Detecting Submerged Objects Using Active Acoustics and Deep Neural Networks: A Test Case for Pelagic Fish

Alberto Testolin, *Member, IEEE*, Dror Kipnis, and Roee Diamant, *Senior Member, IEEE*

Abstract—The accurate detection and quantification of submerged targets has been recognized as a key challenge in marine exploration, one that traditional census approaches cannot handle efficiently. Here we present a deep learning approach to detect the pattern of a moving fish from the reflections of an active acoustic emitter. To allow for real-time detection, we use a convolutional neural network, which provides the simultaneous labeling of a large buffer of signal samples. This allows to capture the structure of the reflecting signal from the moving target and to separate it from clutter reflections. We evaluate system performance both on synthetic (simulated) data, as well as on real data recorded over 50 sea experiments in a variety of sea conditions. When tested on real signals, the network trained on simulated patterns showed non-trivial detection capabilities, suggesting that transfer learning can be a viable approach in these scenarios, where tagged data is often lacking. However, training the network directly on the real reflections with data augmentation techniques allowed to reach a more favorable precision-recall trade-off, approaching an ideal detection bound. We also evaluate an alternative model based on recurrent neural networks which, despite exhibiting slightly inferior performance, could be applied in scenarios requiring on-line processing of the reflection sequence.

Index Terms—Signal processing, signal detection, deep learning, convolutional neural networks, long-short term memory networks, underwater acoustics, detection of pelagic fish, environmental monitoring

1 Introduction

In attempting to understand the ever-changing oceans, biota and atmosphere are two of the greatest global challenges. One of the key aspects in marine monitoring is the assessment of marine life diversity, which must involve efficient and autonomous surveys aimed at creating large tagged databases using both mobile and fixed platforms [1]. One specific challenge is evaluating the quantities and abundance of pelagic fish. Pelagic fish directly affect coastal recreation, tourism, fisheries, aquaculture, and coastal industries. Due to the absence of reliable methods for the classification and biomass evaluation of pelagic fish, current research is often performed manually [2], and is thus highly limited in regard to space and time. The need for better solutions has recently led the European Union (EU) Commission to issue a series of large-scale projects [3] aimed at developing new technologies for automatic marine monitoring. In this paper, we offer our solution for the efficient detection of pelagic fish using active acoustics. Since pelagic fish move through large areas, it is expected that, from a stationary observatory, detection of such fish would be scarce. As such, we aim for a detection system that notifies the presence of a fish. In our project, this serves as a first crucial step towards a full acoustic-optic chain for fish detection and classification. Different than current fish detection sonar systems that can detect low power acoustic reflections by applying directionality from an array of receivers, we aim for omni-directional detection by a single transceiver. Such system can cover larger areas and is simpler to handle, but also requires to handle lower signal-to-clutter ratios (SCRs).

Active acoustic detection involves the transmission of acoustic signals and an analysis of the received reflections. Current acoustic-based census for evaluating marine biota can only provide limited solutions. In particular, acoustic imaging techniques for fish finders produce a very narrow high frequency beam to detect objects located directly below a surveying vessel [4]. As a result, these techniques do not cover a large volume of water. Alternatively, techniques like continuous active sonar (CAS) [5] or tracking approaches [6] allow omni-directional detection, but assume either hard constraints [7] or a statistical model on the target's motion [8], and may thus suffer from model mismatches. Furthermore, these techniques rely on arrays of receivers, consume large computational power, and usually require long processing times. Instead, we are interested in developing a low-cost monitoring system, including a single transceiver capable of providing detection in real-time. Further, observable features directly extracted from the reflected signal (such as target strength and frequency offset) largely depend on the fish species, size, and orientation with respect to the transceiver, while we are interested in a robust detection approach, where non-linear hidden features are flexibly extracted from the data.

- *Alberto Testolin is with the Department of Information Engineering and the Department of General Psychology, University of Padova, 35122 Padova, Italy. E-mail: alberto.testolin@unipd.it.*
- *Dror Kipnis and Roee Diamant are with the Department of Marine Technologies, University of Haifa, Haifa 3498838, Israel. E-mail: dkipnis@campus.haifa.ac.il, roee.d@univ.haifa.ac.il.*

Manuscript received 7 Apr. 2020; revised 2 Nov. 2020; accepted 9 Dec. 2020. Date of publication 14 Dec. 2020; date of current version 1 July 2022.
(Corresponding author: Alberto Testolin.)
Digital Object Identifier no. 10.1109/TMC.2020.3044397

In the context of detection at low SCR levels, machine learning techniques can offer high potential, providing an efficient, data-driven approach for solving complex acoustic detection problems, even in challenging underwater sea environments [9]. In particular, deep learning enables training artificial neural networks composed of many processing layers that learn high-level representations of the data by exploiting multiple levels of abstraction [10]. In other words, these techniques can automatically discover the relevant features needed to solve a certain task directly from the raw input, thus dispensing with the need for human expert knowledge and heavy data pre-processing. Most importantly, the impressive performance of deep learning methods is evident not only in human-like cognitive tasks, such as image and speech recognition [11], [12], but also in challenging domains such as drug discovery [13], genomics [14], network optimization [15] or the detection of rare particles in high-energy physics [16], where difficult signal-versus-background classification problems need to be solved.

While fish classification and biomass estimation is our ultimate goal, acoustic detection of the presence of a fish is the first and crucial step. Here, detection must be preformed in real-time to allow triggering optical cameras as well as higher resolution acoustic sensors. With the aim of providing accurate real-time detection for a low-energy sea platform, our system is based on the analysis of reflections from the emission of a single-carrier short signal by a single mono-static acoustic transceiver. To conserve energy, the signal is transmitted over a large period interval (e.g., 60 s), thus allowing detection based only on a single signal. To allow real-time robust detection, our detection scheme is based on a Convolutional Neural Network (CNN). The network receives as input a buffered sequence of raw acoustic measurements, and produces as output the probability that each time step contains a reflection generated by a moving fish. In this scenario the prediction is performed over a buffer of recorded samples, in order to allow the network to better capture the specific structure of a moving target and to separate it from clutter reflections. However, as we shall discuss later more in detail, we also implemented an alternative *recurrent* deep learning architecture, where the prediction is performed sequentially. Such model could be preferred for embedded systems with limited memory and/or on-line processing requirements.

We trained our networks using two data sets. The first is a synthetic data set of reflections from waves and from moving targets generated by a standard numerical model of acoustic propagation. The second is built from recordings we made over 50 sea experiments performed in different sea environments, including recordings of verified clutter and of hand labeled fish reflections. This second data set allowed the production of a relatively small but diverse enough set of 4,239 recorded reflections from fish, and 5,106 recorded clutter-only reverberations. Similar to the approach adopted in [17], these recordings are used for data augmentation. In particular, the recordings were combined to generate a data set of approximately 20k training and test examples, whose SCR, as well as number of target fish can be parametrically tuned. The recordings used for generating our data set are freely available to download for reproducibility and further testing.

Our acoustic detection system is part of the EU-funded SYMBIOSIS project, whose goal is to develop a low-cost opto-acoustic monitoring observatory for long-term census of pelagic fish. The detection method reported here will serve as the first step in the detection chain. Thus, it is specifically designed to be of light computational complexity and power consumption, while at the same time able to provide favourable precision and recall trade-offs. Our contribution is twofold:

1) A novel, deep learning-based approach for the real-time mono-static detection of pelagic fish using a single active acoustic transceiver.
2) A shared data augmented database including labeled reflections from pelagic fish recorded in multiple sea environments.

The proposed approach is validated against the theoretical performance of an energy detection algorithm that only detects reflections and is thus considered an upper bound on performance, as well as against the performance of another popular machine-learning method - support vector machines [18]. Results show that our methods based on neural networks can be successfully employed in this challenging scenario, thus paving the way for a systematic application of deep learning in underwater acoustic signal processing.

The remainder of the paper is organized as follows. In Section 2, we review the related work. The acoustic detection problem is formulated in Section 3, and the details related to the generation of the two datasets are presented in Section 4. Our deep learning approach and the baseline models are described in Section 5 and validated in Section 6. Section 7 concludes the paper by discussing the strengths and weaknesses of the current system, proposing possible improvements and further validation scenarios.

2 Related Work

While passive acoustic methods are used to detect marine mammals (through their vocalization), and marine fauna and ships (through their emitted noise), remote detection of underwater targets is mostly performed via active acoustics. Clearly, the shape of the pulse and the carrier frequency used for active acoustic detection entail a trade-off between the desired resolution and detection range: the higher the carrier frequency, the better the resolution [19]. Fish-finding SONAR at medium frequencies takes advantage of the resonance frequency of the fish's swimming bladder in order to detect schools of fish [20]. Range estimation can be performed using the widely-used matched filter (MF), which correlates the reflections with the template of a transmitted wide-band signal [21], [22].

In order to estimate the velocity of a target, one can transmit a train of short pulses, and observe the delay between the reflected pulses and the transmitted pulses. This method, which is used in pulse-Doppler radar [23], requires that the propagation time to the target and back will be shorter than the pulse repetition interval (PRI), which is not applicable for long-range underwater acoustics. Another option is to directly measure the Doppler-shifted frequency of the reflected pulse [24]. For this purpose, the higher the frequency of the transmitted pulse, the greater the Doppler effect is. However, for long-range sensing using medium acoustic frequencies, the Doppler shift is relatively small, and a long, continuous wave (CW) probing pulse is required in order to detect the target's Doppler shift [25, ch.19]. A relatively new approach to overcome the slow update-rate problem is the concept of continuously active SONAR (CAS) [26], [27], [28], which applies processing on different frequency sub-bands of a long, wide-band acoustic transmission, thus achieving more detection opportunities, compared to the conventional pulse active SONAR.

Besides these well-established algorithmic solutions, statistical learning methods based on neural networks have also proven effective for SONAR signal processing. Approaches based on deep learning are now being proposed, mostly taking advantage of visual structure in SONAR images [29], micro-Doppler spectrograms [30] or cepstrum data [31], which enables use of the popular 2D convolutional neural network (CNN) architecture [32], but also requires intensive data preprocessing. The use of CNN also showed advantage for reducing the high false alarm rates in low-frequency active sonar operating in shallow water [33]. Current solutions include post-detection classification [34], classifying the output of the bearing-time matched filter using a deep neural network [35], of spectrogram classification of suspected targets by CNN [36].

While these neural network-based classifiers show remarkable results, their effective application requires large annotated datasets, which, at sea, are hard to get [35]. To circumvent this limitation, data augmentation [35] or generation of synthetic datasets [36], [37] can be used. The solution we propose here is to train the deep network using synthetic samples, or samples from real sea experiments for which ground truth information is known. If this data realistically reproduces the main features of the sea measurements, the system should be able to generalize to novel scenarios (transfer learning). This approach is also becoming popular in modern computer vision systems based on deep learning, which handle the variability of real-world data by synthetically manipulating lighting, position, and object textures in the training patterns [38], [39]. For example, in [40] retinal color images were synthesized by applying techniques based on adversarial learning, indicating that the resulting images are substantially different from the real ones, but are anatomically consistent and display a reasonable level of visual quality. A similar approach has been recently exploited to create a state-of-the-art earthquake detection system based on deep networks, which were pre-trained on a large dataset of synthetic seismic sequences [41].

3 Problem Formulation

3.1 System Model

We base our system model on the SYMBIOSIS framework, through which we aim to detect the presence of an approaching fish at a range of a few hundred meters. To determine the type of detected fish, if close enough, SYMBIOSIS uses a set of optical cameras, triggered by the acoustic system, which are used to classify the approaching fish. Hence, acoustic detection should be fast enough to allow processing before the fish moves away. Another restriction is energy-preservation, which allows system deployment at sea for several months. This dictates a low-complexity solution for the detection problem (i.e., matched filter processing or CAS are not viable options). Considering these constraints, we base our detection scheme on the emission of single-frequency short pulses. Specifically, we use a carrier frequency of $F_c = 12$ kHz, and a pulse duration of $T_s = 10$ ms. Considering a sound speed of $c \approx 1530$ m/s, this enables the detection of fish larger than roughly 0.1 m at a minimum range of roughly 7.5 m. To conserve energy, the signals are emitted at PRIs of 40 s, such that sequence-based detection is not possible.

The transceiver is omni-directional and aims to detect fish from all directions. The fish are assumed to be moving at unknown speeds, which can vary between 2 m/s for Atlantic mackerel or up to 40 m/sec for swordfish. All target pelagic fish have a swim bladder, which should reflect well for the emitted acoustic signals. Previous works reported a target strength ranging from -23 dB for Albacore Tuna to -60 dB for Mediterranean horse mackerel [42]. Considering this, the system emits the signals at a source level of 183 dB Re 1μPa @1m, allowing detection up to a range of 500 m by active acoustics. However, our measurements show that, at ranges exceeding 100 m, the signal-to-clutter ratio (SCR) is extremely low and approaches 0 dB.

3.2 General Problem Definition

Our detection scheme is aimed to detect the presence of fish, either an individual one or a school of fish. Its usage should be a first chain in a detection and tracking effort to lock onto individual targets. Our approach is based on the Doppler shift experienced for signals reflected from a moving fish. Sound reflected from a moving underwater target has a Doppler frequency shift of

$$\Delta f = f_c \cdot v/c \, , \tag{1}$$

where v is the target's velocity, relative to the receiver. Due to the low sound speed in water, the resolution needed for estimating the Doppler shift is high. For example, for $v = 10$ m/s, we have $\Delta f = 80$ Hz. However, for the above parameters, the frequency resolution obtained is only $1/T_s = 100$ [Hz]. Parabolic or Gaussian frequency interpolation can be applied to increase the resolution by more than an order of magnitude [43], but it cannot distinguish two frequency components that occupy the same frequency bin. Parametric methods, such as multiple signal classification (MUSIC) [44], can also be used for high-resolution spectral estimation. However, this method has a drawback: the number of spectral components is assumed to be known in advance, and to be uncorrelated (which may not apply to the signal and its reverberations).

To bypass this inherent resolution problem and preserve low computational complexity, we employ a multi-layer neural network with convolutional layers for the classification of reflected echoes. Convolutional neural networks (CNN) are extensively used for image classification tasks (e.g., [11]), due to their ability to learn complex spatial filters, which can effectively capture signal features in the frequency domain [45]. Here we adopt a 1-dimensional CNN architecture, and we consider three different classification tasks:

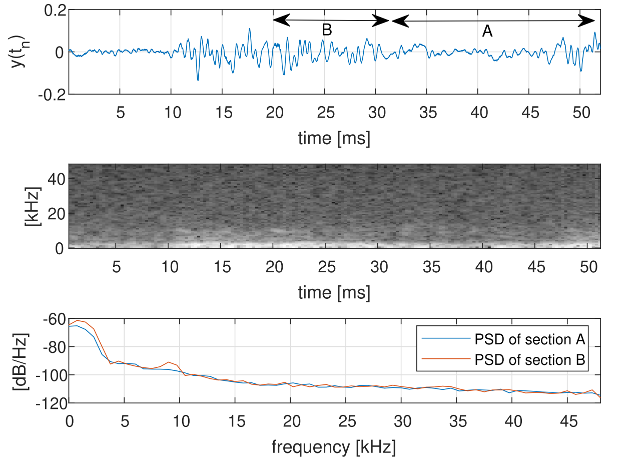

Fig. 1. An example of a recorded buffer of a raw acoustic signal from a sea experiment including active transmissions. The buffer includes a reflection from a moving fish identified between time instances 20-30 ms (i.e., Section B), and clutter reflections between time instances 10-20 ms and after 30 ms (i.e., Section A).

1) To distinguish between reflections coming from moving objects and clutter.
2) To detect the existence of a fish within a recorded buffer of reflections.
3) To detect the precise location of fish-related echoes within a recorded buffer.

Clearly, the last objective is of the highest interest, as it allows not only the detection of a fish, but also the estimation of its range, relative to the receiver. However, the first two approaches are also important: the first objective makes it possible to greatly reduce processing for detection of Doppler shift within reflected echoes. The second objective offers a rough detection for the existence of a nearby fish that can trigger the activation of other sensors and further signal processing.

The input to the CNN is a buffer of raw acoustic samples, $y[i]$ $i = 1, \dots, N$, recorded immediately after the emission of the signal. To maintain low computational complexity, the buffer samples are only pre-processed by a band-pass filter considering the carrier frequency of the emitted signals and of the frequency band determined by the Doppler shift for an assumed maximal target velocity. For all of the above classification objectives, a binary hypothesis test of $\{\mathcal{H}_0, \mathcal{H}_1\}$ is performed for each sample i such that

$$y[i] = \begin{cases} n[i] + c[i], & \mathcal{H}_0 \\ n[i] + x[i], & \mathcal{H}_1 \end{cases}, \tag{2}$$

where $n[i]$ is an additive noise sample, $c[i]$ is a clutter sample (i.e., reflection from a sea boundary, volume scatterer, or a stationary target) and $x[i]$ is a signal reflected from a moving target (e.g., a fish). An example of such a recorded buffer from a sea experiment, including a reflection from a fish, is shown in Fig. 1. By showing thew similarities between the power spectral density of fish-based reflection samples (Section B) and clutter-based reflections (Section A), we demonstrate the challenge of detecting fish reflections from the raw acoustic signal. In particular, we observe the complexity of the reflected signal, which is hard to model, and thus difficult to simulate, and the low SCR, which makes detection particularly challenging.

4 Construction of Data Sets

In this section, we describe the construction of two data sets used to train the deep networks. The data sets are built as an ensemble of vectors including time-domain raw acoustic signals. Signals are band-passed to include only the frequency range for the expected fish-related reverberated signal; that is, taking into account the carrier frequency of the single-tone transmitted signal, and the expected Doppler shift for the fish's upper bound speed. Both noise-only and target-included vectors are generated. For the latter, the vectors are annotated with the start and end indexes of the fish-related reflection. Our database is designed to fulfill the following requirements:

1) Be large enough to allow robust training of a relatively deep neural network.
2) Provide an equalized number of noise-only signals and target-included acoustic vectors.
3) Be diverse enough to include synthetic signals or real recordings from multiple simulated or measured sea environments, respectively.

The data sets account for echoes reflected from either a moving pelagic fish, static objects or clutter. Since the deep network operates on raw acoustic data, we limit the size of the created buffers to 0.7 s buffers divided into seven 0.1 s long chunks of 4800 samples. The preparation stages for these buffers are outlined below.

4.1 Synthetic Data Set

To form realistic reverberations, we simulate channel impulse responses for a moving target, h_{mobile}, and for static targets, h_{static}. The simulations are formed by a two-stage reuse of the widely used Bellhop model [46] whose input are environmental parameters such as bathymetry, sound speed profile, and bottom and surface roughness, and whose output is a complex channel impulse response for a given transmitter and receiver location pair.

To simulate clutter reflections from the sea surface and bottom, we consider a receiver stationed on the sea boundaries. To simulate reflections from stationary targets, we place the receiver on randomly located rock reflectors, and attenuate the signals received by the receiver by an additional typical target strength from rocks. The same procedure is performed to mimic a moving pelagic fish, but here also the signal is interpolated using a randomly determined Doppler shift uniformly distributed between 2 and 40 m/s added to the received signal.[1] Then, in the second stage, the receiver becomes the emitter for the Bellhop simulations, and transmits back the received signals as reflections towards the location of the original transmitter. This results in reflections of three kinds: wave clutter, stationary targets, and mobile target. Convolving a synthetic single-carrier frequency signal with all three channels, we obtain a clutter pattern as well as a multi-path structure for mobile and static reflections. The above separation to three reflectors' types allows also to control the SCR of the received signal. Finally, to reduce the size of the neural network, we shift

[1] We note that this limitation is used only in the construction of the simulation setup and not in the data analysis process nor in the processing of the real recordings.

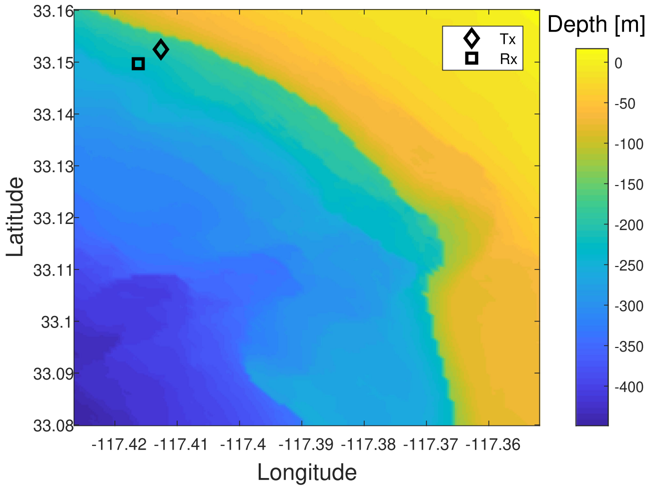

Fig. 2. Bathymetry data used for the Bellhop simulations. Examples of the transceiver and target locations are marked.

the outcome to base-band and decimate the result by a factor of 2.

The considered bathymetry is shown in Fig. 2. We sampled 18000 Monte-Carlo scenarios. In each scenario, the locations of the transmitter and target are randomized within the considered area with a maximal distance of 500 m. The depths of the transmitter and target are also randomly uniformly selected. Then, to demonstrate robustness, the channel's bathymetry is randomized in each simulation run. The flow of the preparation of a scenario buffer is illustrated in Fig. 3, where the ambient noise is denoted by $n(t_n)$. An example of the preparation steps is shown in Fig. 4, where $x_1(t), x_2(t)$ represent the mobile and static targets, respectively, and $x(t)$ represents the joint simulated signal after the addition of noise.

4.2 Real Recordings From Sea Experiments for Data Augmentation

Our second data set includes data augmentation obtained from recordings during sea experiments of clutter noise and confirmed target fish. Different than ambient noise which is considered low compared to the received signal's reflections, the former is a vector of samples formed by recordings of reflections from the sea boundaries, volume scatterers, and static targets, and is confirmed not to include a moving target. In the training process, we both mix and isolate clutter sample vectors from different sea environments to explore the robustness of our detector. The latter is a vector of samples identified to belong to a moving target, and is much shorter than clutter vectors. The recordings are taken from 50 different sea experiments we have conducted in open water in the Mediterranean Sea and in the Red Sea. For transmitting we used the EvoLogics 7-17 software defined acoustic modem, whose omni-directional transmitter emits signals sampled at 62.5 ksps for a frequency range of 7-17 kHz at a source level of 182 dB Re 1μPa. For reception we used an omni-directional self-made recorder based on a Taskam recorder, which continuously recorded 24 bit resolution signals sampled at 96 ksps with a fixed pre-amp gain of 3 dB. Fig. 5 shows a picture of the system at sea. Each of these experiments included an acoustic transceiver emitting 10 ms narrow-band signals of 12 kHz, and recording their reflections for 0.7 s. The choice of a single tone is to allow the network to lock onto frequency offsets caused by the Doppler shift effect which is more observable in single tone signals. The choice of 12 kHz as a carrier frequency is to match the transmission to the resonance frequency of our acoustic projector. As will become clear below, to offline identify if these recordings include reflections from moving targets, we also emitted 10 ms wideband linear chirp signals of 7-17 kHz. An example of a received signal that contain both chirp (LFM) and narrow band (CW) transmissions is shown on Fig. 7. Each experiment lasted at least an hour, and hence resulted in hundreds of clutter-based and target-based vectors.

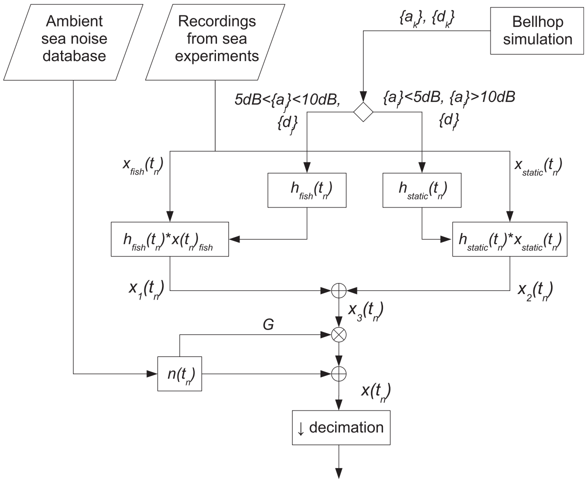

Fig. 3. Preparation flow of a signal for the synthetic dataset.

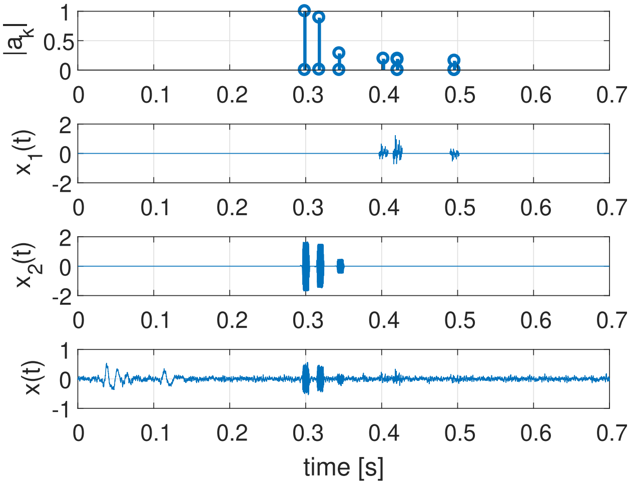

Fig. 4. Buffer preparation stages. Top: magnitudes and delays of the merged channel impulse response. Second panel: simulated mobile target. Third panel: simulated static target. Bottom: merged signal with ambient noise.

To combine the clutter and target vectors, we randomly uniformly pick a window of samples within the clutter vector and replace these clutter-based samples with target-based samples.[2] The target-based samples is chosen as a 20 ms long signal section around the position identified to

[2] Note that this replacement is different than what is usually used in additive noise channels. The reason is that we consider the model in (2) where either clutter- or target-based reflections exists.

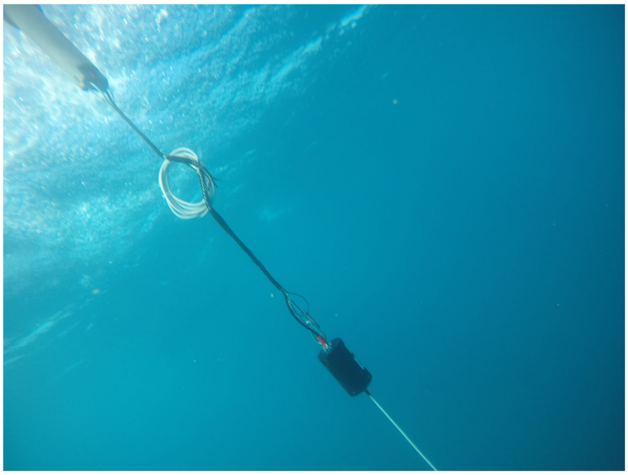

Fig. 5. A picture from our sea experiments showing the transmitting modem.

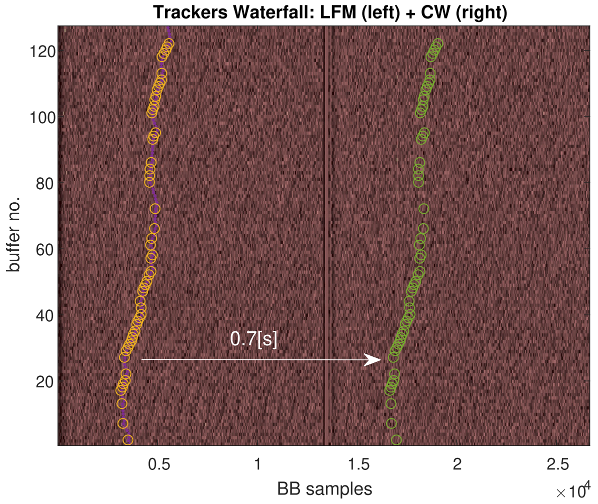

Fig. 6. Waterfall plot of the matched filter response to the chirp signal.

contain reflection from a verified target. This yield 1921 samples sampled at 96 kHz, which we convert to base-band and decimate to be 960 samples. This is approximately one fifth of the total buffer length. The level of the resulting target-originated signal section was normalized for a desired SCR. Labeling of each of these clutter or target vectors is done per time-sample, i.e., only samples withing the boundaries of the 20 ms identified target reflection are labeled as 'target' and all other time-samples are labeled as 'no-target'. Here too we either combine clutter and target from the same sea environment or mix different environments. The repository of real target-related reflections contains 4,239 samples.[3]

The following tasks summarize the process of obtaining the data augmented database:

1) Within an acoustic recording, detect samples associated with reflections from a fish and clutter-only (using the method in [22]).
2) Form buffers of 0.1 s of clutter-only samples, and buffers of 20 ms of samples from fish reflections.
3) Randomly pick buffers of clutter and target, and merge the two by replacing samples within the clutter-only buffer with the samples of the fish reflection. Merging is performed in a ratio set by a desired SCR.
4) Shift buffer to baseband, filter around the bandwidth of the emitted signal, and decimate.

The process of identifying or excluding a target to form our data base is based on the track-before-detect procedure described in [22], and is performed offline. The key idea of this method is reported here for completeness. The method works by forming a time-distance waterfall from a sequence of recorded reflections of wide-band chirp signals, followed by a single carrier signal. The waterfall is constructed by stacking the outputs of a normalized matched filter for each of the recorded reflections. As shown in the example of Fig. 6, the mobile target is indicated by the two curved lines of time-varying location. The left line marks the range that is identified by matched-filter processing of the LFM signal, and the right line – the estimated range corresponding to the reflection of the single-tone signal. The left line is identified through a constraint Viterbi algorithm run. For the latter, the states and observations are represented by the columns (distance) and rows (time), respectively, while the emissions are the normalized matched filter outputs and the transition matrix is set to allow change between consecutive observations only up to a maximal allowed distance. Once the mobile target is identified through the above process, the corresponding buffer of target reflections from the single-carrier frequency signal are identified and traced back to the raw acoustic signal to form the above target vector.

Finally, a validation process is executed for the identified vector of target's reflected samples by calculating the target's speed. Here, we consider only reflections identified to include Doppler shift ratios for target velocities greater than 1 m/s. Due to the sample resolution problem, we calculate this speed by measuring the time-of-arrival offset of the identified target's buffer between consecutive detections. The target identified in the example Fig. 6 corresponds to a reflection from a Albacore Tuna fish recorded during a sea experiment in Northern Israel. The velocity of the Tuna fish is estimated to be roughly 3.8 m/s.

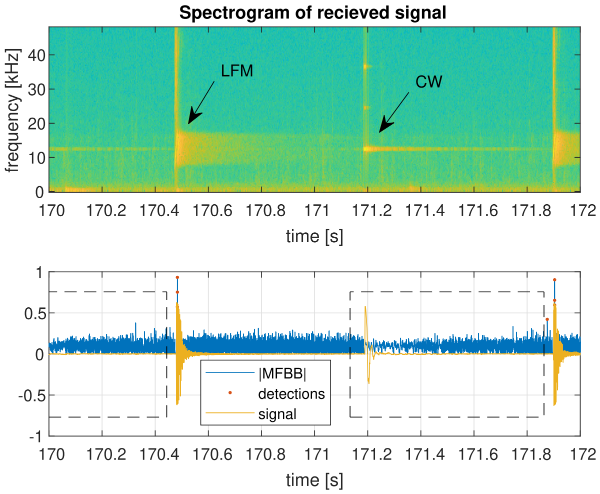

Fig. 7. Received signal, recorded during a sea experiment.

[3] The complete repository of clutter and reflection signals can be found in https://drive.google.com/drive/folders/1E8gV4yV8VnhJgne24bTLHf8nFzlnU4ey

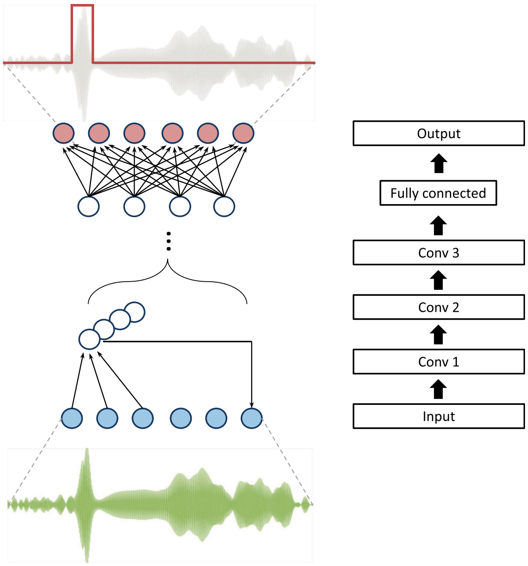

Fig. 8. Graphical representation of the CNN architecture.

5 Detection Methodology

5.1 Convolutional Deep Learning Model

Our preferred detection system is implemented as a deep convolutional neural network (CNN) receiving as input the whole vector of acoustic signals on a set of 4,800 visible units. The input is then elaborated through a series of hidden layers, with the aim of non-linearly mapping each input value to the correct underlying class (i.e., signal *versus* noise). The network is composed of three 1D convolutional layers (without pooling) followed by a final, fully-connected layer. The hyperparameters defining the network architecture (number of filters, filter sizes) and the learning procedure (initial learning rate, dropout factor) were tuned using a random-search optimization procedure.[4] This resulted in three convolutional layers composed of 32, 64, and 128 filters, with kernel sizes of 5, 8, and 20, respectively. The final fully-connected layer contains 100 units (see Fig. 8 for a graphical representation). Rectified linear units are used in all layers. A dropout factor of 0.6 is applied to the fully-connected layer as a regularizer [47], and the initial learning rate is set to 0.001. Overall, the CNN has approximately 900k trainable parameters. Note that the system operates in a purely feed-forward manner and does not require any iterative computation.

5.1.1 CNN Training

The loss function to be minimized is the cross-entropy between the correct class (ground truth labels) and network prediction. Due to the possible unbalancing of output classes (i.e., there might be many more noise samples than signal samples), a weighted form of cross-entropy is used, where the positive class weight is estimated according to frequency of occurrence in the training set.

For both simulations and real recordings data sets, the network is trained on a random sample containing 70 percent of the data, and final performance is tested on the remaining 30 percent of the data. For the data set of real channel recordings from the sea experiments, the test patterns are sampled from different sea conditions in order to test model generalization on novel scenarios. The rationale for selecting the test patterns in the real recordings data set is as follows:

- In Configuration 1, we train using clutter data from all available environments, and test using clutter data from only a single environment (*SHARK*). Both training and test echoes are selected from all environments.
- In Configuration 2, only data from a specific sea experiment (*EILAT*) is considered to see if testing and training on the same environment would increase performance.

In both cases, examples for the test and training set are taken from different WAV files.

Training proceeds for a maximum of 10,000 epochs and learning rate is dynamically adjusted using the ADAM optimizer [48]. In order to prevent overfitting, an early-stopping criterion is adopted: the loss is constantly monitored on a separate validation set (30 percent of patterns randomly chosen from the training set) and training stops if no improvements are observed during the last 1,000 epochs. The CNN is trained off-line using high-performance graphic processing units; once trained, it can be effectively deployed in real-time using low-performance computing hardware.

5.1.2 CNN Testing

Model performance is assessed by computing precision and recall metrics, which are used to produce receiver operating characteristic (ROC) curves [49]

$$\begin{aligned}
&Precision = \frac{TP}{TP + FP};\\
&Recall = \frac{TP}{TP + FN};\\
&TruePositiveRate = \frac{TP}{TP + FN};\\
&FalsePositiveRate = \frac{FP}{FP + TN},
\end{aligned} \tag{3}$$

where TP indicates True Positives, TN True Negatives, FP False Positives and FN False Negatives. This way, we can more faithfully evaluate classification confidence as a function of the binarization threshold imposed over the CNN predictions. Though the model achieving the best area under the curve in ROC space is not guaranteed to also have the best area under the curve in precision-recall space, we always found perfect agreement between these two measures in our analysis. We also perform a random visual inspection on the CNN predictions in order to qualitatively evaluate the detection capability on a variety of input signals.

5.2 Recurrent Deep Learning Model

An alternative detection system is implemented as a recurrent long-short term memory (LSTM) neural network,

[4] Search values and intervals: number of convolutional layers: $\{1, 2, 3, 4\}$; number of kernels: $[32 : 32 : 160]$; kernel size: $[4 : 1 : 30]$; number of fully-connected units: $[50 : 50 : 200]$; dropout: $[0.1 : 0.1 : 0.7]$; initial learning rate: $\{0.0001, 0.001, 0.01, 0.1\}$. Random search was carried out by sampling 300 hyperparameter configurations.

sequentially receiving as input the vector of acoustic signals on a single visible unit. Also in this case, the input is elaborated through a series of hidden layers with the aim of non-linearly mapping each input value to the correct underlying class. The network consists of a stack of two LSTM layers and a final output unit with sigmoid activation function. Due to the computational complexity of this model, hyperparameters were optimized using a grid-search procedure with a limited range of values for each parameter.[5] This resulted in a network composed of two layers with 128 units, a dropout factor of 0.2, and initial learning rate of 0.01. Overall, the LSTM model has approximately 140k trainable parameters.

5.2.1 LSTM Training and Testing

The training and testing setup for the LSTM is the same of the CNN. However, the LSTM struggled in learning such long sequences, hence the patterns in the training set were segmented into shorter sequences of 480 elements. The sequences in the test set remained unaltered. The loss function to be minimized is the cross-entropy between the correct class (ground truth labels) and the network prediction. Training proceeds for a maximum of 1,000 epochs and learning rate is dynamically adjusted using the ADAM optimizer. In order to prevent overfitting, also in this case an early-stopping criterion is adopted: the loss is constantly monitored on a separate validation set (30 percent of patterns randomly chosen from the training set) and training stops if no improvements are observed during the last 100 epochs.[6]

5.3 Benchmark Models

5.3.1 Support Vector Machines

For the binary fish detection task, the performance of deep learning models is validated against Support Vector Machines (SVMs) [18] — a more traditional popular class of machine-learning models that have also been successfully applied to underwater signals (e.g., [9]). In particular, we test two types of SVMs with both linear and non-linear kernels (polynomial and radial basis function), trained through sequential minimal optimization [50] for a maximum of 30 objective evaluations using a 4-fold cross-validation scheme. Hyper-parameters are optimized using a Bayesian optimization criterion [51] and ROC curves are produced by ranking the data according to the predicted score.

5.3.2 Energy Detector

To further validate our models, we consider the energy detector (ED) for an integration time window of T s over the band-passed raw acoustic signals. Considering (2), we operate an ED for a two-hypothesis problem, namely, a target-included hypothesis $\mathcal{H}_1$, and a noise-only hypothesis $\mathcal{H}_0$, and model the output of the the ED as

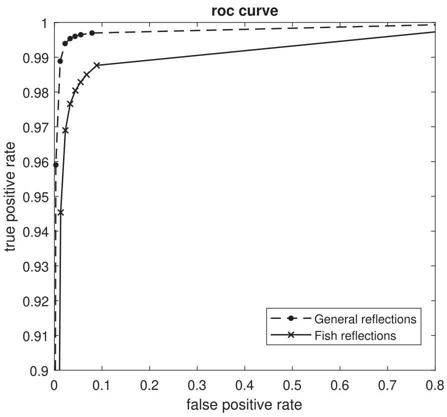

Fig. 9. ROC curve for the simulated data, distinguishing between the easier task of detecting and localizing general reflections (dashed line) and the harder task of detecting and localizing fish-specific reflections (solid line).

$$\mathcal{H}_0 : \mathrm{z_T} = n; \quad z_{\mathrm{noise}} = n \tag{4a}$$

$$\mathcal{H}_1 : \mathrm{z_T} = P + n; \quad z_{\mathrm{noise}} = n \ , \tag{4b}$$

where z_T and z_{noise} denotes the ED's output for noise and target, respectively, n is an i.i.d Gaussian noise of variance N_0,[7] and P is the amplitude of the reflection. By (4), the ED performs a much easier task than separating reflections of a mobile target from clutter. Hence, we consider its performance as an upper bound that could be obtained by the deep learning detectors.

For a received signal z, we consider a normalized ED decision parameter, set by [52]

$$d = \frac{z - E[z_{\mathrm{noise}}]}{\sqrt{\mathrm{var}(z_{\mathrm{noise}})}} \ , \tag{5}$$

where z_{noise} is identified by an initialization process or by accumulating signals z determined as noise. Let d_{noise} be the decision parameter (5) for the case of noise only. Since d_{noise} is normalized, its mean and variance are

$$\begin{aligned} E[d_{\mathrm{noise}}] &= 0 \ , \\ \mathrm{var}(d_{\mathrm{noise}}) &= 1 \ . \end{aligned} \tag{6}$$

Hence, the relation between false-alarm probability and the detection threshold, d_{T}, can be written as [53]

$$P_{\mathrm{fa}} = \frac{1}{2}\mathrm{erfc}\left(\frac{d_{\mathrm{T}} - E[d_{\mathrm{noise}}]}{\sqrt{2\mathrm{var}(d_{\mathrm{noise}})}}\right) = \frac{1}{2}\mathrm{erfc}\left(\frac{d_{\mathrm{T}}}{\sqrt{2}}\right) \ , \tag{7}$$

where erfc is the complementary error function.

By (5)

$$E\big[d_{\mathrm{signal}}\big] \cong \frac{PT}{N_0\sqrt{WT}} \ , \tag{8a}$$

$$\mathrm{var}(d_{\mathrm{signal}}) \cong 1 + \frac{2E}{N_0 WT} \ , \tag{8b}$$

[5] Search values: number of layers: $\{1, 2\}$; number of units: $\{64, 128, 256\}$; dropout: $\{0.1, 0.2, 0.3, 0.4\}$; initial learning rate: $\{0.001, 0.01, 0.1\}$.

[6] Learning in the LSTM required fewer epochs to converge, however it should be noted that the segmented training set size increases tenfold compared to the CNN training set. Overall, training times were significantly longer for the LSTM, also due to the more limited parallelization exploitable by recurrent models. Average training times (workstation equipped with an NVIDIA Titan Xp GPU card) were 04:40±00:10 minutes for the CNN and 18:20±00:20 minutes for the LSTM.

[7] We justify the Gaussian assumption since the integration period, T, is relatively long.

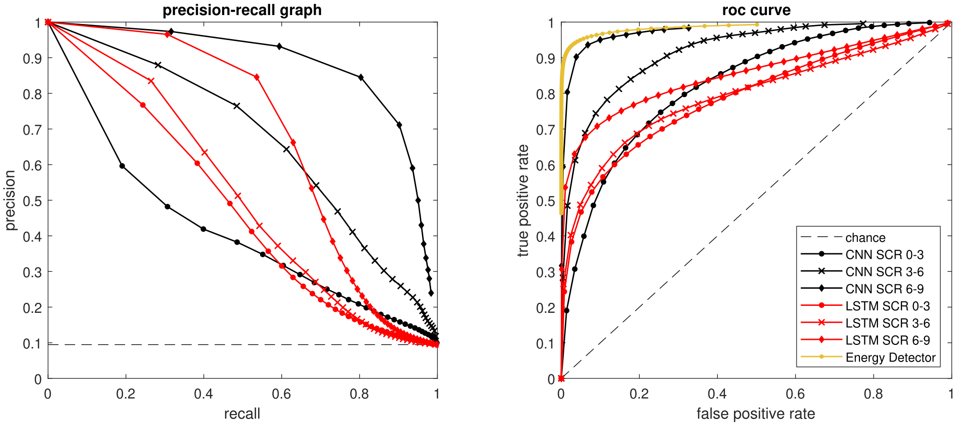

Fig. 10. Precision-recall and ROC curves comparing the CNN (black) and LSTM (red) performance on the first configuration for real sea recordings, for the task of detecting and localizing fish-specific reflections. Different SCR levels are shown using different markers showing high impact of clutter. Dashed lines report chance level, while the yellow curve correspond to the Energy Detector upper-bound.

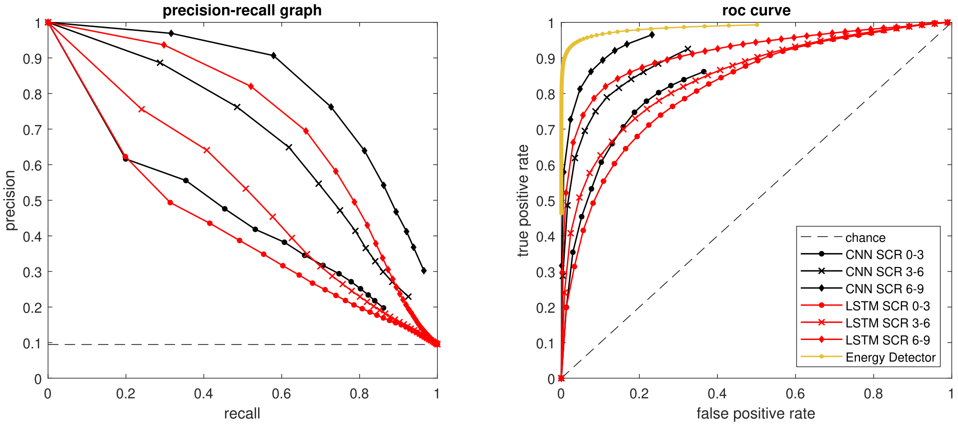

Fig. 11. Precision-recall and ROC curves comparing the CNN (black) and LSTM (red) performance on the second configuration for real sea recordings, for the task of detecting and localizing fish-specific reflections. Different SCR levels are shown using different markers showing high impact of clutter. Dashed lines report chance level, while the yellow curve correspond to the Energy Detector upper-bound.

where W is the frequency band of the band-passed signal, and N_0 is the noise variance. Thus, the detection probability can be approximated by

$$P_{\mathrm{d}} = \frac{1}{2}\mathrm{erfc}\left(\frac{d_{\mathrm{T}} - E\left[d_{\mathrm{signal}}\right]}{\sqrt{2\mathrm{var}(d_{\mathrm{signal}})}}\right) = \frac{1}{2}\mathrm{erfc}\left(\frac{d_{\mathrm{T}} - \frac{PT}{N_0\sqrt{WT}}}{\sqrt{2\left(1+\frac{2PT}{N_0WT}\right)}}\right) . \tag{9}$$

Since our detector performs per-sample decisions, we define the SNR as the power ratio

$$\mu = \frac{P}{N_0 W} . \tag{10}$$

Thus, we get from (7) and (9) the ROC

$$\mu = \sqrt{\frac{2}{WT}}\left(\mathrm{erfc}^{-1}(2P_{\mathrm{fa}}) - \mathrm{erfc}^{-1}(2P_{\mathrm{d}})\right) . \tag{11}$$

6 Results

6.1 Performance on Simulated Data

For the simulated data, it turns out that distinguishing between general reflections coming from moving objects and clutter is quite straightforward for the CNN. As can be appreciated by looking at the dashed line in Fig. 9, classification accuracy is extremely high. The ROC curve is almost at ceiling, which means that the CNN is able to accurately distinguish general reflection patterns from background noise. This result confirms that, contrary to fully-connected models, convolutional networks are good for capturing periodic patterns in the input signal, making them perfect candidates for performing efficient frequency analysis [45]. At the same time, the lower performance achieved in the harder task of detecting fish-specific reflections (solid line in Fig. 9) suggests that the CNN cannot always reliably discriminate between a general signal reflection and a reflection produced specifically by a fish.

6.2 Performance on Data Augmentation Dataset

For the data augmentation dataset, we only consider detection tasks related to fish-specific reflections. As shown in Figs. 10 and 11 both models exhibit good classification accuracy, and the CNN in particular achieves a remarkable performance.[8]

[8] To maintain readability of the Figures, we only show curves related to the *test set* performance. However, it should be noted that the gap between training and test set was relatively small, suggesting that the models achieved an acceptable bias-variance trade-off.

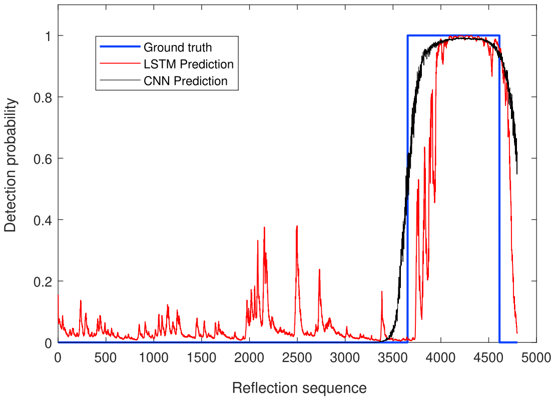

Fig. 12. Example of prediction for the CNN (black) and LSTM (red) on a pattern from the data augmentation dataset from real sea recordings. Ground truth is represented by the blue curve.

As expected, performance of both models is modulated by the level of noise injected in the signals: for high levels of noise (SCR in the range between 0 and 3) the accuracy is far from ceiling, but still satisfactory. For low noise levels (SCR in the range between 6 and 9) the CNN performance approaches that of the Energy Detection bound (see yellow curve, computed with a signal-to-noise level of 2 dB), thus demonstrating that noise level is critical for system performance. It should be noted that different training and testing conditions (Figs. 10 *versus* 11) lead to different performance. This is because the considered environments are fundamentally different, ranging from shallow sandy beach, deep reef, and open water. When the SCR is low, this complexity may confuse the network leading to better results for configuration 2 (which contains data from only one environment). However, for higher SCR, the diversity may assist to better distinguish the fish reflection from the clutter. This is because, being a stationary signal, the former has a more organized structure than the latter, leading to better performance in the configuration 1.

As shown in the test example reported in Fig. 12, the classification operated by the CNN is much more sharp and accurate than that produced by the LSTM. The advantage of the CNN stems from the fact that such model can exploit information from all the 4,800 reflections contained in the buffer in order to classify each time step: indeed, although the kernels of early layers only process a limited portion of the input during each convolution, the resulting feature maps are subsequently processed using increasingly larger kernels, and the final fully-connected layer eventually processes information spanning the entire input space. Vice versa, the LSTM model processes the sequence on-line, and can thus only exploit past information for making the current time step prediction. This phenomenon is visible especially at the beginning of the ground truth portion of the sequence, where the LSTM needs to gradually accumulate a sufficient amount of evidence before increasing the output detection confidence.

Interestingly, even the CNN previously trained on the simulated data is able to partially succeed in detecting fish on the realistic data set obtained from the sea experiments (Fig. 13), despite the fact that the overall accuracy is lower compared to the CNN trained directly on the real recordings. This suggests that transfer learning might be a viable approach to tackling system complexity, since the synthetic signal can be more easily augmented with ground-truth information. However, the Bellhop model should likely be further enriched in order to better match the features of real sea measurements, for example by incorporating a physical model for the complex signal distortion occurring within the fish's body.

The CNN also achieves a good performance on the binary task of detecting the existence of a fish within the recorded buffer of reflections (Fig. 14). This result is expected, given that the binary detection task can be considered a simplified sub-type of the more general task of estimating the precise location of fish reflections. Indeed, the ROC curves shown in Fig. 14 are well aligned with those in Figs. 10 and 11. Also in this case, we observe the clear impact of noise on the detection performance: at challenging noise levels (SCR in the range 0-3) accuracy drops for both testing configurations.

Crucially, none of the benchmark SVM models were able to achieve satisfactory performance on the binary detection task. PR and ROC curves for the best performing SVM model (polynomial kernel of degree 3) in both configurations are shown in Fig. 15. A comparison between the CNN and the SVM in terms of classification accuracy is instead reported in Table 1. This result confirms that traditional machine-learning approaches may not be applicable to the

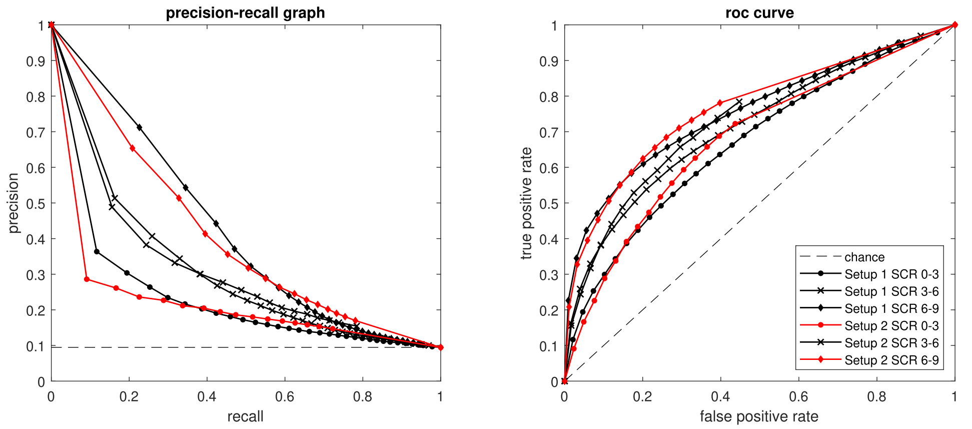

Fig. 13. Precision-recall and ROC curves of the data augmentation dataset from real recordings for the task of detecting and localizing fish-specific reflections in a transfer learning setting. Dashed lines report chance level.

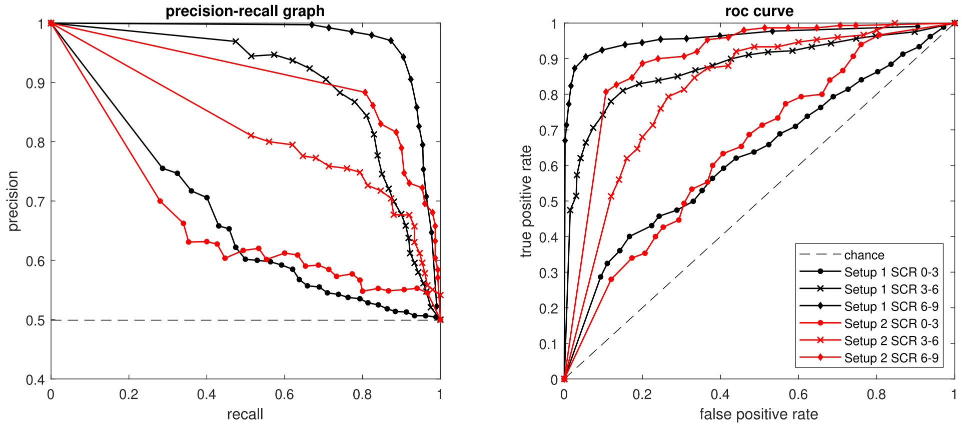

Fig. 14. Precision-recall and ROC curves for the data set of real recordings for the binary task of detecting the existence of a fish within the recorded buffer of reflections, on data set Configurations 1 (black) and 2 (red). Different SCR levels are shown using different markers. Dashed lines report chance level.

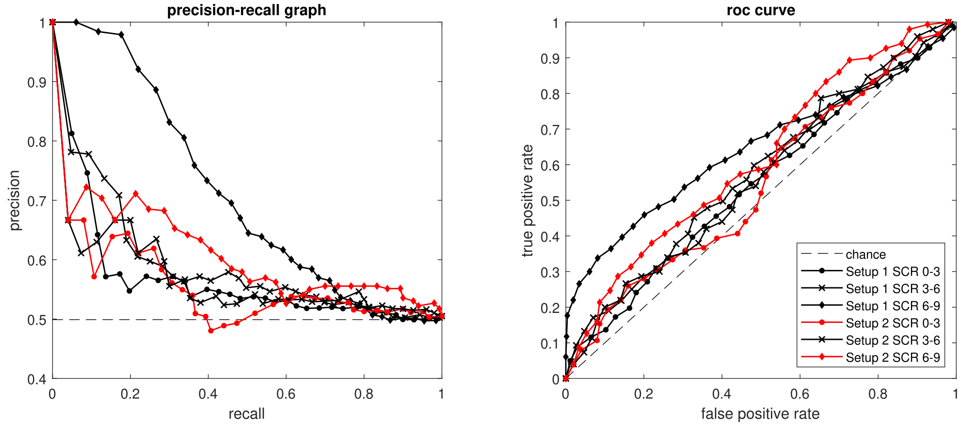

Fig. 15. Precision-recall and ROC curves for the best performing SVM model, on data set Configurations 1 (black) and 2 (red). Different SCR levels are shown using different markers. Dashed lines report chance level.

present setting: the high-dimensionality and the complexity of the acoustic signals require using more efficient and scalable approaches, such as those based on neural networks. It should also be noted that the computational cost to perform real-time predictions on the test data is much higher for SVMs compared to the CNN, which further motivates the use of deep networks to guarantee efficient deployment on sea platforms like SYMBIOSIS.

Finally, for a qualitative assessment of the results we also report some detection samples for the CNN in Fig. 16. Chunks of band-passed acoustic signals are also shown by highlighting the portion corresponding to a fish reflection (green color). As can be appreciated by looking at the curves representing the corresponding CNN detection confidence, the network is able to precisely discriminate the portion of the signal containing a fish reflection from those containing only noise or clutter. In line with the performance shown in the PR and ROC curves, detection confidence is slightly higher for the network trained with the first configuration (top panels) compared to the second configuration (bottom panels).

TABLE 1
Classification Accuracy for the Binary Task of Detection Fish Events Within the Whole Buffer of Reflections

	Configuration 1			Configuration 2		
	scr 6-9	scr 3-6	scr 0-3	scr 6-9	scr 3-6	scr 0-3
CNN	0.89	0.81	0.59	0.84	0.77	0.60
SVM	0.61	0.55	0.54	0.58	0.54	0.50

7 Conclusion

In this paper, we presented a deep learning approach for the task of real-time detection of fish via active acoustics in low signal-to-clutter ratio conditions. Considering the challenging and hard-to-model underwater acoustic environment, our approach takes advantage of both numerical models and field experiments performed in a variety of sea environments to form a freely shared database of real reflection patterns that is used to train and test the neural networks. The labeling procedure for the recorded data was based on a chain of detection verification, starting from a track-before-detect method and ending with the evaluation of speed to detect motion. The performance of two deep learning models, namely convolutional neural networks and long-short-term-memory networks, was systematically evaluated using different databases and tasks, and benchmarked against a more traditional machine-learning approach (Support Vector Machines) and an upper bound based on an energy detector. The results show a favourable trade-off between precision and recall for both deep learning models, far exceeding the performance of the SVM. Our analysis also

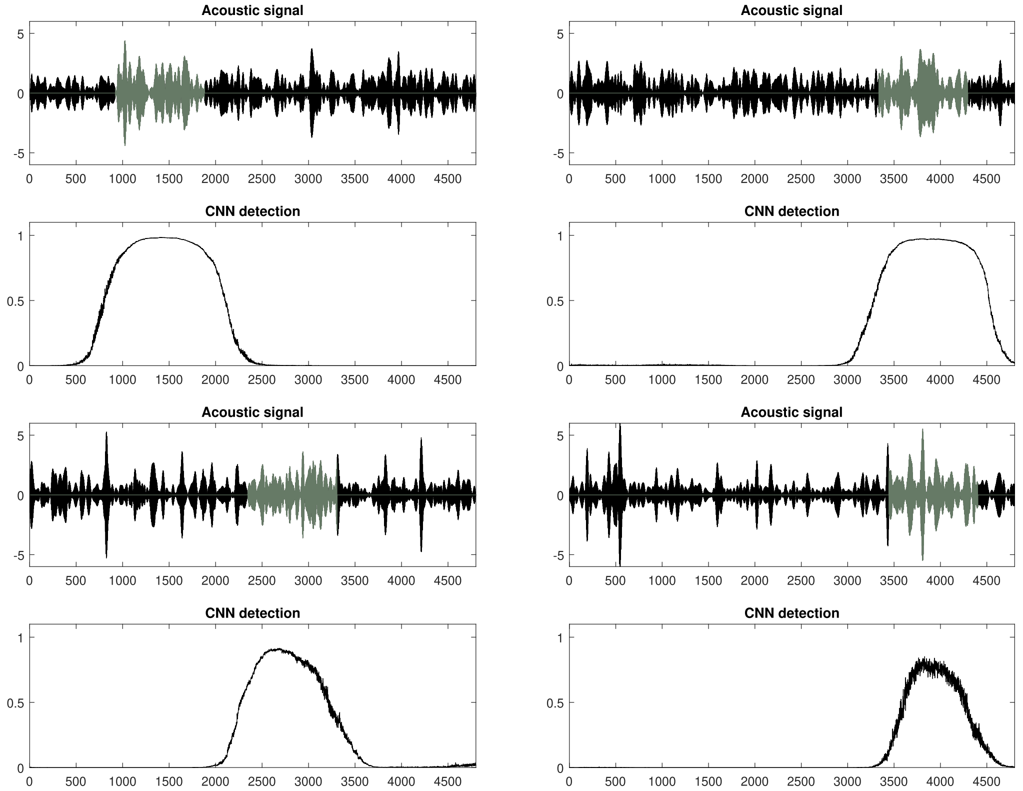

Fig. 16. Samples of CNN detection corresponding to the first data set of real recordings (top) and to the second data set (bottom) at an SCR level ranging from 3 to 6 dB.

highlighted a clear advantage for the CNN model over the LSTM model, which stems from the possibility to use all information available in the buffer to carry out the detection.

Future research should further extend the capabilities of the proposed model, for example by investigating the possibility of predicting motion type and fish species, thus paving the way for the implementation of a completely automatic and efficient system for monitoring marine ecosystems.

Acknowledgments

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 773753 (SYMBIOSIS), and by the Israeli Ministry of Energy Under Grant Number 219-17-013. The work of Alberto Testolin was supported by the NVIDIA Corporation for the donation of a Titan Xp GPU used for this research.

References

[1] E. Pikitch *et al.*, "Ecosystem-based fishery management," *Science*, vol. 305, pp. 346–347, 2004.

[2] K. Colbo, T. Ross, C. Brown, and T. Weber, "A review of oceanographic applications of water column data from multibeam echosounders," *Estuarine Coastal Shelf Sci.*, vol. 145, pp. 41–56, 2014.

[3] European Commission Maritime affairs Integrated maritime policy. Accessed: Apr. 2, 2019. [Online]. Available: https://ec.europa.eu/maritimeaffairs/policy/blue_growth_en

[4] M. Nagaso, K. Mizuno, A. Asada, K. Kobayashi, and M. Matsukawa, "Development of the three-dimensional visualization method for the inner structure of small size fish using 25 MHz acoustic profile measurement," in *Proc. OCEANS - San Diego*, 2013, pp. 1–4.

[5] A. Munafò, G. Canepa, and K. D. LePage, "Continuous active sonars for littoral undersea surveillance," *IEEE J. Ocean. Eng.*, vol. 44, no. 4, pp. 1198–1212, Oct. 2019.

[6] S. Schoenecker, P. Willett, and Y. Bar-Shalom, "Resolution limits for tracking closely-spaced targets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 6, pp. 2900–2910, Dec. 2018.

[7] S. Schoenecker, P. Willett, and Y. Bar-Shalom, "ML-PDA and ML-PMHT: Comparing multistatic sonar trackers for VLO targets using a new multitarget implementation," *IEEE J. Ocean. Eng.*, vol. 39, no. 2, pp. 303–317, Apr. 2014.

[8] S. Schoenecker, P. Willett, and Y. Bar-Shalom, "The effect of K-distributed clutter on trackability," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 475–484, Jan. 2016.

[9] R. Diamant *et al.*, "On the relationship between the underwater acoustic and optical channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8037–8051, Dec. 2017.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015, Art. no. 436.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[12] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 6645–6649.

[13] E. Gawehn, J. A. Hiss, and G. Schneider, "Deep learning in drug discovery," *Mol. Informat.*, vol. 35, no. 1, pp. 3–14, 2016.

[14] H. Y. Xiong *et al.*, "The human splicing code reveals new insights into the genetic determinants of disease," *Science*, vol. 347, no. 6218, 2015, Art. no. 1254806.

[15] M. Zorzi, A. Zanella, A. Testolin, M. D. F. De Grazia, and M. Zorzi, "Cognition-based networks: A new perspective on network optimization using learning and distributed intelligence," *IEEE Access*, vol. 3, pp. 1512–1530, 2015.

[16] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nat. Commun.*, vol. 5, 2014, Art. no. 4308.

[17] B. Tang, Y. Tu, Z. Zhang, and Y. Lin, "Digital signal modulation classification with data augmentation using generative adversarial nets in cognitive radio networks," *IEEE Access*, vol. 6, pp. 15 713–15 722, 2018.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[19] A. Popper and M. Hastings, "The effects of anthropogenic sources of sound on fishes," *J. Fish Biol.*, vol. 75, no. 3, pp. 455–489, 2009.

[20] T. K. Stanton, D. Chu, J. M. Jech, and J. D. Irish, "New broadband methods for resonance classification and high-resolution imagery of fish with swimbladders using a modified commercial broadband echosounder," *ICES J. Marine Sci.*, vol. 67, no. 2, pp. 365–378, 2010.

[21] D. A. Abraham and P. K. Willett, "Active sonar detection in shallow water using the page test," *IEEE J. Ocean. Eng.*, vol. 27, no. 1, pp. 35–46, Jan. 2002.

[22] R. Diamant, D. Kipnis, E. Bigal, A. Scheinin, D. Tchernov, and A. Pinhasi, "An acoustic track-before-detect approach for finding underwater mobile targets," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 104–119, Mar. 2019.

[23] M. Skolnik, *Radar Handbook*, 3rd ed. New York, NY, USA: McGraw-Hill, 2008.

[24] C. Jauffret, A.-C. Pérez, P. Blanc-Benon, and H. Tanguy, "Doppler-only target motion analysis in a high duty cycle sonar system," in *Proc. 19th Int. Conf. Inf. Fusion*, 2016, pp. 2163–2170.

[25] R. P. Hodges, *Underwater Acoustics: Analysis, Design and Performance of Sonar*. Hoboken, NJ, USA: Wiley, 2010.

[26] S. M. Murphy and P. C. Hines, "Sub-band processing of continuous active sonar signals in shallow water," in *Proc. OCEANS Genova*, 2015, pp. 1–4.

[27] G. Canepa, A. Munafò, M. Micheli, L. Morlando, and S. Murphy, "Real-time continuous active sonar processing," in *Proc. OCEANS Genova*, 2015, pp. 1–6.

[28] J. R. Bates, D. Grimmett, G. Canepa, and A. Tesei, "Towards doppler estimation and false alarm rejection for continuous active sonar," *J. Acoustical Soc. America*, vol. 143, no. 3, pp. 1972–1972, 2018.

[29] D. P. Williams, "Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 2497–2502.

[30] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, Jan. 2016.

[31] E. L. Ferguson, R. Ramakrishnan, S. B. Williams, and C. T. Jin, "Deep learning approach to passive monitoring of the underwater acoustic environment," *J. Acoustical Soc. America*, vol. 140, no. 4, pp. 3351–3351, 2016.

[32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[33] J. I. Vestgården, K. T. Hjelmervik, D. H. S. Stender, and H. Berg, "Sonar scattering from the sea bottom near the norwegian coast," in *Proc. OCEANS Anchorage*, 2017, pp. 1–5.

[34] I. Seo, S. Kim, Y. Ryu, J. Park, and D. S. Han, "Underwater moving target classification using multilayer processing of active sonar system," *Appl. Sci.*, vol. 9, no. 21, 2019, Art. no. 4617.

[35] H. Berg and K. T. Hjelmervik, "Classification of anti-submarine warfare sonar targets using a deep neural network," in *Proc. OCEANS MTS/IEEE Charleston*, 2018, pp. 1–5.

[36] G. de Magistris *et al.* "Automatic object classification for low-frequency active sonar using convolutional neural networks," in *Proc. OCEANS MTS/IEEE SEATTLE*, 2019, pp. 1–6.

[37] T. S. Såstad and K. T. Hjelmervik, "Synthesizing realistic, high-resolution anti-submarine sonar data," in *Proc. OCEANS-MTS/IEEE Kobe Techno-Oceans*, 2018, pp. 1–5.

[38] J. Tremblay *et al.*, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 969–977.

[39] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2242–2251.

[40] P. Costa *et al.*, "End-to-end adversarial retinal image synthesis," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 781–791, Mar. 2018.

[41] Z. E. Ross, Y. Yue, M.-A. Meier, E. Hauksson, and T. H. Heaton, "PhaseLink: A deep learning approach to seismic phase association," *J. Geophys. Res.: Solid Earth*, vol. 124, no. 1, pp. 856–869, 2019.

[42] A. Bertrand and E. Josse, "Acoustic estimation of longline tuna abundance," *ICES J. Marine Sci.*, vol. 57, no. 4, pp. 919–926, 2000.

[43] M. Gasior and J. Gonzalez, "Improving FFT frequency measurement resolution by parabolic and Gaussian spectrum interpolation," in *Proc. AIP Conf. Proc.*, 2004, pp. 276–285.

[44] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[45] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9482–9491.

[46] M. B. Porter, "The bellhop manual and user's guide: Preliminary draft," Heat, Light, and Sound Research, Inc., La Jolla, CA, USA, Tech. Rep. 260, 2011.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.

[49] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.

[50] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.*, vol. 6, no. Dec., pp. 1889–1918, 2005.

[51] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. New York, NY, USA: Springer, 2001.

[52] G. Ginolhac and G. Jourdain, "Detection in presence of reverberation," in *Proc. OCEANS MTS/IEEE Conf. Exhib.*, 2000, pp. 1043–1046.

[53] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, Apr. 1967.

Alberto Testolin (Member, IEEE) received the MSc degree in computer science and the PhD degree in cognitive sciences from the University of Padova, Padua, Italy, in 2011 and 2015, respectively. He is currently an assistant professor at the Department of Information Engineering and Department of General Psychology, University of Padova. He is broadly interested in artificial intelligence, machine learning, and cognitive neuroscience. His main research interests include statistical learning theory, predictive coding, sensory perception, cognitive modeling and applications of deep learning to signal processing, networking, and optimization. He is also an active member of the IEEE Task Force on Deep Learning.

Dror Kipnis received the BSc degree in electrical engineering and the BA degree in physics from the Technion, Haifa, Israel, in 2003, and the MSc degree from the Department of Electrical Engineering - Physical Electronics, Tel-Aviv University, Tel Aviv, Israel, in 2007. He is currently working toward the PhD degree from the Department of Marine Technologies, University of Haifa, Haifa, Israel. His research interests include acoustic detection and classification of pelagic fish and dolphins.

Roee Diamant (Senior Member, IEEE) received the BSc and MSc degrees from the Technion, Israel Institute of Technology, Haifa, Israel, in 2002 and 2007, respectively, and the PhD degree from the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada, in 2013. From 2001 to 2009, he worked with Rafael Advanced Defense Systems, Israel, as a project manager and systems engineer, where he developed a commercial underwater modem with network capabilities. In 2015 and 2016, he was a visiting professor with the University of Padova, Italy. In 2009, he received the Israel Excellent Worker First Place Award from the Israeli Presidential Institute. In 2010, he received the NSERC Vanier Canada Graduate Scholarship. He has received three best paper awards, and serves as an associate editor of the *IEEE Ocean Engineering*. Currently, he is the coordinator of the EU H2020 project SYMBIOSIS (BG-14 track), and leads the underwater Acoustic and Navigation Laboratory (ANL), Department of Marine Technologies, University of Haifa. His research interests include underwater acoustic communication, underwater navigation, object identification, and classification.