PennState
Applied Research Laboratory

Aural Based Scene Understanding for Sonar Applications

Final Report

Daniel C. Brown

28 November 2023

Applied Research Laboratory
P.O. Box 30
State College, PA 16804-0030

Sponsored by: U.S. Office of Naval Research
Grant No.: N00014-19-1-2679

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
31-07-2023	Final	01 Aug 2019 to 31 Jul 2023

4. TITLE AND SUBTITLE
Aural Based Scene Understand for Sonar Applications

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-19-1-2679

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Daniel Brown

5d. PROJECT NUMBER
27653

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Pennsylvania State University Applied Research Labotatory
Office of Sponsored Programs
110 Technology Center Building
University Park, PA 16802-7000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 North Randolph Street
Arlington, VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
321

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
"Approved for Public Release; Distribution is Unlimited."

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Acoustic returns scattered from either surfaces or objects are frequently beamformed or image-formed into a representation that is intuitive for the human visual system, enabling both human operators and computer vision algorithms to analyze the acoustic data. However, the generation of these representations from raw sonar data can be computationally burdensome; in fact, beamforming is the processing bottleneck in some embedded synthetic aperture sonar (SAS) systems. It is hypothesized that in certain situations, SAS image formation may have limited benefits or be completely unnecessary, suggesting an adaptive signal processing chain that achieves enhanced efficiency by computing only the necessary representations of data in a given context.
The objective of this research is to investigate approaches using alternative data representations to directly exploit the information content of raw or near-raw sonar data. In particular, this effort explored how one can encode and exploit streaming sonar data as it is received, and the associated performance tradeoffs when beamforming is omitted.

15. SUBJECT TERMS
sonar; synthetic aperture sonar; scattering; modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		15	Daniel Brown
u	u	u			19b. TELEPONE NUMBER (*Include area code*) 814-865-1193

Applied Research Laboratory
State College, PA 16804

28 November 2023

Sponsored by: U.S. Office of Naval Research

Aural Based Scene Understanding for Sonar Applications

Final Report

Daniel C. Brown

1 Project Goals and Objectives

1.1 Summary Statistic Feature Analysis

Acoustic returns scattered from either surfaces or objects are frequently beamformed or image-formed into a representation that is intuitive for the human visual system, enabling both human operators and computer vision algorithms to analyze the acoustic data. However, the generation of these representations from raw sonar data can be computationally burdensome; in fact, beamforming is the processing bottleneck in some embedded synthetic aperture sonar (SAS) systems. It is hypothesized that in certain situations, SAS image formation may have limited benefits or be completely unnecessary, suggesting an adaptive signal processing chain that achieves enhanced efficiency by computing only the necessary representations of data in a given context.

The objective of this research is to investigate approaches using alternative data representations to directly exploit the information content of raw or near-raw sonar data. In particular, this effort explored how one can encode and exploit streaming sonar data as it is received, and the associated performance tradeoffs when beamforming is omitted.

The problem of seabed texture characterization is employed as a testbed for this investigation. The approach herein examines the information in individual and combined sonar echo returns as they pertain to forming auditory objects that combine to form a scene. In particular, we investigated the application of summary statistics as a means for representing underlying scene characteristics such as sand ripples, gravel beds, grass, rocks, sediment, and sea grass. Summary statistics were drawn from the sonar performance estimation community and from communities that study aural perception systems, and the features they represent range from nearly stationary to episodic whether spatially or temporally.

This effort focused on the analysis of side-looking synthetic aperture sonars used for mine countermeasures (MCM). The ability of mine-hunting systems to separate target signatures from the surrounding environment varies with the quality of processed sensor output and the scene complexity. Recently, ONR has supported a number of programs directed toward Performance Estimation (PE) of Automated Target Recognition systems. Recent work in this area has identified image summary statistics that are predictors of both human-assessed and machine-assessed complexity. Early results detailing the process of directly measuring human assessed complexity and the discovery of informative image summary statistics are provided in [1]. Similarly, Midtgaard et al. [2], Gips and Williams [3], and Williams [4] have each separately used image-based summary statistics to predict the performance of Automated Target Recognition Systems.

The long-term goal for the proposed work in this area is to establish predictive links between the raw, sensed data and ATR-based detection and classification performance. Unlike the existing ONR performance estimation programs, which develop image quality and image complexity metrics, this program investigated summary statistics calculated on the raw sensor time series. This work sought to identify raw-data (or near-raw data) summary statistics that are sensitive to both the presence of manmade objects and the clutter-scale texture of the observed scene. The sensitivity of implemented metrics to the clutter-scale scene texture

are investigated through analysis of near-raw data collected by asynthetic aperture sonar (SSAM2) [5]. The results shown herein suggest that while texture classification on the near-raw time-series does not have irrelevant performance, image-based classification still shows substantially superior performance on a small, curated dataset from SSAM1.

1.2 Citizen Hydrophone

A secondary goal, which was identified during the execution of this research program, was led by Dr. Thomas Blanford. This project seeks to improve the accessibility of passive acoustic sensing. Inexpensive passive sensors have served important public outreach roles in other acoustic communities. The Raspberry Shake is a seismic sensor consisting of a geophone, signal conditioning and digitizing electronics, and a Raspberry Pi. Due to their low cost and ease of use, they are widely deployed throughout the world in homes, schools, museums, and well as remote seismic research stations. There are not currently any inexpensive digital hydrophones that can be widely deployed like the Raspberry Shake. COTS solutions that connect hydrophones to digital audio recorders are available, but these still cost nearly $500 per unit.

The objective of the Citizen Scientist Hydrophone effort is to develop a passive digital hydrophone with a price point of less than $100 that is still a scientifically meaningful device. This target price is intended to make it affordable and accessible to students, schools, and individuals interested in science. Under the current effort, we initiated investigating instrumentation approaches by designing and building a prototype system in order to identify tradeoffs and limitations in the hardware.

2 Accomplishments Towards Achieving Goals

This program initially pursued adaptation of a feature extraction technique proposed by Young and Hines to the problem of texture segmentation from streaming sonar data [6]. These initial efforts, which are summarized in Section 2.1, characterized the performance of these algorithms for a high-resolution synthetic aperture sonar system. Later, in the interest of simplifying the data processing pipeline, the Tunable-Q Wavelet Transform (TQWT) and other summary statistics were employed to provide both “aurally inspired” and standard data representations. These results are detailed in Section 2.2.

2.1 Aural Features for Time Series - Young and Hines

The initial goal pursued this year was the identification of summary statistics, inspired by the human auditory system, that are sensitive to the presence of manmade objects. The method employed closely follows the aural classification algorithm of Young and Hines [6]. Their approach, inspired by musical acoustics literature, models the human perceptual sensitivity to timbre. The sequential data conditioning and feature extraction steps of the Young and Hines Method is briefly summarized here.

Their method operates on raw time series recorded by a sonar system and processes this

through a gammatone filter bank to create a dense set of sub-bands. These filters are intended to mimic the spectral response of the human auditory system. For each of the sub-banded waveforms, the envelope is calculated from the Hilbert transform and a series of perceptual features are extracted. A set of these features enumerate the shape of the rising and falling signal envelope surrounding a target echo. These envelop-derived features measure the attack time (time to reach the envelope peak), the attack slope (the slope of the envelope leading to this peak), the decay time (time to fall from the peak to the end of the return, and the decay slope (the slope of the envelope following the peak). These attack related features are calculated across each extracted sub-band. The resulting high-dimensional feature space is collapsed by calculating the minimum, maximum, and mean of each attack parameter. Additionally, the sub-band frequency of the minimum and maximum attack parameters is extracted. Finally, a set of perceptual spectral signal features are extracted that include measures of the peak loudness frequency, peak loudness level, and additional measures characterizing the spectral character. The Young and Hines method of feature extraction produces a thirty-dimensional feature space in which to classify an acoustic return.

2.1.1 Manmade Target Analysis

A set of cylindrical objects were procured by Dr. Daniel Park under his "Alternative Representations of Information for Acoustics" ONR research grant. The objects all have the same external geometry with a length of 8 inches and a diameter of 2 inches, but their internal geometries vary. The objects include solid cylinders, hollow cylinders, hollow cylinders with a small vent hole (<1 mm), and pipes. The hollow cylinders and pipes have wall thicknesses of 0.032 in, 0.065 in and 0.120 in. The cylinder materials include steel, copper, aluminum, and wood. The specific combinations of targets available in the set are provided in Table 1.

The cylindrical targets were analyzed in air by collecting their radiated signature when excited with a force hammer. The radiated signature was collected using a calibrated reference microphone and digitized with a National Instruments data acquisition system. Each of these objects were measured for five separate strikes with the hammer. This data set will be referred to as the "tap data".

The hollow cylindrical targets (hollow, hollow with hole, and pipe) were sub-selected for analysis with the Young and Hines aural feature set. This subset was selected because they are expected to exhibit similar ringing from the excitation of elastic waves due to their thin walls. Feature extraction across the tap data for these objects produced a total of 65 measurements (13 objects with 5 measurements each). The thirty-dimensional feature space has been further reduced to using principal component analysis (PCA). The first two dimensions from this projection are shown in Figure 1. Without supervision, the objects in the test set have clustered according to material type. This provides a qualitative indication that the selected aural feature set is sensitive to the composition of these manmade objects.

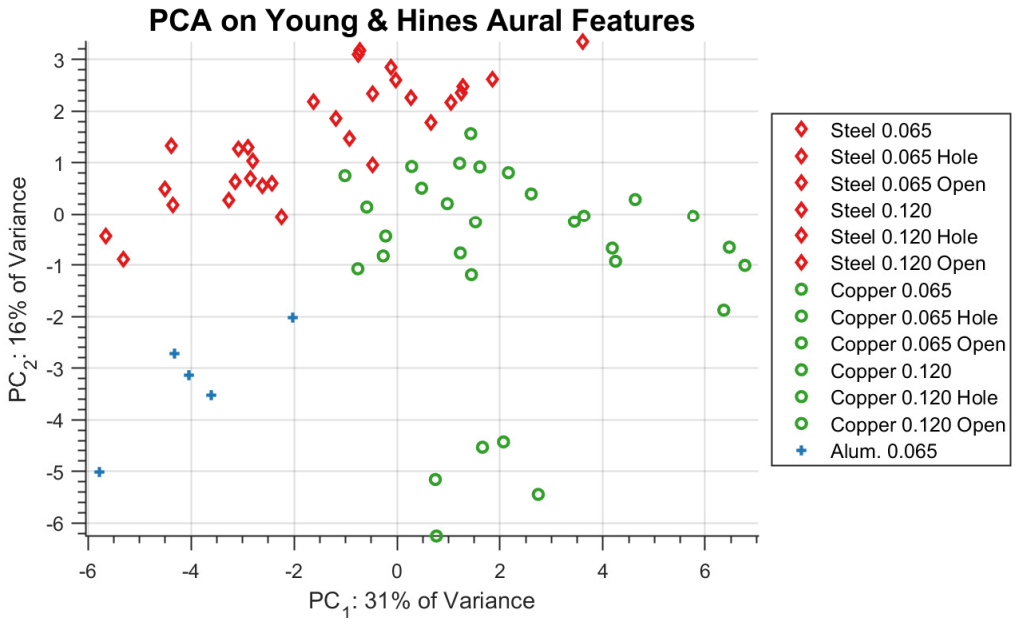


Figure 1: The features extracted from the tap data collected from the cylindrical targets exhibits unsupervised clustering along the first to dimensions in PCA analysis.

Table 1: Several cylindrical targets with length 0 cm and diameter 0 cm were analyzed for this project

Object Number	Object Type	Outer Diameter [in]	Length [in]	Wall Thickness [in]	Note
1	Stainless Steel Pipe	2	8	0.065	Open sides
2	Stainless Steel Pipe	2	8	0.065	Closed sides
4	Stainless Steel Pipe	2	8	0.12	Open sides
5	Stainless Steel Pipe	2	8	0.12	Closed sides
6	Copper Pipe	2	8	0.065	Open sides
7	Copper Pipe	2	8	0.065	Closed sides
8	Copper Pipe	2	8	0.032	Open sides
9	Copper Pipe	2	8	0.032	Closed sides
10	Stainless steel solid cylinder	2	8	-	-
11	Copper solid cylinder	2	8	-	-
12	Wood solid cylinder	2	8	-	-
13	Aluminum pipe	2	8	0.065	Open sides

2.1.2 Near-Raw Sonar Data Analysis

The same aural-based feature extraction technique has been applied to data collected by the SSAM2 system [7] at a recent field exercise that exhibited a wide range of bottom textures. The SSAM2 system is a dual-band SAS with an upper frequency band operating at frequencies greater than 100 kHz as reported in [8]. Three seafloor textures were selected from this experiment for analysis. The first texture is flat and relatively featureless. There are some small (less than 5 cm localized scatterers, but no large-scale clutter or bathymetric features. This texture is labeled **"smooth"**. The second texture consists of a seafloor that has a series of regular, circular, mound-like features. The mound-like features do not have sharp edges that would indicate they are part of a rock formation. Instead, they seem to be made of some accumulation of sediment with a smooth transition up from the surrounding seafloor. The primary visual characteristic of this imagery is the presence of shadows that are created by the mounds. The mounds are roughly 1 m to 2 m in diameter and they have a height of approximately 50 cm. A single 5000 m^2 scene contains more than 100 of these mound-like features. This texture is labeled **"mound"**. The third texture is drawn from surveyed scenes with large rock outcrops. These outcrops span a large fraction ($> 20\%$) of the image for scenes selected for this texture. The rock outcrops produce bright highlights with sharp and rapid rises in received level. This texture is labeled **"rock"**. The total number of SSAM2 pings labeled for each texture type is: smooth 1279, mound 1110, and rock 952.

The raw sensor data acquired by the SSAM2 system must be preconditioned prior to feature extraction using the aural-based representation. The major preconditioning steps are:

Replica correlation The raw time series are replica correlated with the analytic transmit waveform. This focuses seabed and target returns and provides increased signal-to-noise ratio.

Range varying normalization The received signal level falls throughout a single ping due to losses associated with sensor directivity, acoustic spreading, acoustic attenuation, and decreased sediment backscattering strength at low grazing angle. A range varying normalization, based on the median received level for a given range, is applied to the data to offset these losses.

Spectral flattening The post-replica-correlated waveform does not have uniform power spectral density over the operating band of the sensor. This is due to temporal windowing applied to the transmit waveform, the sensitivity of the acoustic projector, and the sensitivity of the acoustic receiver. The spectral response is estimated from the receive data and a spectral normalizer is applied in the frequency domain. Care is taken to regularize the normalization, which reduces errors that may be introduced at the edges of the operating band.

Coherent averaging The SSAM2 receive array consists of multiple hydrophones. Within each ping, the signals from this set of hydrophones are coherently averaged. This coherent averaging is equivalent to processing the sensor as a real-aperture sonar. This step provides modest focusing of the data in the along-track direction at minimal

processing cost.

Spectral mapping The final processing stage performs maps the analytic waveform into a real-valued time series for the full operating bandwidth of the sensor.

Echo detection For each ping, a detector is used to search for the start and stop indices of distinct target return. In some cases, the detector does not detect an echo and these pings are discarded.

Aural-based feature extraction is performed on the SSAM2 data after the preconditioning and the target echo detection steps. The steps are like those used in Section 2.1.1. The SSAM2 system has a larger bandwidth than that recorded in the tap data, and the gammatone filter band was adjusted to account for this. The thirty-dimensional feature space has been further reduced to using principal component analysis (PCA). The first two dimensions from this projection are shown in Figure 2. Without supervision, the rock texture is clustered with modest separability from the smooth and mound textures. The smooth and mound textures do not appear to be distinguishable in this data representation. If we remember, the aural-based feature extraction is focused on characterizing the attack of the signal envelope. It is unsurprising to see that the rock texture, which has temporally rapid variability, is distinct from the smooth texture in this feature space. The mound texture has a number of shadow like features that do not seem to be captured by the selected aural-based features for two reasons: 1) the aural-based feature attempts to characterize the signal attack, which is weak for the mound and 2) the shadows created by these mounds are quite weak in the near-raw data. The aural-based feature set developed by Young and Hines shows modest promise for identifying textures in streaming data where clutter produces strong echo signatures.

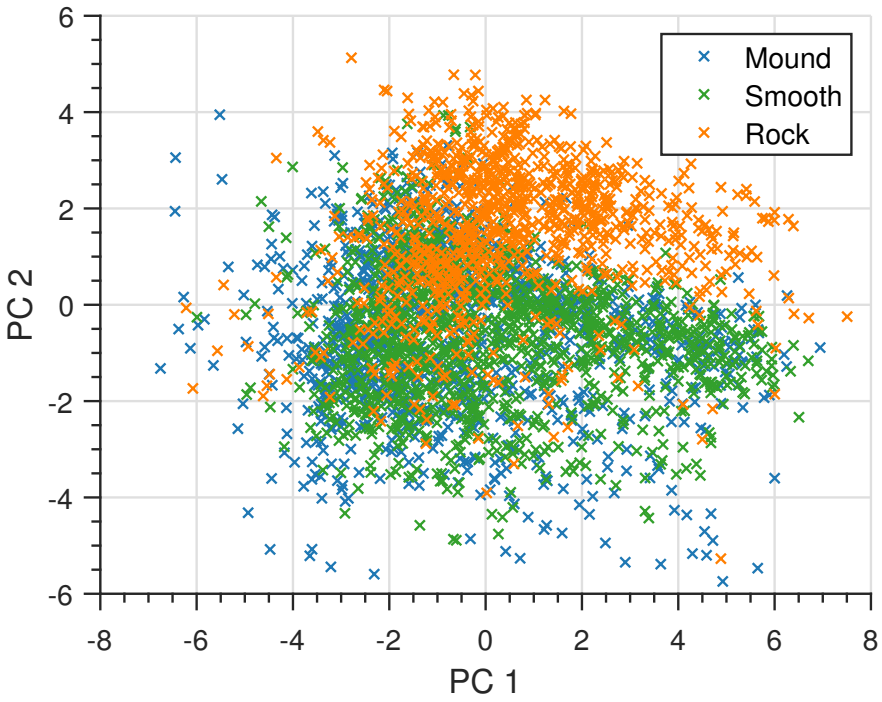


Figure 2: The aural-based features show natural clustering of the rock texture from the smooth and mound textures. The smooth texture is not easily distinguished from the mound texture in this example.

2.2 Alternative Aural-Inspired Summary Statistics

The initial goal pursued was identification of summary statistics, inspired by the human auditory system, following the aural classification algorithm of Young and Hines [6]. To achieve modest results with these features required a significant amount of data preconditioning. The conditioning steps included: (1) replica correlation, (2) range varying normalization, (3) spectral whitening, (4) coherent hydrophone averaging, (5) spectral mapping to real time series, and finally (5) echo detection. This multi-stage preconditioning process was driven by the design of the Young and Hines method, which extracts features from target or events within the recorded time series. To apply this method, the data must be preconditioned so that a detector can identify key events from which to extract features.

In the interest of avoiding the complex preconditioning described above, we compute two alternative sets of aural-inspired summary statistics for seafloor texture classification. Taking inspiration from [6], we employ the gammatone filter bank to provide the total power (in the natural logarithm) in each of 10 logarithmically-spaced bands as a 10-D feature.

The second approach is based on applying a modern wavelet transform to the input time series. As noted above, the analysis stage of the human auditory system can be described as a constant-Q filter bank, where the filters are logarithmically spaced [9]. The bases of a wavelet transform can be defined to meet these criteria. In fact, past work has developed a wavelet transform based on the gammatone filter [10]. In this work we apply the tunable-Q wavelet transform (TQWT) to the time series generated by the SSAM1 system. TQWT

provides an overcomplete basis where the frequency response of the constant-Q filters may be adjusted to the specific task [11]. This representation shares some the key aural-inspired properties of the gammatone filter bank – logarithmically-spaced filters with a constant-Q frequency response. The flexibility of the TWQT has allowed this transform to be applied to a wide variety of acoustic data, including assessment of human speech to detect the onset of Parkinson's disease [12], and decomposition of vibration signals for gearbox fault detection [13].

A small-scale dataset has been curated to assess seafloor texture identification from raw time series. A set of four textures have been labeled from SSAM1 field trials. These textures (as well as a "heterogenous" texture *not* used in training) are shown in Figure 3, where imagery is shown in 3a-3e and the associated matched filter output magnitude in 3f-3j. All images in Figure 3 are shown on a 60 dB scale individually indexed to the largest value. From each texture, four such SSAM1 scenes were selected that each have relatively continuous texture type (trivializing the labeling process). The imagery and corresponding time series were then partitioned into non-overlapping chips (4 m×4 m) as demonstrated in Figure 4. A set of features are extracted from each chip. A brief summary of each is provided below.

Scintillation Index - Image and Time Series The scintillation index is calculated across a range of scales for both the imagery and the time series.

Two-Dimensional Morlet Wavelet - Image and Time Series This two-dimensional continuous wavelet transform is calculated for each chip across a range of wavelet scales. Features are generated by calculating the power associated with each scale.

Tunable-Q Wavelet Transform - Time Series Only The TQWT is applied to the time series from each image chip. The Q of the filters is selected to maximize the number of bands extracted from the time series.

Gammatone Filter Bank - Time Series Only The gammatone filterbank is used to segment the time series into ten logarithmically-spaced bands. The natural logarithm of the power of each band is provided as a feature.

Patch Mean - Image Only This value is simply the mean value of the chip. While uninteresting, past results have indicated it can be an effective support feature for clustering seafloor textures.

The total feature count was 10 for imagery and 30 for time series. These were fed into the minimum redundancy maximum relevance (MRMR) dimensionality reduction algorithm, which uses label supervision to determine the predictive relevance of each feature. The results are shown in Figure 5. Classification was performed using only the top n Principal Components, where n is sufficient to explain 97% of the variance amongst the training data (4 for imagery, 6 for time series).

Across the four textures a total of 3640 chips were produced for both the imagery and time series. One image (364 chips) was held-out as a test set. Amongst the remaining 9 images (3276 chips), five folds were randomly sampled. Hyperparameters for a K-nearest-neighbors-based classifier were optimized in the manner of 5-fold cross validation (independently between the

time series and image chip datasets). The performance of the time series and image representations evaluated on the hold-out test set are compared in Figure 6. Confusion matrices are shown for the imagery representation in Figure 6a and for the time series representation in Figure 6b. The imagery representation shows superior performance; however, meaningful separation is possible for the time series representation.

2.3 Citizen-Scientist Hydrophone Development

Drawing inspiration from the Raspberry Shake system, an initial proof of concept design for the system was built around a Raspberry Pi 4. By leveraging audio hardware compatible with the Raspberry Pi, the system would be able to monitor and record underwater acoustic data within (and above) the audible range, approximately 20 Hz to 50 kHz. By spanning the audible range, end users can listen to their data. Listening to underwater sounds can help non-scientists gain intuition and understanding of underwater environments. The Raspberry Pi also allows the user to integrate non-acoustic sensors and customize the data acquisition for their particular experiment.

A block diagram for the proof-of-concept system is shown in Fig. 7. The Raspberry Pi platform has a wide variety of hardware and sensor systems that are designed for simple integration. A COTS battery and charging system and two channel audio analog to digital conversion are leveraged for this system. The preamplifier, however, was custom designed for the system because no suitable COTS solutions exist. Although there are audio preamplifier circuits compatible with the Raspberry Pi, their input impedance, gain, and noise floor are not well matched for a piezoelectric hydrophone. The system is capable of recording signals from two hydrophones simultaneously.

Two EDO spherical hydrophones were repurposed from another project to use as the transducers in this system. All of the electronics are assembled into an aluminum housing for protection during testing. The user interacts with the system through switches and LED indicators to start and stop measurements. A photograph of the prototype system is shown in Fig. 8. The functionality of the unit has been verified through testing in the ARL/PSU anechoic tank facility.

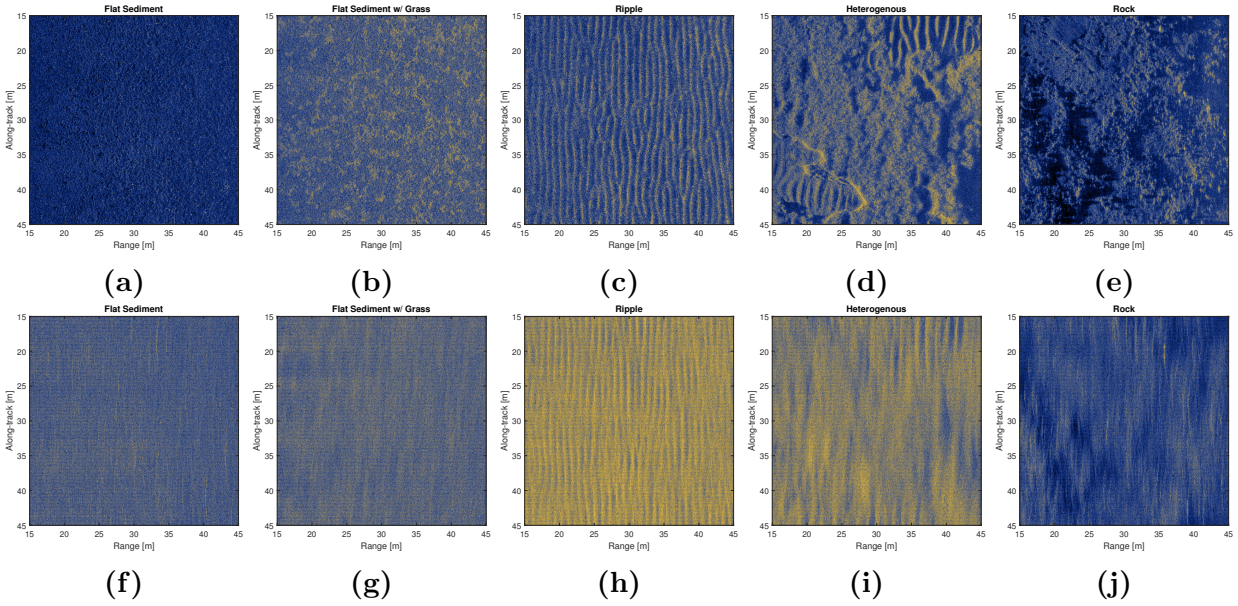


Figure 3: Five textures (flat sand, flat sand with grass, ripple, heterogenous, and rock are shown as imagery (a)-(e) and the associated matched filter output (f)-(j).

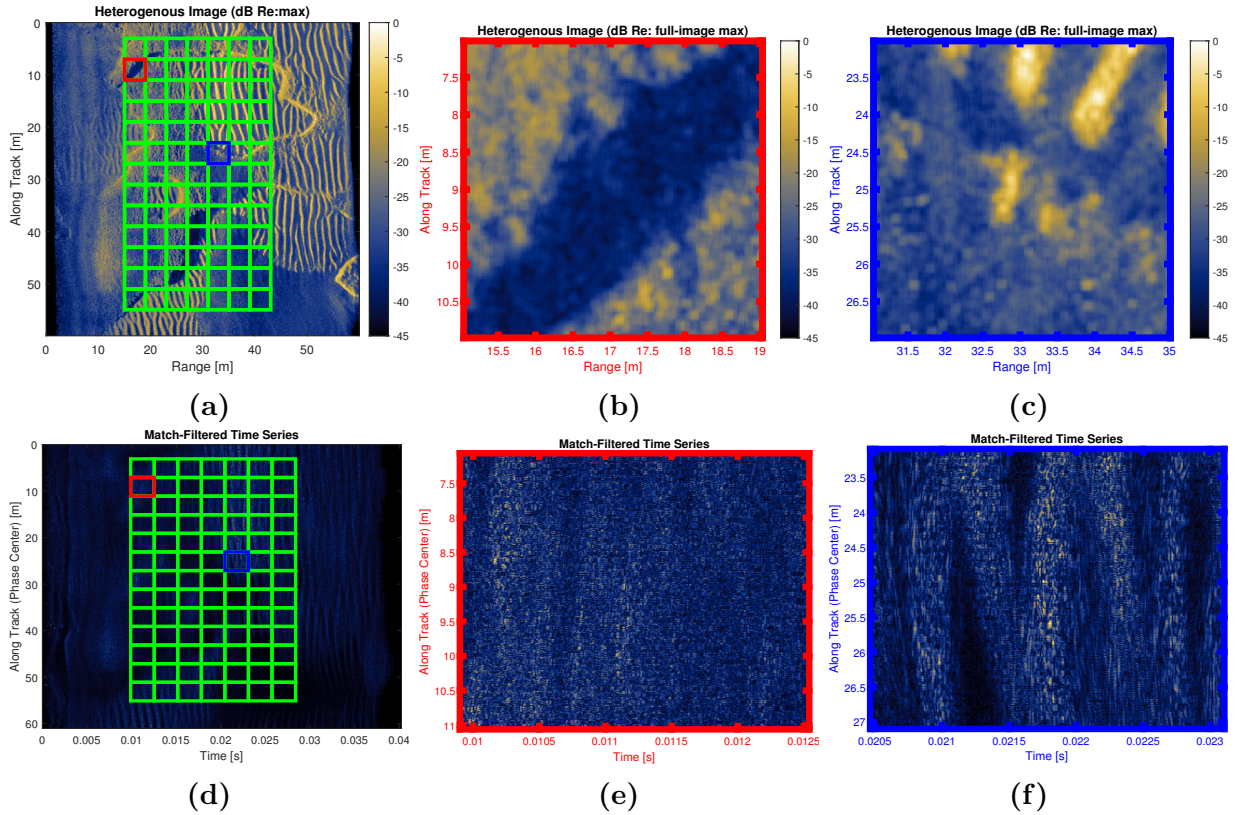
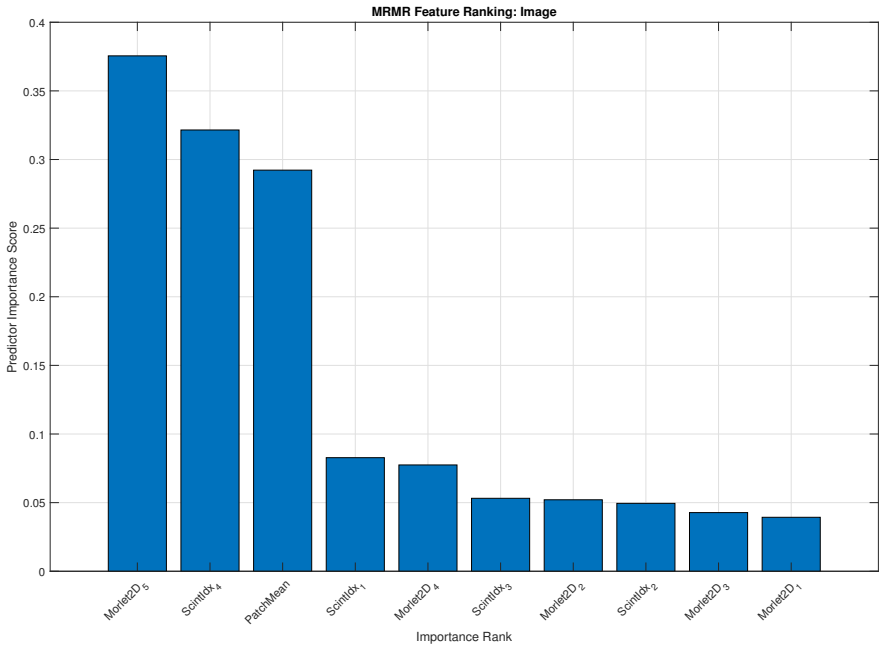
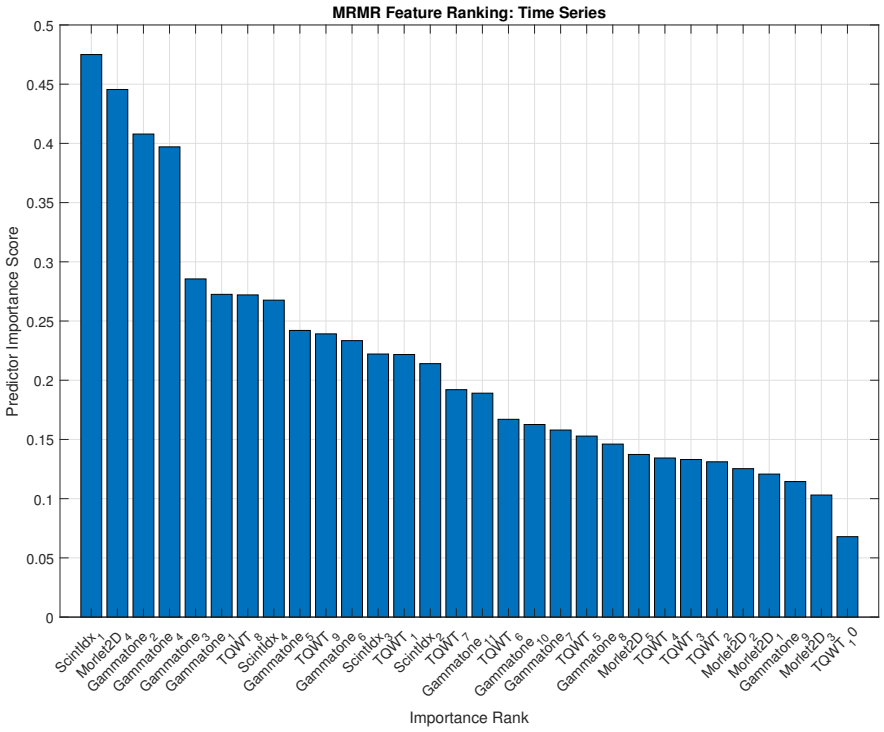


Figure 4: (a) a 60 m×60 m SSAM1 image with a 7×13 grid depicting all partitioned chips; (b)-(c) 4 m×4 m chip examples, with borders colored to match corresponding colored grid blocks in (a); (d)-(f) the matched filter time series corresponding to panels (a)-(c). Spatial gating is used to ensure artifact-free data and prevent data leakage between images.

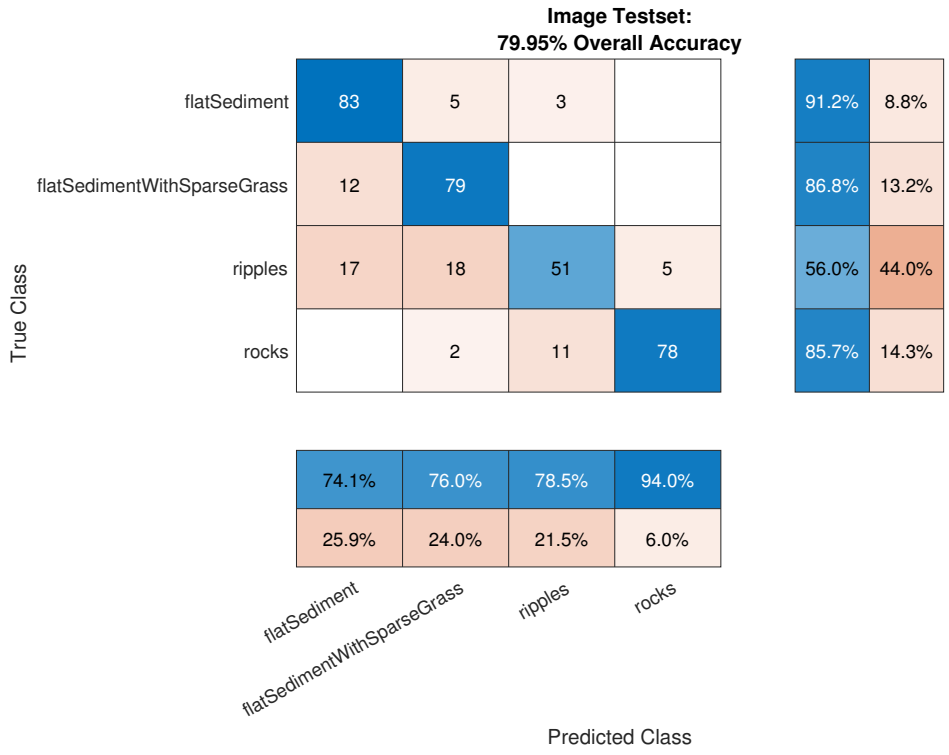


(a)

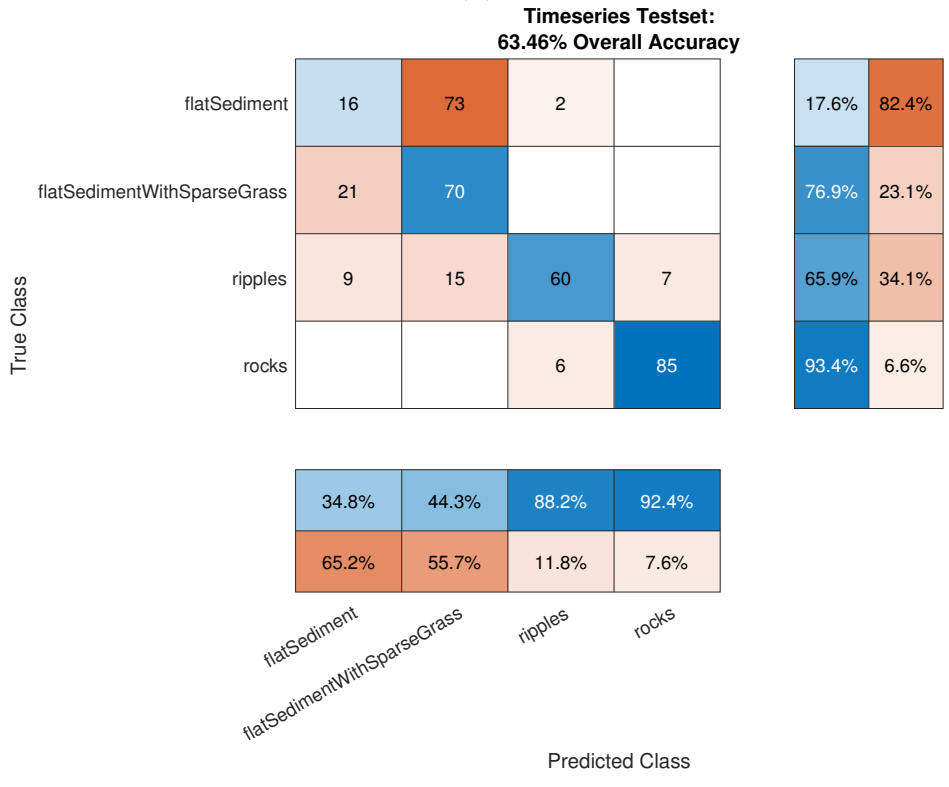


(b)

Figure 5: Results of the minimum redundancy maximum relevance (MRMR) dimensionality reduction algorithm applied to (a) the image chip dataset, and (b) the time series chip dataset. The most useful features are spread across features derived from 2D Morlet wavelets, scintillation index, and the simple patch mean. The most useful time series features are a scintillation index metric, a Morlet wavelet, then a few of the gammatone band power metrics.



(a)



(b)

Figure 6: Confusions matrices for seabed texture classification based on analysis of imagery (a) and time series (b) are provided. While imagery outperforms time series as a data representation, there is a meaningful separation possible between the classes for imagery.

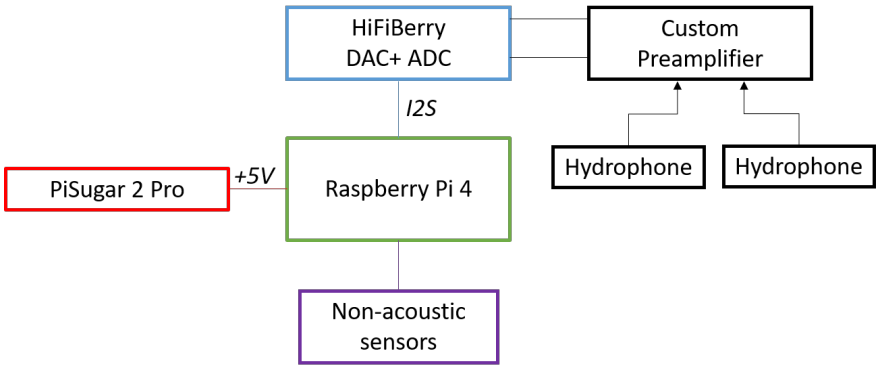


Figure 7: The prototype hydrophone system is built around a Raspberry Pi platform.The Pi Sugar 2 Pro, a commercially available battery and charging solution, provides power.A HiFiBerry DAC+ ADC is used for two channels of analog to digital conversion.Hydrophones and the preamplifier are custom designs for the system.Non-acoustic sensors (such as thermocouples or GPS antennas) may be integrated digitally with the Raspberry Pi in order to customize the sensor for a particular scientific application.

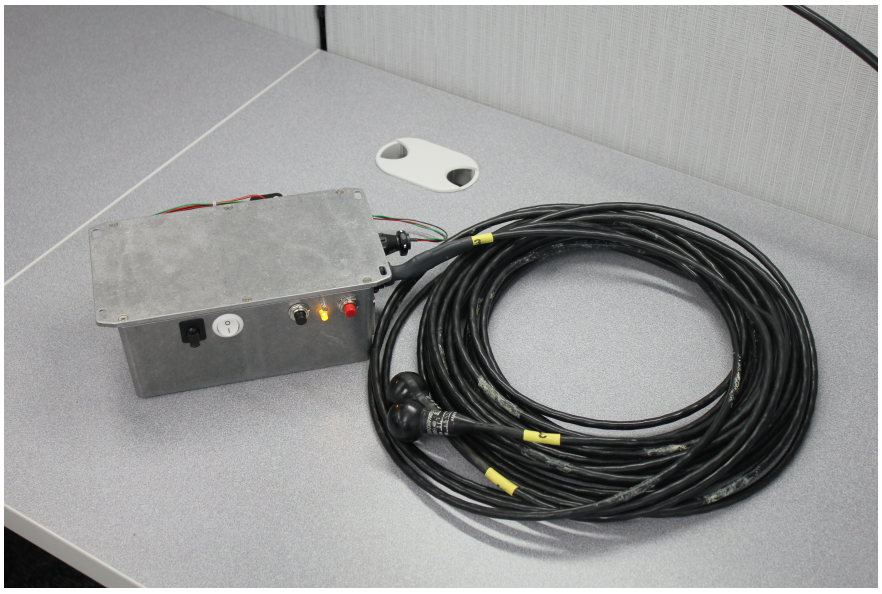


Figure 8: All of the prototype electronics are assembled into an aluminum housing.Switches and LED indicator lights allow the user to interact with the device to start and stop measurements.

3 Dissemination of Results

The results from this effort are being presented at the upcoming Acoustical Society of America meeting in a presentation titled “Aural based scene understanding for sonar applications”.

4 Technology Transfer

We have held numerous discussions with Dr. John DiCecco at the Naval Undersea Warfare - Newport regarding the structure of signals and how they inform information extraction.

References

[1] B. T. Reinhardt, I. D. Gerg, D. C. Brown, and J. D. Park, "Measuring human assessed complexity in synthetic aperture sonar imagery using the Elo rating system," in *Proc. Institute of Acoustics*, vol. 20, 2018, pp. 227–234.

[2] Ø. Midtgaard, I. Alm, T. Sæbø, M. Geilhufe, and R. E. Hansen, "Performances assessment tool for AUV based minehunting," in *Proc. of Institute of Acoustics*, vol. 36, no. 1, 2014.

[3] B. Gips and D. P. Williams, "Through-the-sensor performance estimation of the Mondrian detection algorithm in sonar imagery," in *MTS/IEEE OCEANS*, Oct 2018, pp. 1–8.

[4] D. P. Williams, "The new muesli complexity metric for mine-hunting difficulty in sonar images," in *MTS/IEEE OCEANS*, 2018, pp. 1–7.

[5] J. S. Stroud, D. Brown, D. Cook, J. Fernandez, K. Commander, D. Kolesar, and T. Montgomery, "Using a dual-band synthetic aperture sonar for imaging various seafloor compositions," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 3143–3143, 2006.

[6] V. W. Young and P. C. Hines, "Perception-based automatic classification of impulsive-source active sonar echoes," *J. Acoust. Soc. Am*, vol. 122, no. 3, pp. 1502–1517, 2007.

[7] D. D. Sternlicht, J. E. Fernandez, R. Holtzapple, D. P. Kucik, T. C. Montgomery, and C. M. Loeffler, "Advanced sonar technologies for autonomous mine countermeasures," in *MTS/IEEE OCEANS Conf.*, Sept 2011, pp. 1–5.

[8] T. G-Michael, B. Marchand, J. D. Tucker, T. M. Marston, D. D. Sternlicht, and M. R. Azimi-Sadjadi, "Image-based automated change detection for synthetic aperture sonar by multistage coregistration and canonical correlation analysis," *IEEE J. Oceanic Eng.*, vol. 41, no. 3, pp. 592–612, July 2016.

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

[10] A. Venkitaraman, A. Adiga, and C. S. Seelamantula, "Auditory-motivated gammatone wavelet transform," *Signal Processing*, vol. 94, pp. 608–619, 2014.

[11] I. W. Selesnick, "Wavelet transform with tunable Q-factor," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3560–3575, 2011.

[12] C. O. Sakar, G. Serbes, A. Gunduz, H. C. Tunc, H. Nizam, B. E. Sakar, M. Tutuncu, T. Aydin, M. E. Isenkul, and H. Apaydin, "A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable Q-factor wavelet transform," *Applied Soft Computing*, vol. 74, pp. 255–263, 2019.

[13] G. Cai, X. Chen, and Z. He, "Sparsity-enabled signal decomposition using tunable q-factor wavelet transform for fault feature extraction of gearbox," *Mechanical Systems and Signal Processing*, vol. 41, no. 1-2, pp. 34–53, 2013.