Sound Signal Classification in the New England Mud Patch

Allison N. Earnhardt, Tracianne B. Neilsen, Mason C. Acree, et al.

Citation: Proc. Mtgs. Acoust. **43**, 070001 (2021); doi: 10.1121/2.0001475

View online: https://doi.org/10.1121/2.0001475

View Table of Contents: https://asa.scitation.org/toc/pma/43/1

Published by the Acoustical Society of America

POMA Proceedings of Meetings on Acoustics

Volume 43 http://acousticalsociety.org/

180th Meeting of the Acoustical Society of America

Acoustics in Focus

8-10 June 2021

Underwater Acoustics: Paper 3aUW4

Sound Signal Classification in the New England Mud Patch

Allison N. Earnhardt
Department of Science, Purdue University College of Science, West Lafayette, IN, 47907; allison.earnhardt@gmail.com

Tracianne B. Neilsen and Mason C. Acree
Department of Physics and Astronomy, Brigham Young University, Provo, UT, 84602; tbn@byu.edu;tbnbyu@gmail.com; mason7acree@gmail.com

William S. Hodgkiss
University of California San Diego Scripps Institution of Oceanography, La Jolla, CA; whodgkiss@ucsd.edu

D. P. Knobles
Knobles Scientific and Analysis, Austin, TX; dpknobles@kphysics.org

In ocean acoustics, finding acoustic signals within long recordings is usually time consuming. In order to optimize this process, this paper explores signal classification using two deep learning models. These models are designed to classify various sources from single-sensor, 60 second time-averaged spectral density levels. The training and testing datasets were taken from 32 channels (on two VLAs) during the Seabed Characterization Experiment 2017 in the New England Mud Patch. A balanced dataset consisted of randomly selected data samples for each of the three classes: 'Tonals', 'Chirps',and 'Other'. A two-layer linear model and a four-layer one-dimensional convolutional neural network (CNN) were trained and then tested on data samples from different times. While the linear model achieved above 90% accuracy on the testing samples, the CNN had higher than 98% accuracy. This work shows the potential for deep machine learning algorithms to help identify underwater sound sources, when different signals are present in long audio files. The results of these tests imply that time averaging spectrograms have potential to improve signal classification.

Published by the Acoustical Society of America

https://doi.org/10.1121/2.0001475

1. INTRODUCTION

For decades, advanced signal processing techniques have been used for signal classification. More recently, machine and deep learning models have been applied to underwater sound signals to make use of their pattern recognition in order to identify the sound signal sources. Machine learning is a type of artificial intelligence which includes systems that can learn from data, identify patterns, and make decisions with minimal human interference. Deep learning is a subset of machine learning which uses neural networks with more than two layers.

The application of deep learning in Underwater signal classification is complicated by the effects of temperature, pressure, and salinity on the speed and refraction of sound in the water, as well as the distortive effects of sediment all of which make up the environment. As a result, noises can sound different in different environments at different times and different depths. In addition, good, sufficiently labeled data can be difficult to acquire, and machine models can require expert calibration to recognize signal sources that are more distant or changeable in time.

The work discussed in this paper applies deep learning algorithms to underwater signals in an attempt to find the optimal conditions under which to classify the signals. The data used were collected during the Seabed Characterization Experiment (SBCEX) 2017 on two vertical line arrays with 16 hydrophones each (Wilson, 2017). During this experiment different kinds of signals were transmitted and the work discussed in this paper uses a CNN to classify the signals from the 32 receivers individually over a several day period in the New England Mud Patch.

2. BACKGROUND

Krizhevsky *et al.* (2017) marked a turning point in using machine learning models to identify large scale visual inputs in an attempt to counteract the issue of lack of labeled data. Interest has also expanded to machine learning being applicable to identify auditory inputs. Many techniques learned from image classification have been applied to audio classification using spectrograms and mel-spectrogram images. Piczak (2015) was able to show that CNNs could be used to identify audio clips translated into low level representation (spectrograms). Salamon and Bello (2017) also applied deep learning to the issue of identifying environmental noise while attempting to mitigate the issue of lack of labeled data by augmenting the audio signals. However, one issue remains in attempting to analyze spectrograms in methods similar to images since spectrograms do not maintain the same translation invariance as images and different frequency bands can belong to different classes (Wyse, 2017). One method of overcoming this difficulty is to present a CNN with heterogeneous pooling. Previous works have used methods such as second-order pooling to keep the higher-level features of the data in the network and constant-Q transforms, which might provide better resolution, rather than Short Term Fourier Transforms to analyze the data (Cal, 2019).

As the success of these experiments shows, interests now turn towards identification of underwater sound signal source classification. Underwater environments especially can be difficult to classify using automated processes due to differing signal-to-noise ratio, the large effect seabed and water characteristics can have, as well as the often non-stationary and impulsive tendencies of naturally present underwater acoustical signals. One option according to Kamal *et al.* (2013) is to use unsupervised CNN models in order to allow the model to be less rigid and adapt to the indistinct and variant features of the underwater environment without needing expert calibration. This unsupervised learning method has been used with success by Ozanich (2021); they were able to identify whale and fish noises from a loud coral reef. This work uses supervised machine learning models to classify man-made sounds.

3. EXPERIMENTAL SETUP

The data used were collected, with a sampling frequency of 25000 Hz, during Seabed Characterization Experiment 2017 (Wilson, 2017) from two vertical line arrays of 16 hydrophones each set in the ocean in the New England mud patch by Dr. William S. Hodgkiss and his crew from Marine Physical Laboratories at the Scripps Institute of Oceanography. Around this area, multiple tracks were assigned over which a boat traveled with a submerged loudspeaker which would play either chirps or tonal noises with a 50% duty cycle at intervals of ten seconds. The 'Chirps' played through an octave of frequencies in a second, repeating ten times, before pausing for ten seconds and repeating. 'Tonals' played five constant frequencies, 1.5, 2, 2.5, 3, 3.5, and 4 kHz continuously for ten seconds before pausing ten seconds and repeating. This controlled experiment allowed for more control over the types of sound signals the network classified, reducing the effect of clutter and the effects of the seabed.

The hydrophone recordings of this signal were split into 60 second segments, and those segments were converted to NumPy files. The background noise of these recordings was not removed, in part because the ultimate goal is to apply this model directly to spectrogram data and receive an output of time stamps when signals occur, but also because the goal is to eventually build a model that will also recognize more transient underwater signals such as ships, whales, and SUS charges in the known environment.

These NumPy files were then loaded, and spectrograms were computed with a sample block size of 2^13. The spectrograms were labeled based upon the time during which they occurred by referencing a data log, written during the data collection, of when various signals were being played. The three labels used were ‘Other’, ‘Tonals’ and ‘Chirps’; an example of each is shown in Figure 1.

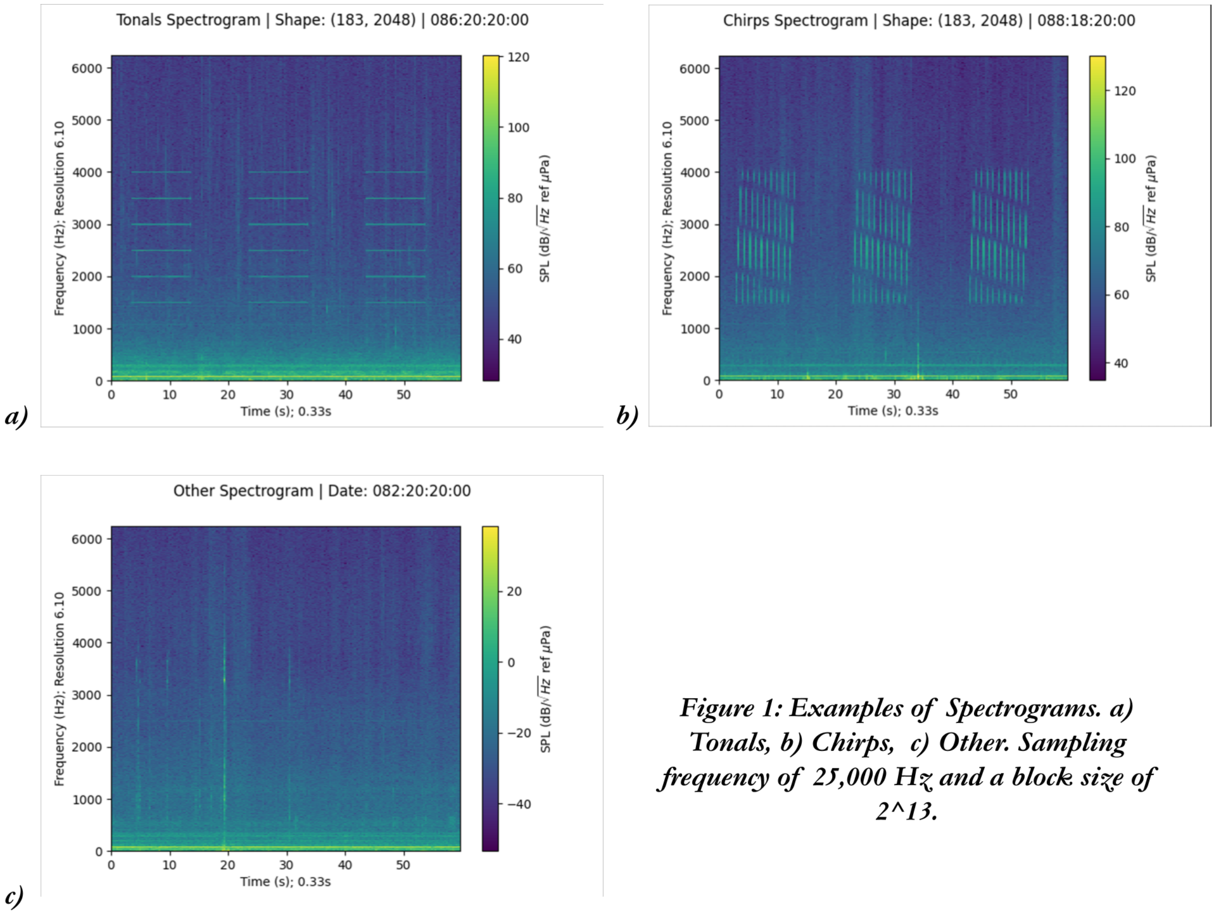


Figure 1: Examples of Spectrograms. a) Tonals, b) Chirps, c) Other. Sampling frequency of 25,000 Hz and a block size of 2^13.

The 60 second spectrograms were then concatenated into longer spectrograms based on the desired time intervals for the training and testing data, shown in Table 3 in the Appendix. These longer spectrograms were saved as HDF5 files along with the time and frequency arrays and the date of the start of the spectrogram time range. This data format allowed greater control over the length of the spectrograms used for training and testing the neural network without increasing the amount of memory or time needed and also reduced the amount of edge effects which might alter the data. Another benefit of the HDF5 format is that each channel can be loaded individually.

The data used to train the network using the HDF5 file format was carefully chosen so that each label, 'Chirps', 'Tonals', and 'Other', had 270 minutes of spectrogram data collected from different times and days, shown in Table 3 in the Appendix. The data were then sliced to produce 60 second spectrograms, with each channel starting at a random time. Each channel was sampled independently, resulting in 32 individual samples for any given time Using spectrogram segments taken from a longer spectrogram allowed for overlap between the data samples, but the random start times also meant there was no control over how evenly the selected times were spread over the total time interval. The time stamp of each data sample was recorded so that the correct label could be identified.

In order to ensure that the proper number of data samples were loaded from each file during training and testing, a dictionary was created that linked the different HDF5 files to the overall time included in them and thus the number of data samples that should be extracted. While this did not guarantee that the entirety of the spectrogram would be covered evenly, it made sure that each spectrogram would be sampled an appropriate number of times given its length.

For the machine learning models discussed in this paper, the author chose to time average the spectrogram inputs for the model. The time-averaging eliminates any issues which might have been posed by the 50% duty cycle when combined with the random start times. Examples of time-averaged spectrograms can be seen in Figure 2, each at the same time as the spectrogram examples in Figure 1. The time-average samples contain distinctive spectral features corresponding to the 'Tonals', 'Chirps', and 'Other' categories.

4. NEURAL NETWORKS

This work used two separate machine learning models in this work in the PyTorch framework. The first machine learning model (illustrated on the left in Figure 3) was a fully connected linear model with two linear layers of 1024 nodes and an output to the same three classes. This second model was built to see how simple a model could successfully categorize the different signals. The second was a four-layer one-dimensional convolutional neural network (illustrated on the right in Figure 3) with a kernel size of 3, padding and stride of 1, an input feature of 1024 and a channel size which changed from 1 to 16 to 32 to 64 to 128. Then two linear layers with 1024 nodes each followed by the output into three different classes, 'Chirps', 'Tonals', and 'Other'.

The networks were trained on the data samples from the times listed in Table 3 in the Appendix and then tested on data from times listed in Table 4. Each data sample was scaled by its individual mean (in Pascals) before being used in the neural network. The training dataset was equally balanced with 270 samples of each type of data for 810 samples per channel, resulting in 25,920 total training samples.

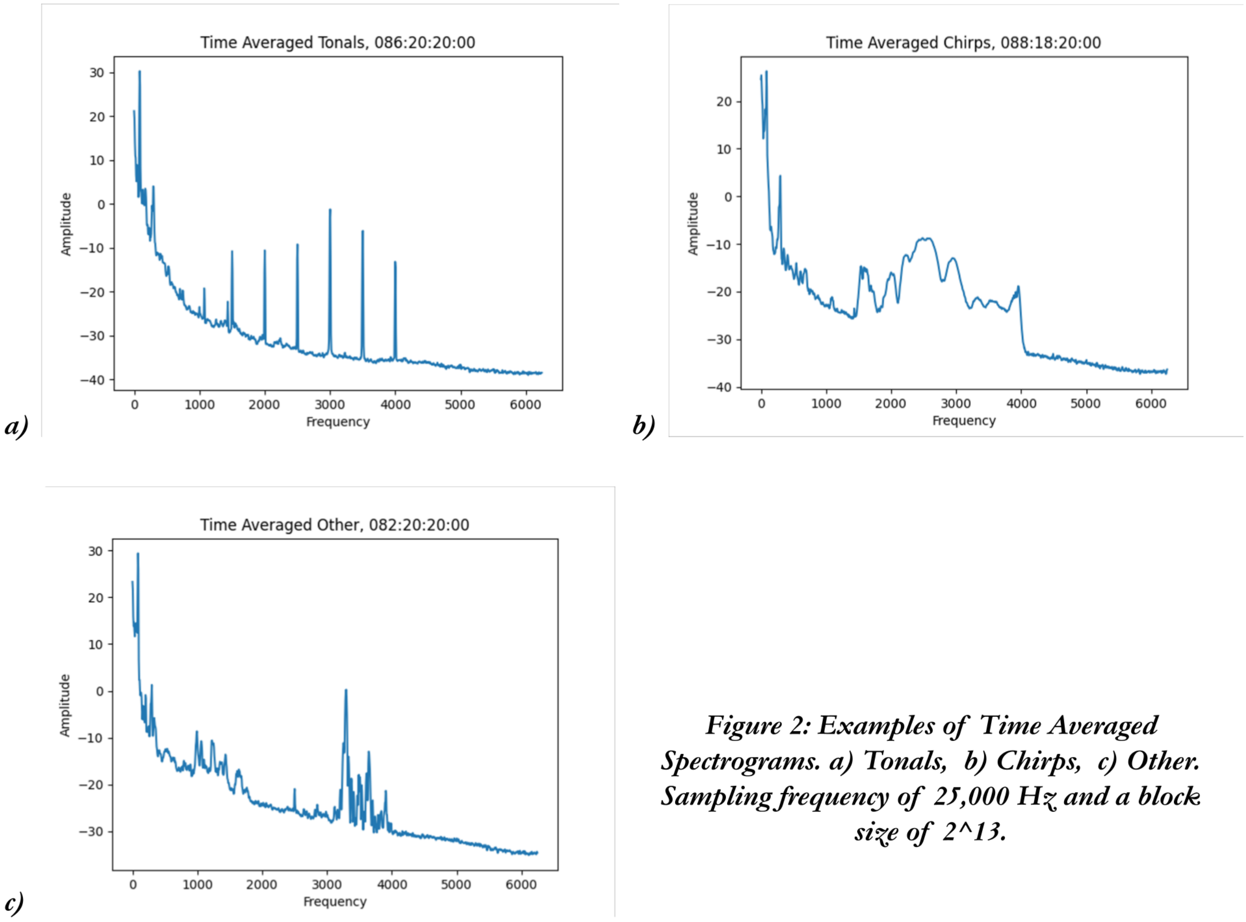
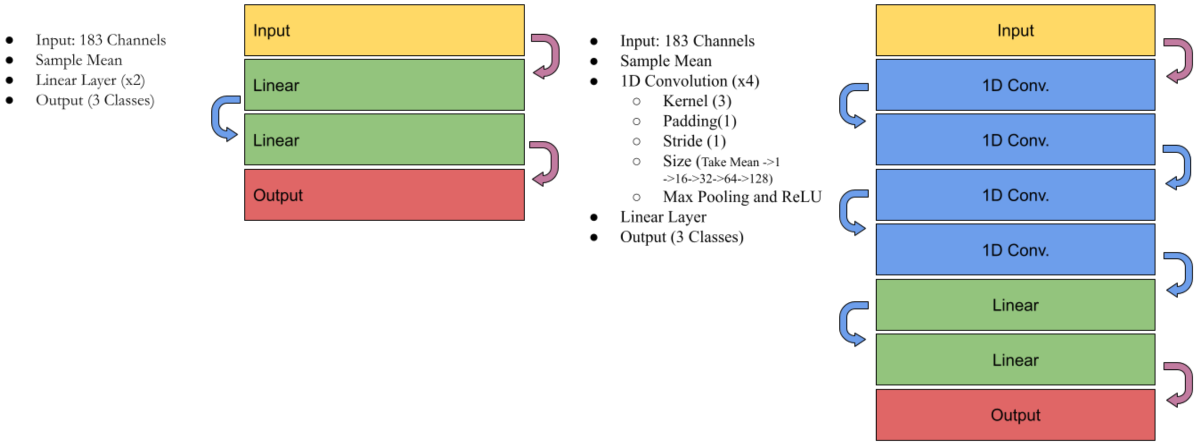


Figure 2: Examples of Time Averaged Spectrograms. a) Tonals, b) Chirps, c) Other. Sampling frequency of 25,000 Hz and a block size of 2^13.



- Input: 183 Channels
- Sample Mean
- Linear Layer (x2)
- Output (3 Classes)

- Input: 183 Channels
- Sample Mean
- 1D Convolution (x4)
  - Kernel (3)
  - Padding(1)
  - Stride (1)
  - Size (Take Mean ->1 ->16->32->64->128)
  - Max Pooling and ReLU
- Linear Layer
- Output (3 Classes)

Figure 3: Diagram of the two neural networks used to classify the sound signal sources: (left) a linear model and (right) a four-layer convolutional neural network.

5. RESULTS

The training and testing results for those two machine learning models can be seen in Tables 1 and 2. They were trained, saved, and then applied to the separate testing dataset. The upper part of each table contains the resulting testing and training accuracy for each class, the false positive, the count for how many samples were incorrectly attributed to each class; the false negatives, the number of samples of each class incorrectly labeled as a different class; and the F1 score, which is the harmonic mean of the precision and the recall, intended to give a measure of the incorrectly classified cases.

On the lower portion of each table, the overall accuracy of both the four-layer one-dimensional CNN and the linear model is displayed. The total accuracy is the accuracy for all of the samples as a whole. Because a different number of testing samples was used for each label, there is a second calculated weighted accuracy which calculates the final accuracy by averaging the testing accuracy of each class. As shown by these overall accuracies, both models achieved an accuracy of above 90%, though the four-layer CNN achieved a higher accuracy of approximately 99%.

Table 1: Training and testing results for the time-averaged input linear model.

Average Linear Layers	Training Accuracy (%)	Testing Accuracy (%)	False Positive Count	False Negative Count	F1 Score
Tonals	61 (N=8639)	79 (N=7205)	35	1529	0.880574
Chirps	96 (N= 8641)	99 (N=10513)	78	25	0.995144
Other	99 (N=8640)	98 (N=4768)	1503	63	0.857247

Total Accuracy	Weighted Accuracy
93%	92.49%

Table 2: Training and testing results for the time-averaged input four-layer one-dimensional CNN model.

Average CNN	Training Accuracy (%)	Testing Accuracy (%)	False Positive Count	False Negative Count	F1 Score
Tonals	99 (N=8639)	99 (N=7205)	51	12	0.995693
Chirps	99 (N= 8641)	99 (N=10513)	15	66	0.996138
Other	98 (N=8640)	98 (N=4768)	62	50	0.98827

Total Accuracy	Weighted Accuracy
99%	99.39%

Another method of evaluating the accuracy is by showing the classification from each model for each channel in time to show consistency over the channels which were at different depths. To illustrate consistency over channels, this work chose a two-hour time block that included all three classes. Figure 4 displays the ideal channel prediction with the channels (on VLA1) in the vertical axis and the time, in this case 120 minutes, as the horizontal axis; the time interval is listed in the figure title in (year-day-hour-minute-second) format. The grey represents times corresponding to the label 'Chirps', the orange for 'Other', and the pink for 'Tonals', while the background was set to green to account for any skipped time (i.e., time not included in the random sampling described in Section 3, which is different for each set of predictions). Figure 5 shows the predictions for the linear model. For the first 30 minutes of the time shown, the predictions for all of the channels agree and are correct; However, starting around minute thirty, some of the channels begin to be confused and classified as 'Other', while missing the real block of others before the 60 minute mark. This confusion continues into the 'Tonals', although many of the channels classify 'Tonals' correctly for much of the time. In Figure 6, the predictions for each channel in time are shown for the four-layer one-dimensional convolutional neural network. Accuracy is greater and agreement between the channels is more consistent for this model. Only a few of the 'Chirps' are predicted to be 'Other', and the section of 'Other' signals before the 60 minute mark were incorrectly labeled as 'Tonals'. Overall, though some channels sometimes incorrectly predicted 'Other' and missed the real 'Other' labels, Figures 5 and 6 show that the independent predictions for each channel tended to choose the correct class.

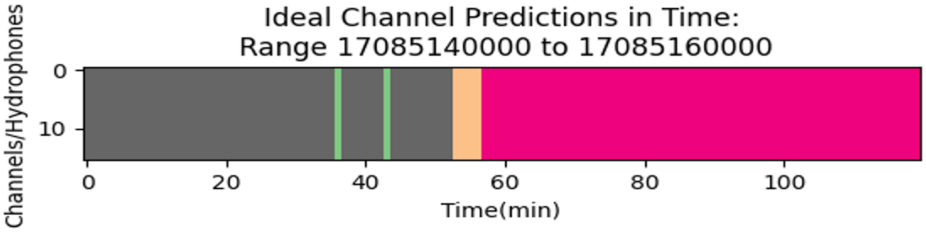


***Figure 4:** Ideal channel output based upon true classification at each time. Plot of all 16 channel predictions in time, orange 'Other', grey is 'Chirps', and pink is 'Tonals', background set to green to account for gaps in time intervals viewed.*

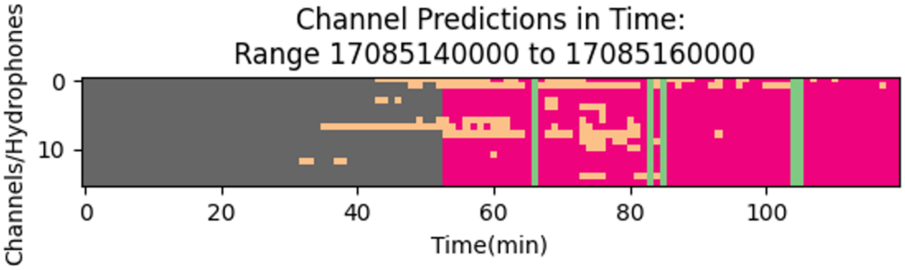


***Figure 5:** Channel output from linear model in time. Similar to Figure 4.*

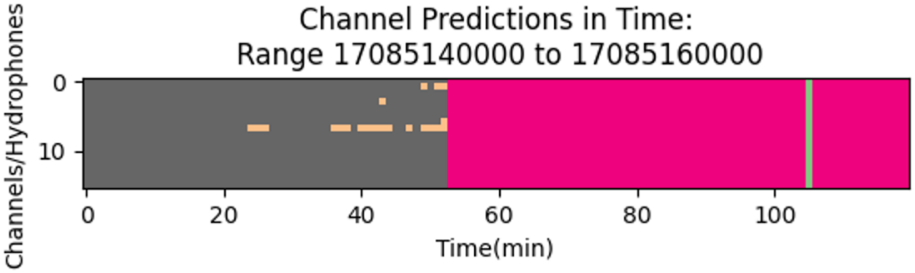


***Figure 6:** Channel output from four-layer one-dimensional CNN in time. Similar to Figure 4.*

6. CONCLUSIONS

One question regarding the application of machine learning techniques to ocean acoustics is the best way to classify signals. The work discussed in this paper used the time average of 60 second spectrograms as the input. These input samples come from the 32 channels on the two VLAs deployed by Marine Physical Laboratories during the Seabed Characterization Experiment 2017. Samples were labeled according to the recording times based on whether or not controlled sources were deployed. The resulting labels were 'Tonals', 'Chirps', and 'Other'. A balanced training dataset was used to train a linear model and a convolutional neural network, which was then tested on samples from times not used during training.

The four-layer one-dimensional convolutional neural network performed better with a signal classification accuracy of greater than 98%. This accuracy is notable in several ways. First, the training and testing data come from different days and when the source was in different locations around the New England Mud Patch, where different thicknesses of mud are present, and each individual channel produced similar classifications (Wilson, 2020). Second, the training and testing data from individual sensors were located at different depths in the water. The success in using 32 individual channels has not been shown previously.

The ideal setting of this study provides insights into areas that require future work. First, the success in using 32 channels individually in training and testing may be tied to the fairly static, almost isovelocity sound speed profile during the time of this experiment (Wilson, 2017). Using individual receivers at multiple depths may not be effective when the sound speed profile varies significantly with depth. In such cases, using data from all channels on the VLA as a single input sample will likely be more effective. Studies should then be done to see if the resulting data (number of channels by number of frequencies) should be input to a one-dimensional or two-dimensional CNN or if a different type of network is better suited to handle the variation in depth. Second, the 'Tonals' and 'Chirps' classes used in this work had distinctive spectral features that were well above the background noise level. Questions remain as to if this time-averaging could work for quieter sources or time-varying sources. Overall, this paper finds great potential in using machine and deep learning for signal classification provided the training data contain the needed variability to represent the dynamic ocean environment.

ACKNOWLEDGMENTS

The first author was supported by NSF grant #1757998 for the Research Experience for Undergraduates hosted virtually by Brigham Young University in 2020. The authors are grateful that the Office of Naval Research funded the data collection efforts of Seabed Characterization Experiment 2017.

APPENDIX A

Table 3: Testing dataset

Start Time (yydddhhmmss)	End Time	Label	Total Time (min)
17086120200	17086142000	Tonals	138
17085180000	17085193000	Tonals	90
17088170500	17088213400	Chirps	269
17089204000	170892014000	Chirps	60
17084013000	17084033000	Other	120
17085220400	17085223300	Other	29

Table 4: Training Dataset

Start Time (yydddhhmmss)	End Time	Label	Total Time (min)
17085150000	17085174400	Tonals	164
17084121600	17084140200	Tonals	106
17084140900	17084163500	Chirps	146
17084163500	17084172400	Chirps	49
17084172400	17084182300	Chirps	59
17088143000	17088144600	Chirps	16
17082200000	17082220000	Other	120
17083230000	17084013000	Other	150

REFERENCES

Altes, R. A. , "Detection, Estimation and Classification using Spectrograms," *The Journal of the Acoustical Society of America,*67(4), 1232. 1979.

Cal, X. , Togneri, R., Zhang, X., and Yu, Y. “Convolutional Neural Network With Second-Order Pooling for Underwater Target Classification,” *IEEE Sensors Journal*, vol. 19, no. 8, April 2019.

Ghosh, J. , Beck, S. D. and Chu, C. "Evidence combination techniques for robust classification of short-duration oceanic signals", Proc. SPIE 1706, Adaptive and Learning Systems, (20 August 1992); https://doi-org.ezproxy.lib.purdue.edu/10.1117/12.139951

Ghosh, J. L. Deuser and Beck, S. D. "A neural network based hybrid system for detection, characterization, and classification of short-duration oceanic signals," in *IEEE Journal of Oceanic Engineering*, vol. 17, no. 4, pp. 351-363, Oct. 1992, doi: 10.1109/48.180304.

Kamal, S., Mohammed, S., and Pillai, P.R.S. S. M.H. “Deep Learning Architectures for Underwater Target Recognition,” Department of Electronics, Cochin University of Science and Technology, India 2013

Krizhevsky, A., I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*. vol. 60, no. 6, June 2017.

Ozanich, E. and Thode, A. and Gerstoft, P. and Freeman, L. A. and Freeman, S. "Deep embedded clustering of coral reef bioacoustics" The Journal of the Acoustical Society of America. Vol. 149, no.4. Doi: 10.1121/10.0004221

Piczak K. J., "Environmental sound classification with convolutional neural networks," *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, 2015, pp. 1-6, doi: 10.1109/MLSP.2015.7324337.

Salamon, J. and Bello, J. P. "Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification," in *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279-283, March 2017, doi: 10.1109/LSP.2017.2657381.

Seabed Characterization Experiment 2017 R/V Endeavor Log 2 21 March – 04 April 2017 W.S. Hodgkiss *Marine Physical Laboratory* Scripps Institution of Oceanography. Unpublished Notebook.

Smith, L. N. "Cyclical Learning Rates for Training Neural Networks," U.S. Naval Research Laboratory, Washington 2015

Srivastava,N. , Hinton, G., Krizhevsky, A., Sutskever, I.,and Salakhutdinov, R. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting" Department of Computer Science, University of Toronto, June 2014.

Wilson, P. S. , Knobles, D. P., and Neilsen, T. B., "Guest editorial an overview of the seabed characterization experiment," 439 IEEE Journal of Oceanic Engineering, vol. 45, no. 1, pp. 1–13, 2020

Wyse L. , "Audio Spectrogram Representation for Processing with Convolutional Neural Networks," *First International Workshop on Deep Learning and Music joint with IJCNN*, May 2017