The Journal of the Acoustical Society of America

Automated extraction of dolphin whistles- a Sequential Monte Carlo Probability Hypothesis Density (SMC-PHD) approach

--Manuscript Draft--

Manuscript Number:	JASA-05519R1
Full Title:	Automated extraction of dolphin whistles- a Sequential Monte Carlo Probability Hypothesis Density (SMC-PHD) approach
Article Type:	Regular Article
Corresponding Author:	Pina Gruden Research Corporation of the University of Hawaii (RCUH) Honolulu, HI UNITED STATES
First Author:	Pina Gruden
Order of Authors:	Pina Gruden
	Paul White, Prof.
Section/Category:	SPECIAL ISSUE ON MACHINE LEARNING IN ACOUSTICS
Keywords:	automated whistle tracking; multi-target Bayesian; probability hypothesis density; sequential Monte Carlo
Abstract:	The need for automated methods to detect and extract marine mammal vocalizations from acoustic data has increased in the last few decades due to the increased availability of long-term recording systems. Automated dolphin whistle extraction represents a challenging problem due to the time-varying number of overlapping whistles present in, potentially, noisy recordings. Typical methods utilize image processing techniques or single target tracking, but often result in fragmentation of whistle contours and/or partial whistle detection. This study casts the problem into a more general statistical multi-target tracking framework, and uses the probability hypothesis density (PHD) filter as a practical approximation to the optimal Bayesian multi-target filter. In particular, a particle version, referred to as a Sequential Monte Carlo PHD (SMC-PHD) filter, is adapted for frequency tracking and specific models are developed for this application. Based on these models, two versions of the SMC-PHD filter are proposed and their performance is investigated on an extensive real-world dataset of dolphin acoustic recordings. The proposed filters are shown to be efficient tools for automated extraction of whistles, suitable for real-time implementation.

Automated extraction of dolphin whistles - a Sequential Monte Carlo Probability Hypothesis Density (SMC-PHD) approach

Pina Gruden[1, a)] and Paul R. White[1]

Institute of Sound and Vibration Research, University of Southampton, Highfield, Hants, SO17 1BJ, UK

The need for automated methods to detect and extract marine mammal vocalizations from acoustic data has increased in the last few decades due to the increased availability of long-term recording systems. Automated dolphin whistle extraction represents a challenging problem due to the time-varying number of overlapping whistles present in, potentially, noisy recordings. Typical methods utilize image processing techniques or single target tracking, but often result in fragmentation of whistle contours and/or partial whistle detection. This study casts the problem into a more general statistical multi-target tracking framework, and uses the probability hypothesis density (PHD) filter as a practical approximation to the optimal Bayesian multi-target filter. In particular, a particle version, referred to as a Sequential Monte Carlo PHD (SMC-PHD) filter, is adapted for frequency tracking and specific models are developed for this application. Based on these models, two versions of the SMC-PHD filter are proposed and their performance is investigated on an extensive real-world dataset of dolphin acoustic recordings. The proposed filters are shown to be efficient tools for automated extraction of whistles, suitable for real-time implementation.

[a)] pgruden@hawaii.edu; Currently at: Research Corporation of the University of Hawaii (RCUH), Honolulu, HI, 96822, US.

I. INTRODUCTION

The detection and extraction of marine mammal calls is a crucial first step in many applications, such as abundance estimation[1], species identification[2–4], behavioural studies[5], and is used in mitigation during industrial activities[6]. These applications can involve data collection over extended periods of time and result in large quantities of data accumulating, in which case automated analysis tools become a necessity. This work proposes and validates a multi-target tracking approach for automated whistle extraction using Sequential Monte Carlo Probability Hypothesis Density (SMC-PHD) filters, including specific models tailored for tracking multiple whistles in a real-world dataset.

When extracting tonal sounds, such as narrowband frequency modulated delphinid whistles, the aim is to describe the contour of each call - *i.e.*, the frequency evolution of a tonal signal through time. This process can be referred to as extraction[7,8], detection[3,9] or tracking[10,11]. The typical signal processing work-flow involves a pre-processing stage, where the effect of the background noise and interfering signals is reduced, followed by the extraction of whistles[3,7,11]. Most methods for automated whistle extraction are based on spectrogram techniques and aim to identify the strongest spectral peaks, which are then connected to form continuous whistle contours[3,7–9,11]. For the purpose of this work, a ”measurement” is defined to be the frequency associated with a single spectral peak identified within a given spectral window. The number of measurements within a spectral window varies between windows and is unknown *a priori*.

Whistle extraction represents a challenging problem for several reasons. One is that the amplitude of a call changes rapidly throughout the whistle duration, which may cause the energy in the whistle to rise above, and then fall below, the detection threshold[3], resulting in sections of the whistle being missed. Further, there are usually many overlapping whistles and other interfering sounds present[3], which can mask the signal being tracked. The end result is a partial extraction and fragmentation of the contours.

This can hamper certain applications, such as classifiers, if they require the extraction of the full whistle contours[2]. While it is still possible to use whistle fragments to identify species[3,12], it is expected that as the length of the detected whistle contour increases, the species specific information contained in that detection improves and therefore enhances the classification. For instance, Ref.[3] found that as the fragment length of the whistle contour increased, the classification performance increased as well. This enhancement can prove significant in situations where a mix of rare and abundant species are present[13].

The goal of this study is thus to improve on the whistle extraction process, by casting it into a multi-target tracking (MTT) framework, which allows for simultaneous tracking of multiple objects of interest from the noisy measurements in the presence of missed detections, and false alarms (*i.e.*, clutter, additional measurements not generated by a whistle). Additionally, in contrast to the majority of automated methods for whistle contour extraction, the MTT accounts for the time-varying number of whistles by modelling their birth (when a whistle starts) and their death (when a whistle ends).

A. Background

A tracker is based on defining a system, whose configuration is defined at time k, by the parameter values in the state vector $\boldsymbol{x}_k$. The dynamics of the system describe how the states evolve with time, and are encapsulated in the state equation (1). The vector, $\boldsymbol{z}_k$, contains the value of the quantities that are measured at time k. The measurements are related to the system states through the measurement equation (2):

$$\boldsymbol{x}_k = F_{\mathrm{s}}(\boldsymbol{x}_{k-1}, \boldsymbol{n}_k)\,, \tag{1}$$

$$\boldsymbol{z}_k = G_{\mathrm{m}}(\boldsymbol{x}_k, \boldsymbol{\eta}_k)\,, \tag{2}$$

where F_{s} is the function which combines the previous state vector and the system noise process $\boldsymbol{n}_k$ to generate the current state and G_{m} is the function computing the measurement, $\boldsymbol{z}_k$, combining the current state vector with a measurement noise process, $\boldsymbol{\eta}_k$. For the whistle tracking problem, we employ a state vector consisting of two parameters - the instantaneous frequency and its derivative (the chirp rate) - whilst the only measurements available are measurements of the instantaneous frequency. The specific form of the state and measurement equations used in this paper are discussed in Section II C.

A Bayesian recursive filtering approach is frequently adopted in single-target tracking problems to estimate the state of a system from a sequence of noisy measurements[14]. The Bayes filter aims to compute the posterior probability density function (pdf) of the state estimate at each time step, and is based on a two stage recursion[14]. The first stage, the prediction step, uses the state dynamics (1) to compute an *a priori* estimate of the state's density function. Whilst the second stage, the update step, updates that density based on the newly available measurement, leading to an estimate of the posterior pdf for the state vector[14]. An analytic solution to the Bayes filter under the assumptions of: a single target, Gaussianity of the noise processes and linearity of the underlying models, can be obtained. The resulting method is the Kalman filter[15]. A more general approach for a single target, avoiding the need for the assumptions of linearity and Gaussianity, is offered by the Sequential Monte Carlo (SMC) filter (or particle filter)[14], which is the basis of much of what follows here.

Finite Set Statistics (FISST) provides a suitable framework within which an MTT Bayesian filter can be constructed[16,17]. FISST models the states of the targets and the measurements using the concept of random finite sets (RFS), and transforms a multi-sensor, multi-target problem into a mathematically equivalent single-sensor, single-target problem. A RFS is an object in which the unordered elements have random values, as in any multivariate random process, but in addition to which the number of elements (the set cardinality) is also random[17].

A multi-target Bayesian filter[17] determines, at each iteration, the full posterior pdf of the multi-target state, which makes it computationally intractable in practice, especially when there are a large number of targets. To overcome this problem, one solution is to use a filter based on the Probability Hypothesis Density (PHD), $v_k(\boldsymbol{x}|\boldsymbol{Z}_{1:k})$ (where $\boldsymbol{Z}_{1:k}$ is the set of measurements $\boldsymbol{z}$ at times 1 to k), which is the first-order moment of the multi-target posterior[18,19]. Note that for compactness and clarity, the dependence of v_k on $\boldsymbol{Z}_{1:k}$ is suppressed in subsequent equations. The majority of the practical applications of the PHD filter involve spatial tracking of moving objects or targets[20–23]. Herein, the PHD filter is applied to track dolphin whistles, and in the following the PHD recursion is outlined in that context.

A PHD is a function whose peaks identify the likely positions of the whistle contours. A whistle with a state $\boldsymbol{x}$ is more likely to be present in the region where the PHD is large. It should be noted that the PHD is a density function but not a pdf: a point made apparent by noting that its integral over the space of its variables is not unity, but is the expected number of whistles.

The PHD filter is implemented in a recursive manner using prediction and update steps. The goal is to determine the number and the states of the whistle contours at each time k. In the prediction step, the predicted PHD, $v_{k|k-1}$, consists of the information regarding newborn whistles and persistent whistles (whistles surviving from the previous time step, represented by the posterior PHD, v_{k-1}). In the update step the predictions of whistles are refined by incorporating the most recent measurements to obtain the posterior PHD, v_k. The prediction and update steps can be written as[17,19]:

$$v_{k|k-1}(\boldsymbol{x}_k) = \gamma_k(\boldsymbol{x}_k) + p_S \int v_{k-1}(\boldsymbol{x}_{k-1}) f_{k|k-1}(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}) d\boldsymbol{x}_{k-1}\,, \tag{3}$$

$$v_k(\boldsymbol{x}_k) = [1-p_D] v_{k|k-1}(\boldsymbol{x}_k) + \sum_{\boldsymbol{z}\in Z_k} \frac{p_D g_k(\boldsymbol{z}|\boldsymbol{x}_k) v_{k|k-1}(\boldsymbol{x}_k)}{\kappa_k(\boldsymbol{z}) + p_D \int g_k(\boldsymbol{z}|\boldsymbol{x}_k) v_{k|k-1}(\boldsymbol{x}_k) d\boldsymbol{x}_k}\,, \tag{4}$$

where $\gamma_k(\boldsymbol{x}_k)$ denotes the PHD of whistle births between time $k-1$ and k (*i.e.*, the integral of $\gamma_k(\boldsymbol{x}_k)$ over a given region gives the expected number of new whistles appearing in that region at a given time); p_S denotes the probability of survival, that is the *a priori* probability that a whistle at time $k-1$ will survive until time k; and $f_{k|k-1}(\boldsymbol{x}_k|\boldsymbol{x}_{k-1})$ denotes single-target state transition density (*i.e.*, probability density of a transition to the state $\boldsymbol{x}_k$ given the state $\boldsymbol{x}_{k-1}$). The probability of detection is denoted by p_D and it represents the probability that a measurement will be detected from a whistle, $\kappa_k(\boldsymbol{z})$ denotes the PHD of clutter, and $g_k(\boldsymbol{z}|\boldsymbol{x}_k)$ denotes the single-target measurement likelihood function (*i.e.*, a likelihood that a measurement $\boldsymbol{z}$ was generated by a whistle with a state $\boldsymbol{x}_k$). Eqs. (3) and (4) have been adapted to exclude the spawning terms[17], since contour splitting is not typically observed in dolphin whistles.

A closed form solution to (3) - (4) can be obtained assuming the PHD is a mixture of weighted Gaussian components leading to the so-called Gaussian Mixture PHD (GM-PHD) filter[24]. Advantages of this method are that it is straightforward to implement, however it requires one to assume a linear model for the system and a Gaussian assumption for the noise processes. Despite this limitation it has been successfully used to track dolphin whistles[11]. A more general approximation to (3) - (4) can be achieved with a particle filter, in what is known as the Sequential Monte Carlo PHD (SMC-PHD) filter[25,26], where weighted particles (random samples) are used to approximate the PHD function. The SMC-PHD filter is a direct generalization of the approach employed for a single-target particle filter, and particles are propagated over time using importance sampling and re-sampling strategies[25,26]. This implementation of the PHD filter imposes no constraints on the underlying models, and is the focus of the current work.

B. Contributions

In this paper we propose a complete multi-target frequency tracking scheme to track frequency modulated narrowband signals from audio recordings using a SMC-PHD filter. To achieve this we develop new models for this application, and a particle labeling scheme to allow associations of the tracks between frames. Further, this paper reports the outcome of performance tests on a real-world dataset, and benchmarks the performance against the previously mentioned GM-PHD filter.

This paper is organized in the following manner. Section II describes the dataset, the SMC-PHD algorithm for dolphin whistle tracking, along with the developed models and optimized parameters, as well as the method's evaluation procedure. Section III contains evaluation of the proposed methods on the real-world dataset comprising of dolphin recordings. The discussion and the conclusions can be found in Section IV and Section V respectively.

II. METHODS

A. Data, pre-processing steps and obtaining the measurements

The dataset for evaluation of the filter's performance was from the 5th Workshop of Detection, Classification, Localization and Density Estimation (DCLDE) conference in 2011, obtained from the MobySound archive (`http://www.mobysound.org`), which has been used in Refs.[3,7,11,27]. For this work, a subset containing raw recordings and hand-annotated files of whistle contours for six delphinid species (*Delphinus capensis*, *Delphinus delphis*, *Peponocephala electra*, *Stenella longirostris*, *Stenella frontalis*, and *Tursiops truncatus*) was used, and is the same dataset used in Ref.[11]. The majority of recordings were sampled at 192 kHz, but a small portion (15%) of the files had higher sampling rates, and were re-sampled to 192 kHz for consistency. The data collection protocols and study areas are given in Refs.[7,28]. In this study, the hand-annotated files supplied with the dataset were used as a ground truth data for the filter's performance evaluation (described in Section II D). Whistles which are 150 ms long and have a Signal to Noise Ratio (SNR) exceeding 10 dB for at least one third of their duration are termed valid[7] and primarily only valid whistles are used in the following analysis. The raw recordings were used for the filter to track the whistles and, in addition, a part of raw data was set aside as a training set for parameter selection in the SMC-PHD filters. This training set consisted of three one minute duration files chosen randomly: they were recordings of *Delphinus capensis*, *Delphinus delphis*, and *Stenella frontalis* that contained 67, 55 and 63 valid whistles, respectively. This training data was not subsequently used in the performance evaluation.

To implement the whistle contour tracking, a set of measurements for each time instance is needed, which is a standard procedure for any multi-target tracking[29]. The measurement sets, that in our application comprise of the spectral peaks, are obtained using established methods[3,11]. This pre-processing reduces the background noise and the impact of interfering signals, and is based on a spectrogram using a sliding window of 2048 points (frequency bin width 93.8 Hz) with 50% overlap. The spectral peaks were identified using an 8 dB threshold applied to the normalized spectrogram, converted to a dB scale. Only spectral peaks in the frequency range (2 - 50 kHz) were selected, since that encompassed most dolphin whistles and their harmonics. The precision of the location of the spectral peaks was improved by fitting a quadratic through points surrounding the peak and using the location of the maximum of that fitted quadratic as the refined peak location. These spectral peaks represent the measurement set from which the whistle contours were tracked. The pre-processing code used to generate the measurement set from the raw files and the measurement set itself are available at https://doi.org/10.5258/SOTON/D0316 to facilitate the comparisons with other algorithms that operate on spectral peaks. Moreover, the SMC-PHD filter implementation is available at https://github.com/PinaGruden/SMCPHD_whistle_contour_tracking.

B. Sequential Monte Carlo PHD (SMC-PHD) filter for the whistle contour detection

The SMC-PHD filter[25,26] consists of the basic prediction and update steps seen in the Bayes filter. Following the principles of sequential Monte-Carlo methods (or particle filters) the underlying integrals are solved recursively using point-wise approximations. These methods rely upon a set of particles and their associated weights which are propagated from one time step to the next.

The standard formulation of the SMC-PHD filter[25,26] suffers from two limitations that relate to initiating newborn particles (*i.e.*, target birth) and to estimating the state[30]. The location where targets are born is typically known *a priori* and a large number of particles are required in that region. Further, state estimates are typically constructed using *ad-hoc* clustering of particles[25,30]. To make the filter more computationally efficient and increase accuracy, a data driven variation of the SMC-PHD filter has been proposed[30–32], which generates new particles based on the measurements. This reduces the number of particles required, and eliminates the need for clustering techniques during the state estimation process by exploiting properties of the PHD update equation. Further, the state estimates do not contain identities (*i.e.*, it is not known which state estimate belongs to which target being tracked), and either particle labeling[33,34] or an external algorithm[22,35] can be used to achieve the temporal association of the estimates and so obtain target tracks.

The algorithm description and the pseudo-code of the SMC-PHD filter used for whistle contour tracking are given in Section II B 1, Alg. 1, the temporal association procedure is detailed in Section II B 2, and the specific models and parameters are presented in Section II C.

1. *The SMC-PHD algorithm*

The SMC-PHD filter propagates through time the weighted particle system $\mathcal{P}_k \equiv \{w_k^{(i)}, \boldsymbol{x}_k^{(i)}\}_{1\le i\le N_k}$, which approximates the PHD function, $v_k(\boldsymbol{x}_k)$[32]. At time k, each whistle contour is represented by a cluster of particles representing the state vectors $\boldsymbol{x}_k^{(i)}$ and the corresponding weights $w_k^{(i)}$. At each time step, the filter produces an estimate of the multi-whistle state, $\hat{\boldsymbol{X}}_k$, which contains state estimates, $\hat{\boldsymbol{x}}_k$, of whistles. Further, the sum of particle weights represents an estimate of the number of whistles[17].

The prediction step in the SMC-PHD filter starts with the persistent particles from the previous time step, along with newborn particles being drawn from a proposal density to form the predicted particles $\boldsymbol{x}_{k|k-1}^{(i)}$. The predicted particle weights, $w_{k|k-1}^{(i)}$, are computed by scaling them using p_S. These estimates are then refined in the update step using the set of measurements, $\boldsymbol{Z}_k$.

The update process consists of multiple elements. First, the weights and particles are partitioned on the basis of the measurement set using the probability that the i-th particle is associated with the j-th measurement, denoted $P_{i,j}$. To allow for the possibility of a missed detection an additional category is added to the measurements, and the probability that the i-th particle is associated with the missed detection is denoted $P_{i,0}$. Partitioning into the clusters is performed by randomly drawing an index for each particle according to $P_{i,j}, j = 0, \cdots, |\boldsymbol{Z}_k|$ (where $|\cdot|$ denotes the cardinality). The computation of $P_{i,j}$ and partitioning are based on Eq. (4), and are detailed in Eq. (50) in Ref.[32]. This partitioning creates a cluster of particles $C_{k|k-1}(z)$, one for each measurement, plus one cluster for the missed detection class $C_{k|k-1}(\emptyset)$. Note that there is the possibility that any cluster could be empty. The method treats the particles associated with the measurements and missed detections in different fashions.

For non-empty clusters associated with a measurement, all the particle weights in the cluster are updated according to the second term in Eq. (4), and particles are resampled through a stratified resampling process[36]. Then a probability of the cluster existing, p_e, is computed, by summing all the weights of particles in that cluster. If that probability is greater than a predefined threshold, η, then the resampled particles within a cluster are averaged to give a state estimate $\hat{\boldsymbol{x}}_k$.

In the cluster corresponding to missed detections, $C_{k|k-1}(\emptyset)$, the particles and associated weights have no measurements on which to base the update. Their weights are scaled by $(1 - p_D)$, *i.e.* probability of missed detection. Only particles whose weights exceed a threshold, ξ, are kept to reduce the computational burden.

Finally, the algorithm takes into account the possibility that a whistle starts at time k, *i.e.* a whistle birth. At the end of each iteration a set of N_b particles are drawn, focussed on regions in the state-space where measurements, not associated with state estimates, denoted $\boldsymbol{Z}_{b,k}$, were made. This process is detailed in Section II C 3.

2. *Temporal association of the whistle estimates*

For each time step, the SMC-PHD filter in Alg. 1 outputs a set of the estimated states ($\hat{\boldsymbol{X}}_k$) that represent whistle contour peaks for that time step. However, these do not have identities associated with them, *i.e.*, one does not know which estimate in one time step links to which estimate in the next time step, something that is required to be able to form continuous whistle tracks (contours).

Two broad approaches for temporal association of the estimated states are used in the literature; one is to use a separate algorithm for the association[22,35], the other is to label the individual particles[33,34]. The particle labeling approach tends to be computationally more efficient since it only requires an additional set of labels to be propagated alongside the weighted particle set, and does not need a separate algorithm. The particle labeling approach was adopted in this work, and each particle i was assigned a label $T^{(i)}$, which was propagated through time. Unlike other labeling approaches, which typically require clustering methods[33,34], the procedure employed in this study grouped particles by exploiting the properties of the PHD update equation[32]. The procedure is outlined below, with reference to the relevant steps in Alg. 1.

On initialization, each particle is assigned a null label $T^{(i)}$. In subsequent prediction steps (line 2 Alg. 1), the set of the predicted particle labels remains the same, $\boldsymbol{T}_{k|k-1} = \boldsymbol{T}_{k-1}$. In the update step, after resampling (line 12 Alg. 1), the resampled particles within a given cluster retain the labels of the particles from which they were derived. After that, the cluster identity is determined, based on the maximum sum of weights of the particles with the same label. If a cluster originates from a newborn whistle, then the cluster is assigned a new identity, with all previously unlabelled particles in this cluster being assigned the new label.

The identity of the state estimate, is determined from the identity of the cluster from which the state estimate was derived (line 16 Alg. 1).

Each newborn particle is assigned the label $T_{b,k}^{(i)} = 0$ (line 27 Alg. 1). The labels are added to the labels of the persistent particles and are predicted and updated together in the next time step.

The individual whistles are then tracked from the estimated states based on their identities. So that all the states with the same identity are linked together into a continuous whistle contour (track). Finally, the duration of the track is examined and if it falls below a threshold then it is rejected[3,7]. This condition is called the track length criterion and various values for the threshold were investigated ranging from 53 to 150 ms (10 to 28 time steps in the spectrogram).

If there is more than one state estimate with the same identity at a given time, this conflict is resolved in the following way: in the case that this is a new whistle (no previous state estimates had this identity), the mean of the states is taken and it becomes the state estimate for that identity. If this is not a new whistle, *i.e.*, there were previously some state estimates with this identity, the last state estimate with the same identity is projected forward to the current time step k, using the system function, Eq. (1). The Mahalanobis distance between the predicted states and those with the conflicting identities is computed. It is the state closest to the prediction which is retained and the other states with conflicting identities are discarded.

C. Models and parameters for the whistle SMC-PHD

As stated in Section I A, the state vectors chosen for this study consist of frequency f [Hz] and chirp rate α (rate of change of frequency, $\dot{f}$ [Hz/s]), so that $\boldsymbol{x}_k = [f, \alpha]^t$, where $[\cdot]^t$ denotes the transpose[10,11]. The following subsections describe the models and parameters for the SMC-PHD filter employed in this study.

1. Measurement model

It is assumed that the only measurement available is the frequency and that model for the noise is additive. Accordingly, the measurement model, (2), can be simplified to[11]:

$$z_k = \boldsymbol{H}\boldsymbol{x}_k + \eta_k = [1, 0]\,\boldsymbol{x}_k + \eta_k\,, \tag{5}$$

where z_k [Hz] denotes the available measurements. The measurement noise, η_k [Hz], is assumed to be Gaussian white noise with a variance R. The value of R is set to the variance of a uniform random variable covering a single frequency bin[11], and is thus dependent on the width of the frequency bin. Specifically, $R = 732$ Hz^2 for the parameters used in this study.

From Eq. (5) the measurement likelihood function, $g_k(z|\boldsymbol{x}_k)$, is determined to be $g_k(z|\boldsymbol{x}_k) = \mathcal{N}(z; \boldsymbol{H}\boldsymbol{x}_k, R)$, where $\mathcal{N}(\cdot; m, \Sigma)$ denotes a Gaussian density with a mean m and a covariance Σ.

2. System model

The system (motion) model describes how the state develops with respect to time. In the case of whistle contour tracking, the motion model should contain information on how the frequency and chirp component of a given whistle evolve with time. The choice of the motion model can be crucial for the performance of the tracking algorithms, and many applications, such as surveillance tracking, may have a good understanding of the underlying dynamics[37]. However, this is not the case for the frequency evolution of whistle contours.

In this work two different motion models are explored. The first model used is a linear model described in Ref.[11], similar to the "nearly-constant-velocity models" in Ref.[37], specifically:

$$\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{x}_{k-1} + \boldsymbol{n}_k = \begin{bmatrix} 1 & \triangle \\ 0 & 1 \end{bmatrix} \boldsymbol{x}_{k-1} + \boldsymbol{n}_k \,, \tag{6}$$

where $\triangle$ denotes the time interval between spectral windows (here 0.0053 s, based on $f_s =$ 192 kHz, window length = 2048, 50% overlap). The system noise, $\boldsymbol{n}_k$, in this model is Gaussian white noise with a covariance matrix $\boldsymbol{Q}$.

While the SMC-PHD allows for non-linear models the linear model in Eq. (6) provides a baseline for comparisons, since it was successfully applied to track dolphin whistles[11] albeit with a different PHD filter.

The covariance matrix, $\boldsymbol{Q}$, is assumed to be diagonal, which is equivalent to assuming that the noise processes acting on the frequency and chirp rate are uncorrelated. In which case there are two degrees of freedom when selecting $\boldsymbol{Q}$, namely the two variances, σ_f^2 and σ_α^2, representing the noise processes driving the frequency and chirp rate, receptively. These were determined experimentally by running the SMC-PHD filter on the training data and selecting the value that gave the best performance (see Section II D for performance metrics description). A range of values were tested, specifically, $\{\sigma_f^2, \sigma_\alpha^2\} = (10, 10^2), (10, 10^3), (10^2, 10^3), (10^2, 10^4), (10^3, 10^4), (10^3, 10^6)$. The best performance was achieved for the pair $(10^2, 10^4)$.

A second motion model was developed based on training a neural network to learn the temporal relationships defining whistle evolution. The training is based on the set of hand-annotated data and the idea is similar to the one used in video tracking[38], where a set of hand-annotated traffic trajectories are used to learn how the objects in the scene typically move, and thus construct a prior to help predict the vehicle motion.

The neural network structure adopted here is that of a Radial Basis Function (RBF) network[39]. This form of network has the advantage of a comparatively simple structure, but retains a good ability to generalize. For our application, the RBF can be expressed as:

$$\boldsymbol{x}_k = \sum_{j=0}^{M} w_j \phi_j(\boldsymbol{x}_{k-1}; \boldsymbol{c}_j, \boldsymbol{Q}_j) + \boldsymbol{n}_k \,, \tag{7}$$

where ϕ_j is the set of $M + 1$ basis functions. Herein we use Gaussian functions which are parametrised by $\boldsymbol{c}_j$ and $\boldsymbol{Q}_j$, these control the location and width of the basis functions respectively (these are closely related to the mean and covariance matrix of a multi-dimensional Gaussian probability density function). A diagonal form for the matrices $\boldsymbol{Q}_j$ allows one to reduce the dimensionality of the model, so limits the amount of training data necessary, but does restrict the ability of a network of a given size to generalise. The basis function for $j = 0$, is included as a special case and represents a bias term, realised by fixing $\phi_0 = 1$.

In this study, the training data comprised 13,688 data points from 185 whistles. The hand annotations only measure the frequencies in whistles. To train the network to learn the full state model it is necessary to know the the chirp rates as well as the frequencies. The chirp rates were estimated from the hand annotations using a one point backward difference formula to approximate the frequency derivative (chirp rate).

For a given network size, M, the network is trained by first using a k-means clustering algorithm[40] to determine a suitable set of centres, $\boldsymbol{c}_j$. The $\boldsymbol{Q}$ matrices are then determined on the basis of the Euclidean distances between those centres. The final step is to compute the weights w_j, which only requires the solution of a linear system of equations[39]. To select a suitable value for M a cross-validation process is used. This cross-validation process randomly sub-divided the training data into thirds. Two thirds were used for training during validation and one third for cross-validation. The network was trained with various choices of M and the process repeated 10 times for different sub-divisions of the training data with the results of the 10 repeats averaged. The value of M yielding the smallest mean squared error was chosen, in this case $M = 60$.

As in the linear case the noise $\boldsymbol{n}_k$ was assumed to be Gaussian, white and uncorrelated between the two state variables. Based on the statistics of the residuals the noise variances $\{\sigma_f^2, \sigma_\alpha^2\}$ were (39 Hz^2, 7326 Hz^2/s^2).

3. Birth model

The birth model defines where in the state space new whistles are likely to appear, and how many appear at each time step, as characterized by the birth PHD, $\gamma_k(\cdot)$. If a whistle appears in a region that is not covered by the birth PHD then the filter may fail to track it[31]. There are two main challenges associated with determining a suitable birth model, one is to determine the birth region (*i.e.*, where do new whistles appear) and the other is to determine the birth magnitude (*i.e.*, how many new whistles appear at each step). Since the birth PHD in the SMC-PHD filter is represented by a cloud of weighted particles, these challenges translate into determining the regions in state space from which the new particles are initiated, and determining their weights.

a. The birth region. In many other tracking applications the birth region is assumed to be a single point or uniform across a region of state space[33,41]. An alternative is to use a data-driven approach, which is effective when the birth region is not known in advance[31]. In this approach every measurement initializes newborn targets, with gating techniques used to divide the measurements into those originating from persistent targets and those originating from newborn targets[21].

The efficiency of the data driven approach arises because it only introduces particles close to measurements where the likelihood of a new target appearing is high. This approach is extended and developed further in the current work. The algorithm used here partitions the particles based on measurements and only the clusters with sufficient weighting are used to estimate states of persistent whistles (lines 15-18 Alg. 1). This defines the set of persistent whistles and the measurements associated with them. Any remaining measurements, denoted $\boldsymbol{Z}_{b,k}$, are then considered when generating newborn particles, denoted $\boldsymbol{x}_{b,k}$.

The process by which newborn particles are generated is as follows. For every $z \in \boldsymbol{Z}_{b,k}$ a fixed number of particles, N_b, is drawn. The frequency and chirp rate components of the state are drawn independently from different distributions. For the frequency element:

$$\{x_{b,k}^{(i)}\}_f \sim \mathcal{N}(x; z^{(j)}, R)\,, \tag{8}$$

Whereas for the chirp rate there is no direct measurement on which to base the initial state estimate. To overcome this, the distribution of chirp rates at the start of a whistle was approximated using a Gaussian Mixture Model (GMM)[42]. The GMM was fitted to the distribution of chirp rates measured in the hand annotated training data set. The starting chirp rates for the annotated dataset were computed based on the difference between the first two frequency samples on a whistle. When fitting the GMM, model order was selected on the basis of the Bayesian Information Criterion[42], leading to a choice of a mixture of three Gaussians. Formally:

$$\{x_{b,k}^{(i)}\}_\alpha \sim \sum_{n=1}^{3} a_n \mathcal{N}(x; \mu_n, \sigma_{\alpha,n}^2)\,, \tag{9}$$

where a_n, μ_n and $\sigma_{\alpha,n}$ are the weights, means and variances of the GMM respectively. For our dataset these parameters were $\boldsymbol{a} = [0.28, 0.02, 0.71]$, $\boldsymbol{\mu} = [1190, -113887, 12999]$ Hz/s, and $\boldsymbol{\sigma}_\alpha^2 = [9.74, 32.6, 1180] \times 10^6$ Hz^2/s^2.

b. The birth magnitude. The birth magnitude, ν_b, is the expected number of object births at a given time[32], and is commonly chosen in an *ad-hoc* manner or based on *a priori* knowledge on the expected number of newborn objects[24,31].

In this study an alternative approach of computing ν_b adaptively was investigated, based on the idea that not all measurements are equally likely to generate newborn whistles. For this purpose, a distribution of the start frequencies of the whistles (the first frequency in each whistle contour) from the training data (see Section II A) was first computed. The start frequencies of the whistles in the training data had a skewed, non-Gaussian distribution (Jarque-Bera test[43], $p = 0.001$ at 5% significance level), with the majority of whistles starting between 8 and 13 kHz. The start frequencies were fitted to a log-normal distribution, $p_{\text{start}}(z)$, with the log of the start frequencies having a mean $\mu = 9.4$ and standard deviation $\sigma = 0.4$. The weight of the i^{th} newborn particle, $w_{b,k}^{(i)}$, is then:

$$w_{b,k}^{(i)} \propto \frac{p_{\text{start}}(z)}{N_b}\,, \tag{10}$$

where N_b denotes the number of particles per newborn whistle. Thus the weights of the newborn particles reflects the *a priori* likelihood of a whistle starting at that frequency.

4. Other parameters

Beside the models discussed in the preceding sections, there are additional parameters required to implement the SMC-PHD filter. Some parameters are determined based on properties of the dataset so the choices here match those used in Ref.[11]. The constant defining the PHD of clutter (κ_k) was chosen to be uniform across the frequency band of interest. The observed mean number of false spectral peaks (clutter) per time step, r, was set to 10, leading to $\kappa_k = r/48000 = 0.0002$. Further, the probability of a whistle surviving from one time step to the next (p_S) was set to $p_S = 0.994$ based on the mean length of whistles observed in the training data.

The particle elimination threshold (ξ) prevents the number of particles increasing without bounds. It needs to be chosen in a way that the particles on the undetected persistent whistles, that are collected in a cluster $C_{k|k-1}(\emptyset)$, are not eliminated. Following recommendations elsewhere[44], this study used $\xi = 100(1 - p_D)/M$ where M denotes the number of particles in $C_{k|k-1}(\emptyset)$.

The remaining parameters were optimized by running the SMC-PHD filter on the training data and choosing the value that resulted in the best performance (defined in Section II D). The parameters evaluated in this way are: the probability of detection (p_D), number of particles per persistent (M_p) and newborn (N_b) whistle, and the threshold used in the state estimation (η). The values used are summarized in Table I.

D. Performance evaluation

To evaluate the SMC-PHD filter's performance, the outputs of the algorithms (which consist of time against frequency peaks for each whistle) were compared to the hand-annotated ground truth data. Only valid whistles (see Section II A) were expected to be detected. A detected whistle was considered a match (true positive) to a ground truth whistle if its timing overlapped with the ground truth whistle and if the mean difference between the detected whistle path and ground truth whistle path did not exceed 3 frequency bins (281 Hz). If the detected whistle exceeded that criteria, it was considered a false positive. It should be noted that detected whistles were matched to ground truth whistles regardless of whether the ground truth whistles were considered valid. However, only the whistles that matched valid ground truth whistles were considered in the evaluation metrics that describe the quality and quantity of matches[7]. Additionally, whilst the algorithm searched for whistles between 2 and 50 kHz, the hand-annotations were only applied to the frequencies between 4.5 and 50 kHz, therefore any detected whistle that had over 40% of its contour below 4.5 kHz was not taken into account in the evaluation process[11].

The performance was measured in terms of recall, precision, fragmentation, mean deviation and coverage[3,7,11]. Recall measures the percentage of the valid whistles that are retrieved, whilst precision measures the percentage of the detections that are correct[7]. A high precision therefore indicates a low false alarm rate and a high recall indicates high detection efficiency[3]. For the detected whistles that matched valid ground truth whistles (true positives), three additional performance metrics were computed that describe the quality of the detections: fragmentation, mean deviation, and coverage. Fragmentation measures the average number of detections per ground truth whistle, mean deviation measures the average frequency deviation between the path of ground truth whistle and its corresponding detection and coverage measures the average percentage of a ground truth whistle that is matched[7].

To further evaluate the SMC-PHD filter, its performance was benchmarked against the GM-PHD filter[11]. The parameters used in the GM-PHD filter were the same as in Ref.[11], namely: $p_S = 0.994$, $p_D = 0.85$, $U = 10$, $T_r = 0.001$, $w_{th} = 0.009$, and $J_{max} = 100$.

The sensitivity of the SMC-PHD filter to the input parameter values in Table I was also investigated. The sensitivity analysis was carried out by drawing 30,000 random samples for each parameter from their respective parameter ranges. Each of these randomly selected parameter sets were then used in the SMC-PHD filter and applied to a single representative 5 s long segment of data containing multiple overlapping whistles, echolocation clicks and echosounder pulses. Note that, since the value of the parameter ξ changes during the recursion automatically, it was not included in the sensitivity analysis. To evaluate the performance of each parameter set, the $F1$ score was computed, which is a harmonic mean of the precision and recall, and reaches its best value at 100:

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \,. \tag{11}$$

Afterwards, a pseudo-marginal distribution of each parameter was obtained by creating equally spaced bins across a given parameter range and computing an average performance in each bin. A pronounced peak in the pseudo-marginal distribution indicates the filter is sensitive to parameter values around the peak location, while a flat distribution indicates no sensitivity for the specified range.

III. RESULTS

In total, 9,192 whistles form six different dolphin species were tracked with the SMC-PHD algorithm. The SMC-PHD was considered using two system models one linear and one based on an RBF (see Section II C 2). The performance was investigated across a range of track length criteria (10 - 28 time steps in the spectrogram; 53 - 150 ms).

The overall performance results, across all species, are summarized in Fig. 1, and it can be seen that the SMC-PHD filter that utilized the RBF motion model appeared to have better precision (with similar recall), for shorter track lengths, compared to the filter using the linear motion model. For longer track lengths there was a trade-off between precision and recall, with the filter using the RBF motion model having higher precision but lower recall compared to the filter using the linear motion model (Fig. 1). There was also a difference in the coverage, fragmentation, and mean deviation between the two filter types. While the filter that used the RBF motion model had slightly lower coverage and slightly higher fragmentation, it had lower mean deviation from the annotated whistle paths (Fig. 1). In both filters the track length criteria appeared to mainly influence the precision, recall, and fragmentation metrics (Fig. 1). A shorter track length criterion resulted in a higher recall, lower precision, and a higher fragmentation compared to when a longer criterion was applied (Fig. 1).

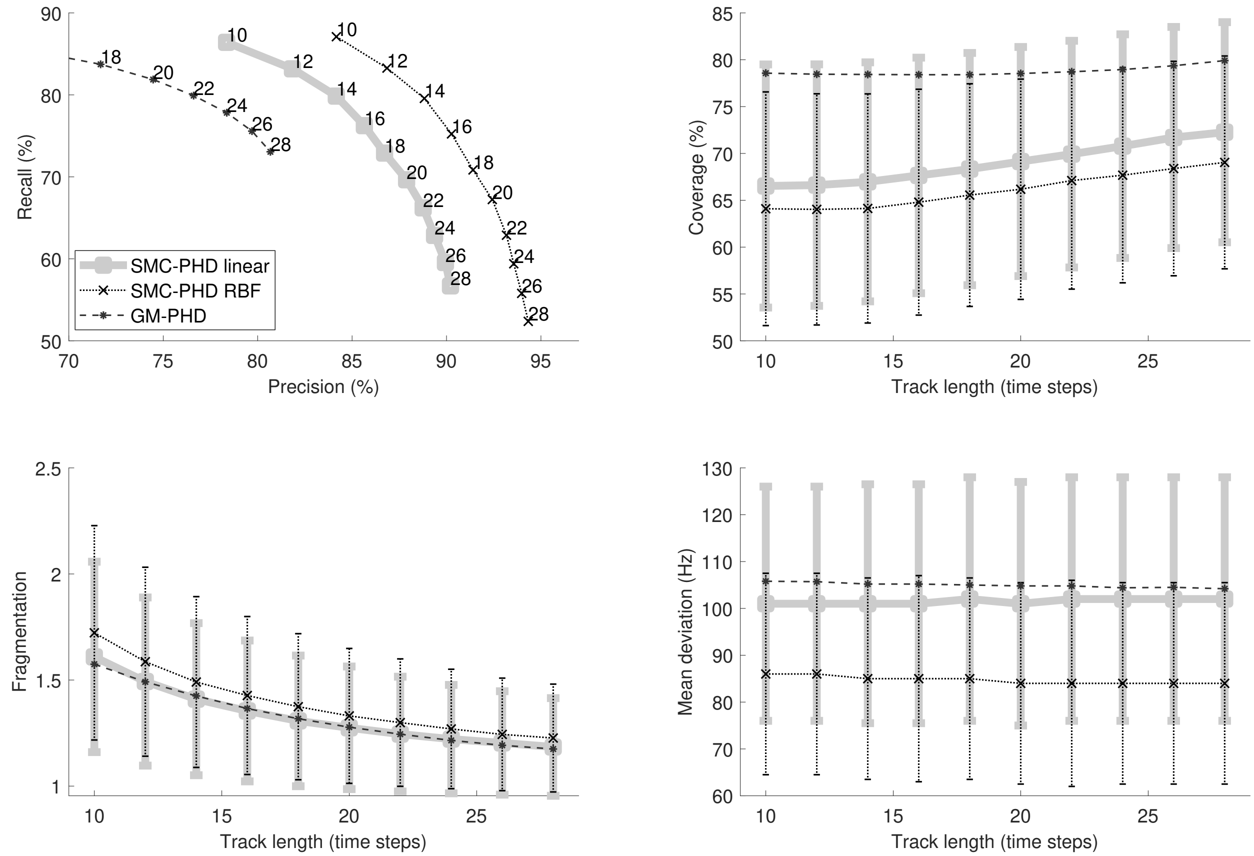

FIG. 1. (Color online) The performance of the SMC-PHD using a linear and RBF motion models across a range of track length criteria (from 10- 28 time steps; 53 - 150 ms). Error bars indicate 1 SD of a given metric. For comparison, the performance of the GM-PHD filter is plotted, but without the corresponding errorbars to preserve the figure's clarity. The performance was computed across all ground truth whistles that met the criteria.

Compared to the GM-PHD filter, both SMC-PHD versions had a better precision, but at the cost of a lower recall when longer track lengths were considered (Fig. 1). The GM-PHD recall and precision values for shorter track lengths are not displayed in Fig.1, to preserve the figure's clarity, but they change smoothly, reaching recall of 90% and precision of 40% for track length 10. This is a comparable recall to both SMC-PHD filters, but at much lower precision. Both SMC-PHD filter versions had a smaller coverage of the individual whistles compared to the GM-PHD filter, but the whistles were tracked more accurately, with a smaller mean deviation between the detection and the ground truth whistle (Fig. 1). All filters had a similar fragmentation rate (Fig. 1).

In terms of computational speed, both versions of the SMC-PHD algorithm were capable of tracking the whistles in real time. For example, a two minute file sampled at 192 kHz, containing 795 hand-annotated whistles, took 92.5 s and 117.5 s to be processed with the SMC-PHD filter with linear motion model and SMC-PHD filter with RBF motion model, respectively (implemented in MATLAB, Release R2016b, on a Mac, Os X, processor 2.7 GHz and 8 GB RAM).

An example of tracking by both versions of the SMC-PHD filter using a short and a long track lengths is shown in Fig. 2. It can be seen that the measurements, from which the filters tracked the whistles, contained a large amount of clutter, *i.e.*, measurements not associated with the whistles (Fig. 2, B). In agreement with the performance results in Fig. 1, it was seen that the SMC-PHD that used a linear motion model produced more false positive detections, for example it detected some of the echosounder pulses when using a track length of 10 time steps (Fig. 2, C) compared to the SMC-PHD that used RBF motion model (Fig. 2, E). It also had higher deviation from the annotated whistle path (for both track lengths) compared to the SMC-PHD filter that used RBF motion model (Fig. 2). However, in some whistles better coverage was achieved (a higher proportion of a given whistle was detected) compared to the SMC-PHD filter that used RBF motion model (Fig. 2). Moreover, in both filters the longer track length criteria resulted in fewer false positives compared to when shorter track length criteria were used (Fig. 2).

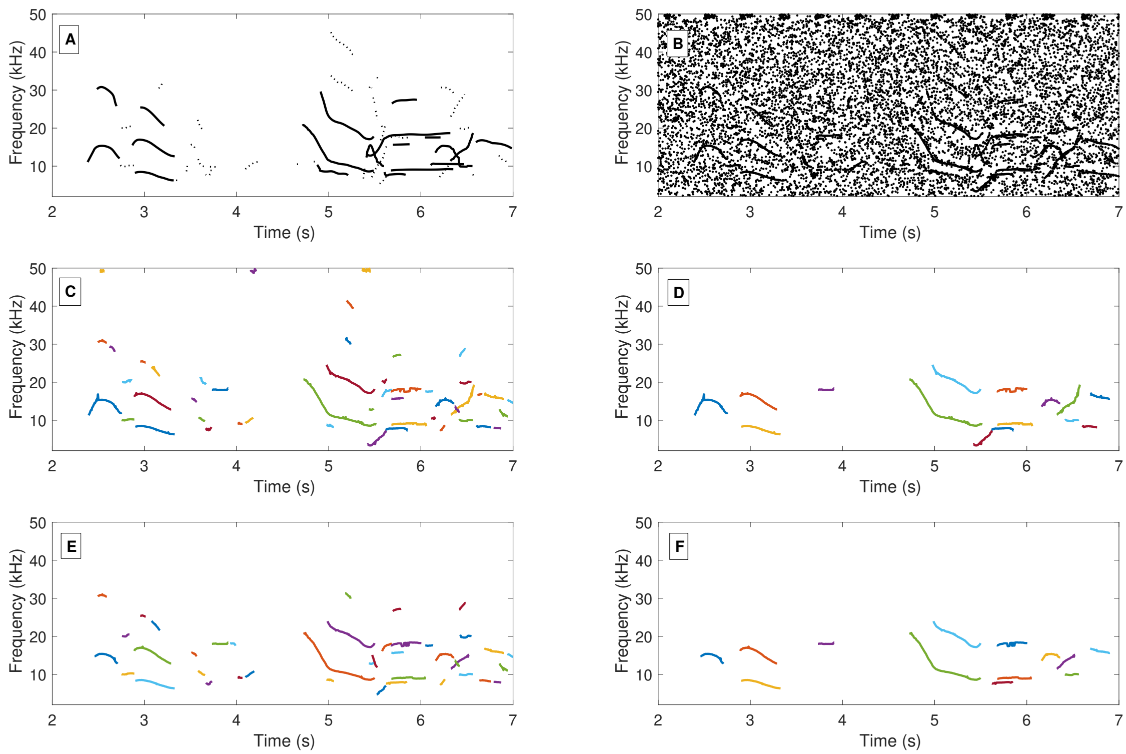

FIG. 2. (Color online) An example of the whistle tracking scenario. (A) Hand-annotated data (solid lines denote valid whistles, dashed lines not-valid whistles; for definition see Section II D). (B) Measurements (spectral peaks) - inputs for the SMC-PHD filters. Extracted whistles with the SMC-PHD filter that utilised linear motion model for track length 10 (C) and 28 (D) time steps. Extracted whistles with the SMC-PHD filter that utilised RBF motion model for track length 10 (E) and 28 (F) time steps.

In order to investigate the sensitivity of the SMC-PHD filter to the values of the input parameters, the best-performing version (using the RBF motion model and track length criteria of 10 time steps) was evaluated on the example shown in Fig. 2. The filter appeared to be insensitive to small deviations from the values in Table I, as seen by the similar performance for bins adjacent to these values in Fig. 3 (indicated with an "x"). However, large changes in some parameters (p_S, p_D, r, η) lead to significant drops in average performance and increase in performance variance. Changes in M_p and N_b did not appear to influence the average F1, but the performance dropped for $M_p < 15$. These results are representative of the behaviour of both SMC-PHD versions, but are not shown here due to space constraints.

It should be noted that the performance distribution in each bin of a given parameter in Fig. 3 represents multiple random draws of all the other parameters and therefore is not optimized. As such, the average F1 is lower than the performance obtained with the optimized parameter values in Table I, which result in F1= 87.6 and F1= 85 for the SMC-PHD filter that uses RBF and linear motion models, respectively.

The false positive detections were also examined in more detail. These were mainly due to interference from an echosounder and burst pulses, which can display a tonal quality in a spectrogram with the resolution chosen here, see Fig. 4.

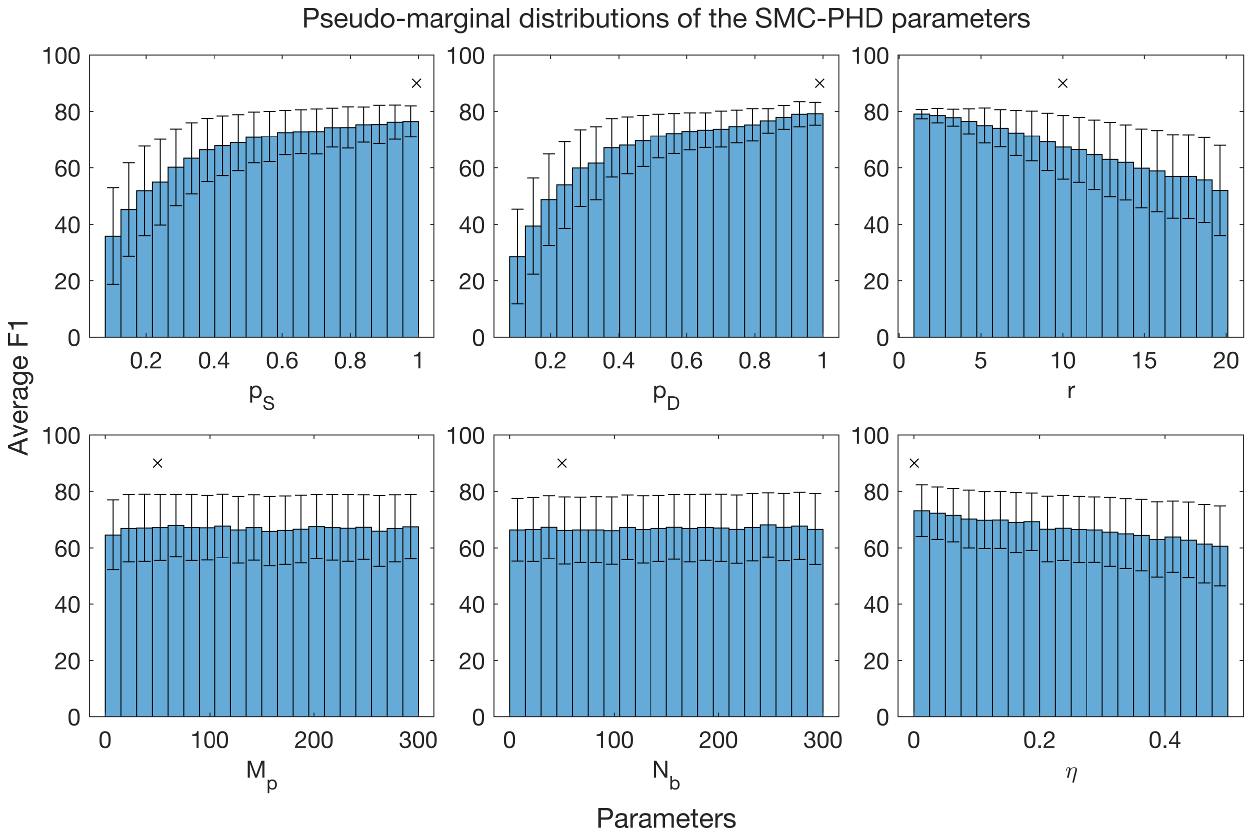

FIG. 3. (Color online) Pseudo-marginal distributions of the SMC-PHD parameters listed in Table I. Average F1 score per bin is shown, with the error bars indicating 1 SD. The values of the paramteres that were used in Table I are denoted by "x".

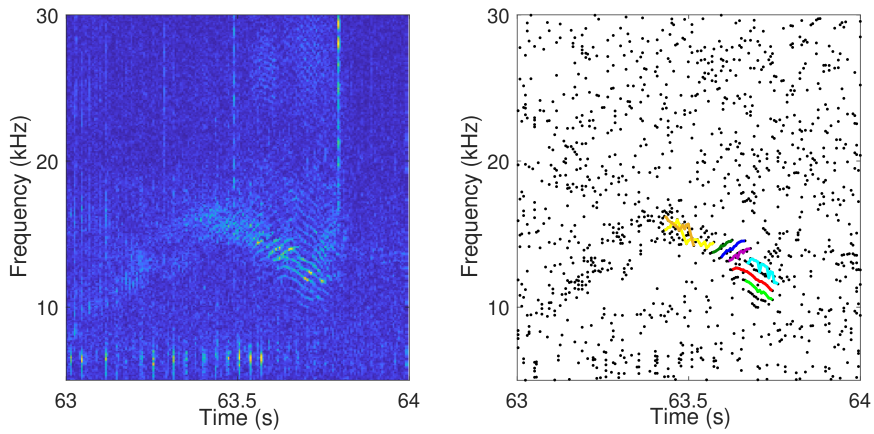

FIG. 4. (Color online) An example of the false positive detection of a burst pulse. A spectrogram of raw data is shown (left) and the measurements (dots) with detected false positives with the SMC-PHD filter (lines) are shown (right).

IV. DISCUSSION

The use of the RFS-based filters is a new approach in the field of the bioacoustics. This paper presents the first attempt to adapt these techniques, specifically the SMC-PHD filter, for the purpose of the frequency tracking of narrowband, frequency modulated signals from the underwater recordings. These methods provide a flexible framework for multi-target tracking, since they impose no restrictions on the form of the character of the underlying motion and measurement models nor do they assume a form for the various noise processes. The underlying models and parameters for this application are developed and the two proposed schemes are tested on a real world dataset comprising of dolphin whistles. The results showed the proposed filters are able to simultaneously extract multiple whistles from complex acoustic environments; they are able to track the whistles from highly cluttered measurements, through crossings with other whistles and points of missed detections; and they are suited for real time implementation. While the proposed filter implementation in Matlab was able to run in real-time on a typical desktop/laptop computer, it is assumed that significant gains in processing speed can be obtained by carefully implementing the filter in a more optimized programming language.

The proposed filters appear to generalize well. While the training data for the filters' parameters and models in Section II C consisted of only three delphinid species, the evaluated performance returned good results for all six delphinid species in the dataset (Fig. 1). Moreover, the recordings used for evaluation contained different whistle types, different noise conditions and amount of interfering signals, and were obtained with different recording equipment. This gives additional evidence that the method generalizes for different delphinid species, noise and recording conditions. However, the optimal values for parameters R, κ_k, and p_D are inherently linked to the choice of pre-processing parameters used to obtain the measurements. For example, the measurement noise variance R depends on the frequency bin resolution used in the spectrogram computation, and the clutter PHD κ_k and the probability of detection p_D depend on the spectrogram amplitude threshold.

The use of the SMC-PHD filter requires the development of specific models and parameters that govern the recursion. The sensitivity of the filter to the input parameter values was evaluated on a representative example that contained multiple overlapping whistles and noise sources, and that was not part of the training data. The filter appeared to be robust to small changes from the trained parameter values. However, large changes in parameters that influence the particle weights (p_S, p_D, r, η) led to a significant drop in performance and increased the variance in the performance results. The value of the parameters that control the number of particles per persistent and newborn whistles, M_p and N_b, did not appear to have a significant influence on the F1 performance. It should be noted, however, that increased number of particles will affect the computational speed of the algorithm, with larger number of particles slowing down the recursion. While the parameter values used in this study appear to give a good performance, there always remains some potential of performance improvement through the selection of a better parameter set. One alternative approach is to modify the filter so that it can adaptively adjust the parameter values during processing[45].

The proposed versions of the SMC-PHD filter were benchmarked against each other and against a different approximation to the PHD filter, the GM-PHD filter[11]. Both versions of the SMC-PHD filter appeared to have better precision and similar recall compared to the GM-PHD for short track length criterion, but at the cost of lower recall when longer track lengths are considered. Both versions of the SMC-PHD filter tracked whistles more accurately, with smaller mean deviation from the annotated whistle path, but at the cost of having smaller coverage of individual whistles compared to the GM-PHD filter.

The performance of the filters depends on the underlying models. A linear motion model describing the evolution of whistle contours was used in the GM-PHD filter and in one of the SMC-PHD filter versions. However, since the true motion model is unknown, it is advantageous to consider learning it from data rather than arbitrarily adopting a linear model. Learning the model from data, as was done for the SMC-PHD filter with RBF motion model, results in a non-linear model and thus requires the use of the SMC-PHD filter. It was seen that the precision of the filter with the non-linear RBF model was better, and this filter tracked individual whistles more accurately (with less deviation) compared to the two filters using linear models. The trade-off was a smaller coverage of individual whistles and slightly higher fragmentation compared to filters using linear models. It should be noted that the non-linear model employed here was trained on a relatively small subset of data, and future studies should consider models trained on larger datasets and consider employing non-Gaussian statistics for the noise processes.

The performance was measured based on the hand annotated ground truth data, that was subjective, as with all hand annotations, but at the same time reflected on the performance of the filters in the practical scenarios. As such the values of the performance of the filters should be taken as a guide, not an absolute measure of performance. For both system models, there was a general trade-off between the precision and the recall depending on the track length criteria (which specifies the minimum whistle contour length before it is classed as a detection). Shorter track lengths led to better recall but lower precision, since the number of false positive detections are increased. A shorter track length criterion also increased fragmentation in both instances. Depending on the requirements of the study, the track length criteria can be chosen appropriately.

To further improve the performance the following could be considered. This study utilized measurements that consisted only of the frequency peaks from a spectrogram, which makes this problem similar to that of bearing-only tracking in other applications. Adding additional information to the measurements, such as the amplitude or the chirp rate, and expanding the measurement model could potentially improve the performance[46] and should be investigated further.

Furthermore, in the present work the particle labeling approach for temporal association was chosen, since it does not add significantly to the computational load of the recursion. With the proposed labeling scheme, the identity conflicts (when multiple estimates were assigned the same identity at a given time step) were resolved outside the main PHD recursion and the particles from conflicting clusters propagated freely with the same labels. Although not reported here, a different approach was also tested, where the particles associated with the estimate that did not get assigned to the track were renamed (assigned a new identity). However, this did not produce better results. Another approach could be that instead of discarding the remainder of the conflicting estimates (estimates that are not assigned to a given track), these estimates would be compared against other tracks (with different labels) and assigned to different tracks as appropriate.

While the proposed filters successfully tracked dolphin whistles, it should be noted that any frequency modulated signals in the measurements would be extracted. On one hand, this can result in some false alarms that lower the precision of the filter. This was seen with the echosounder and burst pulses, which displayed a tonal quality due to the temporal resolution adopted in this study. It may be possible to remove these false alarms in post-processing steps. On the other hand, having the ability to detect burst pulses could be beneficial in certain applications. Moreover, these filters can be adapted for the extraction of baleen whale sounds or other frequency modulated sounds of interest.

V. CONCLUSIONS

This study considered the frequency tracking of dolphin whistle contours in the context of multi-target tracking. This was achieved with the use of the SMC-PHD filter, a practical approximation to the multi-target Bayesian filter. The filter was adapted and extended for the purpose of frequency tracking and specific models were introduced, resulting in two versions of the filter. The proposed SMC-PHD filters successfully tracked a time-varying number of overlapping whistles from highly cluttered measurements in the presence of false alarms and missed detections. The high degree of flexibility provided by these methods, allied to acceptable computational requirements, means that they are well-suited to real-time tracking of narrowband frequency modulated signals.

In addition, to facilitate comparisons of different methods, the measurement sets, a list of all raw audio files used in this study, as well as MATLAB implementation of the method for obtaining spectral peak measurements are openly available from the University of Southampton repository at https://doi.org/10.5258/SOTON/D0316. Moreover, the SMC-PHD filter implementation is available at https://github.com/PinaGruden/SMCPHD_whistle_contour_tracking.

ACKNOWLEDGMENTS

We would like to thank MobySound archive, DCLDE committee and associated analysts for providing the datasets and hand annotations used to test detectors performances in this study. We would also like to thank Slovene human resources development and scholarship fund (Ad futura) for funding this research.

REFERENCES

[1]T. A. Marques, L. Thomas, J. Ward, N. DiMarzio, and P. L. Tyack, "Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainvilles beaked whales," J. Acoust. Soc. Am. **125**(4), 1982–1994 (2009).

[2]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).

[3]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).

[4]P. Gruden, P. R. White, J. N. Oswald, Y. Barkley, S. Cerchio, M. Lammers, and S. Baumann-Pickering, "Differences in oscillatory whistles produced by spinner (*Stenella longirostris*) and pantropical spotted (*Stenella attenuata*) dolphins," Marine Mammal Sci. **32**(2), 520–534 (2016).

[5]N. J. Quick and V. M. Janik, "Whistle rates of wild bottlenose dolphins (*Tursiops truncatus*): influences of group size and behavior," J. Comp. Psychol. **122**(3), 305–311 (2008).

[6]C. R. Weir and S. J. Dolman, "Comparative review of the regional marine mammal mitigation guidelines implemented during industrial seismic surveys, and guidance towards a worldwide standard," J. Int. Wildlife Law Policy **10**(1), 1–27 (2007).

[7]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).

[8]A. Mallawaarachchi, S. H. Ong, M. Chitre, and E. Taylor, "Spectrogram denoising and automated extraction of the fundamental frequency variation of dolphin whistles," J. Acoust. Soc. Am. **124**(2), 1159–1170 (2008).

[9]D. K. Mellinger, S. W. Martin, R. P. Morrissey, L. Thomas, and J. J. Yosco, "A method for detecting whistles, moans, and other frequency contour sounds," J. Acoust. Soc. Am. **129**(6), 4055–4061 (2011).

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

[11] 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).

[12] S. Rankin, F. Archer, J. L. Keating, J. N. Oswald, M. Oswald, A. Curtis, and J. Barlow, “Acoustic classification of dolphins in the california current using whistles, echolocation clicks, and burst pulses,” Marine Mammal Sci. **33**(2), 520–540 (2017).

[13] M. Caillat, L. Thomas, and D. Gillespie, “The effects of acoustic misclassification on cetacean species abundance estimation,” J. Acoust. Soc. Am. **134**(3), 2469–2476 (2013).

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

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

[16] R. Mahler, “”Statistics 101” for multisensor, multitarget data fusion,” IEEE Aerosp. Electron. Syst. Mag. **19**(1), 53–64 (2004).

[17] R. P. Mahler, *Statistical multisource-multitarget information fusion* (Artech House, Inc., Norwood, MA, USA, 2007), p. 856.

[18] 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, US (2000), pp. 99–117.

[19]R. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," IEEE Trans. Aerosp. Electron. Syst. **39**(4), 1152–1178 (2003).

[20]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).

[21]Y.-D. Wang, J.-K. Wu, A. A. Kassim, and W. Huang, "Data-driven probability hypothesis density filter for visual tracking," IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1085–1095 (2008).

[22]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).

[23]T. M. Wood, C. A. Yates, D. A. Wilkinson, and G. Rosser, "Simplified multitarget tracking using the PHD filter for microscopic video data," IEEE Trans. Circuits Syst. Video Technol. **22**(5), 702–713 (2012).

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

[25]B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo implementation of the PHD filter for multi-target tracking," in *Proceedings on International Conference on Information Fusion* (2003), pp. 792–799.

[26]T. Zajic and R. P. Mahler, "Particle-systems implementation of the PHD multitarget-tracking filter," in *Proceedings of SPIE* (2003), Vol. 5096, pp. 291–299.

[27]K. E. Frasier, E. Elizabeth Henderson, H. R. Bassett, and M. A. Roch, "Automated identification and clustering of subunits within delphinid vocalizations," Marine Mammal

Science **32**(3), 911–930 (2016).

[28]S. Baumann-Pickering, S. M. Wiggins, J. A. Hildebrand, M. A. Roch, and H.-U. Schnitz, "Discriminating features of echolocation clicks of melon-headed whales (*Peponocephala electra*), bottlenose dolphins (*Tursiops truncatus*), and Gray's spinner dolphins (*Stenella longirostris longirostris*)," J. Acoust. Soc. Am. **128**(4), 2212–2224 (2010).

[29]B.-N. Vo, M. Mallick, Y. Bar-Shalom, S. Coraluppi, R. Osborne III, R. Mahler, and B.-T. Vo, "Multitarget tracking," in *Wiley Encyclopedia of Electrical and Electronics Engineering* (John Wiley and Sons, Inc., 2015), pp. 1–23.

[30]B. Ristic, D. Clark, and B.-N. Vo, "Improved SMC implementation of the PHD filter," in *13th Conference on Information Fusion 2010*, IEEE (2010), pp. 1–8.

[31]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).

[32]B. Ristic, M. Beard, and C. Fantacci, "An overview of particle methods for random finite set models," Inf. Fusion **31**, 110–126 (2016).

[33]K. Panta, B.-N. Vo, and S. Singh, "Improved probability hypothesis density (PHD) filter for multitarget tracking," in *Third International Conference on Intelligent Sensing and Information Processing, 2005. ICISIP 2005.*, IEEE (2005), pp. 213–218.

[34]D. E. Clark and J. Bell, "Multi-target state estimation and track continuity for the particle PHD filter," IEEE Trans. Aerosp. Electron. Syst. **43**(4), 1441 – 1453 (2007).

[35]K. Panta, B.-N. Vo, and S. Singh, "Novel data association schemes for the probability hypothesis density filter," IEEE Trans. Aerosp. Electron. Syst. **43**(2), 556–570 (2007).

[36]T. Li, M. Bolic, and P. M. Djuric, "Resampling methods for particle filtering: classification, implementation, and strategies," IEEE Sign. Process. Mag. **32**(3), 70–86 (2015).

[37]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).

[38]L. Wang, L. Zhang, and Z. Yi, "Trajectory predictor by using recurrent neural networks in visual tracking," IEEE Trans. Cybern. **47**(10), 3172–3183 (2017).

[39]C. M. Bishop, *Neural networks for pattern recognition* (Oxford University Press, Inc., New York, NY, USA, 1995), pp. 164–191.

[40]D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics (2007), pp. 1027–1035.

[41]B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multitarget filtering with random finite sets," IEEE Trans. Aerosp. Electron. Syst. **41**(4), 1224–1245 (2005).

[42]C. M. Bishop, *Pattern recognition and machine learning* (Springer Science & Business Media, New York, NY, USA, 2006), p. 738.

[43]C. M. Jarque and A. K. Bera, "A test for normality of observations and regression residuals," Int. Stat. Rev. **55**(2), 163–172 (1987).

[44]B. Ristic, "Efficient update of persistent particles in the SMC-PHD filter," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015*, IEEE (2015), pp. 4120–4124.

[45]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).

[46]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).

Algorithm 1 Pseudo-code of the SMC-PHD filter for whistle contour tracking (adapted based on Ref.[32])

Input $\mathcal{P}_{k-1} \equiv \{w_{k-1}^{(i)}, \boldsymbol{x}_{k-1}^{(i)}\}_{1 \le i \le N_{k-1}}; \boldsymbol{Z}_k$

Step 1 Prediction

Draw particles from proposal density to obtain $\boldsymbol{x}_{k|k-1}^{(i)}$ ▷ see Section II C 2

Compute their weights: $w_{k|k-1}^{(i)} = p_S w_{k-1}^{(i)}$

Step 2 Update, Resampling, State Estimation

Partition $\{w_{k|k-1}^{(i)}, \boldsymbol{x}_{k|k-1}^{(i)}\}_{1 \le i \le N_{k-1}}$ to form clusters $C_{k|k-1}(z), z \in \boldsymbol{Z}_k \cup \emptyset$

Initialize $\mathcal{P}_k = \emptyset$, $\hat{\boldsymbol{X}}_k = \emptyset$

for every $z \in \boldsymbol{Z}_k$ **do**

  if $C_{k|k-1}(z) \neq \emptyset$, it consists of M weighted particles $\{w_{k|k-1}^{(m)}, \boldsymbol{x}_{k|k-1}^{(m)}\}_{1 \le m \le M}$ **then**

    Update their weights: $\hat{w}_k^{(m)} = \frac{p_D g_k(z|\boldsymbol{x}_{k|k-1}^{(m)}) w_{k|k-1}^{(m)}}{\kappa_k + p_D \sum_{n=1}^{M} g_k(z|\boldsymbol{x}_{k|k-1}^{(n)}) w_{k|k-1}^{(n)}}$

    Compute probability of cluster's existence: $p_e(z) = \sum_{m=1}^{M} \hat{w}_k^{(m)}$

    Resample based on $\hat{w}_k$ to generate M_p particles $\boldsymbol{x}_k^{(l)}, l = 1, \cdots, M_p$

    Set the resampled particle weights to $w_k^{(l)} = p_e(z)/M_p, l = 1, \cdots, M_p$

    $\{w_k^{(l)}, \boldsymbol{x}_k^{(l)}\}_{1 \le l \le M_p}$ represent updated cluster $C_k(z)$, and $\mathcal{P}_k = \mathcal{P}_k \cup C_k(z)$

    if $p_e(z) > \eta$ **then** ▷ η is a threshold determined in Section II C 4

      Estimate whistle state $\hat{\boldsymbol{x}}_k$ from $C_k(z)$: $\hat{\boldsymbol{x}}_k = 1/M_p \sum_{l=1}^{M_p} \boldsymbol{x}_k^{(l)}$

      $\hat{\boldsymbol{X}}_k = \hat{\boldsymbol{X}}_k \cup \{\hat{\boldsymbol{x}}_k\}$

    end if

  end if

end for

for every pair $(w_{k|k-1}, \boldsymbol{x}_{k|k-1}) \in C_{k|k-1}(\emptyset)$ **do**
  if $w_{k|k-1} > \xi$ **then** ▷ ξ is a threshold determined in Section II C 4
    Update weights as: $w_k = (1 - p_D) w_{k|k-1}$
    And add the weighted particles to $\mathcal{P}_k$
  end if
end for
Step 3 Whistle birth
for every $z \in \boldsymbol{Z}_{b,k}$ **do**
  Generate N_b particles and compute their weights ▷ see Section II C 3
  Add the newborn weighted particles to $\mathcal{P}_k$
end for
Output: $\mathcal{P}_k \equiv \{w_k^{(i)}, \boldsymbol{x}_k^{(i)}\}_{1 \le i \le N_k}; \hat{\boldsymbol{X}}_k$

TABLE I. Summary of the parameters used in the SMC-PHD filter for dolphin whistle tracking. p_S and p_D denote the probabilities of survival and detection respectively; r denotes the average number of clutter measurements per time step; M_p and N_b denote the number of particles per persistent and newborn whistle respectively; η denotes state estimation threshold; ξ denotes particle elimination threshold, where M is the number of particles in cluster $C_{k|k-1}(\emptyset)$.

p_S	p_D	r	M_p	N_b	η	ξ
0.994	0.99	10	50	50	0.0005	1/M

FIGURE CAPTIONS

Fig.1 (Color online) The performance of the SMC-PHD using a linear and RBF motion models across a range of track length criteria (from 10- 28 time steps; 53 - 150 ms). The performance is computed across all ground truth whistles that met the criteria and is not the average of file or species performances. Each error bar indicates one standard deviation of a given metric.

Fig.2 (Color online) An example of the whistle tracking scenario. (A) Hand-annotated data (solid lines denote valid whistles, dashed lines not-valid whistles; for definition see Section II D). (B) Measurements (spectral peaks) - inputs for the SMC-PHD filters. Extracted whistles with the SMC-PHD filter that utilised linear motion model for track length 10 (C) and 28 (D) time steps. Extracted whistles with the SMC-PHD filter that utilised RBF motion model for track length 10 (E) and 28 (F) time steps.

Fig.3 (Color online) Pseudo-marginal distributions of the SMC-PHD parameters listed in Table I. Average F1 score per bin is shown, with the error bars indicating 1 SD. The values of the paramteres that were used in Table I are denoted by "x".

Fig.4 (Color online) An example of the false positive detection of a burst pulse. A spectrogram of raw data is shown (left) and the measurements (dots) with detected false positives with the SMC-PHD filter (lines) are shown (right).