Automated extraction of odontocete whistle contours

Marie A. Roch[a)]
San Diego State University, Department of Computer Science, 5500 Campanile Drive, San Diego, California 92182-7720

T. Scott Brandes
Signal Innovations Group, Incorporated, 4721 Emperor Boulevard, Suite 330, Research Triangle Park, North Carolina 27703

Bhavesh Patel
San Diego State University, Department of Computer Science, 5500 Campanile Drive, San Diego, California 92182-7720

Yvonne Barkley
Southwest Fisheries Science Center, National Oceanic and Atmospheric Administration, 3333 North Torrey Pines Court, La Jolla, California 92037

Simone Baumann-Pickering
Scripps Institution of Oceanography, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0205

Melissa S. Soldevilla
Duke University Marine Laboratory, 135 Duke Marine Lab Road, Beaufort, North Carolina 28516

(Received 7 March 2011; revised 11 July 2011; accepted 25 July 2011)

Many odontocetes produce frequency modulated tonal calls known as whistles. The ability to automatically determine time × frequency tracks corresponding to these vocalizations has numerous applications including species description, identification, and density estimation. This work develops and compares two algorithms on a common corpus of nearly one hour of data collected in the Southern California Bight and at Palmyra Atoll. The corpus contains over 3000 whistles from bottlenose dolphins, long- and short-beaked common dolphins, spinner dolphins, and melon-headed whales that have been annotated by a human, and released to the Moby Sound archive. Both algorithms use a common signal processing front end to determine time × frequency peaks from a spectrogram. In the first method, a particle filter performs Bayesian filtering, estimating the contour from the noisy spectral peaks. The second method uses an adaptive polynomial prediction to connect peaks into a graph, merging graphs when they cross. Whistle contours are extracted from graphs using information from both sides of crossings. The particle filter was able to retrieve 71.5% (recall) of the human annotated tonals with 60.8% of the detections being valid (precision). The graph algorithm's recall rate was 80.0% with a precision of 76.9%.
 [DOI: 10.1121/1.3624821]

PACS number(s): 43.80.Cs, 43.60.Uv [WA] Pages: 2212–2223

I. INTRODUCTION

The identification and description of individual marine mammal tonal calls is a task that has numerous applications. Contour description is useful for describing species' vocal repertoires (i.e., Wang *et al*., 1995) or can be a preliminary step in a signal processing chain to determine information about which species generated a set of whistles (Oswald *et al*., 2007). Population densities can be estimated from call detections when the call production rate is known (Marques *et al*., 2009), and finally identifying a call recorded at multiple hydrophones can be used to solve the call correspondence task in localization applications.

Over the last two decades, several groups have worked on methods to automate the description of whistles. Whistles are tonal calls produced by many species of odontocetes. Some of these are semi-automated, such as the approaches used by Buck and Tyack (1993) and Lammers *et al*. (2003), where a user is required to provide information to assist the algorithm such as starting and ending points of the whistles. Most fully automated algorithms, including the techniques described in this work, focus on extracting contour ridges from time × frequency representations of a signal. The time × frequency representation typically consists of a spectrogram computed from sequences of Fourier transforms of overlapping windowed audio data.

Examples of this type of algorithm include the work of Datta and Sturtivant (2002) which uses edge-detection techniques from image processing to find and connect areas of the spectrogram with sharp transitions in intensity. Other

[a)]Author to whom correspondence should be addressed. Electronic mail: marie.roch@sdsu.edu

 0001-4966/2011/130(4)/2212/12/$30.00

strategies include identifying peaks in the spectra and connecting them in a coherent manner. For algorithms that identify whistles from time × frequency peaks, the challenge lies in determining when peaks are in close enough proximity to be part of the same whistle contour and being able to disambiguate individual whistles when they cross one another. Halkias and Ellis (2006) detected short segments of whistles and then decided whether or not to connect nearby segments based upon likelihood models trained on the mean frequency curvature and energy. Pulsed killer whale calls appear to have a harmonic structure when analyzed with long windows relative to the interpulse interval [see Watkins (1967) for a discussion of this relationship]. Brown and Miller (2007) as well as Shapiro and Wang (2009) have exploited this structure to determine call tracks.

An alternative approach is to consider whistle discovery within the context of Bayesian filtering, with the time-varying whistle frequency serving as a hidden state that is estimated from the noisy sound field. Mallawaarachchi *et al*. (2008) used Kalman filters, a closed-form solution of the Bayesian filtering problem, to adaptively predict the next time × frequency peak along a whistle path, thus solving the disambiguation problem by retaining state information about each whistle. Other approaches that use predictions based on partial detections include the whistle detectors in Ishmael (Mellinger, 2001) and PamGuard (Gillespie *et al*., 2008).

Some groups have proposed techniques that are not based on sequences of short-time Fourier transforms. Adam (2008) extracted calls of killer whales using the Hilbert-Huang transform. Ioana *et al*. (2010) proposed a method to extract tonals when the signal's phase track can be approximated by a polynomial. The strongest signal is estimated by the product high-order ambiguity function (Barbarossa *et al*., 1998), subtracted, and the process is iterated to find the next strongest signal.

In this work, we consider two methods for the automatic annotation of whistles. The first uses a particle filter, which overcomes several shortcomings of Kalman filters which constrain the probability distributions to Gaussians and only permits linear state update equations. The second method uses the formalism of a graph to connect spectral peaks and permits delayed decisions about which paths are associated with which whistle (if any) until more information than simply the next detected peak is available.

Kalman filters become limited in real-world scenarios and have difficulty when distributions are non-Gaussian, such as when distributions become multimodal at contour intersections or when background noise increases suddenly. As a robust alternative to Kalman filters in more complex environments, particle filters provide a sequential Monte Carlo solution for Bayesian filtering that works in nonlinear and non-Gaussian settings (Doucet *et al*., 2001, pp. 3–14; Arulampalam *et al*., 2002). In previous work, particle filters have been used in formant tracking for human speech (Shi and Chang 2003); however, extensions are needed for detecting odontocete whistles since their approach requires that formants remain uninterrupted by other sounds or formants. White and Hadley (2008) showed that particle filters have the potential for use in cetacean whistle extraction by showing that a simple particle filter can extract a single short whistle. In the work presented here we extend their approach by accommodating a more sophisticated particle filter specifically designed for odontocete whistle extraction in a complex acoustic environment with numerous overlapping whistles from multiple individuals.

As an alternative approach, we also consider the construction of graph representations of whistle networks. Graphs are commonly used to represent the interconnections between nodes and have applications in search, path-finding, etc. (Nilsson, 1980, Chap. 2). Spectral peaks in the time × frequency space are examined and either appended to existing graphs or form new graphs. Graphs may contain multiple whistles, and a disambiguation step analyzes each graph to extract the multiple whistles that may lie within.

This work uses a common signal processing front-end to compare the results of particle-filter and graph based methods for detecting whistle contours. Nearly one hour of recordings has been hand-annotated for five different species of odontocetes, and metrics have been defined to determine the efficacy of the algorithms not only for retrieval and false detections, but also to characterize the quality of the detections. Both techniques are capable of real-time whistle extraction on current-generation workstations such as the Phenom™ II X4 940 (Advanced Micro Devices, Sunnydale, CA), or Xeon™ X3360 (Intel, Santa Clara, CA) with multiple gigabytes of RAM.

II. METHODS

A. Data collection

Data sampled at 192 kHz with 16 or 24 bit quantization were collected for five species of odontocetes. Calls from short-beaked and long-beaked common dolphins (*Delphinis delphis* and *D. capensis*, respectively), as well as bottlenose dolphins (*Tursiops truncatus*) were collected in the Southern California Bight between 2004 and 2006. Additional bottlenose dolphin recordings along with recordings from melon-headed whales (*Peponocephala electra*) and spinner dolphins (*Stenella longirostris longirostris*) were collected during 2006 and 2007 at Palmyra Atoll. Two types of hydrophones were used, the ITC 1042 (Intl. Transducer Corp., Santa Barbara, CA), and the HS150 (Sonar Research and Development Ltd., Beverly, UK), both of which have flat frequency responses (±3 dB) between 1–100 kHz. Hydrophones were dipped or towed from small boats, the stationary platform R/P FLIP (Fisher and Spiess 1963), and the R/V David Starr Jordan. Hydrophone depths were typically 10 to 30 m.

Trained visual observers confirmed the identity of each species. Recordings were made only in the presence of single-species groups when no other groups were sighted. Limitations of the data collection include differences in the ability to sight other species due to observation platform height as well as similarities between long-beaked and short-beaked common dolphins (Heyning and Perrin, 1994) that make identification of these species more difficult. The sightings, recording durations, and specific files of the 56 m 39 s subset of the data used in this study are reported in Tables I and II.

TABLE I. Summary of recordings. Abbreviations: CalCOFI—California Cooperative Oceanic Fisheries Investigations oceanographic survey, SCI—San Clemente Island small boat survey, SOCAL—SOuthern CALifornia Instrumentation cruises on the R/V Sproul, FLIP—R/P FLIP moored recordings, and Palmyra—Palmyra Atoll small boat recordings.

	Sighting						
	1		2		3		
Species	Duration	Expedition	Duration	Expedition	Duration	Expedition	Total duration
Bottlenose dolphin	4 m 13 s	SCI	6 m 39 s	Palmyra			10 m 52 s
Long-beaked common dolphin	5 m 0 s	CalCOFI	3 m 54 s	SOCAL	5 m 0 s	FLIP	13 m 54 s
Melon-headed whale	6 m 57 s	Palmyra	1 m 5 s	Palmyra	3 m 18 s	Palmyra	11 m 20 s
Short-beaked common dolphin	2 m 30 s	SCI	6 m 11 s	SCI	4 m 47 s	SCI	13 m 28 s
Spinner dolphin	2 m 23 s	Palmyra	2 m 5 s	Palmyra	2 m 37 s	Palmyra	7 m 5 s
						grand total	56 m 39 s

B. Signal processing front-end

Spectrograms are formed from the log magnitude spectra of Hamming-windowed data frames computed every Δ_t ms. In this work we use a frame length of 8 ms and a frame advance of $\Delta_t = 2$ ms, resulting in a per frequency bin bandwidth of $\Delta_f = 125$ Hz. Only frequency bins between 5 and 50 kHz are processed as most calls of the species of interest lie within this range. Spectrograms are smoothed by using a median filter over a 3×3 time-frequency grid followed by a per frequency bin spectral means subtraction over a 3 s window.

Spectral peaks are identified for each frame by noting all frequency bins whose normalized magnitude exceeds 10 dB rel. and have no higher-energy neighbors within 250 Hz ($\pm 2\Delta_f$), which is roughly where the first side band of a Hamming window occurs for the 8 ms window used by the algorithm. Part of the motivation for suppressing close peaks is to prevent detections of peaks from echoes with very short delay which would have a similar trajectory such as those that might occur when an animal is near the surface. Regions of broad band energy, such as those produced by impulsive sounds such as snapping shrimp or echolocation clicks, are detected by checking to see if the percentage of frequency bins identified as peaks has increased dramatically from the previous frame. When this increases by more than 1%, we consider it unlikely that the new peaks are attributable to the start of new whistles, and the frame is not processed. Subsequent frames use the number of peaks from the last accepted frame when determining the percentage increase in peaks, and the algorithm initializes this value to 5% of the frequency bins at the start of processing. The thresholds for this ad hoc method will be shown to produce good results for the species in this study, and would need to be adjusted for any species that started to chorus in large numbers within Δt s. The specified analysis and growth rate parameters result in the admission of up to eighteen calls in the first analysis frame and up to three new calls within any 2 ms period.

C. Whistle extraction

We compare two competing methods of determining whistle (tonal) contour patterns from the detected peaks. The first method employs particle filters to model the trajectory of hypothesized peaks and incrementally builds candidate tonal contours. The second method assembles peaks that meet criteria into a graph representation. No attempt is made

TABLE II. Audio files corresponding to the summary data of Table I. Files are publicly available in the Moby Sound archive as part of the 2011 Detection, Classification, and Localization of Marine Mammals Using Passive Acoustic Monitoring conference dataset.

Species	Sighting	File(s)
Bottlenose dolphin	1	Qx-Tt-SCI0608-N1-060814-121518.wav
	2	palmyra092007FS192-070924-205305.wav and palmyra092007FS192-070924-205730.wav
Long-beaked common dolphin	1	Qx-Dc-CC0411-TAT11-CH2-041114-154040-s.wav
	2	Qx-Dc-SC03-TAT09-060516-171606.wav
	3	QX-Dc-FLIP0610-VLA-061015-165000.wav
Melon-headed whale	1	palmyra092007FS192-070925-023000.wav
	2	palmyra092007FS192-071004-032342.wav
	3	palmyra102006-061020-204327_4.wav
Short-beaked common dolphin	1	Qx-Dd-SCI0608-N1-060815-100318.wav
	2	Qx-Dd-SCI0608-Ziph-060817-100219.wav
	3	Qx-Dd-SCI0608-Ziph-060817-125009.wav
Spinner dolphin	1	palmyra092007FS192-070927-224737.wav
	2	palmyra092007FS192-071011-232000.wav
	3	palmyra102006-061103-213127_4.wav

to disambiguate crossing tonals until after a graph has been completed. This allows information from both sides of a crossing to be considered when disambiguating multiple whistles that cross.

For both methods, false detections are reduced by discarding detections of less than 150 ms duration that are frequently due to noise. While some whistles may be shorter than 150 ms, results reported by Oswald *et al.* (2003) for nine species of odontocetes in the eastern tropical Pacific (four of which are covered in this study) had mean durations from 0.3 to 1.4 s, with the species producing the shortest duration whistles having a standard deviation of 0.3.

1. Particle filter

If we have a sequence of detected spectral peaks $\boldsymbol{s}_{1:t}$ from time index 1 to t that can be used to model a sequence of contour estimates $\boldsymbol{c}_{0:t}$ as a general Markovian process, Bayes' theorem describes the posterior distribution (that of the estimated contour given the spectral peaks) at any time t as

$$p(\boldsymbol{c}_{0:t}|\boldsymbol{s}_{1:t}) = \frac{p(\boldsymbol{s}_{1:t}|\boldsymbol{c}_{0:t})p(\boldsymbol{c}_{0:t})}{\int p(\boldsymbol{s}_{1:t}|\boldsymbol{c}_{0:t})p(\boldsymbol{c}_{0:t})d\boldsymbol{c}_{0:t}}, \tag{1}$$

where the initial contour estimate, $\boldsymbol{c}_0$ is set to the first spectral peak encountered that is not associated with another whistle. This joint distribution of the posterior can be written as a recursive pair of prediction and updating equations using the Chapman-Kolmogorov equation (Papoulis, 1991, p. 193) and Bayes' theorem

$$\begin{aligned} \text{Prediction} &: p(\boldsymbol{c}_t|\boldsymbol{s}_{1:t-1}) = \int p(\boldsymbol{c}_t|\boldsymbol{c}_{t-1})p(\boldsymbol{c}_{t-1}|\boldsymbol{s}_{1:t-1})d\boldsymbol{c}_{t-1}, \\ \text{Updating} &: p(\boldsymbol{c}_t|\boldsymbol{s}_{1:t}) = \frac{p(\boldsymbol{s}_t|\boldsymbol{c}_t)p(\boldsymbol{c}_t|\boldsymbol{s}_{1:t-1})}{\int p(\boldsymbol{s}_t|\boldsymbol{c}_t)p(\boldsymbol{c}_t|\boldsymbol{s}_{1:t-1})d\boldsymbol{c}_t}. \end{aligned} \tag{2}$$

This recursion describes a Bayesian filtering process, where the posterior that is estimated in one time step is used as the prior distribution (our belief about the previous frequency of the whistle) in the subsequent time step. If all of these distributions are Gaussian and the state updates are linear, then Kalman filtering provides a closed-form solution to this recursion. When this constraint does not hold, as in many systems of interest, sequential Monte Carlo methods such as particle filtering can be used to find estimates of this posterior.

The particle filter estimates the posterior update with a weighted collection of N point samples or particles $\boldsymbol{c}_t^i$,

$$p(\boldsymbol{c}_t|\boldsymbol{s}_{1:t}) \approx \sum_{i=1}^{N} w_t^i \delta(\boldsymbol{c}_t - \boldsymbol{c}_t^i), \tag{3}$$

where the weights w_0^i are each initialized as $1/N$. Here, the continuous posterior is approximated as a discrete distribution using the Dirac δ function $\delta(\cdot)$ over each of the i particles, and the particle weights w_t^i are normalized. Since the shape and peak of the posterior are unknown, we generate point samples by using a distribution we define. This distribution is referred to as an importance density, $q(\boldsymbol{c}_t|\boldsymbol{s}_{1:t})$, and the particle weights are set proportionally, $w_t^i \propto p(\boldsymbol{c}_t|\boldsymbol{s}_{1:t})/q(\boldsymbol{c}_t|\boldsymbol{s}_{1:t})$. This can be written recursively using Bayes' theorem as

$$w_t^i \propto w_{t-1}^i \frac{p(s_t|\boldsymbol{c}_t^i)p(\boldsymbol{c}_t^i|\boldsymbol{c}_{t-1}^i)}{q(\boldsymbol{c}_t^i|\boldsymbol{c}_{t-1}^i, \boldsymbol{s}_t)}. \tag{4}$$

By setting the importance density as the product of particle weight in the previous time step and the state update prior, $q(\boldsymbol{c}_t^i|\boldsymbol{c}_{t-1}^i, \boldsymbol{s}_t) = w_{t-1}^i p(\boldsymbol{c}_t^i|\boldsymbol{c}_{t-1}^i)$, the particle weight becomes

$$w_t^i \propto p(\boldsymbol{s}_t|\boldsymbol{c}_t^i). \tag{5}$$

In this way, the particle weights are resampled at each time step and normalized to sum one in a process referred to as sampling importance resampling (Gordon *et al.*, 1993). In the work presented here, the likelihood function $p(\boldsymbol{s}_t|\boldsymbol{c}_t^i)$ takes the form of a normal distribution.

To improve performance, systematic resampling (Kitagawa, 1996) is implemented with particle replacement at each time step. During each recursion, particles with a low weight are extinguished and replacement particles are regenerated near particles with a large weight. Particles far from the peak of the posterior are removed and more particles are added near the peak so that particles have a better chance of being distributed within the informative parts of the posterior. This is done within a continuous resampling space and works much like a regularized particle filter (Musso *et al.*, 2001).

In the predictive step, the particle locations are updated according to the motion model for the whistle update. When the whistle estimate is at least seven samples long, it is used to approximate the first and second order time derivatives of the whistle frequency at time $t-1$. These rates of change are used with a standard second order equation of motion (6) to estimate the new location for each particle at time t. As a way to further increase the chances that the particles will be well distributed throughout the measureable space spanned by the posterior and to avoid being gradually herded off course by perpetuating state estimate errors, a small random noise is included in the prediction step model. This adjustment is accounted for as a Gaussian random walk and applied within the motion model. The random adjustment ϵ is drawn from a Gaussian distribution such that 95% of the draws are within about 40 Hz of the prediction (zero mean, variance of 375 Hz, or a few frequency bins) and added into the particle motion model to update particle positions c_t^i,

$$\begin{aligned} &\epsilon \sim N(\mu = 0, \sigma^2 = 375), \\ &c_t^i = c_{t_0}^i + \dot{f}(t - t_0) + \frac{1}{2}\ddot{f}(t - t_0)^2 + \epsilon. \end{aligned} \tag{6}$$

Here, the estimated first and second derivatives of the whistle contour at an initial time step are $\dot{f}$ and $\ddot{f}$, respectively. Prediction is typically for a single time step $(t_0 \equiv t - 1)$; however, larger time steps can occur when trying to reacquire whistle contours that have been lost due to brief signal masking such as echolocation clicks.

The likelihood function $p(\boldsymbol{s}_t|\boldsymbol{c}_t^i)$ describes how likely the updated estimates are with the observations, and can take the form of a Gaussian distribution without restricting the ability of the particle filter to estimate non-Gaussian posteriors. In cases where the particles span multiple sound intensity peaks, treating the likelihood as a sum of Gaussians using each observation instead of taking the "best" data point increases the power of the particle filter to navigate a whistle contour when the observations present a more complicated scene. Once the weights are determined by calculating the likelihood for each particle update, the center of mass of the particles represents the best estimate for the peak of the posterior, and is used as the estimated frequency of the whistle contour at time t. To improve performance, the whistle contour is modeled as a three dimensional feature, including not only frequency, but also the first and second order derivatives of the contour, $\boldsymbol{c} = [f, \dot{f}, \ddot{f}]^T$. The added shift in frequency described in Eq. (6) is applied directly to the first component of $\boldsymbol{c}$, and the first and second derivatives of the contour are estimated using the last five samples of the contour estimate. The likelihood function is chosen as a multivariate normal distribution with the mean defined by the location of each particle, a zero covariance, and the diagonal variance $\boldsymbol{\Sigma} = \Delta_f \cdot [3,\ 1.5,\ 1]$. This can be written as $p(\boldsymbol{s}_t|\boldsymbol{c}_t^i) = N(\boldsymbol{s}_t|\boldsymbol{c}_t^i, \sum)$, where $\boldsymbol{s}_t = [s_{1,t}, s_{2,t}, s_{3,t}]^T$ is generated from a spectral peak at time t. The first component $\boldsymbol{s}_{1,t}$ is the frequency of the spectral peak, and the second $\boldsymbol{s}_{2,t}$ and third $\boldsymbol{s}_{3,t}$ components represent the rates of change $\dot{f}$ and $\ddot{f}$, respectively, as determined by treating $\boldsymbol{s}_{1,t}$ as the subsequent contour update. Early in contour finding, before enough contour updates have been found to approximate $\dot{f}$ or $\ddot{f}$, a lower dimensional likelihood function is used that is scaled to the number of features available. In this way, the best matches of both the particles and the spectral peaks are found based on the available information.

A set of particles is used to estimate a single whistle contour, as shown in Fig. 1. When there are multiple whistle contours present, each contour is estimated with its own set of particles in each time step. New whistles are initiated each time a sound level peak occurs that cannot be associated with an existing set of particles describing current whistle contours. This can happen due to spectral separation, on the order of 500 Hz difference between a sound level peak and a current whistle contour using the given likelihood function, or it can occur due to having more numerous sound level peaks than current whistle contours. Groups of particles are not considered to be a whistle contour until a minimum number of time updates occur. The threshold is based on a user specified minimum whistle duration.

2. Graph detection of tonals

The graph search algorithm maintains two sets of graphs to organize candidate detections. The first set is the *fragment* set and in general contains small fragments of whistles that are identified. As these fragments grow, they are migrated to the *active* set which consists of longer sets of time×frequency peaks without any attempt to disambiguate tonal crossings. Each graph has a set of endpoints that may be extended as new time × frequency peaks are discovered. After a period of inactivity where no new elements are added to an active graph, it is removed from the *active* set and a disambiguation algorithm extracts individual whistles.

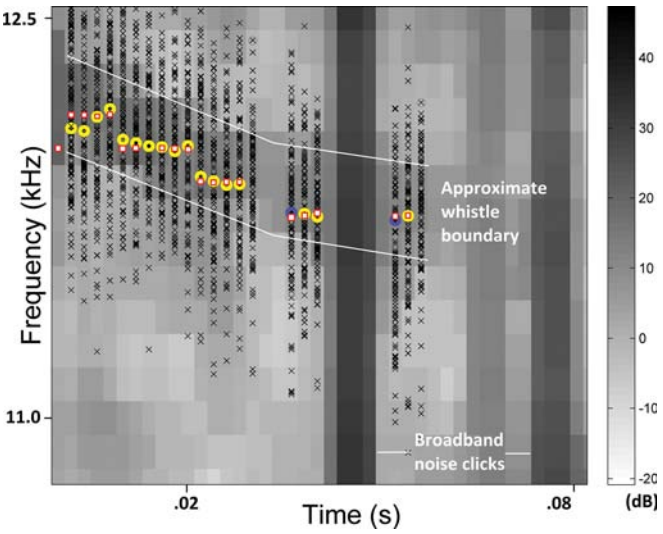

FIG. 1. (Color online) Particle filter performance in whistle discovery as shown with a spectrogram. Approximate boundary of an odontocete whistle is marked by the solid lines. Detected peaks of whistle are shown as squares. Particles in each time step are shown as ×'s, and the center of mass of the particles is depicted as circles. These are the effective measures of the whistle contour in each time step. Tracking continues even through time frames without nearby whistle peaks allowing whistle contour detection to resume once peaks are detected again. Locations where the whistle contour detection is resumed are denoted with darker circles.

The two major operations for each frame of the spectrogram consist of graph extension and graph pruning. Extension consists of examining the peaks and determining if they are appropriate for extending an existing graph in the *fragment* or *active* sets. The pruning step identifies graphs whose endpoints are too far away from time t to be extended, and identifies the whistles contained therein. Throughout this section, t denotes start times of spectrogram frames.

a. Graph extension. Criteria for graph extension are based on an adaptive polynomial fit of a recent portion (25 ms) of the path to be extended. When the multiple paths are possible due to recent whistle crossing, each possible path is fit. The fit uses an ordinary least squares criterion (Press, 1992, Chap. 15.4). The goodness of the fit is measured by an adjusted R^2 coefficient (Dillon and Goldstein, 1984, Chap. 6.3.2), which penalizes the fitness measure by a function of the number of parameters and data points:

$$\hat{R}^2 = \frac{\dfrac{\sum_t \left(s_t - \hat{p}_f(t)\right)^2}{N - (\text{degree}(\hat{p}_f) - 1)}}{\dfrac{\sum_t (s_t - \mu_s)^2}{N - 1}}, \tag{7}$$

where t varies over the N regression samples, μ_s is the mean of the regression sample frequencies, and $\hat{p}_f(\cdots)$ is a prediction polynomial of order $\text{degree}(\hat{p}_f)$. The fit is initially tried with a first order polynomial. A heuristic that accounts for

the sensitivity of polynomial prediction to quantization noise along with a check for goodness of fit and quantity of estimation data is used to determine whether or not a higher order polynomial should be applied. Letting $\sigma_{\hat{p}_f}$ denote the standard deviation of the squared residuals, poor fits are re-estimated with the next higher order polynomial when the heuristic

$$\hat{R}^2 < 0.6, \quad \sigma_{\hat{p}_f} > 2\Delta_f, \quad \text{and} \quad N > 3\,\text{degree}(\hat{p}_f) \qquad (8)$$

is satisfied.

A new peak, s_t, extends one or more paths in an existing graph(s) if it is within 50 ms of the path's endpoint and $|\hat{p}_f(t) - s_t| \leq 1000$ (Fig. 2). Connections are first tried in the *active* set to favor well established paths. If no match is found, the *fragment* set is searched. Should an appropriate path from a *fragment* set graph be found, the duration of the newly extended path is examined and the graph is moved to the *active* set if the longest possible path exceeds 50 ms. When no viable extensions of existing paths are feasible, a new graph consisting of the detected peak is added to the *fragment* set.

There are two special cases that merit discussion. It is possible for a peak to be added to more than one graph. An example of this occurs when tonal contours cross. In this case, the graphs are merged using the union-find algorithm (Cormen *et al.*, 1990, Chap. 21) which permits efficient merging of sets with near constant-time performance. By merging the graphs, we delay the decision about which path should be taken on the other side of the crossing until the graph has been completed, allowing information from both sides of the crossing to be used. The second case arises when two tonals are in close proximity to one another and share a similar slope. Such spectral peaks are typically within the tolerance range of the predicted path, and the ability to connect multiple peaks to the same graph endpoint can result in a lattice structure where two roughly parallel segments are bridged many times. To prevent this, the same peak is not permitted to be joined to two endpoints that are part of the same graph.

b. Graph pruning. After each graph extension, graphs are pruned. When a graph has no end points that are within 50 ms of the current frame, it is no longer possible to extend the graph. Consequently, the graph is removed from the *active* or *fragment* set. When the time difference between the first and last nodes of the graph are less than 150 ms, the graph is discarded. An example of graphs produced by analyzing common dolphin whistles can be seen in the third panel of Fig. 3.

Graphs that are retained are subjected to a disambiguation step. Conceptually, graph paths are reduced to a set of nodes that are either start/termination points for a candidate whistle or intersection points. Each intersection is resolved into one or more contour segments by examining each possible pairing between arcs leading into and out of an intersection node. Nodes with longer paths associated with them are more likely to be important and are processed first. Ordering is established by multiplying the length of the longest input and output paths associated with each node.

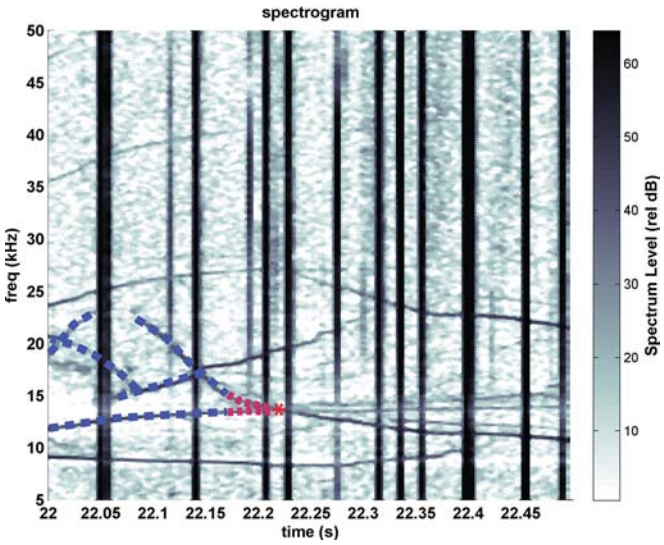

FIG. 2. (Color online) Graph extension. Dashed curves depict an active graph. A peak is depicted by an asterisk and ordinary least squares regression curves are fit along the closest 25 ms of paths near the peak as indicated by the change in shade and dash pattern. Peaks that are within 1 kHz of the path predicted by the polynomial fits will be added to the graph.

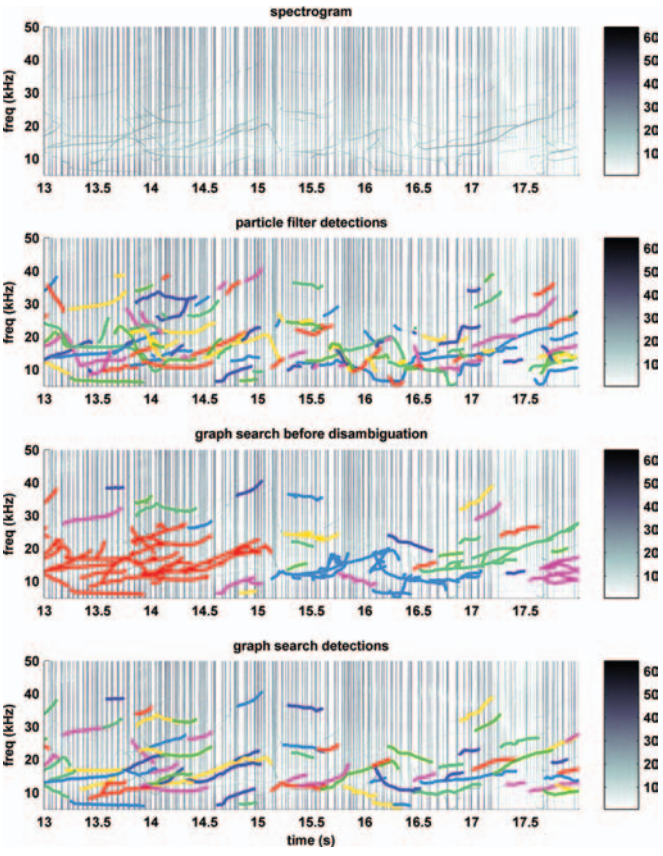

FIG. 3. Whistle detection algorithm performance amid the interference of odontocete echolocation clicks. The uppermost panel shows a spectrogram of 5 s of long-beaked common dolphin call data (analysis bandwidth 125 Hz) with relative dBs of signal to noise ratio encoded by gray levels. The second panel shows the whistles detected by the particle filter algorithm. The last two panels show the whistle graphs and extracted whistles as detected by the graph search algorithm.

Input and output path pairs are assigned scores based on a heuristic derived from the adaptive polynomial fit used in the graph extension step. Forward and backward average squared prediction errors from up to 300 ms of the incoming and outgoing paths are summed to determine the feasibility of each pairing:

$$\text{penalty}(\text{in}_{i,k},\text{out}_{j,k}) = \frac{1}{\left|\text{path}_{k,\text{in}_{i,k}}\right|}\sum_{t_{\text{in}}\epsilon\text{path}_{k,\text{in}_{i,k}}}\left(s_{t_{\text{in}}}-\hat{p}_f^{\text{out}_{j,k}}(t_{\text{in}})\right)^2 + \frac{1}{\left|\text{path}_{k,\text{out}_{j,k}}\right|}\sum_{t_{\text{out}}\epsilon\text{path}_{k,\text{out}_{j,k}}}\left(s_{t_{\text{out}}}-\hat{p}_f^{\text{in}_{i,k}}(t_{\text{out}})\right)^2, \quad (9)$$

where t_{node} is the time in s associated with the junction node, $\text{in}_{i,k}$ and $\text{out}_{j,k}$ represent the ith input and jth output edges, respectively, from intersection node k, $\text{path}_{p,k} = \{\text{all nodes along path } p \leq 0.3\,\text{s from intersection node } k\}$. The prediction polynomials $\hat{p}_f^{\text{in}_{i,k}}$ and $\hat{p}_f^{\text{out}_{j,k}}$ are estimated from .3 s of data or to the nearest intersection node along the input and output paths. An example can be seen in the crossing whistles of Fig. 4. A total of four penalties will be computed, the first two of which are between one of the incoming edges and the two outgoing ones (highlighted). The first of these penalties is formed by estimating predictor polynomials for edges $\overrightarrow{DA}$ and $\overrightarrow{AB}$: $p_f^{\overrightarrow{DA}}$ and $p_f^{\overrightarrow{AB}}$. The average squared error of the predictions of $p_f^{\overrightarrow{DA}}$ onto the closest 0.3 s of $\overrightarrow{AB}$ and vice-versa are summed. This is repeated for the other three possible combinations, and a greedy algorithm connects the paths with the lowest penalties. When no more pairs can be processed, the next intersection node is examined. The processing of intersection nodes is ordered by the lengths of the longest possible pair of paths to favor longer whistles. As will be shown empirically, spurious detections tend to have higher false positive rates, and the rationale is to build on what are likely to be better detections first.

Due to the frequency quantization in discrete Fourier transforms, an optimization was added to address whistles with similar slopes. When this occurs, the whistles' paths may fall in the same time × frequency bin for multiple frames. This results in two intersection nodes that may have a single output and input path between them that corresponds to both whistles. When two intersection nodes share a single path with multiple inputs on one side and outputs on another, we permit the path that bridges the two nodes to belong to multiple whistles (Fig. 5).

The disambiguation algorithm results in a new set of graphs where each candidate tonal has no crossings, but may have one or more extraneous edges. These are removed by a final pass that filters out short edges of less than 5 ms, which are not part of an interior path.

D. Ground truth and metrics

An important component of an automated detection system is the ability to measure its performance. This must be done by comparing the system output with a set of known detections, referred to as ground truth information. Although spectrograms of whistles can be subjective, humans typically perform well on visual separation tasks and a trained analyst (author Y.B.) used custom software that permitted the user to interactively specify tonal contours. The analyst placed points along a whistle through which cubic B-spline curves were fit. B-splines consist of multiple piecewise Bezier curves that are constrained to have smooth transitions between certain points through which the B-spline must pass (Dierckx, 1993, Chap. 1). It is important to recognize that while efforts were made to carefully record accurate ground truth information (including inspection of randomly sampled

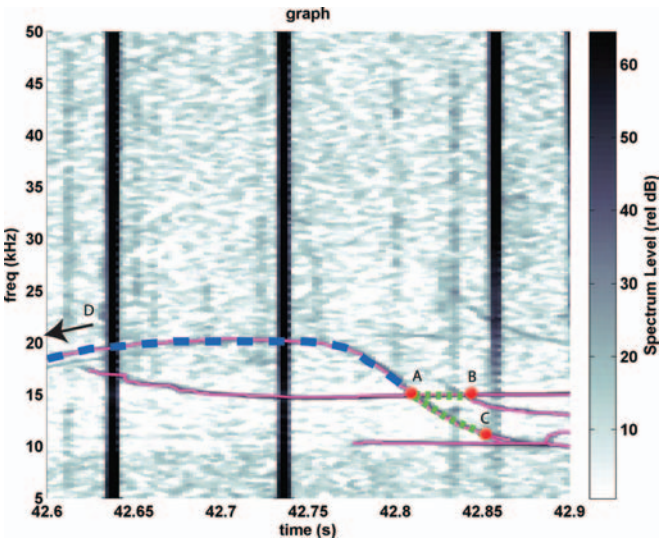

FIG. 4. Graph disambiguation. When deciding whether the incoming arc $\overrightarrow{DA}$ should be joined with the outgoing arc $\overrightarrow{AB}$ or $\overrightarrow{AC}$, polynomials $\hat{p}_f$ are estimated for all three arcs. The sum of the squared prediction errors of pairs of incoming and outgoing edges [Eq. (9)] is used to determine which pairs should be joined. In this example, $\overrightarrow{DA}$ is joined to $\overrightarrow{AC}$.

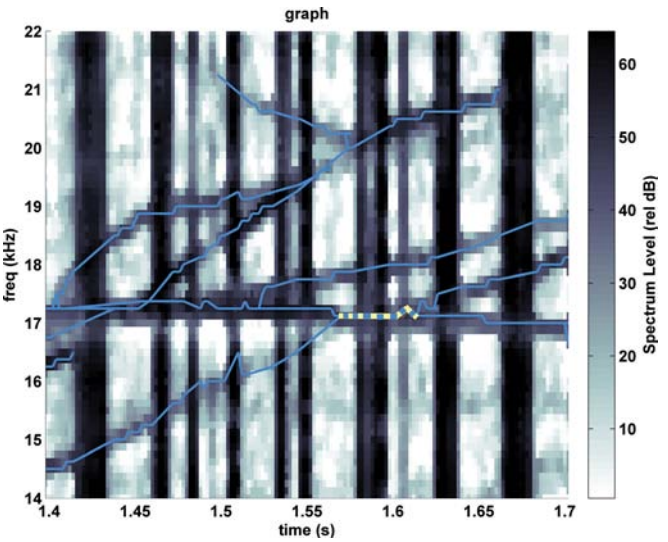

FIG. 5. (Color online) Common subpaths. The graph for these common dolphin whistles shows a dashed segment that is shared between two whistles. The intersection nodes are characterized by having multiple inputs on one side and multiple outputs on the other, joined by a single segment. When this occurs, the disambiguation algorithm permits the segment to be used in more than one whistle, permitting both whistles in the figure above to be recognized.

segments for quality control), some decisions are subjective and some amount of error is nearly inevitable.

In general, the analyst worked on short segments of 3–5 s of recording and would adjust the spectrogram contrast and brightness to most favorably display the tonal contours. Complete tonals as well as fragments were noted, regardless of their length or signal-to-noise ratio (SNR). Stepped whistles were recorded as single tonals, while harmonics were recorded separately. When echoes could be clearly distinguished they were not recorded.

The analyst-specified ground truth information was compared to the detected whistles using a series of metrics and selection criteria. The metrics are designed to measure the correctness and quality of detections. The selection criteria are used to determine which tonals were expected to be detected, and are based on SNR and length metrics. As the SNR of tonal calls can vary depending upon the part of the call, tonals are only expected to be detected when a certain percentage of the contour exceeds a specified SNR. A second criterion rejects tonals that are less than a minimum duration. We set the selection criteria to be appropriate for the types of signals that could possibly be detected based on the thresholds used in our algorithms: whistles of 150 ms or longer with a third of the whistle having a SNR ≥ 10 dB.

For each tonal in the ground truth tonal list, we examine the set of detected tonals that overlap the start and end time of the detected tonal. This is done regardless of whether or not the tonal meets the selection criteria. All ground truth tonals are processed so that it can be determined whether or not a detection matches some ground truth tonal that failed the selection criteria. In such cases, the matched tonal will not be included in the metrics that describe the quality and quantity of matches, but neither will it be considered to be a false positive (bad match).

As the cubic spline interpolations may have minor deviations from the actual tonal path, the recorded frequencies are quantized to the nearest 125 Hz (based on an 8 ms analysis window) and a search is conducted within ± 500 Hz (± 4 bins) for the frequency bin with maximal energy. For each overlapping point between a detected tonal and a specific current ground truth tonal, the absolute frequency difference between the detection and ground truth peak is computed. If the mean difference is ≥ 350 Hz (a few frequency bins away), the detected tonal is rejected as a false positive. Otherwise, it is marked as a valid detection.

Measurements of system performance describe the system's ability to retrieve tonals as well as the quality of the retrieved matches. The primary system metrics are recall and precision. Recall measures the percentage of the expected detections that were retrieved,

$$\text{recall} = \frac{\sum_{g \epsilon \text{ground}_c} \text{match}(\text{detections}, g)}{|\text{ground}_c|} \times 100, \quad (10)$$

where ground_c is the set of ground truth tonals subject to the aforementioned selection criteria, and $\text{match}(t_1, t_2)$ is an indicator function that returns one if tonal t_2 has one or more valid detections in t_1, and zero otherwise. Precision is a metric that measures the percentage of detections that are correct:

$$\text{precision} = \frac{\sum_{d \epsilon \text{detections}} \text{match}(d, \text{ground}_c)}{|\text{detections}|} \times 100, \quad (11)$$

and the false positive rate is simply 100-precision.

Several other metrics are defined to assess the quality of matches. Coverage is an indication of the average percentage of a ground truth tonal that is matched and is truncated at 100% to prevent artificial inflation of the coverage statistic should a detection be slightly longer than a ground truth tonal. As multiple detections may cover a single ground truth tonal, fragmentation is a measure of the average number of detections per ground truth tonal. Deviation is a measure of the average frequency deviation between the path of ground truth tonal and its corresponding detection(s). Metrics are summarized in Fig. 6.

III. RESULTS

Over three thousand ground truth whistles met the selection criteria that tonals must be at least 150 ms in duration and that at least a third of the tonal had to have a SNR of ≥ 10 dB. The number of whistles meeting these criteria and the metrics associated with their detections are summarized by sighting and species in Table III. The particle filter was able to retrieve 71.5% (recall) of the 3372 ground truth tonals with a precision of 60.8%. The graph algorithm showed a recall rate of 80.0% with a precision of 76.9%. The average deviation from the ground truth frequency was low for both algorithms (particle filter 161 Hz, $\sigma = 51$, graph search 70 Hz, $\sigma = 76$). Both algorithms performed reasonably well on the coverage (particle filter 79.7% $\sigma = 23.2$, graph search 86.0% $\sigma = 20.5$) and fragmentation (1.2 detections per tonal for both algorithms) metrics. Sample

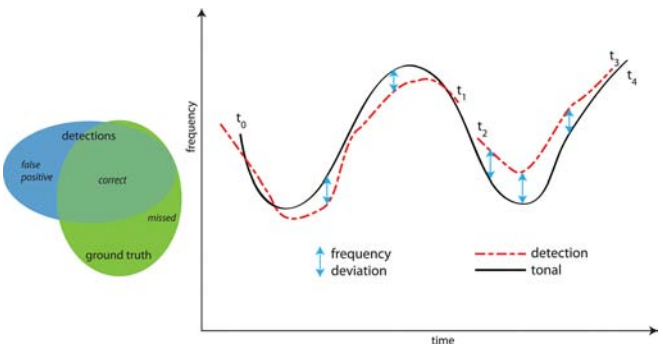

FIG. 6. (Color online) Metrics used to characterize detections. The Venn diagram on the left shows the overlap between the detected tonals and ground truth data. Recall computes the percentage of correct detections relative to the ground truth while precision is the percentage of detections that were correct. The exaggerated caricatures of a call and associated detections on the right illustrate the quality metrics. Average deviation is the mean frequency deviation between the tonal call and detection(s). As systems may detect a call in multiple pieces, or fragments, the number of fragments per call is recorded. Coverage is an indication of the percentage of the tonal that was detected and in this case would be $\{[(t_1 - t_0) + (t_3 - t_2)]/(t_4 - t_0)\}100$. Call and detection data are caricatures with exaggerated frequency deviation.

TABLE III. Performance comparison of graph and particle filter algorithms for the detection of odontocete whistle contours. Summary statistics are computed across all ground truth tonals meeting SNR and duration selection criteria (see text) and are not averages of sighting statistics. When given, $\pm\sigma$ indicates standard deviation.

			Particle filter					Graph search				
Species	Sighting	Tonals	Precision	Recall	μ deviation $\pm\sigma$ Hz	Coverage $\pm\sigma$ %	Fragments	Precision	Recall	μ deviation $\pm\sigma$ Hz	Coverage $\pm\sigma$ %	Fragments
Bottlenose dolphin	1	89	69.9	79.8	170 ± 53	84.1 ± 21.2	1.3	67.6	84.3	44 ± 59	83.1 ± 21.6	1.3
	2	265	95.9	82.6	141 ± 51	76.4 ± 22.8	1.2	95.5	82.6	128 ± 51	77.0 ± 22.3	1.3
	all	354	87.2	81.9	148 ± 53	78.3 ± 22.7	1.2	86.4	83.1	106 ± 65	78.5 ± 22.2	1.3
Long-beaked common dolphin	1	300	11.4	26.3	173 ± 64	57.5 ± 32.0	1.5	18.0	20.3	148 ± 71	71.0 ± 25.0	1.2
	2	10	84.6	90.0	138 ± 27	70.3 ± 24.7	1.2	100.0	80.0	94 ± 15	78.1 ± 24.7	1.5
	3	247	92.5	86.6	148 ± 52	84.3 ± 21.2	1.3	93.6	86.6	44 ± 64	88.1 ± 18.3	1.2
	all	557	29.9	54.2	154 ± 56	76.9 ± 27.2	1.3	49.4	50.8	68 ± 78	84.1 ± 21.2	1.2
Melon-headed whale	1	90	78.5	67.8	140 ± 52	74.8 ± 22.0	1.0	81.2	71.1	40 ± 51	79.0 ± 23.3	1.1
	2	78	21.8	69.2	166 ± 46	78.4 ± 18.7	1.1	17.6	64.1	100 ± 35	80.8 ± 17.4	1.1
	3	170	86.5	74.1	151 ± 51	78.6 ± 23.6	1.2	88.2	72.9	108 ± 54	81.0 ± 20.0	1.2
	all	338	52.7	71.3	151 ± 50	77.6 ± 22.2	1.1	48.5	70.4	88 ± 58	80.4 ± 20.4	1.1
Short-beaked common dolphin	1	92	73.5	78.3	155 ± 52	79.6 ± 20.0	1.2	66.9	83.7	137 ± 71	73.6 ± 23.5	1.1
	2	1112	66.8	64.4	166 ± 64	81.9 ± 23.5	1.1	96.7	90.5	18 ± 51	95.0 ± 15.0	1.1
	3	233	76.3	86.3	146 ± 42	83.2 ± 20.5	1.2	79.2	89.7	46 ± 63	85.8 ± 20.5	1.3
	all	1437	69.1	68.8	161 ± 60	82.0 ± 22.7	1.1	90.7	89.9	30 ± 61	92.2 ± 17.6	1.1
Spinner dolphin	1	357	85.4	88.2	177 ± 50	76.0 ± 22.7	1.2	88.8	89.1	130 ± 59	77.6 ± 21.8	1.4
	2	146	87.9	81.5	162 ± 45	76.6 ± 22.0	1.1	86.4	82.9	127 ± 56	77.2 ± 18.5	1.2
	3	183	86.2	84.2	175 ± 53	86.6 ± 19.2	1.3	83.3	82.0	141 ± 59	83.4 ± 21.1	1.5
	all	686	86.1	85.7	173 ± 50	78.9 ± 22.1	1.2	86.8	85.9	132 ± 58	79.0 ± 21.1	1.4
Overall		3372	60.8	71.5	161 ± 56	79.7 ± 23.2	1.2	76.9	80.0	70 ± 76	86.0 ± 20.5	1.2

detections for various levels of acoustic clutter can be seen in Figs. 3 and 7.

IV. DISCUSSION

Both algorithms demonstrate the ability to extract whistles from very complex auditory scenes with many animals vocalizing simultaneously. The precision associated with both algorithms deserves further analysis, as the values indicate that both algorithms produce a fair number of false positives. The majority of these false positives are quite short, as seen in the cumulative distribution function for false positives with respect to length (Fig. 8). They occur most often in regions with strong noise and in areas where the noise floor rises suddenly. Examples of phenomena that can give rise to this include increases in wind velocity, rainfall, and anthropogenic sources. The large number of false positive in the second melon-headed whale recording is directly attributable to

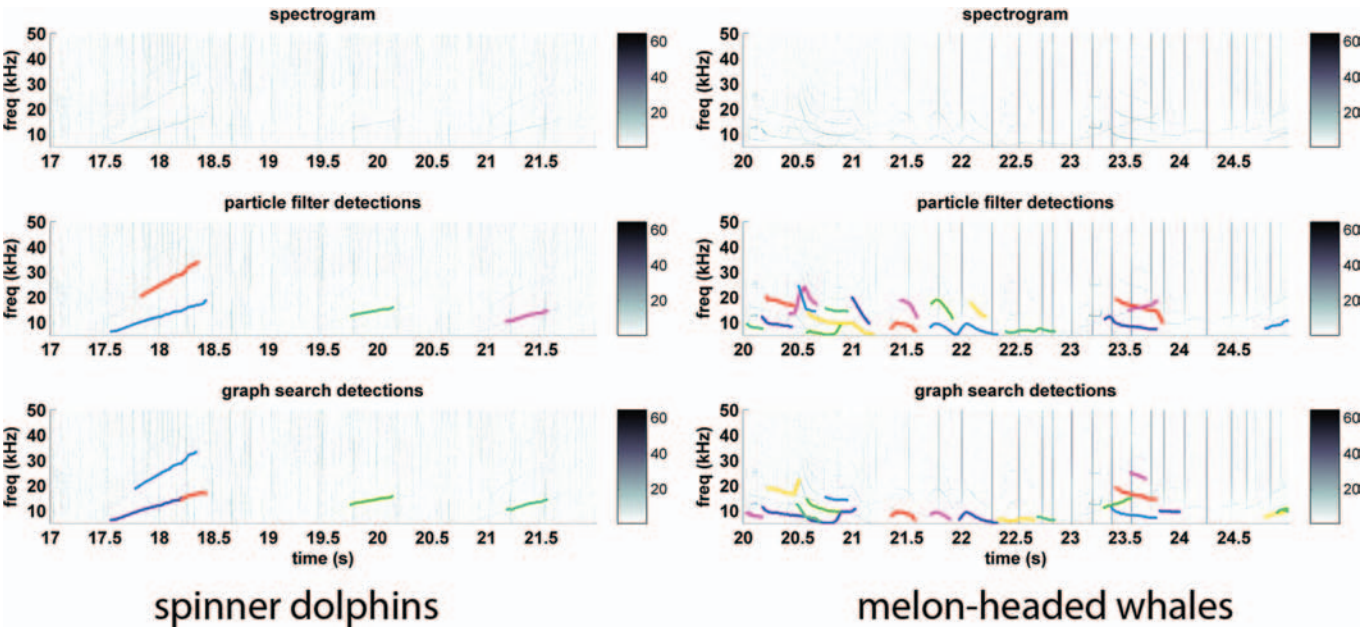

FIG. 7. Sample detections of acoustic scenes with differing degrees of clutter.

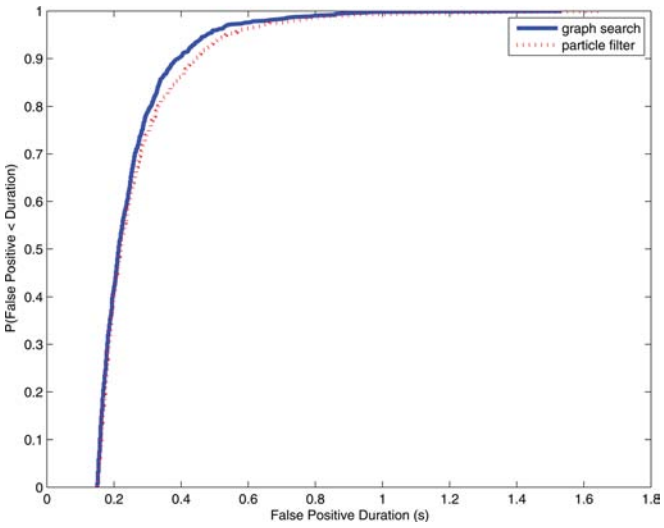

FIG. 8. (Color online) Cumulative density function for incorrect detections whose duration is less than or equal the duration indicated on the False Positive Duration axis. Both algorithms require that a hypothesized tonal have a duration $\geq$150 ms to be reported as a detection. The vast majority of false positive detections for both algorithms have short duration.

broadband hydrophone tow noise between 5–25 kHz that occurred when the tow vessel executed tight turns. A third contributor to erroneous detections is echo sounder pings that produce chains of peaks that the algorithms organize into tonals (Fig. 9). This is the major cause for the poor precision observed in the first long-beaked common dolphin sighting.

Just as transitions into higher noise regions can cause false positives, transitions into lower noise regions can result in misses due to low signal to noise estimates in the peak detection algorithm. Both types of errors suggest that improvements to the noise estimation and removal portion of the common signal processing chain could be a productive area for future performance gains. Finally, missed detections also occur in regions of very high impulsive noise density such as occur in strong burst pulsed calls which are series of echolocation clicks produced with a very short interclick interval.

There are a number of situations where a single whistle will commonly result in multiple detections. As the system does not track harmonics or associate echoes with the first arrival, these are seen as separate events. Similarly, stepped whistles are tracked as separate entities when the step size is large. Some of these events have the potential to be associated during post-processing analysis; however, this will require non-trivial effort due to phenomena such as incomplete detections and propagation loss at higher frequencies. A final type of duplicate detection occurs in the particle filter detector only. Occasionally, when spectral peaks are in close proximity, one of the peaks will be used to form a new hypothesis instead of updating an existing hypothesis. Subsequent peaks alternate between the two hypotheses, leapfrogging one another and forming two tonal paths instead of one. Improving the rules in governing the update of whistle paths might alleviate this problem, particularly since whistle paths that missed an update have first priority when new peaks are presented.

For the whistles that are correctly detected, performance is overall quite good. Detected tonals follow the human analyst's ground truth track closely, and typically cover 80–85% of the whistle as recorded by the analyst. The majority of times, whistles are detected as single contours, although the fragmentation rate of 1.2 indicates that this is not always the case.

Data from other common and bottlenose dolphin sightings collected using the same equipment and methods were used in the development of the algorithms, and there was no significant tuning of algorithm parameters for the data reported in these experiments. As developed, these algorithms are quite effective for determining presence/absence of animals and should be able to provide reasonable estimates of contour statistics for longer calls. When visual observations are available, the extracted contours are suitable for development of species recognition algorithms as well as the exploration of associations between behavioral state and whistle content. The minimum length threshold of 150 ms along with the propensity for false detections in

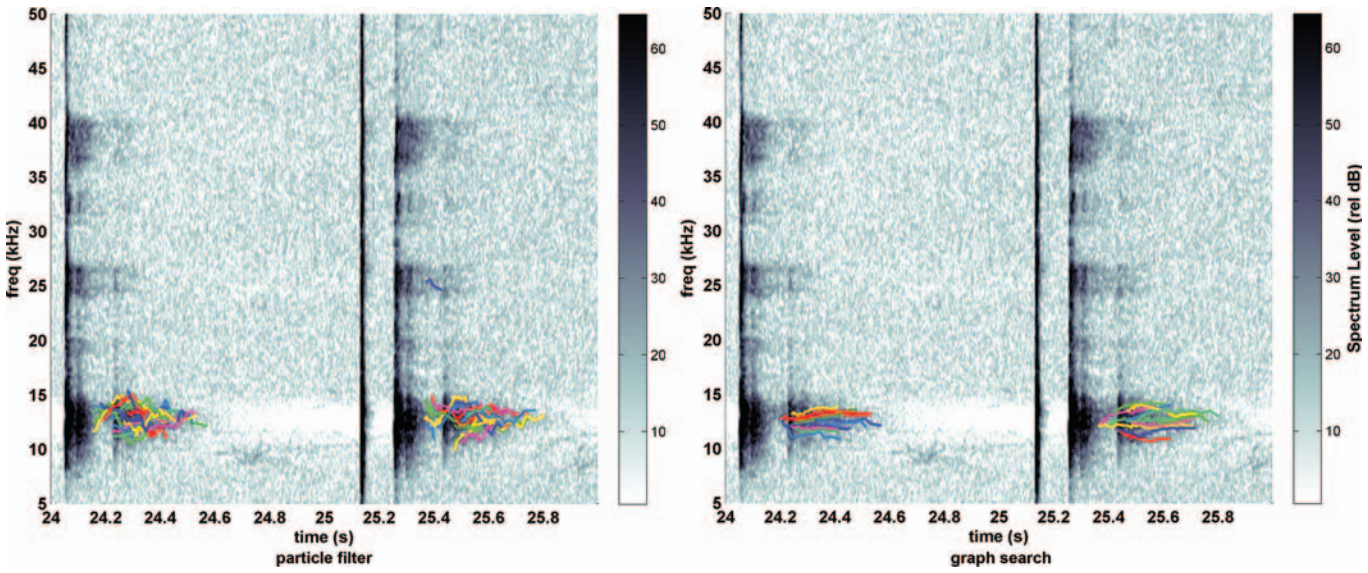

FIG. 9. Example of false positive detections caused by echosounders in both algorithms.

shorter whistles could impact behavioral studies, and future work should investigate additional noise reduction techniques to more reliably extract shorter whistles.

V. CONCLUSIONS

Both the particle filter and the graph search algorithms show the ability to extract whistles from complex auditory scenes from five different species containing multiple overlapping simultaneous whistles. This is demonstrated on a diverse five species dataset consisting of nearly one hour of recorded data with 3372 ground truthed (analyst detected) calls meeting retrieval criteria of having a relative SNR $\geq$10 dB for at least one third of the call and a duration $\geq$150 ms.

The algorithms are capable of retrieving tonal contours at a speed of several times real-time on modern computer architectures. The graph search algorithm outperformed the particle filter, retrieving 80.0% of the whistles versus 71.5% by the particle filter. A higher percentage of the detections from the graph search algorithm (76.9%) matched ground truth calls than those produced by the particle filter (60.8%), and in both cases the false positives were dominated by short duration detections. Correct matches were typically within one to two frequency bins of the ground truth tonal. Approximately 80% or more of each tonal was detected (79.7% particle filter, 86.0% graph search) and whistles were on average split into 1.2 detections indicating that most tonals were not split. The most challenging environments for either algorithm include those with echo sounders, heavy burst pulse call activity, and regions of noise state transition, all of which are areas for further development of the spectral peak detector.

Direct comparisons with other algorithms are difficult due to differences in data sets, and we avoid making any claims about our algorithms versus others for this reason. In an effort to encourage such comparisons, the audio data from these experiments have been made available to the bioacoustics community in the Moby Sound archive (Heimlich *et al.*, 2011) as part of the Fifth International Detection, Classification, and Localization Workshop dataset. The ground truth information will be released to the Moby Sound archive after the workshop (August 22–25, 2011 in Portland, OR).

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their helpful comments on an earlier version of this manuscript. Numerous people contributed to the collection of the data used in this work. We would like to thank our colleagues at Cascadia Research Collective, the Scripps Whale Acoustics Lab, and The National University of Singapore's Marine Mammal Research Laboratory who provided visual confirmations on our sightings, especially John Calambokidis, Dominique Camacho, Greg Campbell, Stephen Claussen, Annie Douglas, Erin Falcone, Greg Falxa, Andrea Havron, Allan Ligon, Megan McKenna, Yeo Kian Peen, Jen Quan, Nadia Rubio, Greg Schorr, Charles Speed, and Michael Smith, also the crews of Cal-COFI, the R/V Sproul, the R/P Flip, and the R/V Zenobia. We also thank Greg Campbell and Liz Henderson for their help with sighting data and array configurations, and Chris Garsha, Brent Hurley, and Sean Wiggins for hardware support. Data collection was conducted with assistance from John Hildebrand and was supported by the U.S. Navy Environmental Readiness Division, Frank Stone and Ernie Young, and analysis and algorithm development was supported by the Office of Naval Research, Mike Weise and Jim Eckman.

Adam, O. (**2008**). "Segmentation of killer whale vocalizations using the Hilbert-Huang transform," EURASIP J. Adv. Signal Process. doi: 10.1155/2008/245936.

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (**2002**). "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Trans. Signal Process. **50**(2), 174–188.

Barbarossa, S., Scaglione, A., and Giannakis, G. B. (**1998**). "Product high-order ambiguity function for multicomponent polynomial-phase signal modeling," IEEE Trans. Signal Process. **46**(3), 691–708.

Brown, J. C., and Miller, P. J. O. (**2007**). "Automatic classification of killer whale vocalizations using dynamic time warping," J. Acoust. Soc. Am. **122**(2), 1201–1207.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for *Tursiops truncatus* signature whistles," J. Acoust. Soc. Am. **94**(5), 2497–2506.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (**1990**). *Introduction to Algorithms* (MIT Press, Cambridge, MA), p. 1028.

Datta, S., and Sturtivant, C. (**2002**). "Dolphin whistle classification for determining group identities," Signal Processing **82**(2), 127–327.

Dierckx, P. (**1993**). *Curve and Surface Fitting with Splines* (Oxford Science Publications, Oxford), p. 285.

Dillon, W. R., and Goldstein, M. (**1984**). *Multivariate Analysis, Methods and Applications* (Wiley, New York), p. 587.

Doucet, A., de Freitas, N., and Gordon, N. (**2001**). "An introduction to sequential Monte Carlo Methods," in *Sequential Monte Carlo Methods in Practice*, edited by A. Doucet, N. De Freitas, and N. Gordon (Springer, New York), p. 581.

Fisher, F. H., and Spiess, F. N. (**1963**). "FLIP-FLoating Instrument Platform," J. Acoust. Soc. Am. **35**(10), 1633–1644.

Gillespie, D., Gordon, J., McHugh, R., McLaren, D., Mellinger, D. K., Redmond, P., Thode, A., Trinder, P., and Deng, X.-Y. (**2008**). "PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans," Proc. Inst. Acoustics.

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (**1993**). Novel-approach to nonlinear non-Gaussian Bayesian state estimation. IEE Proc. F **140**(2), 107–113.

Halkias, X. C., and Ellis, D. P. W. (**2006**). "Call detection and extraction using Bayesian inference," Appl. Acoust. **67**(11-12), 1164–1174.

Heimlich, S., Klinck, H., and Mellinger, D. K. (**2011**). *The Moby Sound Database for Research in the Automatic Recognition of Marine Mammal Calls*, http://www.mobysound.org/ (Last viewed on April 1, 2011).

Heyning, J. E., and Perrin, W. F. (**1994**). "Evidence for two species of common dolphins (genus *Delphinus*) from the eastern North Pacific," Contr. Sci (Los Angeles) **442**, 1–35.

Ioana, C., Gervaise, C., Stéphan, Y., and Mars, J. I. (**2010**). "Analysis of underwater mammal vocalizations using time-frequency-phase tracker," Appl. Acoust. **71**(11), 1070–1080.

Kitagawa, G. (**1996**). "Monte Carlo filter and smoother for non-gaussian nonlinear state space models," J. Comput. Graph. Stat. **5**(1), 1–25.

Lammers, M. O., Au, W. W. L., and Herzing, D. L. (**2003**). "The broadband social acoustic signaling behavior of spinner and spotted dolphins," J. Acoust. Soc. Am. **114**(3), 1629–1639.

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**(2), 1159–1170.

Marques, T. A., Thomas, L., Ward, J., DiMarzio, N., and Tyack P. L. (**2009**). "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales," J. Acoust. Soc. Am. **125**(4), 1982–1994.

Mellinger, D. K. (**2001**). Ishmael 1.0 User's Guide. NOAA PMEL, Seattle, OAR-PMEL-120, p. 30.

Musso, C., Oudjane, C., and LeGland, F. (**2001**). "Improving regularized particle filters," in *Sequential Monte Carlo Methods in Practice*, edited by A. Doucet, N. De Freitas, and N. Gordon (Springer, New York), pp. 247–721.

Nilsson, N. J. (**1980**). *Principles of Artificial Intelligence* (Tioga, Palo Alto, CA), p. 476.

Oswald, J. N., Barlow, J., and Norris, T. F. (**2003**). "Acoustic identification of nine delphinid species in the eastern tropical Pacific ocean," Mar. Mammal Sci. **19**(1), 20–37.

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**(1), 587–595.

Papoulis, A. (**1991**). *Probability, Random Vriables, and Stochastic Processes* (McGraw-Hill, New York), p. 666.

Press, W. H. (**1992**). *Numerical Recipes in C: the Art of Scientific Computing* (Cambridge University Press, Cambridge), p. 994.

Shapiro, A. D., and Wang, C. (**2009**). "A versatile pitch tracking algorithm: From human speech to killer whale vocalizations," J. Acoust. Soc. Am. **126**(1), 451–459.

Shi, Y., and Chang, E. (**2003**). "Spectrogram-based formant tracking via particle filters," Intl. Conf. Acoust., Speech, Signal Proc. (ICASSP), Hong Kong, China, pp. I-168–I-171.

Wang, D., Wursig, B., and Evans, W. (**1995**). "Comparisons of whistles among seven odontocete species," in *Sensory Systems of Aquatic Mammals*, edited by R. A. Kastelein, J. A. Thomas, P. E. Nachtigall, (De Spil, Woerden, NL), pp. 299–323.

Watkins, W. A. (**1967**). "The harmonic interval: Fact or artifact in spectral analysis of pulse trains," in *Symposium on Marine Bio-Acoustics*, edited by W. N. Tavolga (Pergamon Press, New York), pp. 15–43.

White, P. R., and Hadley, M. L. (**2008**). "Introduction to particle filters for tracking applications in the passive acoustic monitoring of cetaceans," Can. Acoust. **36**(1), 146–152.