ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446338

BIOMIMETIC MAPPINGS FOR ACTIVE SONAR OBJECT RECOGNITION IN CLUTTER

Sangwook Park[1], *Angeles Salles*[2], *Kathryne Allen*[3], *Cynthia Moss*[3], *Mounya Elhilali*[4]

[1]Department of Electronic Engineering, Gangneung-Wonju National University
[2]Department of Biological Sciences, University of Illinois Chicago
[3]Department of Psychological and Brain Science, Johns Hopkins University
[4]Department of Electrical and Computer Engineering, Johns Hopkins University

ABSTRACT

SONAR technology plays a pivotal role in terrain exploration and specifically identification of objects of interest. However, it grapples with a recurring challenge of clutter and noise which limits the performance of target recognition models. The challenge of noisy observations renders the choice of robust signal representations critical. Inspired by mammalian representation in the midbrain of echolocating bats, the present study evaluates the robustness of a decomposition of echo measurements that matches the statistics of natural vocalizations. This representation is contrasted with equally rich generic mappings as well as digital sonar images based on time-frequency representations. The study shows the clear advantage of the naturally optimized representation for object recognition in presence of background noise and clutter, and further underscores the potential of bio-inspired approaches in advancing SONAR technology.

Index Terms— artificial midbrain, echo representation, active sonar, target identification

1. INTRODUCTION

SONAR is a critical technology for exploring and understanding the underwater world; though its applications spread to a wide range of fields including communication systems, navigation technologies, autonomous vehicles, and robotics [1, 2, 3, 4]. Unlike traditional visual or optical recognition systems, which rely on light to identify objects, SONAR utilizes sound waves to achieve the same goal, making it a valuable technology in settings with reduced visibility or darkness; and also, where specific sound profiles can travel or penetrate further. By emitting sound pulses into the environment and analyzing the returning echoes, SONAR systems can discern and recognize objects in their vicinity. In complex environments, object recognition using SONAR presents several challenges that can complicate the accurate identification and classification of objects. These challenges include: i) varied object shapes and materials making it difficult to create a one-size-fit-all recognition algorithm [5]; ii) background noise and clutter often distort or mask echoes from target objects, making them harder to detect and recognize [6]; iii) variations in orientation which significantly impact how pulse are reflected by targets hence complicating recognition [7]; iv) limited data due to challenges of collecting comprehensive and diverse SONAR data for training large recognition systems, particularly for specific objects or environments; v) ambiguity and environmental factors which stem from unique challenges in different settings (e.g., open water, coastal areas) and environmental compositions (e.g. dense or cluttered areas).

In contrast, animals that rely on active sensing are able to negotiate complex environments, navigate during motion or flight, hunt for prey and avoid collisions with obstacles and other animals; all using lightweight, low-power biological sonar strategies. Echolocating bats, for example, transmit sonar signals and process auditory information carried by returning echoes to guide behavioral decisions [8]. The present work explores benefits of sonar signal representations inspired by biological principles for the goal of robust target recognition. The approach hones in specifically on the representation in the Inferior Colliculus (IC), a midbrain structure that serves as a crucial nexus for auditory information processing, where incoming sensory signals from the ears are integrated with feedback from the auditory cortex, thus playing a pivotal role in auditory perception. One of the hallmarks of signal processing in the bats' inferior colliculus is that neural filtering appears to match the statistical properties of bat vocalizations [9]. In earlier work, we developed a bio-inspired autoencoder model constrained to match natural statistics of animal calls [10]. This network, referred to as *Biomimetic Network* (BioNet), closely emulates the auditory characteristics of IC neurons in echolocating bats. In the present study, we examine the benefits of this naturally optimized representation for robust object recognition in presence of background noise and clutter.

To contrast, we also examine the robustness of other biomimetic representations that are equally rich but not explicitly optimized to natural environments. We specifically explore the representation in the mammalian auditory cortex whereby incoming sound signals undergo extensive decom-

This work was supported by ONR N00014-23-1-2050 and N00014-23-1-2086

positions along time and frequency, hence representing the incoming sound signal with varying resolutions [11]. This decomposition is akin to convolutions commonly used in deep neural networks whereby a signal or image is mapped to an embedding space by filtering incoming information along different axes and resolutions [12]. In lieu of a data-driven optimized representation, this cortical representation uses a wavelet decomposition with Gabor-like filters that closely emulate neural characteristics in the mammalian auditory cortex [13, 14]. This representation, referred to as *cortical wavelet analysis* (CorWav) results in a rich feature space that represents spectrotemporal dynamics of the sonar signal. Finally, we also examine the informative content of a more generic sonar image using time-frequency mappings. In order to align with the complexity afforded by the BioNet and CorWav representations, we adopt a biomimetic spectrographic representation whereby the signal undergoes a number of linear and nonlinear transformations including basilar membrane filtering, hair-cell transduction, and lateral inhibition mimicking sharpening of spectral information observed in the cochlear nucleus [15]. This approach results in a time frequency image referred to as *Auditory Spectrogram* (Spec).

Building upon these bio-inspired methodologies, this study examines the robustness of these different mappings for sonar object recognition. This task is explored using various physical shapes across different orientations, by recording sonar signals in controlled uncluttered and noisy backgrounds. Clutter is introduced by adding foliage to the environment at different distances from the target object. These recordings are then evaluated through different analysis methods to assess the recognition of target shapes using Fisher distance as benchmark. Given the limitations in training data, a direct discriminability measure is favored to a deep-learning classifier in order to directly evaluate the behavior of different sonar representations in term of graceful degradation in noise clutter. In the rest of the paper, we provide detailed descriptions of the three representations considered in this paper (BioNet, CorWav and Spec). Section 3 describes the experimental setup, data collection and evaluation methods, and section 4 presents the final evaluation and analysis of different signal representations. The conclusions and discussion are presented in the last section.

2. SONAR REPRESENTATIONS

Typical bat vocalizations span up to 150KHz, all signals are analyzed using a sampling rate of 300KHz.

2.1. Auditory Spectrogram (Spec)

A biomimetic time-frequency spectrogram is obtained by adapting the approach originally proposed by Chi et al. [16] to the ultrasonic range. The signal waveform is first analyzed through a cochlear filterbank composed of 128-channels with overlapping constant-Q ($Q_{10dB} \approx 3$) bandpass filters, whose center frequencies are uniformly placed in logarithmic scale. The original filters are shifted in the current work to extend to 150KHz. The next stage, mimicking limited peripheral phase-locking and nonlinear transduction is achieved using a sigmoid nonlinear compression ($\alpha = 0.1$) and first-order recursive low-pass filter ($b = 1, a = [1, -0.97]$). Further frequency sharpening is introduced by performing a differential between adjacent frequency channels then half-wave rectification. Finally, temporal integration is performed using a windowing operation. The implementation of the original paper is available in the NSL toolbox [17]. In the current study, we use the following settings: 0.2 ms frame length without overlap, 24-channels per octave (i.e., 128 channels over 5.33octaves). This analysis maps each echo signal into 128xT feature space S, where T is the number of frames.

2.2. Cortical Wavelet Analysis (CorWav)

Mammalian cortical processing is modeled using a wavelet-like decomposition which builds on the nonlinear decompositions of the auditory spectrogram S and a set of complex-valued filters $\Gamma_{\Omega,\omega}$ to perform a wavelet decomposition [11]:

$$C_{\Omega,\omega}[f] = \sum_{t,\delta,\tau} S[\delta,\tau]\, \Gamma_{\Omega,\omega}[f-\delta, t-\tau] \qquad (1)$$

where t, f, Ω and ω represent time, frequency, spectral modulations, and temporal modulations, respectively. C is the cortical response obtained from this mapping driven by an input spectrogram S. The wavelet decomposition Γ is a set of 2D complex-valued Gabor functions that decompose the signal along multiple resolutions:

$$\begin{aligned} &\Gamma_{\Omega,\omega}[f,t] = W_{\Omega,\omega}[f,t]\; e^{j2\pi(\frac{\Omega}{N_f}f + \frac{\omega}{N_t}t)} \\ &\text{where } W_{\Omega,\omega}[f,t] = \tfrac{1}{2\pi\sigma_f\sigma_t} e^{-\frac{1}{2}\frac{[f-f_0]^2)}{\sigma_f^2} + \frac{[t-t_0]^2}{\sigma_t^2}} \end{aligned} \qquad (2)$$

where N_f and N_t are the number of bins for frequency and time in spectrotemporal representation, respectively. f_0 and σ_f is the center of frequency axis and the corresponding variance, respectively. Similarly, t_0 and σ_t are for the time axis. By parameterizing the filters Γ along spectral and temporal modulations (Ω and ω), the analysis results in a multi-scale complex-valued mapping that reflects the magnitude and phase of the modulation space. We set spectral modulations from 0 to 3 cycle/octave with 0.3 cycle/octave step and temporal modulations from -384 to 384 Hz with 16 Hz step.

From eq. (1), the cortical analysis returns a tensor representation along temporal modulation (ω) and spectral modulation (Ω) varying along frequency (f). Principal Component Analysis (PCA) is then used to reduce dimensionality along the frequency axis to capture the most variance. Flattened, this reduced representation spans 3300 dimensional space.

2.3. BioNet: Artificial Midbrain Model

Using the auditory spectrogram described in the Section 2.1 as input, BioNet simulates neural response of the bat's mid-

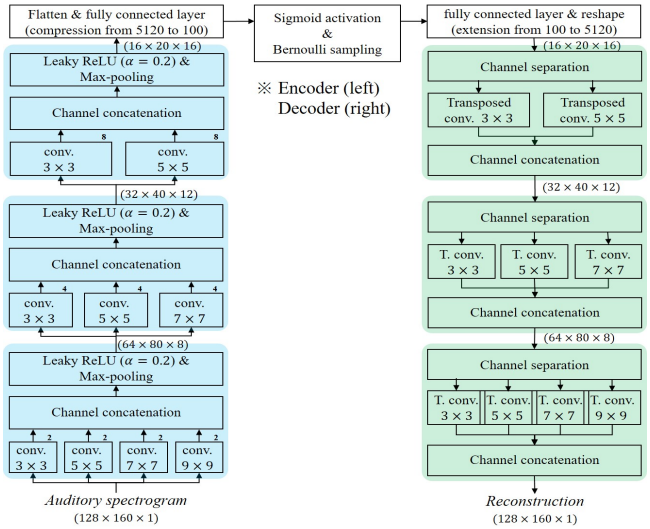


Fig. 1: Autoencoder architecture for training BioNet: Blue and green boxes represent the encoder and the decoder, respectively. After training, only encoder is used for the BioNet.

brain. Conceptually, BioNet is designed using an autoencoder framework which performs a compression and reconstruction using an encoder/decoder pipeline (Fig. 1). The embedding representation employs a Bernoulli sampling on the sigmoid activation of encoder outputs, similar to principles of a variational autoencoder [18]. This sampling process allows the network to emulate typical neural activity including excitation and inhibition.

Concretely, the model is optimized to minimize total loss:

$$L = \sum_{n}[(x_n - D(E(x_n)))^2 + \lambda(\rho - \sum_{i}\sigma(E^i(x_n)))^2] \quad (3)$$

where x is for training data sample with data index n. E and D means encoder and decoder, respectively. The sigmoid activation, $\sigma(.)$ provides a prior probability for Bernoulli sampling. Network nodes are indexed by i, and λ is a regularization coefficient to restrict the number of active nodes to ρ hence allowing a sparsity constraint on the network activity. The network is trained using natural bat vocalizations from 17,713 calls (about 10 min of data). Full details of the architecture are provided in [10] and reveal that network filters converge on characteristics that match those reported in the midbrain of biological neurons recorded in bats [9]. In this study, BioNet is trained with $\lambda = 0.0001, \rho = 10$. The output of the Encoder is used as the representation feature of BioNet, which is a 100-dimensional vector.

3. EXPERIMENTS

3.1. Database: Echo sound collection

A sensor set which consisting of a microphone and a speaker is placed on 50 cm far from a target object in anechoic room (Fig. 2a-b). Echo sounds are manually recorded with five

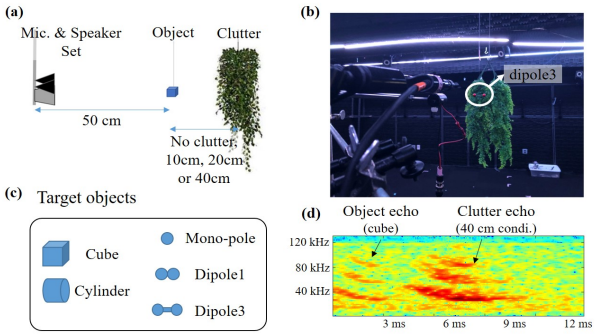


Fig. 2: About echo recordings: (a) illustration of recording condition, (b) picture of recording setup, (c) target objects, (d) an example of echoes

types of objects: cube, cylinder, mono-pole(sphere), dipole1, and dipole3 (Fig. 2c). All objects are made by 3D printer with 3cm size for cube and cylinder while the diameter of mono-pole is set to 1cm. The dipoles are made by connecting two mono-poles with a thin bar in different gap, 1cm (dipole1) and 3cm (dipole3). Note that two mono-poles are adjacent to each other for dipole1.

For each object, recordings are performed from three-different viewpoints: 0°, 45°, and 90°. 0° references the flat surface of the cube and cylinder, as well as the dipole position to observe two mono-poles with the gap. Additional recordings are performed in four-different conditions. Clutter is introduced using artificial bush-like leaves. Each recording configuration (object type, viewpoint, and clutter condition) is repeated 10 times.

A call sound to induce an echo is synthesized by mimicking big brown bat's echolocating call, using second-order Frequency Modulation (FM) sweeps [19]. The synthetic call consists of two FM sweep for 2 ms: A 1^{st} FM sweep from 50 kHz to 25 kHz and a 2^{nd} from 100 kHz to 50 kHz.

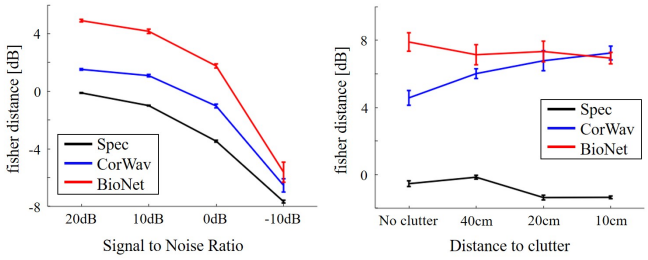
3.2. Analysis Pipeline

To investigate discriminability using different signal representations, echoes from objects and orientations are categorized into 12 classes. Note that mono-pole echoes from 3-different views are always the same because it is a sphere. Cube echoes at 0° and 90° are the same as well.

This study opts for a direct measure of discriminability across classes using Fisher distance, which is defined as:

$$fisher\ distance(i,j) = \frac{trace(C_{i,j})}{trace(C_i) + trace(C_j)} \quad (4)$$

where C_i is a covariance matrix for representation features of class i, and $C_{i,j}$ is a cross covariance matrix for the features of two classes i and j. This fisher distance is a ratio of variance in between classes to variance of within class. Then, average of fisher distances for all combinations (${}_{12}C_2 = 66$) is applied. The same procedure is repeated under different noise configurations by evaluating all models using noisy echoes



(a) Under noisy condition (b) Under clutter condition

Fig. 3: fisher distance in noisy and clutter condition.

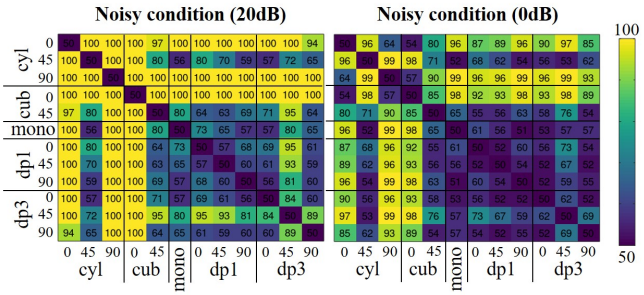


Fig. 4: Confusion matrix, a result of echo discrimination based on nearest neighbor for BioNet.

which are synthesized by adding Gaussian white noise at different SNR values. For statistical comparison, 8:2 cross validation is performed 10 times, and the results are summarized using the mean and standard deviation.

3.3. Object Recognition Results

For an overall evaluation, we analyze outputs of all models using their original mappings, without dimensionality reduction (5120-D for Spec, 100-D for BioNet and 3300-D for CorWav). Discriminability results in noisy and cluttered environments of these representations are contrasted in Fig. 3(a)-(b), respectively. As anticipated, discriminability diminishes with increased background noise (panel a), though BioNet shows a clear robustness all the way to -10dB. In contrast, the clutter condition reveals interesting patterns in echo analysis (panel b). On the one hand, BioNet shows nearly consistent discriminability across clutter conditions, potentially reaching a plateau of performance. In contrast, the CorWav representation appears to benefit from the presence of clutter behind the object as the clutter gets nearer to the object (from $40cm$ to $10cm$). An interpretation of these results supports some speculations in the community that sound waves bouncing off a background may provide a back mirror view of an object hence potentially informing of its identity and improving its recognition. Finally, The results from the Spec representation show very weak discriminability in presence of clutter, where clutter echoes seem to truly mask the object identity in the time-frequency spectrogram.

Fig. 4 shows the confusion matrix resulting from echo discrimination based on nearest neighbor with BioNet feature in example noisy conditions. As illustrated in the figure, BioNet shows perfect discriminability between the cube and cylinder relative to other shapes. The discriminability definitely diminishes from weak-power echoes (small dipole, mono-pole, and edge view of cube and cylinder), though it is higher for rotated large dipole. This trend is maintained in the 0dB condition where strong-power echoes from flat or curved views are still discriminable against others.

Finally, since the different representation vary widely in their dimsensionality, we wanted to control for this element by projecting all representations onto the same number of dimensions via PCA. Fisher distance is recalculated each time on the controlled subspace. Table 1 shows that BioNet results in the best discriminability among the methods regardless of dimensionality.

Table 1: fisher distance [dB] of features in controlled dimension under noisy condition.

20dB	100-dim.	50-dim.	25-dim.
Spec	0.35 ± 0.01	2.48 ± 0.01	3.80 ± 0.16
CorWav	5.96 ± 0.10	6.25 ± 0.12	5.96 ± 0.11
BioNet	$\mathbf{6.62 \pm 0.05}$	$\mathbf{6.77 \pm 0.09}$	$\mathbf{6.53 \pm 0.14}$

0dB	100-dim.	50-dim.	25-dim.
Spec	-2.40 ± 0.01	-1.18 ± 0.02	0.38 ± 0.05
CorWav	2.60 ± 0.03	3.13 ± 0.01	2.39 ± 0.01
BioNet	$\mathbf{3.09 \pm 0.03}$	$\mathbf{3.71 \pm 0.03}$	$\mathbf{3.71 \pm 0.02}$

4. CONCLUSION

Inspired by bat biosonar processing, this study models echo sound representations in biological systems. These approaches capture both rich generic representations (time-frequency spectrogram and high-dimensional wavelet-like mapping) and data-driven representations (autoencoder matching natural statistics of bat vocalizations). These representations are evaluated on controlled recordings of echo sounds. The results reveal a clear advantage of the optimized representation using the BioNet model. Embeddings from this model extract signal characteristics that match those from the auditory midbrain of bats and rely on natural statistics as inference principles to constrain this mapping. The generic but representation based on cortical wavelet decomposition also reveals some advantages in clutter though never outperforms the BioNet features. Overall, while all the features explored in this work span high-dimensional spaces and combine linear and nonlinear transformations, the spectro-temporal statistics of bat calls appear to underlie noise invariance in both biological and artificial systems. Looking ahead, these results provide a foundation for expanding sonar object recognition to 3D objects with a focus on integrating a sequence of echoes from multiple views of an object. This direction will further improve our understanding of echo-based object recognition and further advance the field.

5. REFERENCES

[1] David P. Williams, Francesco Baralli, Michele Micheli, and Simone Vasoli, "Adaptive underwater sonar surveys in the presence of strong currents," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 5 2016, pp. 2604–2611, IEEE.

[2] Matteo Franchi, Alessandro Ridolfi, and Benedetto Allotta, "Underwater navigation with 2D forward looking SONAR: An adaptive unscented Kalman filter-based strategy for AUVs," *Journal of Field Robotics*, vol. 38, no. 3, pp. 355–385, 5 2021.

[3] Arvind Kumar, Rampravesh Kumar, Mahesh Chandra, and Kamlesh Kishore, "Study of under-water Sonar System for change in propagation speed, depth of water, bottom loss and estimating optimal PDFs," in *2023 6th International Conference on Information Systems and Computer Networks, ISCON 2023*. 2023, Institute of Electrical and Electronics Engineers Inc.

[4] Yulin Tang, Liming Wang, Shaohua Jin, Jianhu Zhao, Chao Huang, and Yongcan Yu, "AUV-Based Side-Scan Sonar Real-Time Method for Underwater-Target Detection," *Journal of Marine Science and Engineering*, vol. 11, no. 4, 4 2023.

[5] Itiel E. Dror, Mark Zagaeski, and Cynthia F. Moss, "Three-Dimensional target recognition via sonar: A neural network model," *Neural Networks*, vol. 8, no. 1, pp. 149–160, 1 1995.

[6] Joseph M. Fialkowski and Roger C. Gauss, "Methods for identifying and controlling sonar clutter," *IEEE Journal of Oceanic Engineering*, vol. 35, no. 2, pp. 330–354, 4 2010.

[7] Hyeonwoo Cho, Jeonghwe Gu, and Son Cheol Yu, "Robust Sonar-Based Underwater Object Recognition Against Angle-of-View Variation," *IEEE Sensors Journal*, vol. 16, no. 4, pp. 1013–1025, 2 2016.

[8] Cynthia F. Moss and Annemarie Surlykke, "Auditory scene analysis by echolocation in bats," *The Journal of the Acoustical Society of America*, vol. 110, no. 4, pp. 2207–2226, 2001.

[9] S. Andoni, N. Li, and G. D. Pollak, "Spectrotemporal Receptive Fields in the Inferior Colliculus Revealing Selectivity for Spectral Motion in Conspecific Vocalizations," *Journal of Neuroscience*, vol. 27, no. 18, pp. 4882–4893, 2007.

[10] Sangwook Park, Angeles Salles, Kathryne Allen, Cynthia F. Moss, and Mounya Elhilali, "Natural Statistics as Inference Principles of Auditory Tuning in Biological and Artificial Midbrain Networks," *eneuro*, vol. 8, no. 3, pp. 0525–20, 5 2021.

[11] M Elhilali, S A Shamma, J Z Simon, and J B Fritz, "A Linear Systems View to the Concept of STRF," in *Handbook of Modern Techniques in Auditory Cortex*, D Depireux and M Elhilali, Eds., pp. 33–60. Nova Science Pub Inc, 2013.

[12] Joakim Anden and Stéphane Mallat, "Multiscale scattering for audio classification," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, 2011, pp. 657–662.

[13] Christoph E Schreiner, "Order and disorder in auditory cortical maps," *Current Opinion in Neurobiology*, vol. 5, no. 4, pp. 489–496, 8 1995.

[14] Edgar Hemery and Jean-Julien Aucouturier, "One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis," *Frontiers in Computational Neuroscience*, 2015.

[15] X Yang, K Wang, and S A Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[16] T Chi, P Ru, and S A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[17] T. Chi and S. Shamma, "NSL Matlab Toolbox," 2005.

[18] Carl Doersch, "Tutorial on Variational Autoencoders," 6 2016.

[19] Kathryne M. Allen, Angeles Salles, Sangwook Park, Mounya Elhilali, and Cynthia F. Moss, "Effect of background clutter on neural discrimination in the bat auditory midbrain," *Journal of Neurophysiology*, vol. 126, no. 5, pp. 1772–1782, 11 2021.