Tracking time differences of arrivals of multiple sound sources in the presence of clutter and missed detections

Pina Gruden, Eva-Marie Nosal and Erin Oleson

Citation: The Journal of the Acoustical Society of America **150**, 3399 (2021); doi: 10.1121/10.0006780
View online: https://doi.org/10.1121/10.0006780
View Table of Contents: https://asa.scitation.org/toc/jas/150/5
Published by the Acoustical Society of America

JASA ARTICLE

Tracking time differences of arrivals of multiple sound sources in the presence of clutter and missed detections[a)]

Pina Gruden,[1,b)] Eva-Marie Nosal,[2] and Erin Oleson[3]

[1]*Joint Institute for Marine and Atmospheric Research, Research Corporation of the University of Hawai'i, Honolulu, Hawaii 96822, USA*

[2]*Ocean and Resources Engineering, University of Hawai'i at Mānoa, Honolulu, Hawaii 96822, USA*

[3]*Pacific Islands Fisheries Science Center, National Oceanic and Atmospheric Administration (NOAA), Honolulu, Hawaii 96818, USA*

ABSTRACT:

Acoustic line transect surveys are often used in combination with visual methods to estimate the abundance of marine mammal populations. These surveys typically use towed linear hydrophone arrays and estimate the time differences of arrival (TDOAs) of the signal of interest between the pairs of hydrophones. The signal source TDOAs or bearings are then tracked through time to estimate the animal position, often manually. The process of estimating TDOAs from data and tracking them through time can be especially challenging in the presence of multiple acoustically active sources, missed detections, and clutter (false TDOAs). This study proposes a multi-target tracking method to automate TDOA tracking. The problem formulation is based on the Gaussian mixture probability hypothesis density filter and includes multiple sources, source appearance and disappearance, missed detections, and false alarms. It is shown that by using an extended measurement model and combining measurements from broadband echolocation clicks and narrowband whistles, more information can be extracted from the acoustic encounters. The method is demonstrated on false killer whale (*Pseudorca crassidens*) recordings from Hawaiian waters.

 https://doi.org/10.1121/10.0006780

(Received 27 January 2021; revised 18 August 2021; accepted 19 August 2021; published online 5 November 2021)

[Editor: Marie Roch] Pages: 3399–3416

I. INTRODUCTION

Passive acoustic monitoring can be an important addition to the traditional visual-based line transect methods for estimating the abundance of marine mammals. This is especially relevant for elusive species when groups are frequently missed by visual observers, species with complex group structure, and species that show behavioral response to the boat presence.[1] Acoustic line transect methods incorporate a towed hydrophone array and often rely on estimating the time difference of arrival (TDOA) of acoustic calls recorded on hydrophone pairs. TDOA tracks are created by connecting multiple TDOA estimates from a given source across multiple time steps, and the resulting tracks are used to estimate the animal position and/or bearing.[2–4] The decision of which track to assign a newly calculated TDOA estimate to is often performed manually and can be especially difficult when multiple vocalizing sources are present. This paper presents a multi-target tracking (MTT) method for automated TDOA tracking from acoustic recordings of both narrowband and broadband signals such as the whistles and echolocation clicks produced by delphinid species. The method is demonstrated on recordings of false killer whales (*Pseudorca crassidens*) obtained during line transect surveys in the Hawaii exclusive economic zone (EEZ), USA.

The Hawaiian Archipelago is home to three distinct false killer whale populations, one of which appears to be at a high risk of extinction.[1,5,6,59] False killer whales are highly vocal, producing broadband echolocation clicks[7,8] and narrowband whistles.[9,10] Due to significant differences in the signal properties, tracking and localization is typically carried out separately based either on whistles or echolocation clicks but generally not on both. However, as the types and rates of vocalizations depend on the animal behavioral context,[11] it is expected that a better understanding of a given acoustic encounter can be gained by performing the tracking task based on combined whistle and click information.

To obtain TDOA estimates, signals of interest are typically first detected, and TDOA estimates are obtained using methods such as the standard cross-correlation (SCC),[12,13] cross-correlation of spectrograms,[14–16] rhythm analysis,[2,3] or direct difference of arrival times.[17] An alternative approach to estimate TDOAs from the data, which avoids the detect-first paradigm, is to construct cross-correlograms and track the TDOAs as slowly varying peaks through time.[17] Cross-correlograms can be computed using generalized cross-correlation (GCC) methods, which offer advantages over SCC when estimating time delays for the narrowband signals.[18]

Depending on the bandwidth of the signal of interest and how many active sources are present, the cross-correlation function will exhibit multiple peaks per time step. Some of these peaks are due to true sources, whereas others are due to spurious peaks called clutter. These

[a)]This paper is part of a special issue on Machine Learning in Acoustics.

[b)]Electronic mail: pgruden@hawaii.edu

spurious peaks can result from incorrect associations between the direct and multipath arrivals of a signal or incorrect associations between direct signals from different sources. A range of methods have been proposed to reduce/eliminate clutter, including classification based on different properties of the direct and multipath clicks;[19] considering the variance of the arrival angles and rhythm analysis;[2,3] assigning clicks from individual animals to click trains;[20] considering the cross-correlation of sequences of calls instead of individual calls;[21] or considering only the TDOAs that achieve a high correlation score.[15]

Once clutter has been eliminated, the remaining peaks in the cross-correlation function are assumed to belong to multiple true sources and must be connected into multiple TDOA tracks. These multiple TDOA tracks have been previously estimated using various approaches such as the manual analysis of bearing-time scatterplots;[22] automated analysis based on properties of the signals of interest;[17] and a traditional MTT approach based on multiple hypothesis tracking.[23] Although the automated methods have advantages over the manual analysis, such as being less time-consuming and producing more objective results, they can suffer from computational limitations when dealing with high clutter density and many closely spaced targets.[24]

We propose to cast the TDOA estimation problem into a more general statistical MTT framework, where clutter (false TDOAs), missed detections (when a target is present but no measurement is collected), and sources' appearance and disappearance are incorporated in the problem formulation.[24,25] This framework is based on random finite sets (RFSs) and is computationally less demanding and more accurate than traditional MTT methods.[24] Although the fundamental ideas of RFS can be applied to a variety of problems, each problem requires a careful engineering of the models and assumptions that drive the filters implemented within the RFS.

A frequently used filter within the RFS framework is the probability hypothesis density (PHD) filter,[24,26] a computationally tractable approximation of the multi-target Bayes filter that propagates only the first-order moment of the multi-target posterior.[24] Detailed discussions can be found in Refs. 24 and 27 The PHD filter has been previously used to track multiple targets in sonar[28,29] and video,[30] multiple speakers in reverberant environments,[25,31] and multiple overlapping dolphin whistles in spectrograms.[32,33]

Although the problems of whistle tracking[32,33] and TDOA tracking appear similar, there are key differences that require the filters to be derived and defined differently. These differences pertain to what is measured: a whistle is produced by only one animal, whereas a TDOA track is produced by a group of likely multiple animals, resulting in a much higher track variance in the TDOA domain. Moreover, in whistle tracking, missed detections are caused by the signal amplitude falling below a threshold and, generally, span only a few time steps. Meanwhile, missed detections in TDOA tracking can also be the result of animals not vocalizing during certain periods of time, resulting in increased track fragmentation due to multiple, prolonged periods of missed detections. Thus, although whistle tracking filters can extract meaningful tracks with less informative measurements based on only one feature (in this case, frequency),[32,33] TDOA tracking filters require more informative measurements with additional features for successful tracking.

It has been shown that an improved tracking performance can be achieved with the PHD filter by incorporating the amplitude information to the measurements.[34,35] However, the formulation presented in Refs. 34 and 35 performs a joint update step for both newly appearing and persistent targets, which can bias the number of estimated targets.[36] In the formulation proposed in Ref. 36, the newborn and persistent targets are updated separately, but it does not incorporate the amplitude information. In this paper, we derive a PHD filter formulation that incorporates the amplitude information for more informative measurements and improved tracking and updates persistent and newborn targets separately for reduced bias in the number of estimated targets.

The novel contributions of this work are the following: (i) we present a modified PHD filter formulation to incorporate the amplitude information and separately update newborn and persistent targets; (ii) we present a target birth model that incorporates the measurement amplitude information to better inform the appearance of new targets; (iii) we propose a framework for combining the measurements from broadband echolocation clicks and narrowband whistles to reduce fragmentation and improve TDOA tracking with the extended PHD filter; and (iv) we apply the proposed extended PHD filter to track multiple sources in simulated and measured towed array data. The paper is structured as follows. Section II describes the preprocessing method to obtain the measurements. Section III describes the proposed filter and models for TDOA tracking. The method is demonstrated and compared to a simpler PHD filter formulation on simulated data in Sec. IV and real data in Sec. V. The discussion and conclusions can be found in Secs. VI and VII, respectively.

II. GCC AND CROSS-CORRELOGRAMS

Estimating TDOAs from the cross-correlation of narrowband signals is challenging for multiple reasons. For example, in Hawaii, false killer whale whistles have a narrow bandwidth of about 600 Hz and occur between about 2.5 to 12 kHz (based on unpublished NOAA data).[60] Because the width of the cross-correlation peak is proportional to the signal bandwidth, W, approximately $1/W$, the peak is wide for narrowband signals. Moreover, the cross-correlation function has a quasiperiodic nature when computed from narrowband signals: the main peak occurs at the time delay, and additional peaks occur with a period of $2\pi/\omega_0$, where ω_0 is the signal center frequency.[38] If the signal bandwidth is a small fraction of the center frequency (i.e., $W/\omega_0 \ll 1$), which is the case in false killer whale

https://doi.org/10.1121/10.0006780

whistles, the adjacent peaks have very nearly equal height, and identifying the largest peak requires either a very large signal-to-noise ratio (SNR) or exceedingly long observation times.[37]

The true peak in the cross-correlation function can be enhanced by choosing an appropriate frequency weighting function in the GCC method such as a smoothed coherence transform (SCOT).[18,38,39] Sharpening the peak is useful, especially in cases with multiple time delays (from multiple sources), because sharper peaks allow better distinction. However, sharp peaks are more sensitive to errors introduced by the finite observation time, particularly in cases of low SNR.[38]

Limiting the frequency bands to those in which the signals of interest occur can improve the SNR. When there are multiple types of overlapping signals present in the recordings, it can also help to treat these separately. For example, when narrowband whistles and broadband echolocation clicks occur together, it can help to filter out the clicks, which otherwise dominate the cross-correlation function because clicks are broadband and high amplitude. For many odontocetes, clicks can be efficiently filtered out by band limiting the frequencies of interest. However, false killer whale echolocation clicks exhibit significant energy below 20 kHz (Ref. 8) and, thus, overlap in frequency with the whistles. Hence, a different filtering strategy is required.

In this work, we first remove echolocation clicks with a three-stage click removal process. Stage one employs a fourth-order bandpass Butterworth filter with cut-off frequencies of 2.5 and 12 kHz. The filter is applied in a forward and reverse fashion to preserve the phase response. Stage two employs an adaptive weighting to the filtered time-domain signal.[32,40] Stage three applies median background subtraction across the frequency bins (61 point median filter) to the frequency domain signal.[32,40] This de-clicking procedure does not affect the time delay estimates, and a result of this procedure is shown in Fig. 1. A sample code for this process is available in Ref. 41.

After de-clicking the full time-domain signal, the generalized cross-correlation smoothed coherence transform (GCC-SCOT) is computed in 1 s long sliding windows. The windows have a 50% overlap, resulting in 0.5 s time steps. The window length was chosen to include full whistle contours; false killer whale whistles are typically 0.44 ($\pm$0.22) s in duration.[9] The TDOA resolution for this method depends on the sampling frequency (f_s) and is $1/f_s$, which for our data ($f_s = 500$ kHz) yields 2 μs.

To improve the SNR, only frequencies between 2.5 and 12 kHz (typical for Hawaiian false killer whales) were used to compute the GCC-SCOT. Other frequencies are set to zero. We used envelopes of the GCC-SCOT, computed with the Hilbert transform,[42] to provide the time delay estimates rather than the raw cross-correlation.

The result of these processing steps is a cross-correlogram consisting of the envelope, A_{xy}, of the GCC-SCOT per each time step for a given sensor pair. Because the cross-correlogram consists of the envelope information, its probability density function (pdf) can be described by a Rayleigh distribution,[42] which is parameterized by its variance σ_r^2. To simplify the expressions for the amplitude pdfs in Sec. III B 4, this cross-correlogram is normalized so that the segments containing background noise only will have $\sigma_r = 1$.

Using cross-correlation methods to estimate the time delay for the broadband signals, such as clicks, is more straightforward than for narrowband signals such as whistles. Because the bandwidth of broadband signals is large, the cross-correlation peak is sharp, and the cross-correlation function is not oscillatory as it is for narrowband signals; this is true even for SCC. Thus, the processing scheme that we developed for whistles also works for clicks with some modification. The signals are first bandpassed with a fourth-order Butterworth filter with cut-off frequencies of 8 and 30 kHz. Cross-correlograms are then computed based on the GCC-SCOT with a 1 s long sliding window and 50% overlap. The window length and overlap were kept the same as

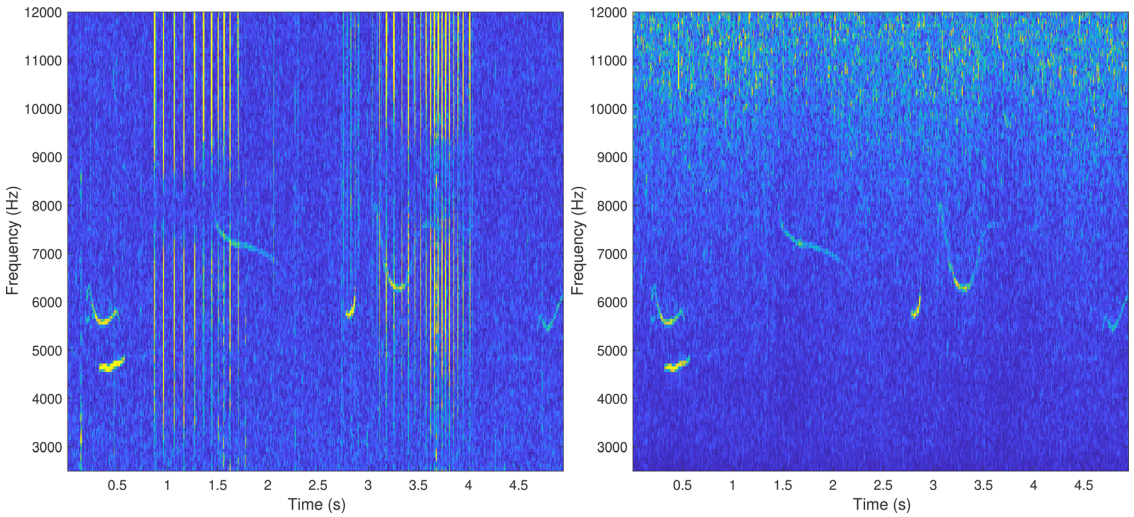

FIG. 1. (Color online) Spectrograms of a false killer whale recording ($f_s = 500$ kHz, 8192 point Hanning window, 50% overlap). The raw signal containing clicks and whistles (left) and the de-clicked signal (right) are shown.

for whistles to allow us to combine the measurements for the joint tracking of the different signal types. Although a 1 s window typically contains multiple clicks in each frame, leading to multiple peaks in the cross-correlation function, only one of the peaks dominates the function for a given source. Moreover, the spurious peaks will not form a coherent track over multiple time steps, whereas the dominant (correct) peaks will. As they were for whistles, the cross-correlograms based on the clicks are normalized so that the Rayleigh parameter for noise σ_r is unitary.

Both of the resulting normalized cross-correlograms (based on whistles and clicks) are used to obtain the measurements from which TDOA tracks are extracted as discussed in Sec. III.

III. GAUSSIAN MIXTURE PROBABILITY HYPOTHESIS DENSITY (GM-PHD) FOR TDOA TRACKING

Assuming Gaussianity of the noise processes and linearity of the underlying models, the analytic solution to the PHD filter[24,26] is the Gaussian mixture probability hypothesis density (GM-PHD) filter.[43] Intuitively, the PHD, $v_{k|k}(\boldsymbol{x})$, is a function whose peaks indicate the likely target positions and whose integral gives the expected number of targets in a given region of the state space.[43] In the GM-PHD approximation to the PHD filter, the PHD function is approximated with a mixture of weighted Gaussian components that are propagated recursively through time. The means and covariances of the components are predicted and updated with the Kalman filter,[44] whereas the weights are predicted and updated with the PHD filter equations.[43] New components, representing new targets, are introduced to the recursion through a birth PHD, which also assumes that the PHD for target births is a Gaussian mixture.[43] Moreover, when new components are initiated based on the available measurements, the PHD equations need to be modified so that the newborn and persistent components are predicted and updated separately.[36] To maintain computational feasibility, pruning and merging techniques are employed at the end of each time step.[43] This is followed by the estimation of multiple target states from the posterior PHD, which is achieved by taking the Gaussian components with the weights above some threshold.[43] To maintain continuity of the target tracks, individual labels are propagated along with the components.[32,45]

The GM-PHD filter requires specification of the underlying models and parameters. These are application specific, and different models will affect the filter's performance. A better performance in terms of distinguishing targets from clutter can be achieved when the GM-PHD filter measurement model is extended to include the amplitude information.[34,35] However, this requires the PHD equations to be re-derived. While Refs. 34 and 35 derive the PHD filter that incorporates the amplitude information, they update newborn and persistent targets jointly, which can bias the number of estimated targets.[36] In Ref. 36, the newborn and persistent targets are updated separately, but they do not derive the updated equations that would incorporate the amplitude information. In this paper, we derive the PHD filter that incorporates the amplitude information and updates persistent and newborn targets separately. The derivation is shown in the supplementary material.[46] For brevity, the discussion in this paper is limited to the parts of the model, prediction, and update steps that are affected by including the amplitude feature. The pruning, merging, and state estimation parts of the filter are not changed from a standard GM-PHD implementation[43] and still controlled by merging U, pruning T_r, and weight w_{th} thresholds and the maximum number of components per time step $J_{\max}$.

This section is organized as follows. The training data for certain GM-PHD filter parameters and models is described in Sec. III A, and the specific models and parameters are discussed in Sec. III B. The extended GM-PHD equations with the amplitude feature, which include a separate update for newborn and persistent targets, are discussed in Sec. III C. The parameters used in the filter are summarized in Table I.

A. Training data

Certain parameters in the GM-PHD need to be specified based on the characteristics of the system under consideration, and training data is used to learn these parameters. For the purpose of obtaining the training data, a 2.2 h long encounter containing false killer whale vocalizations was considered from the data collected with a linear towed array (see Sec. V for details). This training encounter is different from the test encounter used in Sec. V and was hand-annotated as follows. A cross-correlogram was formed as described in Sec. II for whistles. The annotations contained 34 TDOA tracks in total with a median track duration of 2.6 min (interquartile range of 10.8 min) and some tracks as long as 61.7 min. Because the TDOAs in these tracks did not occur at regular, fixed intervals and they are noisy, they were first fitted with a polynomial and then interpolated to obtain a sample for every time step. The polynomial model

TABLE I. Summary of parameters used in the GM-PHD-SA filter for TDOA tracking. p_S and p_D^{λ} denote the probabilities of the survival and detection respectively; U, T_r, and w_{th} denote the merging, pruning, and weight thresholds respectively; $J_{\max}$ denotes the maximum allowed number of Gaussian components in one iteration; ν_b denotes the expected number of newborn targets; σ_v^2 and R denote the system and measurement noise variances, respectively; r_k denotes the clutter rate; (d_1,d_2) denote the lower and upper expected SNR values for the targets (linear scale); and the symbol "*" denotes parameters learned from the training data.

p_S	p_D^{λ}	U	T_r	w_{th}	$J_{\max}$	ν_b	σ_v^2 [(s/s^2)^2]*	R [s^2]*	r_k*	(d_1,d_2)*
0.99	0.4	4	0.001	0.1	100	0.0005	1.3×10^{-9}	4.5×10^{-8}	1	(3.16,100)

https://doi.org/10.1121/10.0006780

order was selected for each track based on the goodness of fit criteria—root mean squared error (RMSE). To prevent overfitting, the lowest model order at which the RMSE stopped changing significantly was selected, the residuals plot was examined, and the significance of the coefficients was evaluated by examining the confidence intervals. The polynomially fitted and interpolated TDOA tracks, denoted by z_p, together with their corresponding cross-correlogram containing the background noise and clutter information became our training data. From this data, the first- and second-order derivatives ($\dot{z}_p$ and $\ddot{z}_p$) of z_p were obtained.

B. Models for TDOA tracking

Most target tracking techniques are model based and rely on two basic models, collectively known as the state space models: a system (or dynamic) model that describes target motion and a measurement (or observation) model that relates the noisy measurements to the target states, i.e., the full descriptions of targets.[47] The PHD filter requires additional models that govern the clutter (false alarms) and birth (appearance) of new targets, which is discussed below.

1. System model

In our application, the aim is to track multiple TDOAs from different sources on a cross-correlogram. The state vector $\boldsymbol{x}$, thus, consists of the TDOA information (τ) and rate of change of TDOA ($\dot{\tau}$),

$$\boldsymbol{x} = [\tau, \dot{\tau}]^T, \tag{1}$$

where $[\cdot]^T$ denotes the transpose. The variables τ and $\dot{\tau}$ can be interpreted as a source position and source velocity in TDOA space, respectively.

It is assumed that the target state develops according to the nearly constant velocity (NCV) model,[47] also referred to as a discrete white noise acceleration model.[48] The second-order derivative of the position (acceleration) is assumed to be a zero-mean random process, and the system model is

$$\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{x}_{k-1} + \boldsymbol{n}_{k-1} = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \boldsymbol{x}_{k-1} + \boldsymbol{n}_{k-1}, \tag{2}$$

where subscripts k and $k-1$ denote the current and previous time steps, respectively, Δ denotes the time interval between the overlapping windows, and $\boldsymbol{n}_{k-1}$ is the zero-mean white noise process with covariance $\boldsymbol{Q}$, which can be expressed as[48]

$$\boldsymbol{Q} = \begin{bmatrix} \frac{1}{4}\Delta^4 & \frac{1}{2}\Delta^3 \\ \frac{1}{2}\Delta^3 & \Delta^2 \end{bmatrix} \sigma_v^2, \tag{3}$$

where σ_v is the standard deviation of the process noise and its physical dimension is that of acceleration.

The choice for the value of σ_v should be on the order of the maximum acceleration magnitude, α_M, and a practical range should be $0.5\alpha_M \leq \sigma_v \leq \alpha_M$.[48] The value of α_M for our application was learned from the hand-annotated data (described in Sec. III A) to be 7.3×10^{-5} and, therefore, $\sigma_v = 0.5\alpha_M = 3.7 \times 10^{-5}$ s/s^2.

2. Measurement model

Although acoustic line transect surveys typically use only timing information for tracking,[4] having more informative measurements can be beneficial. For example, adding amplitude information as a measurement improved the tracking performance in radar and sonar applications.[34,35,49] When track formation is based on the consistency of amplitude returns (in addition to the consistency of target motion), better distinction between targets and clutter is achieved.[35] Although our application is based on passive (not active) acoustics, it is still reasonable to expect that the amplitude of the peaks in the cross-correlation associated with the targets will be relatively consistent and higher in amplitude than peaks resulting from clutter.

The measurements in this study are obtained by finding all of the local maxima in the cross-correlograms (computed as described in Sec. II) above a threshold λ in each time step k. Thus, the measurements consist of the measured TDOA information, z, and amplitude, a, of A_{xy}:

$$\tilde{\boldsymbol{z}} = [z, a]^T, \tag{4}$$

and the measurement model can be written as

$$\tilde{\boldsymbol{z}}_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & \sigma_r^2 \end{bmatrix} \tilde{\boldsymbol{x}}_k + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \eta_k, \tag{5}$$

where σ_r^2 is the background noise variance (normalized to unity), $\tilde{\boldsymbol{x}}$ denotes the augmented state, $\tilde{\boldsymbol{x}} = [\boldsymbol{x}^T, d]^T$, and d denotes the expected SNR. Note that $\tilde{\boldsymbol{x}}$ is constructed here for the sake of derivation; however, only $\boldsymbol{x}$ is directly propagated through time with the GM-PHD filter. As d is not known, in practice, it is marginalized out from the final equations as shown below. The measurement noise, η_k, is assumed to be independent Gaussian white noise with a variance R.

For certain applications, the variance of the TDOA measurement can be expressed analytically as a function of the SNR, bandwidth, integration time, and center frequency of the signal.[50] For the towed arrays, the uncertainty in the TDOA measurement is also affected by the uncertainty in the hydrophone position.[51] If the hydrophone displacements can be measured or estimated, the variance of the TDOA can be computed analytically.[51] However, estimating the displacement requires position sensors in the array, which increases the cost and complexity of the system and is often not available. Moreover, delphinid groups consist of multiple closely swimming animals, which also affect the accuracy of the TDOA measurement. Thus, R was learned from the hand-annotated data, described in Sec. III A, as follows. First, the absolute difference between the hand-annotated

and training data was computed. Second, the median absolute deviation (MAD), which is robust to outliers, was computed to be 2.1×10^{-4} s. This resulted in $R = \text{MAD}^2 = 4.5 \times 10^{-8}$ s^2.

To determine how likely it is that a measurement is due to a target, the target measurement likelihood function is required and is determined from Eq. (5). If we assume that a is independent of z, then the target likelihood function $\tilde{g}_k(\tilde{z}|\tilde{\boldsymbol{x}})$ is[35]

$$\tilde{g}_k(\tilde{z}|\tilde{\boldsymbol{x}}) = g_k(z|\boldsymbol{x})g_a(a|d), \tag{6}$$

where $g_k(z|\boldsymbol{x})$ denotes the target likelihood based on the TDOA information and $g_a(a|d)$ denotes the target amplitude likelihood function. For clarity, the subscript k is left implicit in the function arguments.

The target amplitude likelihood $g_a(a|d)$ is defined in Sec. III B 4, and $g_k(z|\boldsymbol{x})$ is

$$g_k(z|\boldsymbol{x}) = \mathcal{N}(z; \hat{z}, S), \tag{7}$$

where $\mathcal{N}(\cdot; \hat{z}, S)$ denotes a Gaussian density function with mean $\hat{z}$ and covariance S.

Further, measurements are obtained given a certain probability of detection, p_D, which can be trained from the data[32,33] when no prior knowledge is available or it can be approximated analytically. In sonar and radar,[35,49] the probability of detection given a threshold λ can be modeled as the function of the SNR only, $p_D^\lambda(d)$. However, in passive acoustic applications based on biological signals, the overall probability of detection is likely a product of two components: detectability and availability.[52] Detectability is the probability that the animal is detected given that it produces a vocalization. This concept is similar to the way in which the probability of detection is used in sonar, and it depends on the SNR. Availability is the probability that an animal will produce a vocalization and requires knowledge of the vocalization rate. Because vocalization rates are not available for false killer whales, we take p_D^λ to be lower than what is expected based on the SNR alone, which would be $p_D^\lambda = 0.63$; see Eqs. (17) and (18) in Ref. 34 for calculation, assuming values $\lambda = 3.7$ and SNR between 5 and 20 dB (see Sec. III B 4). We used $p_D^\lambda = 0.4$ as an empirically determined conservative estimate.

3. Clutter model

As measurements originate from targets and clutter, the PHD filter framework requires a model for clutter that defines how likely it is that a measurement is the result of clutter (i.e., clutter likelihood function), how much clutter is expected, and how clutter is distributed across the observation space.

Assuming that a is independent of z, the likelihood function for clutter $\tilde{c}_k(\tilde{z})$ is[35]

$$\tilde{c}_k(\tilde{z}) = c_k(z)c_a(a), \tag{8}$$

where $c_k(z)$ denotes the clutter likelihood based on the TDOA information, and $c_a(a)$ denotes the clutter amplitude likelihood function.

The clutter amplitude likelihood $c_a(a)$ is discussed in Sec. III B 4. In this work, we assume a uniform distribution of clutter in TDOA space, therefore, the clutter is not dependent on the measurement, i.e., $c_k(z) = c_k$, and c_k is

$$c_k = \mathcal{U}(-\tau_m, \tau_m), \tag{9}$$

where $\mathcal{U}(A, B)$ denotes a uniform distribution between the parameters A and B, and $\tau_m = d_s/c$ with sensor separation d_s and speed of sound c (1500 m/s).

The number of clutter points per time step is assumed to be Poisson distributed and is drawn from a Poisson distribution parameterized by a clutter rate, r_k. The clutter rate was learned from the training data (Sec. III A) to be $r_k = 1$.

4. Amplitude feature likelihood models

To use the amplitude likelihoods in practice, the type and shape of the distribution must be determined. It can be shown that the envelope of a narrowband random signal is Rayleigh distributed[42] and is a function of only its variance σ_r^2. Thus, a Rayleigh distribution can be used to describe $g_a(a|d)$ and $c_a(a)$.[34,35] Assuming that the background noise is normalized ($\sigma_r = 1$) and the measurements are obtained by considering peaks (local maxima) in the A_{xy} above a threshold λ, one can express $c_a(a)$ as[35]

$$c_a^\lambda(a) = a \exp\left(\frac{\lambda^2 - a^2}{2}\right), \quad a \geq \lambda. \tag{10}$$

The threshold, λ, is typically determined for a specified value of false alarms, p_{FA}, and given the assumptions above (normalized background and Rayleigh distribution), the threshold is $\lambda = \text{sqrt}(-2 * \log(p_{FA}))$ [obtained by rearranging Eq. (5) in Ref. 34]. In this work, we chose $p_{FA} = 0.001$; thus, $\lambda = 3.7$.

The amplitude likelihood for targets, $g_a(a|d)$, is dependent on the expected SNR, d. However, when the expected SNR is not known, the parameter d can be marginalized over a range of possible values $[d_1, d_2]$ and the expression for the amplitude likelihood for the targets, which is not dependent on d, can be written as[34,35]

$$g_a(a) = \frac{2\left(\exp\left(\frac{-a^2}{2(1+d_2)}\right) - \exp\left(\frac{-a^2}{2(1+d_1)}\right)\right)}{a(\ln(1+d_2) - \ln(1+d_1))}. \tag{11}$$

In this work, the values for $[d_1, d_2]$ are learned from the data as follows. The training data, described in Sec. III A, contains TDOA and amplitude information from a normalized cross-correlogram. The SNR is computed as $\text{SNR} = a/E_N$, where a denotes the target cross-correlation amplitudes (a squared quantity proportional to the hydrophone signal energy), and E_N is the total background noise energy. The background noise energy is estimated by computing the median value across all time steps t of the cross-correlogram, A_{xy}, resulting in a vector of noise energies per

https://doi.org/10.1121/10.0006780

TDOA. The total background noise energy, E_N, is then obtained as the median value of this vector,

$$E_N = \underset{\tau}{\text{median}}\left(\underset{t}{\text{median}}\left(A_{xy}(t,\tau)\right)\right). \tag{12}$$

The resulting SNR values were distributed between 5 and 20 dB, which is equivalent to $[d_1, d_2] = [3.16, 100]$.

5. Appearance of new targets

New targets are introduced to the recursion with a birth PHD. The birth PHD is a density with peaks that corresponds to the positions that the targets are likely to appear in at a given time step. It should be defined in a way that covers the region in which targets are expected to appear.[36] In some applications, the birth region is assumed to be known *a priori* and is concentrated to small specific areas.[43] When the birth regions are unknown, it is advantageous to base the birth PHD on the measurements.[36] Basing the birth PHD on the measurements ensures that the components get introduced where the likelihood of a new target appearing is high. Moreover, it is often useful to assume that not every measurement is equally likely to initialize a newborn target.[32,33] In this study, the measurements that are more likely to initialize a newborn target are the ones with a higher amplitude.

We obtain newborn Gaussian components from the measurements as follows. A Gaussian mixture, $p(\boldsymbol{x})$, is constructed at each time step, k, from the measurements, $\tilde{\boldsymbol{z}} \in \tilde{\boldsymbol{Z}}_k$, and prior information on the rate of change of the TDOA, $\dot{z}_p$, which was obtained from the training data. The weights, $w^{(j)}$, of the mixture are set proportional to the measured amplitudes, $a^{(j)}$, such that $w^{(j)} = a^{(j)} / \sum_{j=1}^{|\tilde{\boldsymbol{Z}}_k|} a^{(j)}$, where $|\tilde{\boldsymbol{Z}}_k|$ denotes the cardinality of the measurement set at time k. The means of the mixture are $\boldsymbol{m}^{(j)} = [z^{(j)}, \dot{z}_p^{(j)}]^T$, where $\dot{z}_p^{(j)}$ denotes the rate of change of TDOA drawn from a prior $f(\dot{z}_p)$. The covariance, $\boldsymbol{P}^{(j)}$, is the same for all components in the mixture and is equal to $\boldsymbol{Q}$ in Eq. (3) with the off diagonal elements set to zero. Formally,

$$p(\boldsymbol{x}) = \sum_{j=1}^{|\tilde{\boldsymbol{Z}}_k|} w^{(j)} \mathcal{N}\left(x; \boldsymbol{m}^{(j)}, \boldsymbol{P}^{(j)}\right). \tag{13}$$

The prior, $f(\dot{z}_p)$, was learned from the training data (Sec. III A) by fitting $\dot{z}_p$ with a Gaussian mixture model (GMM).[53] The model order was selected based on the Bayesian information criterion,[53] and $f(\dot{z}_p)$ was determined to be a mixture of four Gaussians,

$$f(\dot{z}_p) = \sum_{n=1}^{4} b^{(n)} \mathcal{N}(x; \mu^{(n)}, \Sigma^{(n)}), \tag{14}$$

where $b^{(n)}$, $\mu^{(n)}$, and $\Sigma^{(n)}$ are the weights, means, and variances of the GMM, respectively. For our dataset, these parameters were $\boldsymbol{b} = [0.17, 0.06, 0.14, 0.63]$, $\boldsymbol{\mu} = [-5.2 \times 10^{-9}, 9.6 \times 10^{-5}, 4.6 \times 10^{-5}, 5.9 \times 10^{-6}]$, and $\boldsymbol{\Sigma} = [3.1 \times 10^{-16}, 4.1 \times 10^{-9}, 4.3 \times 10^{-10}, 2.9 \times 10^{-11}]$.

The birth PHD is constructed by summing N_b weighted Gaussian components and has the form

$$\gamma_k(\boldsymbol{x}) = \sum_{i=1}^{N_b} w_{k,b}^{(i)} \mathcal{N}\left(x; \boldsymbol{m}_{k,b}^{(i)}, \boldsymbol{Q}^{(i)}\right). \tag{15}$$

Note that the weights of the newborn components, $w_{k,b}^{(i)}$, sum to the expected number of newborn targets per time step, ν_b, thus, γ_k is a density but not a pdf. The means, $\boldsymbol{m}_{k,b}^{(i)}$, of newborn components are drawn from $p(\boldsymbol{x})$ [Eq. (13)]. The covariances are equal to $\boldsymbol{Q}$ [Eq. (3)]. Their weights, $w_{k,b}^{(i)}$, are proportional to the weights in $p(\boldsymbol{x})$ and, thus, also proportional to the amplitudes, and set to the sum of the expected number of newborn targets, which we set to $\nu_b = 0.0005$. Note that this weight assignment makes the filter biased toward higher amplitude measurements, which are more likely originating from the targets. Hence, ν_b can be set to a lower value and still allows for successful tracking.

C. GM-PHD filter with amplitude feature (GM-PHD-SA)

The PHD and, consequently, the GM-PHD equations need to be modified when using an extended measurement model with the amplitude feature as demonstrated in Refs. 34 and 35. Additionally, when the birth PHD is based on the measurements, the prediction and update steps need to be performed separately for the persistent and newborn targets[36] to avoid biasing the number of estimated targets. We present combined GM-PHD equations that take into account both the extended measurement model and the separate prediction and update for the newborn and persistent targets, henceforth, referred to as the GM-PHD-SA filter ("S" stands for the separate prediction and update for the newborn and persistent targets, and "A" stands for the amplitude). Detailed derivations are given in the supplementary material.[46]

The PHD, $v_{k|k}(\boldsymbol{x})$, can be expressed as two PHDs, one for persisting targets, $v_{k|k,p}(\boldsymbol{x})$, and one for newborn targets, $v_{k|k,b}(\boldsymbol{x})$. In this paper, the subscripts $k|k-1$ and $k|k$ are used to indicate the predicted and updated elements, respectively. The subscripts p and b are used to denote the persistent and newborn targets, respectively. The probability of a target surviving from one step to another is assumed to be state independent and the same for the persistent and new targets, $p_{S,p}(\boldsymbol{x}) = p_{S,b}(\boldsymbol{x}) = p_S$.

The prediction step of the GM-PHD-SA filter stays unmodified from the standard filter.[43] The predicted PHD for the persistent, $v_{k|k-1,p}(\cdot)$, and newborn, $v_{k|k-1,b}(\cdot)$, targets consists of Gaussian components and can be written as[34,36,43]

$$v_{k|k-1,p}(\boldsymbol{x}) = \sum_{i=1}^{J_{k-1}} w_{k|k-1}^{(i)} \mathcal{N}\left(x; \boldsymbol{m}_{k|k-1}^{(i)}, \boldsymbol{P}_{k|k-1}^{(i)}\right), \tag{16}$$

$$v_{k|k-1,b}(\boldsymbol{x}) = \gamma_k(\boldsymbol{x}), \tag{17}$$

where $w^{(i)}_{k|k-1} = p_S w^{(i)}_{k-1}$, the means, $\boldsymbol{m}^{(i)}_{k|k-1}$, and covariances, $\boldsymbol{P}^{(i)}_{k|k-1}$, of the persistent targets are predicted with a Kalman filter,[43,44] $\gamma_k(\cdot)$ denotes the birth PHD defined in Eq. (15), and J_{k-1} denotes the number of Gaussian components (persistent and newborn from the previous time step).

The update step of the GM-PHD-SA filter is modified to incorporate amplitude information and is performed separately for the persistent and newborn targets. The updated PHD for persistent targets, $v_{k|k,p}(\cdot)$, becomes

$$v_{k|k,p}(\boldsymbol{x}) = \left[1 - p^{\lambda}_D\right] v_{k|k-1,p}(\boldsymbol{x}) + \sum_{\tilde{z}\in\tilde{\boldsymbol{Z}}^{\lambda}_k} \sum_{j=1}^{J_{k-1}} w^{(j)}_{k|k}(\tilde{z}) \mathcal{N}\left(x; \boldsymbol{m}^{(j)}_{k|k}(\tilde{z}), \boldsymbol{P}^{(j)}_{k|k}\right), \quad (18)$$

where $1 - p^{\lambda}_D$ represents a probability of the missed detection, and $\tilde{z}$ denotes a measurement in a measurement set, $\tilde{\boldsymbol{Z}}^{\lambda}_k$, above a threshold, λ. The means, $\boldsymbol{m}_{k|k}$, and covariances, $\boldsymbol{P}_{k|k}$, of the updated mixture are calculated from $v_{k|k-1,p}(\boldsymbol{x})$ with the Kalman filter update,[43,44] and the updated weights $w^{(j)}_{k|k}(\tilde{z})$ are calculated as

$$w^{(j)}_{k|k}(\tilde{z}) = \frac{w^{(j)}_{k|k-1} g_a(a) g^{(j)}_{k,p}(z|\boldsymbol{x})}{\mathcal{L}(\tilde{z})} \quad (19)$$

and

$$\mathcal{L}(\tilde{z}) = r_k c_k c^{\lambda}_a(a) + g_a(a) \sum_{l=1}^{N_b} w^{(l)}_{k,b} g^{(l)}_{k,b}(z|\boldsymbol{x}) + g_a(a) \sum_{l=1}^{J_{k-1}} w^{(l)}_{k|k-1} g^{(l)}_{k,p}(z|\boldsymbol{x}), \quad (20)$$

where the first term in Eq. (20) relates to the clutter with the clutter rate, r_k, and the clutter likelihood, c_k, and amplitude likelihood for clutter, $c^{\lambda}_a(a)$, are defined in Eqs. (9) and (10), respectively. The second term relates to N_b newborn targets, where $g_a(a)$ is the amplitude likelihood for the targets defined in Eq. (11). $g_{k,b}(\cdot)$ denotes the target likelihood function for the newborn targets in Eq. (7), where $\hat{z} = \boldsymbol{H}\boldsymbol{m}^{(l)}_{k,b}$ is the predicted measurement, $S = \boldsymbol{H}\boldsymbol{Q}^{(l)}\boldsymbol{H}^T + R$ is the innovation, and $\boldsymbol{H} = [1, 0]$. The third term relates to J_{k-1} persistent targets, where $g_{k,p}(\cdot)$ denotes the target likelihood for the persistent targets in Eq. (7), where $\hat{z} = \boldsymbol{H}\boldsymbol{m}^{(l)}_{k|k-1}$ and $S = \boldsymbol{H}\boldsymbol{P}^{(l)}_{k|k-1}\boldsymbol{H}^T + R$.

Because the newborn targets are initiated based on the measurements, they are assumed to be always detected, i.e., $p^{\lambda}_D = 1$. The update for newborn targets, $v_{k|k,b}(\boldsymbol{x})$, becomes

$$v_{k|k,b}(\boldsymbol{x}) = \sum_{\tilde{z}\in\tilde{\boldsymbol{Z}}^{\lambda}_k} \sum_{i=1}^{N_b} w^{(i)}_{k|k,b}(\tilde{z}) \mathcal{N}\left(x; \boldsymbol{m}^{(i)}_{k|k,b}(\tilde{z}), \boldsymbol{P}^{(i)}_{k|k,b}\right). \quad (21)$$

For the newborn targets, the means and covariances are computed from $\gamma_k(\boldsymbol{x})$ with the Kalman filter update,[44] and the updated newborn weights $w^{(i)}_{k|k,b}$ are calculated as

$$w^{(i)}_{k|k,b}(\tilde{z}) = \frac{w^{(i)}_{k,b} g_a(a) g^{(j)}_{k,b}(z|\boldsymbol{x})}{\mathcal{L}(\tilde{z})}. \quad (22)$$

Intuitively, what is gained by using the GM-PHD-SA filter can be understood if one considers the terms that relate to the amplitude likelihoods in Eq. (19) and rewrites the equation as

$$w^{(j)}_{k|k}(\tilde{z}) = \frac{w^{(j)}_{k|k-1} g^{(j)}_{k,p}(\cdot)}{\frac{c^{\lambda}_a(a)}{g_a(a)} r_k c_k + \sum_{l=1}^{N_b} w^{(l)}_{k,b} g^{(l)}_{k,b}(\cdot) + \sum_{l=1}^{J_{k-1}} w^{(l)}_{k|k-1} g^{(l)}_{k,p}(\cdot)}, \quad (23)$$

where the amplitude likelihood ratio, $c^{\lambda}_a(a)/g_a(a)$, controls which term dominates the denominator. When the amplitude of a measurement is large, $a \gg 1$ (i.e., when the measurement originates from a target), then $c^{\lambda}_a(a)/g_a(a) \ll 1$, and the terms related to the newborn and persistent targets dominate. In this case, the weight is mainly determined by $g_k(\cdot)$, i.e., the TDOA part of the measurement only. When the amplitude of a measurement is very small, $a \ll 1$ (i.e., when the measurement originates from a clutter), then $c^{\lambda}_a(a)/g_a(a) \gg 1$, and the clutter term dominates. In this case, the weight is determined based on both the amplitude and TDOA information.

IV. SIMULATIONS

We use simulation to demonstrate that the GM-PHD-SA filter described in Sec. III can extract multiple TDOA tracks immersed in clutter and evaluate its performance against a simpler version of the filter (discussed below). The number of targets in the simulation varies randomly between one and seven, and they are observed in the clutter over the region of TDOA space $[-0.02, 0.02]$ s, which corresponds to a 30 m sensor separation. Note that negative TDOA values correspond to sources ahead of the array, and positive TDOA values correspond to the sources behind the array. The target states are TDOA and rate of change of the TDOA [Eq. (1)] and are assumed to evolve according to the constant velocity (CV) model in Eq. (2), where $\sigma^2_v = 1.3 \times 10^{-9}$. At each time step, each target has the survival probability $p_S = 0.99$ and probability of detection $p^{\lambda}_D = 0.4$. The targets are assumed to appear according to the model described in Sec. III B 5.

Measurements contain two features, TDOA and amplitude information, and are obtained every $\Delta = 0.5$ s. The TDOA part of the measurements for the targets is simulated based on the real-world hand-annotated TDOA tracks with the noise variance set to $R = 4.5 \times 10^{-8}$. The TDOA part of the measurements for the clutter is simulated based on a uniform distribution in TDOA space [Eq. (9)]. The number of clutter points has a Poisson distribution, and two different clutter rates, r_k, are considered, $r_k = 1$ and $r_k = 10$, per time step to simulate the lower and higher clutter scenarios. The amplitude part of the measurements for the targets and

https://doi.org/10.1121/10.0006780

clutter is simulated by randomly drawing from the pdfs in Eqs. (11) and (10), respectively, with $\lambda = 3.7$ and the expected target SNR bounds $[d_1, d_2] = [5, 20]$ dB.

One hundred different cases are simulated for a given clutter rate, and the filter performance is quantified by precision, recall, coverage, fragmentation, and mean deviation as described in Ref. 54. These are typically used to evaluate the whistle tracking performance[32,33,40,54] and measure the quantity and quality of the detected tracks. Briefly, precision measures the percentage of the detections that are correct, recall measures the percentage of the expected detections that are retrieved, coverage measures the average percentage of a ground truth track that is detected, fragmentation measures the average number of detections per ground truth track, and mean deviation measures the average deviation between the ground truth track and its corresponding detection(s).

Filter performance is also benchmarked against a simpler implementation of the GM-PHD filter, henceforth referred to as the GM-PHD-S filter, which includes separate predictions and updates for the persistent and newborn targets but does not use the amplitude feature in the measurements.[32] Note that because the GM-PHD-S filter has no amplitude information to inform the newborn weights, these are set to be the same for all newborn targets, i.e., $w_{k,b}^{(i)} = \nu_b / N_b$. Unlike the GM-PHD-SA, the newborn weights in the GM-PHD-S are evenly distributed between all of the measurements (targets and clutter). As a consequence, if ν_b is set too low, all of the newborn tracks will decay faster and not pass the state estimation stage. Therefore, GM-PHD-S requires a higher value of ν_b than GM-PHD-SA does to successfully extract the tracks. Note, if ν_b was left the same as for the GM-PHD-SA filter, GM-PHD-S would extract no tracks. Thus, the parameter ν_b had to be adjusted to $\nu_b = 0.005$ to improve the performance. The rest of the parameters remained the same as those for the GM-PHD-SA filter.

The results of the tracking on the 100 simulated cases for each r_k are shown in Table II, and it can be seen that the GM-PHD-SA filter outperforms the GM-PHD-S filter. When the clutter rate is low ($r_k = 1$), the GM-PHD-SA filter has a recall, coverage, and mean deviation similar to that of the GM-PHD-S filter, but precision is significantly better and fragmentation is lower. When the clutter rate is high ($r_k = 10$), the performance of the GM-PHD-SA filter remains stable, but the performance of the GM-PHD-S filter deteriorates. Two examples, one for each clutter rate, and corresponding tracking by both filters are shown in Fig. 2. In each example, six TDOA tracks are present, some of which overlap. As expected from the results in Table II, the GM-PHD-SA filter's performance remains stable regardless of the amount of clutter in the measurements [Figs. 2(b) and 2(e)], whereas the GM-PHD-S filter's performance deteriorates significantly with higher clutter rate [Figs. 2(c) and 2(f)].

V. REAL DATA

This section applies the proposed framework to field acoustic recordings and combines the measurements originating from different signal types to improve the tracking and insight about the acoustic encounter.

Data were collected during the Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS) in 2017.[55] A linear hydrophone array sampling at $f_s = 500$ kHz was towed at 10 kn, 335 m behind the vessel, at depths between 10 and 15 m. We used two hydrophones separated by $d_s = 31.1$ m in this study (type HTI-96-min, Long Beach, MI, and a combined sensitivity with custom built preamplifiers of -144 dB $\pm$ 5 dB re 1 V/μPa from 2 to 100 kHz and approximately linear roll-off to -156 dB $\pm$ 2 dB re 1 V/μPa at 150 kHz). A representative encounter with false killer whales was chosen, which included multiple subgroups of animals that were reasonably well separated in the TDOA space and contained a good representation of the different types of TDOA tracks. This encounter is different from the encounter used for the filter parameter training in Sec. III A.

Data were processed as described in Sec. II to obtain the normalized cross-correlograms based on the whistles and clicks. Then the measurements, $\tilde{z}_k = [z_k, a]$, were obtained by finding all of the peaks in the normalized cross-correlograms above the threshold $\lambda = 3.7$. In addition to two measurement sets (one based on the whistles and one on the echolocation clicks), a third measurement set was formed by combining the extracted peaks from the whistle and click cross-correlograms. The combined measurements are shown in Fig. 3.

The measurement sets were hand annotated to obtain the ground truth information. In the example considered here, three subgroups are present, two in which animals whistle and echolocate and one in which animals echolocate only [Fig. 3(a)]. In some subgroups, the measurements show a large variance around the mean group trajectory. The TDOA track from the echolocate-only subgroup crosses the

TABLE II. Performance [median (interquartile range)] of the GM-PHD-S and GM-PHD-SA filters on the 100 cases of simulated data for two different clutter rates (r_k). *R* denotes recall, *P* is the precision, *Cover* denotes the coverage, *Frag* denotes the fragmentation, and *μDev* denotes the mean deviation from the ground truth data.

Filter type	r_k	*R* (%)	*P* (%)	Cover (%)	Frag	μDev (s)
GMPHD-S	1	100 (0)	6.2 (4.3)	93.6 (7.6)	2.6 (1.1)	1.3×10^{-4} (5.5×10^{-5})
GMPHD-SA		100 (0)	100 (13.4)	93.8 (6.8)	1 (0.3)	1.3×10^{-4} (5.7×10^{-5})
GMPHD-S	10	7.1 (33.3)	50 (100)	48.7 (99.6)	0.5 (1)	5.7×10^{-5} (3.5×10^{-4})
GMPHD-SA		100 (0)	100 (0)	92.7 (10.8)	1.2 (0.3)	1.4×10^{-4} (5.4×10^{-5})

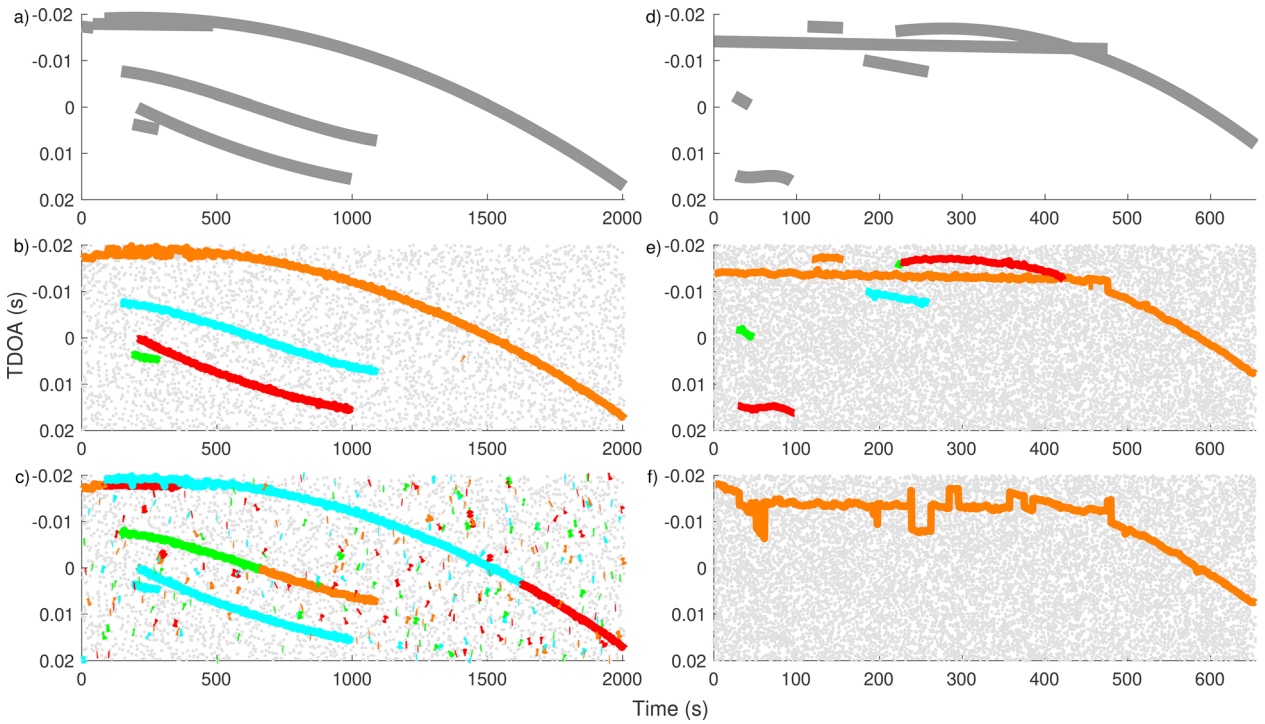

FIG. 2. (Color online) Two example simulations for different clutter rates r_k and the corresponding GM-PHD-SA and GM-PHD-S detections. [(a)–(c)] Simulation with $r_k = 1$ and [(d)–(f)] simulation with $r_k = 10$. Six ground truth tracks (gray lines) are present in both examples and are shown in (a) for $r_k = 1$ and (d) for $r_k = 10$. GM-PHD-SA detections (colored lines) are shown in (b) for $r_k = 1$ and (e) for $r_k = 10$. Note that different colors denote different detections (i.e., track fragments). GM-PHD-S detections (colored lines) are shown in (c) for $r_k = 1$ and (f) for $r_k = 10$. All TDOA measurements (clutter and target) are denoted by gray dots.

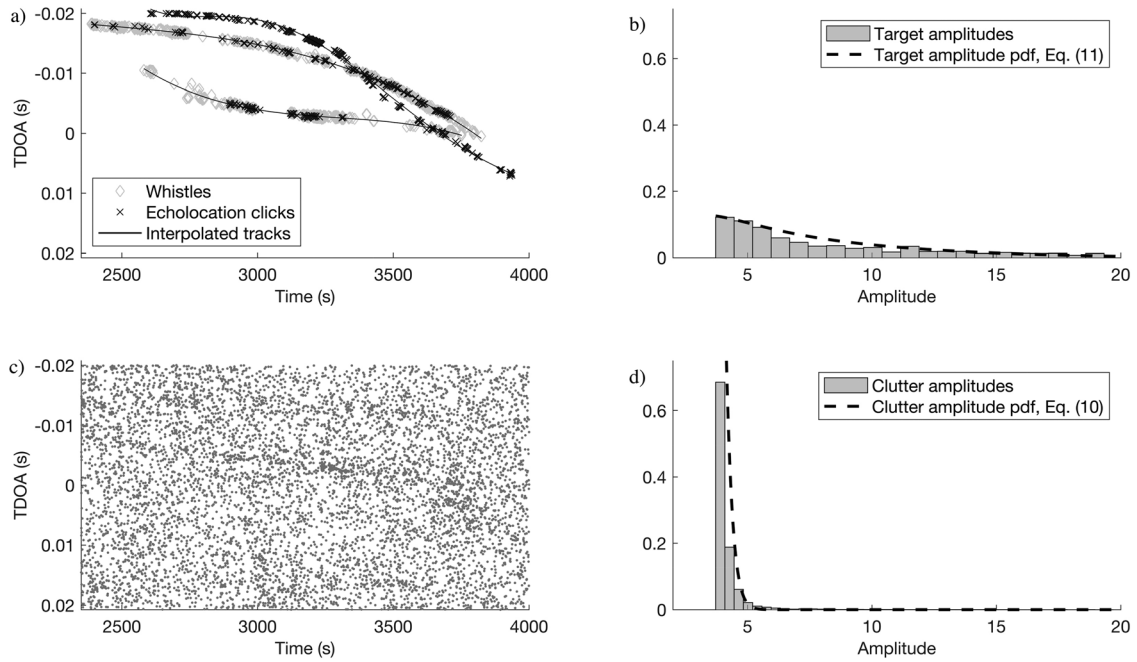

FIG. 3. Measurements obtained from hand-annotating cross-correlograms (based on clicks and whistles) of a false killer whale encounter. The top row shows the measurements associated with the targets. (a) TDOA measurements derived from whistle (gray diamonds) and echolocation click (black ×) cross-correlograms with corresponding interpolated ground truth tracks (black lines); and (b) amplitudes of the cross-correlation measurements with corresponding target amplitude pdf [Eq. (11)]. The bottom row shows the measurements associated with clutter. (c) TDOA measurements derived from both types of cross-correlograms; and (d) amplitude of the cross-correlation measurements with corresponding clutter amplitude pdf [Eq. (10)].

https://doi.org/10.1121/10.0006780

tracks of the other two subgroups. Each hand-annotated track was fitted with a polynomial and interpolated to obtain a ground truth TDOA track for each subgroup. The polynomial orders examined were between one and nine. The best polynomial order for each track was selected as discussed in Sec. III A. The best order ranged between two and five, depending on the measurement set under consideration. For the combined measurement set, the order was three for the two subgroups that produced clicks and whistles and five for the subgroup that only echolocated. The interpolated ground truth TDOA tracks based on the combined measurements are shown in [Fig. 3(a)]. The measurements contain a high amount of clutter [Fig. 3(c)]. The amplitude measurements are also shown [Figs. 3(b) and 3(d)], and they follow the assumed amplitude pdfs from Eqs. (10) and (11), respectively.

The GM-PHD-SA filter (Sec. III, Table I) was used to track TDOAs of the subgroups from the measurements. The filter performance was quantified using precision, recall, coverage, fragmentation, and mean deviation from the ground truth track.[54] To investigate the feasibility of the filter for TDOA tracking from acoustic recordings, four scenarios are considered. To facilitate comparisons with the other filters, the measurements, annotated, and interpolated data for each of these scenarios are included in the supplementary material.[46]

In the first scenario, the TDOA measurements from a single subgroup of the whistling and clicking animals were augmented with simulated clutter; i.e., a single target immersed in simulated clutter. The clutter was simulated with a uniform distribution in TDOA space [Eq. (9)], where the number of clutter points have a Poisson distribution with a clutter rate of $r_k = 1$ per time step. The clutter amplitude was simulated by randomly drawing from the pdf in Eq. (10), where $\lambda = 3.7$. The GM-PHD-SA tracking results showed that while the subgroup was tracked when only measurements from either whistles and (to some extent) clicks were available, better results were achieved when both measurements were combined (Fig. 4). In all of the cases, the subgroup was detected in fragments during which the animals were vocally active, but in the combined measurement case, the extracted fragments of the track were longer and better coverage of the TDOA track was achieved compared to when separate measurements were considered (Table III).

In the second scenario, the TDOA measurements from all three of the subgroups were augmented with simulated clutter; i.e., multiple targets immersed in simulated clutter. The clutter was simulated as in the first scenario. The GM-PHD-SA filter successfully tracked all three of the subgroups (Fig. 5). Note that in the whistle measurements, only two subgroups were present, whereas in the click and combined measurements, all three of the subgroups were present. When tracking was based on combined measurements, the extracted fragments of the three subgroups were longer and the precision was higher compared to the tracking based on the measurements from the echolocation clicks only (Table III). When tracking was based on whistle measurements, the precision was better and the coverage was similar compared to the performance on the click and combined measurements (Table III). However, because one of the subgroups did not whistle, the information on that group was missed in this case. In general, the GM-PHD-SA filter maintains the tracks through shorter periods of silence (which

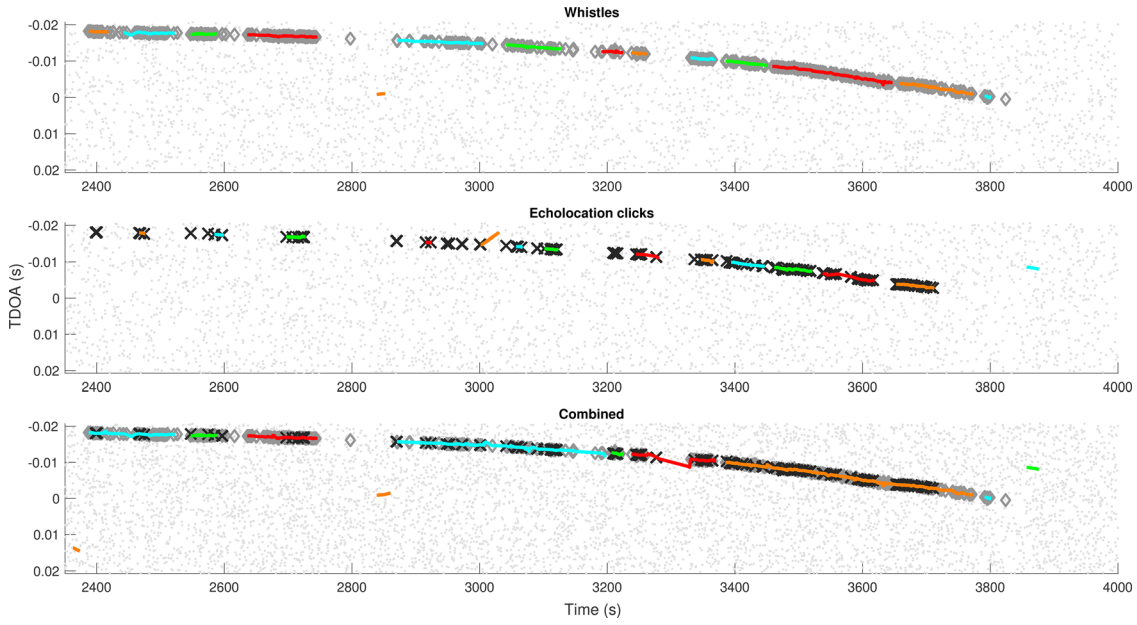

FIG. 4. (Color online) GM-PHD-SA tracking of TDOAs from a subgroup of false killer whales immersed in simulated clutter (scenario 1). Tracking from the whistle (top), echolocation click (middle), and combined (bottom) cross-correlogram measurements. Measurements are denoted by light-gray dots, measurements originating from the subgroup of animals are denoted by dark-gray diamonds (whistles) and black × (clicks), and the estimated GM-PHD-SA tracks are denoted as colored lines. Note that the different colors denote different detections (i.e., track fragments).

TABLE III. Performance of the GM-PHD-SA filter for different scenarios (described in Sec. V). The performance is shown for the measurements originating from whistle (*W*) and click (*C*) cross-correlograms and for the combined measurements (*Comb*). *R* denotes recall, *P* is the precision, *Cover* denotes the coverage, *Frag* denotes the fragmentation, and μDev denotes the mean deviation from the ground truth data. For scenarios 2 and 4, which contain multiple ground truth tracks, the metrics *Cover*, *Frag*, and μDev are stated as the mean value ± standard deviation.

Scenario	Type	*R* (%)	*P* (%)	Cover (%)	Frag	μDev (s)
1	*W*	100	92.9	49.0	13	1.3×10^{-4}
	C	100	92.9	19.0	13	1.5×10^{-4}
	Comb	100	72.7	54.8	8	1.4×10^{-4}
2	*W*	100	100	33.7 ± 18	12.5 ± 2.1	$(1.5 \pm 0.3) \times 10^{-4}$
	C	100	86.8	22.6 ± 6	11.7 ± 7	$(2.2 \pm 1.6) \times 10^{-4}$
	Comb	100	90.5	34.1 ± 16	13.3 ± 2	$(2.2 \pm 1.1) \times 10^{-4}$
3	*W*	100	66.7	46.0	14	1.3×10^{-4}
	C	100	62.5	21.6	15	1.8×10^{-4}
	Comb	100	36.8	51.0	14	1.6×10^{-4}
4	*W*	100	71.1	35.6 ± 13.4	13.5 ± 2.1	$(1.7 \pm 0.5) \times 10^{-4}$
	C	100	67.3	28.0 ± 4.9	12.7 ± 4.5	$(2.5 \pm 0.8) \times 10^{-4}$
	Comb	100	49.3	36.6 ± 18.6	12.3 ± 4.2	$(2.5 \pm 0.8) \times 10^{-4}$

can be considered similar to missed measurements from a filtering perspective), but fragmentation occurs when the period of silence is longer.

In the third scenario, TDOA measurements from a single subgroup of whistling and clicking animals were augmented with real clutter from the corresponding cross-correlogram. Real clutter was defined as all of the measurements $\tilde{z}_k$ above $\lambda = 3.7$, which were not associated with measurements from any of the subgroups. Besides the main subgroup, additional tracks are detected and appear to be false positives (Fig. 6). This lowered the precision compared to when clutter was simulated in the first scenario (Table III). On further investigation, we found that most of the potential false positive tracks were measurements originating from the subgroups that were missed in the hand annotations. The amount of the extracted track was again higher when combined measurements were considered (Table III).

In the fourth scenario, TDOA measurements from all three of the subgroups were augmented with real clutter from the corresponding cross-correlograms, with real clutter obtained as in the third scenario. All three of the subgroups are tracked well (Fig. 7), but the precision was lower compared to when the clutter was simulated in the second scenario (Table III). On further investigation, we again found that most of the potential false positive detections were measurements originating from the animals (some from whistle and some from click cross-correlograms) missed in the hand-annotation process. Tracking based on combined measurements extracted longer portions of

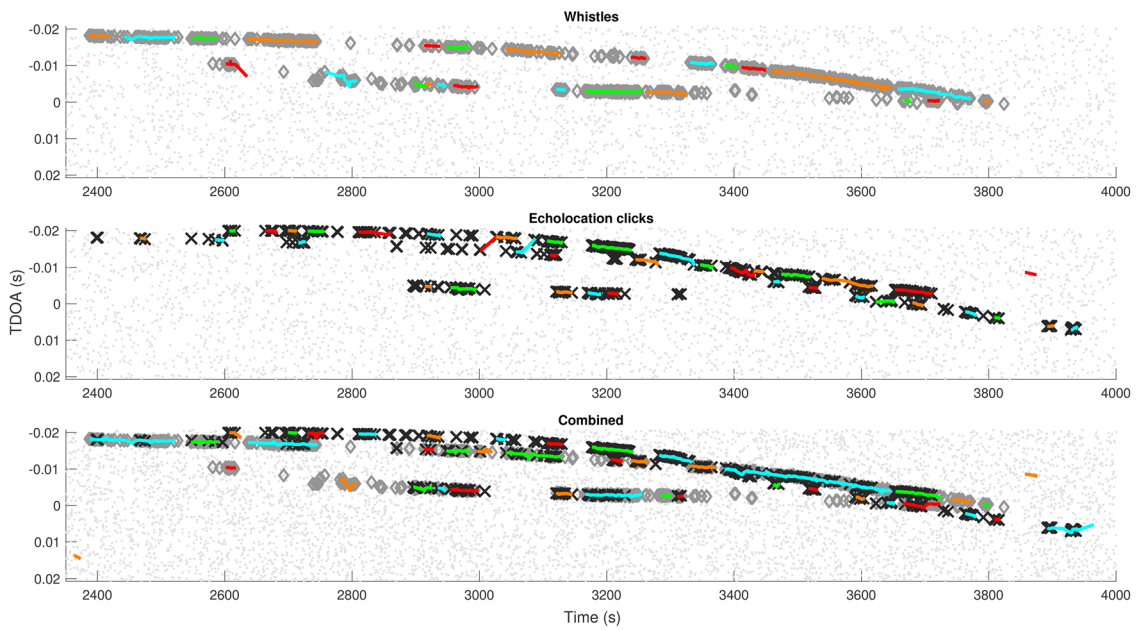

FIG. 5. (Color online) GM-PHD-SA tracking of TDOAs from three subgroups of false killer whales immersed in simulated clutter (scenario 2). The symbols and colors follow the same definitions as in Fig. 4.

https://doi.org/10.1121/10.0006780

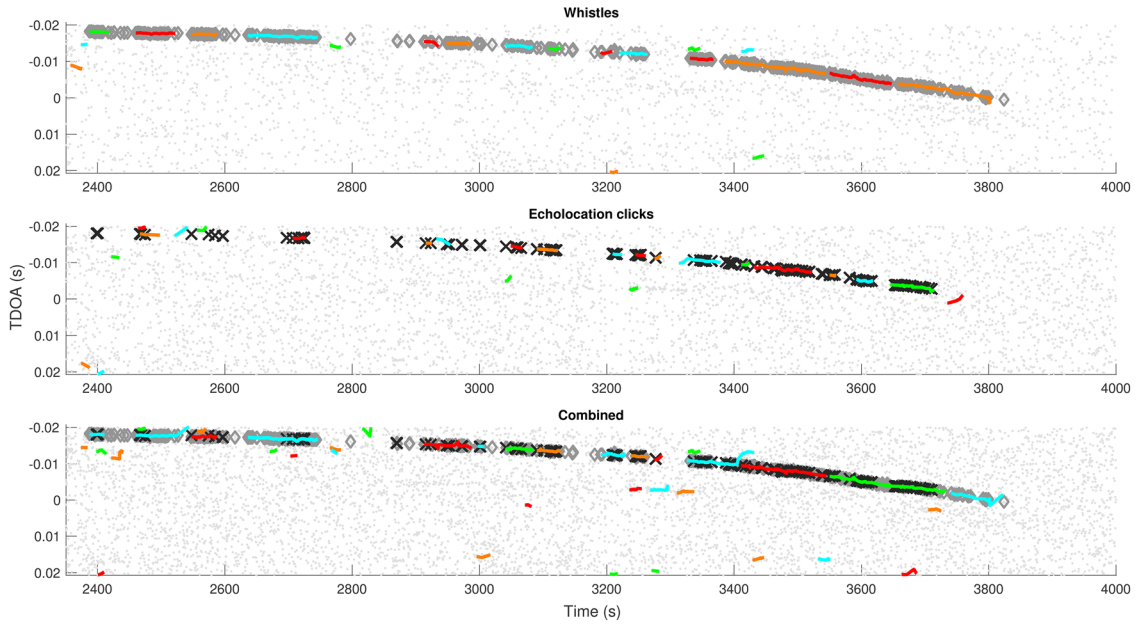

FIG. 6. (Color online) GM-PHD-SA tracking of TDOAs from a subgroup of false killer whales immersed in real clutter (scenario 3). The symbols and colors follow the same definitions as in Fig. 4.

the three subgroups tracks compared to tracking based on measurements from the echolocation clicks only (larger coverage, Table III). An example of track switching can be observed at 3800 s in the combined measurements, where one subgroup ends and another subgroup is in close proximity.

Filter performance was benchmarked against the GM-PHD-S filter. As described in Sec. IV, the newborn weights for the GM-PHD-S filter were set the same for all newborn targets, and the expected number of newborn targets was adjusted to $\nu_b = 0.005$ for better performance. The rest of the parameters remained the same as those for the GM-PHD-SA (Table I).

For all of the scenarios, precision was much lower (due to more false detections) and fragmentation was much higher for the GM-PHD-S filter compared to the GM-PHD-SA filter (Table IV). Recall and mean deviation from the ground truth tracks were similar and coverage was comparable, with GM-PHD-S performing slightly worse in most cases (Table IV). An example of GM-PHD-S tracking for the fourth scenario is shown in Fig. 8, and additional figures

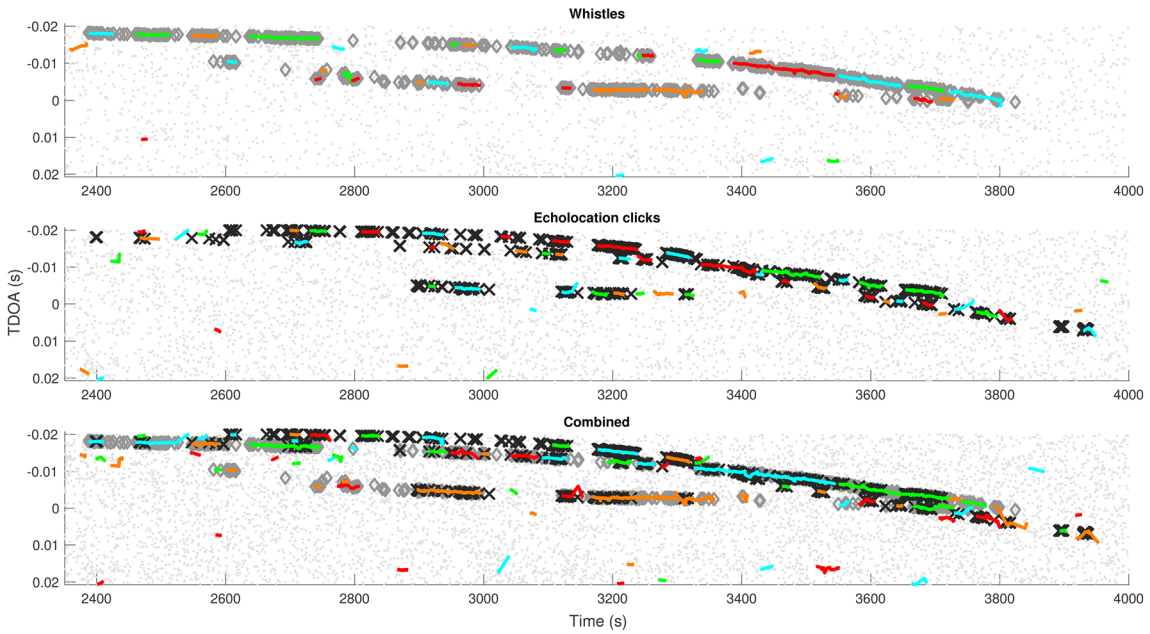

FIG. 7. (Color online) GM-PHD-SA tracking of TDOAs from three subgroups of false killer whales immersed in real clutter (scenario 4). The symbols and colors follow the same definitions as in Fig. 4.

TABLE IV. Performance of the GM-PHD-S filter for different scenarios (described in Sec. V). For a description of the column names and symbols, see Table III.

Scenario	Type	R (%)	P (%)	Cover (%)	Frag	μDev (s)
1	W	100	61.8	46.6	21	1.4×10^{-4}
	C	100	37.0	13.6	10	1.5×10^{-4}
	Comb	100	16.9	55.1	23	1.5×10^{-4}
2	W	100	70.8	33.9 ± 17.8	17.0 ± 5.6	$(1.5 \pm 0.2) \times 10^{-4}$
	C	100	66.7	19.9 ± 6.2	10.7 ± 7.6	$(2.2 \pm 1.1) \times 10^{-4}$
	Comb	100	33.9	39.7 ± 14.1	20.0 ± 3.6	$(2.4 \pm 1.0) \times 10^{-4}$
3	W	100	56.3	40.9	18	1.4×10^{-4}
	C	100	39.5	16.4	17	1.8×10^{-4}
	Comb	100	23.0	52.9	23	1.7×10^{-4}
4	W	100	55.2	37.0 ± 8.0	18.5 ± 0.7	$(1.6 \pm 0.3) \times 10^{-4}$
	C	100	45.0	29.5 ± 6.7	17.3 ± 9.0	$(2.2 \pm 0.7) \times 10^{-4}$
	Comb	100	25.4	44.7 ± 12.0	20.3 ± 1.2	$(2.7 \pm 0.8) \times 10^{-4}$

for the rest of the scenarios can be found in the supplementary material.[46]

VI. DISCUSSION

This study developed a GM-PHD-SA filter to track the TDOAs of subgroups of animals from their acoustic recordings. The cross-correlation amplitude was added as an additional feature to TDOA, and the resulting GM-PHD-SA filter successfully tracked multiple subgroups from the towed array recordings. In contrast to other common approaches used in marine mammal tracking, no prior detection of the signals of interest was required. Moreover, the proposed scheme incorporates multiple vocalization types (whistles and echolocation clicks), allowing longer tracks to be extracted and providing more detailed information about the acoustic encounters. Simultaneously analyzing information from both clicks and whistles could help improve our understanding of the vocalization behavior within subgroups of delphinid species.

By incorporating the amplitude as one of the measurement features and, thus, making the measurements more informative, the performance of the GM-PHD-SA filter was greatly improved compared to that of the benchmark GM-PHD-S filter, which only considered the TDOA information. GM-PHD-SA reported many fewer false targets (Tables III and IV). The difference in performance was even more pronounced on the simulated data (Table II) because as the clutter rate in the measurements increased, the GM-PHD-SA filter performance was not impacted, whereas GM-PHD-S performance deteriorated (recall decreased significantly). In general, the amount of clutter in the measurements will depend on the threshold λ, which is used to obtain the measurements from A_{xy}. The lower the λ, the more clutter

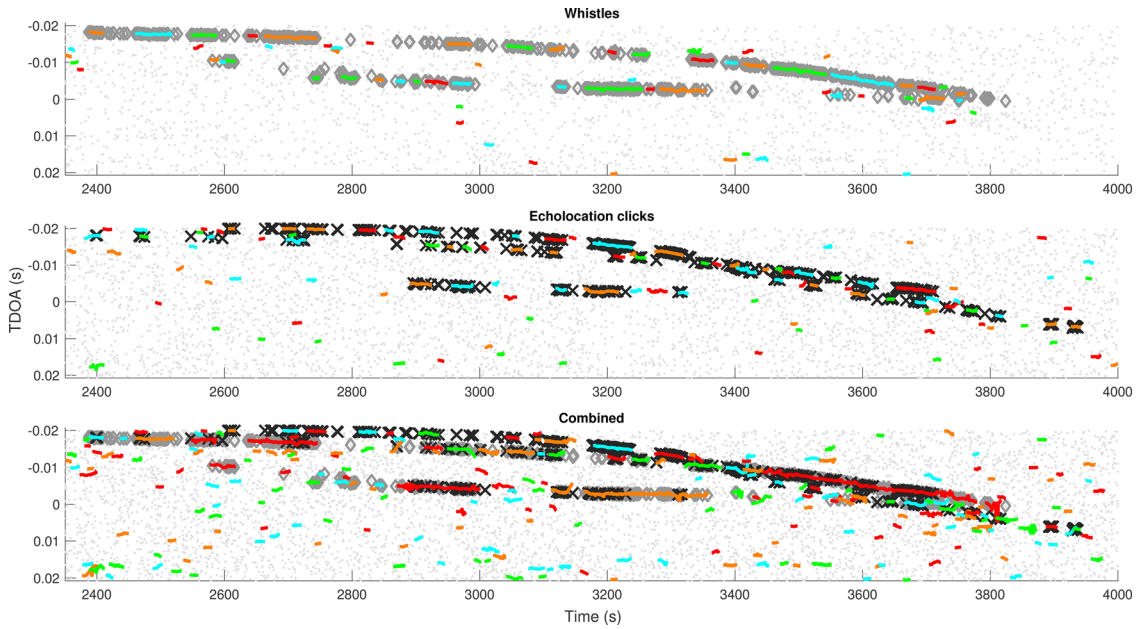

FIG. 8. (Color online) GM-PHD-S tracking of TDOAs from three subgroups of false killer whales immersed in real clutter (scenario 4). The symbols and colors follow the same definitions as in Fig. 4.

https://doi.org/10.1121/10.0006780

measurements, but also more target related measurements will be present. Having more target measurements could improve the coverage (amount of a track extracted), whereas having more clutter measurements has a potential to decrease precision due to increased false positive detections. However, when more simulated clutter measurements were present in the data, the GM-PHD-SA precision did not change (Tables II and III, scenario 2, where measurements based on the combined data contain twice as much clutter compared to the measurements based on either whistles or clicks). On the contrary, having more simulated clutter measurements present decreased the precision of the GM-PHD-S filter significantly (Table IV, scenario 2) and deteriorated its recall (Table II). On the other hand, if higher λ would be imposed on A_{xy}, this would result in fewer measurements overall (both clutter and target). Having fewer target related measurements would result in a potential decrease in the performance for both of the filters. Having fewer (or no) clutter related measurements would result in a comparable performance between the two filters because the amplitude feature would no longer be meaningful (if all measurements originate from targets, all have high amplitudes).

In this study, the amplitudes of the measurements were assumed to follow Rayleigh pdfs, and a good agreement between the analytical pdfs and real data was obtained (Fig. 3). However, it should be noted that the real clutter contained some measurements that had high amplitude values, and this resulted in decreased precision of the GM-PHD-SA filter in the scenarios with real clutter (Table III, scenarios 3 and 4) compared to scenarios with simulated clutter (Table III, scenarios 1 and 2). Some of these higher amplitude measurements resulted from targeted animals that were missed during the annotation process, and some could also be the result of other correlated sources in the environment.

Certain model parameters were obtained from the training data (Table I), which was relatively small in the current study. In the future, larger training datasets should be considered, either by the hand annotation of the field data or, potentially, by constructing them from existing data with the help of the deep learning procedures.[56] The parameters reported in this paper were trained based on the whistle cross-correlograms and used for tracking all of the measurement types, including echolocation clicks and combined clicks and whistles. The parameters related to the evolution of the TDOA tracks are expected to be similar between the whistle or click cross-correlograms. However, certain parameters, such as the clutter rate and SNR range, might differ significantly when different measurement types are considered. Performance on the click and combined measurements may have suffered as a consequence. Although the proposed parameters appear to perform sufficiently well, a detailed optimization is still needed and would ideally include all of the GM-PHD-SA filter parameters (Table I). Future studies could use the Bayesian optimization or similar to determine the optimal parameter values by minimizing a performance metric. While in the present study, the performance was quantified with metrics typically used for whistle contour tracking,[32,40,54] it is difficult to optimize the parameters based on multiple competing metrics. Thus, a single metric, such as an optimal sub-pattern assignment (OSPA)[57] and its variants, could be used. The OSPA is typically used for MTT based on RFSs and combines the information on cardinality, localization, and labelling errors, which, in essence, includes all of the metrics measured in this study. In addition to the detailed parameter optimization, a sensitivity study[33] of the filter to these optimized parameter values should be performed in future work. Future studies should also evaluate GM-PHD-SA performance on a larger variety of acoustic encounters.

Comparison to ground truth data is required to evaluate filter performance. However, hand-annotated ground truth data is often subjective and prone to errors as seen in Sec. V, Figs. 6 and 7. Moreover, the definition of the ground truth data must be carefully considered as there is no unique understanding of what ground truth should be, and favoring one definition over others implies some expectations on the filter performance. For example, when tracking subgroups of animals containing multiple individual animals, should the ground truth constitute the mean TDOA trajectory of all of the animals in the subgroup or something else? Nor is it clear whether one ground truth track should consist of multiple, separate track segments when the animals are vocally active or whether it should incorporate multiple vocalizations and the large periods of silence in between them as a single track. This translates to whether it is reasonable to expect the filter to track through long gaps (on the order of tens of seconds) during which the animals do not vocalize but move significantly, or whether it is more reasonable to expect the filter to interrupt tracking through these long silent periods. In the present study, the ground truth data were defined as the mean trajectory of the group, obtained by fitting and interpolating the analyst hand-annotated measurements, including through periods of silence. Although this definition of ground truth captures well our intuition of what the "ideal" track should be, in practice, the large silent gaps included in this definition resulted in the low coverage of individual tracks even though the filter successfully tracked through periods when animals were vocally active and through some brief missed detections within those periods. Hence, under this definition of ground truth, the performance metric "coverage" does not adequately describe the filter's ability to continue tracking through the brief missed detections as it is typically understood in the target tracking literature. Instead, the performance is affected by the filter's inability to continue tracking targets through long silent periods, which is a much more demanding task. For comparison, when the tracking was performed on the simulated data (Sec. IV), where p_D^{λ} was low but constant through time (i.e., the animals were assumed to vocalize with a constant rate), the coverage was very high (Table II).

Several challenges were encountered in the current application. All of the major MTT algorithms assume that the detection profile is known *a priori*, and the GM-PHD filter implementation requires the probability of detection to be constant across time.[27] Contrary to typical target tracking

applications where the probability of target detection only depends on the SNR of the target returns, in biological applications, the probability of detection is also a function of animal vocalization rates, which are typically unknown. In addition, the measurements originating from animal subgroups typically occurred in temporal clusters and thus the assumption that p_D^λ is constant in time was not met. This resulted in severe track fragmentation, where the filter was able to track the subgroups through sparse missed detections and shorter periods of silence but did not continue the tracks through longer periods of silence. By combining the measurements based on whistles and clicks, some of these gaps were filled and longer track segments could be extracted. The performance might be further improved by considering a nonconstant probability of detection with a "background-agnostic" PHD filter.[58]

Another challenge resulting from the limited spatial resolution of the measurements was that the individually tracked targets (subgroups) actually consisted of multiple individuals swimming in close proximity. As the animals moved and vocalized, their corresponding TDOA measurements consisted of multiple point detections spread around the subgroup mean position. This meant that the measurement model had to consider a greater noise variance of TDOA measurements than what would be expected if a single source or multiple well separated sources were present. This larger noise variance also makes the tracks more sensitive to clutter and nearby subgroups, leading to tracks diverging from the true positions.

The GM-PHD-SA filter and processing framework were presented for towed arrays with two sensors and applied to false killer whales in this work, but it is suitable for the TDOA tracking of any biological or nonbiological, broadband, or narrowband sources. TDOA tracking is often the first step in the overall signal processing chain that leads to localization. We hope that our method of obtaining automated and improved TDOA estimates will correspond to improved localization estimates in future work. Our approach does not require a specific localization strategy because after the TDOA tracks are obtained, different localization procedures can be used to obtain the animal locations. Future efforts could investigate the extension of the method to multiple sensors.

VII. CONCLUSIONS

This paper described an automated method that simultaneously tracked multiple TDOAs from acoustically active sources in the presence of clutter and missed detections. The measurements were based on narrowband whistles and broadband echolocation clicks, and incorporated cross-correlation amplitude as an additional feature to TDOA. The proposed scheme tracked multiple subgroups based on their TDOAs and amplitudes well and extracted more information compared to cases that considered only one signal type. This is an important step toward automating source tracking with hydrophone arrays.

ACKNOWLEDGMENTS

The data used in this study was collected as part of the Hawaiian Islands Cetacean and Ecosystem Assessment Survey in 2017, part of the Pacific Marine Assessment Program for Protected Species (PacMAPPS) funded by the National Marine Fisheries Service, the U.S. Navy, and the Bureau of Ocean Energy Management. PG was funded by the Pacific Islands Fisheries Science Center and the NMFS National Take-Reduction Program. We would like to thank Jennifer McCullough for supplying insight into false killer whale behavior and data that enabled the initial exploration of TDOA measurement variances, Yvonne Barkley for supplying information on false killer whale whistle characteristics, and Fabio Casagrande Hirono for helpful comments and suggestions on the method and the manuscript. We would also like to thank three anonymous reviewers for their comments, which helped strengthen this manuscript.

NOMENCLATURE

A_{xy}	Envelope of GCC-SCOT
a	Measured amplitude
$c_a(a)$	Clutter amplitude likelihood
$c_a^\lambda(a)$	Thresholded $c_a(a)$ with a threshold λ
$\tilde{c}_k(\tilde{\boldsymbol{z}})$	Likelihood function for clutter
$c_k(z) \overset{\text{abbr}}{=} c_k$	Clutter likelihood based on the TDOA
d	Expected SNR
d_1, d_2	Lower and upper values for the expected SNR
η_k, R_k	Measurement noise process and its covariance matrix, respectively
F	State transition (system) matrix
$g_a(a\|d) \overset{\text{abbr}}{=} g_a(a)$	Target amplitude likelihood
$\tilde{g}_k(\tilde{\boldsymbol{z}}\|\tilde{\boldsymbol{x}})$	Likelihood function for the targets
$g_k(z\|\boldsymbol{x})$	Target likelihood based on the TDOA
$\gamma_k(\boldsymbol{x})$	Birth PHD
J_{k-1}	Number of persistent targets derived from the previous time step $k-1$
λ	Threshold imposed on A_{xy}
$m_{k,b}$	Newborn component mean
$\boldsymbol{n}_{k-1}$ and Q	System noise process and its covariance matrix, respectively
ν_b	Expected number of newborn targets per time step
$p_D^\lambda(d) \overset{\text{abbr}}{=} p_D^\lambda$	Probability of detection given the threshold λ
p_S	Probability of the target's survival from one time step to another
r_k	Clutter rate
σ_r^2	Variance of a Rayleigh distribution
σ_v	Standard deviation of the system noise
$\tau, \dot{\tau}, \ddot{\tau}$	TDOA and its derivatives
$v_{k\|k-1}(\boldsymbol{x})$	Predicted PHD
$v_{k\|k}(\boldsymbol{x})$	Posterior PHD
$v_{k\|k-1,p}(\boldsymbol{x})$, $v_{k\|k,p}(\boldsymbol{x})$	Predicted and posterior PHDs for the persistent targets, respectively
$v_{k\|k-1,b}(\boldsymbol{x})$, $v_{k\|k,b}(\boldsymbol{x})$	Predicted and posterior PHDs for the newborn targets, respectively

https://doi.org/10.1121/10.0006780

w	Weight
$w_{k,b}$	Newborn component weight
$\boldsymbol{x}, \tilde{\boldsymbol{x}}$	State vector and extended state vector, respectively
z	Measured TDOA
$\tilde{\boldsymbol{z}}$	Extended measurement vector
$\tilde{\boldsymbol{Z}}_k$	Extended multi-target measurement set at time k

[1]A. L. Bradford, K. A. Forney, E. M. Oleson, and J. Barlow, "Accounting for subgroup structure in line-transect abundance estimates of false killer whales (*Pseudorca crassidens*) in Hawaiian waters," PloS One **9**(2), e90464 (2014).

[2]A. Thode, "Tracking sperm whale (*Physeter macrocephalus*) dive profiles using a towed passive acoustic array," J. Acoust. Soc. Am. **116**(1), 245–253 (2004).

[3]A. Thode, "Three-dimensional passive acoustic tracking of sperm whales (*Physeter macrocephalus*) in ray-refracting environments," J. Acoust. Soc. Am. **118**(6), 3575–3584 (2005).

[4]A. I. DeAngelis, R. Valtierra, S. M. Van Parijs, and D. Cholewiak, "Using multipath reflections to obtain dive depths of beaked whales from a towed hydrophone array," J. Acoust. Soc. Am. **142**(2), 1078–1087 (2017).

[5]R. W. Baird, "A review of false killer whales in Hawaiian waters: Biology, status, and risk factors," Technical Report for the U.S. Marine Mammal Commission, Order No. E40475499, Cascadia Research Collective, Olympia, WA (2009).

[6]E. M. Oleson, C. H. Boggs, K. A. Forney, M. B. Hanson, D. R. Kobayashi, B. L. Taylor, P. R. Wade, and G. M. Ylitalo, "Status review of Hawaiian insular false killer whales (Pseudorca crassidens) under the endangered species act," Technical Memorandum NOAA-TM-NMFS-PIFSC-22, U.S. Dept. Commerce, NOAA, 2010).

[7]W. W. Au, J. L. Pawloski, P. E. Nachtigall, M. Blonz, and R. C. Gisner, "Echolocation signals and transmission beam pattern of a false killer whale (*Pseudorca crassidens*)," J. Acoust. Soc. Am. **98**(1), 51–59 (1995).

[8]S. Baumann-Pickering, A. E. Simonis, E. M. Oleson, R. W. Baird, M. A. Roch, and S. M. Wiggins, "False killer whale and short-finned pilot whale acoustic identification," Endangered Species Res. **28**(2), 97–108 (2015).

[9]J. N. Oswald, S. Rankin, J. Barlow, and M. O. Lammers, "A tool for real-time acoustic species identification of delphinid whistles," J. Acoust. Soc. Am. **122**(1), 587–595 (2007).

[10]Y. M. Barkley, E. M. Oleson, J. N. Oswald, and E. C. Franklin, "Whistle classification of sympatric false killer whale populations in Hawaiian waters yields low accuracy rates," Front. Mar. Sci. **6**, 1–15 (2019).

[11]E. Henderson, J. Hildebrand, M. Smith, and E. Falcone, "The behavioral context of common dolphin (*Delphinus* sp.) vocalizations," Mar. Mammal Sci. **28**(3), 439–460 (2012).

[12]S. M. Wiggins, M. A. McDonald, and J. A. Hildebrand, "Beaked whale and dolphin tracking using a multichannel autonomous acoustic recorder," J. Acoust. Soc. Am. **131**(1), 156–163 (2012).

[13]M. Gassmann, E. Elizabeth Henderson, S. M. Wiggins, M. A. Roch, and J. A. Hildebrand, "Offshore killer whale tracking using multiple hydrophone arrays," J. Acoust. Soc. Am. **134**(5), 3513–3521 (2013).

[14]C. W. Clark and W. T. Ellison, "Calibration and comparison of the acoustic location methods used during the spring migration of the bowhead whale, *Balaena mysticetus*, off Pt. Barrow, Alaska, 1984–1993," J. Acoust. Soc. Am. **107**(6), 3509–3517 (2000).

[15]C. O. Tiemann, M. B. Porter, and L. N. Frazer, "Localization of marine mammals near hawaii using an acoustic propagation model," J. Acoust. Soc. Am. **115**(6), 2834–2843 (2004).

[16]S. M. Wiggins, K. E. Frasier, E. Elizabeth Henderson, and J. A. Hildebrand, "Tracking dolphin whistles using an autonomous acoustic recorder array," J. Acoust. Soc. Am. **133**(6), 3813–3818 (2013).

[17]E.-M. Nosal, "Methods for tracking multiple marine mammals with wide-baseline passive acoustic arrays," J. Acoust. Soc. Am. **134**(3), 2383–2392 (2013).

[18]G. C. Carter, "Coherence and time delay estimation," Proc. IEEE **75**(2), 236–255 (1987).

[19]P. R. White, T. G. Leighton, D. C. Finfer, C. Powles, and O. N. Baumann, "Localisation of sperm whales using bottom-mounted sensors," Appl. Acoust. **67**(11-12), 1074–1090 (2006).

[20]P. M. Baggenstoss, "An algorithm for the localization of multiple interfering sperm whales using multi-sensor time difference of arrival," J. Acoust. Soc. Am. **130**(1), 102–112 (2011).

[21]T. A. Helble, G. R. Ierley, G. L. D'Spain, and S. W. Martin, "Automated acoustic localization and call association for vocalizing humpback whales on the Navy's Pacific Missile Range Facility," J. Acoust. Soc. Am. **137**(1), 11–21 (2015).

[22]B. Miller and S. Dawson, "A large-aperture low-cost hydrophone array for tracking whales from small boats," J. Acoust. Soc. Am. **126**(5), 2248–2256 (2009).

[23]P. M. Baggenstoss, "A multi-hypothesis tracker for clicking whales," J. Acoust. Soc. Am. **137**(5), 2552–2562 (2015).

[24]R. P. Mahler, *Statistical Multisource-Multitarget Information Fusion* (Artech House Inc., Norwood, MA (2007), p. 856.

[25]W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," IEEE Trans. Sign. Process. **54**(9), 3291–3304 (2006).

[26]R. Mahler, "A theoretical foundation for the Stein-Winter 'Probability Hypothesis Density (PHD)' multitarget tracking approach," in *Proceedings of the 2000 MSS National Symposium on Sensor and Data Fusion*, DTIC Document, San Antonio, Texas (2000), pp. 99–117.

[27]R. P. Mahler, *Advances in Statistical Multisource-Multitarget Information Fusion* (Artech House Inc., Norwood, MA, 2014), p. 1128.

[28]D. E. Clark, I. Ruiz, Y. Petillot, and J. Bell, "Particle PHD filter multiple target tracking in sonar image," IEEE Trans. Aerosp. Electron. Syst. **1**(43), 409–416 (2007).

[29]A.-A. Saucan, T. Chonavel, C. Sintes, and J.-M. Le Caillec, "CPHD-DOA tracking of multiple extended sonar targets in impulsive environments," IEEE Trans. Sign. Process. **64**(5), 1147–1160 (2016).

[30]E. Maggio, M. Taj, and A. Cavallaro, "Efficient multitarget visual tracking using random finite sets," IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1016–1027 (2008).

[31]C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *2015 IEEE International Conference on Digital Signal Processing (DSP)*, IEEE, Singapore (2015), pp. 1206–1210.

[32]P. Gruden and P. R. White, "Automated tracking of dolphin whistles using Gaussian mixture probability hypothesis density filters," J. Acoust. Soc. Am. **140**(3), 1981–1991 (2016).

[33]P. Gruden and P. R. White, "Automated extraction of dolphin whistles—A sequential Monte Carlo probability hypothesis density approach," J. Acoust. Soc. Am. **148**(5), 3014–3026 (2020).

[34]D. Clark, B. Ristic, and B.-N. Vo, "PHD filtering with target amplitude feature," in *2008 11th International Conference on Information Fusion*, IEEE, Cologne, Germany (2008), pp. 1–7.

[35]D. Clark, B. Ristic, B.-N. Vo, and B. T. Vo, "Bayesian multi-object filtering with amplitude feature likelihood for unknown object SNR," IEEE Trans. Sign. Process. **58**(1), 26–37 (2010).

[36]B. Ristic, D. Clark, B.-N. Vo, and B.-T. Vo, "Adaptive target birth intensity for PHD and CPHD filters," IEEE Trans. Aerosp. Electron. Syst. **48**(2), 1656–1668 (2012).

[37]A. Weiss and E. Weinstein, "Fundamental limitations in passive time delay estimation—Part I: Narrow-band systems," IEEE Trans. Acoust. Speech Sign. Process. **31**(2), 472–486 (1983).

[38]C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Sig. Process. **24**(4), 320–327 (1976).

[39]K. Scarbrough, N. Ahmed, and G. C. Carter, "Comparison of two methods for time delay estimation of sinusoids," Proc. IEEE **70**(1), 90–92 (1982).

[40]D. Gillespie, M. Caillat, J. Gordon, and P. R. White, "Automatic detection and classification of odontocete whistles," J. Acoust. Soc. Am. **134**(3), 2427–2437 (2013).

[41]The click removal sample code is available at https://github.com/PinaGruden/Click_removal_sample_code (Last viewed June 14, 2021).

[42]J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, Fourth ed. (Wiley, Hoboken, NJ, 2010), p. 604.

[43]B.-N. Vo and W.-K. Ma, "The Gaussian mixture probability hypothesis density filter," IEEE Trans. Sign. Process. **54**(11), 4091–4104 (2006).

[44]S. M. Bozic, *Digital and Kalman Filtering* (Edward Arnold Ltd., London, UK, 1979), p. 157.

[45]K. Panta, D. E. Clark, and B.-N. Vo, "Data association and track management for the Gaussian mixture probability hypothesis density filter," IEEE Trans. Aerosp. Electron. Syst. **45**(3), 1003–1016 (2009).
[46]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0006780 for the derivation of the PHD filter, which incorporates the amplitude information and updates persistent and newborn targets separately; measurements, annotated, and interpolated data for scenarios 1–4; and additional figures of the GM-PHD-S tracking performance, respectively.
[47]X. R. Li and V. P. Jilkov, "Survey of maneuvering target tracking. Part I: Dynamic models," IEEE Trans. Aerosp. Electron. Syst. **39**(4), 1333–1364 (2003).
[48]Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation* (Wiley, New York, 2001), p. 558.
[49]D. Lerro and Y. Bar-Shalom, "Automated tracking with target amplitude information," in 1990 American Control Conference, IEEE, San Diego, CA (1990), pp. 2875–2880.
[50]A. Quazi, "An overview on the time delay estimate in active and passive systems for target localization," IEEE Trans. Acoust. Speech Sign. Process. **29**(3), 527–533 (1981).
[51]A. von Benda-Beckmann, S. Beerens, and S. Van Ijsselmuide, "Effect of towed array stability on instantaneous localization of marine mammals," J. Acoust. Soc. Am. **134**(3), 2409–2417 (2013).
[52]D. A. McCallum, *A Conceptual Guide to Detection Probability for Point Counts and Other Count-Based Survey Methods* (U.S. Dept. of Agriculture, Forest Service, Pacific Southwest Research Station, Asilomar, CA, 2005).
[53]C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer Science nd Business Media, New York, 2006), p. 738.
[54]M. A. Roch, T. S. Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, and M. S. Soldevilla, "Automated extraction of odontocete whistle contours," J. Acoust. Soc. Am. **130**(4), 2212–2223 (2011).
[55]K. M. Yano, E. M. Oleson, J. L. Keating, L. T. Ballance, M. C. Hill, A. L. Bradford, A. N. Allen, T. W. Joyce, J. E. Moore, and A. Henry, "Cetacean and seabird data collected during the Hawaiian Islands Cetacean and Ecosystem Assessment Survey (HICEAS), July–December 2017," Technical Memorandum NOAA-TM-NMFS-PIFSC-72, U.S. Dept. of Commerce, NOAA (2018).
[56]P. Li, X. Liua, K. J. Palmer, E. Fleishman, D. Gillespie, E.-M. Nosal, Y. Shiu, H. Klinck, D. Cholewiak, T. Helble, and M. A. Roch, "Learning deep models from synthetic data for extracting dolphin whistle contours," in 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, Glasgow, UK (2020), pp. 1–10.
[57]B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo, "A metric for performance evaluation of multi-target tracking algorithms," IEEE Trans. Sign. Process. **59**(7), 3452–3457 (2011).
[58]R. P. Mahler, B.-T. Vo, and B.-N. Vo, "CPHD filtering with unknown clutter rate and detection profile," IEEE Trans. Sign. Process. **59**(8), 3497–3513 (2011).
[59]A. L. Bradford, E. A. Becker, E. M. Oleson, K. A. Forney, J. E. Moore, and J. Barlow, "Abundance estimates of false killer whales in Hawaiian waters and the broader central Pacific," Technical Memorandum NOAA-TM-NMFS-PIFSC-104, U.S. Dept. of Commerce, NOAA (2020).
[60]Y. Barkley (personal communication, 2020).