Acoustic detector for multiple vocalizing marine mammal individuals

Simen Hexeberg*, Hari Vishnu*, Koay Teong Beng*, Abel Ho*,
Wu Yusong*, Mandar Chitre*, Karenne Tun+, Karen Lim+
*Acoustic Research Laboratory, Tropical Marine Science Institute, Singapore
+ National Parks Board, Singapore

***Abstract*—Marine mammals play an important role in the marine ecosystems but are increasingly threatened by human activities. To better protect these species it is helpful to understand their spatio-temporal visiting patterns and to estimate their population sizes. In this work we present a non-invasive method to aid conservation efforts, using a multi-channel passive acoustic monitoring system. We test the system in Singapore's noisy coastal waters and show that the processing developed herein can automatically classify marine mammal whistles depending on shape, estimate their direction of arrival and indicate whether a set of detected whistles that occur temporally nearby are likely to have arisen from multiple vocalizing individuals.**

***Index Terms*—Passive Acoustic Monitoring (PAM), whistle detection, computer vision, DOA estimation, TDOA**

I. Introduction

Dolphins are intelligent apex predators who play an important role in keeping the marine ecosystem in balance. Several of its species, including the Indo-Pacific Humpback Dolphin found in Singapore waters [1], are classified as vulnerable by the International Union for Conservation of Nature [2]. Monitoring their locations and population sizes can aid conservation efforts to protect these species. As dolphins are vocal animals who depend on acoustics for tasks such as echolocation, foraging and communication, it is possible to detect them by means of passive acoustic monitoring (PAM) systems. PAM enables continuous monitoring over long periods of time and without interfering with the natural habitat of these animals.

Many methods are developed over the past few years to detect marine mammals by means of passive acoustics. Most approaches use spectrogram-representations to reveal patterns that are visually harder to detect in waveforms. These spectrogram-based methods can be broadly separated into two classes. In the first class are vocalizations detected and classified by searching in fixed time windows [3], [4], but without knowing where in time and frequency the detection occurred within that window. The second class of methods obtain accurate time-frequency information by searching at the level of individual time-frequency bins to detect signal profiles. This class of methods, however, are mainly applied to whistle detection by exploiting that whistles are narrowband frequency modulated (FM) signals. Some approaches includes threshold-based edge-detectors, e.g. [5], particle-filter based detectors, e.g. [6], and more recently, convolutional neural network based detectors [7]. This class of methods has the advantage of easily generalizing to unseen whistles, as they are not trained to detect specific whistle profiles. This is useful in environments with rich, complex whistles, but it has the downside of potential high false positive rate (FPR) if other non-biological FM signals are present, which is often the case in Singapore's busy waters.

In order to identify potential marine mammal vocalizations as events-of-interest (EOIs) from raw acoustic recordings, we have developed a machine learning (ML) detector to detect whistles [8]. Given these EOIs, the natural next step would be to use them to get an insight on population numbers or density estimates. In the current work, given the EOIs by the ML detector, we attempt to detect when more than one marine mammal are present by estimating the direction of arrival (DOA) of detected vocalizations. We accomplish this by leveraging recent advances in computer vision and applying it on multi-channel acoustic data. While we focus on dolphins in this work, the approach can be used for any vocalizing marine mammal. Since the proposed algorithm requires temporally dense whistles in the recordings, and since we cannot guarantee that all dolphins in an area vocalize at similar times, the estimate of this method is a lower bound on the number of dolphins present at a given time. It gives us some idea about the occurrence of dolphin groups and their preferred locations and times, thus shedding light on their behaviour and aiding conservation efforts. We choose to focus on whistle sounds because they are more straightforward to detect in the time-frequency representations, despite the limited bandwidth of these signals (as compared to e.g. echolocation clicks) which makes DOA estimation more challenging.

In Section II of this paper, we outline in detail the methodology used for detecting multiple vocalizing marine mammal individuals, including the hardware setup and the different components of the detector. In Section III, we present results based on this technique using data recorded in Singapore waters, and discuss them. In Section IV, we describe the envisioned experimental setup where this system is planned to be used, and in Section V we conclude the paper.

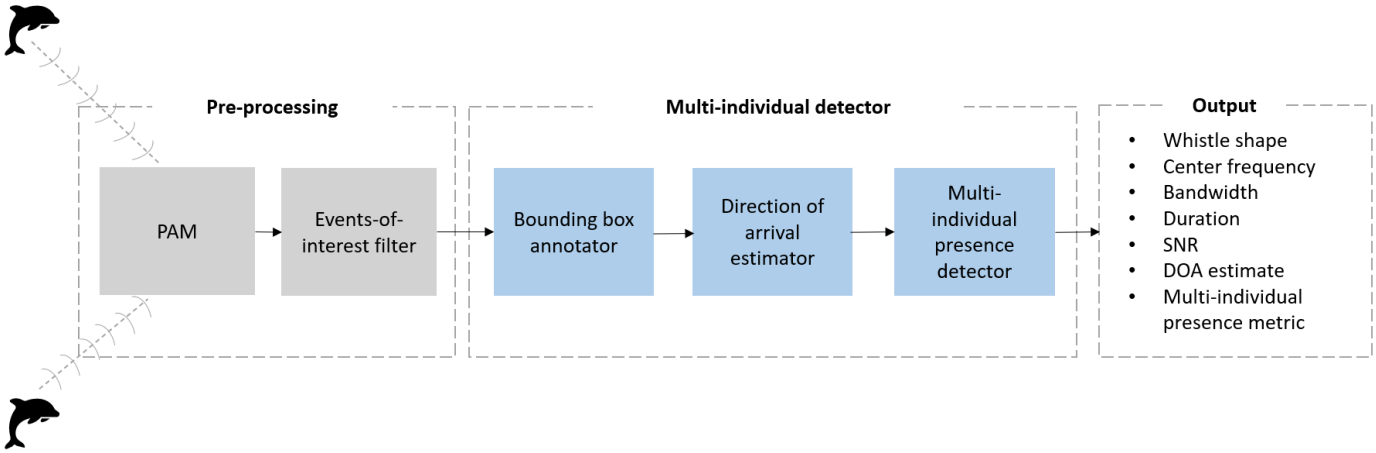

Fig. 1. High-level multi-individual detector pipeline.

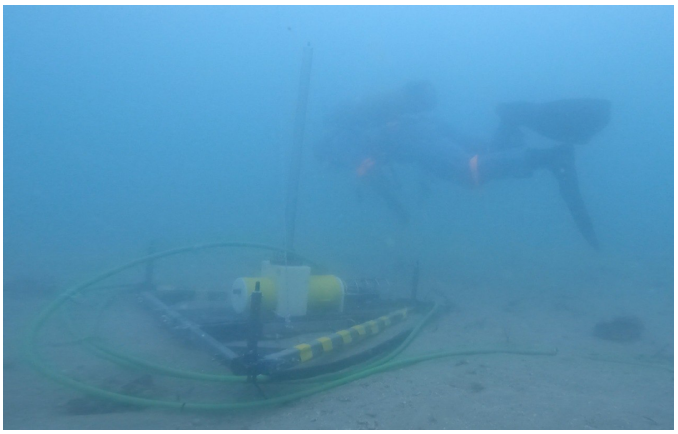

Fig. 2. Image of the 4-channel PAM array used in this work. The PAM is placed on the seabed at approximately 15 m depth and the hydrophones are arranged in a tetrahedron with 1 m spacing between each pair.

II. Methodology

The multi-individual detector consists of several modules, illustrated in Fig. 1 and explained in more details in the following sections.

A. Hardware setup

We use a 4-channel PAM array [9] to detect the presence of multiple dolphins acoustically. The hydrophones are arranged in a tetrahedron with 1 m spacing between each pair and shown in Fig. 2. This 3-D geometry enables us to estimate signal directions in 3-D space, which we leverage to detect the presence of multiple vocalizing individuals. This system will be tested as part of the marine environmental sensing network being developed for monitoring across Singapore waters.

B. Events-of-interest filter

First, an ML-detector is used to obtain a list of EOIs from raw long-duration acoustic recording timeseries. The EOIs are short audio clips containing one or more marine mammal vocalizations. The ML detector is a binary classifier convolutional neural network trained on signal and noise samples [8], but its details are not discussed in this paper. Given these EOIs, the aim of the current work is to detect scenarios amongst these where the vocalizations may be attributed to multiple dolphins (rather than a single individual).

C. Bounding box annotator

The EOI filter is followed by an automatic bounding-box annotator in the time-frequency domain, which uses as its base model YOLO [10], a computer vision model commonly used for object detection tasks. We make us of a model pre-trained on the COCO data set [11] and custom train it on an application-specific semi-synthetic data set. This data set consists of computer-generated FM signals designed to mimic whistles of biological origin added on top of real ambient noise recordings from Singapore waters. The FM signals are generated with variation in shape, duration, bandwidth, center frequency and signal-to-noise ratio (SNR) similar to known dolphin whistles in the literature [12], [13]. The annotator model is trained to detect and classify six different type of whistles based on the signal's shape, start frequency and end frequency. These classes are upsweeps and downsweeps of linear, convex and concave shaped FM signals.

Acoustic recordings are split into 3-second-long windows, converted to images via spectrograms and fed to the annotator. Only data from one of the four hydrophones are used in this part. The annotator searches each image for the pre-trained classes and outputs bounding boxes around detected signals if the confidence level is above a pre-defined threshold. Two examples of detections from Singapore waters can be seen in Fig. 3. The bounding box annotator naturally provides useful meta data such as duration, bandwidth, and center frequency of detected signals. We use these meta data to detect harmonics in order to separate individual calls. Further, and more importantly, the bounding boxes enable effective signal

denoising in the time-frequency domain by removing all noise surrounding the boxes and, as a result, significantly increase the accuracy in the downstream DOA algorithm.

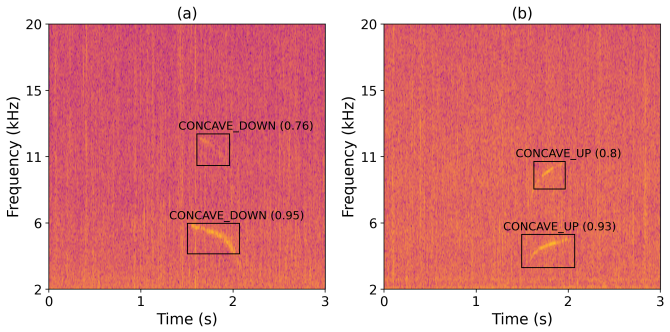

Fig. 3. Examples of two events detected by the bounding box annotator in Singapore waters. The predicted class is above each predicted box, followed by the detection confidence given as a value between 0 (low) and 1 (high). All four signals in these two examples are correctly classified by the annotator according to their shape and the harmonics are correctly identified by the harmonic detector.

D. DOA estimator

A sensor array with $n+1$ sensors has $n(n+1)/2$ distinct time differences of arrival (TDOAs) between sensor pairs, denoted here the *complete set*. This TDOA information can be used to localize the sources. The complete set has redundancy as linear combinations of only n independent TDOAs are sufficient to obtain all remaining TDOAs, given a noise-free environment. The condition of independence of the TDOAs ensures that the n TDOAs used cover measurements from $n+1$ sensors. In a challenging and noisy environment such as the one encountered in Singapore, however, one or more of the TDOAs may be prone to errors. TDOA errors tend to be divided into two classes, namely errors caused by (1) measurement noise (typically small) and (2) errors caused by multipath or noise interference, often referred to as TDOA outliers and typically larger. Including TDOA outliers in DOA estimates often leads to large errors, which we wish to avoid. Instead, we aim to achieve estimates that are more robust to TDOA outliers by exploiting the redundancy of the complete set, for which some approaches already exist in the literature. Some techniques, such as [14], use cross-correlation peaks assisted by auto-correlation peaks and cyclic zero-sum conditions[1] to identify individual sources in a multi-source multipath environment. Others use heuristics to implement cross-correlation peak picking strategies [15]. A third class of techniques identifies outliers by evaluating how much each TDOA contributes to a residual of a designed cost function [16], [17]. We present next a voting-based algorithm for DOA estimation, building on the principle of TDOA redundancy in the complete set, which we thereafter expands with a simple cross-correlation peak picking method.

[1]A necessary (but not sufficient) requirement for error-free TODAs is that they sum up to zero along a cyclic path, i.e., a path starting and ending at the same sensor.

1) DETSAC: The Deterministic sample consensus (DETSAC) algorithm presented in this work is a voting-based DOA estimation algorithm inspired by the Random sample consensus (RANSAC) algorithm [18]. Since 3 independent TDOAs are sufficient to estimate the DOA in 3D-space, and the PAM array used here has 4 hydrophones and hence 6 independent TDOAs in total, some redundancy is available to exploit for robustness against TDOA outliers. Direction ambiguity is introduced if using a set of only 2 independent TDOAs, resulting in two possible solutions of which only one can be correct. These smaller TDOA sets can be exploited to increase accuracy despite this ambiguity, which is the main motivation behind DETSAC. In short, DETSAC works by (i) collecting a pool of direction estimates from smaller sets of independent TDOAs from the complete set, (ii) cluster the directions in the pool, and (iii) use the centroid direction from the largest cluster as the final DOA esitmate. The full algorithm is presented in algorithm 1. For the direction clustering in step 8 we use the Density-based spatial clustering of applications with noise (DBSCAN) [19] algorithm with minimum cluster size of 1 and distance parameter[2] $\epsilon = 3°$. DETSAC is based on the following intuition: in a system with some error-free TDOAs and some outliers, the sets of independent TDOAs containing no outliers should produce more consistent DOA estimates than the sets where one or more outliers are present. Hence, even if the error-free sets are outnumbered by sets with errors, they may still yield the largest number of consistent directions and hence form the largest cluster.

Algorithm 1 DETSAC

Input: Complete TDOA set
Output: DOA estimate $\hat{\theta} = [\hat{\alpha}, \hat{\epsilon}]$

1: $\mathcal{D} \leftarrow \emptyset$ ▷ Start with an empty pool of directions
2: $\mathcal{T}_2 \leftarrow$ All independent TDOA sets of size 2
3: $\mathcal{T}_3 \leftarrow$ All independent TDOA sets of size 3
4: **for each** TDOA set $\in \mathcal{T}_2$ **do**
5: Obtain directions θ_1 and θ_2
6: $\mathcal{D} \leftarrow [\theta_1, \theta_2]$ ▷ Add directions to pool
7: **end for**
8: $C_1, ..., C_n \leftarrow$ Cluster directions in $\mathcal{D}$
9: **if** the two largest clusters C_1, C_2 are of equal size **then**
10: Repeat step 4-8 with $\mathcal{T}_3$
11: **end if**
12: $\hat{\theta} \leftarrow$ Get centroid direction of the largest cluster C_1

Including pre-processing steps, the following sequence is applied to each detected bounding box to estimate DOA in terms of azimuth (α) and elevation (ϵ) angles:

1) Compute short-time Fourier transform (STFT) of the data.
2) Perform time-frequency data denoising by zeroing out everything outside the bounding box in the STFT.

[2]Two points are considered neighbors if the angular distance between the two points is below the threshold ϵ.

3) Inverse transform the denoised signal from time-frequency domain to time domain (timeseries).
4) Obtain the complete set of TDOAs by locating the cross-correlation peaks.
5) Run DETSAC with the complete set of TDOAs to obtain the DOA estimate.

The main purpose of the time-frequency denoising step is to reduce the effect of noise on the estimate. A secondary effect is that isolating each detected signal makes the system robust to multiple sources (given no or minimal source overlap in both time and frequency).

2) Flexible ArgMax: Next, we implement a simple cross-correlation peak picking method which we refer to as *Flexible ArgMax*. The sample delay between two sensors is typically estimated by locating their cross-correlation peak. The highest peak, however, may not always be the correct one. This problem increases as signal bandwidth or SNR is reduced. The Flexible ArgMax algorithm adds flexibility in the correlation peak selection by allowing either the 1st or 2nd peak to be used. With 6 TDOAs in the complete set, this result in 2^6 complete set combinations. We apply DETSAC to each combination, cluster the resulting pool of DOA estimates and set the final estimate to the centroid of the largest cluster. It should be noted that this approach is computationally expensive and does not scale well, neither in terms of the number of sensors nor in terms of the number of peaks to search for. Reducing the computational cost while allowing the same flexibility is likely possible but has not been attempted in this work.

3) Ensemble algorithm: Testing shows that DETSAC with Flexible ArgMax does not perform strictly better than standard DETSAC. It is more accurate for some inputs and less accurate for others. An ensemble of the two algorithms can therefore achieve higher performance than any one of the algorithms alone. The ensemble selects the DOA estimate with the highest confidence between the two algorithms. The estimation confidence is described next.

4) DOA confidence model: For any complete set of TDOAs, DETSAC returns a DOA estimate which may or may not be contaminated by large errors. Consequently, it is imperative to assign a confidence level to the DOA estimate as a measure of uncertainty. We train a random forest binary classifier to predict the estimation confidence by feeding it features representing constraints arising out of the array geometry (e.g. number of cyclic zero-sum paths) and features related to the sizes of the DOA clusters (e.g. size gap between the two largest clusters and the percentage of all DOA estimates in the largest cluster).The classifier is trained on semi-synthetic data and outputs a value between 0 (low confidence) and 1 (high confidence). The error cut-off to separate the two classes is set to 2° and one confidence model is trained for each DOA algorithm (DETSAC with and without Flexible ArgMax).

TABLE I
PARAMETER LIST FOR THE MULTI-INDIVIDUAL PRESENCE ALGORITHM

Parameter	Symbol	Value	Comment
Time between detections	Δt		Measured
Angular separation	$\Delta\hat{\theta}$		Estimated
Elevation angle	ϵ_i		Estimated
Swim speed	$\hat{v}$	1.4 m/s	Assumed
Maximum detection range	$R_{\max}$	250 m	Assumed
Depth of PAM array	S_d	15 m	Measured

E. Evaluation of multi-individual presence

Lastly, the multi-individual detector searches for high-probability examples of multiple individuals by leveraging the output from the previous stages. The algorithm assumes the PAM array rests on the seabed. Given two temporally close detections, the algorithm computes a 'multi-individual presence metric' (MIPM) indicating how likely the two calls originate from two individuals (rather than one). The MIPM varies between 0 (improbable) and 1 (very likely) and is computed based on the parameters in table I. Note that the maximum detection range and swim speed are assumed quantities. The latter is set to the swim speed for Indo Pacific Humpback Dolphins as reported by [20].

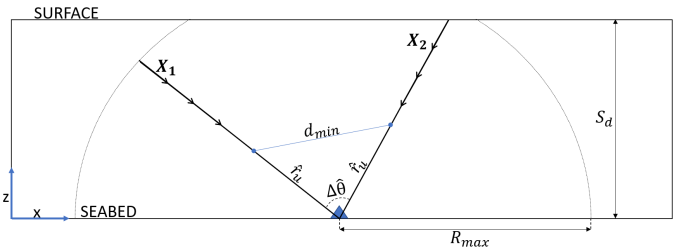

Fig. 4. Illustration of the multi-individual detector.

A basic intuition behind the method is that two temporally close detections with large angular separation are likely to have originated from different individuals. However, this becomes less likely for higher elevation angles, in which case the array-to-source range is upper bounded by the array-to-surface distance at that angle. For example, if a whistle arrives at $\epsilon = 90^\circ$ (from straight above the array), the individual's range is limited by the array depth S_d (15 m). Consequently, for a given swim-speed, the dolphin may easily change its relative direction to the PAM by a large angle in a short time. Likewise, low elevation angles correspond to a longer feasible array-to-source range, limited by the detection range of our algorithm rather than the array-to-surface distance, as illustrated in Fig. 4.

We denote a detected vocalization as $\mathbf{X_i} = [\hat{\theta}_\mathbf{i}, t_i, \hat{r}^i_{\max}]$ where $\hat{\theta}_\mathbf{i} = [\hat{\alpha}_i, \hat{\epsilon}_i]$ is the estimated azimuth and elevation angles in radians, respectively, t_i is the time of the detection

in seconds, and $\hat{r}^i_{\text{max}}$ is an estimate of the longest feasible array-to-source distance, given by

$$\hat{r}^i_{\text{max}} = \min\left(\frac{S_d}{\sin(\epsilon_i)}, R_{\text{max}}\right). \quad (1)$$

Note that the elevation in (1) is defined in a global reference frame where $\epsilon = 0°$ lies in the plane perpendicular to the gravitational force and increases in the upward direction. Now, given two detected vocalizations $\mathbf{X_i}$ and $\mathbf{X_j}$, we calculate $\hat{r}_u$ as the array-to-source upper bound distance at which a single individual can generate both vocalizations if swimming the shortest possible path that results in the angular distance $\Delta\hat{\theta}$ in Δt time:

$$\hat{r}_u = \frac{d_{\text{min}}}{2\sin(\frac{\Delta\hat{\theta}}{2})}, \quad (2)$$

where $d_{\text{min}} = \hat{v}\Delta t$ is the distance along the shortest path when swimming with constant speed $\hat{v}$. Finally, the MIPM is given by

$$MIPM = \begin{cases} 1 - \frac{\hat{r}_u}{\hat{r}_{\text{max}}}, & \text{if } \hat{r}_{\text{max}} \geq \hat{r}_u \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where $\hat{r}_{\text{max}} = \min\left(\hat{r}^i_{\text{max}}, \hat{r}^j_{\text{max}}\right)$ is the smallest of the two detections' maximum estimated ranges, ensuring the MIPM is invariant to their temporal order.

To illustrate, lets first assume that two detected vocalizations originate from the same individual and that the estimated maximum ranges are $\hat{r}^1_{\text{max}} = 100$ m and $\hat{r}^2_{\text{max}} = 200$ m. Consequently, the individual must have been somewhere in the range of 0-100 m from the array at some point in time between the two vocalizations, i.e., $\hat{r}_{\text{max}} = 100$ m. Lets further assume that the array-to-source upper bound distance was found to be $\hat{r}_u = 30$ m, meaning that the individual must have been at most 30 m from the array to have sufficient time to move location between the two calls. Assuming the probability of an individual being located at a given distance from the array is uniformly distributed[3] over the feasible range $\hat{r}_{\text{max}}$, we can estimate the probability that the two vocalizations originated from a single individual as $P_1 = \frac{\hat{r}_u}{\hat{r}_{\text{max}}} = \frac{30 \text{ m}}{100 \text{ m}} = 0.3$. The probability that the calls originate from two distinct individuals is then simply $MIPM = 1 - P_1 = 0.7$.

III. Results and Discussion

The detector has been tested in Singapore waters over a period of 8 months, searching for whistles in the 2-20 kHz frequency band. The EOIs are first identified using a ML detector [8] (not discussed here), and then further shortlisted through manual checking. The resulting data set contains 592 whistles. The performances of the multi-individual detector's different modules are presented next. Modules are either evaluated on semi-synthetic simulated data, real data, or both.

[3]Since detection rate drops with lower SNR it may be more accurate with a probability distribution that decays with increasing distance.

A. Bounding box annotator

The bounding box annotator has a detection rate of 62.0% on the real data set with a mean prediction confidence of 55.8%. It picks up non-biological FM signals from the background in a few cases, yielding a false positive rate of 5.0% but with a lower mean prediction confidence of 29.7%. The missed detections arise in cases with either very low SNR or where the signals differ significantly from the synthetic signals seen during training. Two such examples can be seen in Fig. 5. Hence, the detection rate may be improved by adding signals resembling the missed detections to the training set.

The bounding box annotator further classifies the signal sweep (up/down) with an accuracy of 97.4% when evaluated on whistles with a distinct upsweep or downsweep (87.3% of data).

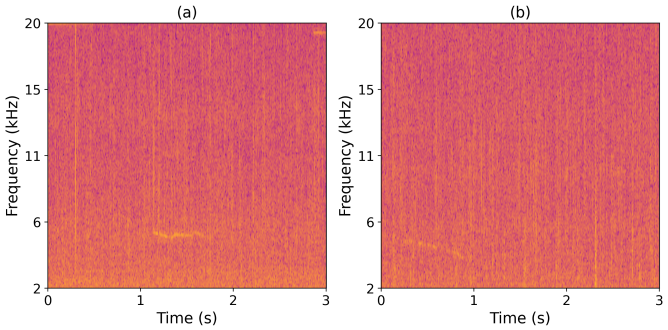

Fig. 5. Examples of two signals not detected by the bounding box annotator. The signal in (a) has sufficiently high SNR but its shape and bandwidth differs too much from the training data, while the signal in (b) has a shape that could be detected but too low SNR. To be clear, the signal in (b) is a downsweep located in the bottom left corner, approximately 1 second long.

B. DOA estimator

In the absence of DOA ground truth for the detected whistles we use a semi-synthetic data set to evaluate the performance of the DOA algorithms. Some key statistics of the data set are summarized in table II. We use 70% of the data to train the two DOA confidence models and the remaining 30% to evaluate these two models as well as the DOA algorithms.

Both DOA confidence models obtain an accuracy above 80% on the test set as shown in table III. A comparison of the different DOA algorithms are further shown in table IV. The 'TDOA3' algorithm is a benchmark estimator using 3 independent but randomly selected TDOAs to obtain the DOA estimate, i.e., this estimator does not exploit the available TDOA redundancy. The overall accuracy of DETSAC and Flexible ArgMax are similar, and significantly higher than the benchmark, but their performance differ on different parts of the data. The ensemble of these two algorithms shows a 2-3% gain in accuracy over each standalone algorithm. To put an upper bound on the effect of using the ensemble,

TABLE II
SEMI-SYNTHETIC DATA SET USED TO TRAIN AND TEST THE DOA CONFIDENCE MODELS AND TO EVALUATE THE DOA ALGORITHMS

Sample size	10,000
Bandwidth	0.1 - 2.7 kHz
Center frequency	3 - 19 kHz
Duration	0.15 - 1.7 s
SNR (25 percentile)	-4.01 dB
SNR (50 percentile)	-2.27 dB
SNR (75 percentile)	-0.47 dB

TABLE III
PERFORMANCE OF THE DOA CONFIDENCE MODELS ON SEMI-SYNTHETIC DATA

Model trained on	**Accuracy**	**FPR**	**FNR**
DETSAC	85.1%	9.1%	5.9%
Flexible ArgMax	81.3%	11.2%	7.5%

TABLE IV
COMPARISON OF DOA ALGORITHMS ON SEMI-SYNTHETIC DATA

Algorithm	**Accuracy**
TDOA3	46.1%
DETSAC	58.4%
Flexible ArgMax	59.3%
Ensemble	61.9%
Oracle - best of two	65.2%

we also compute the performance if the best estimate amongst the two methods were selected, denoted 'Oracle - best of two'. These accuracy levels should be considered an upper bound on the performance of similar SNRs and bandwidths on real data, as the semi-synthetic data set only incorporates real ambient noise, but not other effects present in a real environment such as multipath or measurement noise.

The ensemble DOA algorithm is next applied to the real data set. Recall that DETSAC obtains a pool of DOA estimates where some estimates are subject to errors due to direction ambiguity in 3D-space when using TDOA sets that only connect 3 sensors. We can resolve this ambiguity to some degree and reduce the number of incorrect estimates in the pool, by exploiting the prior information that the PAM array is resting on the seabed. Consequently, DOAs cannot have negative elevation angles with respect to a local reference frame attached to the PAM (where $\epsilon = 0^\circ$ lies in the plane spanned by the three bottom sensors). Therefore, we reject all single estimates with $\epsilon < S$ where the slack $S = -5^\circ$ is added to account for uncertainties due to bathymetric variation, measurement inaccuracies and inaccuracies in the DOA estimate. Among the signals detected by the bounding box annotator, the ensemble DOA estimator obtains 190 (40.9%) estimates with a confidence score above 0.5, i.e., the confidence model predicts these as correct estimates with an error below 2°. Accounting for the presence of harmonics and treating them together with the fundamental tonals, we obtain 168 calls (43.6%) with correct estimates. These numbers are substantially lower than the 61.9% accuracy obtained on the semi-synthetic data set, suggesting that the additional challenges posed by real-environment effects are significant. Part of the difference, however, may be explained by the difference in SNR and bandwidth between the two data sets. The detections in the real data set has on average 0.44 dB lower SNR and only 0.1 kHz higher bandwidth. A second factor is how well the DOA confidence models generalize to the real data.

C. Multi-individual detector

DOA estimates with confidence levels above 0.5 are lastly fed to the multi-individual detector. The algorithm searches for temporally close detections and finds 3 examples of MIPM above 0.6 of which one is shown in Fig. 6. Several factors may explain the low number of high MIPM detections. For one, the detector searches for temporally dense but angularly spread detections. These do not necessarily occur even when multiple dolphins are present, as dolphins often operate in pods giving rise to angularly concentrated detections. Also, we impose a requirement to use DOA estimates with high certainty, further reducing the potential MIPM candidates to a small subset of the initial set of whistle detections. Lastly, and perhaps the most important reason for so few certain detections of multiple vocalizing individuals, is the lack of a proper range estimate which introduces large uncertainty margins in the detector,

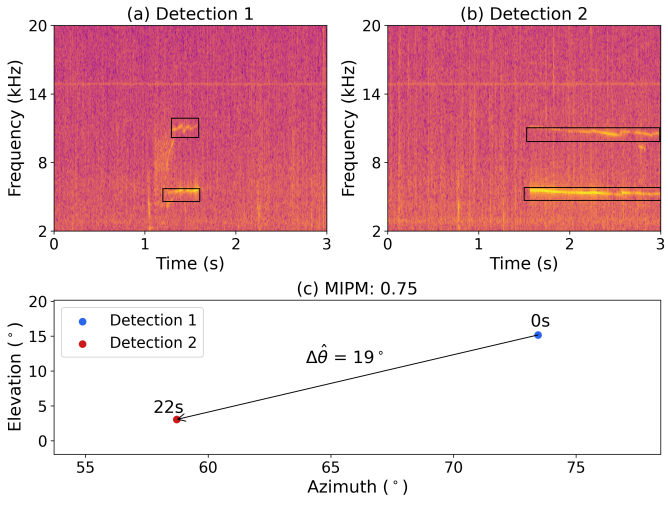

Fig. 6. (a) and (b) Spectrograms of two detected whistles in the recorded data, and (c) their DOA estimates on an azimuth-elevation map showing that they occurred 22 seconds apart with an angular difference of 19°, indicating a high MIPM (0.75) that they originated from two different individuals.

significantly limiting the potential search space for MIPMs.

IV. Experiment setup in Singapore waters

The algorithm and PAM system are being integrated into an advanced scientific buoy[4] located at Singapore's southern waters as part of the test-bedding initiatives under the Marine Environmental Science Sensor Network (MESN) program. The objective is to study the feasibility of leveraging the MESN buoy's onboard edge computer and its environmental sensors to detect and correlate the activities of marine mammals to marine science observations.

The MESN buoy is a unique scientific observatory platform that allows onsite expansion and removal of physical sensors and software functionalities through well-defined mechanical, electrical and software interfaces. This makes it suitable for test-bedding new ideas with little operational cost. In this setup, the buoy would collect more than thirty environmental parameters, such as meteorological, hydrodynamics, waves, nutrients, biological productivity, dissolved gases etc. The PAM array will be installed onto the buoy as part of a special section called the 'peripheral module' that would take payloads that can not be fitted into a conventional moonpool. Meanwhile, the multi-individual detector would be installed onboard the edge computation node as part of software agents in the framework. Lastly, the detections would be labelled and sent to the cloud along with the estimated source direction and metadata. Fig. 7 shows the picture of the actual buoy at the site and an overlay of the render of the underwater sensors.

The buoy is located at the border of a marine park, where a stretch of corals is situated to the East and a more barren, deeper shipping channel to the West. This may provide a further opportunity to gather data on the effect of anthropogenic activities on the megafauna apart from natural environmental effects.

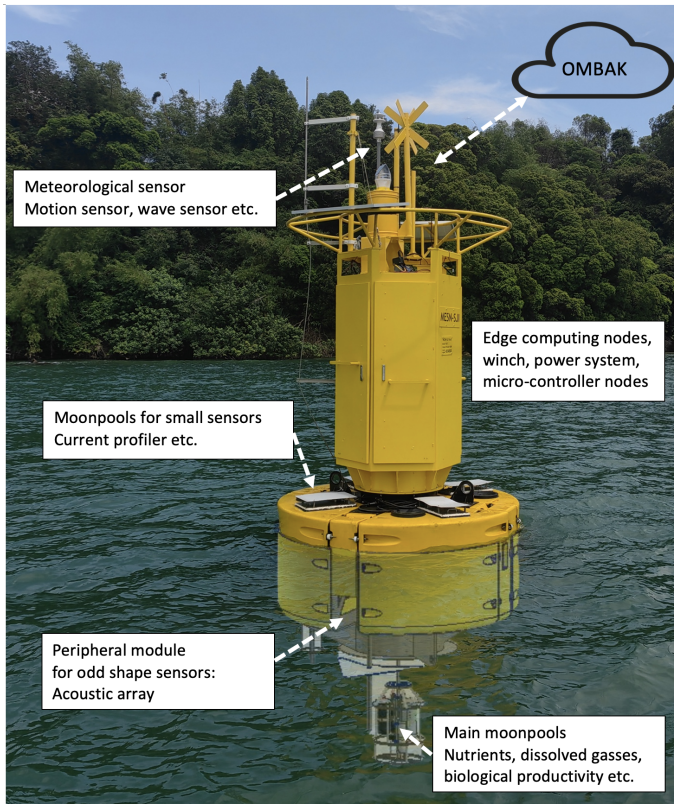

Fig. 7. Experiment setup on the MESN marine scientific buoy. Note the buoy provides multiple ways to easily add sensors while keeping it protected within the buoy's body.

V. Conclusions

In this paper we have presented a non-invasive method to automatically detect presence of multiple marine mammal individuals in noisy environments. The method consist of a chain of detectors and estimators. The bounding box annotator is trained on synthetic signals with real noise and shows good performance on real data. It accurately bounds most whistles and classifies them according to signal shape. It further enables effective signal-denoising in the time-frequency domain. The majority of undetected whistles have either low SNR or shapes not seen during training. The detection rate may be increased by adjusting the training set, e.g. by augmenting it with real whistles or by further increase the variation of the synthetic signals.

The multi-individual detector relies on an accurate DOA estimate. The DOA estimator is shown to exploit the available TDOA redundancy in the PAM to reduce errors. It has been validated in simulations on synthetic data with real noise, but further validation at sea, ideally using whistle-like signals with ground truth, would help to better understand its performance and limitations.

The multi-individual detector detects only a few high MIPM examples. Several factors may explain this, as discussed in section III. Perhaps the main limitation with the current detector is the lack of a robust range estimate. The multi-individual detector could be applied in a network of 2 or more spatially distributed PAMs to obtain more accurate range estimates using triangulation of DOAs and consequently make it possible to detect multiple individuals in scenarios which is not possible with the current detector. A simpler (but likely less accurate) range estimator using a single PAM could be implemented by making use of reported whistle source levels for Indo-Pacific Humpback Dolphins [21] and backtracking the received SNRs using a transmission loss model. These proposed improvements could be interesting to explore in future work.

Acknowledgment

The acoustic detector development was supported by the National Parks Board, Singapore. The test bedding is supported by the MESN program funded by the National Research Foundation (NRF), Singapore. We also thank the NRF for

[4]https://ombak.mesn.sg/

access to facilities at the St. John's Island National Marine Laboratory.

REFERENCES

[1] T. Jefferson, B. Smith, G. Braulik, and W. Perrin. (2017) Sousa chinensis. [Online]. Available: https://www.iucnredlist.org/species/82031425/123794774#geographic-range

[2] The iucn red list of threatened species. [Online]. Available: https://www.iucnredlist.org/search?query=dolphin&searchType=species

[3] F. Caruso, L. Dong, M. Lin, M. Liu, Z. Gong, W. Xu, G. Alonge, and S. Li, "Monitoring of a nearshore small dolphin species using passive acoustic platforms and supervised machine learning techniques," in *Frontiers in Marine Science*, 2020.

[4] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds. Springer International Publishing, 2020, pp. 290–305.

[5] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, p. 39–47, 2004.

[6] M. Roch, T. Brandes, B. Patel, Y. Barkley, S. Baumann-Pickering, and M. Soldevilla, "Automated extraction of odontocete whistle contours," *The Journal of the Acoustical Society of America*, vol. 130, pp. 2212–23, 2011.

[7] C. Jin, M. Kim, S. Jang, and D.-G. Paeng, "Semantic segmentation-based whistle extraction of indo-pacific bottlenose dolphin residing at the coast of jeju island," *Ecological Indicators*, vol. 137, 2022.

[8] H. Vishnu, V. R. Soorya, M. Chitre, T. B. Koay, A. Ho, Y. M. Too, K. Tun, and K. Lim, "Acoustic detection of marine mammal vocalizations in snapping-shrimp infested noisy waters," *9th International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals using Passive Acoustics, Oahu, Hawaii*, 2022.

[9] A. Ho, T. B. Koay, W. Yusong, H. Vishnu, and K. Tun, "A small cloud enabled passive acoustic monitoring array for real-time detection of vocalising marine megafauna," *9th International Workshop on Detection, Classification, Localization, and Density Estimation of Marine Mammals using Passive Acoustics, Oahu, Hawaii*, 2022.

[10] G. Jocher *et al.* Yolo v5. [Online]. Available: https://github.com/ultralytics/yolov5/

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, pp. 740–755.

[12] J. Seekings, K. P. Yeo, Z. P. Chen, S. C. Nanayakkara, J. Tan, P. Tay, and E. Taylor, "Classification of a large collection of whistles from Indo-Pacific humpback dolphins (Sousa chinensis)," *OCEANS'10 IEEE Sydney, OCEANSSYD 2010*, pp. 3–7, 2010.

[13] J. M. Hoffman, L. S. Ponnampalam, C. C. Araújo, J. Y. Wang, S. H. Kuit, and S. K. Hung, " Comparison of Indo-Pacific humpback dolphin (Sousa chinensis) whistles from two areas of western Peninsular Malaysia ," *The Journal of the Acoustical Society of America*, vol. 138, no. 5, pp. 2829–2835, 2015. [Online]. Available: http://dx.doi.org/10.1121/1.4934254

[14] J. Scheuing and B. Yang, "Disambiguation of tdoa estimates in multi-path multi-source environments (datemm)," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006.

[15] A. Canclini, P. Bestagini, F. Antonacci, M. Compagnoni, A. Sarti, and S. Tubaro, "A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1563–1575, 2015.

[16] J. S. Picard and A. J. Weiss, "Time difference localization in the presence of outliers," *Signal Processing*, vol. 92, 2012.

[17] E.-E. Jan and J. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 1996, pp. 1321–1324 vol.3.

[18] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, p. 381–395, 1981.

[19] J. S. Martin Ester, Hans-Peter Kriegel and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, p. 226–231 vol.96.

[20] G. J. Parra and G. J. Ross, "Humpback Dolphins," in *Encyclopedia of Marine Mammals*. Elsevier, 2009, pp. 576–582. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780123735539001346

[21] Z.-T. Wang, W. Au, L. Rendell, K.-X. Wang, H.-P. Wu, Y.-P. Wu, J.-C. Liu, G.-Q. Duan, H.-J. Cao, and D. Wang, "Apparent source levels and active communication space of whistles of free-ranging indo-pacific humpback dolphins (sousa chinensis) in the pearl river estuary and beibu gulf, china," *PeerJ*, vol. 4, pp. 1–38, 2016.