Discovering Biological Hotspots with a Passively Listening AUV

Seth McCammon[1], Stewart Jamieson[2], T. Aran Mooney[1] and Yogesh Girdhar[1]

***Abstract*— We present a novel system which blends multiple distinct sensing modalities in *audio-visual surveys* to assist marine biologists in collecting datasets for understanding the ecological relationship of fish and other organisms with their habitats on and around coral reefs. Our system, designed for the CUREE AUV, uses four hydrophones to determine the bearing to biological sound sources through beamforming. These observations are merged in a Bayesian Occupancy Grid to produce a 2D map of the acoustic activity of a coral reef. Simultaneously, the AUV uses unsupervised topic modeling to identify different benthic habitats. Combining these maps allows us to determine the level of acoustic activity within each habitat. We demonstrated the system in field trials on reefs in the U.S. Virgin Islands, where it was able to autonomously discover the favored habitats of snapping shrimp (genus *Alpheus*).**

I. INTRODUCTION

Coral reefs are one of the most biologically active environments on earth, with over 25% of all known marine animals spending some portion of their life cycle on a reef [1]. Existing methods for surveying biological activity on coral reefs rely heavily on vision, whether using human divers counting organisms [2], [3], or automated with autonomous underwater vehicle (AUV) surveys [4]. However, in the underwater domain, relying purely on vision provides only a partial picture of the environment, since effects like optical backscatter and light attenuation diminish the effectiveness of vision-based approaches developed for terrestrial applications [5], [6]. Hotspots of biological activity are usually far easier to detect acoustically than visually because many of the animals contributing to the reef soundscape, such as snapping shrimp (genus *Alpheus*) use the structure of the reef to hide and are not visible in visual surveys [7].

While active acoustics are a widely-used tool among AUVs and ROVs for navigation, mapping, and communications tasks, passive acoustics have seen comparatively limited use. Yet, they can also provide valuable information in an underwater environment. Recent studies have shown that a reef's soundscape is an important indicator of its overall health, and has links to fish abundance [3], and larvae settlement [8], [9]. Existing methods for measuring sound on reefs use fixed recorders which can be deployed for months at a time [10] or uncontrolled drifters that can drift across a reef [11]. Single-channel recorders collect only point data, which is often assumed to generalize across the entire reef. Multi-channel recorders gain a small amount of spatial context, however such hydrophone-based observations are bearing-only, meaning that these recorders typically cannot precisely localize sound sources in two dimensions on a reef. AUVs such as the Curious Underwater Robot for Ecosystem Exploration (CUREE) [12], shown in Fig. 1, allow acoustic sources to be precisely localized, since they can act as a mobile sensor platform, moving both hydrophones around a reef, allowing the AUV to build a map of sound sources.

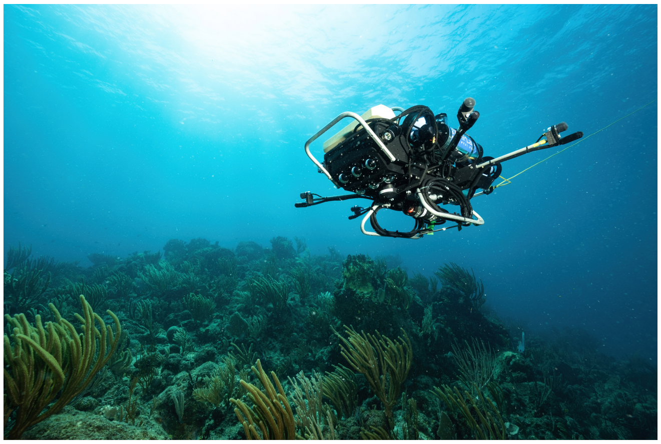

Fig. 1: The Curious Underwater Robot for Ecosystem Exploration (CUREE) AUV equipped with four hydrophones surveying a reef. The hydrophones are mounted on arms extending from the vehicle's body to increase the aperture width of the hydrophone array and consequently the angular resolution with which the AUV can localize sound sources.

The primary contribution of this paper is a new multi-sensor method for locating hotspots of biological activity on a coral reef using an AUV. Our approach, the *audio-visual survey*, uses passive acoustics to map the distribution of biological activity on a reef and vision to identify the distinct types of benthic habitats that make up the reef. As a sensing modality, passive acoustic observation is an excellent compliment to vision underwater. Cameras provide information-dense observations, but their effective range underwater is limited. In contrast, passive acoustic observations offer more limited information (e.g. they only provide the bearing to a target, not a full localization), however they can collect data across much longer ranges. While acoustic hotspot maps and visual habitat maps are individually useful datasets for marine biologists, the true utility of an audio-visual survey lies in combining the two sensing modalities. This allows the AUV to begin to act as a partner to human scientists in ecological studies, autonomously investigating questions like "Which habitats on a coral reef are the most biologically active?"

This work was supported in part by NSF Grant 2133029 and by NOAA Ocean Exploration Grant NA19OAR4320072.R

[1] S. McCammon, T. A. Mooney, and Y. Girdhar are with the Woods Hole Oceanographic Institution (WHOI), Woods Hole, MA. {smccammon, amooney, ygirdhar}@whoi.edu

[2] S. Jamieson is with the MIT-WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering, and the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology (MIT).

The remainder of this paper is organized as follows. Section II provides an overview of related work in autonomous reef monitoring and audio-visual sensing. In Section III we provide a detailed description of the two primary components of an audio-visual survey: the passive acoustic mapping, and the unsupervised benthic habitat topic model. In Section IV we demonstrate the audio-visual survey on reefs around St. John in the U.S. Virgin Islands, where we use the CUREE AUV to autonomously identify habitats where snapping shrimp tend to gather. Finally, in Section V we offer some final conclusions and directions for future work.

II. Background and Related Work

Many methods have been proposed that use autonomous vehicles to measure biodiversity and bioactivity underwater, and specifically to estimate the number and type of species present in a marine environment. Such methods include environmental DNA (eDNA) sampling for population estimation [13], active acoustics for measuring biomass in the water column [14], and visual detection, counting, and tracking of individual organisms [6], [15], [16]. However, eDNA requires laboratory analysis, often taking significant time and presenting a barrier to its use in autonomous applications. Active acoustics sensors are typically expensive and have considerable power requirements, adding an additional burden on the already limited battery of a small, reef-capable AUV, such as the platform used in this work. Finally, vision-based methods are less effective underwater compared to terrestrial deployments, have a significant error rate, and suffer from a double-counting problem where fish swim in and out of frame, making them a less-reliable source of bioactivity estimation [6]. Where visual methods have seen the most success underwater is mapping the static components of the underwater environment. Spatiotemporal topic modeling has been used to identify semantically distinct habitat types [17]. The chief advantage that topic models have over other vision-based approaches is that they do not rely on any pre-training or large datasets.

An increasingly popular sensing modality for studying underwater ecosystems is passive acoustics. Like vision, passive acoustics is inexpensive and non-invasive, making it easier to observe an underwater without disturbing it with physical sampling (eDNA) or significant acoustic energy (active acoustics). Biologists have used passive acoustic recorders to measure changes in biological activity on coral reefs over time [3], [8], [10], but focus has often been temporal patterns; these methods have been traditionally limited in their ability to measure spatial variation. While passive acoustics can localize a sound source [18], in the typical far-field source model, only information about the bearing to the source is recovered (Fig. 2). A limited number of methods exist that can reconstruct the range information in the far-field [19]. However, these approaches rely on assumptions about the propagation of sound over long distances, and are not well-suited to localization on coral reefs, where the ranges are short and the bathymetry is complex and unknown. Triangulation allows for a set of bearing-only observations to reconstruct range information by moving the bearing-only sensor, and methods to accomplish this have been developed for domains like radiation source detection [20], indoor SLAM using WiFi signals [21], and locating targets underwater [22]–[24]. However, these methods typically assume a small number of targets, or uniquely identifiable targets, which enables detection-based methods. In contrast, a reef sounsdscape is made up of a multitude of sources, each creating a signal with relatively low information content, making detection-based source localization infeasible.

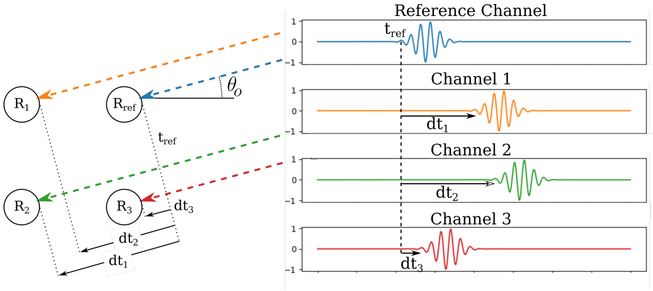

Fig. 2: Sound source localization methods use the time delays $(dt_1, ..., dt_N)$ and the known geometry of the N receivers to compute the angle of arrival of a sound source, θ_o. These delays are computed relative to a chosen reference channel by applying a steering vector $\mathbf{a}(\theta)$ of phase delays to the multi-channel signal and compute the resultant energy.

In terrestrial robotics applications, sound and vision have been combined to give robots a greater awareness of their surroundings. In particular, these methods have been used in human-robot applications to track the movement of people and objects [25], [26], to identify unique speakers in the canonical "cocktail party problem" [27]–[29], or to improve in navigation and mapping [30], [31]. However, to our knowledge, applications of audio-visual methods in the marine domain are limited to static sensors [32], and have not been explored for mobile sensing.

III. Methods

As its name suggests, an audio-visual survey builds two parallel maps of the coral reef. One map uses using a set of passive acoustic observations to map the distribution of biological acoustic sources on the reef, and the other uses visual observations to characterize the different habitats that compose the reef. Independently, each map is valuable to marine biologists, however when combined, they allow us to discover which habitats are hotspots of acoustic, and therefore biological, activity. In this work, we make several simplifying assumptions about the distribution of sound sources on the reef. The first of these assumptions is that the number of sound sources is large and spatially varying. The second is that acoustic sources are distributed in only two dimensions across the reef. While some fish, particularly larger predators, inhabit the water column above the reef, the majority of organisms are close to the bottom.

A. Passive Acoustic Mapping

To build an acoustic map of the coral reef, the AUV must collect a set of acoustic observations $\Omega^a = \{o_1^a, o_2^a, ...o_{n_a}^a\}$. Unlike visual observations, which are collected continuously as the AUV moves, audio observations can only be made when the robot is stationary due to the noise created by the robot's thrusters. To collect a single acoustic observation, o_i^a, CUREE stops its thrusters and drifts for 10 to 15 seconds to allow it to observe the reef soundscape.

The origin vector of a planar sound wave is defined as the vector from the robot position x_r to the sound source x_o where $\mathbf{r}_o = x_o - x_r$. It is typically defined in spherical coordinates with origin azimuth θ_o and origin elevation ϕ_o,

$$\mathbf{r}_o = \hat{x}\cos(\phi_o)\sin(\theta_o) + \hat{y}\sin(\phi_o)\sin(\theta_o) + \hat{z}\cos(\theta_o), \quad (1)$$

where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are the unit vectors along the x, y, and z axes. However, x_o is unknown, and so to estimate the angle of arrival we use the Bartlett Beamformer [18], which computes the origin vector using

$$\theta_o, \phi_o = \underset{\theta\in\Theta,\phi\in\Phi}{\operatorname{argmax}}\, \mathbf{a}(\theta,\phi)^H \times \mathbf{R} \times \mathbf{a}(\theta,\phi), \quad (2)$$

where $\Theta = [0, 2\pi)$ and $\Phi = [-\frac{\pi}{2}, \frac{\pi}{2}]$. Here $\mathbf{a}(\theta,\phi)$ is the steering vector of phase delays associated with a particular azimuth and elevation, ^{H} is the Hermetian operator, and $\mathbf{R}$ is the spatial covariance matrix which captures the pairwise correlations between the signals arriving at the array elements. This matrix is defined as

$$\mathbf{R} = \begin{bmatrix} \langle x_1, x_1^* \rangle & \langle x_1, x_2^* \rangle & \cdots & \langle x_1, x_N^* \rangle \\ \langle x_2, x_1^* \rangle & \langle x_2, x_2^* \rangle & \cdots & \langle x_2, x_N^* \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_N, x_1^* \rangle & \langle x_N, x_2^* \rangle & \ldots & \langle x_N, x_N^* \rangle \end{bmatrix}, \quad (3)$$

where the $\langle\cdot\rangle$ operator computes the time averaged power of a pair of signals over a window of length T using

$$\langle x(t_0), x(t_0)^* \rangle = \frac{1}{T}\int_{t_0}^{t_0+T} x(t)x(t)^* dt. \quad (4)$$

The Bartlett beamformer is a narrowband beamformer, since the phase delays in $\mathbf{a}(\theta,\phi)$ are dependant on the carrier frequency. In our system, the exact frequency of the sound source is unknown since individual organisms can produce sounds over a range of frequencies. To account for this, we use multiple Bartlett beamformers evenly spaced over the expected frequency range, with their power output averaged. For beamforming the sound produced by snapping shrimp, we use 10 beamformers over the frequency range of 5 kHz to 20 kHz. In this work we use the Bartlett beamformer provided in the ARL Tools Python package [33].

The output of the beamformer is the maximum-power azimuth and elevation angles. For a single audio source, solving the optimization problem presented in Eq. 2 would be sufficient to determine the origin vector. However, on a reef we are presented with an unknown but assumed large number of individual sound sources. Thus, instead of computing a single origin vector, we are interested in computing the distribution of sources, $p(s) \in \mathcal{P}(\Theta \times \Phi)$. To compute this distribution, we subdivide each observation window into a set of snapshots 0.1 seconds in length, and solve Eq. 2 for each. An example of this distribution is shown in Fig. 3. While the beamformer does produce a two-dimensional output (Θ, Φ), we assume that the sources we are mapping are on the seafloor, and therefore we can discard the elevation angle and simplify $p(s)$ to a single-dimensional distribution over Θ. Since there can be multiple hotspots on a reef, $p(s)$ can be multimodal with an unknown number of components. To represent $p(s)$ we use a Bayesian Gaussian Mixture Model (GMM), which can represent an arbitrary probability distribution using a sum of M weighted Gaussian components $o_a = \sum_{i=1}^{M} \psi_i N(\mu_i, \sigma_i)$. The advantage of using a Bayesian GMM over the standard GMM formulation is that the number of components is not fixed but rather varies according to a Dirichlet distribution [34]. For brevity, we omit including the full Bayesian GMM formulation here, but it can be found in [34], and the implementation used in this work is part of the scikit-learn Python toolkit [35].

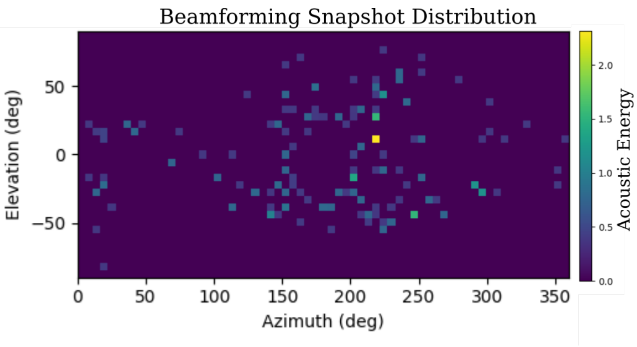

Fig. 3: Distribution of acoustic energy across 169 0.1s windows taken during a single acoustic observation.

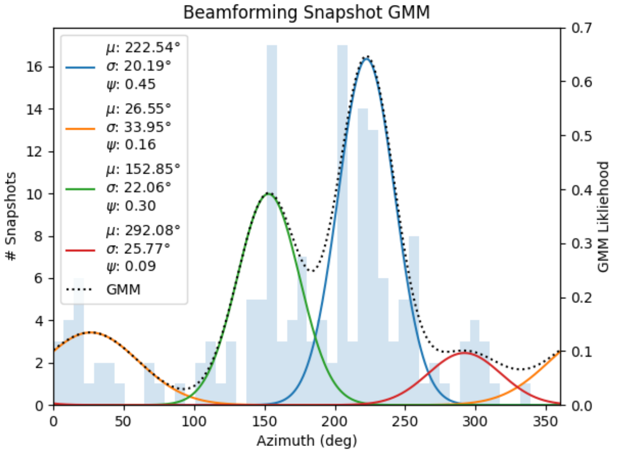

Fig. 4: Histogram of azimuths in the snapshots shown in Fig. 3 (left axes), and corresponding Bayesian GMM with 4 components (right axes).

The other challenge with representing $p(s)$ with a GMM is that $p(s)$ is a circular distribution. While there has been recent work on fitting a GMM in a circular distribution [36], we chose a simpler approach by embedding $p(s)$ on a manifold in $\mathbb{R}^2$ by changing the domain from Θ to $\sin(\Theta)$ and $\cos(\Theta)$. One the GMM is converged, the component means, $[\mu_{cos(\Theta)}, \mu_{sin(\Theta)}]$, and covariances, Σ_i, can be converted back to the original domain by

$$\mu_\Theta = \arctan2(\mu_{cos(\Theta)}, \mu_{sin(\Theta)}), \quad (5)$$

$$\sigma_{\Theta} = \sqrt{\frac{\ln((1 - 2\Sigma_{cos,cos})(1 - 2\Sigma_{sin,sin}))}{-2}}. \tag{6}$$

The weights of the GMM components are unaffected by the return to the single-dimension basis. An example of our angle-wrapped Bayesian GMM is shown in Fig. 4.

The final step to produce an audio map of a reef is to reverse the forward sensor model described in Eq. 1, and combine the one-dimensional (bearing only) audio observations described by the GMM components into a two-dimensional map of source locations. To accomplish this, we use a Bayesian Occupancy Grid (BOG). BOGs have seen significant prior use in mapping underwater sound sources [23], [24]. The BOG discretizes the world into a grid of uniform cells, and with each observation, performs a bayesian update of the likelihood that a source exists in that cell. In this work, we chose this discretization level to be 0.1 m. We adopt the formulation for a BOG given in [23] modified to account for multimodal observation distributions. For a single component of a visual observation $(\mu_i, \sigma_i, \psi_i) \in o^a$, the source likelihood of a location x_j is

$$p(x_i = 1|\mu_i, \sigma_i, \psi_i) = \frac{\psi_i||x_j - x_r||_2}{l\sqrt{2\pi\sigma_i^2}} \exp\left(\frac{\Delta(\mu_i, x_j, x_r)^2}{\sigma_i^2}\right), \tag{7}$$

for a robot located at x_r where $\Delta(\mu, x_j, x_r)$ is the minimum angle between the vector defined by $x_r - x_j$ and μ. The parameter l accounts for the lower signal-to-noise ratio with increased distance to a source in the presence of background noise. Based on empirical data we set l to be 20 m.

B. Habitat Identification

As CUREE moves through the world, it continuously builds a set of visual observations of the seafloor, $\Omega^v = \{o_1^v, o_2^v, ...o_{n_v}^v\}$, using its downward-facing cameras. Each $o_i^v \in \Omega^v$ is a bag-of-words of visual feature "words" $w \in V$, where V is a "vocabulary" comprised of quantized ORB features as well as quantized hue and lightness features from the HSL color model. To minimize the impact of light attenuation and backscatter on the distribution of these features, we color-correct the images before word extraction using DeepSeeColor [5]. Each word is associated with the spatiotemporal coordinates of the time and location at which it was observed. In order to fuse the observations Ω^v into a single spatial model mapped over our region of interest, we employ Realtime Online Spatial Topic Modeling (ROST) [17]. While in this work V is comprised exclusively by visual features, in future work it could be expanded to include measures of 3D terrain complexity, such as rugosity, or oceanographic variables such as temperature or salinity

At a high level, ROST performs unsupervised clustering of the $o^v \in \Omega^v$ based on both the distribution of words within o^v, as well as the spatiotemporal proximity of o_v to other observations in Ω^v. It divides the region of interest into a grid of cells such that each cell contains all of the words observed in a distinct spatiotemporal volume. ROST assumes that every cell has a latent distribution $Z \in \Pi^K$ over a space of K "topics",[1] and that a cell's latent distribution is correlated to its neighbors'. ROST uses Monte Carlo sampling techniques to simultaneously learn these topics and assign each feature observation to a particular topic in realtime. Without any pretraining, this approach learns topics which reflect visually-distinct ecological habitats in the environment, such that $Z(\mathbf{x})$ represents the mixture of habitats observed at location $\mathbf{x}$.

C. Data Synthesis

To learn the mapping between the learned visual topics and acoustic activity, we used Bayesian Linear Regression [35], [37], since it balances the simplicity of Ordinary Linear Regression while also providing estimates of model and predictive uncertainty. The model assumes there is weight vector $\psi \in \mathbb{R}^K$ such that

$$A(\mathbf{x}) = \psi^\top Z(\mathbf{x}) + \epsilon, \tag{8}$$

where $Z(\mathbf{x})$ is the topic map, $A(\mathbf{x})$ is the audio map, and $\epsilon \sim \mathcal{N}(0, \sigma)$. Bayesian Linear Regression finds the most likely weight vector $\hat{\psi} = \max_\psi \Pr(\psi \mid Z(\mathbf{x}), A(\mathbf{x}))$, which describes the mean acoustic activity level of each topic.

One key advantage of Bayesian Linear Regression over more expressive models, such as Gaussian Process Regression, is that Bayesian Linear Regression is significantly less computationally expensive, scaling at $\mathcal{O}(|\mathbf{x}|)$ as opposed to the $\mathcal{O}(|\mathbf{x}|^3)$ [38]. Experimentally, we found that the expressive power we lost by using the simpler Bayesian Linear Regression was a minor issue more than compensated by the reduced computational requirements.

IV. Field Experiments

In two expeditions to St. John in the U.S. Virgin Islands, we conducted multiple audio-visual surveys on three different reefs. The first of these reefs, Joel's Shoal, is a highly compact and isolated reef approximately 30 m in diameter, surrounded by mostly sand. The second reef, Booby Rock, is a larger reef which wraps around its namesake rock. It features a mix of sandy channels, rocky and soft coral, seagrass beds, and on the eastern edge is a 5 m cliff, beyond which the seafloor drops off to a relatively flat sandy bottom. The final reef, Tektite, has a similar structure to BR, with a patchy mix of sand, seagrass, and hard and soft corals. Work was conducted under the National Park Service Scientific Research and Collecting Permit #VIIS-2022-SCI-0005.

The CUREE platform [12], shown in Fig. 1, was used to conduct all surveys. It was equipped with four calibrated HTI Min-96 hydrophones mounted on the end of aluminum arms 30 cm long and arranged in a planar array. The maximum separation between hydrophones along the vehicle's sway axis was 89 cm, and 36 cm along the surge axis.

A. Synthetic Sound Source Localization

To evaluate the accuracy of our acoustic mapping approach, we conducted a series of tests using an underwater speaker deployed on a sandy seafloor as a synthetic acoustic hotspot. During these tests, the speaker either played acoustic chirps (a pure tone linearly varying in frequency from 2 kHz

[1] Note $\Pi^K = \{\mathbf{p} \in \mathbb{R}^K_{>0} : ||\mathbf{p}||_1 = 1\}$ denotes the probability simplex.

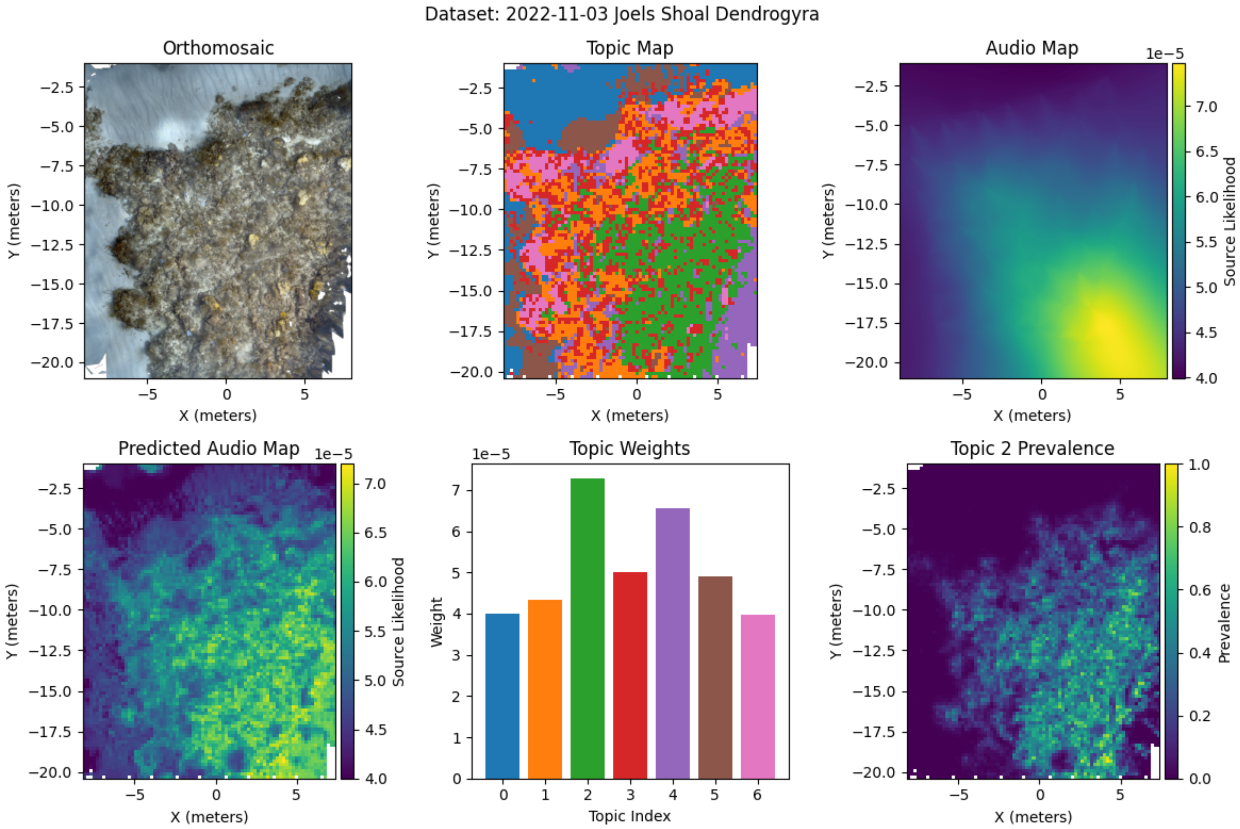

Fig. 5: Dataset collected in a grid survey on the edge of Joel's Shoal near St. John, USVI. The top-left image shows a color-corrected reconstructed orthomosaic of the Joel's Shoal site around a dead *Dendrogyra cylindrus* (Pillar Coral). The pillar coral is located at approximately (-4m, -7.5m). Seven semantically distinct topics were identified by the unsupervised topic model discussed in Section III-B (top-center). The corresponding audio map is constructed from 56 10 second drifting periods (top-right). Using Bayesian Linear Regression, we can choose a set of weights (bottom-center) that best predict the audio map from the topic map (bottom-left). The distribution of the highest-weighted topic covers rocky coral regions (bottom-right).

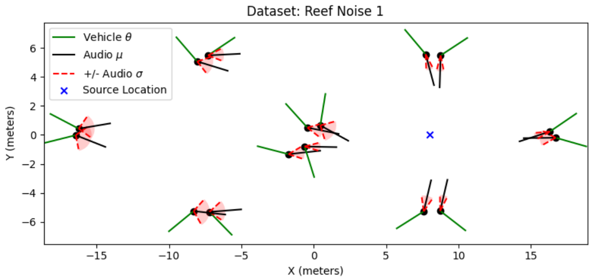

Fig. 6: Audio sample positions and o_a from a synthetic sound source figure-eight experiment using the Bartlett beamformer. Audio samples taken closer to the sound source have lower variance in estimated direction compared to samples taken farther away.

to 20 kHz), or a recording of a healthy reef soundscape recorded at Tektite. An example mission is shown in Fig. 6 In each synthetic localization mission, CUREE tracked a figure-eight pattern with loops 8 m in diameter, and the acoustic source positioned at the center of one loop. CUREE went around each figure-eight once clockwise and once counter-clockwise per mission. We evaluated two different acoustic recording systems, a Teensy4.1 Audio board, sampling at 44.1 kHz, and a OceanInstruments Soundtrap sampling at 288 kHz. Ultimately, we found no meaningful advantage to the higher sampling rate offered by the OceanInstruments Soundtrap, and the additional samples require additional computation time to process. We also evaluated the performance of four common beamformers: the Bartlett beamformer described in Sec. III-A, as well as the Capon, MUSIC, and Delay-and-Sum beamformers. The results of this anaylysis are shown in Table I. In each experiment we evaluated both the average angular uncertainty (i.e., the variance of snapshot angle within each observation window), and the localization error of the estimated sound source location. We found that the Bartlett beamformer provided the best balance of localization accuracy and angular uncertainty.

TABLE I: Beamformer Evaluation with Synthetic Sound Source

Beamformer	Domain	Average Angular Uncertainty	Final Localization Error
Delay-and-Sum	Time	**38.296°**	2.151 m
Bartlett	Frequency	44.668°	1.699 m
Capon	Frequency	54.305°	**1.211** m
MUSIC	Frequency	58.064°	2.221 m

B. Acoustic Hotspots on Coral Reefs

Across the two field expeditions and three reefs, we conducted 17 audio-visual surveys. Seven surveys were conducted at Joel's Shoal, six at Booby Rock, and four at Tektite. Each survey was conducted using a pre-planned

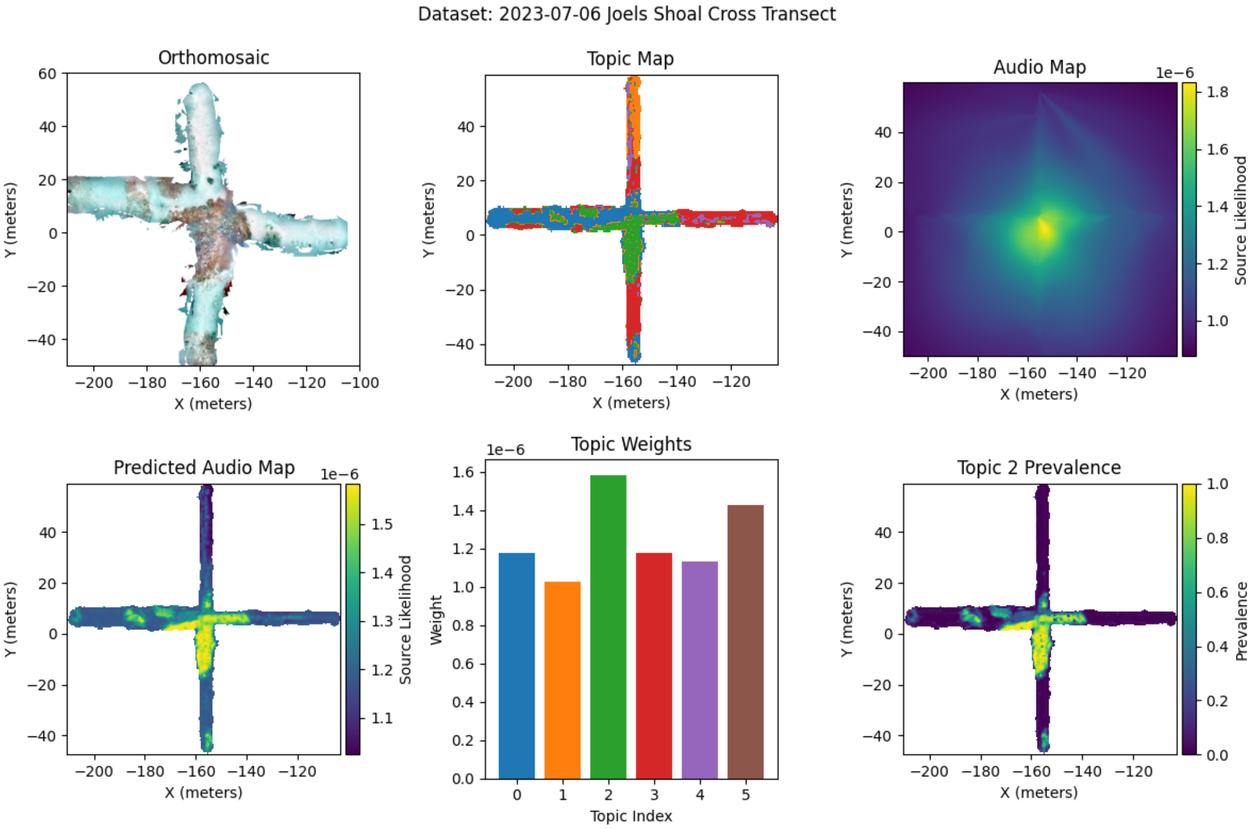

Fig. 7: Dataset collected in a transect survey over Joel's Shoal. The maximum-weighted topic, Topic 2, covers the rocky coral that forms the main mass of Joel's Shoal reef, while Topics 0, 1, and 3 cover the sandy surrounding regions. The difference between the topic map and orthomosaic geometry is due to errors in the AUV localization that are corrected when the visual imagery is stitched into the orthomosaic.

trajectory at approximately 1.5 meters altitude. The surveys can be broadly categorized into two types. Dense grid surveys cover small regions of a reef (20 to 30 m squares) with a lawnmower pattern. Transect surveys cover much larger areas, however they do not achieve complete coverage, and are instead focused on observing a breadth of habitats. Examples of grid and transect surveys at Joel's Shoal are shown in Fig. 5 and Fig. 7, respectively. In both surveys, the primary source of acoustic activity is the rocky coral at the center of the reef. While we lack a ground truth count of snapping shrimp populations on Joel's Shoal, this result matches studies at other sites, where researchers found that snapping shrimp tend to prefer complex bathymetry where they can safely hide from predators beneath rocks and coral [7]. This result was replicated in surveys conducted at the other sites. At Booby Rock, five of the six surveys were transect surveys which followed the same 300 m long path over the reef on the northern side of the rock. In each survey, the primary acoustic activity was located in shallow coral-covered regions, similar to the most acoustically active portions of Joel's Shoal shown in Figs. 5 and 7.

V. Conclusion

In this paper we demonstrated a new system for conducting multi-sensor surveys of coral reefs. Our system uses Bartlett beamforming to measure the distribution of acoustic energy in the soundscape of a coral reef and fused using a Bayesian Occupancy Grid. Simultaneously, we use Realtime Spatiotemporal Topic Modelling to cluster benthic habitat types based on visual features. By combining these maps, we were able to identify which habitat types on a coral reef are biological "hotspots" and responsible for most of the acoustic energy in the reef soundscape. We demonstrated the system in field experiments on reefs off of St. John in the US Virgin Islands, where we autonomously identified the preferred habitat of snapping shrimp.

An implicit assumption of our system is that the survey domain is small enough that the AUV can achieve complete coverage of the environment. In future work, we plan to relax this assumption and use the audio map in an informative path planning architecture, where the AUV optimizes its trajectory to both produce the most accurate acoustic map of the reef and maximize its time spent visually surveying the biological hotspots it discovers. Another avenue for improvement is expanding the range of organisms the AUV targets in the audio-visual survey. In this work we targeted snapping shrimp due to their ubiquity and ease of detection however, with neural network trained to identify fish calls in lower frequency bands, we could apply the same techniques outlined to map the distribution of specific species of fish. Finally, we also plan to expand the variables used for habitat identification to include temperature, salinity, and 3D measures of structure, such as rugosity.

References

[1] N. Knowlton, R. E. Brainard, R. Fisher, M. Moews, L. Plaisance, and M. J. Caley, "Coral reef biodiversity," *Life in the world's oceans: diversity distribution and abundance*, pp. 65–74, 2010.

[2] A. M. Friedlander and J. D. Parrish, "Habitat characteristics affecting fish assemblages on a hawaiian coral reef," *Journal of experimental marine biology and ecology*, vol. 224, no. 1, pp. 1–30, 1998.

[3] M. B. Kaplan, T. A. Mooney, J. Partan, and A. R. Solow, "Coral reef species assemblages are associated with ambient soundscapes," *Marine Ecology Progress Series*, vol. 533, pp. 93–107, 2015.

[4] S. B. Williams, O. Pizarro, M. Jakuba, and N. Barrett, "Auv benthic habitat mapping in south eastern tasmania," in *Field and Service Robotics: Results of the 7th International Conference*. Springer, 2010, pp. 275–284.

[5] S. Jamieson, J. P. How, and Y. Girdhar, "DeepSeeColor: Realtime adaptive color correction for autonomous underwater vehicles via deep learning methods," 2023.

[6] L. Cai, N. E. McGuire, R. Hanlon, T. A. Mooney, and Y. Girdhar, "Semi-supervised visual tracking of marine animals using autonomous underwater vehicles," *arXiv preprint arXiv:2302.07344*, 2023.

[7] J. Butler, M. J. Butler IV, and H. Gaff, "Snap, crackle, and pop: Acoustic-based model estimation of snapping shrimp populations in healthy and degraded hard-bottom habitats," *Ecological Indicators*, vol. 77, pp. 377–385, 2017.

[8] E. Parmentier, L. Berten, P. Rigo, F. Aubrun, S. Nedelec, S. D. Simpson, and D. Lecchini, "The influence of various reef sounds on coral-fish larvae behaviour," *Journal of Fish Biology*, vol. 86, no. 5, pp. 1507–1518, 2015.

[9] A. Lillis, A. Apprill, J. J. Suca, C. Becker, J. K. Llopiz, and T. A. Mooney, "Soundscapes influence the settlement of the common caribbean coral porites astreoides irrespective of light conditions," *Royal Society Open Science*, vol. 5, no. 12, p. 181358, 2018.

[10] A. Lillis and T. A. Mooney, "Snapping shrimp sound production patterns on caribbean coral reefs: relationships with celestial cycles and environmental variables," *Coral Reefs*, vol. 37, no. 2, pp. 597–607, 2018.

[11] A. Lillis, F. Caruso, T. A. Mooney, J. Llopiz, D. Bohnenstiehl, and D. B. Eggleston, "Drifting hydrophones as an ecologically meaningful approach to underwater soundscape measurement in coastal benthic habitats," *Journal of Ecoacoustics*, vol. 2, no. 10.22261, 2018.

[12] Y. Girdhar, N. McGuire, L. Cai, S. Jamieson, S. McCammon, B. Claus, J. E. S. Soucie, J. E. Todd, and T. A. Mooney, "CUREE: A curious underwater robot for ecosystem exploration," in *proc. IEEE International Conference on Robotics and Automation (ICRA) to appear*, 2023.

[13] K. M. Yamahara, C. M. Preston, J. Birch, K. Walz, R. Marin III, S. Jensen, D. Pargett, B. Roman, W. Ussler III, Y. Zhang *et al.*, "In situ autonomous acquisition and preservation of marine environmental dna using an autonomous underwater vehicle," *Frontiers in Marine Science*, vol. 6, p. 373, 2019.

[14] K. Benoit-Bird, T. Patrick Welch, C. Waluk, J. Barth, I. Wangen, P. McGill, C. Okuda, G. Hollinger, M. Sato, and S. McCammon, "Equipping an underwater glider with a new echosounder to explore ocean ecosystems," *Limnology and Oceanography: Methods*, vol. 16, no. 11, pp. 734–749, 2018.

[15] M. Sung, S.-C. Yu, and Y. Girdhar, "Vision based real-time fish detection using convolutional neural network," in *OCEANS 2017-Aberdeen*. IEEE, 2017, pp. 1–6.

[16] S. Zhang, X. Yang, Y. Wang, Z. Zhao, J. Liu, Y. Liu, C. Sun, and C. Zhou, "Automatic fish population counting by machine vision and a hybrid deep neural network model," *Animals*, vol. 10, no. 2, p. 364, 2020.

[17] Y. Girdhar, P. Giguere, and G. Dudek, "Autonomous adaptive exploration using realtime online spatiotemporal topic modeling," *The International Journal of Robotics Research*, vol. 33, no. 4, pp. 645–657, 2014.

[18] D. H. Johnson and D. E. Dudgeon, "Array signal processing: Concepts and methods," 1992.

[19] J. Bonnel, A. M. Thode, S. B. Blackwell, K. Kim, and A. Michael Macrander, "Range estimation of bowhead whale (balaena mysticetus) calls in the arctic using a single hydrophone," *The journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 145–155, 2014.

[20] H. Ardiny, S. Witwicki, and F. Mondada, "Autonomous exploration for radioactive hotspots localization taking account of sensor limitations," *Sensors*, vol. 19, no. 2, p. 292, 2019.

[21] A. Arun, R. Ayyalasomayajula, W. Hunter, and D. Bharadia, "P2slam: Bearing based wifi slam for indoor robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3326–3333, 2022.

[22] J. Gebbie, M. Siderius, P. L. Nielsen, and J. Miller, "Passive localization of noise-producing targets using a compact volumetric array," *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 80–89, 2014.

[23] G. Ferri, A. Tesei, P. Stinco, and K. D. LePage, "A bayesian occupancy grid mapping method for the control of passive sonar robotics surveillance networks," in *OCEANS 2019-Marseille*. IEEE, 2019, pp. 1–9.

[24] G. Ferri, A. Faggiani, R. Petroccia, P. Stinco, and A. Tesei, "A robotic cooperative network for localising a submarine in distress: Results from repmus21," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3088–3094.

[25] J. Fritsch, M. Kleinehagenbrock, S. Lang, G. A. Fink, and G. Sagerer, "Audiovisual person tracking with a mobile robot," in *Proc. Int. Conf. on Intelligent Autonomous Systems*, 2004, pp. 898–906.

[26] Y. Masuyama, Y. Bando, K. Yatabe, Y. Sasaki, M. Onishi, and Y. Oikawa, "Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 4848–4854.

[27] Y.-m. Qian, C. Weng, X.-k. Chang, S. Wang, and D. Yu, "Past review, current progress, and challenges ahead on the cocktail party problem," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 40–63, 2018.

[28] S. Lathuilière, B. Massé, P. Mesejo, and R. Horaud, "Neural network based reinforcement learning for audio–visual gaze control in human–robot interaction," *Pattern Recognition Letters*, vol. 118, pp. 61–71, 2019.

[29] S. Majumder, Z. Al-Halah, and K. Grauman, "Move2hear: Active audio-visual source separation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 275–285.

[30] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman, "Soundspaces: Audio-visual navigation in 3d environments," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 2020, pp. 17–36.

[31] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman, "Audio-visual floorplan reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1183–1192.

[32] X. Mouy, R. A. Rountree, F. Juanes, and S. E. Dosso, "Passive acoustic localization of fish using a compact hydrophone array," *The Journal of the Acoustical Society of America*, vol. 141, no. 5_Supplement, pp. 3863–3863, 2017.

[33] "Beamforming and array processing," 2023. [Online]. Available: https://arlpy.readthedocs.io/en/latest/bf.html

[34] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian approaches to gaussian mixture modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1133–1142, 1998.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] N. Ito, S. Araki, and T. Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1153–1157.

[37] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[38] D. Duvenaud, "Automatic model construction with gaussian processes," Ph.D. dissertation, University of Cambridge, 2014.