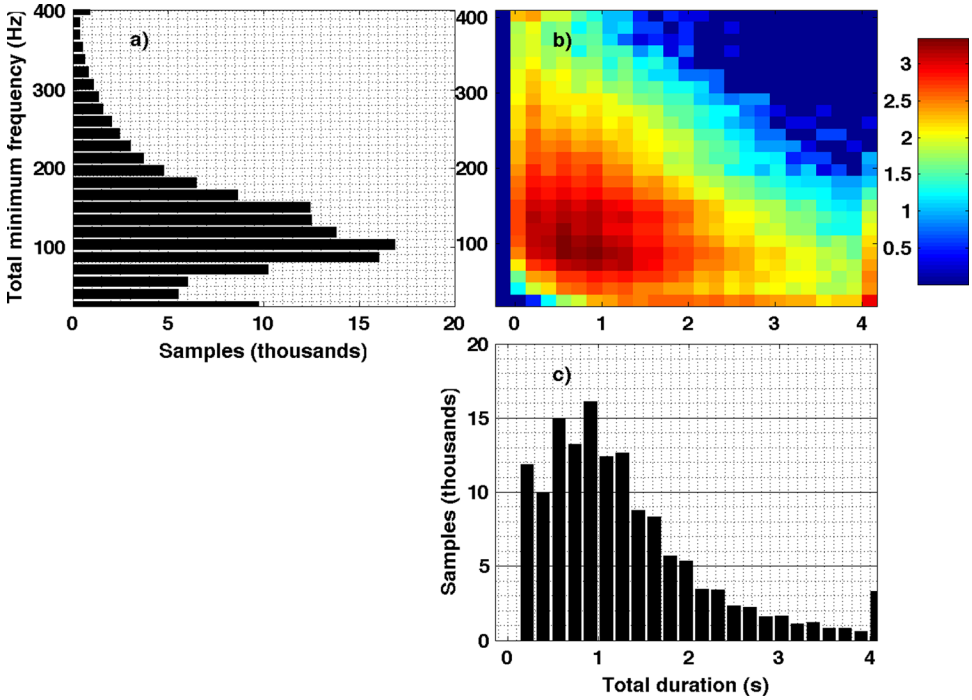

FIG. 7. (Color online) Distributions of the duration (s) and minimum frequency (Hz) of 141 796 whale call feature vectors obtained in 2008 from all sites, used to train the first neural network. (a) Marginal distribution for total minimum frequency; (b) 2D distribution of both parameters, with the intensity scaled to the log number of calls in a bin; (c) marginal distribution for total duration.

Several frequently used terms in this section are now defined. An "excess call fraction" (also known as "false discovery rate," or one minus the "precision") is the fraction of total automated transients that are excess transients, not whale transients. Thus, the initial training data for the first network had an excess call fraction of 0.89. Next, the "missed call fraction" (also known as one minus the "recall") is defined as the fraction of manually detected calls on individual DASARs that do not correspond to any automated transient, using the 50% time and frequency bandwidth overlap criteria of Eq. (6). Similar terms must be defined for call sets. A call set obtained manually is considered missed by the automated method if it shares fewer than two DASARs in common with every automated call set, using the criteria from Eq. (6). Thus, a "missed call set fraction" can be defined as the fraction of manually analyzed call sets that are missed by the automated method. A given automated call set is considered an "excess set" if it shares fewer than two DASARs with any manually derived call set; consequently, an "excess call set fraction" is the fraction of automated call sets that do not correspond with any manually derived call set. Defining these terms separates the question of accurate localization of call events from the question of whether calls on different DASARs were correctly matched.

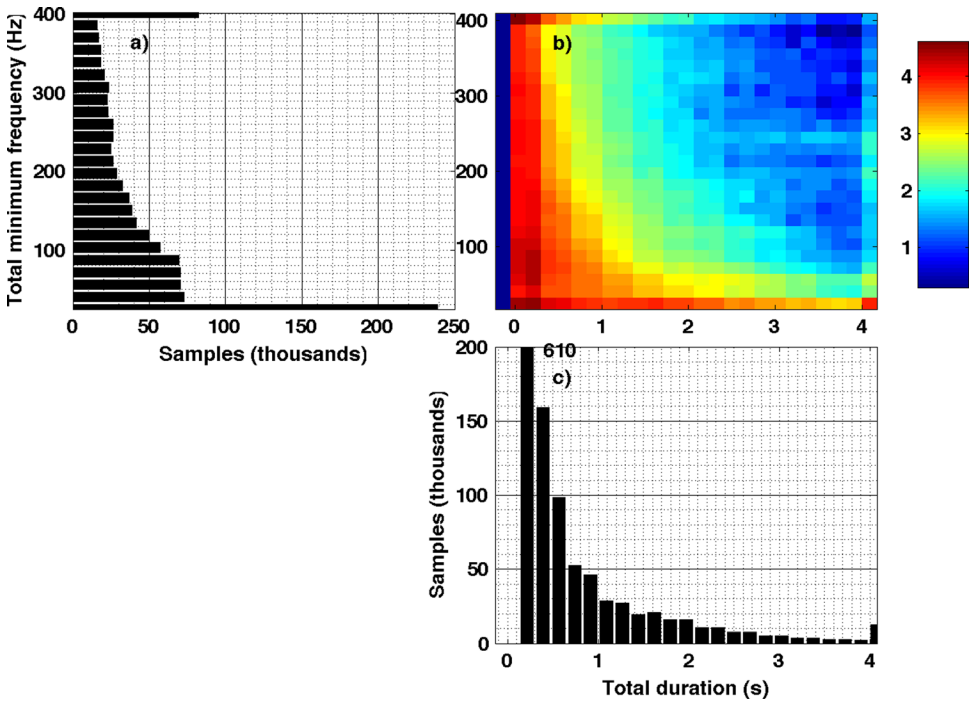

FIG. 8. (Color online) Same as Fig. 7, but showing the distributions of 1 153 506 feature vectors not associated with whale calls, used to train the first neural network. The first bin in (c) has about 610 000 samples, but the y-axis is limited to 200 000 samples to improve the visibility of the rest of distribution.

Once the networks had been trained, four years of acoustic data obtained between 2007 and 2010 were processed two different ways: using a "bulk processing" approach, and using a "threshold test" approach.

For the bulk processing runs, the complete algorithm was applied to all days of acoustic data obtained between 2007 and 2010, using the parameter values in Table II, which were estimated from trial and error tests on portions of the 2008 data. The bulk processing runs all used the same neural network thresholds, which were fixed using a criterion that only 10% of whale transients in the training data could be rejected by a given network. This criterion was an initial estimate as to what fraction of the total whale calls could be safely eliminated without significantly altering the final temporal and spatial distributions of the calls. The thresholds thus derived for both networks were −0.8 and 0.8, respectively; this combination (−0.8, 0.8) will be designated the "default thresholds." When the training data sets are applied to the networks using these default thresholds, the respective excess call fractions of the cascaded network outputs are 0.6 and 0.01. The first network thus reduces the excess call fraction of the training set by 33% (1−0.6/0.89) at the cost of rejecting 10% of transients associated with manually detected calls in the training data. Stated another way, the first network cuts the number of excess transients per manually detected call from 8 [0.89/(1−0.89)] to 1.5 [0.6/(1−0.6)]; a five-fold reduction.

Next, a series of "threshold test" runs were executed. The initial four stages of the algorithm (no cross-DASAR matching or localization) were applied to the subset of days that had been manually analyzed from each of the four years (Table I), using 121 different combinations of threshold values for the neural networks. These runs were compared with the manual analyses in order to examine the tradeoff between excess call fraction and missed call fraction.

The 2007 and 2010 manual analyses contain no data that were used to train the algorithm, while the 2008 and 2009 analyses do contain data used to train the neural network stage. No attempt was made to optimize other parameters of the algorithm listed in Table II.

B. Example of bulk processing using default network thresholds

Figure 9 displays how the various stages of the algorithm strip away the initial event detections during bulk processing of the 2008 season. Both heavy whale acoustic activity and close-range airgun surveys occurred during this season. Each subplot represents a different site, and the automated outputs from all the DASARs at that site are summed to produce the total number of detections surviving each stage at that site per day. As mentioned in Sec. IV D, the image processing stage can subdivide an "event" from the first stage into multiple "transients." If this stage outputs more transients than the initial input events, then the numbers of detections in the first two stages of Fig. 9 were increased accordingly for visual consistency. In Sec. VI this figure will be used to examine the relative importance of various stages in winnowing detections.

C. Comparing manual and automated performance before localization stage

Figure 10 compares the "bulk processing" and "threshold test" runs with manually analyzed results from the dates shown in Table I. The top subplot shows the combined performance from all five sites, while Figs. 10(b) and 10(c) show the performance at a single site both close to and far from local seismic airgun activity in 2007 and 2008 (sites 3 and 5, respectively). Each year of data is represented by a different symbol (e.g., diamond for 2010).

Each subplot shows three different comparisons between the manual and automated data. First, Fig. 10(a) shows that, across all sites, the missed call percentages range between 30% (2007 and 2009) and 40% (2008), with excess call percentages between 30% (2009 and 2010) and 55% (2007). Recall that, in the bulk processing runs, the default thresholds were set such that 20% of the "whale transients" would be removed by the fourth stage alone, so a missed call percentage of at least 20% is expected.

The relative impact of cross-DASAR matching (fifth) stage can be seen by comparing the large filled symbol to the corresponding small filled symbol: the missed call set fraction increases by a few percentage points relative to the missed call fraction, while the excess call set fraction shows a corresponding decrease relative to the excess call fraction.

The final comparison is shown by the curves connecting the hollow symbols displayed in Fig. 10, which consolidate the results of the threshold test runs. Threshold pairs that produce the same missed call fraction (to within ±0.05) have been grouped together, and the threshold pair from that group that yields the lowest excess call fraction is plotted in the figure. Thus, the curves display the performance of the first four stages when the optimal network threshold combinations are used. The default thresholds used in the bulk processing runs are close to optimal for 2009 and 2010, as their corresponding small filled symbols lie close to their associated optimal performance curves. The default thresholds are suboptimal for 2007 and 2008, as the small filled symbols for those years lie well above their corresponding curves. The optimal performance curves are similar for three of the four years. The year 2007 shows a higher excess call fraction for a given missed call fraction, for reasons to be discussed in Sec. VI.

D. Example of whale call spatial distributions from manual and automated processing

Figure 11 shows the spatial distribution of all whale calls localized between noon and midnight over the dates in 2009 listed in Table I. The top subplot maps the call distributions of the manual analyses, while the bottom shows the corresponding automated call distributions using the default network thresholds, which are shown to be close to optimal for 2009 in Fig. 10(a). As the manual results shown here were taken from the *last* 12 hours of each day, none of them were used to train the neural networks or adjust the parameters in Table II, because the primary author only had access to the *first* 12 hours of each day when training the software. This figure shows gross similarity between the manual and automated analyses over these coarse spatial scales.

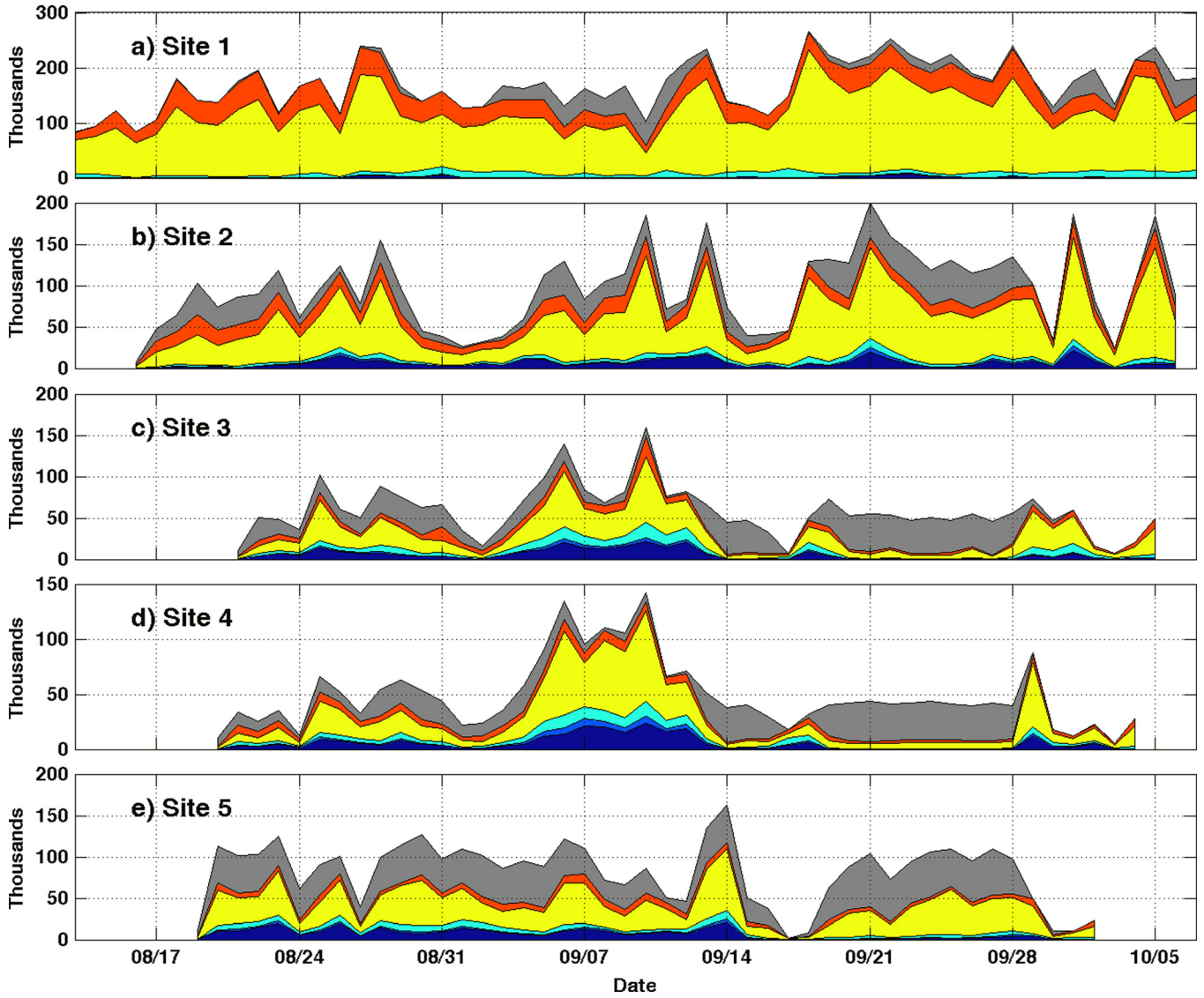

FIG. 9. (Color online) Progression of the bulk 2008 processing through each automated stage. Each subplot (a) through (e) represents sites 1–5, respectively. The output of each stage for all DASARs at a given site has been added together. The vertical extent of each color/shade indicates the number of events that have been removed on that date by the corresponding stage. Substantial local seismic survey activity took place near sites 3 and 4 between September 18 and 28. Starting from the top, the colors/shades show detections removed by interval filter (gray), image processing stage (orange), neural network stage (yellow), cross-DASAR matching (cyan), and localization (light blue). The dark blue area (bottom shaded area) shows final call counts.

A detailed evaluation of the statistical similarities between the manual and automated call localization distributions over four years requires temporal-spatial statistical analyses that lie beyond the scope of this paper. However, Fig. 12 shows some gross statistics of the call locations for both the manual and automated processing. The left column shows the number of DASARs that contribute to a given call location, and the right column shows a measure of the location uncertainty in terms of an "effective radius," or the radius of a circle that shares the same area as the 90% confidence ellipse. Figure 12 shows that the automated locations generally have fewer DASARs contributing to their locations, and fail to locate 13% of call sets, vs 5% for the manual results. However, automated locations derived from 2 or more DASARs display a positional uncertainty very similar to the manually analyzed locations.

The distribution of manually analyzed calls in Fig. 12(a) seems unusual, in that manually analyzed data from DASARs deployed for other studies (e.g., Blackwell *et al.*, 2008), find that calls localized with many DASARs are relatively less common than calls localized with fewer DASARs, as is shown in the automated result, and as would be expected from the sonar equation. By contrast, Fig. 12(a) shows a relatively even distribution of calls across DASARS, and a relatively high number of call locations that used seven DASARs (vs six DASARs).

This pattern seems to arise from systematic differences between the 26 manual analysts used to generate the manual dataset. In particular, three analysts generated 49% of all seven-DASAR localizations. One might argue that the three analysts happened to analyze days of low ambient noise levels and high call volumes; however, sites 3 and 5 produced 90% of the seven-DASAR locations, and site 4 only 10% of these kinds of locations, even when using the same days, and thus the same ambient noise conditions for all sites. Different analysts, however, analyzed different sites on the same

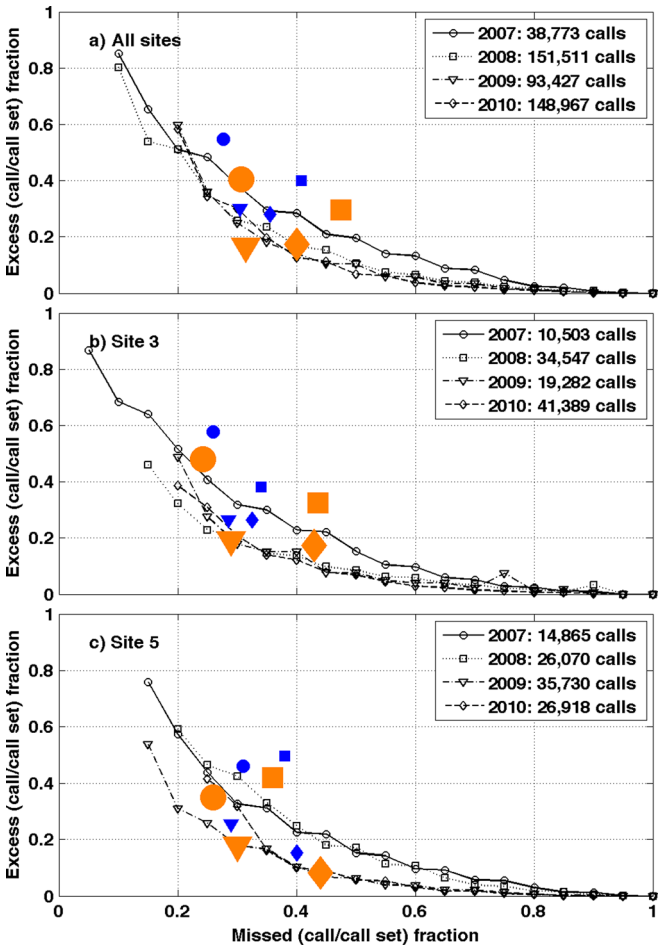

FIG. 10. (Color online) Comparison between manual analyses and fourth- and fifth-stage automated results, conducted on dates listed in Sec. V A. Data from 2007 (circles), 2008 (squares), 2009 (triangles), and 2010 (diamonds) are shown. Manual analyses from 2008 and 2009 were used to train portions of the automated classifier. The solid small blue symbols indicate the fourth-stage (neural network) automated performance, expressed in terms of excess call fraction (also known as "false discovery rate," or one minus the "precision") and missed call fraction (also known as one minus the "recall"), using default network thresholds discussed in Sec. V A. The large solid symbols indicate the corresponding performance of the fifth-stage (cross-DASAR matching) automated results, expressed in terms of excess call set fraction and missed call set fraction, using default network thresholds. The curves connecting hollow symbols show the neural network performance using optimized network thresholds, as discussed in the text. Subplots are comparisons between (a) all sites; (b) site 3 only (close to airgun surveys in 2007 and 2008); and (c) site 5 only (distant from airgun surveys). The legend indicates the number of individual call detections obtained by manual analyses for each year.

day, with more experienced analysts often assigned to days and sites with heavier call activity. We thus tentatively conclude that different manual analysts tend to link different numbers of calls together, and the aggregate effect is to flatten out the distribution shown in Fig. 12(a).

VI. DISCUSSION

A. Relative importance of various stages during bulk processing

The bulk processing output in Fig. 9 provides insight into the relative contributions of the processing stages. As expected, the neural network stage (yellow) generally plays the prominent role in winnowing candidate detections, reducing the number of transients by at least a factor of 4. Other stages also provide assistance in discriminating bowhead whale calls from other transients. For example, the interval filtering stage can strip up to 90% of initial events detected by the first stage during times of heavy seismic activity, as can be seen in Fig. 9 between September 21 and 28 on sites 3 and 4.

Both Figs. 9 and 10 show that the cross-DASAR matching stage can also play an important role in reducing excess call detections. For example, Fig. 9 reveals that the acoustic environment at site 1, the shallowest of the sites, substantially differs from the others. On average this site produces twice as many event detections as the others, and few of the events are associated with airguns. During at least half of the 2008 deployment, over 90% of the site 1 "calls" that survive the neural network stage cannot be matched to at least one other DASAR, making the call-matching stage at least as important as the neural network stage in removing extra detections from this site. A random review of spectrograms from DASARs at Site 1 show the presence of numerous low-level transient narrowband pulses, uncorrelated between DASARs, consistent with what can be deduced from the algorithm's performance. The origin of these pulses is uncertain; they may be biological in origin or may arise from constant mechanical disturbances of the DASAR sensors in this shallow, high-current region.

B. Sensitivity of neural network performance to training set size and feature selection

Another practical question of interest is how small the training set could be while still reproducing the performance curves shown in Fig. 10. The original training data set from 2008 was sampled randomly without replacement, generating smaller training samples that contained 0.1, 1, 10, 20, and 50% of the number of feature vectors of the final 1.3 million-example training set (141 796 whale examples and 1.153 million false examples). The ratio of false to true whale samples was kept constant (such that the 10% training set contained 14 180 whale examples and 115 350 false examples, etc.). Three independent training sets from each size category were subsampled from the large training data set.

Every sub-sampled set was then used to train an additional neural network, using the same architecture and training protocol for the original network, and using the same pseudo-random number sequence to initialize the weights. In particular, 60% of the subsample was used to train the network, 20% was used to determine when to stop training, and 20% was used to validate performance, as discussed in Sec. IV E 2. The resulting performance curve was then compared to the performance curve of the first network only. The analysis found that a network trained with 10% the size of the original training set converged to virtually the same performance curve as a network trained with the entire set. There is another aspect to the story, however.

A frequent criticism of neural networks is that they are "black boxes" that provide no insight into the relative

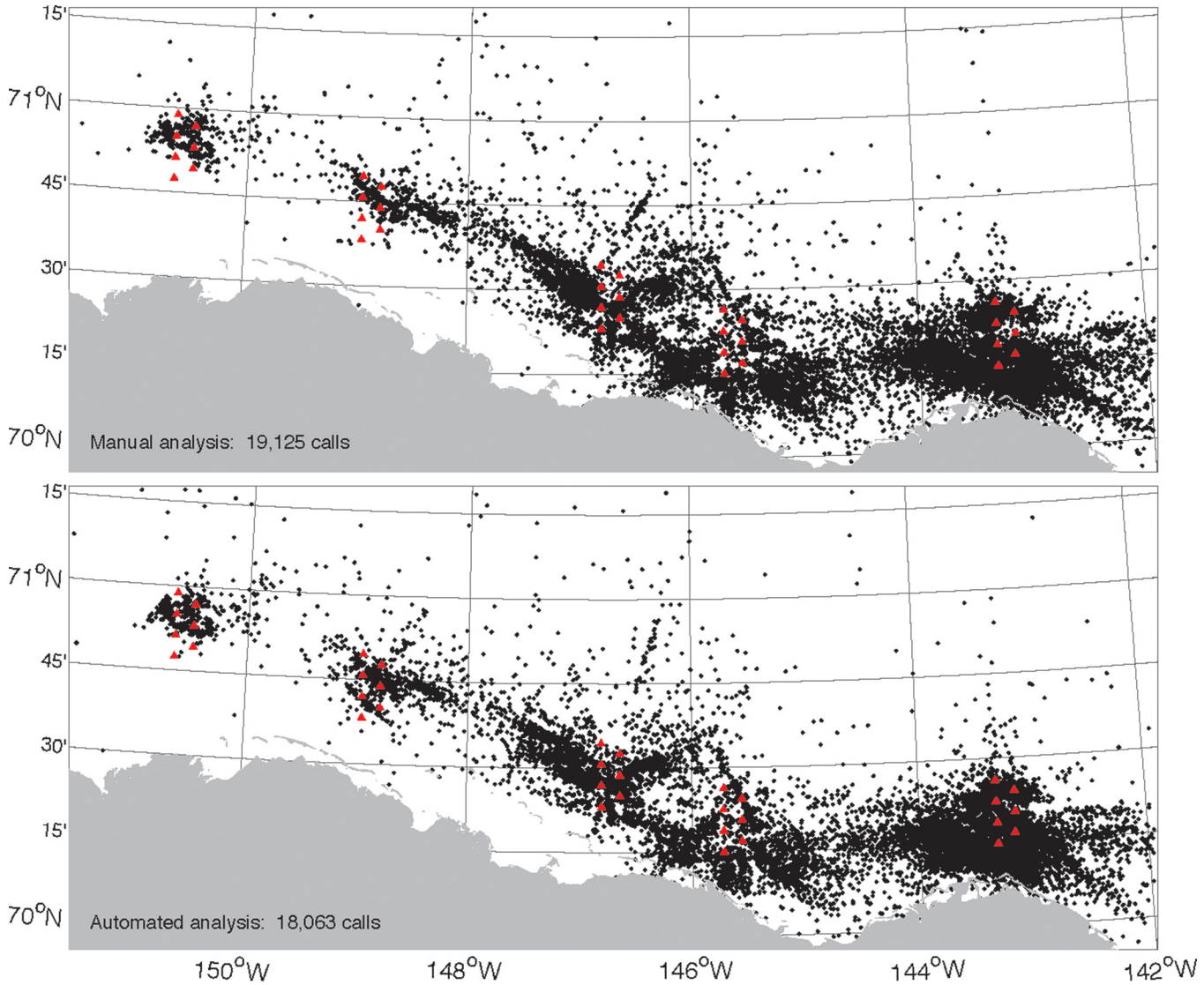

FIG. 11. (Color online) Top: Spatial distribution of 19 125 whale call locations obtained by manual analysis of the last 12 h of 8 non-contiguous days in 2009. Bottom: distribution of 18 063 whale call locations computed by automated detection algorithm over the same time periods. No manual data from the top subplot were used to train the software.

importance and interactions of the feature vectors' components. To address this concern, an additional series of neural networks was trained using 10% of the original training data set, after removing a portion of the feature space from the training vectors. The following groups of parameters from Table III were systematically removed from the training sets as a block: those related to duration (three parameters), minimum frequency (three parameters), maximum frequency (three parameters), bandwidth (four parameters), SEL/SNR (four parameters), orientation (one parameter), eccentricity (one parameters), solidity (one parameter), time-bandwidth product (two parameters), kurtosis (one parameter), and presence of harmonics (one feature). Thus for a "duration" scenario, the three parameters in Table III related to signal duration are excised, and the remaining 22 features were used to train the network.

For each group of excised parameters, the three subsampled training sets were used to train three networks, in order to determine whether the relative contributions of the feature space were consistent. Each of the trained networks used a unique sequence of psuedo-random numbers to assign the initial weights, and these same sequences were used when testing reduced feature sets on a particular network. That way, random fluctuations in how the initial weights were seeded could be held fixed during the comparisons.

As expected, no one feature dominated the performance; indeed, the three sets of trained networks apparently achieved similar performance while assigning different relative marginal weights to the feature vectors. The relative importance of the features became much more consistent when the analysis was repeated with 20% of the training dataset. Under those circumstances the feature sets associated with both duration and SNR/SEL reduced the excess call fraction (for a given missed call fraction) by 0.02–0.03, the orientation feature reduced the fraction by 0.01–0.03, and the relative importance of the other parameters seemed to be roughly equal, with their relative impacts varying between the three networks, even when 20% of the total training set was sampled.

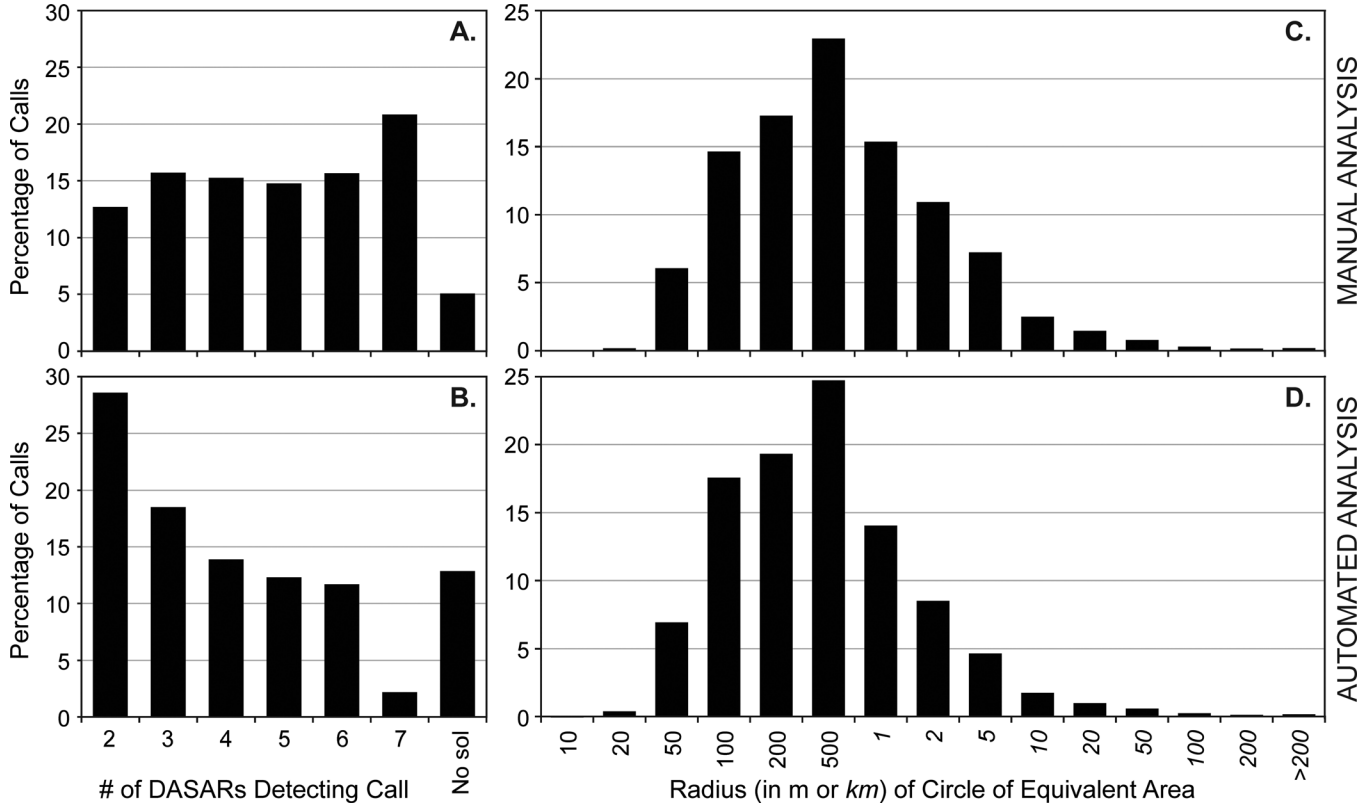

FIG. 12. Statistics of whale call locations for manual and analyzed positions shown in Fig. 11. Left column: Distribution of number of DASARS used to localize calls in (a) manual and (b) automated analysis. The "no sol" category indicates percentage of call sets that yield no localization solution (e.g., no crossed bearings, or failure to obtain bearings from enough DASARS). Right column: Distribution of area of 90% confidence ellipse for call locations, expressed in terms of radius of a circle of equivalent area, for (c) manual and (d) automated analysis. Italicized x-axis labels are in units of km; otherwise, units are meters. The right column only uses positions derived from two or more DASARs.

In summary, only 10% of the original training set (tens of thousands of whale calls and hundreds of thousands of false examples) seems required to reproduce the gross performance curve of the primary neural network, but at least 20% of the original training set is needed to obtain a consistent internal representation of the weights, in terms of relative feature contributions.

C. Interpreting excess call fraction

Figure 10(a) shows how the cascaded networks produce excess call fractions between 0.3 and 0.6 during the threshold test runs, with values in 2007 and 2008 greater than those in 2009 and 2010. The optimal performance curves also reveal a similar trend, with 2007 exhibiting greater excess call fractions than later years, across all missed call fractions.

It is tempting to suggest that the local airgun surveys in 2007 and 2008 are responsible for the higher excess call fractions during those years, but Fig. 10(c) shows that the discrepancy between years exists at sites relatively distant from the airgun surveys as well. Furthermore, a day-by-day review of the results found that the automated excess call fraction *decreases* during days that local airgun surveys are present, and the missed call fraction increases. An alternative explanation for the observed differences between years is that the manual analysis procedure has become more standardized over time. For example, in 2007, the first year of the large-scale SEPCO study, manual analysts often used evidence of localization convergence to assess whether a detection was a whale call, and would reject weak SNR signals that did not seem to contribute substantially to the localization. The DASAR sensors used in 2007 also had less accurate bearing estimation capability than subsequent years. (The DASAR sensors were improved in 2008 and remained unchanged through 2010.) Detections with inferior localization performance (e.g., due to low SNR or inaccurate bearings which resulted in poorer convergence characteristics), were typically ignored by the manual analysts. Thus some valid whale calls would have been unmatched with the manual results, producing an excess call count.

Again, the issue of missed manual call detections arises when comparing the results in Fig. 10(a) to the spatial distribution maps in Fig. 11. In 2009 call sets generated using the default network thresholds show an excess call set fraction of 0.18 and missed call fraction of 0.3. One would thus expect that the total number of automated call localizations in 2009 would be 83% (0.7×1.18) of the number of manually analyzed localizations (19 125 in Fig. 11), or 15 873 automated localizations. In fact, 18 063 automated locations are found (95% of the number of manual detections).

Furthermore, there is an unexpectedly good correspondence between the gross spatial distributions of the automated and manual results, given the excess call set fractions shown

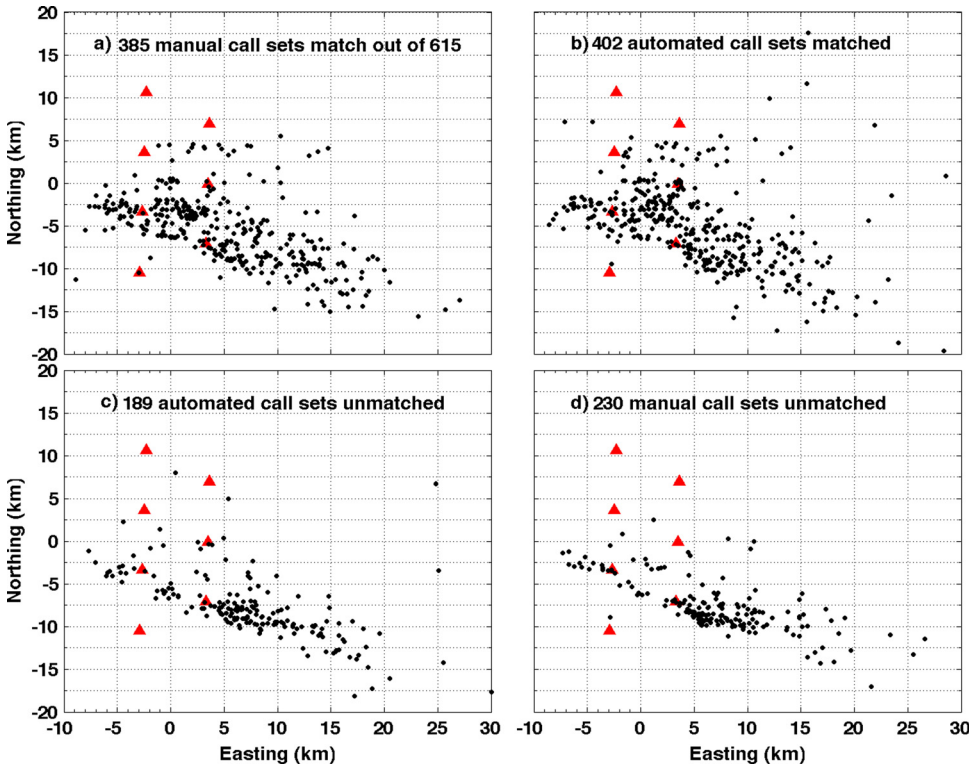

FIG. 13. (Color online) Detailed comparison between manual and automated call set spatial distributions for site 2, August 28, 2008, between midnight and 4 a.m. (a) Spatial distribution of manual call sets that match an automated call set. A "match" between an automated and manual call set occurs when they share at least two DASARs in common. (b) Spatial distribution of automated call sets that match the call sets shown with (a). There are more locations plotted here than in (a) because a manual call set encompassing four or more DASARs may match two two-DASAR automated call sets. (c) Spatial distribution of "excess" automated call sets that do not match any manual call set; (d) manual call sets that do not match any automated call sets.

in Fig. 10. This apparent discrepancy was consistent across all years and led to further suspicions that at least some of the excess call sets were indeed true whale calls that had somehow not been matched with a manual result.

To examine this speculation, the manual and automated locations were compared in greater detail in Fig. 13, during a period of heavy whale calling activity. Figure 13(c) plots the localizations derived from the excess call sets and demonstrates how closely the resulting spatial distribution corresponds to the manual distributions in both Fig. 13(a) and 13(d). The similarity between the distributions in Fig. 13 indicates that the excess call fractions shown in Fig. 10 incorporate actual whale calls missed by the manual analysts, and not just false alarms.

A manual review of the results shown in Fig. 13(c) reveals two primary reasons for the existence of automated call sets unassociated with the manual call sets. First, the original manual analysis missed substantial numbers of brief and/or weak calls: 53% of the localizations in Fig. 13(c) arise from whale calls that were not detected on any DASAR by the original manual analysis. It is not surprising that the manual analysis is biased against brief calls, as analysts typically review 60 s of acoustic data at a time in a spectrogram, a time scale large enough that weak calls less than a second long could be missed. Second, an additional 25% of the localizations in Fig. 13(c) correspond to matched-call sets that actually contain some calls flagged by the manual analysts on individual DASARs, using the 50% overlap criteria defined in Eq. (6). The automated matched call sets, however, use different combinations of DASARs than those used by the manual analysts. The automated algorithm seems biased against high SNR calls that generate substantial amounts of reverberation, which are exactly the calls which manual analysts excel at flagging. The analysts, in turn, tended to ignore the weaker examples of a given call on more distant DASARs. Manual and automated matched call sets were thus found that shared fewer than two DASARs in common, even though they were localizing the same call event, as could be inferred by comparing the modulation patterns, bearings, and relative timing of the calls in both matched sets. Finally, the remaining 22% of the automated locations in Fig. 13(c) seem to be true false alarms, in that they are localizations of signals that are clearly not whale calls, or localizations of signals that clearly do not arise from a common source event. Thus in summary, over 75% of the excess call sets shown in Fig. 13(c) are likely to be legitimate whale positions.

Similar logic lies behind the explanation for the 230 locations missed by the automated algorithm, shown in Fig. 13(d). Slightly over 20% of the missed manual call sets are actually localized by the automated algorithm, but the algorithm produced call sets using different DASARs than the corresponding manual call set. Of the remaining missed manual locations, 79% were detected on three or fewer DASARs, and 48% were detected on just two DASARs. By contrast, 54% of all the manual localizations used three or fewer DASARs, and 30% used two DASARs. Thus, calls that are present on fewer DASARs are more likely to be missed by the automated algorithm.

Similar trends were observed in other years. Therefore, the excess call fractions shown in Fig. 10 likely overestimate the true "false alarm" rate, so the performance curves in Fig. 10 should be considered an upper bound on the underlying false alarm fraction. Rather than considering manual analyses as "ground truth," these findings suggest that both manual and automated processing miss legitimate calls,

complicating the challenge of comparing large-scale datasets generated by both techniques.

VII. CONCLUSION

An automated detection, classification, and localization scheme for bowhead whale calls has been applied to four years of data from large spatial arrays of 35 autonomous recorders. In two of those years, substantial numbers of close-range airgun signals were also present. Image processing methods, neural network classifiers, and spectrogram correlation methods were combined to detect and localize arbitrarily modulated bowhead whale calls. Hundreds of thousands of whale calls and over a million false samples were used to train the networks, using data from two of the four years (2008 and 2009). The input features used here are not conditioned spectrogram time-frequency bins, but are smaller sets of quantitative features extracted through the image processing techniques. Only 10% of the training set seemed required to obtain the observed performance, but at least 20% of the training set is needed to produce a consistent pattern in the relative weighting of the features, with features related to duration, SNR, and orientation having the largest marginal impact on classifier performance. The optimized performance curve from the reliable validating data year (2010) indicates that the networks did not have to be retrained with new data; simply adjusting the output thresholds with new data was sufficient. While neural networks were effective and easy to implement, other standard pattern-recognition classifiers should presumably work.

The large-scale spatial and temporal distributions of the calls are similar for manual and automated methods (Fig. 11), despite the excess call set fractions displayed by the automated algorithms in Fig. 10. Various lines of evidence (e.g., Fig. 13) suggest that the manual analyses miss legitimate whale calls, inflating the apparent false detection fraction. This interpretation, if true, helps explain the relatively poorer performance of the algorithm in 2007, a year when manual analysis procedures and instrument hardware were in the process of being standardized. It also suggests that automated missed call and call set fractions could be decreased further by adjusting the appropriate network thresholds, with little deterioration in the quality of the resulting call set spatial distributions.

The algorithm could be improved in several ways. None of the parameters in Table II have been systematically optimized, other than the neural network thresholds. The merging of widely separated harmonic components into a single "transient" event could be improved further (e.g., Heller and Pinezich, 2008). There are also indications that each site should have its own dedicated neural network, trained with data from that site, instead of applying a common network trained with data from all sites. Site 1, in particular, seems to experience a different acoustic environment from the other sites and may benefit from a dedicated neural network, or at least a different set of network thresholds.

The algorithm also shows vulnerability to the presence of ships, which tend to generate numerous short FM-type signals in equalized spectrograms. Adding additional features into the classifier that characterize the frequency spectrum of the ambient noise background, including cepstral or entropy-based (Burg spectrum) measures (Erbe and King, 2008), could mitigate the impact of vessel noise.

ACKNOWLEDGMENTS

This work has been supported by Shell Oil Exploration and Production Company (SEPCO). Kristin Otte and Sara Tennant, among many others, helped derive the manual training sets used in this effort. Bill McLennan developed the acoustic data format used by the automated processor, and devised a nonlinear calibration scheme used in the 2007 data set. Steve Eddins of Mathworks provided background on connected component labeling algorithms. Professor Garrison Cottrell at UCSD provided advice on the training and implementation of neural networks, and Dr. Sergio Belongie of UCSD provided advice on image processing methods. Melania Guerra and Delphine Mathias assisted with field deployments, and they, along with Diana Ponce, also assisted with data analysis and fact-checking. Julien Delarue of JASCO provided references on walrus sounds.

Asitha, M., Ong, S. H., Mandar, C., and Elizabeth, T. (**2008**). "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," J. Acoust. Soc. Am. **124**, 1159–1170.

Bahoura, M., and Simard, Y. (**2010**). "Blue whale call classification using short-time Fourier and wavelet packet transforms and artificial neural network," Digital Signal Process. **20**, 1256–1263.

Bishop, C. M. (**1995**). *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford), p. 475.

Bishop, C. M. (**2006**). *Pattern Recognition and Machine Learning* (Springer, New York), p. 738.

Blackwell, S. B., Kim, K. H., Burgess, W. C., Greene, Jr., C. R., and Aerts, A. M. (**2008**). "Acoustic localization of migrating bowhead whales near Northstar, Autumn 2008," in Monitoring of industrial sounds, seals, and bowhead whales near BP's Northstar Oil Development, Alaskan Beaufort Sea, 2008: Annual Summary Report, edited by L. A. M. Aerts and W. J. Richardson, LGL Report P1081, 2009, Chap. 4, pp. 4–5.

Blackwell, S. B., Richardson, W. J., Greene, Jr., C. R., and Streever, B. (**2007**). "Bowhead whale (*Balaena mysticetus*) migration and calling behaviour in the Alaskan Beaufort sea, Autumn 2001–04: An acoustic localization study," Arctic **60**, 255–270.

Clark, C. W., and Ellison, W. T. (**2000**). "Calibration and comparison of the acoustic location methods used during the spring migration of the bowhead whale, *Balaena mysticetus*, off Pt. Barrow, Alaska, 1984–1993," J. Acoust. Soc. Am. **107**, 3509–3517.

Clark, C. W., and Johnson, J. H. (**1984**). "The sounds of the bowhead whale, *Balaena mysticetus*, during the spring migrations of 1979 and 1980," Can. J. Zool. **62**, 1436–1441.

Datta, S., and Sturtivant, C. (**2002**). "Dolphin whistle classification for determining group identities," Signal Process. **82**, 251–258.

Deecke, V. B., Ford, J. K. B., and 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.

Delarue, J., Laurinolli, M., and Martin, B. (**2009**). "Bowhead whale (*Balaena mysticetus*) songs in the Chukchi Sea between October 2007 and May 2008," J. Acoust. Soc. Am. **126**, 3319–3328.

D'spain, G. L., Kuperman, W. A., Clark, C. W., and Mellinger, D. K. (**1995**). "Simultaneous source ranging and bottom geoacoustic inversion using shallow water, broadband dispersion of fin whale calls," J. Acoust. Soc. Am. **97**, 3353.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification* (Wiley, New York), p. 654.

Erbe, C., and King, A. R. (**2008**). "Automatic detection of marine mammals using information entropy," J. Acoust. Soc. Am. **124**, 2833–2840.

Ganchev, T., and Potamitis, I. (**2007**). "Automatic acoustic identification of singing insects," Bioacoustics **16**, 281–328.

Gillespie, D. (**2004**). "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," Can. Acoust. **32**, 39–47.
Gonzalez, R., and Woods, R. (**2002**). *Digital Image Processing* (Prentice-Hall, Upper Saddle River, NJ), p. 793.
Greene, C. R., Jr., McLennan, M. W., Norman, R. G., Mcdonald, T. L., Jakubczak, R. S., and Richardson, W. J. (**2004**). "Directional frequency and recording (DIFAR) sensors in seafloor recorders to locate calling bowhead whales during their fall migration," J. Acoust. Soc. Am. **116**, 799–813.
Haralick, R. M., and Shapiro, A. D. (**1992**). *Computer and Robot Vision* (Addison-Wesley, Reading, MA), pp. 40–48.
Hecht-Nielsen, R. (**1989**). "Theory of the backpropagation neural network," in *Proceedings of the International Joint Conference on Neural Networks*, *IEEE*, Vol. 1, pp. 593–605.
Heimlich, S. L., Mellinger, D. K., Klinck, H., Stafford, K., Moore, S. E., Berchok, C., and Nieukirk, S. L. (**2009**). "Detecting bowhead whale (*Balaena mysticetus*) sounds in the Beaufort Sea: Confounding sounds in a cacophony of noise," in *16th Biennial Conference on the Biology of Marine Mammals*, Quebec City, Canada, pp. 110–111.
Heller, J. R., and Pinezich, J. D. (**2008**). "Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm," J. Acoust. Soc. Am. **124**, 1830–1837.
Huber, P. J. (**1964**). "Robust estimation of a location parameter," Ann. Math. Stat. **35**, 73–78.
Kùrková, V. (**1992**). "Kolmogorov's theorem and multilayer neural networks," Neural Comput. **5**, 501–506.
Lammers, M. O., Au, W. W. L., and Herzing, D. L. (**2003**). "The broadband social acoustic signaling behavior of spinner and spotted dolphins," J. Acoust. Soc. Am. **114**, 1629–1639.
Lampert, T. A., and O'Keefe, S. E. M. (**2010a**). "A survey of spectrogram track detection algorithms," Appl. Acoust. **71**, 87–100.
Lampert, T. A., and O'Keefe, S. E. M. (**2010b**). "An active contour algorithm for spectrogram track detection," Pattern Recogn. Lett. **31**, 1201–1206.
LeCun, Y., Bottou, L., Orr, G. B., and Muller, K. R. (**1998**). "Efficient backprop," in *Neural Networks: Tricks of the Trade* (Springer, New York), pp. 9–50.
Lenth, R. V. (**1981a**). "On finding the source of a signal," Technometrics **23**, 149–154.
Lenth, R. V. (**1981b**). " Robust measures of location for directional data," Technometrics **23**, 77–81.
Levanon, N. (**1988**). *Radar Principles* (Wiley, New York), p. 320.
Madhusudhana, S. K., Oleson, E. M., Soldevilla, M. S., Roch, M. A., and Hildebrand, J. A. (**2008**). " Frequency based algorithm for robust contour extraction of blue whale B and D calls," in *OCEANS 2008-MTS/IEEE Kobe Techno-Ocean*, pp. 1–8.
Madsen, P. T. (**2005**). "Marine mammals and noise: Problems with root mean square sound pressure levels for transients," J. Acoust. Soc. Am. **117**, 3952–3957.
Mellinger, D. K. (**2002**). "Ishmael 1.0 User's Guide," PMEL-120, NOAA/PMEL Technical Memo.
Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529.
Moller, M. F. (**1993**). "A scaled conjugate gradient algorithm for fast supervised learning," Neural Networks **6**, 525–533.
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.
Morrissey, R. P., Ward, J., DiMarzio, N., Jarvis, S., and Moretti, D. J. (**2006**). "Passive acoustic detection and localization of sperm whales (*Physeter macrocephalus*) in the tongue of the ocean," Appl. Acoust. **67**, 1091–1105.
Mouy, X., Leary, D., Martin, B., and Laurinolli, M. (**2008**). "A comparison of methods for the automatic classification of marine mammal vocalizations in the Arctic," in *New Trends for Environmental Monitoring Using Passive Systems*, 2008, pp. 1–6.
Nitzberg, R. (**1986**). "Clutter Map CFAR Analysis," IEEE Trans. Aerospace Electron. Syst. **AES-22**, 419–421.
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.
Potter, J. R., Mellinger, D. K., and Clark, C. W. (**1994**). "Marine mammal call discrimination using artificial neural networks," J. Acoust. Soc. Am. **96**, 1255–1261.
Pozzi, L., Gamba, M., and Giacoma, C. (**2010**). "The use of artificial neural networks to classify primate vocalizations: A pilot study on Black Lemurs," Am. J. Primatol. **72**, 337–348.
Rabiner, L. R. (**1989**). "A tutorial on hidden Markov-models and selected applications in speech recognition," Proc. IEEE **77**, 257–286.
Risch, D., Clark, C. W., Corkeron, P. J., Elepfandt, K. M., Kovacs, K. M., Lydersen, C., Stirling, I., and van Parijs, S. M. (**2007**). "Vocalizations of male bearded seals, *Erignathus barbatus*: Classification and geographical variation," Animal Behavior **73**, 747–762.
Roch, M. A., Soldevilla, M. S., Burtenshaw, J. C., Henderson, E. E., and Hildebrand, J. A. (**2007**). "Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California," J. Acoust. Soc. Am. **121**, 1737–1748.
Rucklidge, W. (**1996**). *Efficient Visual Recognition Using the Hausdorff Distance* (Springer, Berlin), p. 178.
Rumelhart, D. E. (**1986**). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, MA), Vol. 1, p. 567.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (**1986**). "Learning representations by back-propagating errors," Nature **323**, 533–536.
Stafford, K. M., Moore, S. E., Laidre, K. L., and Heide-Jorgensen, M. P. (**2008**). "Bowhead whale springtime song off West Greenland," J. Acoust. Soc. Am. **124**, 3315–3323.
Sturtivant, C., and Datta, S. (**1995**). "Techniques to isolate dolphin whistles and other tonal sounds from background noise," Acoust. Lett. **18**, 189–193.
Thode, A. (**1999**). "Localization, inversion, and source signal recovery of blue whale sounds using matched field processing," Ph.D. thesis, Scripps Instituition of Oceanography, University of California, San Diego.
Top, P. (**2009**). "Kalman filter tracking of dolphin whistle contours," J. Acoust. Soc. Am. **126**, 2165–2165.
Ward, J., Fitzpatrick, M., DiMarzio, N., Moretti, D., and Morrissey, R. (**2000**). "New algorithms for open ocean marine mammal monitoring," in *OCEANS 2000 MTS/IEEE Conference and Exhibition*, Conference Proceedings Cat. No. 00CH37158, IEEE, Vol. 3, pp. 1749–1752.
Ward, J., Morrissey, R., Moretti, D., DiMarzio, N., Jarvis, S., Johnson, M., Tyack, P., and White, C. (**2008**). "Passive acoustic detection and localization of *Mesoplodon densirostris* (Blainville's beaked whale) vocalizations using distributed, bottom-mounted hydrophones in conjunction with a Digital Tag (DTag) recording," Can. Acoust. **36**, 60–66.
Weisburn, B. A., Mitchell, S. G., Clark, C. W., and Parks, T. W. (**1993**). "Isolating biological acoustic transient signals," in *ICASSP-93*, *IEEE*, Vol. 1, pp. 269–272.
Wiggins, S. M., McDonald, M. A., Munger, L. M., Moore, S. E., Hildebrand, J. A. (**2004**). "Waveguide propagation allows range estimates for North Pacific right whales in the Bering Sea," Can. Acoust. **32**, 146–154.