1-2013

Automated Extraction and Classification of Time-Frequency Contours in Humpback Vocalizations

Hui Ou
Portland State University

Whitlow W.L. Au
University of Hawaii

Lisa M. Zurk
Portland State University, zurkl@pdx.edu

Marc O. Lammers
University of Hawaii

Citation Details

Ou, H., Au, W. W., Zurk, L. M., & Lammers, M. O. (2013). Automated extraction and classification of time-frequency contours in humpback vocalizations. The Journal of the Acoustical Society of America, 133(1), 301-310.

Automated extraction and classification of time-frequency contours in humpback vocalizations

Hui Ou[a)]
Department of Electrical and Computer Engineering, Northwest Electromagnetics and Acoustics Research Laboratory (NEAR-Lab), Portland State University, Portland, Oregon 97201

Whitlow W. L. Au
Marine Mammal Research Program, Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, Hawaii 96744

Lisa M. Zurk
Department of Electrical and Computer Engineering, Northwest Electromagnetics and Acoustics Research Laboratory (NEAR-Lab), Portland State University, Portland, Oregon 97201

Marc O. Lammers
Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, Hawaii 96744

(Received 19 April 2012; accepted 20 November 2012)

A time-frequency contour extraction and classification algorithm was created to analyze humpback whale vocalizations. The algorithm automatically extracted contours of whale vocalization units by searching for gray-level discontinuities in the spectrogram images. The unit-to-unit similarity was quantified by cross-correlating the contour lines. A library of distinctive humpback units was then generated by applying an unsupervised, cluster-based learning algorithm. The purpose of this study was to provide a fast and automated feature selection tool to describe the vocal signatures of animal groups. This approach could benefit a variety of applications such as species description, identification, and evolution of song structures. The algorithm was tested on humpback whale song data recorded at various locations in Hawaii from 2002 to 2003. Results presented in this paper showed low probability of false alarm (0%–4%) under noisy environments with small boat vessels and snapping shrimp. The classification algorithm was tested on a controlled set of 30 units forming six unit types, and all the units were correctly classified. In a case study on humpback data collected in the Auau Chanel, Hawaii, in 2002, the algorithm extracted 951 units, which were classified into 12 distinctive types. © *2013 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4770251]

PACS number(s): 43.60.Bf, 43.60.Np, 43.80.Ka, 43.30.Sf [JAS] Pages: 301–310

I. INTRODUCTION

Various aspects of humpback whale song have been studied from the 1970s (Winn *et al.*, 1970; Payne and McVay, 1971) to the present time (Lammers *et al.*, 2011). Helweg *et al.* (1992) have written an excellent review on the understanding of humpback whale songs. The various aspects of humpback whale songs that have been studied include geographic and seasonal variations (Helweg *et al.*, 1998; Au *et al.*, 2000; Cerchio *et al.*, 2001), the evolution of song structure throughout the year and between years (Payne *et al.*, 1983), size of singers (Spitz *et al.*, 2002), and the behavioral and spatial distribution of singers (Frankel *et al.*, 1995; Tyack, 1981), to name a few.

Payne and McVay (1971) were one of the first to analyze and describe humpback whale songs. They described songs as being made up of different units arranged in sequences. Units were described as "the shortest sounds in the song which seem continuous to the human ear," and they are burst of sounds that typically last between 1 and 3 s (Au *et al.*, 2006). According to Payne and McVay (1971) various units are organized in a specific pattern to make up a phrase. Phrases are further organized into a pattern to make up a theme. And finally, themes are organized into a pattern to make up a song. Songs are often repeated for periods lasting from several minutes to hours. Because units are the most elementary components of humpback whale songs, they are generally the basis of any studies on the characteristics of songs.

The importance of studying humpback songs during normal conspecific interaction has been pointed out in NRC (2003), suggesting that such information might lead to identifying anomalies when whales are exposed to anthropogenic interactions. Past efforts have generally classified humpback whale songs by listening to individual units and/or visually inspecting spectrograms. Au *et al.* (2006), in studying the source levels of different units, separated nine units by aural discrimination along with spectrographic examination. However, this approach is slow. Worse, the set of units selected by different analysts from the same data could be quite

[a)]Author to whom correspondence should be addressed. Present address: Hawaii Institute of Marine Biology, University of Hawaii, Kaneohe, HI 96744. Electronic mail: ou@hawaii.edu different, and the number of unit types could be an issue of much debate. By creating an automated unit extraction algorithm, we can mathematically classify the units into clusters and produce a deterministic set of unit types that can be regenerated without ambiguity.

The development of an automated extraction algorithm also benefits other applications, such as species identification. From a signal processing standpoint, analyzing marine mammal vocalizations includes detecting sounds from ambient noise, extracting the signals (or, units) and analyzing their features. These methods build a foundation for further classification analysis. For example, species identification could be approached by comparing the signals produced by an unknown species with a library of "template units" from several species. Similarly, an analysis could be performed on songs produced by singers recorded at various locations and times to determine if they are the same group of animals. However, it is both unnecessary and computationally costly to perform such an analysis on all the units extracted from the entire recording because most are usually repetitions of a few basic types with slight variations. This can be solved by applying a unit extraction algorithm prior to the classification. Automated detection and classification of humpback whale calls (and vocalizations of whale species in general) has received significant research interest. Energy detectors such as ISHMAEL (Mellinger, 2001), XBAT (Figueroa, 2007), and PAMGUARD (Gillespie *et al.*, 2008) are the among the most popular humpback detectors built into acoustic analysis packages. However, these methods generally require high signal-to-noise ratio (SNR) to avoid high false detection rates. The recent development of a power-law detector (Helble *et al.*, 2012) works well with low SNR recordings contaminated by shipping noise. It also extracts features such as the start/end time of each call. However, these parameters are not sufficient for call description because additional features need to be extracted to build a classifier. Algorithms based on spectrogram analysis provide more information about the calls for later classification. Spectrogram correlation (Mellinger and Clark, 2000; Abbot *et al.*, 2010) is one of the more popular methods. It compares the spectrogram of recorded signal with a library of calls for detection and classification. Frequency contour tracking is another approach that extracts time-frequency signatures of whale calls (Oswald *et al.*, 2007; Roch *et al.*, 2011; Mohammad and McHugh, 2011; Mallawaarachchi *et al.*, 2008). This approach considers the signal's frequency modulation over time and extracts features such as the contour track, the start/end frequency, the number of up/down sweeps, the duration, etc., for call description. Many classification schemes have been developed to work in conjunction with these feature extraction methods to establish species identify. Leading methods include: The use of classification trees (Oswald *et al.*, 2007), a support vector machine classifier (Mohammad and McHugh, 2011), *k*-means clustering (Brown and Miller, 2007), hidden Markov models (Brown and Smaragdis, 2009; Rickwood and Taylor, 2008; Datta and Sturtivant, 2002), and neural networks (Potter *et al.*, 1994; Mellinger, 2008).

The approach taken in this paper is a synergy of contour extraction and spectrogram correlation with new developments on both sides. Frequency contours are extracted by applying image edge detection filters on the spectrogram of humpback sounds. It is followed by a unit-pairwise comparison that calculates the correlation between contour pixels and assigns weights according to the unit frequency span. An unsupervised learning algorithm divides the units into clusters, and for each cluster, it selects a unit representing the cluster center. Thus the algorithm automatically detects, extracts, classifies, and selects the distinctive unit types for a large dataset. The rest of this paper is divided into four parts. Section II discusses the considerations taken into designing the detection and classification algorithm. Section III explains the method for unit detection with statistics of false alarms and missed detections under different noise conditions. The learning algorithm is introduced in Sec. IV. The algorithm is demonstrated on humpback whale song data recorded during the 2002 Hawaiian Winter season. Finally, conclusions and future research directions are given in Sec. V.

II. DETECTION AND CLASSIFICATION DESIGN CONSIDERATIONS

The main objective of automated unit extraction is to identify distinctive patterns in humpback vocalizations. Detecting humpback units from noisy environment is a necessary first step of the analysis. However, it is more important that the units extracted from the background should possess high qualities (such as high SNR and low time-frequency distortion) so that they can be used for unit type description or as template units for group/species identification. It is possible to apply noise reduction methods on the units, but noise filters usually introduce distortions in the time-frequency domain and reduce the unit's quality as classification templates. Thus achieving a low probability of missed detection is not a concern when the data are noisy. Another argument is that most of the units are repeated many times during a recording, and the chance of missing all the units of one type is low. If most of the recording is of poor quality, we suggest applying noise reduction methods with low time-frequency distortion (Ou *et al.*, 2011) before the analysis.

Reducing the probability of false alarms is necessary. If a noise pattern (such as the frequency tones of motorized boats) is falsely detected as a humpback unit, it will most likely introduce a false unit cluster in the final results. We discriminate humpback units from boat noise on the time-frequency domain by detecting the frequency contours and rejecting the events that lasts more than 5 s. An alternative approach is to reject all the data contaminated by boat noise to ensure high quality.

The same type of units could look slightly different when repeated or in a year-to-year comparison (Au *et al.*, 2006). Therefore the classifier should provide a certain level of tolerance to allow variations of units. Thus, the framework of clustering analysis is used when developing the units classifier. The cluster-based classifier groups similar units in one cluster and assigns a cluster center, i.e., the unit with minimal averaged distance to the rest of the units in the cluster.

The data used in this paper were collected over several years at different sites in Hawaii. The first dataset was collected in the Auau Channel between the islands of Maui, Lanai, Kahoolawe, and Molokai, during the winter season of 2002 and 2003. The data were recorded by divers with a Sony digital audio tape (DAT) recorder encased in an underwater housing at close range to each singer. The experiment was described in Au *et al.* (2006), which showed different data collected with a vertical hydrophone array. The second dataset was recorded using an ecological acoustic recorder (EAR) anchored near French Frigate Shoals (FFS) in the northwestern Hawaiian Islands. A description of the EAR hardware can be found in Lammers *et al.* (2008).

III. DETECTION OF VOCALIZATION UNITS

Vocalization units are detected by their time-frequency contour lines. The detector has been tested with humpback units selected from both the Hawaiian datasets under various noise conditions.

A. Time-frequency contour extraction

In the literature, different methods have been proposed to extract the contours of vocalization units or whistles (of dolphin species) from the spectrogram. For example, Mohammad and McHugh (2011) iteratively learned the shape of contour lines with spectrogram segmentation, and Roch *et al.* (2011) built a regression model for the trajectory of contour lines with particle filters. Our approach is closer to Mohammad and McHugh (2011) as we analyze the spectrogram as an image. Instead of using an iterative method, we apply edge detection filters on the image to search for gray-level discontinuities, which are then connected into contour lines.

The spectrogram of the acoustic time series is calculated using the short time Fourier transform (STFT) with a Hanning window. The window size should be of 2^k-points in length to be able to use the fast Fourier transform (FFT) algorithm for its computational advantage. It is also important to obtain a balanced time-frequency resolution for the vocalization units, such that the matrix (or image) representing a unit should be roughly of equal dimension on both time and frequency. Under these constraints, we apply a 1024-point window with 75% overlap on the time series re-sampled at 10 kHz. With these parameters, a typical one-second humpback song unit with a frequency span of 350 Hz is represented by a 36 × 36 time-frequency matrix.

A smoothing filter is applied on the spectrogram to enhance the quality of the image. This method connects the weak pixels (pixels with low gray levels) and increases the contrast between contour patterns and the background. The filter is implemented in the frequency domain of the image as a two-dimensional low-pass filter. Thus another advantage is that it eliminates high frequency pixels that do not form into any contour shape and are usually caused by broadband noise (such as Gaussian noise or snapping shrimp noise). We emphasize the difference between "high frequency pixels" and "high frequency content of the signal." The former refers to the two-dimensional (2D) discrete Fourier transform (DFT) of the spectrogram image, whereas the later refers to the frequency content of the acoustic data. Let $\bar{S}(x, y)$ denote the spectrogram image, where x and y represent the index of pixels along the time and frequency axis. The 2D-DFT on the spectrogram image has the following expression (Gonzalez and Woods, 2001):

$$F(u,v) = \frac{1}{XY}\sum_{x=0}^{X-1}\sum_{y=0}^{Y-1}\bar{S}(x,y)e^{-2\pi j(ux/X+vy/Y)}. \quad (1)$$

Note that the spectrogram image has been normalized to the range [0,1] before calculating the 2D-DFT. A 2D Gaussian low-pass filtering mask is given by:

$$H(u,v) = e^{-D^2(u,v)/2\sigma^2}, \quad (2)$$

where $D(u, v)$ is the distance from the origin of the Fourier transform. The low-pass filtering is conducted on $F(u, v)$, which represents the frequency domain of the image; the result is then inversed back to the spatial domain to obtain the frequency-enhanced image. Implementation of these steps has been discussed in detail by Gonzalez and Woods (2001), and therefore will not be repeated here.

Figure 1(a) shows six example units produced by humpback whales. The examples shown in this graph were taken from the Auau Channel 2002 dataset. These data were collected using Sony DAT recorder operated by divers in close range to the singers. The whale signals recorded in this experiment have high SNR. However, in tropical waters, it is common to have snapping shrimp noise in the background. To illustrate the image enhancement in the aspect of noise reduction, we added a small amount of white Gaussian noise to the data with an SNR of 20 dB. The SNR is calculated using:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_i g^2(t_i)}{\sum_i w^2(t_i)}, \quad (3)$$

where $g(\cdot)$ is the (discrete) recorded signal, and $w(\cdot)$ is the additive noise. Figure 1(b) shows the enhanced spectrogram after applying a low-pass Gaussian filter on the image. The Gaussian mask used to produce this result is a 7 × 7 squared matrix with standard deviation of $\sigma = 0.9$.

The next step is detecting gray-level discontinuities or "edge lines" in the image. We calculate the second-order derivative of the image to identify the edge points, which are the points of high gray-level transitions comparing with the neighboring points. The discrete second-order derivative is calculated using the gradient operators. Popular choices include Roberts, Prewitt, or Sobel (Gonzalez and Woods, 2001). Sobel operators are selected for this application because they provide extra smoothing by giving higher weight to points closer to the center of the mask. The following matrices give the east-west (on the x dimension) and north-south (on the y dimension) Sobel operators:

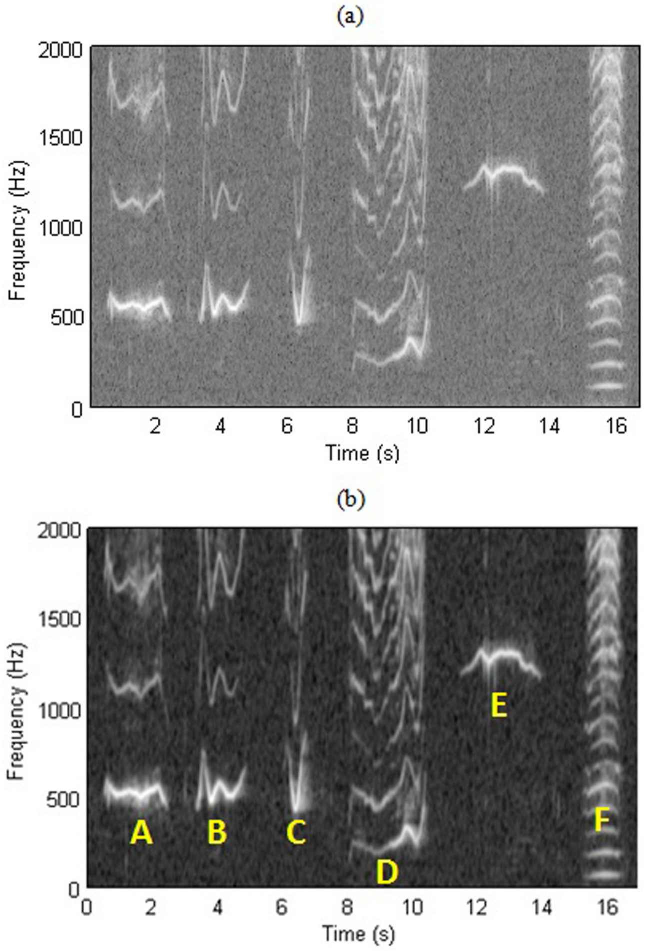

FIG. 1. (Color online) Spectrogram enhancement with a two-dimensional Gaussian filter. (a) A spectrogram showing six humpback whale calls recorded in the Auau chanel, Hawaii, during February to April of 2002. These units are labeled A-F from left to right. White Gaussian noise was added to the data to illustrate the effect of image enhancement. (b) Enhanced spectrogram using a 7 × 7 Gaussian filtering mask, with standard deviation $\sigma = 0.9$.

$$Sl_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \tag{4}$$

and

$$Sl_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}. \tag{5}$$

The gradient magnitude is defined as:

$$\| G \| = \sqrt{(Sl_x * \bar{S}_{xy})^2 + (Sl_y * \bar{S}_{xy})^2}, \tag{6}$$

and the direction of the gradient is given by:

$$\Theta = \arctan\left(\frac{Sl_y * \bar{S}_{xy}}{Sl_x * \bar{S}_{xy}}\right), \tag{7}$$

where * is the convolution operator.

The edge-tracing is then computed using the non-maximum suppression approach of the Canny algorithm (Canny, 1986). The algorithm starts with a search of edge points based on the gradient magnitudes and directions. For example, a zero-degree (i.e., $\Theta = 0$) north-south edge point identified as $\|G\|$ is greater than its east-west neighboring points. The search is repeated for eight directions with $\Theta = 0$, $\pm\pi/4$, $\pm\pi/2$, $\pm 3\pi/4$, and π. Discontinuous (hence noisy) edge points are eliminated by applying a threshold. Figure 2 demonstrates the edge points extracted from the spectrogram shown in Fig. 1.

Edge points detected using the preceding method are connected to form the contour lines. The higher harmonics are not used in unit detection/classification because they are often distorted compared with the contour at the fundamental frequency. Thus the contour- linking algorithm only applies to the points at the fundamental frequency. The contour linking is performed as follows. The binary matrix indicating the location of edge points (such as the image in 2) is summed with respect to the frequency axis. The local maxima of the summation give a rough estimate of unit locations on the time axis. The algorithm searches for the first edge point with a fixed time index (which corresponds to a local maximum) while increasing the frequency index. This edge point is used as the starting point of the contour line. A mask of ones with size 5 × 5 centering at the starting point is applied on the binary image. The direction that gives the maximum product is identified as the next point along the contour line. This computation is repeated until the summation of products becomes one (thus, no more edge points except for the center point) or when the contour line grows back to its starting point. The intuition of using a 5 × 5 mask instead of 3 × 3 is to allow discontinuous edge points to be joined when forming the contour line.

The contour extraction algorithm gives the following results: A binary matrix outlining the shape of the contour, the time duration, and the minimal/maximum frequency of the unit. A contour is considered not-a-unit if the time duration is outside the range $0.3s \le \tau \le 3s$, which should include all typical humpback song units (Au and Hastings, 2008).

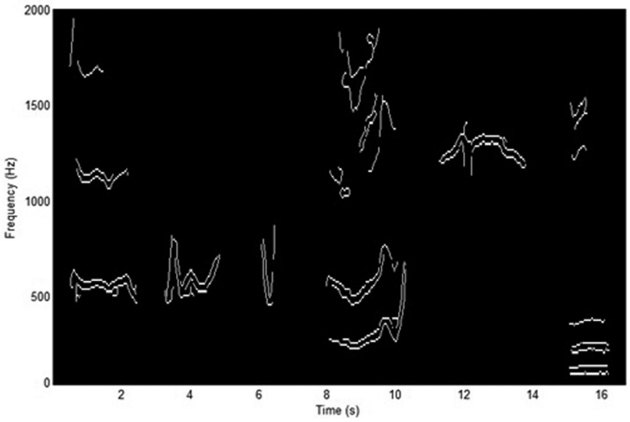

FIG. 2. Edge points extracted from the spectrogram shown in Fig. 1. The edge points are connected to make contour lines. Note the all the units have been correctly detected at the fundamental frequency.

The detector uses only the time duration to determine the presence/absence of a unit, whereas the classifier described in Sec. IV is built based on all these results.

B. Detector performance

Monte Carlo simulations are conducted to quantify the detector performance with known signals for a wide range of SNRs. The signals are the six units shown in Fig. 1(b). Two sets of noise, snapping shrimp and motorized boat noise, have been added to the signals, respectively. The snapping shrimp noise was recorded in Kaneohe Bay, Hawaii, on March 2010. There are no visible humpback whales in range during the recording. The sound has been manually inspected to ensure absence of whale songs. The boat noise was recorded in the Willamette River, Oregon, on February 2010. Besides boat noise, the main ambient noise source came from traffic noise coupling in the water from a nearby bridge. Each noise clip is 40 min in length. The snapping shrimp noise was recorded continuously, whereas the boat noise was picked from a 2 h recording with seven boats. Signals are added in the time domain to a random segment of noise that has been amplified to obtain various SNR levels for the simulation. The SNR has been calculated for each unit using Eq. (3).

Table I shows the probability of false alarm (P_{FA}) versus the probability of missed detection (P_{MD}) for unit types A to F (as labeled in Fig. 1) with varying SNR and noise types. The results are obtained for 6000 trials per statistic (with 1000 trials per unit). Case 1 in the table represents the snapping shrimp noise, and Case 2 represents the boat noise. The resulting P_{FA} remains zero for all the four SNR levels in Case 1, and it is between 0% and 4% for the Case 2 simulations. Results of P_{MD} vary for each unit type. In Case 1, units with fewer harmonics (type A, B, C, E) are less likely to be missed (with $P_{MD} = 0\%$ for SNRs between -3 and 3 dB), whereas the units with more harmonics (type D and F) yield much higher P_{MD} under noisy conditions. Simulations for Case 2 generally have higher P_{MD} than Case 1. It is especially difficult to detect unit F under boat noise: The P_{MD} is more than 80% for all the SNR levels tested. We explain this poor performance in two ways. First, boat noise consists of frequency tones with fundamental frequency between tens of hertz to a few hundred hertz, and their harmonics could reach up to a few kilohertz (Ogden *et al.*, 2011). If the humpback song unit shares a similar fundamental frequency (such as unit F), its contour line could overlap with the frequency tones of a boat, and this makes it extremely difficult to detect using the spectrogram. Second, unit F has many harmonic tones with its energy distributed among them. Its SNR at the fundamental frequency (which is the contour used for detection) is much lower than the overall SNR defined in Eq. (3), thus the statistics are worse compared with other units.

TABLE I. Probability of false alarm (P_{FA}) versus probability of missed detection (P_{MD}) using the contour extraction algorithm for humpback unit detection. Case 1 is with snapping shrimp noise, and Case 2 is with boat noise.

		P_{MD} per unit (%)						
SNR (dB)	Noise	A	B	C	D	E	F	P_{FA} (%)
3	Case 1	0	0	0	0	0	0	0
	Case 2	1	0	30	21	13	84	0
0	Case 1	0	0	0	0	0	0	0
	Case 2	7	10	49	37	33	90	2
−3	Case 1	0	0	0	7	0	20	0
	Case 2	16	23	68	54	42	92	4
−6	Case 1	0	0	0	20	10	36	0

As discussed in Sec. II, the objective of automated contour detection is to extract units with low time-frequency distortion so that these units can be used to describe group/species identities. With low P_{FA} and reasonable P_{MD}, the detector has achieved its intentions.

IV. LEARNING THE HUMPBACK UNITS

A quantification for unit pairwise comparison is introduced based on the contour shape and frequency span. An unsupervised learning algorithm is developed to divide the units into classes. These methods are verified with Monte Carlo simulations, they are also tested on the Auau Channel 2002 dataset.

A. Pairwise comparison between time-frequency contours

The similarity score between two units is quantified using their time-frequency contours, although it is implemented under several assumptions. First, a unit could be repeated (by the same or another singer) with slightly different time duration. We assume that the precise length of duration in the time domain does not discriminate a unit from another if their frequency modulation matches. For this reason, the unit contour with shorter time duration is padded to match the length of the longer unit. The second assumption is that the gray-level of pixels within the unit contour do not add extra information to the unit identity. The time-frequency energy distribution varies slightly when the singer repeats a unit. The information is contained in the frequency modulation within the time-frequency support region, which is the time-frequency contour. We believe the unit identify should be determined by the shape of the contour rather than the waveform amplitudes. Last, units of similar contour shapes but with slight variations in the frequency range are assumed to be the same type. Because our understanding of communications between humpback whales is extremely limited, we can not provide proof of these assumptions. However, for readers who might hold different opinions, this quantification method could still be applied with simple modifications.

Let $B(x, y)$ be the binary matrix defining the shape of a unit contour with x and y being the time and frequency indices. The matrix B is generated by padding ones inside the contour line. Let Δf denote the frequency span of the contour, i.e., $\Delta f = \max(f) - \min(f)$. The similarity score $V_{i,j}$ between units i and j is directly set to zero if any of the following conditions are satisfied:

$$\frac{\min(\Delta f_i, \Delta f_j)}{\max(\Delta f_i, \Delta f_j)} < \eta \tag{8}$$

or

$$\frac{1}{2}\,|\,\Delta f_i \,-\, \Delta f_j\,| \,>\, \xi_c, \text{ with } \Delta f_{i,j} \le \xi_l, \tag{9}$$

where $0 < \eta \le 1$ is a threshold that compares the frequency span between two units, and ξ_c and ξ_l are thresholds in hertz. Equation (8) discriminates units with very narrow Δf from the wider ones. The threshold η is fixed at 0.2 in later computations. Equation (9) discriminates units of narrow Δf but in a very different frequency range. The thresholds are fixed at $\xi_c = 200$ Hz, and $\xi_l = 100$ Hz. For the units that passed these pre-conditions, their similarity score is calculated by cross-correlating B_i and B_j with respect to the frequency axis:

$$V_{i,j} = \frac{\max_{y'}\left\{\sum_{x,y} \bar{B}_i(x, y + y') \cdot B_j(x, y)\right\}}{\sqrt{\sum_{x,y} \bar{B}_i \cdot \sum_{x,y} B_j}}, \tag{10}$$

where $\bar{B}_i$ is the contour matrix of unit i padded to match the time duration of unit j, assuming that unit i is shorter among the two.

B. Unsupervised learning and clustering

Given a database of all the units extracted from songs made by a group of singers, the learning algorithm automatically identifies the distinctive unit types and classifies each unit to a type. Each unit type is a class, denoted by C_k, $k = 1, 2,\ldots, K$ with K being the number of classes. It consists of a cluster of units, denoted by $\mathcal{U}_k^i$, $i = 1, 2,\ldots, N_k$ with N_k being the number of units in the kth class. Units classification is performed under the following conditions:

(1) The pairwise similarity of any two units in the same class should be higher than a threshold δ_i:

$$V(\mathcal{U}_k^i, \mathcal{U}_k^j) \ge \delta_i, \quad \text{for all } \mathcal{U}_k^{i,j} \in \mathcal{C}_k, \tag{11}$$

where $V(\mathcal{U}_m^i, \mathcal{U}_n^j)$ is calculated using Eq. (10).

(2) A unit U^i is assigned to class C_k if Condition 1 is satisfied after adding the unit to the kth class and:

$$\tilde{V}(\mathcal{U}^i, \mathcal{C}_k) \ge \tilde{V}(\mathcal{U}^i, \mathcal{C}_m), \quad \text{forr all } m \ne k, \tag{12}$$

where $\tilde{V}$ $(\mathcal{U}^i, \mathcal{C}_k)$ gives the similarity score between the ith unit and the kth class. The unit-class similarity score is defined as the average of unit pairwise similarity scores:

$$\tilde{V}(\mathcal{U}^i, \mathcal{C}_k) \equiv \frac{1}{N_k}\sum_j V(\mathcal{U}^i, \mathcal{U}_k^j). \tag{13}$$

If a unit is not assigned to any existing class due to Condition 2, a new class is created for this unit. Thus the threshold δ_i determines the size and number of classes. It is selected between $0.5 < \delta_i < 1$. Higher value of δ_i leads to smaller classes. However, the selection of δ_i presents a dilemma. The algorithm generates more classes using a high value of δ_i with each class consisting of a few units with high pairwise similarity scores. It is shown later in Sec. IV C that humpback units exhibit different levels of variations on the time-frequency domain when the same unit type is being repeated. It suggests that units with low pairwise similarity scores could sometimes belong to the same class. Using a lower δ_i could account for unit variations, but it would also lead to higher rate of misclassification due to the loose threshold. Instead the issue of unit variation is solved using an alternative method by introducing the following constraint on class-wise similarity scores:

(3) The similarity score between any two classes should be lower than a threshold δ_c:

$$\bar{V}_{m,n} \equiv \frac{1}{N_m N_n}\sum_{i,j} V(\mathcal{U}_m^i, \mathcal{U}_n^j) \le \delta_c, \tag{14}$$

where $\bar{V}_{m,n}$ defines the similarity score between $\mathcal{C}_m$ and $\mathcal{C}_n$.

If two classes have $\bar{V}_{m,n} > \delta_c$, the units in these two classes are merged into one class. This constraint is applied on the classification results given by the previous two conditions to generate the final set of classes. Note that $\bar{V}_{m,n}$ is defined as the average unit pairwise similarity scores between two classes. By applying Eq. (14), classes with high similarity scores (which are most likely caused by unit variations) are merged without decreasing δ_i. The threshold δ_c is selected between $0 < \delta_c < \delta_i$.

Table II outlines the unit clustering algorithm. It starts with the unit pairwise similarity score matrix **V** such that $\mathbf{V}[i, j] = V(\mathcal{U}^i, \mathcal{U}^j)$ for all $i \ne j$. Lines 1–9 conduct a search on **V** for clusters among the top 5% high-score units. This step generates an initial set of classes containing units that are closest to the center of clusters. Lines 10–13 assign the remaining units to classes according to the two conditions. New classes are created for units that are outside of the "radius" (specified by δ_i) of any existing clusters. In the final step specified by lines 14–17, high-similarity classes are merged according to Eq. (14). Lastly, the algorithm returns the class arrays $\{\mathcal{C}_k\}$, $k = 1, 2,\ldots, K$, with K being the number of classes.

A representative unit is selected for each unit type. It is equivalent to finding the unit that is closest to the cluster center for each class. The center unit has the highest average similarity score compared to all other units in the class. This score is similar to Eq. (13) except that the center unit should be excluded from the class:

$$\tilde{V}_c \equiv \frac{1}{N_k - 1}\sum_{j \ne c} V(\mathcal{U}_k^c, \mathcal{U}_k^j). \tag{15}$$

The unit that gives maximum $\tilde{V}_c$ is selected as the representative unit for the class.

C. A controlled learning test on selected Hawaiian humpback data

This section describes a controlled test of the unit clustering algorithm. From the Auau data and the FFS data, 30 units collected from different singers at different

TABLE II. Implementation of the unit clustering algorithm. The algorithm returns the class arrays $\{\mathcal{C}_k\}$, $k = 1, 2, \ldots, K$, with K being the number of classes.

Algorithm: Unit Clustering

```
Start with units obtained using the automated contour extraction algorithm: $\mathcal{U}^i$,
$i = 1, 2, \ldots, N$. Calculate their pairwise similarity score matrix $\mathbf{V}$, with
$\mathbf{V}[i, j] = V(\mathcal{U}^i, \mathcal{U}^j)$ for all $i \neq j$.
  A quick search for high-score unit clusters:
Initialize the class array $\mathcal{C} = \{\ \}$, with number of classes $K = 0$.
do calculate the maximum value of $\mathbf{V}$ such that $\mathbf{V}[i, j] = \max\{\mathbf{V}\}$
    if $\mathcal{U}^i$(or $\mathcal{U}^j$) $\in \mathcal{C}_k$ and $\mathcal{U}^j$(or $\mathcal{U}^i$) $\in \mathcal{C}_k$
        add $\mathcal{U}^i$ (or $\mathcal{U}^j$) to $\mathcal{C}_k$;
    else
        create a new class $\mathcal{C}_{K+1}$, add $\mathcal{U}^i$ and $\mathcal{U}^j$ to it;
        update the number of classes $K = K + 1$;
    set $\mathbf{V}[i, j] = 0$;
while max $\{\mathbf{V}\} > 0.99$
  Assign the remaining units to classes:
for $i = 1, 2, \ldots, N$
    if $\mathcal{U}^i$ is not assigned to any existing class
        add $\mathcal{U}^i$ to the class specified by Conditions 1
        & 2, or if such a class does not exist, create a
        new class $\mathcal{C}_{K+1}$ for $\mathcal{U}^i$ and update $K = K + 1$;
end for
  Merge classes with high similarity scores:
do
    calculate the class-wise similarity score matrix $\bar{\mathbf{V}}$
    according to (14);
    merge the two classes $\mathcal{C}_m$ and $\mathcal{C}_n$ such that
    $\bar{\mathbf{V}}\ [m, n] = \max\{\bar{\mathbf{V}}\}$, and update $K = K - 1$;
while $\max\{\bar{\mathbf{V}}\} > \delta_c$
```

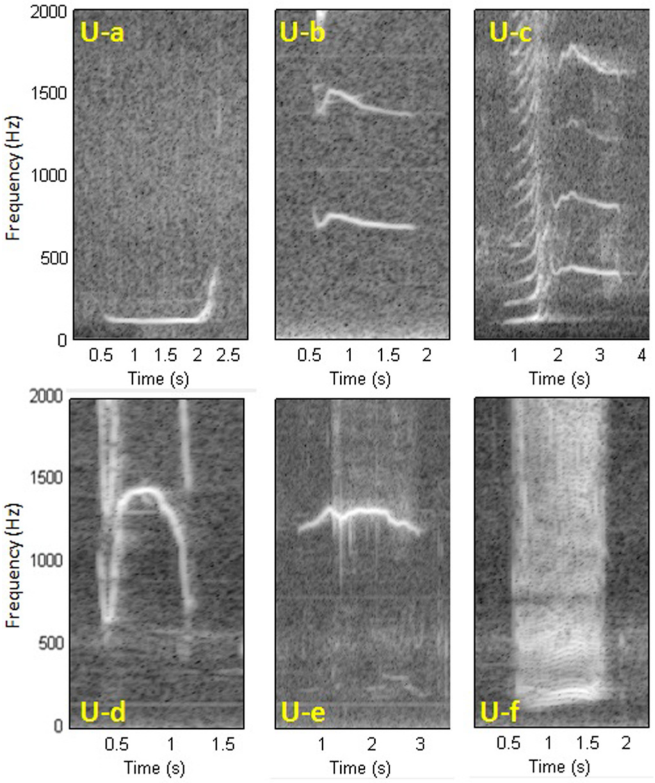

FIG. 3. (Color online) Six unit types selected from the Auau data and the FFS data, for the controlled test of unit clustering algorithm. Each class consists of five units.

locations in 2 yr were selected for the test. These units classify into six unit types by visually inspecting their spectrogram and by listening to the sound clips.

Figure 3 shows the spectrogram of the six unit types. They are labeled as U-a to U-f for convenience. Each type consists of five units by either one singer or multiple singers recorded during different time frames. The number of singers and recording site are listed in Table III, which also gives the mean and standard deviation of the time duration (τ) and frequency range ($f_{\mathrm{min}}, f_{\mathrm{max}}$) for each unit type. These parameters exhibit high standard deviation. For example, Fig. 4 shows three U-d type units produced by the same singer. The three spectrogram plots clearly depict the variations of time-frequency contour in terms of its shape, energy distribution (indicated by the gray-level in the plots), time duration, and frequency span. All these factors have been addressed in the previous section when quantifying the pairwise unit contour similarity score.

In the unit clustering algorithm, the two thresholds δ_i and δ_c control the number of clusters as well as the degree of flexibility allowed for one cluster. By setting $0.50 \leq \delta_i \leq 0.75$ and $0.30 \leq \delta_c < \delta_i$, the algorithm correctly extracted all the six unit types. Further, all the 30 units have been correctly classified under these parameters. Increasing δ_i to more than 0.75 would produce more classes than expected. For example, the third unit shown in Fig. 4 would be identified as a different type if the threshold δ_i is too high. Decreasing δ_c would merge classes that do not necessarily consist of units of high pairwise similarity scores. For example, units of U-a and U-f would be merged into one class if δ_c is too low. Because the clustering results largely depend on the threshold levels, it is important to calibrate them in a controlled test such as shown here. If the method is to be adapted to analyze sounds of different marine mammal species, it is necessary to conduct a calibration test to determine the optimum clustering thresholds.

D. Test on Auau 2002 humpback data

The automated unit extraction and clustering algorithm is applied on the Auau 2002 humpback data. The length of this data is 193 min, which contain the recordings from 12 singers. A total of 951 humpback units are extracted from the data using the automated contour extraction method. The

TABLE III. Composition of the six unit types for the controlled test. The frequency range f_{min} and f_{max} are specified for the time-frequency contour measured at the fundamental frequency.

Unit	τ(s)	f_{min} (Hz)	f_{max} (Hz)	Singers	Site
U-a	1.7 ± 0.3	110 ± 5	430 ± 25	2	FFS
U-b	1.7 ± 0.2	590 ± 50	770 ± 55	1	Auau
U-c	3.0 ± 0.3	90 ± 15	460 ± 50	4	FFS
U-d	1.0 ± 0.1	570 ± 90	1450 ± 85	1	Auau
U-e	2.5 ± 0.3	1190 ± 50	1390 ± 40	2	Auau
U-f	1.4 ± 0.2	100 ± 15	170 ± 30	2	Auau

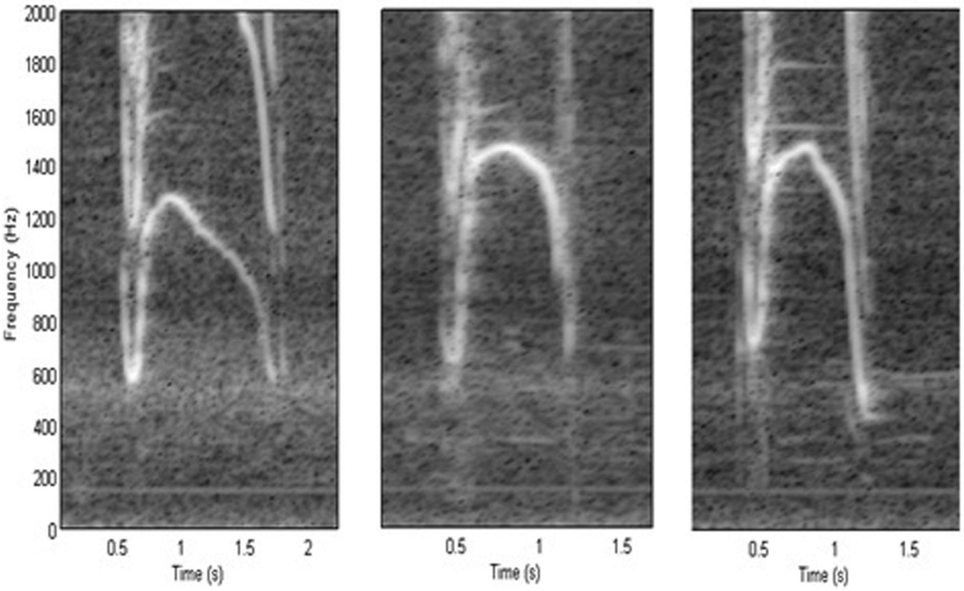

FIG. 4. Examples of the U-d type units which illustrates the variation of time- frequency contour of the same unit type. These units were repeated by the same singer for about 2 min.

unit length τ ranges between 0.4 and 3.7 s. Among all the units, 53.5% of them have $\tau < 1s$, 36.8% of them have $1s \leq \tau < 2s$, and 9.7% have $\tau \geq 2s$.

Twelve unit types are identified by the clustering algorithm using the calibrated thresholds $\delta_i = 0.60$ and $\delta_c = 0.35$. Figure 5 shows the spectrogram of the cluster-center unit for each unit types, which have been labeled U-1 to U-12. According to the range of their fundamental frequency, these unit types can be further divided into four categories. The fundamental frequency of types U-1 through U-4 are the lowest, ranging from tens of hertz to less than 350 hertz. These four unit types constitute 71.0% of all the units extracted from the entire data. Mid-range unit types U-7 to U-10 rank second in quantity with a combined share of 16.6%. Their fundamental frequency ranges from approximately 300 Hz to less than 1000 Hz, whereas their frequency

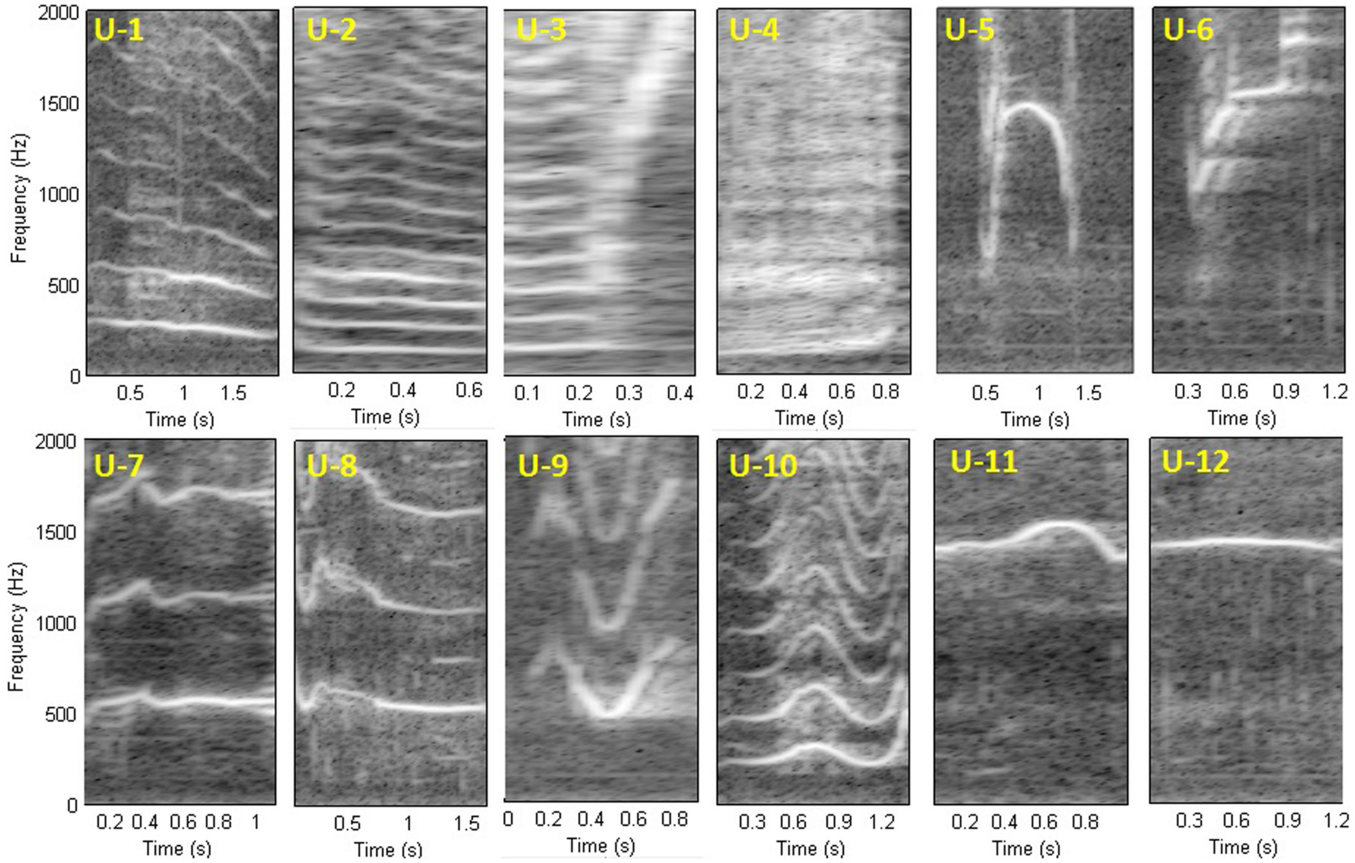

FIG. 5. (Color online) Twelve unit types extracted from the Auau 2002 data. The spectrogram displays the cluster-center unit of each unit type.

span ranges between 120 Hz and approximately 600 Hz. Next are the two high-frequency types U-11 and U-12, which make up 8.3% of the total. Their fundamental frequency is generally above 1300 Hz with the highest being 1870 Hz. The remaining two types U-5 and U-6 both have wide frequency span, ranging between 800 Hz and more than 1000 Hz. Though their time-frequency contour shapes bear little resemblance. These two unit types make up 2.7% and 1.4% of the total, respectively.

V. CONCLUSIONS

Vocalization units of humpback whales are identified by their time-frequency contour in the spectrogram image. Using the contour detecting and classification algorithms introduced in this paper, selection of distinctive unit types can be conducted in an automated analysis. The method has been tested in the presence of snapping shrimp noise and boat noise with humpback data collected at various sites in Hawaii. Using the technique of image edge detection, the algorithm is capable of capturing time-frequency contours of all types of humpback units while maintaining a low probability of false alarms.

The classification step also uses the time-frequency contour of each unit as its signature, thus eliminating the additional feature extraction step and providing a faster solution. The classification is implemented with a unit clustering algorithm, which is capable of recognizing units of the same type but with slight variations in terms frequency modulation and time duration. The clustering algorithm has been validated in a case study containing 30 units of six types with all of them correctly classified. The entire tool of unit detection and classification has been applied to process the deployment data recorded in the Auau Channel in Hawaii in 2002. The results showed 951 humpback song units of 12 distinctive types.

Although the algorithms here have been developed with the motivation of analyzing humpback whale vocalizations, the contour detection and clustering methods have the potential of being extended to capture the characteristics of many other types of marine mammal vocalizations. Further, the unit clustering algorithm is likely to be useful for classifications based on other similarity score quantifications, and can be extended to broader applications.

ACKNOWLEDGMENTS

The authors would like to thank Adam A. Pack of the Department of Biology, University of Hawaii at Hilo, for collecting the acoustic data in Hawaii.

Abbot, T. A., Premus, V. E., and Abbot, P. A. (**2010**). "A real-time method for autonomous passive acoustic detection-classification of humpback whales," J. Acoust. Soc. Am. **127**, 2894–2903.

Au, W. W. L., and Hastings, M. C. (**2008**). *Principles of Marine Bioacoustics* (Springer Science + Business Media, New York), pp. 444–469.

Au, W. W. L., Mobley, J., Burgess, W. C., Lammers, M. O., and Nachtigall, P. E. (**2000**). "Seasonal and diurnal trends of chorusing humpback whales wintering in waters off western Maui," Marine Mammal Sci. **16**, 530–544.

Au, W. W. L., Pack, A. A., Lammers, M. O., Herman, L. M., Deakos, M. H., and Andrews, K. (**2006**). "Acoustic properties of humpback whale songs," J. Acoust. Soc. Am. **120**, 1103–1110.

Brown, J. C., and Miller, P. J. O. (**2007**). "Automatic classification of killer whale vocalizations using dynamic time warping," J. Acoust. Soc. Am. **122**, 1201–1207.

Brown, J. C., and Smaragdis, P. (**2009**). "Hidden Markov and Gaussian mixture models for automatic call classification," J. Acoust. Soc. Am. **125**, EL221–EL224.

Canny, J. (**1986**). "A computational approach for edge detection," IEEE Trans. Pattern Anal. Mach. Intell. **8**, 679–698.

Cerchio, S., Jacobsen, J. K., and Norris, T. F. (**2001**). "Temporal and geographical variation in songs of humpback whales, *Megaptera novaeangliae*: Synchronous change in Hawaiian and Mexican breeding assemblages," Anim. Behav. **62**, 313–329.

Datta, S., and Sturtivant, C. (**2002**). "Dolphin whistle classification for determining group identities," Signal Process. **82**, 127–327.

Figueroa, H. (**2007**). XBAT. v5., Technical Report (Cornell University Bioacoustics Research Program, Ithaca, NY).

Frankel, A. S., Clark, C. W., Herman, L. M., and Gabriele, C. M. (**1995**). "Spatial distribution, habitat utilization, and social interactions of humpback whales, *Megaptera novaeangliae*, off Hawaii, determined using acoustic and visual techniques," Can. J. Zool. **73**, 1134–1146.

Gillespie, D., Gordon, J., Mchugh, R., Mclaren, D., Mellinger, D., Redmond, P., Thode, A., Trinder, P., and Deng, X. (**2008**). "PAMGUARD: Semi-automated, open source software for real-time acoustic detection and localization of cetaceans," in *Proceedings of the Institute of Acoustics*, Vol. 30, Pt. 5.

Gonzalez, R. C., and Woods, R. E. (**2001**). *Digital Image Processing*, 2nd ed. (Prentice-Hall, Upper Saddle River, NJ), pp. 567–585.

Helble, T. A., Ierley, G. R., D'Spain, G. L., Roch, M. A., and Hildebrand, J. A. (**2012**). "A generalized power-low detection algorithm for humpback vocalizations," J. Acoust. Soc. Am. **131**, 2682–2699.

Helweg, D. A., Cato, D. H., Jenkins, P. F., Garrigue, C., and McCauley, R. D. (**1998**). "Geographic variation in South Pacific humpback whale songs," Behav. Ecol. **135**, 1–27.

Helweg, D. A., Frankel, A. S., Mobley, J., and Herman, L. M. (**1992**). "Humpback whale song: Our current understanding," in *Marine Mammal Sensory Systems*, edited by J. A. Thomas, R. A. Kastelein, and A. S. Supin (Plenum, New York), pp. 459–483.

Lammers, M. O., Brainard, R. E., Au, W. W. L., Mooney, T. A., and Wong, K. B. (**2008**). "An ecological acoustic recorder (ear) for long-term monitoring of biological and anthropogenic sounds on coral reefs and other marine habitats," J. Acoust. Soc. Am. **123**, 1720–1728.

Lammers, M. O., Fisher-Pool, P. I., Au, W. W. L., Meyer, C. C., Wong, K. C., and Brainard, R. E. (**2011**). "Humpback whale (*Megaptera novaeangliae*) wintering behavior in the northwestern Hawaiian islands observed acoustically," Mar. Ecol. Prog. Ser. **423**, 261–268.

Mallawaarachchi, A., Ong, S. H., Chitre, M., and Taylor, E. (**2008**). "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," J. Acoust. Soc. Am. **124**, 1159–1170.

Mellinger, D. (**2001**). "ISHMAEL" 1.0 Users Guide, Technical Report, NOAA Technical Memorandum OAR PMEL-120 (NOAA/PMEL7600, 98115-6349).

Mellinger, D. (**2008**). "A neural network for classifying clicks of Blainville's beaked whales (*Mesoplodon densirostris*)," Can. Acoust. **36**, 55–59.

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

Mohammad, B., and McHugh, R. (**2011**). "Automatic detection and characterization of dispersive North Atlantic right whale upcalls recorded in a shallow-water environment using a region-based active contour model," IEEE J. Ocean Eng. **36**, 431–440.

NRC (**2003**). *Ocean Noise and Marine Mammals* (National Academy, Washington, DC), pp. 1–204.

Ogden, G. L., Zurk, L. M., Jones, M. E., and Peterson, M. E. (**2011**). "Extraction of small boat harmonic signatures from passive sonar," J. Acoust. Soc. Am. **129**, 3768–3776.

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.

Ou, H. H., Allen, J. S., and Syrmos, V. L. (**2011**). "Frame-based time-scale filters for underwater acoustic noise reduction," IEEE J. Ocean Eng. **36**, 285–297.

Payne, K., Tyack, P., and Payne, R. (**1983**). "Progressive changes in the songs of humpback whales (*Megaptera novaeangliae*): A detailed analysis of two seasons in Hawaii," in *Communication and Behavior of Whale*, edited by R. Payne (Westview, Boulder, CO), pp. 9–57.

Payne, R. S., and McVay, S. (**1971**). "Songs of humpback whales," Science **173**, 585–597.

Potter, J., Mellinger, D., and Clark, C. (**1994**). "Marine mammal call discrimination using artificial neural networks," J. Acoust. Soc. Am. **96**, 1255–1262.

Rickwood, P., and Taylor, A. (**2008**). "Methods for automatically analyzing humpback song units," J. Acoust. Soc. Am. **123**, 1763–1772.

Roch, M. A., Brandes, T. S., Patel Y. B., Baumann-Pickering, S., and Soldevilla, M. S. (**2011**). "Automate extraction of odontocete whistle contours," J. Acoust. Soc. Am. **130**, 2212–2223.

Spitz, S. S., Herman, L. M., Pack, A. A., and Deakos, M. H. (**2002**). "The relation of body size of male humpback whales to their social roles on the Hawaiian winter grounds," Can. J. Zool. **80**, 1938–1947.

Tyack, P. L. (**1981**). "Interactions between singing Hawaiian humpback whales and conspecifics nearby," Behav. Ecol. Sociobiol. **8**, 105–116.

Winn, H. E., Perkins, P. J., and Poulter, T. (**1970**). "Sounds of the humpback whale," in *Proceedings of the 7th Annual Conf. Biological Sonar* (Stanford Research Institute, Menlo Park, CA), pp. 39–52.