2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA) | 979-8-3503-1104-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/I3DA57090.2023.10289193

Towards the evaluation of marine acoustic biodiversity through data-driven audio source separation

Michele Mancusi
GLADIA Research Group
Sapienza University of Rome
Rome, Italy
mancusi@di.uniroma1.it

Nicola Zonca
Native Prime
Bologna, Italy
nicola@nativeprime.com

Emanuele Rodolà
GLADIA Research Group
Sapienza University of Rome
Rome, Italy
rodola@di.uniroma1.it

Silvia Zuffi
IMATI-CNR
National Biodiversity Future Center, NBFC
silvia@mi.imati.cnr.it

***Abstract*—The marine ecosystem faces alarming changes, including biodiversity loss and the migration of tropical species to temperate regions. Monitoring underwater environments and their inhabitants is crucial, but challenging in vast and uncontrolled areas like oceans. Passive acoustics monitoring (PAM) has emerged as an effective method, using hydrophones to capture underwater sound. Soundscapes with rich sound spectra indicate high biodiversity, soniferous fish vocalizations can be detected to identify specific species. Our focus is on sound separation within underwater soundscapes, isolating fish vocalizations from background noise for accurate biodiversity assessment. To address the lack of suitable datasets, we collected fish vocalizations from online repositories and captured sea soundscapes at various locations. We propose an online generation of synthetic soundscapes to train two popular sound separation networks. Our study includes comprehensive evaluations on a synthetic test set, showing that these separation models can be effectively applied in our settings, yielding encouraging results. Qualitative results on real data showcase the model's generalization ability. Utilizing sound separation networks enables automatic extraction of fish vocalizations from PAM recordings, enhancing biodiversity monitoring and capturing animal sounds in their natural habitats.**

***Index Terms*—bioacoustics, soundscape ecology, deep learning source separation.**

I. Introduction

The oceans cover 71% of the Earth's surface and represent the natural habitat of numerous marine species. The biodiversity present in these environments is impressive, and tracking the activity and quantity of all existing species is essential for monitoring the whole marine ecosystem. In fact, today more than ever, the environmental issue is of crucial importance, and the oceans, like the whole planet Earth, are facing drastic and dramatic changes due to human activity, among which overfishing and ocean warming [1]. These changes, in addition to damaging the marine ecosystem, mainly affect the species that inhabit the sea: monitoring biodiversity is of vital importance, to understand the trend of the abundance of marine fauna, identify the most vulnerable areas, and take action to safeguard endangered species [2]. However, monitoring marine animals is challenging because many of the methods used on the Earth's surface for tracking, such as photos and videos, are often ineffective in the marine environment, due to the limited accessibility to many areas and poor light and visibility conditions. Furthermore, a significant amount of data that can be collected on physical quantities, such as temperature, salinity and pressure, may not reflect the biodiversity in a specific location. Therefore, there is a need for tools capable of overcoming these challenges and providing an accurate assessment of the marine habitat biodiversity. Underwater, instead of relying on optical signals, sound can be used to monitor biodiversity [3]. Indeed, the acoustic environment faithfully reflects the traits of the fauna present in a specific location and its behavior [4] [5]. One of the most popular and effective methods for monitoring marine biodiversity is passive acoustics monitoring (PAM), which employs hydrophones to capture underwater sound. Many aquatic organisms produce species-specific sounds, and mod-

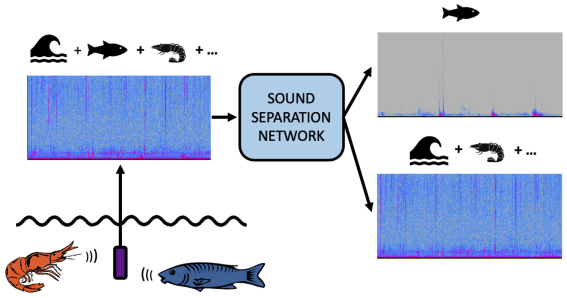

Fig. 1. We consider the problem of separating the sound produced by fishes from the background sound of the sea.

MM and ER are supported by the ERC grant no. 802554 (SPECGEO). SZ is funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.4 - Call for tender No. 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU; Award Number: Project code CN00000033, Concession Decree No. 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP B83C22002930006, project title: National Biodiversity Future Center - NBFC

979-8-3503-1104-4/23˜©˜2023 IEEE

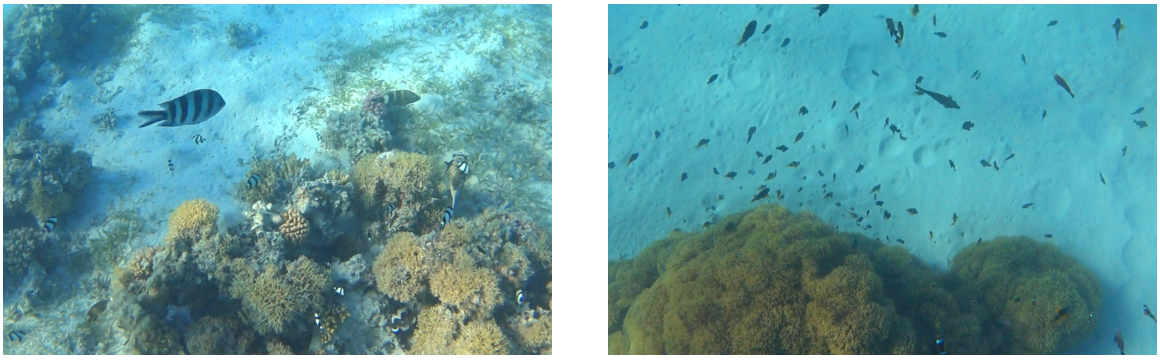

Fig. 2. Frames from the video captured at Marsa Alam showing the presence of soniferous fish.

ern technologies are becoming more and more convenient and precise, allowing for very accurate and careful data acquisition. Acoustic indices were initially used to assess biodiversity from PAM recordings [6] [5]. These indices are used to estimate richness, amplitude, heterogeneity and evenness of an acoustic environment. Some of them are: the acoustic entropy index (H), which indicates how much the amplitude of a signal is uniform in time and frequency; the acoustic complexity index (ACI) which takes into account the variation of a signal in different frequency bins over time and then averages over the entire frequency range [7]. While easy to apply, a drawback of acoustic indices is that they are not learned from data, and they are not, therefore, discriminative for animal sounds with respect to sounds with similar patterns, but of a different origin. Therefore, in order for these techniques to be applied, only the soundscape produced by natural sources should be considered. When the objective is to detect fish vocalizations, the PAM audio signal is usually visualized as a spectrogram and visually examined by an expert. This approach exploits the fact that fish vocalize within a relatively narrow range of low frequencies and often produce repetitive sounds. In this work, our aim is to present a solution for distinguishing the sound produced by fish from the background noise, with the goal of automating fish sound detection process and facilitating the use of soundscape analysis for biodiversity assessment. We employ recent advances in sound separation for human speech and music to the problem of separating fish vocalizations in PAM recordings. Machine learning techniques, and deep learning, in particular, require a large amount of data. In supervised learning, data must be annotated to provide ground truth information for training neural networks. Data annotation typically involves the manual identification of the attributes one wants to automatically recover. Obtaining annotated data for the task of sound separation, given a mixed signal, is clearly challenging. A preferable strategy is to generate training data by combining individual audio sources. This approach has been largely exploited for speech, music and anthropic sound. But while for human speech and music there is an abundance of data, this is not the case for fish vocalizations. At present, to the best of our knowledge, no datasets exist that include many examples of fish vocalization examples. Nonetheless, it is widely recognized that this is required for future progress [8]. Therefore, we collected a dataset of fish vocalizations from the internet. It is worth noting that the website https://fishsounds.net/ did not exist when we started this project. In most cases, we obtained a single sound example for each species. This limits the possibility of applying AI techniques to automatically classify the fish species from sound, as several recordings for each species would be necessary. In addition, with about 35000 known species of fish, the number of known soniferous species is quite limited, and while some sources have been identified, the majority of fish sounds remain unidentified [8]. Nevertheless, vocalizations from different fish species share similarities and posses distinctive characteristics that enable training a network that can separate fish-produced sounds from the background noise, typically consisting from the sound of waves and snapping shrimps (members of the *Alpheidae* family). We created a sound separation dataset by randomly overlapping fish vocalizations with sea backgrounds recorded at various locations on the Greek island of Nisyros. We use this dataset creation process to train two recent and popular architectures for sound separation: Conv-TasNet [9] (which we will call TasNet) and Demucs (version 2) [10]. We quantitatively evaluate the performance of these networks on two synthetic test sets, one obtained with backgrounds recorded in Nisyros, and one generated with backgrounds recorded in Favignana (Italy). We qualitatively show performance on a few examples of recordings performed in Marsa Alam, Egypt (Fig. 2). Our quantitative evaluation shows that the sound emitted from fishes can be successfully recovered from recordings with noisy background. The network performance visually slightly decreases in the in-the-wild setting, but still it is possible to recover the predominant fish sounds even if the networks have been trained with data captured with different devices in different locations. To our knowledge, this is the first work that applies modern sound separation techniques to PAM.

II. Related Work

The assessment of marine biodiversity through acoustic techniques is evolving rapidly and although several methods are used, at the moment none of these is considered the ideal tool for investigating the marine environment and its diversity. Some existing methods are discussed below.

A. Classical approaches

Spectrograms are valuable tools to analyze the temporal variation of an audio signal amplitude at different frequencies. They represent, through the Fourier transform, the 1D audio signal with a 2D image, with time and frequency as axes and pixel intensity as amplitude. By analyzing spectrograms, it is possible to identify the presence of some marine species through the identification of image patterns that are indicators of audio features in the species-specific vocalizations. For example, it is possible to observe whether the sound emitted is rhythmic or more smooth and harmonious. These patterns can be associated to known soniferous species or unidentified fish sources. While the examination of spectrograms as images enables the approximate detection of spectral patterns, it is insufficient for precise identification of the intricate modulation characteristics of underwater animal sounds. Automatic systems based on pattern recognition in images are often not sufficient for detecting fish vocalizations, and spectrograms needs to be inspected manually. However, studies focused on comprehending the drivers of marine biodiversity changes typically rely on prolonged audio recordings, spanning months or years. Conducting a manual analysis of such data is impractical. Unsupervised modeling techniques have been used to analyze spectrograms, with clustering being among the most widely adopted [11]. Assuming that the data has underlying patterns, clustering allows grouping elements with similar characteristics. Hence, large amounts of audio data can be modeled as a few audio clusters and these can be exploited to assess biodiversity by measuring per-cluster acoustic metrics. Unfortunately, this type of analysis can easily fail when non-biological sources contaminate the collected data [12]. In addition, it is still necessary to individually analyze the sources that are part of each cluster to understand the key elements that contribute to marine biodiversity.

B. Data-driven approaches

Several studies [13]–[22] have been carried out to trace, recognize, and isolate the biological sound sources present in nature. The use of machine learning has been fundamental to obtaining significant results. In particular, in [19], [20], the authors propose using deep learning to detect *odontocete echolocation* and bird sounds, respectively. In [22], Clink et al. introduce a workflow for the automated detection and classification of female gibbon calls, testing supervised and unsupervised approaches. In [21], Li et al. propose to use generative adversarial networks (GANs) to generate training data for learning to extract of toothed whales' whistles from time-frequency spectrograms. Recently, Sun et al. [**?**] have introduced a toolbox for soundscape information retrieval based on non-negative matrix factorization.

C. Source separation

The work that has led to significant progress in this field is mainly in music and speech. In these contexts, the separation task is particularly challenging due to the inherent complexity of overlapping harmonics, temporal and spectral variability, and unpredictable background noise. Deep learning has brought significant advances to source separation by leveraging the ability of neural networks to model complex, non-linear relationships and learn high-level abstract features from data. This paradigm has provided a robust, data-driven approach to the source separation problem, outperforming traditional signal processing methods. One of the seminal works in speech separation is [9], where they propose an end-to-end, fully-convolutional time-domain audio separation network that significantly outperformed traditional frequency-domain methods. While, for music source separation, in [10], a model is proposed that relies on depthwise separable convolutions and bidirectional LSTMs (Long Short-Time Memory), leading to improved performance over previous state-of-the-art methods. Further advances for music source separation have been made in [23], where a novel Bayesian method for unsupervised source separation is introduced.

III. Method

The problem of separating audio sources consists of breaking down a mixture of signals $y(t) \in \mathbb{R}^T$ into its n components $c_1(t), \ldots, c_n(t) \in \mathbb{R}^T$, where,

$$y(t) = \sum_{i=1}^{n} c_i(t). \tag{1}$$

The mixture is represented as a vector in the waveform domain. In our case, we consider $n = 2$ sources: fish and background. In order to perform sound separation, we employ the two aforementioned networks, TasNet and Demucs. These two types of networks are trained in a supervised manner, and while both have an encoder-decoder structure and act directly on the audio waveform, they are fundamentally different: TasNet learns a mask to be applied to the mixture to filter the desired source signal, whereas Demucs learns to directly synthesize the required signals without using any filtering. We train both network with supervised training, on the same dataset. Critical for the success of the separation networks is the availability of a large training dataset with overlapped and separated sources. The availability of such dataset for the specific case of fish vocalizations poses several challenges. Here, we contribute with a novel synthetic dataset that we define as follows. We collected a large set of recorded vocalizations from online sources, these will be the basis of the foreground fish source. At the same time, we recorded a set of diverse sea recordings that constitute the data to represent background sound. Details on the collected data are reported in the Experiments section. During training, at each epoch we create random combinations, with randomized amplitude, of fish and background audio data. In this way, despite the limited number of sound sources, in particular for the fish data, we prevent the networks from overfitting on a fixed training dataset. Audio data is loaded from the network as a set of audio chunks of length 44160, obtained by splitting the audio data with an overlap fraction of 0.25. The foreground fish vocalization dataset is also loaded as a set of samples with 0.25 overlap, where each sample is a chunk of size 44160.

The synthetic data for training is created as follows. At each epoch, for each foreground sample i, we define the two audio sources s_0 (foreground) and s_1 (background) as follows:

$$s_0 = k_f \alpha_f x_f$$
$$s_1 = (1 + k_b) x_b,$$

where x_f is the sample with index i, and x_b is a random background chunk; k_f and k_b are two random coefficients sampled from a uniform distribution, while α_f is a fixed attenuation factor for the fish audio, required to model relative amplitude in real conditions. In this way, at each epoch, every fish sample is combined differently with a random background. We set $\alpha_f = 0.1$.

A. TasNet

TasNet is a convolutional audio separation model in the time domain, composed by an encoder, a separation module and a decoder, as shown in figure 3 (A). The encoder transforms small overlapping fragments of the mixture into feature vectors in an intermediate latent space. Using this representation, the separation module calculates a mask for each source. Each mask, multiplied by the respective intermediate representation of the mixture, generates the latent features of the relative source. Finally, the decoder converts each latent representation into a time-domain waveform, thus obtaining the desired separated signals. In figure 3 (B) we report the entire system flowchart from [9].

1) Encoder: Initially, the input mixture is divided into N overlapping parts $\mathbf{x}_i \in \mathbb{R}^L$, where $i = 1, \ldots, N$, each of length L. Each $\mathbf{x}_i$ is transformed by the encoder into the corresponding vector in the latent domain $\mathbf{z}_i \in \mathbb{R}^M$ through a 1D convolution operation (formally expressed by a matrix multiplication) followed by a ReLU activation function $\mathcal{G}(\cdot)$:

$$\boldsymbol{z}_i = \mathcal{G}(\boldsymbol{x}_i \boldsymbol{S}), \tag{2}$$

where $\boldsymbol{S}$ is a $L \times M$ matrix of convolution coefficients.

2) Separation module: The actual separation of each fragment of the mixture occurs in the separation module, in which n mask vectors $\mathbf{m}_i \in \mathbb{R}^M$ are estimated, where $i = 1, \ldots, n$ and n is the number of signals to be separated. Each of these vectors, being masks, must necessarily be $\mathbf{m}_i \in [0, 1]$. The vector representation in the latent space $\mathbf{b}_i \in \mathbb{R}^M$ of each signal is calculated by multiplying the relative mask $\mathbf{m}_i$ by the mixture $\mathbf{z}_i$,

$$\mathbf{b}_i = \mathbf{z}_i \odot \mathbf{m}_i \tag{3}$$

where $\odot$ denotes element-wise multiplication. This module is a temporal convolutional network (TCN) [24], which is fully convolutional and consists of stacked 1D dilated convolutional blocks with increasing dilation factors. These factors make it possible to gradually capture increasingly broad contexts, thus exploiting long-range dependencies within the signal. Here, with respect to the architecture described in [9], we do not make use of the skip connections in the 1D convolutional blocks.

3) Decoder: The reconstruction of each source is computed by the decoder. The latter takes as input $\mathbf{z}_i$ and returns a vector $\hat{\mathbf{x}}_i$ in the waveform domain by applying a 1D transposed convolution operation,

$$\hat{\mathbf{x}}_i = \mathbf{z}_i \mathbf{T} \tag{4}$$

where $\hat{\mathbf{x}}_i \in \mathbb{R}^L$ is the reconstruction of $\mathbf{x}_i$ and $\mathbf{T}$ is a $M \times L$ matrix of convolution weights.

B. Demucs

Demucs is an autoencoder model made of a convolutional encoder and a convolutional decoder linked with skip U-Net connections and a 2-layers bidirectional LSTM. The size of the latent space is $C_B = 6$.

1) Encoder: As illustrated in figure 4, the encoder consists of $B = 6$ stacked convolutional layers, and the number of output channels C_i in each layer equals the number of input channels C_{i+1} in the next layer. From the second layer onwards, the output channels are twice the number of input channels. All these stacked layers have the task of compressing the information in order to obtain a compact representation of the training data. The input channels in the first layer are $C_0 = 2$ and the output channels are $C_0 = 100$. The output channels in the last layer are $C_B = 3200$, which is the hidden size of the LSTM.

2) Decoder: Since LSTM outputs a tensor with $2C_B$ channels, a linear layer is needed to reduce the number of channels to C_B. The decoder is built essentially like the encoder, but with the convolutional layers put in reverse order and transposed convolutions instead of the regular convolutions. The decoder has the task of expanding the dimensions of the compressed vectors in the latent space to regain vectors with sizes equal to those of the input space. The last layer returns tensors with $N \cdot C_0$ channels, synthesizing the N sources present, initially, in the input mixture.

3) U-network: In this architecture, the encoder layers are connected to the decoder layers with the same index through skip connections, as happens in the Wave-U-Net [25]. The objective of these connections is to connect the various decoder layers with those of the encoders to transfer information directly from ones to the others in such a way as to facilitate reconstruction. Compared to Wave-U-Net, Demucs skip connections use transposed convolutions instead of linear interpolations, since they require less memory and computational time.

IV. Experiments

A. Fish Vocalization Data

We collected 191 audio files corresponding to the vocalization of 143 different species. Most of the recordings were downloaded from FishBase[1]. The collected data often exhibit unnatural noise, since in many cases the recordings are performed in fish tanks. In order to create a dataset that can be used to synthesize realistic audio data, we preprocessed for

[1] www.fishbase.org

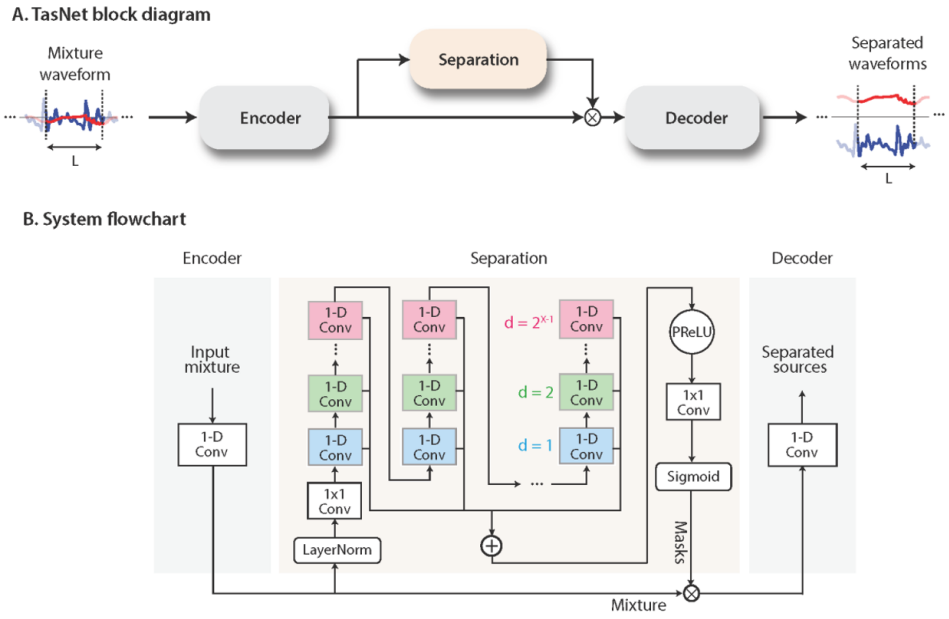

Fig. 3. (A): TasNet block diagram. A piece of the input signal is projected into a multidimensional hidden space through the encoder. Then, a separation module calculates an estimated mask for each individual source. Ultimately, a decoder converts these masked encoded features back into waveform domain signals. (B): System flowchart. The encoder consists of a 1D convolutional module that maps the mixture into the features space. A temporal convolutional network (TCN) calculates the mask vectors, and the decoder reconstructs the separated signals by a 1D transposed convolution operation. In the separation module, different dilation factors in each 1D Conv block are highlighted with different colors. This figure is taken from [9].

noise removal, and normalized for a peak amplitude of -1 dB. For this purpose, we used the open source software Audacity. Figure 5 illustrates examples before and after preprocessing. We employ the fish vocalizations for the online creation of training samples by combining them with recorded sea backgrounds ad described previously, and for creating a synthetic testset with ground-truth separated signals.

B. Sea Recordings

We performed sea recordings at the Greek island of Nisyros and at the Italian island of Favignana. Greek recordings were performed in October 2019, April 2021, August 2021, and October 2021 at different sites around the island, both near the coast and in the open sea; whether the Italian ones in June 2023. Data were captured with an Aquarian Scientific AS-1 hydrophone (linear range 1Hz to 100kHz ±2dB, operating depth 200 mt). Sea recordings in Nisyros are used as backgrounds for training and test. In addition, we collected a sound video dataset at Marsa Alam (Egypt) using an action camera Sony HDR-AS50 Full HD. The audio channel from this data is used for a qualitative evaluation.

C. Networks Training

During training, we considered different 11 sea recordings for creating backgrounds, captured in different locations and different times, and of various duration. The first 5 files were captured with sample rate of $192K$ and were converted to $44K$. Recordings with length greater than 3 minutes were divided in multiple files of smaller duration ($1-2$ minutes) and constitute a dataset of background chunks. We have a total of 133 files for background. A couple of recordings that we use for representing backgrounds were manually filtered for removing fish sounds. This was not necessary for most of the recordings, where fish sounds were harder to find. We use the 80% of the fish vocalization data for training. We fed both the networks with samples of size 44160 and trained both the networks with a learning rate of 0.0001 and a number of epochs equal to 200. TasNet employs an autoencoder with 512 filters (N), each of length 256 (L). The bottleneck has 256 channels (B), and the convolutional blocks contain 512 channels (H). For convolutional operations, we use a kernel size of 3 (P) across 8 blocks (X) that are repeated 4 times (R). The network is designed to separate inputs into 2 distinct "speakers" (C).

D. Evaluation

For the quantitative evaluation, we generated a set of synthetic inputs using the 20% fish vocalizations that were not used for training, combining these sounds at random with background chunks also not used for training. For the qualitative evaluation of the data recorded in Egypt, we trained the network using the whole vocalization dataset. We apply the trained TasNet and Demucs networks to the synthetic testset and quantify the sound separation performance, computing an SDR (Source to Distortion Ratio) score [27], is considered an excellent metric to assess sound quality, between recovered and ground truth fish and background audio sources. In order to compute the SDR score, the reconstruction $\hat{s}_i$ of a source s_{target} is assumed of consisting of four components:

$$\hat{s}_i = s_{target} + e_{interf} + e_{artif} + e_{noise}$$

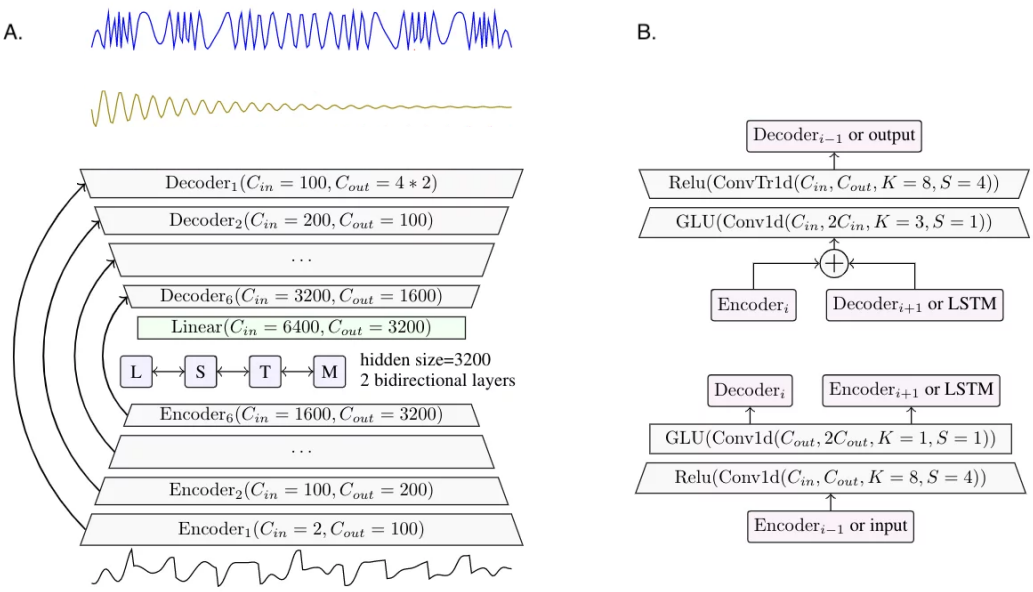

Fig. 4. (A): Demucs model with the input mixture and the two output sources, all in the waveform domain. (B): Encoder/decoder block architecture. In each encoder block, there is a convolution with kernel size $K = 8$ (to have dependencies with adjacent time steps) and stride $S = 4$ followed by a ReLU activation function. The result is given as input to another convolution with kernel size $K = 1$ and stride $S = 1$, in order to increase the expressivity of the network with little additional computation. In the end, a gated linear unit (GLU) activation function [26] is applied. The decoder block is constructed in reverse order with respect to the encoder, and it consists of a convolution with kernel size $K = 3$ and stride $S = 1$, followed by a GLU and then a transposed convolution with kernel size $K = 8$ and stride $S = 4$, followed by a ReLU. This figure is taken from [10].

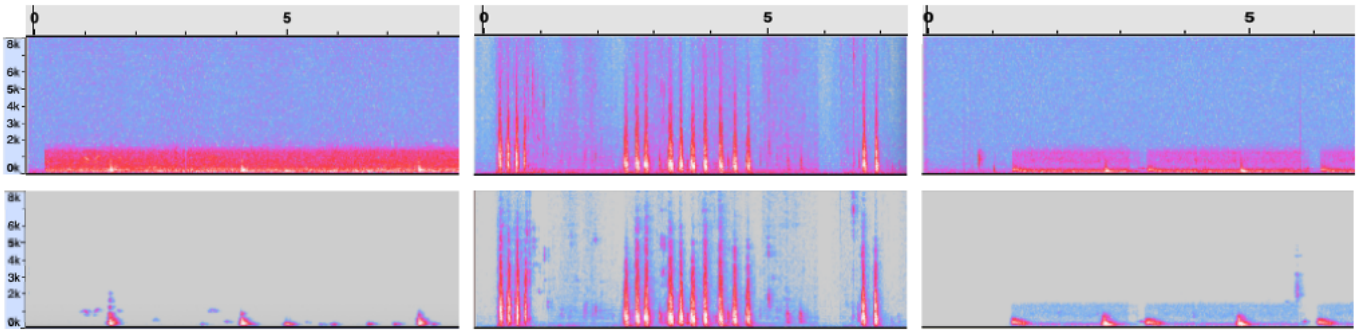

Fig. 5. Fish vocalization before (top) and after (bottom) noise removal and normalization. Time in seconds.

where e_{interf}, e_{artif} and e_{noise} are respectively error terms for interference, artifacts and noise [27]. Using these terms, the SDR is expressed as:

$$\text{SDR} := 10 \log_{10} \left(\frac{\|s_{target}\|^2}{\|e_{interf} + e_{artif} + e_{noise}\|^2} \right).$$

Table I reports our results (the higher, the better).

We note from Table I that the Tasnet network performs significantly better than Demucs. The former reaches an SDR score equal to 10.59 on the separation of the sound of the fish and 17.60 on the background, while the latter obtains just 2.65 of SDR on the background and even a negative score on fish, equal to -5.96 of SDR. Figure 6 and 7 show two randomly selected examples of separation. Although the separations produced by Demucs appear to be perceptibly better, it can be seen how they show artifacts; in particular, vertical lines are introduced that are repeated periodically, while in Tasnet, this behavior is not present. Furthermore, it is possible to notice how, on the synthetic data, in correspondence with the sounds of the fish, Demucs generates fictitious frequencies that are

TABLE I
QUANTITATIVE EVALUATION ON THE SYNTHETIC TESTSET USING NISYROS BACKGROUNDS.

Metric	**TasNet**	
	Channel	*Value*
SDR	Fish	**10.60 ± 9.00**
SDR	Background	**17.60 ± 7.04**
Metric	**Demucs**	
	Channel	*Value*
SDR	Fish	-3.71 ± 2.03
SDR	Background	2.65 ± 4.05

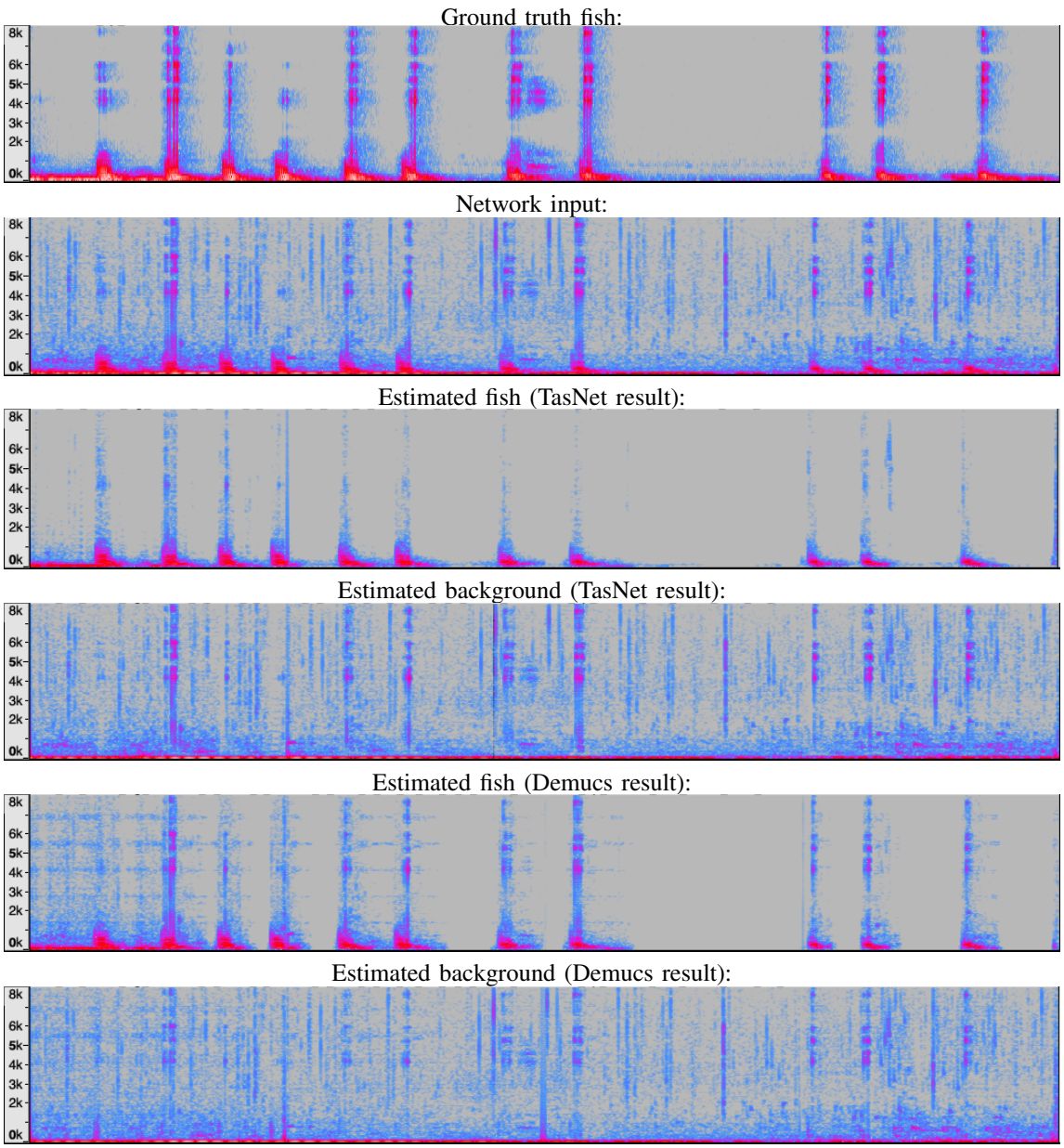

Fig. 6. Synthetic testset example. From top: fish vocalization (*Prionotus*) (4 seconds); overlap with sea background; TasNet fish and background separation; Demucs fish and background separation. Vertical axis is frequency, horizontal axis is time.

TABLE II
QUANTITATIVE EVALUATION ON THE SYNTHETIC TESTSET USING FAVIGNANA BACKGROUNDS.

Metric	**TasNet**	
	Channel	*Value*
SDR	Fish	**8.11 ± 15.48**
SDR	Background	**6.27 ± 5.14**
Metric	**Demucs**	
	Channel	*Value*
SDR	Fish	-5.27 ± 7.92
SDR	Background	-2.81 ± 2.14

not present in the TasNet separations. This is probably due to the fact that Demucs is a network that does not separate the signal by filtering it, but by directly synthesizing the requested source, not performing well with the data of our dataset. Instead, Tasnet, a more classical network that filters the desired signal from the mixture, appears to be more robust and performs better with the data in our possession. Results in Table II, obtained with a set of backgrounds captured at different locations in relation to the data used for training, further confirm the above discussion. Figure 8 and 9 show two examples of separation applied to the data recorded in Marsa Alam. Note that the performance of the networks are here qualitatively lower than on the synthetic dataset, and

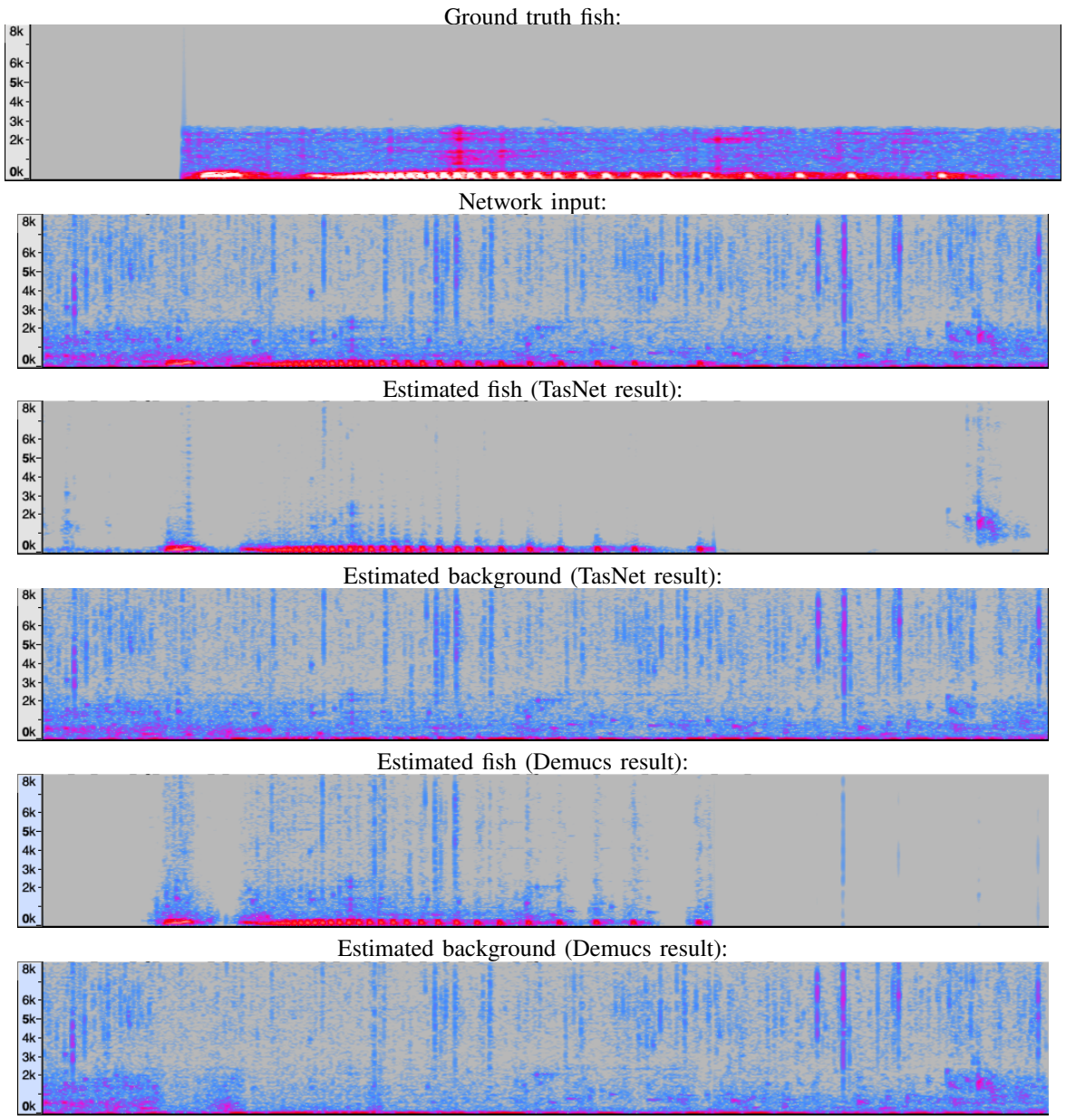

Fig. 7. Synthetic testset example. From top: fish vocalization (*Epinephelus guttatus*) (2 seconds); overlap with sea background; TasNet fish and background separation; Demucs fish and background separation. Vertical axis is frequency, horizontal axis is time.

this can be due in particular to the distribution shift between the background data: while in the Aegean Sea we noticed a consistent presence of clicks sounds emitted by shrimps, these are not present in the Marsa Alam dataset. Moreover, the latter data includes a significant sensor noise. Despite these differences, both networks are able to identify sounds that we can attribute to the fish species observed in the video channel of the captured data.

V. Conclusion

With this study, we demonstrate the effective application of deep learning techniques for source separation of marine data. We achieve this by applying the most effective source separation architectures to the problem of isolating fish vocalizations from sea background, obtaining competitive signal-to-distortion ratio (SDR) scores on a synthetic test set generated composing real animal and background sources. Notably, as observed by experts, our trained networks also perform qualitatively well on in-the-wild data, captured with a different device in a different environment. We attribute this generalization ability to our online training strategy, where a new synthetic training set is generated at each epoch. We hope these results will pave the way for new methods of studying the marine environment and contribute to developing new automatic PAM techniques for monitoring marine biodiversity and, possibly, accurately tracking fauna in the oceans.

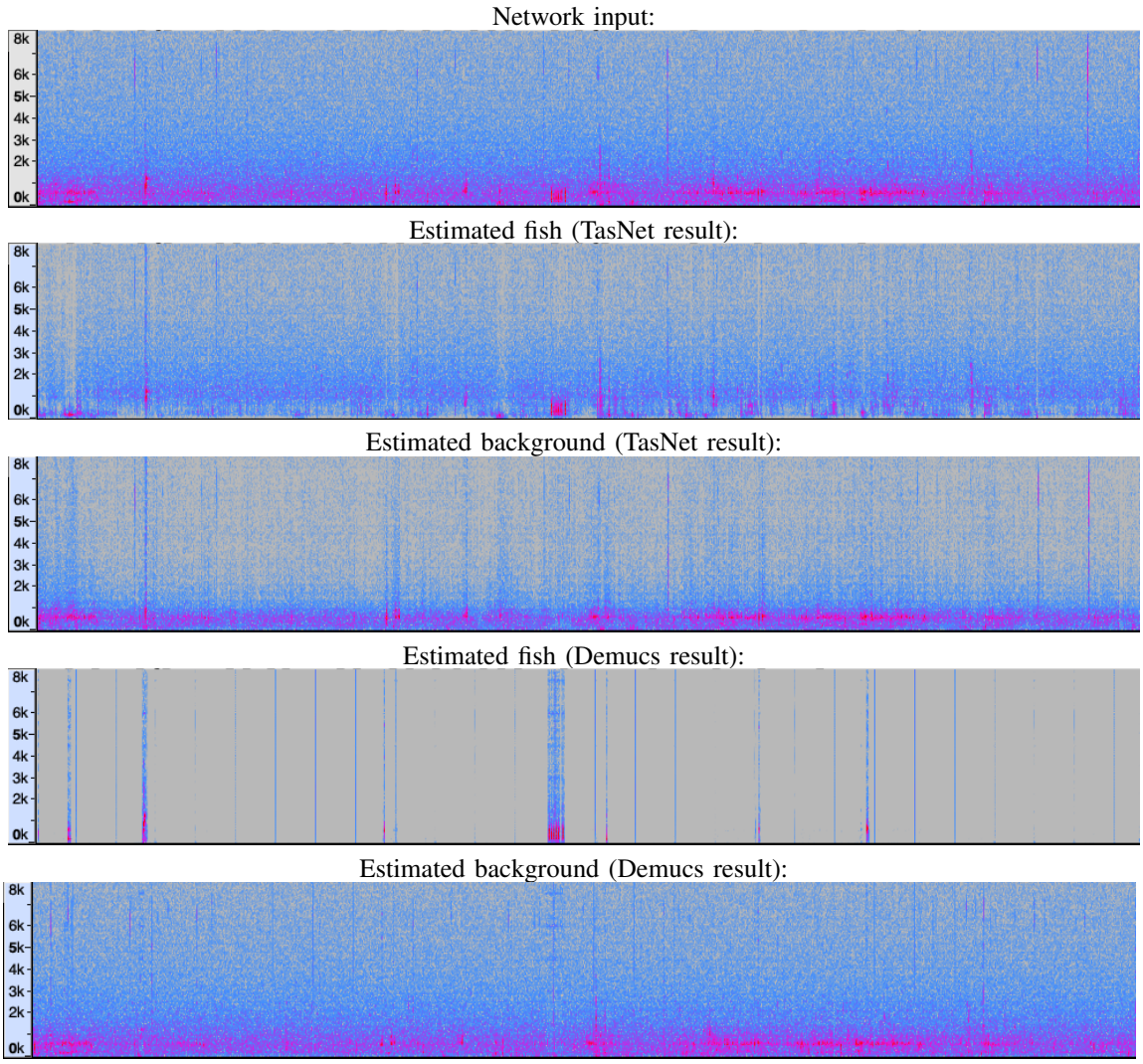

Fig. 8. In-the-wild experiment at Marsa Alam (20 seconds). From top: sea recording; TasNet fish and background separation; Demucs fish and background separation. Vertical axis is frequency, horizontal axis is time.

REFERENCES

[1] H. C. et al., "Emerging marine diseases—climate links and anthropogenic factors," *Science*, p. 1505–1510, 1999.

[2] B. P, P. AB, B. N, H. JS, N. T, R. D, and S. B., "Qantifying the evidence for biodiversity effects on ecosystem functioning and services." *Ecol. Lett.*, p. 1146–1156, 2006.

[3] P. BC, V.-R. LJ, D. SL, F. A, K. BL, N. BM, G. SH, and Pieretti, "Soundscape ecology: the science of sound in the landscape." *Bioscience*, p. 203–216, 2006.

[4] S. M, "The sounds: our sonic environment and the tuning of the world." *Behav.*, p. 49–70, 1969.

[5] T. A. Mooney, L. Di Iorio, M. Lammers, T.-H. Lin, S. L. Nedelec, M. Parsons, C. Radford, E. Urban, and J. Stanley, "Listening forward: approaching marine biodiversity assessments using acoustic methods," *Royal Society Open Science*, vol. 7, no. 8, p. 201287, 2020. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rsos.201287

[6] J. Sueur, S. Pavoine, O. Hamerlynck, and S. Duvail, "Rapid acoustic survey for biodiversity appraisal," *PloS one*, vol. 3, no. 12, 2008.

[7] N. Pieretti, A. Farina, and D. Morri, "A new methodology to infer the singing activity of an avian community: The acoustic complexity index (aci)," *Ecological Indicators*, vol. 11, no. 3, pp. 868–873, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X10002037

[8] M. J. G. Parsons, T.-H. Lin, T. A. Mooney, C. Erbe, F. Juanes, M. Lammers, S. Li, S. Linke, A. Looby, S. L. Nedelec, I. Van Opzeeland, C. Radford, A. N. Rice, L. Sayigh, J. Stanley, E. Urban, and L. Di Iorio, "Sounding the call for a global library of underwater biological sounds," *Frontiers in Ecology and Evolution*, vol. 10, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fevo.2022.810156

[9] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint 1911.13254*, 2019.

[11] C. van Rijsbergen, *Information retrieval*. London, UK: Butterworth, 1979.

[12] T.-H. Lin, Y. Tsao, and T. Akamatsu, "Comparison of passive acoustic soniferous fish monitoring with supervised and unsupervised approaches," *The Journal of the Acoustical Society of America*, vol. 143, no. 4, pp. EL278–EL284, 2018.

[13] S. Innami and H. Kasai, "Nmf-based environmental sound source separation using time-variant gain features," *Computers & Mathematics with Applications*, vol. 64, no. 5, pp. 1333–1342, 2012.

[14] T.-H. Lin and Y. Tsao, "Listening to the deep: Exploring marine

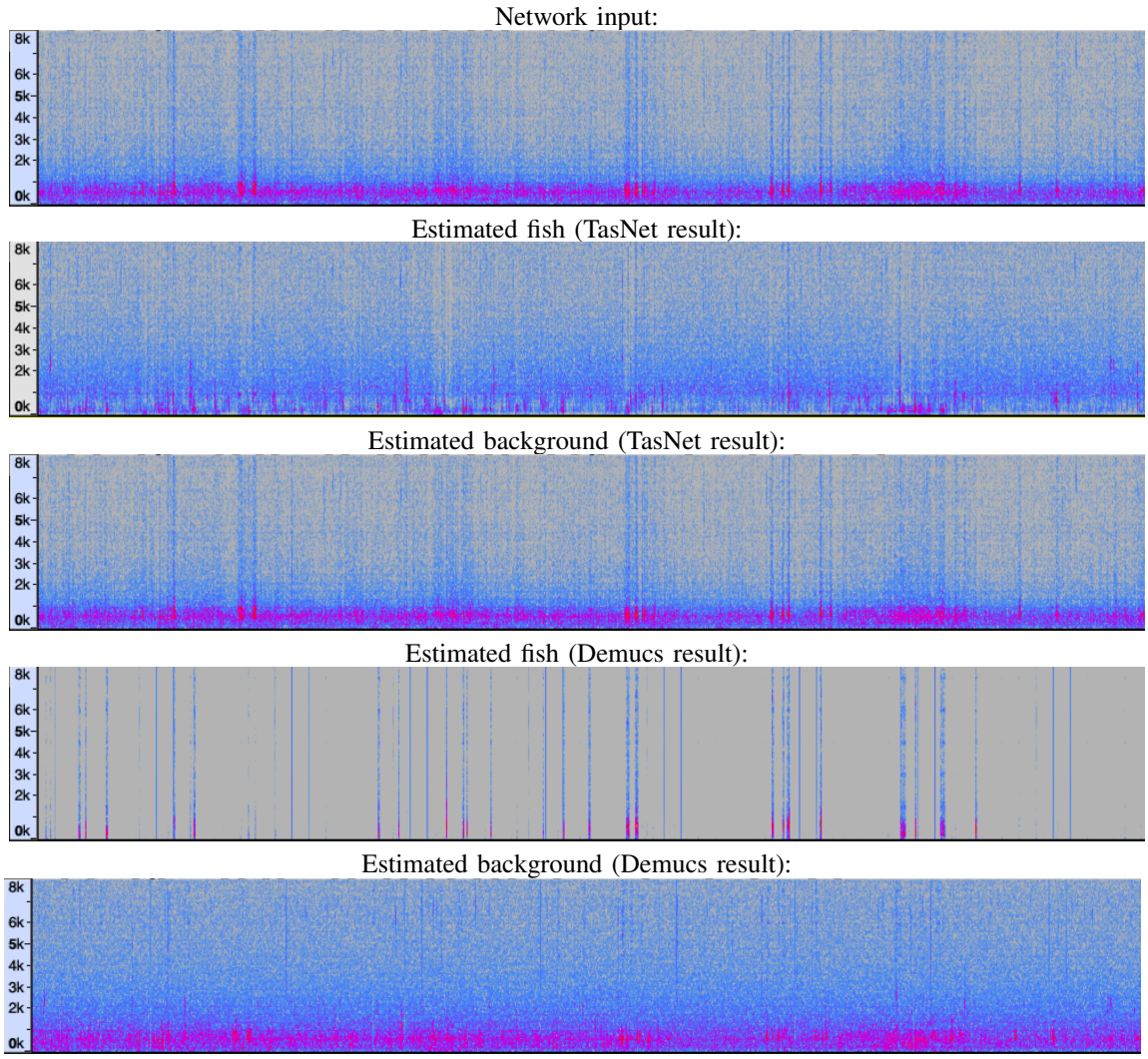

Fig. 9. In-the-wild experiment at Marsa Alam (60 seconds). From top: sea recording; TasNet fish and background separation; Demucs fish and background separation. Vertical axis is frequency, horizontal axis is time.

soundscape variability by information retrieval techniques," in *2018 OCEANS-Mts/IEEE Kobe Techno-Oceans (OTO)*. IEEE, 2018, pp. 1–6.

[15] D. Gillespie, D. K. Mellinger, J. Gordon, D. McLaren, P. Redmond, R. McHugh, P. Trinder, X.-Y. Deng, and A. Thode, "Pamguard: Semi-automated, open source software for real-time acoustic detection and localization of cetaceans." *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2547–2547, 2009.

[16] Z. Zhang and P. R. White, "A blind source separation approach for humpback whale song separation," *The Journal of the Acoustical Society of America*, vol. 141, no. 4, pp. 2705–2714, 2017.

[17] J. Xie, M. Towsey, J. Zhang, and P. Roe, "Adaptive frequency scaled wavelet packet decomposition for frog call classification," *Ecological Informatics*, vol. 32, pp. 134–144, 2016.

[18] J.-j. Jiang, L.-r. Bu, F.-j. Duan, X.-q. Wang, W. Liu, Z.-b. Sun, and C.-y. Li, "Whistle detection and classification for whales based on convolutional neural networks," *Applied Acoustics*, vol. 150, pp. 169–178, 2019.

[19] W. Luo, W. Yang, and Y. Zhang, "Convolutional neural network for detecting odontocete echolocation clicks," *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. EL7–EL12, 2019.

[20] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[21] P. Li, M. A. Roch, H. Klinck, E. Fleishman, D. Gillespie, E.-M. Nosal, Y. Shiu, and X. Liu, "Learning stage-wise gans for whistle extraction in time-frequency spectrograms," *IEEE Transactions on Multimedia*, 2023.

[22] D. J. Clink, I. Kier, A. H. Ahmad, and H. Klinck, "A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings," *Frontiers in Ecology and Evolution*, vol. 11, p. 28, 2023.

[23] E. Postolache, G. Mariani, M. Mancusi, A. Santilli, L. Cosmo, and E. Rodolà, "Latent autoregressive source separation," in *Proc. AAAI*, ser. AAAI Press, 2023.

[24] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[25] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *ISMIR*, 2017.

[26] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.