Automatic acoustic estimation of sperm whale size distributions achieved through machine recognition of on-axis clicks

Wilfried A. M. Beslin[a] and Hal Whitehead
Department of Biology, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada

Shane Gero
Zoophysiology, Institute for Bioscience, Aarhus University, Aarhus, Denmark

(Received 20 June 2018; revised 28 October 2018; accepted 14 November 2018; published online 26 December 2018)

The waveforms of individual sperm whale clicks often appear as multiple pulses, which are the product of a single pulse reverberating throughout the spermaceti organ. Since there is a relationship between spermaceti organ size and total body size, it is possible to estimate a whale's length by measuring the inter-pulse intervals (IPIs) within its clicks. However, if a click is recorded off-axis, the IPI corresponding to spermaceti organ length is usually obscured. This paper presents an algorithm for automatically estimating the "true" IPIs of sperm whales in a recording by measuring them from on-axis clicks only. The routine works by classifying detected clicks with a support vector machine, assessing the stability of their IPIs, and then clustering the stable IPIs using Gaussian mixture models. Results show that the routine is very accurate in obtaining reliable IPIs, but has a high false negative rate. Nonetheless, since sperm whales click very frequently, it is possible to obtain useful IPI distributions with only a few minutes of recording. This algorithm makes it possible to estimate the body lengths of multiple sperm whales automatically with only one hydrophone. An implementation is available for download at http://whitelab.biology.dal.ca/CABLE/cable.htm.
 https://doi.org/10.1121/1.5082291

[WWA] Pages: 3485–3495

I. INTRODUCTION

Passive acoustic monitoring (PAM) has become a popular means of studying whales and dolphins over the past several years. With better recording equipment, sound analysis tools, and the realization that cetaceans are more easily observed acoustically than visually, PAM is increasingly being used to supplement or replace traditional visual surveys (Thomas *et al.*, 1986; Mellinger *et al.*, 2007). The sperm whale is very well suited to study through PAM, since this species spends most of its time foraging at depth (Watwood *et al.*, 2006), during which it typically produces loud clicks (Backus and Schevill, 1966; Whitehead and Weilgart, 1990). The incorporation of passive acoustics into sperm whale surveys has significantly increased the range and sensitivity of detection (Barlow and Taylor, 2005).

Sperm whale clicks also possess an interesting feature: a single click is composed of multiple pulses (Backus and Schevill, 1966). According to the accepted "bent-horn" model of sperm whale sound production (Norris and Harvey, 1972; Møhl, 2001), these pulses are the product of reverberations between air sacs at the front and back of the spermaceti organ. As a consequence, the inter-pulse interval (IPI) is directly related to the length of the spermaceti organ. Since there is also an allometric relationship between spermaceti organ length and body length (Nishiwaki *et al.*, 1963; Clarke, 1978; Gordon, 1991), it is possible to estimate a whale's body length simply by measuring its IPI (Norris and Harvey, 1972; Møhl *et al.*, 1981; Adler-Fenchel, 1980; Gordon, 1991; Rhinelander and Dawson, 2004; Growcott *et al.*, 2011). This feature makes PAM especially informative for sperm whales.

Unfortunately, however, most sperm whale clicks from typical far-field recordings do not display a clear structure suitable for IPI calculation. They often appear with extra pulses at variable locations, making the pulse interval irregular. These extra pulses arise because of directionality: sperm whale clicks are highly directional, and their structure in both frequency and time appears different based on the position of the receiver relative to the whale's acoustic axis (Møhl *et al.*, 2003; Zimmer *et al.*, 2005). Only clicks recorded on-axis (i.e., directly in front or behind) display the characteristic multi-pulse structure representative of the spermaceti organ size. Clicks recorded off-axis are confounded by omnidirectional reflections from the air sacs (Zimmer *et al.*, 2005). As a consequence, IPI calculation is actually a difficult task, because it requires that on-axis clicks be separated from the off-axis ones.

Most studies that have used IPIs have worked around the directionality problem by manually searching for and removing off-axis clicks (e.g., Adler-Fenchel, 1980; Gordon, 1991; Drouot *et al.*, 2004; Rendell and Whitehead, 2004; Rhinelander and Dawson, 2004; Schulz *et al.*, 2011). This method is effective, but can quickly become impractical, as just 1 h of recording can yield over 4000 clicks per whale. Another approach is to average every click in a sequence (Teloni *et al.*, 2007; Antunes *et al.*, 2010). Averaging works because the stability of the true IPI allows it to emerge above the noise, but this assumes that each click was produced by the same whale. Thus, this method is only reliable if individual

[a]Electronic mail: wilfried.beslin@dal.ca

click trains can be separated, which is difficult and impractical in many situations.

One approach to IPI compilation that has not been tested until now is using automatic classification to isolate on-axis clicks. A great advantage to this approach is that it does not require knowledge of which whale produced which clicks, so click trains do not need to be resolved. The goal of this research was to produce a software tool capable of compiling reliable IPI distributions automatically, based on machine classification of clicks. This tool is designed to be as simple to use as possible, requiring only a single-channel audio recording file as input. The output consists of filtered IPI distributions, with estimates of how many whales are present, what their true IPIs are, and ultimately their body sizes. Such a tool could greatly enhance the effectiveness of PAM for sperm whales.

This paper describes how the tool was developed, its underlying algorithms, and its performance. The tool itself can be downloaded at http://whitelab.biology.dal.ca/CABLE/cable.htm.

II. METHODS

A. Data collection

The routine was developed using a primary dataset of recordings collected off the west coast of the island of Dominica in the Eastern Caribbean from February to April 2015. These recordings were collected as part of a long-term behavioural research program on female sperm whale societies (see Gero *et al.*, 2014). A secondary dataset collected off the Galápagos Islands from January to May 2014 was also used for additional testing.

In both regions, the same equipment and protocols were used. Female and immature sperm whales were followed aboard a 12-m auxiliary sailing vessel. Acoustic recordings were made using a custom-built towed hydrophone (Benthos AQ-4 elements, frequency response 0.1–30 kHz) and a filter box with high-pass filters up to 1 kHz. This resulted in a recording chain with a flat frequency response across a minimum of 2–20 kHz. Audio data were collected through a computer-based recording system, with a sampling rate of either 48 or 96 kHz, and 16-bit resolution. All recordings were stored in WAVE format. Recordings were categorized into two types based on how they were obtained, which are referred to as "first-click" and "standard". The Dominica dataset included both types, while the Galápagos dataset consisted of only standard recordings.

In the first-click protocol, acoustic recording was initiated immediately after a whale began a foraging dive. The research vessel remained stationary. In this scenario, since the whale is near and facing almost directly away from the research vessel during its descent, the echolocation clicks it produces are likely to be perceived clearly and on-axis. The purpose of first-click recordings was to obtain samples of on-axis sperm whale clicks. Since the animals of interest were facing away from the hydrophone in these recordings, only backward on-axis clicks could be characterized. However, forward clicks should not pose a problem (see Appendix A in the supplementary material[1] for justification). A total of 7 first-click recordings from Dominica were used for this purpose. These lasted between about 3.5 and 7.5 min (36 min total). Due to the social nature of female sperm whales, most first-click recordings captured more than one animal diving at the same time. Based on photographic identification of the flukes of diving whales (Arnbom, 1987), these seven recordings contained clicks from at least ten adult female-sized individuals in total.

In the standard protocol, acoustic recording was initiated at predetermined time periods (usually 1 h intervals) during days when sperm whales were encountered. In some cases, the research vessel was stationary, while in others it was sailing or motoring at low speed. In this scenario, the location, orientation, number, and identity of whales immediately surrounding the hydrophone is usually unknown. Since sperm whales spend most of their time foraging, any whales present during standard recordings are likely to produce echolocation clicks. However, these clicks may be perceived from any angle, and the majority are typically off-axis. Standard recordings were normally run for 4 min, although on a few occasions, this varied between 3 and 15 min. Standard recordings were used to obtain click samples typical of most PAM situations. A total of 174 standard recordings were used from Dominica, representing 14 h total. The Galápagos dataset consisted of 141 standard recordings, representing 10 h and 12 min total.

B. Software design overview

The routine takes digital audio files as input (WAVE format), filters the contents automatically for on-axis sperm whale clicks through a series of steps, and outputs the IPIs of the filtered clicks along with estimates of animal counts, their IPIs, and body lengths. All analysis is conducted at 48 kHz. If the original sampling rate is different, then the recording is resampled automatically. Recordings can have any number of channels, but only one is used. This program uses MATLAB version R2015a with the Signal Processing Toolbox, the Statistics and Machine Learning Toolbox, the Curve Fitting Toolbox, and the Parallel Computing Toolbox (The MathWorks, Inc., Natick, Massachusetts). Sections II C–II H describe each step in the algorithm, with further details on certain steps expanded upon in Appendix A in the supplementary material.[1] A complete graphical representation of the algorithm is also available in Appendix B in the supplementary material.[1]

The routine uses several parameters that can be adjusted. With the exception of two key parameters (discussed later), a complete sensitivity analysis was beyond the scope of this work. However, each parameter has a default value that was established based on published information, data observations, and/or robustness to various signal-to-noise ratios (SNRs). Thus, defaults should be reliable and widely applicable. Parameters are described in Appendix C in the supplementary material.[1]

C. Audio loading and preprocessing

The program extracts the recorded sound pressure waveform from one channel (the first by default) of an input audio file. If the waveform needs to be resampled, a finite impulse

response anti-aliasing filter is applied. This filter uses a Kaiser window and has an order of $50 \times \max(p, q)$, where p/q is the reduced resampling ratio. The waveform is then noise-filtered using a 2–12 kHz Butterworth bandpass filter, run in both directions to avoid non-linear frequency-dependent delay (i.e., zero-phase filtering). Although sperm whale clicks contain some energy outside this band, IPIs were generally more stable when limited to 2–12 kHz. Filter order after zero-phase filtering is 12.

D. Click detection

Candidate sperm whale clicks are detected within the time series by a custom click detection algorithm. This algorithm is based on the Page test (Page, 1954), a method commonly used to isolate cetacean clicks. The particular implementation used here is similar to the ones described by Miller (2010) and Zimmer (2011), and used by the open-source PAM software PAMGUARD (Gillespie *et al.*, 2008). It was adapted to capture the multi-pulsed structure of sperm whale clicks as accurately as possible.

The detector essentially consists of two steps. The first step is the Page test, which finds regions in the time series that correspond to potential sperm whale clicks. This is followed by a validation step, which edits these regions and establishes the start and end periods of each click. Clicks are detected based on the SNR. Signal and noise power are computed based on the square of the waveform envelope, where envelope is computed as the absolute value of the analytic signal (obtained using the Hilbert transform). Details of the click detection process are included in Appendix A in the supplementary material.[1]

E. Feature extraction

After candidate clicks have been detected, the program computes a set of spectral and temporal features from each click. The purpose of these features is to provide information from which on-axis sperm whale clicks can be differentiated from off-axis clicks and other transients. A description of each feature and the feature selection process is included in Appendix A in the supplementary material.[1]

Most features depend on information that must be computed beforehand, including the location of individual pulses within a click, the frequency spectra of clicks, and exponential curve fits. Sections II E 1–II E 3 describe the calculation of these dependencies in more detail.

1. Pulse detection

The isolation of individual pulses within a sperm whale click is particularly tricky, because the noise level within clicks is often highly variable, and later pulses may be fainter than the average noise. Thus, conventional click detection does not perform well at this resolution, because a fixed SNR threshold risks rejecting many pulses, or detecting many spurious ones. Therefore, a different approach was used. This approach involves signal smoothing, followed by the detection of local maxima. Details on this procedure are in Appendix A in the supplementary material.[1]

2. Frequency spectrum calculation

To compute spectra, a Tukey window is applied to each click, where the flat portion always encompasses the entire click. The Fast Fourier Transform (FFT) is then applied to each windowed click. The number of points used in FFT is the number of samples within the window of the longest click in the file, rounded up to the next integer power of 2. Thus, the number of FFT points is consistent for each click in a file, but can vary between files.

3. Exponential fitting

This process depends on pulse detection and is intended to describe the amplitude decay of pulses in on-axis clicks. It involves the least-squares fitting of exponential curves of the form

$$y(t) = \alpha e^{\beta t}, \tag{1}$$

where y corresponds to the peak pulse amplitudes (measured from the waveform envelope), and t is time. For every click, this equation is fit to the peaks of all pulses composing the click. To standardize these fits and ensure that the coefficients are comparable across clicks, each click undergoes two transformations before the fit is applied. First, the whole click is scaled in amplitude such that its tallest peak is equal to one. Second, it is scaled along the time axis so that the first pulse's peak occurs at $t = 0$, and the mean delay between consecutive peaks is equal to one. This is done in an attempt to standardize the IPI in a manner that is robust to variability in the number of pulses detected within clicks.

F. Click classification

The next step uses the extracted features to automatically classify each click as being an on-axis sperm whale echolocation click ("Good") or not ("Bad"). Only echolocation clicks are considered Good. Coda clicks, which are used for communication (Watkins and Schevill, 1977; Weilgart and Whitehead, 1993), are considered Bad in this case, because they differ slightly from echolocation clicks in their structure (Madsen *et al.*, 2002) and IPI (Schulz *et al.*, 2011; Bøttcher *et al.*, 2018). Classification is performed by a support vector machine (SVM) that uses a quadratic kernel. Using Platt's (1999) method, SVM scores for each click are modified to estimate the probability that the click is Good. The routine accepts clicks as Good if their probability scores are above a certain threshold (discussed in Sec. II I).

The SVM was trained using a dataset of clicks automatically detected from all seven first-click recordings. Clicks were labeled as being Good or Bad by an observer (W.B.), resulting in a dataset of 487 Good clicks and 6499 Bad clicks. Good clicks were identified based on clear multi-pulsed waveforms characteristic of on-axis clicks, as described by Zimmer *et al.* (2005). The accuracy of the SVM was assessed based on tenfold cross-validation. Further details on classifier training are included in Appendix A in the supplementary material.[1]

G. IPI calculation and validation

After each click has been automatically classified as Good or Bad, the routine computes IPIs for all Good clicks. This is done using the two methods proposed by Goold (1996): autocorrelation analysis and cepstral analysis. Thus, each Good click initially has two IPI estimates. For both methods, the program constrains IPI calculation between 2 ms, and either 9 ms or the click duration, whichever is shorter. The upper bound of 9 ms is a limit for the IPIs of large male sperm whales, while the lower bound of 2 ms is used to avoid confusion from high correlations within wide first pulses (Marcoux *et al.*, 2006). However, this lower bound may exclude clicks produced by young calves (Tønnesen *et al.*, 2018).

For cepstral analysis, the power cepstrum is computed as

$$C_q = |\mathrm{FFT}(\log_{10}(|\mathrm{FFT}(x_t)|^2))|. \tag{2}$$

To get a good signal in the cepstrum, it is best if all pulses have the same amplitude. To facilitate this, clicks are windowed before the first FFT, where the window function consists of chi-squared probability densities as suggested by Goold (1996),

$$f_w(n) = \frac{1}{2^{k/2}\Gamma(k/2)} n^{(k/2)-1} e^{-(n/2)}, \tag{3}$$

where Γ is the gamma function for positive integers

$$\Gamma(k) = (k-1)! \tag{4}$$

In all cases, k is set to 4, as this value appeared most appropriate based on visual inspection of several windowed clicks. The second FFT uses a Tukey window with the flat part spanning 2–12 kHz. The number of samples used for each window is the maximum of either the smallest integer power of 2 that is larger than the number of samples within the longest click, or the smallest integer power of 2 such that the "Nyquist quefrency" is greater than the upper IPI limit.

For each click, a final IPI is obtained by averaging the autocorrelation and cepstral IPIs. The point of using both methods is to improve confidence in the IPI estimate. Neither method on its own is perfect (Antunes *et al.*, 2010; Bøttcher *et al.*, 2018), but if they both return the same number, then the final IPI is likely to be reliable. Therefore, the next step in the routine rejects all clicks whose two IPI estimates deviate from the average by more than 0.05 ms by default, as in Schulz *et al.* (2011).

After each IPI has been calculated and validated for precision, a final validation step is performed. This involves searching for IPI repetitions. Since sperm whales emit echolocation clicks in trains at short, regular intervals, it is expected that the same IPI will be recorded more than once within a few seconds. The program exploits this property to further validate the IPIs it has measured. For each Good click with a precise IPI, the routine scans the time series locally about the click's time of occurrence, in both directions. The target of this scan is another click with the same IPI as the focal click, within tolerance ($\pm$0.05 ms by default). If the scan is successful, then a new scan is performed about the repeated click. This cycle continues for as many repetitions as specified. To reduce confusion, a "repetition" is explicitly defined as being one recurring instance (within tolerance) of an IPI within a neighboring click. Based on this definition, "zero repetitions" means that a click has no neighbors with a similar IPI, "one repetition" means that a click has one neighbor with a similar IPI, and so on.

When searching for the first repetition, the scan is conducted within a broad range of typical sperm whale interclick-interval (ICI) values from the original click (0.25–1.5 s by default). For subsequent repetitions, the range is narrowed such that only clicks with the same ICI as that separating the previous two clicks ($\pm$0.2 s by default) are considered. If more than one click is found within range, then each click is used to search for successive repetitions until the required number of repetitions has been met. All clicks with an insufficient number of successive IPI repetitions within ICI range are removed. Those with enough repetitions contribute to the final IPI distribution.

H. Animal count and length estimation

The number of whales present and their body lengths are estimated through cluster analysis of the filtered IPI distribution. This is accomplished using Gaussian mixture models (GMMs). IPI measurements from individual whales have been found to be quite stable, often within $\pm$0.05 ms from the mean (Schulz *et al.*, 2011; Antunes *et al.*, 2010; Growcott *et al.*, 2011). Therefore, output IPI distributions are expected to contain mixtures of narrow peaks, where each peak corresponds to an individual whale (assuming each whale present in the recording has a distinct IPI). Mixture modeling is thus a suitable approach for resolving the composition of IPI distributions.

GMMs are fitted through the Expectation-Maximization (EM) algorithm, which is an iterative process for estimating the most likely parameter values. It requires that the number of clusters, k, be specified beforehand, and may also take estimates of cluster means, variances, and proportions to accelerate convergence. In this case, GMMs are initialized based on two Gaussian kernel density estimates (KDEs). One kernel uses a wide bandwidth (0.0333 by default), while the other has a narrow bandwidth (0.0167 by default). GMMs are run for every k within the range $\max(1, N_{\mathrm{wide}} - 1)$ to $N_{\mathrm{narrow}} + 1$, where N_{wide} and N_{narrow} represent the number of peaks found within the wide and narrow bandwidth KDE functions, respectively. Initial estimates of cluster means and proportions are also based on KDE peaks. In the case where $k = N_{\mathrm{wide}} - 1$, a peak is removed at random. Likewise, for $k = N_{\mathrm{narrow}} + 1$, a peak is added at random. For values of k in-between, the peaks to use are decided based on sparseness, where the most isolated peaks are added first as k increases. The initial standard deviation varies based on the value of k and the bandwidths of the two KDEs. To avoid numerical instabilities, and also account for IPI quantization to some degree, a value of $(1/Fs)/4$ is added to each standard deviation during the EM estimation, where Fs is in kilohertz. By default, standard deviation is

constrained to be identical for all clusters in a model. GMMs with different numbers of clusters are compared to one another using the Bayesian Information Criterion (BIC), where smaller values represent better support. When analyzing an IPI distribution through mixture modeling, the routine returns models for all k's tested, ranked by order of smallest BIC.

The results of GMM clustering provide insight into how many whales are present, and what their IPIs are. The means of each cluster in a GMM are estimates of each whale's true IPI. From these measures, body lengths can be estimated using equations such as those published by Gordon (1991) and Growcott *et al.* (2011).

I. Performance analysis

The performance of the automatic IPI filter was assessed primarily by examining its output for standard recordings from Dominica. The routine was run multiple times for each recording, where two parameters were changed for each run: the minimum probability at which a click is considered Good, and the number of times each IPI must be repeated in succession. Minimum probability was tested for values of 0.1, 0.3, 0.5, 0.7, and 0.9, with IPI repetition fixed at 1. IPI repetition was tested for values of 0 through 4, with minimum probability fixed at 0.7. Ideally, to measure routine performance, the "true" probability distribution of IPIs for each whale present during the standard recordings would need to be known. Unfortunately, this information is extremely difficult to obtain in the field and was not available for this analysis. Thus, performance was assessed using two alternative approaches. One of these is called "peak definition," which measures the stability of variance among clusters of IPIs in a given recording. "Peak" in this sense refers to areas of high density in IPI distributions that appear roughly normally distributed, and presumably correspond to the IPIs of individual whales. The other measure of performance is referred to as "accuracy," with respect to manually compiled IPI distributions. Accuracy in this sense quantifies how two probability distributions are similar to one another.

To measure peak definition, mixture modeling was applied. For each distribution, two GMMs with an equal number of clusters were compared, where one model required all clusters to have the same variance, and the other did not. The constrained model consisted of the "best" model output by the same procedure used for estimating whale lengths. The unconstrained model was obtained by running the EM algorithm for the same k as the constrained model, with initial parameter estimates also equal to the constrained model values. Peak definition was measured as the log likelihood ratio between these two models,

$$\log(\Lambda) = \log\left(\frac{\mathcal{L}_{\text{constrained}}}{\mathcal{L}_{\text{unconstrained}}}\right). \tag{5}$$

The argument for using this measure is that individual whales are not expected to differ greatly in their IPI variation, so models with shared variance should fit reasonably well. If the likelihood of unconstrained (and thus potentially overfit) models is considerably better, then this suggests that individual clusters may not be clear. Distributions with clear peaks should have a $\log(\Lambda)$ close to zero. Since $\log(\Lambda)$ is necessarily zero when $k = 1$, those cases were ignored.

To measure accuracy (with respect to manually-compiled IPI distributions), nine standard recordings from Dominica were selected, which varied in quality from good to poor. Quality was indicated by the mean peak SNR of each detected click. Manual IPI calculations were performed by one observer (W.B.); details on this process are included in Appendix A in the supplementary material.[1] Manual and automatic distributions were compared through shared-variance GMMs, which were fit using the same process as for length estimation. The accuracy of the automatic method was measured as the total overlap in area between the probability density functions of the two GMMs.

Since the automatic routine was developed entirely using clicks obtained from Dominica, it is of interest to examine how it performs under different scenarios. To this end, automatic IPI distributions from the 141 standard Galápagos recordings were also examined. In addition to oceanographic differences, the Galápagos differs significantly from Dominica in that sperm whales are typically grouped in much greater numbers (Whitehead *et al.*, 2012).

III. RESULTS

A. Classifier performance

Overall accuracy of the SVM in classifying on-axis sperm whale clicks versus other transients, as estimated by tenfold cross-validation before Platt transformation, is 98.8%. When adjusted to account for the imbalance in frequency between each class, accuracy is 94.3%. Sensitivity (a.k.a., true positive rate, or recall) is 89.1%, specificity (a.k.a., true negative rate) is 99.5%, and precision is 92.7%.

B. Output IPI distributions

In many cases, the routine filtered out all IPIs, resulting in empty distributions. This occurred with greatest frequency when the number of required IPI repetitions was high (e.g., two repetitions or higher). However, based on visual inspection of each distribution, many of those that were not empty appeared as expected, in the sense that they contained narrow peaks at values appropriate for Caribbean sperm whales (Fig. 1). GMMs also appeared to detect these peaks quite accurately in most cases. Those distributions that were not as clear either had very few IPIs, had many small clusters (most likely noise), or had many peaks very close to each other. The vast majority of noisy distributions occurred when the filter did not require IPIs to be repeated. In general, for distributions with many IPIs, increasing the required number of repetitions resulted in clearer patterns, but at the cost of missing peaks.

Peak definition, as assessed by the likelihood ratio between GMMs with shared and unshared variance, was often worse when IPIs did not need to be repeated. For all cases where IPIs did need to be repeated, peak definition was usually quite good in comparison, and did not change greatly with the number of repetitions [Fig. 2(a)] or the probability

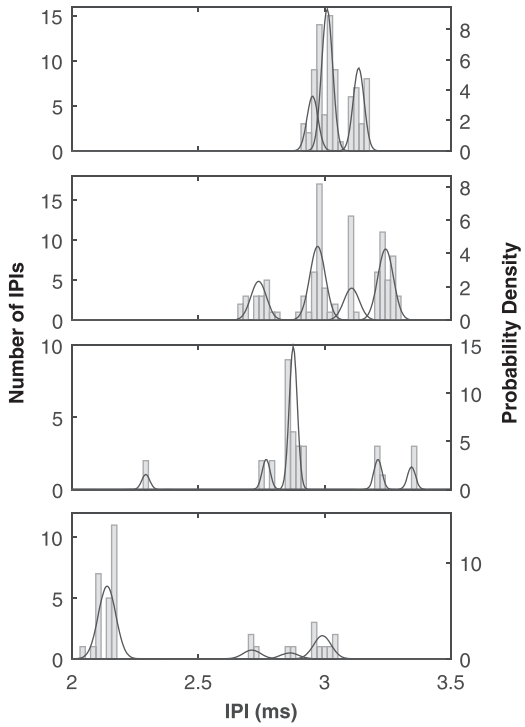

FIG. 1. Example IPI distributions output from four standard recordings from Dominica. Filtration parameters were set at 1 IPI repetition, and a goodness probability threshold of 0.7. Black lines represent probability density functions of clusters from the best GMMs, according to BIC. Bin width = 1/Fs.

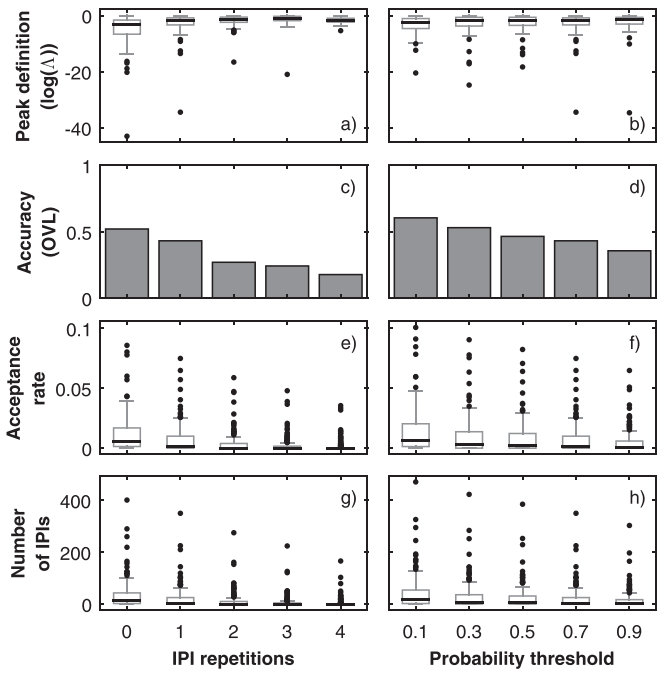

FIG. 2. Performance of automatic IPI compilation for 174 standard recordings from Dominica. The left column shows the effect of variable numbers of IPI repetitions, with the goodness probability threshold held constant at 0.7. The right column shows the effect of variable goodness probability thresholds, with the number of IPI repetitions held constant at 1. log (Λ) = log likelihood ratio between mixture models with shared and unshared variance. OVL = overlapping coefficient between probability density functions of mixture models fit to manually and automatically compiled IPI distributions.

threshold [Fig. 2(b)]. Accuracy, as measured by the amount of overlap between the probability density functions of GMMs fit to manually- and automatically-compiled IPIs, showed a consistent decrease as both filter parameters became more selective [Figs. 2(c) and 2(d)]. Observing the manual and automatic distributions themselves showed fairly good agreement in the detection of peaks, although some peaks in the automatic distribution appeared to be missing [Fig. 3(a)]. The acceptance rate is also much lower with automatic filtration [Fig. 3(b)]. These indicate a high false negative rate.

The proportion of detected clicks that are accepted into the final distributions is always very low, below 0.01 in the overwhelming majority of cases. Not surprisingly, it decreases consistently as filter parameters become more selective, but this is much more pronounced with IPI repetition [Fig. 2(e)] than with probability threshold [Fig. 2(f)]. Acceptance rates of zero are common, and usually represent the majority of cases when IPIs need to be repeated twice or more (depending slightly on probability threshold). However, there was a fair amount of variation in all cases. Though uncommon, it was possible for some distributions produced by the least selective filter to be empty. Likewise, some distributions produced by the most selective filter were larger than average [Figs. 2(g) and 2(h)].

C. Differences between recording scenarios

As expected, the frequency of click detections was much greater for the Galápagos (mean = 1390 clicks/min)

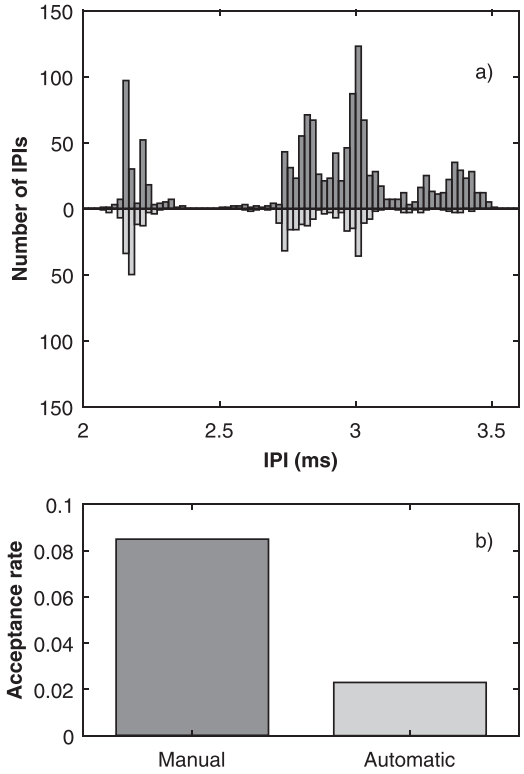

FIG. 3. Automatic IPI compilation compared to manual compilation for nine standard recordings. Automatic compilation for the case shown here required 1 IPI repetition and a goodness probability threshold of 0.7. The dark shade corresponds to manual data. (a) Comparison of IPI distributions, with manual counts on top and automatic counts on the bottom. Bin width = 1/Fs. (b) Click acceptance rates.

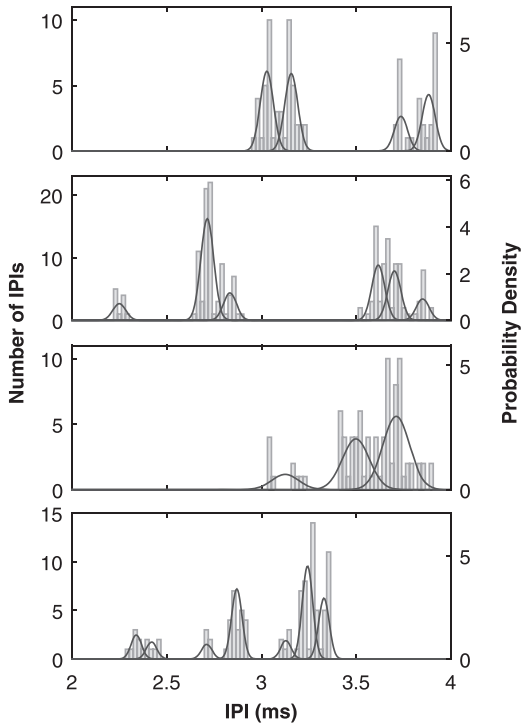

FIG. 4. Example IPI distributions output from four standard recordings from the Galápagos. Filtration parameters were set at 1 IPI repetition and a goodness probability threshold of 0.7. Black lines represent probability density functions of clusters from the best GMMs, according to BIC. Bin width = 1/Fs.

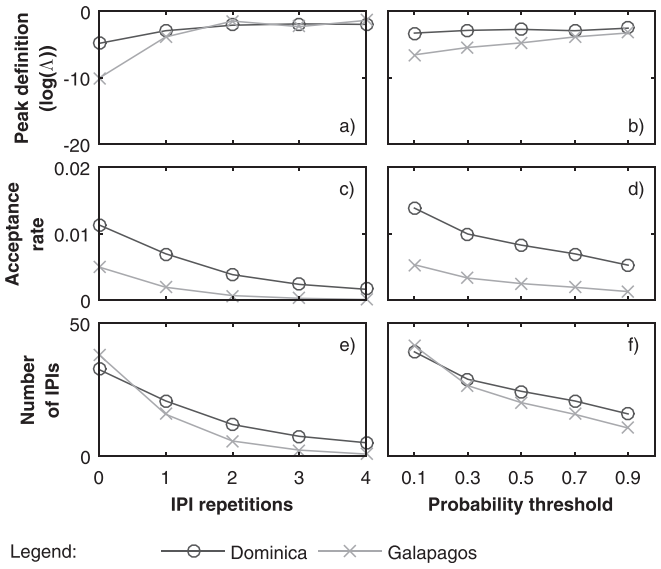

FIG. 5. Performance of automatic IPI compilation compared between 174 standard recordings from Dominica, and 141 standard recordings from the Galápagos. Points represent mean values. The left column shows the effect of variable numbers of IPI repetitions, with the goodness probability threshold held constant at 0.7. The right column shows the effect of variable goodness probability thresholds, with the number of IPI repetitions held constant at 1. log (Λ) = log likelihood ratio between mixture models with shared and unshared variance.

than for Dominica (mean = 489 clicks/min). IPI distributions from the Galápagos were similar to Dominica in that they often contained narrow peaks when enough IPIs were present; however, the density of peaks was generally higher, and individual clusters tended to be more ambiguous (Fig. 4).

Compared to Dominica, Galápagos distributions showed similar trends in peak definition, but were generally of lower quality. On average, the Galápagos peak definition was comparable to Dominica when IPIs needed to be repeated twice or more, but became progressively worse below two repetitions [Fig. 5(a)]. When the probability threshold was varied, Galápagos distributions showed a slight increase in average peak definition, but the level remained inferior to Dominica [Fig. 5(b)]. In summary, it appears that peak definition is generally worse for Galápagos distributions, but it improves at a faster rate than for Dominica as filter parameters become more selective.

Regarding click acceptance rate, Galápagos distributions showed the same decreasing trends as for Dominica with both number of IPI repetitions and probability threshold. However, Galápagos distributions consistently had smaller acceptance rates on average than for Dominica [Figs. 5(c) and 5(d)]. Despite this though, Galápagos distributions contained relatively similar numbers of IPIs as in Dominica [Figs. 5(e) and 5(f)].

IV. DISCUSSION

A. Performance

The automatic IPI compilation algorithm presented here was overall successful in producing reliable IPI distributions straight from single-hydrophone recordings of foraging sperm whales. The SVM was shown to be very effective in distinguishing between on-axis sperm whale echolocation clicks and other click types. The full routine, when applied to approximately 4-min long recordings of sperm whales in Dominica, often produced IPI distributions that contained precise peaks at values that were reasonable for these whales, provided that the IPI filtration parameters were not extreme. Furthermore, the distributions produced by the routine were similar to those obtained through manual click filtration, suggesting that they are similarly reliable. Gaussian mixture modeling appears to be an effective method for detecting IPI peaks automatically. If several candidate models are available for a given distribution, the model with the smallest BIC should usually be the most plausible, but this may not always be the case: for example, in the top distribution of Fig. 1, a 2-cluster model may have been more appropriate than the 3-cluster one selected. Therefore, alternative models with slightly less BIC support should not be discounted.

IPI repetition was shown to be a very influential parameter. When IPIs do not need to be repeated, the resulting distributions are likely to contain more IPIs, but in many cases they are unsuitable for analysis. This is evident from the peak definition measure. Most distributions produced without requiring IPI repetition had a distinctively noisy appearance, so it is likely that this noise is responsible for poor peak definition. GMMs, as implemented here, attempt to cluster every sample, so outliers can be problematic. Outliers may be grouped into very small, possibly single-sample clusters, or they may be included with other samples to form wide clusters. In either case, models with shared variance usually do not fit well, because the erroneous clusters are likely to have large differences in variance.

Comparing automatic IPI distributions with manually compiled ones showed that the automatic method becomes less accurate as filter parameters become more selective. This may seem surprising, but it is easily explained. Increasing filter selectiveness results in fewer false positives, but this comes at the cost of more false negatives, or misses. The rate at which the number of misses increases is much greater than the rate at which the number of false positives decreases, which causes overall accuracy to decrease. A consequence of this imbalance is that it is not necessarily desirable to attain maximum accuracy. As shown by peak definition, a modest number of false positives can make it difficult to analyze an IPI distribution. In contrast, false negatives are a nuisance, but they do not complicate analysis to the same degree as false positives. Therefore, a balance needs to be found between the two, with greater weight placed on reducing false positives. In light of this, disabling IPI repetition checks is still not a good option, even though this yields the highest overall accuracy.

This brings up the greatest weakness of automatic IPI compilation: acceptance rate. For all recordings, even the least selective filtration criteria resulted in very small distributions, relative to the total number of clicks that were detected. To a certain extent, this is expected, given the rarity of clicks with clear multi-pulse structures. Essentially, individual clicks are only suitable for IPI calculation if three criteria are met: (1) the hydrophone must be aligned with the whale's acoustic axis; (2) the click must not coincide with other clicks or echoes; and (3) the click must be significantly louder than background noise. Clearly, the probability that all of these conditions will be true for any click is small, particularly with far-field PAM recordings. Some recordings may be more likely to meet them than others, depending on factors such as noise level, distance from the whales, number of whales, and reflective profile of the environment. Whale orientation, however, is a more random factor, and some recordings may simply be more fortunate than others in the amount of time that whales are aligned with the hydrophone; this would explain why acceptance rate is so variable. However, the rarity of on-axis clicks alone does not explain the routine's particularly low acceptance rate. Recall that there is a large discrepancy in acceptance rate between manually and automatically compiled IPI distributions [Fig. 3(b)], which reflects a highly aggressive filter.

While it is not ideal to reject so many positives, this should not be debilitating in practice. The main reason for this is because of the high click rates of sperm whales: 1.2 clicks/s while foraging according to Whitehead and Weilgart (1990). Thus, good IPI distributions can successfully be obtained from just a few minutes of recording, as is evident from many of the distributions obtained here (e.g., Fig. 1). Another reason is that it does not take many IPIs to resolve peaks. Since there are few false positives, each IPI in a distribution is very likely to be a true one, at least when each IPI is required to be repeated at least once. Thus, automatic IPI distributions are likely to be reliable even when clusters contain relatively few samples, in that those clusters likely represent the IPIs of some whales in the area. Click acceptance rate can also be improved with simple techniques. For instance, maintaining distance between the hydrophone and reflective surfaces (notably the sea surface) should help, since direct-path clicks will be less likely to overlap with their echoes. However, it is important to remember that click acceptance rate is inherently variable, due to its dependence on whale orientation.

On the software end, the only way to substantially improve acceptance rate, aside from relaxing filtration criteria, is by improving the classifier. With an estimated sensitivity of 89.1%, the SVM is quite good at recognizing on-axis clicks, but this is perhaps not enough. Since each IPI must be repeated to be valid, any on-axis clicks missed by the SVM can further invalidate surrounding clicks by creating gaps in repetition chains. This explains why acceptance rate decreases rapidly as the number of required repetitions increases. Thus, improving the SVM's sensitivity would greatly reduce this problem. However, it should not be done with detriment to specificity, otherwise peak definition may decrease. This may seem difficult, but it should be possible. One factor that likely contributes to classifier confusion is binary classification. This is a problem for two reasons. First, there is no hard separation between on- and off-axis clicks. Ideas for solving this problem include fuzzy labelling (i.e., weighting) of training instances, or using a semi-supervised learning approach where "Unsure" clicks are included as unlabeled instances (Schwenker and Trentin, 2014). For the present classifier, such clicks were simply removed from the dataset (see Appendix A in the supplementary material[1]). The second problem is that there are several click types which may exhibit features that overlap with those of the targeted type. For example, on-axis coda clicks and clear surface reflections could share some features with direct-path on-axis echolocation clicks. This can be addressed by using more than two classes. In this case, it might work best if classification is done hierarchically, where clicks are given multiple labels. For example, each click could be classified as being a usual, coda, or other click type, being on-axis or not, and being a direct-path or reflected click. This would be especially useful for studies that focus on codas (e.g., Pavan *et al.*, 2000; Rendell and Whitehead, 2004; Marcoux *et al.*, 2006; Schulz *et al.*, 2008; Antunes *et al.*, 2011; Schulz *et al.*, 2011; Gero *et al.*, 2016b).

Another factor that could contribute to a low acceptance rate is if the dataset used to train the SVM does not fully capture the complete range of possibilities. The Good clicks in the training dataset consisted mainly of clicks from whales that started a foraging dive, in the first few minutes of their dive. However, some features, notably spectral ones, are known to change with depth (Thode *et al.*, 2002). Thus, the SVM might potentially have difficulty recognizing clicks from deeper whales. Individual variation in click features might also cause some difficulty, as there were only ten diving whales in the training dataset. However, the high accuracy reported by cross-validation suggests that this is a relatively minor problem.

Using the current classifier, checking for one repetition is likely the best trade-off in most cases, at least for Dominica surface recordings. If many IPIs are available, the number of repetitions may be increased to further improve

the quality of the distribution. In contrast, if few IPIs are available, one technique to improve the acceptance rate might be to disable IPI repetition, and then ignore potentially erroneous IPIs (i.e., very small clusters). This would be possible, for example, by applying mixture modeling techniques that are robust to outliers (McNicholas, 2016). Of course though, these distributions would likely not be as precise as when IPI repetition is enforced.

As evident from Fig. 2, the "goodness" probability threshold does not impact peak definition or acceptance rate as drastically as IPI repetition (between 0.1 and 0.9 at least). This might be a consequence of the binary nature of the SVM, and also the fact that ambiguous clicks were not used to train it. Nevertheless, a value of 0.7 is recommended as a default.

1. Differences between recording scenarios

Compared to Dominica, IPI distributions from the Galápagos generally had more peaks, which were often closely spaced and more ambiguous to interpret. This fits with groups being considerably larger off the Galápagos (Whitehead *et al.*, 2012). As for the generally poorer peak definition and lower acceptance rate, the most likely explanation is that clicks recorded off the Galápagos were of poorer quality overall, in the sense that few of them had clear multi-pulsed structures. More poor quality clicks would necessarily result in a lower acceptance rate. Peak definition could also be impacted, due to a higher number of false positives: if more poor-quality clicks are present, the SVM has more opportunities to misclassify Bad clicks as Good. A likely reason why the Galápagos might have poorer clicks is because of its higher click density: when more whales are clicking together, the clicks have a higher chance of overlapping with one another, resulting in a greater proportion of unusable clicks.

Another explanation for the apparent inferiority of Galápagos IPI distributions could be that the SVM does not recognize on-axis clicks from the Galápagos as easily as it does for Dominica. This could happen if the distribution of classifying features differs somehow between Dominica and Galápagos clicks. Such a difference might occur if, for example, sound does not propagate the same way between regions, or the recording setup differed in some way that was not identified. Another plausible cause is that the "voices" of the whales encountered are significantly different between regions. Noise is not likely a factor, since the recordings analyzed here did not differ significantly between regions in this regard.

Ideally, to deal with potential differences between recording scenarios, the classifier should be trained using clicks from each scenario. Unfortunately, this is a time-consuming procedure that must be done by someone who is skilled at recognizing on-axis sperm whale clicks. A simpler but less effective workaround is to try adjusting the IPI filtration parameters. Parameter values could be increased if there appear to be many false positives, for example, or perhaps decreased if the number of accepted IPIs is overwhelmingly low.

B. Note on IPI variability

Recent work by Bøttcher *et al.* (2018) concluded that IPI estimates from individual sperm whales are not consistent between recordings and may change significantly with depth. Their results showed that IPI distributions from individual whales frequently appear bimodal overall, with a spread of about 0.2 ms. This seems inconsistent with the high IPI precision reported in the previous literature (Schulz *et al.*, 2011; Antunes *et al.*, 2010; Rhinelander and Dawson, 2004). The reason for this is likely because the IPI compilation approach used by Bøttcher *et al.* (2018) was fundamentally different. Specifically, Bøttcher *et al.* (2018) performed very little filtration of individual clicks or IPIs. While the recordings used in that study were obtained such that the hydrophone was mostly aligned with the target whale's acoustic axis, such recordings still include a considerable number of Bad clicks, as defined here. Thus, the IPI distributions presented by Bøttcher *et al.* (2018) necessarily had greater variability. It is possible that the bimodal patterns they observed were influenced not only by depth, but also by systematic bias in IPI estimates due to slightly off-axis clicks recorded in succession. Such a bias is not expected in the results presented here, since only clear on-axis clicks were retained.

IPI variation due to depth change has been hypothesized before (Goold, 1996) and is not unexpected. However, this should have minimal impact on our routine in most cases. A typical sperm whale foraging dive consists of a short descent phase during which echolocation begins, a long foraging phase at relatively consistent depth, and a short ascent phase (Watwood *et al.*, 2006). Consequently, the vast majority of clicks are produced, and therefore recorded, during the bottom phase, where depth change is limited. Furthermore, any clicks recorded from the descent phase would most likely be off-axis, unless deliberately recording in the slick of a diving whale. Nevertheless, if IPI sample sizes are very low, it may be possible to obtain multimodal distributions from individual whales as they change depth. Therefore, extra care should be taken when interpreting mixture models that have clusters containing few samples in close proximity to one another (e.g., centroids within 0.2 ms). Determining a reliable minimum number of IPIs per cluster may require further research on IPI variability in Good clicks only.

C. Applications

The ability to measure IPIs automatically should be a great addition to sperm whale PAM. One of the primary goals of marine mammal passive acoustic surveys is abundance estimation, since this is essential for ecosystem and management studies (Mellinger *et al.*, 2007). Abundance estimates depend on the number of animals detected, which can be difficult to obtain through acoustics alone. For sperm whales, counting the number of peaks in IPI distributions could be one way of doing this, with the caveat that only whales of different size will be detected. If similar-sized whales are present, then this count would represent an underestimate, unless additional information is available (e.g., bearing or location). IPIs would also provide information

that is usually impossible to get from standard passive acoustic surveys, notably the size of each animal, and to a certain extent, sex (mature males can be identified). This important information must usually be obtained from visual surveys, which are expensive and prone to limitations such as weather and time of day. Thus, through IPIs, acoustic information could be used to compare length and sex distributions between areas, seasons, and different time periods, as well as between social units (Best, 1979; Whitehead *et al.*, 1991) and clans (Rendell and Whitehead, 2003; Gero *et al.*, 2016a).

Since IPI is variable between individuals this measure could also be used to some extent to track the movements of individual whales or social units. For example, if multiple sensors with the ability to determine IPIs are deployed in an area, IPI "hits" could be compared between sensors over time. If a particular IPI peak is detected at some hydrophone X, and again later at another hydrophone Y, then one could infer that the same whale has traveled from X to Y. Of course, this kind of IPI-based telemetry would be limited by the number of whales in an area that have similar IPIs. It could be particularly useful, though, in areas where whales travel in social units with stable memberships. In this case, the signature of a unit would be a set of IPI peaks. These peak distributions might contain a fair amount of information that could be used to discriminate units with some confidence.

V. CONCLUSIONS

On-axis sperm whale clicks can quite accurately be recognized by an automatic classifier. From this, an algorithm capable of automatically compiling and analyzing reliable sperm whale IPI distributions directly from acoustic recordings has been developed. The method works with only one audio channel, and does not require knowledge of how many whales are present, or how they are oriented with respect to the hydrophone. Examination of the output IPI distributions shows that they often contain clear peaks, and are comparable with manually compiled IPIs. However, the method rejects many more clicks than expected by manual compilation. Fortunately, given the high click rates of sperm whales, even relatively short recordings are likely to yield enough IPIs to produce clear distributions, although the actual number of measures will depend heavily on whale orientation. Based on the current implementation, filtration parameters may need to be adjusted to accommodate different recording scenarios. In the long term, expansion of the classifier's training dataset with a wider variety of clicks (e.g., from mature males, bottom-mounted hydrophones, new regions, etc.) may enable it to perform better under a wider variety of scenarios. Modifications to the classification model could also potentially improve acceptance rate.

The software should be a useful extension to sperm whale PAM. The ability to obtain IPIs, and consequently body length estimates from sperm whales without the need to tag or even see them, should be a great advantage for studying their abundance, movements, and behaviour.

ACKNOWLEDGMENTS

Many thanks go to the crews of *Balaena* during the Dominica 2015 and Galápagos 2014 field seasons for data collection, to Luke Rendell for tips on data and performance analyses, to Charlotte Dunn for providing bug reports and feedback on the software, to Stan Matwin and Erico Neves De Souza for thoughts on classification and machine learning, to Gabrielle Macklin and Marie Ryan for their volunteer work during early conception of the click classifier, and to Kristian Beedholm for sharing click detection source code during preliminary software development. Brian Miller, Alex Hay, Luke Rendell, and two anonymous reviewers provided valuable comments on earlier drafts of the manuscript. The fieldwork in Dominica was performed under scientific research permits from the Fisheries Division of the Ministry of Agriculture and Fisheries and funded through grants to S.G., including an expedition grant from the Carlsberg Foundation, an FNU fellowship for the Danish Council for Independent Research supplemented by a Sapere Aude Research Talent Award, and supplementary funding from a FNU Large Frame Grant to Peter Madsen. The fieldwork in the Galápagos was performed under permits from the Ministerio de Defensa Nacional, the Ministerio del Ambiente, and the Dirección del Parque Nacional Galápagos. S.G. was supported by a technical and scientific research grant from the Villum Foundation. H.W. received funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery and Equipment grants, the National Geographic Society, the International Whaling Commission, the Whale and Dolphin Conservation Society, Cetacean Society International, and the Green Island Foundation. W.A.M.B. was supported by a NSERC Canada Graduate Scholarship, the Nova Scotia Graduate Scholarship, and the Dr. Patrick Lett Fund. This study emanates from The Dominica Sperm Whale Project: http://www.thespermwhaleproject.org Follow: @DomWhale.

[1]See supplementary material at https://doi.org/10.1121/1.5082291 for: Appendix A, which contains some additional information on the methodology; for Appendix B, which contains a detailed graphical representation of the entire algorithm; and for Appendix C, which describes all modifiable parameters of the routine.

Adler-Fenchel, H. S. (**1980**). "Acoustically derived estimate of the size distribution for a sample of sperm whales (*Physeter catodon*) in the Western North Atlantic," Can. J. Fish. Aquat. Sci. **37**, 2358–2361.

Antunes, R., Rendell, L., and Gordon, J. (**2010**). "Measuring inter-pulse intervals in sperm whale clicks: Consistency of automatic estimation methods," J. Acoust. Soc. Am. **127**, 3239–3247.

Antunes, R., Schulz, T., Gero, S., Whitehead, H., Gordon, J., and Rendell, L. (**2011**). "Individually distinctive acoustic features in sperm whale codas," Anim. Behav. **81**, 723–730.

Arnbom, T. (**1987**). "Individual identification of sperm whales," Rep. Int. Whaling Comm. **37**, 201–204.

Backus, R. H., and Schevill, W. E. (**1966**). "*Physeter* clicks," in *Whales, Dolphins, and Porpoises*, edited by K. S. Norris (University of California Press, Berkeley, CA), pp. 510–527.

Barlow, J., and Taylor, B. L. (**2005**). "Estimates of sperm whale abundance in the northeastern temperate pacific from a combined acoustic and visual survey," Mar. Mammal Sci. **21**, 429–445.

Best, P. B. (**1979**). "Social organization in sperm whales, *Physeter macrocephalus*," in *Behaviour of Marine Animals*, edited by H. E. Winn and B. L. Olla (Plenum Press, New York), Vol. 3, pp. 227–290.

Bøttcher, A., Gero, S., Beedholm, K., Whitehead, H., and Madsen, P. T. (**2018**). "Variability of the inter-pulse interval in sperm whale clicks with

implications for size estimation and individual identification," J. Acoust. Soc. Am. **144**, 365–374.
Clarke, M. R. (**1978**). "Structure and proportions of the spermaceti organ in the sperm whale," J. Mar. Biol. Assoc. U.K. **58**, 1–17.
Drouot, V., Gannier, A., and Goold, J. C. (**2004**). "Diving and feeding behavior of sperm whales (*Physeter macrocephalus*) in the northwestern Mediterranean Sea," Aquatic Mammals **30**, 419–426.
Gero, S., Bøttcher, A., Whitehead, H., and Madsen, P. T. (**2016a**). "Socially segregated, sympatric sperm whale clans in the Atlantic Ocean," Royal Soc. Open Sci. **3**, 160061.
Gero, S., Milligan, M., Rinaldi, C., Francis, P., Gordon, J., Carlson, C., Steffen, A., Tyack, P., Evans, P., and Whitehead, H. (**2014**). "Behavior and social structure of the sperm whales of Dominica, West Indies," Mar. Mammal Sci. **30**, 905–922.
Gero, S., Whitehead, H., and Rendell, L. (**2016b**). "Individual, unit and vocal clan level identity cues in sperm whale codas," Royal Soc. Open Sci. **3**, 150372.
Gillespie, D., Mellinger, D. K., Gordon, J., McLaren, D., Redmond, P., McHugh, R., Trinder, P. W., Deng, X. Y., and Thode, A. (**2008**). "PAMGUARD: Semiautomated, open-source software for real-time acoustic detection and localisation of cetaceans," in *Proceedings of the Institute of Acoustics*, Vol. 30, Pt. 5.
Goold, J. C. (**1996**). "Signal processing techniques for acoustic measurement of sperm whale body lengths," J. Acoust. Soc. Am. **100**, 3431–3441.
Gordon, J. C. D. (**1991**). "Evaluation of a method for determining the length of sperm whales (*Physeter catodon*) from their vocalizations," J. Zool. Lond. **224**, 301–314.
Growcott, A., Miller, B., Sirguey, P., Slooten, E., and Dawson, S. (**2011**). "Measuring body length of male sperm whales from their clicks: The relationship between inter-pulse intervals and photogrammetrically measured lengths," J. Acoust. Soc. Am. **130**, 568–573.
Madsen, P. T., Payne, R., Kristiansen, N. U., Wahlberg, M., Kerr, I., and Møhl, B. (**2002**). "Sperm whale sound production studied with ultrasound time/depth-recording tags," J. Exp. Biol. **205**, 1899–1906.
Marcoux, M., Whitehead, H., and Rendell, L. (**2006**). "Coda vocalizations recorded in breeding areas are almost entirely produced by mature female sperm whales (*Physeter macrocephalus*)," Can. J. Zool. **84**, 609–614.
McNicholas, P. D. (**2016**). "Model-based clustering," J. Classif. **33**, 331–373.
Mellinger, D. K., Stafford, K. M., Moore, S. E., Dziak, R. P., and Matsumoto, H. (**2007**). "An overview of fixed passive acoustic observation methods for cetaceans," Oceanography **20**, 36–45.
Miller, B. S. (**2010**). "Acoustically derived growth rates and three-dimensional localisation of sperm whales (*Physeter macrocephalus*) in Kaikoura, New Zealand," Ph.D. thesis, University of Otago, New Zealand.
Møhl, B. (**2001**). "Sound transmission in the nose of the sperm whale *Physeter catodon*. A post mortem study," J. Comp. Physiol. A **187**, 335–340.
Møhl, B., Larsen, E., and Amundin, M. (**1981**). "Sperm whale size determination: Outlines of an acoustic approach," FAO Fisheries Ser. **5**, 327–332.
Møhl, B., Wahlberg, M., Madsen, P. T., Heerfordt, A., and Lund, A. (**2003**). "The monopulsed nature of sperm whale clicks," J. Acoust. Soc. Am. **114**, 1143–1154.
Nishiwaki, N., Oshumi, S., and Maeda, Y. (**1963**). "Changes in form of the sperm whale accompanied with growth," Sci. Rep. Wh. Res. Inst. Tokyo **17**, 1–13.
Norris, K. S., and Harvey, G. W. (**1972**). "A theory for the function of the spermaceti organ of the sperm whale (*Physter catodon L.*)," in *Animal Orientation and Navigation*, edited by S. R. Galler, K. Schmidt-Koenig, G. J. Jacobs, and R. E. Belleville, SP-262 (NASA, Washington, DC), pp. 397–417.
Page, S. E. (**1954**). "Continuous inspection schemes," Biometrika **41**, 100–115.
Pavan, G., Hayward, T. J., Borsani, J. F., Priano, M., Manghi, M., Fossati, C., and Gordon, J. (**2000**). "Time patterns of sperm whale codas recorded in the Mediterranean Sea 1985–1996," J. Acoust. Soc. Am. **107**, 3487–3495.
Platt, J. (**1999**). "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers* (MIT Press, Cambridge, MA), pp. 61–74.
Rendell, L., and Whitehead, H. (**2003**). "Vocal clans in sperm whales (*Physeter macrocephalus*)," Proc. Royal Soc. London B **270**, 225–231.
Rendell, L., and Whitehead, H. (**2004**). "Do sperm whales share coda vocalizations? Insights into coda usage from acoustic size measurements," Anim. Behav. **67**, 865–874.
Rhinelander, M. Q., and Dawson, S. M. (**2004**). "Measuring sperm whales from their clicks: Stability of interpulse intervals and validation that they indicate whale length," J. Acoust. Soc. Am. **115**, 1826–1831.
Schulz, T. M., Whitehead, H., Gero, S., and Rendell, L. (**2008**). "Overlapping and matching of codas in vocal interactions between sperm whales: Insights into communication function," Anim. Behav. **76**, 1977–1988.
Schulz, T. M., Whitehead, H., Gero, S., and Rendell, L. (**2011**). "Individual vocal production in a sperm whale (*Physeter macrocephalus*) social unit," Mar. Mammal Sci. **27**, 149–166.
Schwenker, F., and Trentin, E. (**2014**). "Pattern classification and clustering: A review of partially supervised learning approaches," Pattern Recogn. Lett. **37**, 4–14.
Teloni, V., Zimmer, W. M. X., Wahlberg, M., and Madsen, P. (**2007**). "Consistent acoustic size estimation of sperm whales using clicks recorded from unknown aspects," J. Cetacean Res. Manage. **9**, 127–136.
Thode, A., Mellinger, D. K., Stienessen, S., Martinez, A., and Mullin, K. (**2002**). "Depth-dependent acoustic features of diving sperm whales (*Physeter macrocephalus*) in the Gulf of Mexico," J. Acoust. Soc. Am. **112**, 308–321.
Thomas, J. A., Fisher, S. R., and Awbrey, F. A. (**1986**). "Use of acoustic techniques in studying whale behavior," Rep. Int. Whaling Commun. (Special Issue) **8**, 121–138.
Tønnesen, P., Gero, S., Ladegaard, M., Johnson, M., and Madsen, P. (**2018**). "First-year sperm whale calves echolocate and perform long, deep dives," Behav. Ecol. Sociobiol. **72**, 165–179.
Watkins, W. A., and Schevill, W. E. (**1977**). "Sperm whale codas," J. Acoust. Soc. Am. **62**, 1485–1490.
Watwood, S. L., Miller, P. J. O., Johnson, M., Madsen, P. T., and Tyack, P. L. (**2006**). "Deep-diving foraging behavior of sperm whales (*Physeter macrocephalus*)," J. Anim. Ecol. **75**, 814–825.
Weilgart, L., and Whitehead, H. (**1993**). "Coda communication by sperm whales (*Physeter macrocephalus*) off the Galápagos Islands," Can. J. Zool. **71**, 744–752.
Whitehead, H., Antunes, R., Gero, S., Wong, S. N. P., Engelhaupt, D., and Rendell, L. (**2012**). "Multilevel societies of female sperm whales (*Physeter macrocephalus*) in the Atlantic and Pacific: Why are they so different?," Int. J. Primatol. **33**, 1142–1164.
Whitehead, H., Waters, S., and Lyrholm, T. (**1991**). "Social organization of female sperm whales and their offspring: Constant companions and casual acquaintances," Behav. Ecol. Sociobiol. **29**, 385–389.
Whitehead, H., and Weilgart, W. (**1990**). "Click rates from sperm whales," J. Acoust. Soc. Am. **87**, 1798–1806.
Zimmer, W. M. X. (**2011**). *Passive Acoustic Monitoring of Cetaceans* (Cambridge University Press, Cambridge).
Zimmer, W. M. X., Madsen, P. T., Teloni, V., Johnson, M. P., and Tyack, P. L. (**2005**). "Off-axis effects on the multipulse structure of sperm whale usual clicks with implications for sound production," J. Acoust. Soc. Am. **118**, 3337–3345.