NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

APPLICATION OF MACHINE LEARNING TECHNIQUES TO IDENTIFY FORAGING CALLS OF BALEEN WHALES

by

Michelle Tanalega

June 2018

Thesis Advisor: John E. Joseph
Co-Advisor: Tetyana Margolina

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** June 2018	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE APPLICATION OF MACHINE LEARNING TECHNIQUES TO IDENTIFY FORAGING CALLS OF BALEEN WHALES	**5. FUNDING NUMBERS**
6. AUTHOR(S) Michelle Tanalega	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

An unsupervised machine learning algorithm has been applied to passive acoustic monitoring datasets to detect and classify foraging calls of blue whales, *Balaenoptera musculus*, and fin whales, *Balaenoptera physalus*. This approach involves using a k-means clustering algorithm to cluster data based on common features, which produces a number of specified centroids. The centroids are then compared to machine-selected candidates for classification. Once divided into initial clusters, further clustering is done to fine-tune results. Preliminary testing of the algorithm yielded promising results. The cross-validation method and the DCLDE 2015 scoring tool were used to estimate out-of-sample performance of the detection algorithm. The automated detector/identifier has been applied to data collected during different seasons, and its performance was analyzed for various types of noise present in data, signal-to-noise ratios, and acoustic environment. The advantages of this approach over traditional manual scanning are increased reliable performance, and time and cost efficiency. This approach could potentially be a faster method of sorting and classifying large acoustic data sets.

14. SUBJECT TERMS machine learning, k-means, baleen whales, foraging calls	**15. NUMBER OF PAGES** 77
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

APPLICATION OF MACHINE LEARNING TECHNIQUES TO IDENTIFY FORAGING CALLS OF BALEEN WHALES

Michelle Tanalega
Lieutenant, United States Navy
BS, U.S. Naval Academy, 2012

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN PHYSICAL OCEANOGRAPHY

from the

NAVAL POSTGRADUATE SCHOOL
June 2018

Approved by: John E. Joseph
Advisor

Tetyana Margolina
Co-Advisor

Peter C. Chu
Chair, Department of Oceanography

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

An unsupervised machine learning algorithm has been applied to passive acoustic monitoring datasets to detect and classify foraging calls of blue whales, *Balaenoptera musculus,* and fin whales, *Balaenoptera physalus.* This approach involves using a k-means clustering algorithm to cluster data based on common features, which produces a number of specified centroids. The centroids are then compared to machine-selected candidates for classification. Once divided into initial clusters, further clustering is done to fine-tune results. Preliminary testing of the algorithm yielded promising results. The cross-validation method and the DCLDE 2015 scoring tool were used to estimate out-of-sample performance of the detection algorithm. The automated detector/identifier has been applied to data collected during different seasons, and its performance was analyzed for various types of noise present in data, signal-to-noise ratios, and acoustic environment. The advantages of this approach over traditional manual scanning are increased reliable performance, and time and cost efficiency. This approach could potentially be a faster method of sorting and classifying large acoustic data sets.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

BRS	Behavioral Response Study
CINM	Channel Islands National Marine Sanctuary
DCLDE	Detection, Classification, Localization and Density Estimation
DCPP	Diablo Canyon Power Plant
FFT	Fast Fourier Transforms
GUI	Graphical User Interface
HARP	High-Frequency Acoustic Recording Package
LTSA	Long Term Spectral Average
PAM	Passive Acoustic Modeling
RL	Receive Level, dB in re 1 µPa
SCB	Southern California Bight
SIO	Scripps Institution of Oceanography
SL	Source Level, dB in re 1 µPa
SOCAL	Southern California
TL	Transmission Loss, dB in re 1 µPa

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I thank my advisors, Professor John Joseph and Dr. Tetyana Margolina, for their patience and guidance throughout this process. I learned a great deal from them during the course of this thesis. I thank my husband for his love and support during the late nights of writing and studying. I thank my cohort and friends for being good teammates and helping me get through some tough classes. I also thank my parents and my brother for their love and encouragement and for being my cheerleaders every day.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Acoustic monitoring marine mammal populations plays an important role in understanding species distribution and conservation efforts. The majority of detection methods rely on human operators to detect vocalizations to infer population densities, migratory patterns, and changes in behavior. Monitoring vocalizations requires processing large amounts of data. In order to decrease the process time of large data sets, auto-detectors have been developed. However, auto-detection in marine bioacoustics is a challenging problem. Distinguishing between different species still presents an issue to researchers. This research presents an unsupervised machine learning technique to classify blue and fin whale calls.

A. NAVAL RELEVANCE

The military has a major presence in the Southern California Bight (SCB) with five naval bases in San Diego. The Navy is required by law to protect marine mammals and their habitats to comply with the Endangered Species Act (ESA) and Marine Mammal Protection Act (MMPA). Precautions such as limiting sonar operations and altering ship's track to avoid marine mammal collisions are taken routinely. Recently, the Pacific Fleet agreed to new operating limitations for active sonar as a direct result of a federal court settlement between the Natural Resources Defense Council and the National Marine Fisheries Services (2015). However, hundreds of different types of platforms conduct a variety of naval and joint activities in the same waterways traveled by blue and fin whales on a daily basis. The development of a reliable auto-classification method would allow for a more efficient conservation effort. Lower costs, less manpower, and more objective results allowing for larger monitored areas would enhance the study of species distribution and aid the Navy in tailoring training and routine operations in the SCB.

As the methods improve, this research could also be applied to the undersea warfare domain, aiding sonar technicians in the classification of challenging signals of interest by reducing the uncertainty and subjectivity introduced by the human factors that affect

operator performance. Additionally, this can be applied to auto-detection algorithms used on unmanned systems to minimize false alarm rates.

B. BLUE WHALES AND FIN WHALES

The two largest cetacean species are blue whales *Balaenoptera musculus* and fin whales *Balaenoptera physalus*. Due to their size, these species were hunted for blubber, oil, and baleen in the 20th century. While blue and fin whales can be found all over the world, this paper will be discussing the populations observed in the SCB and off the coast of Central California. The SCB is the coastline from San Diego to Point Conception where strong upwelling occurs creating a highly productive ecosystem (Smith & Eppley 1982) which attracts blue and fin whales to the area (Barlow & Forney 2007). This region is also very popular for recreation, economic and military utilization. These activities pose threats to all marine mammals in the area. Evaluating these threats to blue whales and fin whales is of particular interest since they have been on the endangered species list since 1970 under the Endangered Species Conservation Act, the predecessor to the Endangered Species Act.

Elevated levels of human activity in a marine mammal habitat can degrade a species' ability to conduct daily life functions or they can cause death due to ship strikes (Carretta et. al 2015, National Research Council, 2003, 2005). Commercial threats consist mainly of ship strikes and fishing mishaps (Rockwood et al. 2017, Berman-Kowalewski et al. 2010, Pace et al. 2014). Commercial shipping noise is another source of sound that affects marine mammal behavior (Melcón et al. 2012). Melcón also notes military operations conducted in this area, mainly training exercises involving active sonar, can affect feeding behaviors.

From previous studies and observations, general patterns of blue and fin whales populations in the SCB have been determined. Blue whales are more migratory than fin whales. Foraging occurs in the SCB from the beginning of June until the end of November when blue whales migrate equatorward towards warmer water for the mating season (Burtenshaw et al. 2004). Although estimating population density remains a challenge, two studies estimate that there is a higher density of blue whales closer inland during feeding season (Barlow & Forney 2007, Becker et al. 2010). More specifically, larger numbers of

blue whales are visually observed in the Santa Barbara Channel (Fiedler et al. 1998). Acoustical studies conducted by Širovic´ and colleagues (2015) also showed a high number of blue whale vocalizations in this area.

Blue whales produce several different sounds for a wide range of activities. Pulsed A-calls and tonal B-calls are produced only by males and have distinguishable characters. Blue whales also produce another gender-neutral call during foraging named the D-call (Wiggins et al. 2005; Oleson et al. 2007a). This research focuses on the highly variable D-call, documented with a maximum pulse length of 4 seconds over frequency ranges of 25-90 Hz (Thompson et al. 1996) and 45-95 Hz (Madhusudhana et al. 2009). The main method used to find these calls is visual scanning, which was also used in this research and will be discussed later as part of data preparation.

In contrast, fin whales inhabit the SCB throughout the year. Several studies show fin whale populations present during all seasons but there are larger numbers at the end of summer and the beginning of fall (Dohl et al. 1980, 1983; Carretta et al. 1995; Barlow 1995; Forney & Barlow 1998, Oleson 2005, Širovic´ et al. 2013). Fin whales also prefer the southern portion of the SCB according to an acoustic survey study (Širovic´ et al. 2015). This study along with visual surveys conducted by Jefferson et al. (2014) shows fin whales concentrate off Redondo Beach and to the west of San Clemente Island, the location of the Navy's Southern California Anti-submarine warfare Range (SOAR).

Fin whales can be detected by two different vocalizations. The 20-Hz call and the 40-Hz call, each named for its center frequency. The 20-Hz call has been detected during all seasons in the SCB, but is more frequent during the winter mating season (Oleson 2005). However, it is only produced by males possible for mating purposes (Croll et al. 2002). This research focuses on the 40-Hz call, a gender-neutral call used as a part of feeding behaviors. These calls are generally down sweeping with a frequency range between 62-48 Hz (Širović et al. 2013).

It is important to understand long-term migratory or non-migratory patterns to most accurately assess the effects of human impact. Commercial maritime companies and military vessels must adhere to a myriad of regulations and instructions when operating

around these environments. However, operating areas and shipping lanes often overlap with marine mammal habitats. Better understanding of whale behavior will help fine tune these regulations to make the ocean a shareable space.

C. CURRENT MARINE MAMMAL OBSERVATION TECHNIQUES

There are several ways to study marine mammal behavior and distribution patterns despite the many challenges such a task presents. Whales can spend up to 95 percent of their time underwater which means ship-based visual observations would have to be compiled over several years to decades to provide any discernible patterns (Wiggins 2010). Such ship-based operations are expensive and time consuming, but methods have been developed to take advantage of the natural sounds used by whales to travel through their environment and communicate. Acoustic monitoring can provide more robust data sets than visual surveys to analyze marine mammal behavior and better predictions of how sound affects their way of life. The two main types of tools used are tags and passive acoustic monitoring (PAM) systems. For purposes of this paper, we will emphasize PAM methods and go into further detail. Autonomous acoustic recorders are used to accomplish PAM. Acoustic recorders typically are used for long periods of times unattended both animal and anthropogenic sounds, contributing to the base development of ecological and behavioral response models (Wiggins 2010).

Acoustic recorders became more capable with the development of the High frequency Acoustic Recording Package (HARP) as described by Wiggins and Hildebrand (2007). They discuss the new advantages of these packages, which include increased data capacity and data collection in broader frequency bands. They go on to explain how the large amount of data, which can be up to 12 TB/year, is processed from the bottom-mounted sensors to create spectrograms. The spectrograms cannot be evaluated in near-real time due to human and processing limitations. HARP data is largely scanned manually by trained marine mammal experts, a process that can take weeks to months.

Due to large amounts of data, the recordings are first averaged into long-term spectral averages (LTSAs) for more manageable processing. LTSAs can provide a representative time-frequency overview of the large data sets and a means of searching and

analyzing data points of interests as described by Wiggins and Hildebrand (2007). Wiggins (2010) further explains that each pixel in a LTSA represents approximately five seconds of averaged spectral data. However, manually going through long-term data sets is both expensive and time-consuming process where improvements are needed. Researchers use Triton, a MATLAB-based program developed at Scripps, for this process. Analysts scan through the LTSAs and can zoom in further on the data to create a shorter-term spectrogram to identify vocalization candidates of any variety. Classification by analysts is somewhat subjective because it depends on the characteristics of the candidate call and the contextual information the analyst observed in the LTSA such as presence of other calls or anthropogenic noise. Due to the nature of subjectivity, this means there is potential that two analysts reviewing the same data will have different results. Moreover, the same analyst on two different days might have different results.

The human factor is one issue when trying to identify D-calls and 40-Hz calls. Another issue is that the calls can exhibit similar characteristics. Both calls exhibit a general down sweeping within a variable frequency range. The calls can also be overlapping in frequency and duration. Additionally, non-animal noises can mask, break, or overlap the foraging calls making them almost impossible to detect in the LTSA setting. To help mitigate some of these issues while processing, researchers have developed automated detectors and continue to advance machine learning processes for these purposes.

D. MACHINE LEARNING APPLICATIONS

Automated detectors have been developed to reduce processing time on large data sets (Wiggins 2010). Wiggins adds that although automated detectors are useful for detail analysis, multiple algorithms might have to be applied to the same set of LTSAs to detect signals of interest among the wide range of sounds present. Previous studies have developed methods for automated detection (Mellinger 1994) and classification (Madhusudhana et al. 2010, Bahoura et al. 2012) of baleen whale vocalizations using match filters or spectrogram correlation. Classification of blue whale D-calls and fin whale 40-Hz calls was a challenge due to an insufficient volume of confirmed calls. In 2015, Scripps Institution of Oceanography (SIO) hosted the Detection, Classification, Location, and

Density Estimation (DCLDE) Workshop to provide an annotated data set focused on these two calls, which allowed for the creation of more complex algorithms to apply machine learning techniques.

Machine learning is an optimal way of scanning and processing the large data sets created by HARP. Machine learning is the concept of constructing algorithms that can learn from data presented and make data-driven predictions or decisions (Kohavi et. al 1998). More precisely, it has two goals: knowledge acquisition from external sources and the enhancement of knowledge representations to better exploit existing knowledge (Briscoe et al. 1996). Machine learning can be broadly classified into the two categories of supervised and unsupervised learning. Supervised learning involves training from data that provides input data and the expected corresponding outcome (Flach 2012). Unsupervised learning is training on data with only input vectors and the system has to find structure in its inputs without instruction or parameters (Flach 2012).

E. PREVIOUS WORK AND RESEARCH OBJECTIVES

In the application of machine learning techniques to identify marine mammal vocalizations, a common issue associated with automated detection is filtering out ambient noise sufficiently without the use of an additional detector to correctly isolate calls of interest (Wiggins 2010). This research is a continuation of previous work that developed a pattern recognition technique applying logistic regression to detect and classify blue whale D-calls and fin whale 40-Hz calls (Huang 2016). A logistic regression algorithm is a supervised learning method that develops a prediction function to classify designated objects, in this case foraging calls of fin and blue whales. This method resulted in high performance with 96% accuracy for the logistic regression classifier, 96% recall and 92% precision for pattern recognition.

While there are several approaches to accomplish unsupervised learning, this research uses k-means clustering techniques. K-means clustering, or Lloyd's algorithm, originally started in signal processing, but is now used for several applications. It is a method of vector quantization that aims to partition a number of observations, n, into a number of clusters, k, that display similar characteristics (Lloyd 1982). The number of

clusters, also called centroids, is determined before the algorithm starts and results in k clusters consisting of observations with the nearest mean dividing the data space into Voronoi cells. The observation points in a particular cell have shorter distance to a particular cluster than to any other cluster in the data space. The process is repeated until each data point is assigned to a cluster.

The advantages of using k-means clusters are that is a widely used method for cluster analysis, the process is not overly complex, and the system trains quickly (Singh et al. 2011). Singh goes on to say that some disadvantages include that is difficult to predict the optimal number of clusters and that outliers can have a large effect on the composition of the clusters that are formed.

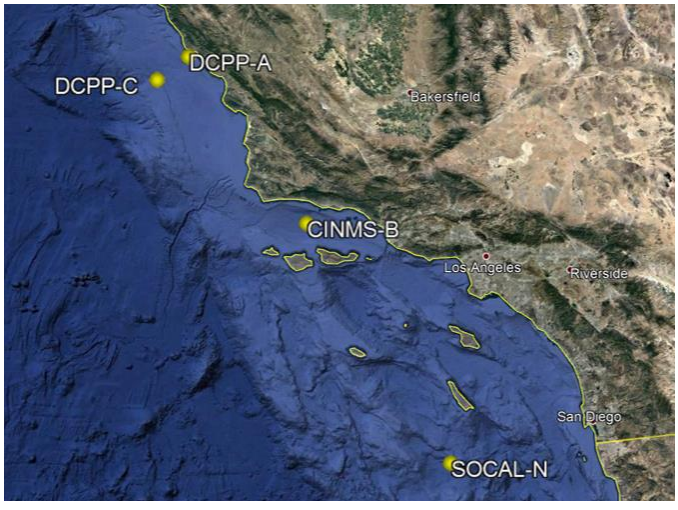
The methods of automated detection for blue and fin whale foraging calls has not yet been perfected. This research aims to contribute to the development of improved and efficient methods of distinguishing similar calls. The main focus of this research is to develop a machine learning based algorithm that can conduct unsupervised learning using clustering to detect foraging calls, D-calls and 40-Hz calls, of blue and fin whales, respectively. This research attempts construct centroids for a variety of noise environments to help classify foraging calls found in large PAM data sets in order to reduce time, money and effort that would otherwise be spent during manual scanning procedures. Specifically, efforts were directed toward four main questions. Is there enough data to train the algorithm? Is the algorithm location/season specific? Do we need to retrain if moving to a different location? What are the main obstacles in automatic detection of baleen whale calls using an unsupervised machine learning technique? Performance will be measured with the same cross-validation method as in Huang 2016.

THIS PAGE INTENTIONALLY LEFT BLANK

II. DATA

A. HARP DATA

The data in this research were provided as part of the DCLDE 2015 Workshop, which provided acoustic data sets from three HARP deployments and a scoring tool to estimate the performance of automated detectors. Two sets of recorded acoustic data were distributed to workshop participants; one for baleen whale calls and the other for toothed whale calls (SIO 2015). Only the baleen whale data set is used for this research. The calls were recorded from 2009-2013 from HARPs located in three different locations with varying depths off the central and southern California coast (Figure 1). This research uses data only from 2011-2013. The sensor identifiers are acronyms of the HARP deployment site location and an associated letter. The sensors deployed at Channel Islands National Marine Sanctuary are labeled CINM; site B was used for this research. Data was also gathered from sensors deployed at the Diablo Canyon Power Plant (DCPP) location (sites A and C). Additionally, data from a Southern California (SOCAL) location (site N) were used to augment the data set from the DCLDE workshop. It was collected from a study conducted by Širović et al. (2012).



Site map of HARP locations used for this research.

Figure 1. HARP locations off Central and Southern California

B. DATA COLLECTION

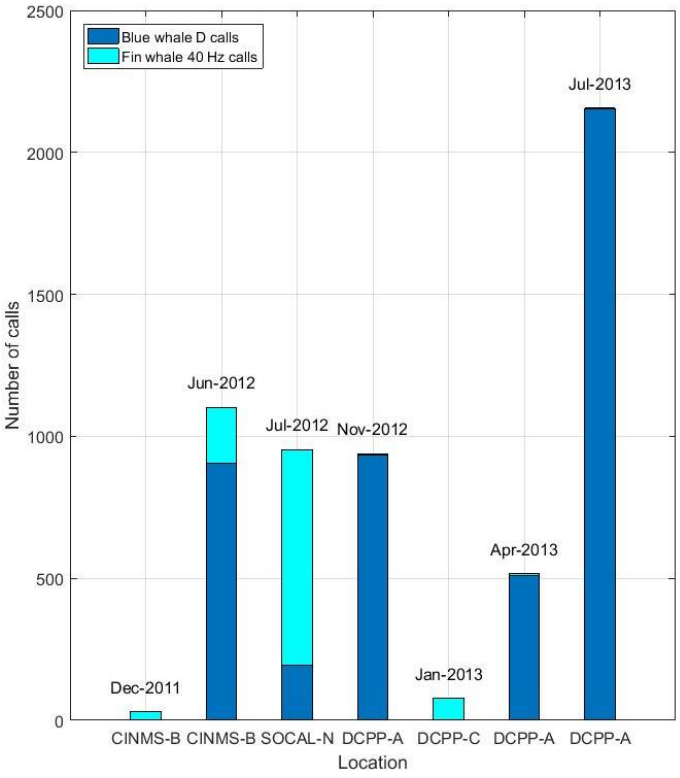
The HARPs continuously record data once recording begins, producing approximately 1.2 TB/month (Wiggins 2010). The DCLDE dataset comprises over 19 acoustic files that were collected at high sampling rates. Data from CINMS-B and DCPP-C were sampled at 200 kHz. Data from DCPP-A had a sampling rate of 320 kHz. To manage data processing more easily, the files were decimated to 1 kHz and 1.6 kHz sampling frequency (DCLDE 2015).

The deployments lasted from five days to two weeks. DCPP-A was recording for the largest amount of time overall with 18 totals days in three seasons. There are data from five days in April, seven days in July, and six days in November. CINMS-B had the second longest time recording but only has data from winter and summer from 13 days of recordings. Four days are from December and nine days are from June. This HARP was also recording for five days in November, but did not have any calls from a possible instrument malfunction. DCPP-C recorded for four days in February and SOCAL-N recorded for five days in July. Overall, the data set contains information on a total of 5,778 calls with 4,697 D-calls and 1,081 40-Hz calls and is summarized in Table 1. The annotated data is labeled with the following six characteristics: project name, site, species, start-time, end-time, and call type.

Table 1. Data summary

Sensor Name	Depth (m)	Sample Rate (kHz)	Latitude	Longitude	Blue Whale Calls	Fin Whale Calls
DCPP-A	65	320	35-36.7N	121-14.5W	3596	16
CINMS-B	600	200	34-17.0N	120-01.7W	907	227
DCPP-C	100	200	35-24.0N	121-33.8W	1	77
SOCAL-N	1250	200	32- 22.2N	118- 33.90W	193	761

For this research, the distribution of data by season was also examined (Figure 2). The HARPs were deployed during all seasons. The majority of blue whale calls were detected by DCPP-A and CINMS-B in April, June, and July. The majority of fin whale calls were detected by CINMS-B and SOCAL-N in summer and winter months. The high volume of calls collected by DCPP-A and CINMS-B is also a function of their greater deployment time. LTSAs from SOCAL-N were scanned by two analysts (including myself) specifically for fin whale calls that were needed to balance the data set so the algorithm could have several hundred fin whale calls to analyze. Due to this focused effort, the relatively high number of fin whale calls is not surprising.



A bar graph depicting the number of calls detected by each sensor. The month and year are noted at the top of each bar.

Figure 2. Distribution of calls by sensor and season

There were some noted deficiencies in the data. There was periodic self-noise believed to be from the sensors recording device (Figure 3a). This self-noise potentially

masked several calls, making it difficult to observe them during visual scanning. There are also periods of broadband noise that made visual detection challenging (Figure 3b). These periods could last for several minutes to several hours depending on the time and location of the HARP.

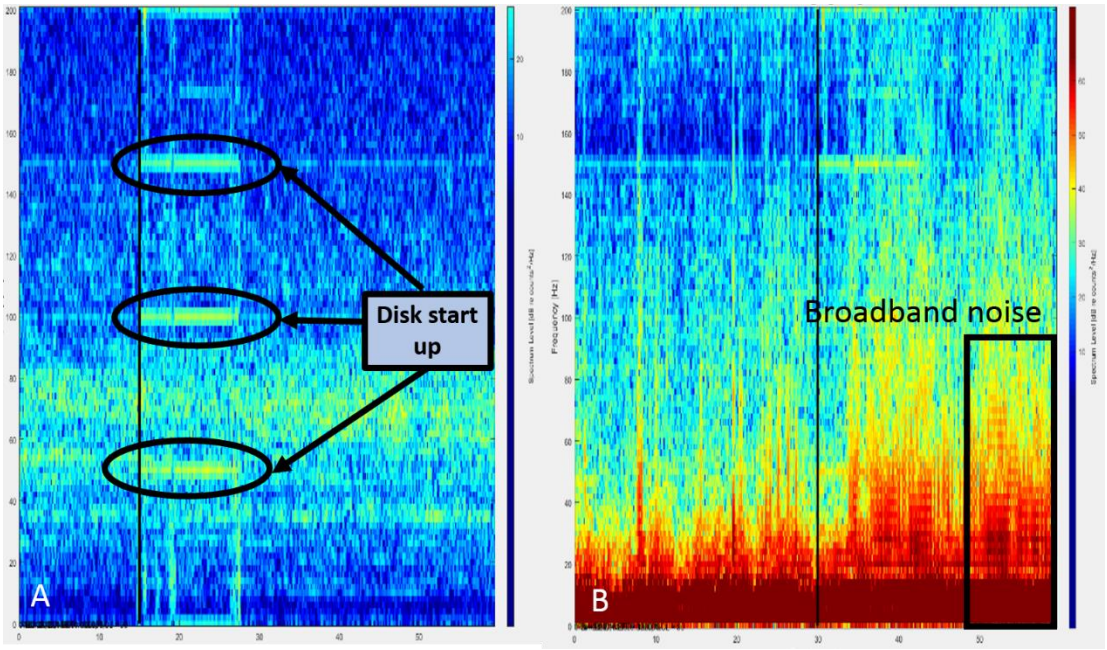
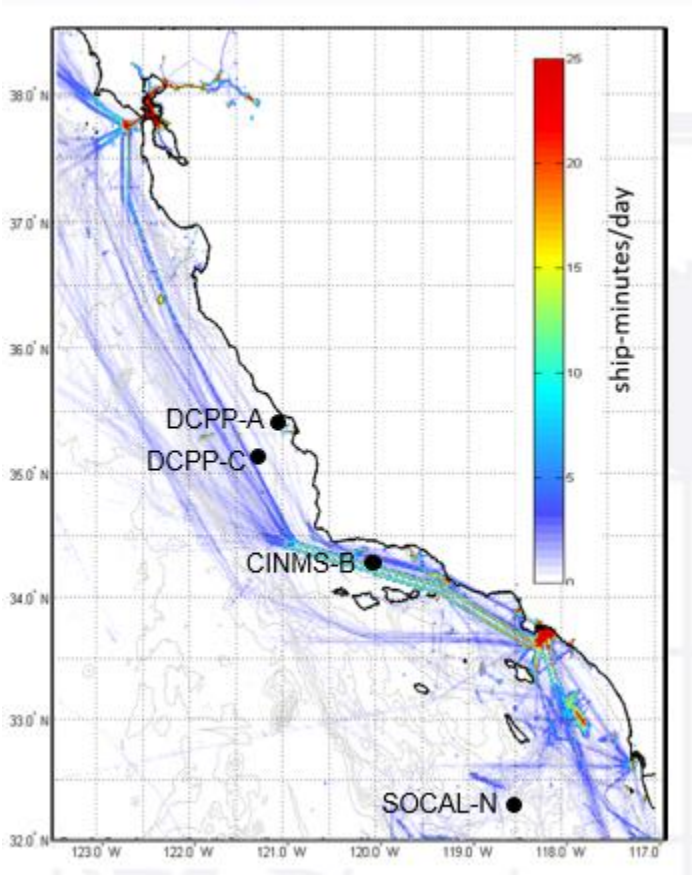


Image A shows the self-noise believed to be from disk start up for the recording device. Image B shows an example of broadband noise present in the data that possibly masked some whale vocalizations.

Figure 3. Self-noise and broadband noise in data during visual scanning

Shipping noise is a significant acoustic source from human activity found in the data set as the nearby Ports of Los Angeles and Long Beach are among the busiest in the world. A shipping pattern map shows major activity near the HARP locations, contributed to the background noise in the data set (Figure 4). Shipping data was obtained from Automatic Identification System (AIS) reports gathered from shipboard broadcasting systems and are updated every two seconds. Previous studies have shown marine mammals have a noted change in behavior when exposed to shipping noise (Castellote 2012, Joseph and Margolina 2015). While the shipping lanes in the SCB have been slightly changed to

accommodate marine mammal habitats, it is not all encompassing and the presence of vessels still affects marine mammals (NOAA 2012).



Shipping routes through known whale habitats in the Southern California Bight. The high-volume shipping traffic is also located near HARP locations, which could mask whale vocalizations.

Figure 4. Shipping density off the coast of California December 2012–March 2013

C. MODELING THE ENVIRONMENT

The Behavioral Response Study (BRS) Modeling Tool developed at NPS was used to model the acoustic environment around four data sites from a shallow low-frequency sound source. The tool has four main components: environmental databases, a library of pre-run model outputs on a regular grid for a set of depths, a model of underwater acoustic

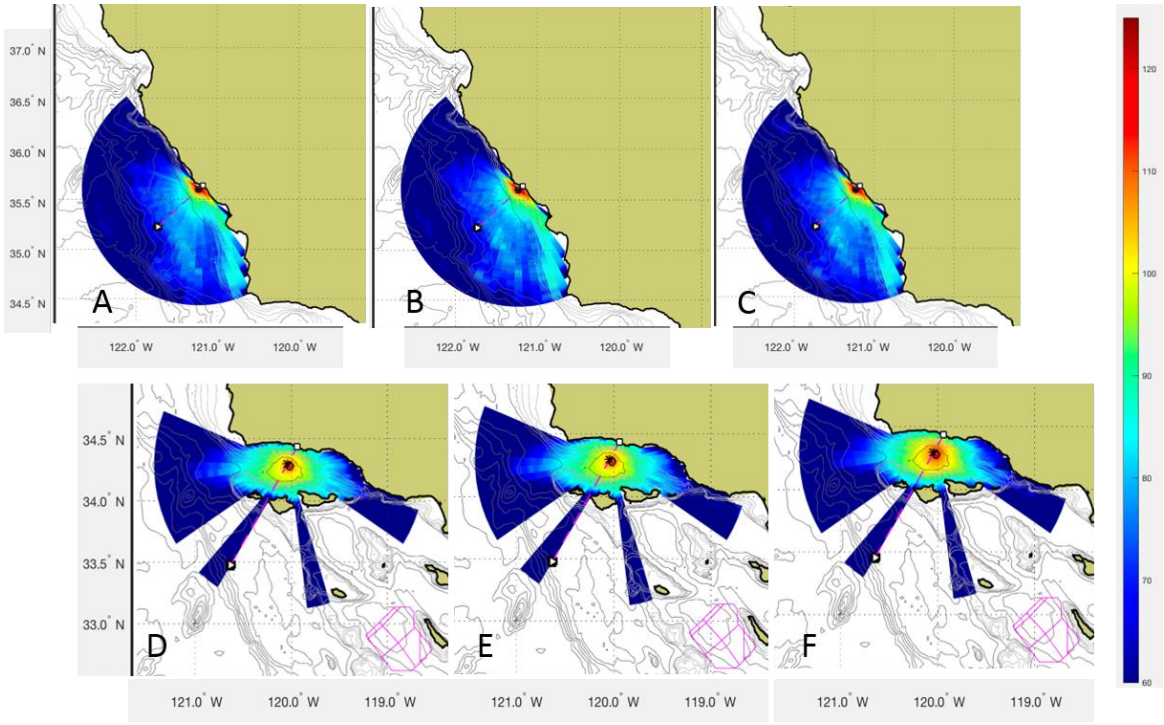
propagation, and a graphical user interface (GUI) providing visualization and interactive capabilities.

The environmental database consists of four main components: bathymetry, monthly sound speed profiles, sea surface wind speed, and bottom sediment type. Bathymetry data has a resolution of ¼ minute latitude and longitude. It is from the Digital Bathymetric Data Base Variable resolution (DBDBV) developed by the Naval Oceanographic Office. The US Navy's Generalized Digital Environmental Model (GDEM) was used monthly ocean temperatures and salinity in order to calculate sound speed. The sea surface wind was collected from the High-Resolution Global Sea Surface Wind Speed and Climatology from NOAA and has a resolution of ¼ degree. The Navy's bottom sediment type database and Global Ocean Sediment Thickness from NOAA are used for plotting bottom type and geo-acoustics.

The model used for estimating underwater acoustic propagation was the Navy Standard Parabolic Equation Model (NSPE). The NSPE is used to predict narrow band, low-frequency transmission loss (TL) based on the ocean environment. Estimating the TL in a particular environment is important factor to determine if a sound source will be detected at a distant receiver. The GUI allows users to set model parameters in order to create model input, retrieve pre-run model outputs, process and visualize model output. Additionally, the GUI allows users to interactively derive transmission loss information from a two-dimensional model output.

The BRS modeling tool was used to model the received level at each HARP site from a sound source located within 70 nautical miles from the hydrophone. The latitude, longitude and depth of each HARP were inputted as the receiver location. Other input parameters include source level (SL) and month. SL was set at 175 dB in re 1 µPa at 1m with a source depth of 30 m based on previous studies by Samaran et al. (2010) and Wiggins (2018). They found 175 dB average SL for both blue whale D-calls and fin whale 40-Hz calls. In general, it appears that the receive level (RL) increase as proximity to the sensor increases, however, differences in bathymetry and time of year can cause unique features in sound propagation patterns.

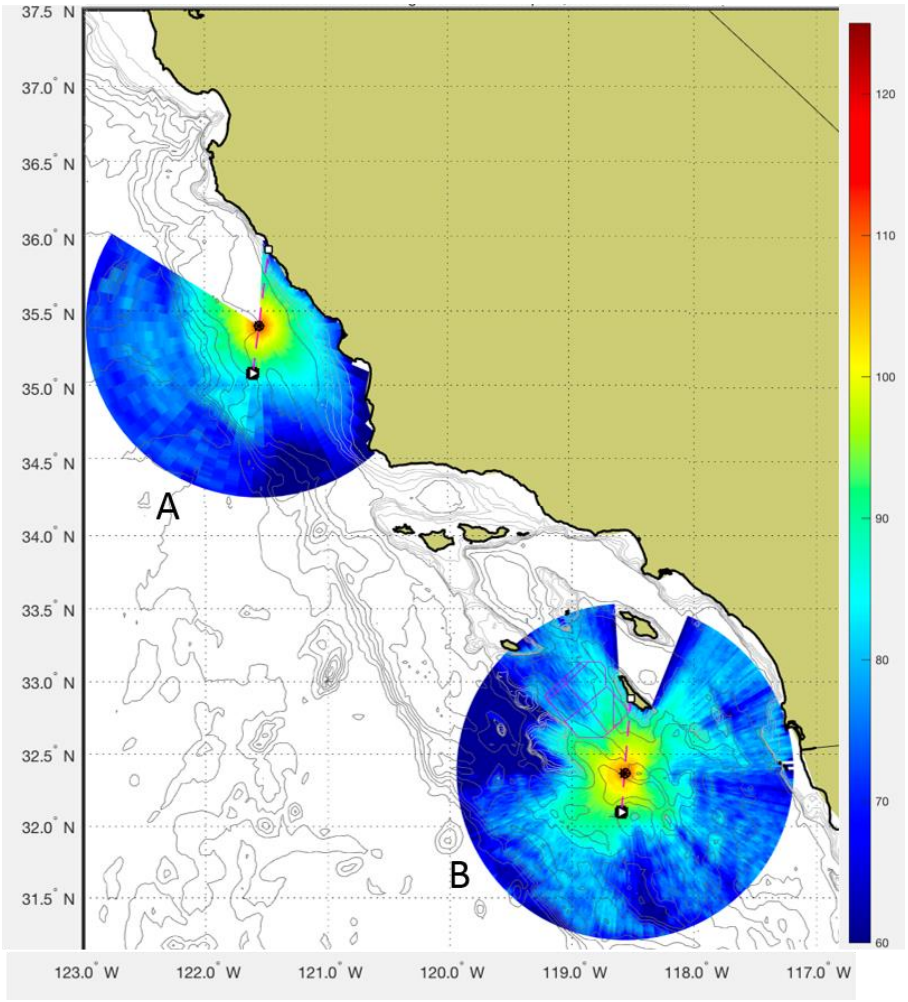
DCPP-A and CINMS-B sites were modeled for the various months they were deployed (Figure 5). DCPP-A was the shallowest sensor and the closest to the coastline. Modeling shows that reception was higher along the coast to the south. Reception to the west and north of the hydrophone was not as high as indicated by the dark blue color (Figure 5 A-C). CINMS-B was located in the SBC basin, which created higher reception levels caused by the bathymetry (Figure 5 D-F). The Santa Barbara Channel Islands blocked reception from contacts to the south.



Images A-C show the DCPP-A site: A: November; B: April; C: July. Images D-E show the CINMS-B site: D: November; E: December; F: July. The pink outline represents the Navy's training area SOAR off San Clemente Island.

Figure 5. Receive level comparison between DCPP-A and CINMS-B sites

The DCPP-C and SOCAL-N sites were located much further away from the coast and were not as restricted by islands for detection purposes (Figure 6). The RL was greater in general than the RL around DCPP-A or CINMS-B. There also appears to be some bottom bounce extending this high RL to the west. SOCAL-N is modeled for July and shows the enhanced propagation as well from sources located nearly 70 nm away from the sensor.



A indicates the DCPP-C site in February. B shows the SOCAL-N site in July.

Figure 6. Receive level comparison between HARP sites DCPP-C and SOCAL-N

III. METHODOLOGY

A. DATA PREPARATION

1. Visual Scanning

The original DCLDE 2015 low-frequency dataset had less than 7% (320 calls) of fin whale 40-Hz calls and 93% (4504 calls) of blue whale D-calls. In order to add to the bank of knowledge for generating fin whale centroids used in the machine learning algorithm, additional fin whale calls were necessary. SIO provided an additional data set from a HARP, called SOCAL-N in previous sections, located near in the same region containing fin whale calls. The data set was manually scanned by two independent parties (Tetyana Margolina and Michelle Tanalega) and results were compared. The best results were then added to the existing database. After scanning, 193 blue whale calls and 761 fin whale calls were added. The combined data set is 19% fin whale calls and 81% blue whale calls.

Visual scanning using the Triton program is a convenient method to create a usable log of annotated foraging calls. Triton displays the data in form of spectrograms with a set of parameters based on user input. Data is uploaded as a XWAV file and is decimated by a factor of 100 to allow the operator to easily focus the analysis on the low frequency bands of the foraging calls of interest. The user must set parameters as desired for viewing data. A logging function is also available in Triton to keep track of calls of interest.

2. Spectrograms

Spectrograms are three-dimensional images that show how a signal frequency changes with time. Time is on the x-axis, frequency is on the y-axis and intensity is represented by a color scale. Huang (2016) created centered spectrograms using adapted techniques from SIO and the Ocean Acoustic Laboratory of the Naval Postgraduate School (Oleson 2007a; Margolina 2010; Širović 2011). These procedures were modified to account for different sampling rates and overlap needed to obtain higher resolution for defining features in the output image. Centered spectrograms are created by first averaging the length of each annotated call in the DCLDE dataset over a time duration of ten seconds with a frequency band from 0-100 Hz. In order to break the signal into its frequency components, a fast Fourier transform (FFT) was conducted using a specified number of

points. This research and previous work by Huang use the sampling rate of each HARP as the number of points to create each FFT segment with a 91% overlap in data. Additionally, other input parameters used by Huang were used in this research. In Triton, parameters include 100% contrast setting, 16% brightness, and the color-map for MATLAB is "jet." The output image is a 9.09-second spectrogram that has a frequency resolution of 1 Hz and a time resolution 0.09 seconds. The centered spectrograms contain 10,100 pixels, or elements, and can be represented as 100x101 matrices.

3. Defining Features

Blue and fin whale foraging calls share some common characteristics. Both calls are produced by both genders in each species and are typically produced at shallow depths with no regular temporal pattern. However, there are some defining features that separate the calls shown in Figure 7. Blue whale D-calls are down sweeping calls for up to 30 Hz occurring in 30-120 Hz frequency band lasting 1-4 seconds. Fin whale 40-Hz calls are also down sweeping for approximately 30 Hz, occurring in the 40-75 Hz frequency band with a duration of approximately one second. These calls are highly variable and can often overlap in duration and frequency bands, making them difficult to distinguish.

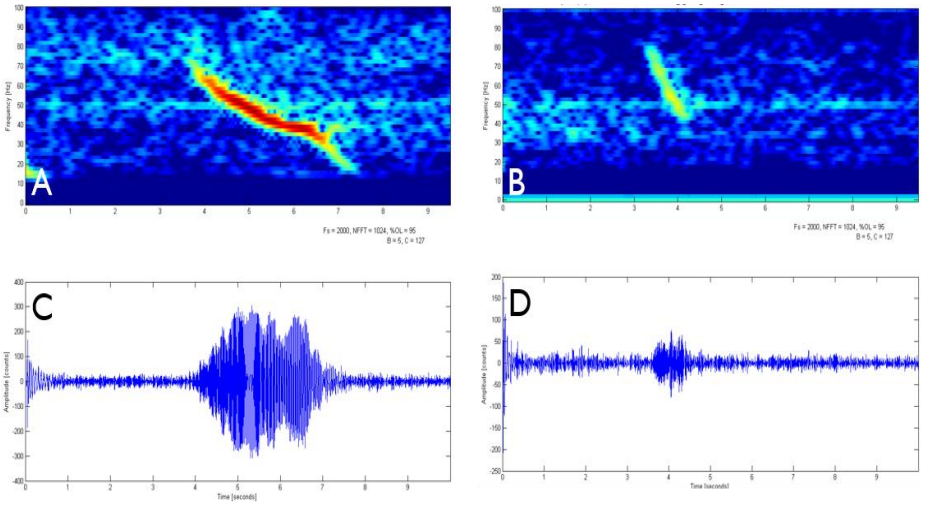
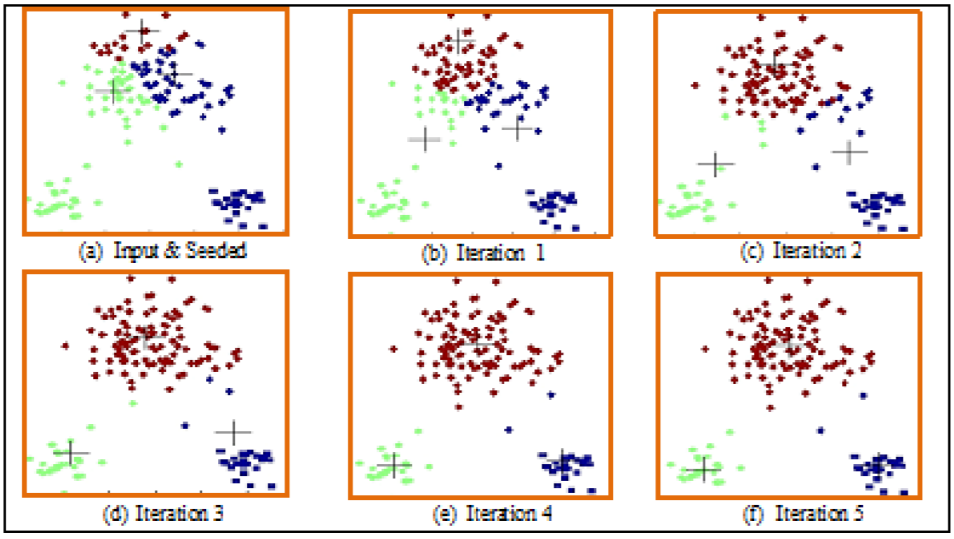


Image A is a spectrogram of a typical blue whale D-call. Image B is a spectrogram of a typical fin whale 40-Hz call. Images C and D are corresponding time series plots to each call in the spectrogram shown above it.

Figure 7. Blue and fin whale foraging calls. Adapted from Margolina (2015).

B. K-MEANS CONCEPTS

The K-means clustering method is an iterative process for clustering a data set into groups containing similar information based on the averages of inputted data (MacQueen 1967, Lloyd 1982). The algorithm has the general process of initialization, assignment, recalculation, and reassignment of cluster means until all samples in a data set are assigned to a stable cluster (Figure 8).



A two-dimensional example of the k-means algorithm. The data set is divided into three clusters by the repetitive process of cluster assignment and reassignment based on least-squared distance to the centroid (+symbol). The location of the centroid also continues to update with each iteration until a stable position is reached.

Figure 8. An illustration of the k-means algorithm. Source: Jain (2008).

During the initialization process, an initial number of k cluster centers are chosen. As a default, MATLAB uses the k-means++ algorithm to establish cluster centers called centroids (MathWorks 2018). The k-means++ algorithm was found to decrease run time and produce better results by using a random selection process rather than the traditional method of placing centroids arbitrarily in the data set (Arthur and Vassilvitskii 2007). The algorithm consists of four steps:

First, an observation, *X*, is chosen uniformly at random to be the first centroid, called c_1. Second, the distance from each data point to the centroid is calculated, represented as *D(x)*. Third, a new centroid, c_2, is chosen also at random with the probability:

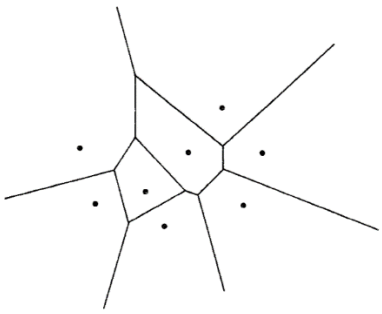
$$\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$$

Fourth, repeat the process until the number of specified centroids are selected.

The next phase assigns all data points to the closest centroid by calculating the distance from each sample to all centroids. This could be done using several types of distance measurements, but this research uses Euclidean method of computing distances using the formula:

$$D(x,c) = (x-c)(x-c)'$$

From this simple calculation, each data point is assigned to the nearest centroid creating Voronoi cells that contain all data points that are closest to a specified centroid (Figure 9).



Simple Voronoi diagram showing eight Voronoi cells.

Figure 9. Diagram of Voronoi cells. Source: Aurenhammer (1991).

The recalculation and reassignment steps in the process are to minimize the sum of the distances from the points to the centroid using a two-phase iterative algorithm detailed on the MathWorks website (MathWorks, 2018). The first phase uses batch updates. During

each iteration, observation points are re-assigned to their nearest cluster centroid all at one time. The reassignment is followed by a recalculation of cluster centroids. Batch updates have greater potential to not converge to a solution that is a local minimum, meaning the data are separated in a way that moving any observation point to a different point will increase the total sum of distances. Computing time of batch updates is relatively fast, but could only supply a starting point for the second phase and is not an overall solution. The second phase uses online updates. Online updates individually re-assigns observation points only if doing so reduces the sum of distances. Each iteration of this phase does one pass through all the points. Similar to the first phase, centroids are recalculated after reassignment. This phase does converge to a local minimum, but there might be other local minima with lower total sum of distances. The number of starting points must be refined to find a global minimum. The update process repositions the location of the centroid in each cluster. The reassignment process continues until the cluster positions stabilize and all points remain in the same cluster (Figure 8f).

C. K-MEANS APPLIED TO DCLDE DATA

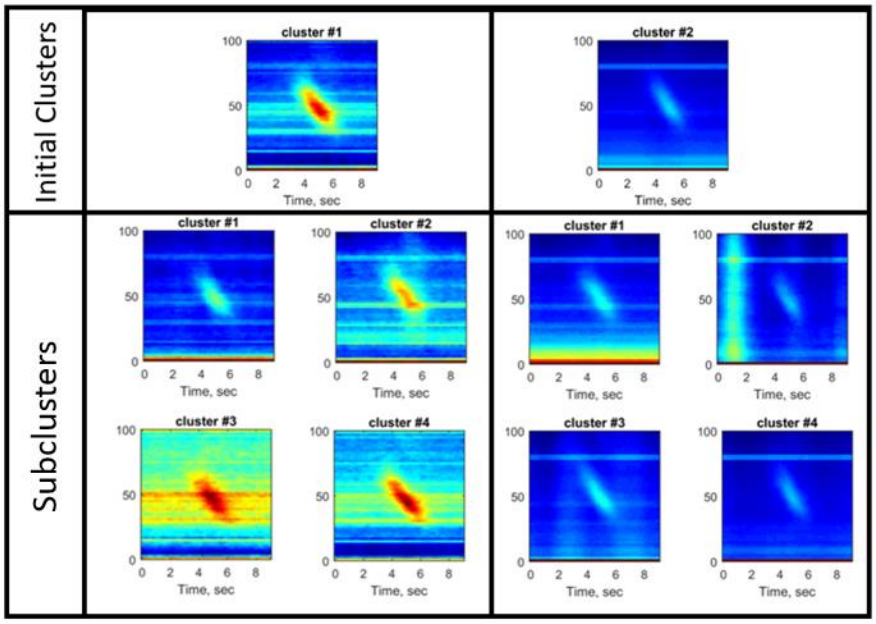
Before explaining how the algorithm works with the DCLDE data, it is important to discuss how MATLAB views an image since the data set is composed of thousands of spectrograms images. How these images are processed in MATLAB is the foundation of how k-means clustering is a possibility for this type of data. In the spectrograms, each pixel is defined with one color based on the intensity of the pixel.

In MATLAB, an image is represented as a matrix of values based on a given parameter, intensity in this case. Each pixel is identified by a matrix location based on its row and column position. In order to apply k-means to images, the images must first be vectorized. This simply means taking the values from a matrix format and reshaping them into a vector that can be more easily compared to other vectors derived from other images. A simple example is presented in Figure 10. A 9-pixel image is presented as a 3 x 3 matrix. Each pixel is given a value based on its color. For this example, red=4, green=2, and blue =1. When the matrix becomes a vector, it is done as column- major order. The order of the values put into a linear array are done so by column rather than by row.

Simple example of matrix vectorization. The matrix is reshaped into a linear vector of values for faster processing.

Figure 10. Matrix vectorization by column-major order. Adapted from Bendersky (2015).

Recall, all spectrogram images contains 10,100 pixels each with an intensity value in matrix format. The values are vectorized into a linear array containing 10,100 components during the application of the k-means algorithm. This process occurs for each spectrogram. After the vector is formed, the initial centroids are chosen uniformly at random and the k-means process begins. The distances from corresponding pixels between arrays is measured and compared to find the closest centroid. After assignment, the centroid location is changed to the true center of the cluster. At this stage, the centroids will not be actual spectrogram images. The centroid images produced are created from finding the center of each cluster based on distance calculations. From these calculations, an image is created where the pixels are representative of the spectrograms in each cluster (Figure 11). The images produced are called centroids and are a representative of clustered, or similar, data.



Initial and secondary clustering example shown. The top two clusters are initial clusters. Numbers of both initial and secondary clusters are determined by user and can be changed to best fit data needs. The spectrograms have frequency on the x-axis and time on the y-axis.

Figure 11. Example of initial and secondary centroids

Once the data has been partitioned, the designated data set is run through the k-means algorithm in MATLAB. Initial clusters are formed using the k-means ++ algorithm. The clusters are formed by comparing values in each vector in order to find common trends in the data. For example, one cluster could have calls with loud shipping noise while another has mostly calls in a relatively "quiet" environment. The centroids are sequentially labeled and do not represent actual foraging calls. Initial centroids are then partitioned again during secondary clustering using the same method. Both the number of initial and secondary clusters are determined by manual input.

While using this approach, the number of initial and secondary clusters were changed and manipulated several times to determine if there was an optimum number of clusters for the data. In some runs, the number of clusters were equal for both species, while in other runs the number of clusters were changed. For example, in one run, the number of initial clusters would be set to three and the number of secondary clusters would be set to three for both blue and fin whales. In another run, the number of initial clusters would be

changed to four and the number of secondary clusters set to five for blue whales, but kept the at three initial and secondary for fin whales. The results were compared to determine the best combination of initial and secondary cluster numbers.

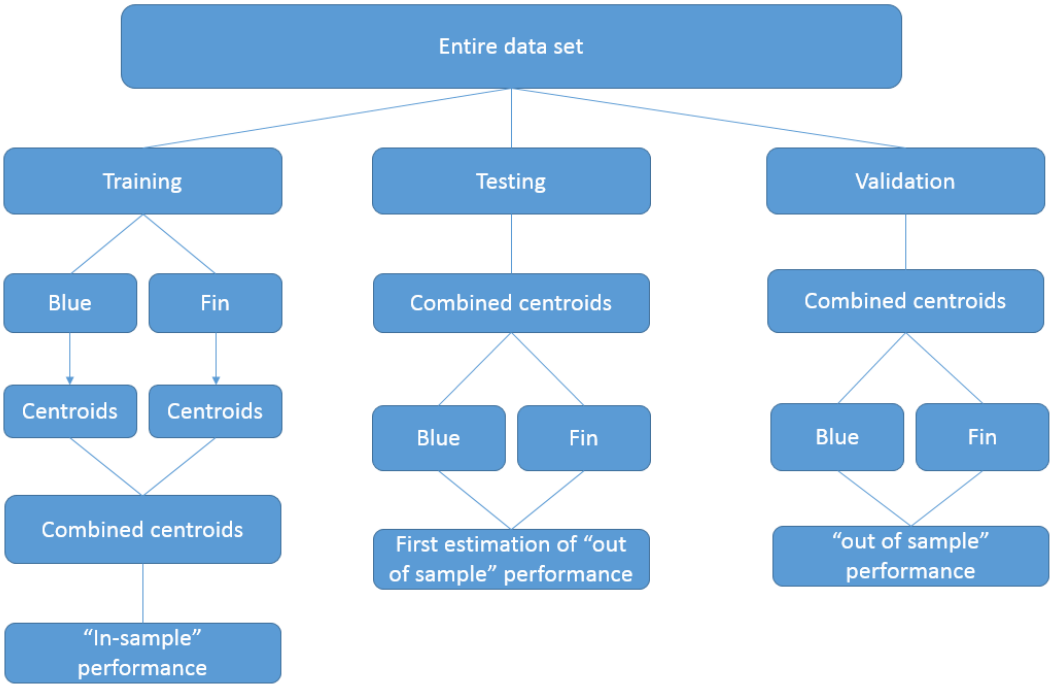
D. CROSS-VALIDATION METHOD

The data set was randomly partitioned into three subgroups to satisfy the cross-validation method (Figure 12). Approximately 60% formed a training data set. The remaining 40% was divided equally into a validation data set (20%) and a testing data set (20%). The partitioning was done as to preserve the original composition of the data set in terms of deployment location and season, and also to preserve the proportion between blue and fin whale calls. The training set is used to fit the model and determine the number of initial and secondary clusters. The training set is separated again by species as identified in the DCLDE data set by expert analysts. The training group of each species is then run through the k-mean algorithm, forming separated centroids. The data are combined again to include both species before it is used to evaluate "in-sample" results. In-sample results are performance metrics of recall, precision and accuracy, which will be defined in the Results section (Table 2). Several runs are completed until the number of initial and secondary clusters yields the best in-sample results.

After satisfactory in-sample results are reached, the algorithm is ready to be evaluated using the testing data set. The testing subgroup is used to estimate an "out-of-sample" performance of the algorithm. Out-of-sample results are also the performance metrics of recall, precision, and accuracy. The testing subgroup of data, which contains both species, is run through the combined centroids previously determined by the training subgroup. The testing subgroup can be used similarly like the training data set. It can be used several times on different combinations of clusters to determine if any trends developed in the data.

If satisfactory results are not obtained from the testing subset, the algorithm can be applied to the validation data set. The validation data set will only be used if training and testing subset must be combined to create a new training subset. Since the two-thirds of the data would be used for new training, the final third would be the only source of validation

data. It is important to conduct several test runs on training and testing subgroups prior to the run on the validation subgroup because it will only be used once. This final run using the validation subset will be the new out-of-sample performance and most accurately describe the true performance of the algorithm. However, if the results from the testing subset are satisfactory, then the validation data set is not used and the out-of-sample performance is based on testing subset performance.



The cross-validation method specific for this data set. Combined centroids represent the data being applied to the centroids developed by the k-means algorithm. At this stage, there is no separation of blue and fin whale centroids.

Figure 12. Cross-validation method flow chart

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS

As in previous work by Huang et al. (2016), the performance of the proposed k-means classification algorithm was evaluated using the DCLDE 2015 workshop-scoring tool. In this chapter, the scoring tool is explained and the species' specific results of the algorithm are displayed.

A. ESTIMATION OF DETECTOR PERFORMANCE

Detector performance was estimated by comparing the annotation file from the workshop to the output of the k-means algorithm, to determine if the algorithm correctly classified calls by species. Performance of the algorithm was estimated for each species separately (Table 2) using the training or testing datasets. Since the datasets contain less fin whale calls than blue whale calls, performance estimation for combined calls would be unrealistically inflated by the blue whale successful detections. For each call within the total training or testing dataset, distances to all species-specific centroids were calculated, and the call was identified as a blue or a fin whale call based on the smallest distance to any of the centroids.

The performance of various combinations of clusters is compared by the metrics of recall, precision and accuracy (Table 2). Recall is the number of successfully retrieved calls of a certain species. Precision is number of calls correctly categorized as a true positive out of positive classifications. Accuracy is the number of calls correctly categorized out of the entire data set. For each species, true positives (TP) are correctly identified calls of this species. False positives (FP) are calls misidentified as calls produced by this species. False negatives (FN) are calls of another species incorrectly identified as produced by this species. True negatives (TN) are correctly identified calls of another species. Column totals are the total number of positive calls (P) and the total number of negative calls (N).

Table 2. Summary of performance metrics

Blue whale performance			
		Annotation	
		Blue	Fin
Identification	Blue	True Positive (TP)	False Positive (FP)
	Fin	False Negative (FN)	True Negative (TN)
Totals		P	N

Fin whale performance			
		Annotation	
		Fin	Blue
Identification	Fin	True Positive (TP)	False Positive (FP)
	Blue	False Negative (FN)	True Negative (TN)
	Totals	P	N

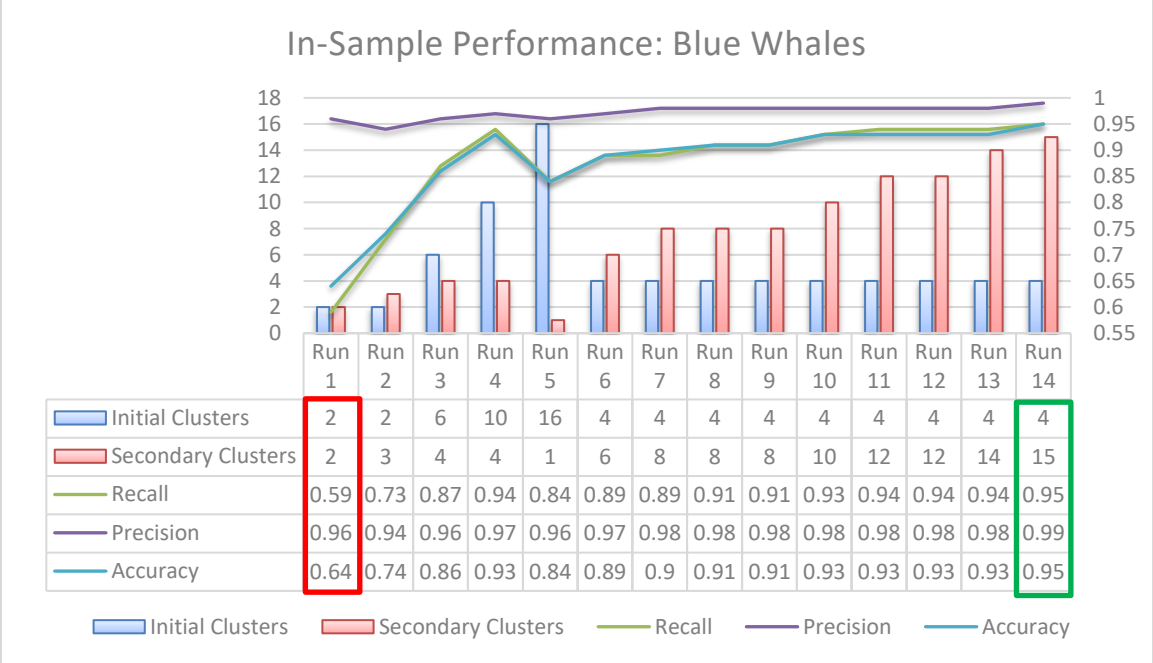
Formulas used:	
Recall	$\frac{TP}{P}$
Precision	$\frac{TP}{TP+FP}$
Accuracy	$\frac{TP+TN}{P+N}$

B. IN-SAMPLE PERFORMANCE

Due to the high variability in both types of calls, determining the appropriate number of clusters to use is challenging and therefore a modeling experiment of varying the combinations of initial and secondary clusters was conducted. Several runs were conducted with different combinations of initial and secondary clusters. The initial runs kept the number of initial clusters the same for both species but the number of secondary clusters were changed. In some instances, the number of initial and secondary runs were kept stable for one species while variations were conducted in the other species to determine if that influenced results.

As more runs were conducted, several trends were noted. Generally, as the number of initial and secondary clusters increased, computing time increased. For performance, it was observed that as the number of secondary clusters increased, blue whale classification performance increased while fin whale performance did not change significantly. However, when the number of initial clusters was kept relatively low and the number of

secondary was increased, the performance increased for both species. Additionally, blue whale data exhibited more variability in precision and recall than in accuracy. Fin whale data exhibited more variability in accuracy than in precision and recall. The overall performance of the training data set for blue and fin whales is shown in Figures 13 and 14, respectively.



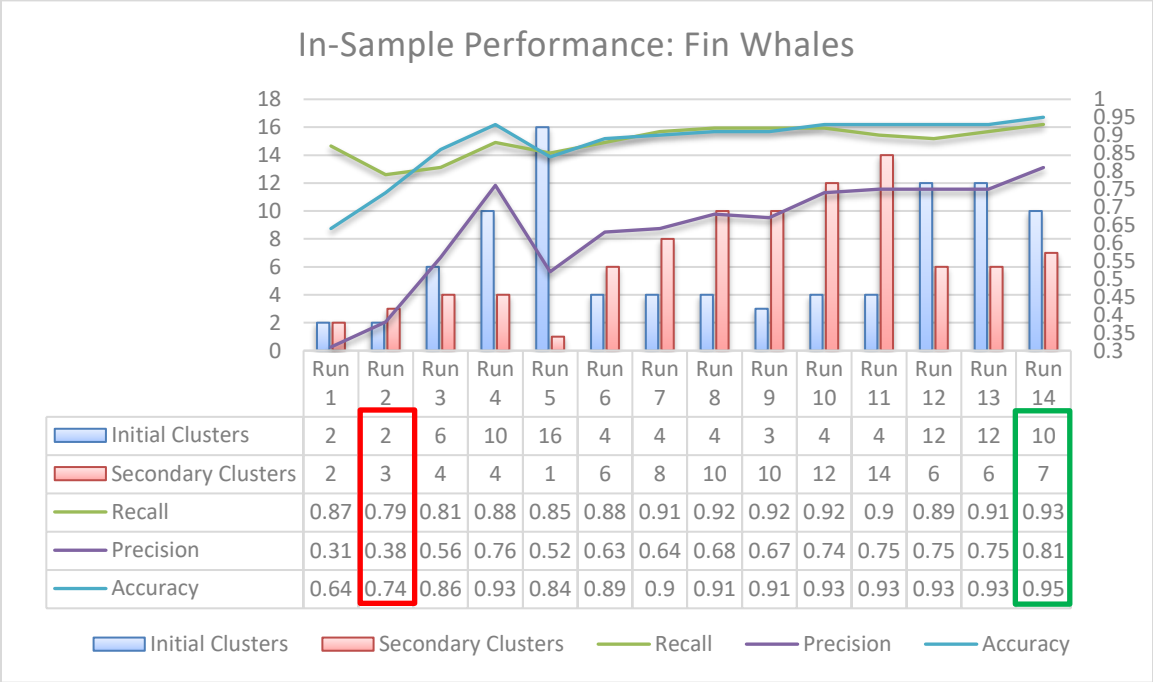
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Run 11	Run 12	Run 13	Run 14
Initial Clusters	2	2	6	10	16	4	4	4	4	4	4	4	4	4
Secondary Clusters	2	3	4	4	1	6	8	8	8	10	12	12	14	15
Recall	0.59	0.73	0.87	0.94	0.84	0.89	0.89	0.91	0.91	0.93	0.94	0.94	0.94	0.95
Precision	0.96	0.94	0.96	0.97	0.96	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99
Accuracy	0.64	0.74	0.86	0.93	0.84	0.89	0.9	0.91	0.91	0.93	0.93	0.93	0.93	0.95

Comparison of in-sample performance for blue whale calls. The red box highlights the least-efficient run, and the green box highlights the most efficient run

Figure 13. Comparison of in-sample performance for blue whale calls

The first five runs of the algorithm on the blue whale data focused on increasing the number of initial clusters. As the number of initial clusters increased from two to ten, there was a significant increase in recall and accuracy while precision remained relatively steady. The largest number of initial clusters generated was 16; however, this resulted in a noticeable decrease in performance. Runs 6-14 focused on maintaining a steady number of initial clusters and varying the number of secondary clusters. Runs 7-9 were kept identical for blue whales as combinations varied for fin whales to see if performance would change for fin whales. As the number of secondary clusters increased, recall and precision

increased considerably while accuracy increased slightly. Run 14 resulted in the optimal combination of four initial and 15 secondary clusters.



	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Run 11	Run 12	Run 13	Run 14
Initial Clusters	2	2	6	10	16	4	4	4	3	4	4	12	12	10
Secondary Clusters	2	3	4	4	1	6	8	10	10	12	14	6	6	7
Recall	0.87	0.79	0.81	0.88	0.85	0.88	0.91	0.92	0.92	0.92	0.9	0.89	0.91	0.93
Precision	0.31	0.38	0.56	0.76	0.52	0.63	0.64	0.68	0.67	0.74	0.75	0.75	0.75	0.81
Accuracy	0.64	0.74	0.86	0.93	0.84	0.89	0.9	0.91	0.91	0.93	0.93	0.93	0.93	0.95

Comparison of in-sample performance for fin whale calls. The red box highlights the least-efficient run, and the green box highlights the most efficient run.

Figure 14. In-sample performance for fin whale calls

As with the blue whale data, the first five runs for fin whales focused on increasing the number the initial clusters until peak performance was attained. Unlike blue whales, fin whale precision showed the greatest increase during runs 1-4 while accuracy increased somewhat and recall varied slightly. Run 5 also had the greatest number of initial clusters with 16, but also showed the same decrease in overall performance as the blue whale detector. Runs 6-11 also focused on varying secondary clusters. This showed a large improvement in accuracy which leveled at run 11. Interestingly, the overall performance was similar between runs 11 and 12 although both have different approaches. Run 11 had a lower number of initial clusters with a high number of secondary clusters while run 12 was the opposite arrangement of clusters. Run 13 had slight improvements over run 12 while blue whale run 13 had a low number of initial clusters and a high number of

secondary clusters. Run 14 was the optimal combination of clusters with 10 initial and 7 secondary clusters.

The results were also compared using a precision-recall (PR) plots to determine the best overall combination of clusters. PR curves plot data with the precision on the y-axis and the recall on the x-axis and ideal performance is achieved in the upper right corner of the plot (1, 1) (Davis 2006). From a PR plot of both species in-sample performance, the least and most efficient runs can be identified (Figure 15). The most efficient run for each species is clearly run 14. For fin whales, the least efficient combination of clusters was run 2 with two initial clusters and four secondary clusters. For blue whales, run 1 was the least efficient with two initial and two secondary clusters. However, since performance of each run will depend on both blue and fin whale centroids, we only consider paired performance with the same run number. The results present a compromise between precision and recall for both species.

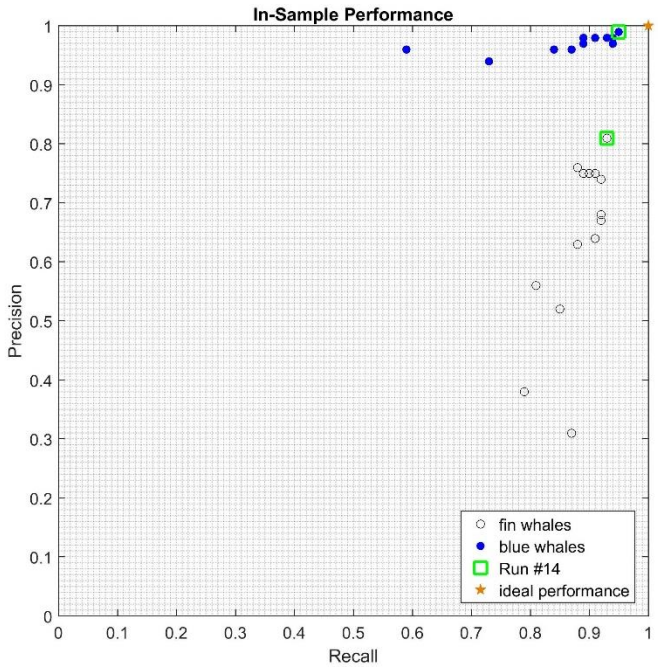


Figure 15. PR plot of the in-sample performance of the algorithm

The composition of the highest and lowest performing combination of clusters were examined to analyze the difference in performance on the training data set. Centroids were analyzed according to their visual appearance. Those with large amounts of background noise are described as loud while those with minimal noise are faint. Strong signals are those with high intensity according to the color bar scale where blue is low intensity and red is high intensity. It is clear that run 1, two initial clusters and two secondary clusters, had the overall lowest performance for blue whales from the PR plots. For this run, precision was high at 96% but accuracy and recall were the lowest at 64% and 59% respectively. Looking at the composition of centroids, cluster #1 has the strongest signals in the separation of data. Subcluster #1 has more ambient noise than subcluster #2. Both subclusters only show horizontal sound patterns and not vertical patterns known to be in the data. Cluster #2 shows quieter background noises. Notably, only subcluster #2 captures the loud signals.

The least efficient combination of clusters for fin whales was run 2 with two initial clusters and three secondary clusters. Performance was lowest at 38% precision, 79% recall and 74% accuracy. Contrary from the blue whale clusters, cluster #1 consists primarily of fainter calls. Subcluster #3 shows the most ambient noise in the cluster. Cluster #2 has signals in louder environments with more ambient noise than cluster #1. None of the clusters capture the broadband noise patterns known to be in data set.

The highest performance for blue whale calls was achieved on run 14 with four initial clusters and 15 secondary clusters. Comparing these centroids to the run 1, the initial clusters have several similarities. Clusters #1, #3, and #4 separated the data the same way and have almost identical initial centroids compared to the lowest performing run. Cluster #3 also shows the loudest part of the data set, but with 15 subclusters accounts for more variety in samples.

The highest performance for the fin whale calls was also achieved on run 14 with 10 initial clusters and 7 secondary clusters. The first six clusters of run 11 look very close to the six initial clusters of run 1. The additional four initial clusters in run 11 seem to account for the large variations in noise in the fin whale data set. Cluster #7 has the faintest centroid and the subcluster centroids barely resemble calls but might account for the

algorithm identifying faint calls as fin whales, which would increase the performance. A summary of the highest performing run, run 14, is given in Table 3.

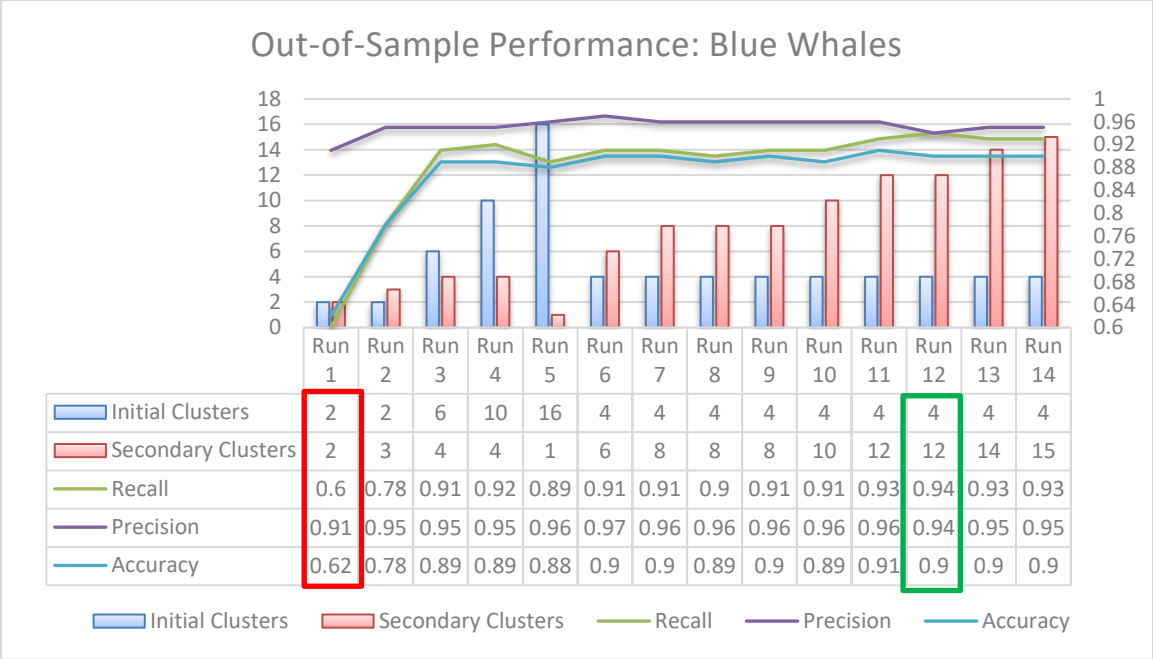
Table 3. In-sample-performance of run 14

Blue whale performance					Fin whale performance				
		Annotations					Annotations		
		Blue	**Fin**	Totals			**Blue**	**Fin**	Totals
Cluster Assignment	**Blue**	2620	37	2657	Cluster Assignment	**Blue**	529	126	655
	Fin	126	529	655		**Fin**	37	2620	2657
	Totals	P=2746	N=566	3312		Totals	P=566	N=2746	3312

C. OUT-OF-SAMPLE PERFORMANCE

Results from the final training run were satisfactory and optimistic enough to begin evaluating the data on the testing data set. Similar results were expected, but were not the case with some combination of clusters. In general, the overall performance of the algorithm was much better with the testing data set. The blue whale data set showed the same steady precision but changes in recall and accuracy. The fin whale data set also showed the same variability in precision as during the training runs, but recall and accuracy were not as closely related.

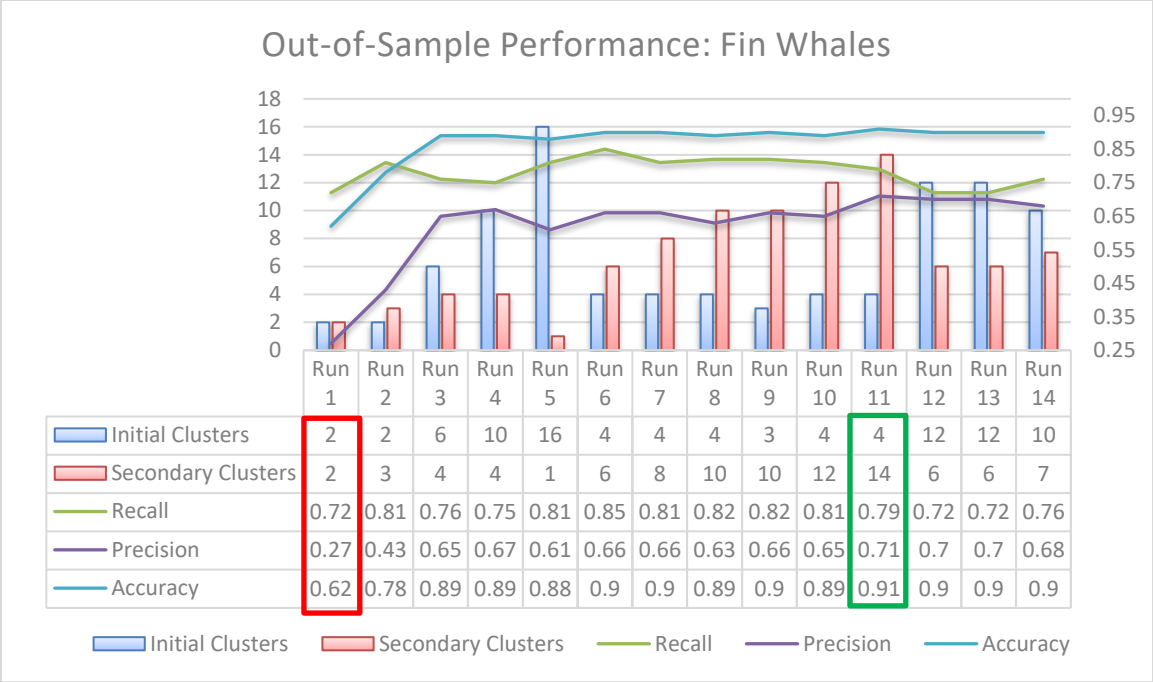
For blue whales, precision stayed fairly constant with around 90% (Figure 16). Similar to the trend in the training runs, recall and accuracy increased when the number of initial clusters increased during the first four runs. Overall performance did decrease on run 5, but not as noticeable as with the training data. Precision had a peak at run 4 with 97% and steadily decreased as the number of secondary clusters increased. Unlike the in-sample performance, precision decreased from 96% in run 11 to 94% in run 12. Accuracy stayed fairly constant around 90% after run 6 and reached a maximum at 91% during run 11. Recall exhibited similar trends with an average of 92% from runs 6-14 and a maximum at 94% during run 12.



Comparison of out-of-sample performance. The green box highlights the overall highest performing run, and the red box highlights the lowest performing run.

Figure 16. Out-of-sample performance for blue whale calls

Although the performance was overall lower than with blue whale data, the performance of the algorithm for fin whales was generally much higher using the testing data (Figure 17). Similar to the trends in the training data, precision and accuracy increased with the first four runs. Recall was variable during the first four runs and decreased as the number of clusters increased. Overall performance decreased on run 5 as before. Recall peaked at run 6 with 85% and steadily decreased as the number secondary clusters increased. Contrarily, precision increased as the number of secondary clusters increased, reaching a peak at run 7. Accuracy remained steady around 90% after run 6 and also reached a peak at run 11 with 91%.



Comparison of out-of-sample cluster performance. The green box highlights the overall highest performing run, and the red box highlights the lowest performing runs.

Figure 17. Comparison of out-of-sample performance for fin whales

A PR plot shows a comparison of the out-of-sample performance for each species (Figure 18). For blue whales, the least efficient run was run 1 again while the most efficient was run 12. For fin whales, the least efficient combination of clusters was also run 1 while the most efficient was run 11 with four initial clusters and 14 secondary clusters. The highest in-sample performance for fin whales was with a cluster combination of higher initial clusters.

Run 14, the highest performing in-sample run, had very good out-of-sample results. We expected this run to do fairly well based on the in-sample performance. While there were runs with greater performance for blue and fin whale data, like runs 12 and 11 respectively, the overall performance is still based on a paired performance for both blue and fin whale data with same run number. For further analysis, the centroids from those higher performing runs were compared to investigate their higher performance.

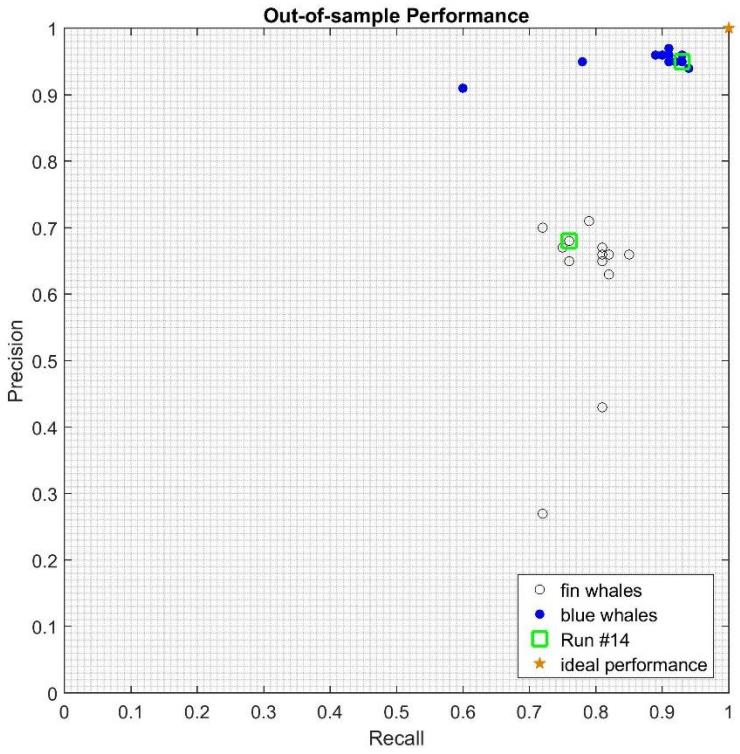


Figure 18. Out-of-sample performance PR plot

The composition of the highest and lowest performing combination of clusters were examined to analyze the difference in out-of-sample performances. It is clear that run 1, two initial clusters and two secondary clusters, had the overall lowest performance for blue and fin whales from the PR plot. Blue whale performance was 60% recall, 91% precision, and 62% accuracy. The centroids for blue whales are the same as the training data.

Fin whale performance for run 1 results in 72% recall, 27% precision, and 62% accuracy. The centroids are slightly different from the least-efficient in-sample centroids. The first cluster is still the quieter of the two but there is no third subcluster with more ambient noise. Only the second cluster captures signal with background noise. There are simply not enough divisions of data to account for the variety of sounds calls are detected.

Fin whale out-of-sample performance was highest during run 11 with four initial clusters and 14 secondary clusters. Recall was 79%, precision 71%, and accuracy was 91%. Different from the least efficient run, the first two centroids of run 11 are the noisier clusters while the last two are quieter and do not have as much background noise. Cluster #2

represents the loudest environments in the data set and cluster #3 represents the faintest calls in the data set. Cluster #4 is a good mix of calls in a quiet environment with calls in the louder environment.

Blue whale out-of-sample performance was highest during run 12 with four initial clusters and 12 secondary clusters. Performance was 94% recall, 94% precision, and 90% accuracy. Initial clusters and secondary clusters are similar to the clusters shown for run 14, the highest in-sample run. Fewer subclusters might have worked better on the testing data because it appears that several of the more noise-cluttered centroids were combined and were more inclusive in the identification process. Despite the greater performance obtained with the clusters from different runs, the out-sample performance for run 14 was expected to give the overall best results. A summary of the out-of-sample performance for run 14 is given in Table 4.

Table 4. Out-of-sample performance of run 14

Blue whale performance					**Fin whale Performance**				
		Annotations					**Annotations**		
		Blue	**Fin**	Totals			**Blue**	**Fin**	Totals
Identification	**Blue**	849	46	895	**Identification**	**Blue**	143	66	209
	Fin	66	143	209		**Fin**	46	849	895
	Totals	P=915	N=189	1104		**Totals**	P=189	N=915	1104

THIS PAGE INTENTIONALLY LEFT BLANK

V. DISCUSSION

While using the k-means clustering approach on the data, there is a possibility of under- or over-tuning the algorithm as adapted from descriptions by Schaathun (2012). Under-tuning can occur if not enough clusters are chosen to train the data and the algorithm cannot glean any insight into the true structures of the data. Under-tuning as in runs 1-3 resulted in low in-sample and out-of-sample performance. Over-tuning can occur when there is an excess of clusters used on the data set and the algorithm cannot process new images unless they look like the training data which results in a poor out-of-sample performance. Over-tuning can cause a decrease in overall performance as displayed during run 5 of both the training and testing data sets. Schaathun (2012) describes this decrease in performance as the *curse of dimensionality* where the model fits too perfectly to the training data and begins to capture overly-specific elements that are unique to samples. When the model is then given new data, it cannot classify it properly because of the introduction of different structures. The cross-validation method was used to mitigate these issues of under- and over-fitting the algorithm.

Returning to the research questions from Section I: Is there enough data to train the algorithm? Results indicate there was sufficient data to train the algorithm based on high in-sample results that translated well to give similarly high out-of-sample performance. Overall, performance of the algorithm applied to blue whale data was higher than when applied to fin whale data. This could have been due to the large volume of blue whale calls or the quality of the fin whale calls.

Is the algorithm location/season specific? Although this research only applied the algorithm to sensors in four nearby locations, it did cover all seasons. The algorithm did not need be retrained when moving to different locations or when processing data from different seasons in this similar geographic location.

Do we need to retrain if moving to a different location to classify blue and fin whale calls? The centroids generated during this research capture sounds possibly unique to the SCB. A different geographic location with a different acoustic environment and perhaps

different variability in blue and fin whales will most likely change the composition of the data set. If the algorithm is applied in a different geographic location, it would most likely need to be retrained to determine the ideal number of clusters.

What are the main obstacles in automatic classification of baleen whale calls using an unsupervised machine learning technique? From the cluster analysis, it appears the largest challenge for classification using this method remains the need to determine the ideal combination of clusters before analysis to avoid under- or over-fitting the model. Additionally, it also matters how the data is clustered and not just the overall number of clusters. As an example, in runs 3 (six initial and four secondary) and 5 (four initial and six secondary), the overall number of clusters were the same at 24. However, different results were obtained from each run.

The results could have been slightly biased because of the large number of blue whale calls and it is easier for an analyst to misidentify fin whale calls due to their short duration. The algorithm did appear to perform better with the larger number of samples. It was able to process a variety of blue whale calls with different levels of intensity and background noise. For blue whale data, it was also noted that the algorithm was more sensitive to changes in recall and precision while accuracy remained relatively steady during the training runs. For fin whale data, on the other hand, the algorithm was more sensitive to changes to accuracy while precision and recall remained relatively stable during training runs.

There are some limitations to this method. It was only performed using annotation obtained from using Triton software. Triton and the k-means algorithm are MATLAB based programs. While many educational facilities have MATLAB, it is not open source software and is not available everywhere. Other programming languages do have the ability to apply k-means clustering such as R, python, and C, which can be adapted to conduct similar clustering but would have to be compatible with current PAM files.

Another limitation that occurred in this research was the training method. The calls were separated by annotated species to create the centroids. It was easier to measure the performance of the algorithm by knowing the annotations of the calls. It was difficult to

determine performance with data containing both blue and fin whale calls because the sample would not be labeled in order to determine how the algorithm performed.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSIONS AND RECOMMENDATIONS

In conclusion, the k-means algorithm was overall effective in classifying the whale vocalizations based on their spectrograms. The best application of the algorithm appears to have a small number of initial clusters, keeping the first partitions relatively large. Then apply a larger number of secondary clustering which helps capture more specific sound features in the data set(s). The highest in-sample performance for blue whales was run 14 with four initial clusters and 15 secondary clusters resulting in 95% recall, 99% precision, and 95% accuracy. Run 14 also had a high out-of-sample performance with 93% recall, 95% precision, and 90% accuracy. The highest in-sample performance for fin whales was obtained during run 14 with ten initial clusters and seven secondary clusters resulting in 93% recall, 81% precision, and 95% accuracy. During testing, run 14 performed satisfactory with 76% recall, 68% precision and 90% accuracy. From the results, the best approach seems to be to have relatively small number of initial clusters and then spilt those clusters into several clusters. Performance metrics seemed to increase when this approach was applied. However, it is challenging to determine the optimal number of clusters without empirically testing various combinations of clusters on the training data sets that are available.

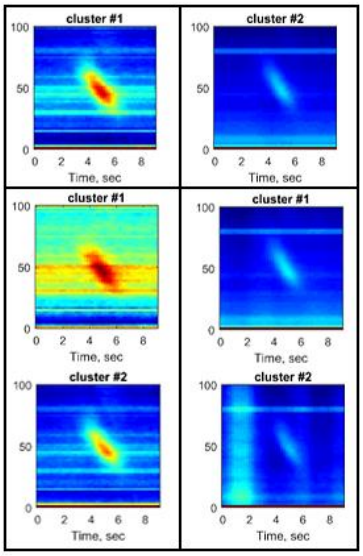
This research leaves many unexplored options in the applications of the k-means clustering algorithm. Follow-on research could include samples of ambient noise with no calls present since this research only included positive calls. Additionally, call samples from other species could be added to the data set or the algorithm could be applied to a data set of other species all together. In naval application, an attempt should be made to apply this method to a set of submarine, ship, or manmade sound sources to determine potential for classification of overlapping or difficult to identify signals. Sailors conducting undersea warfare operations can have a challenging time detecting and classifying signals of interests from ambient noise. This method might aid in the classification process of those signals by decreasing subjectivity and the amount of time a sailor might have to spend analyzing so they can move on to other targets. K-means clustering is not the only unsupervised machine learning method, but has shown

promising results in its application of identifying marine mammal baleen whales and future potential to be applied in other fields.

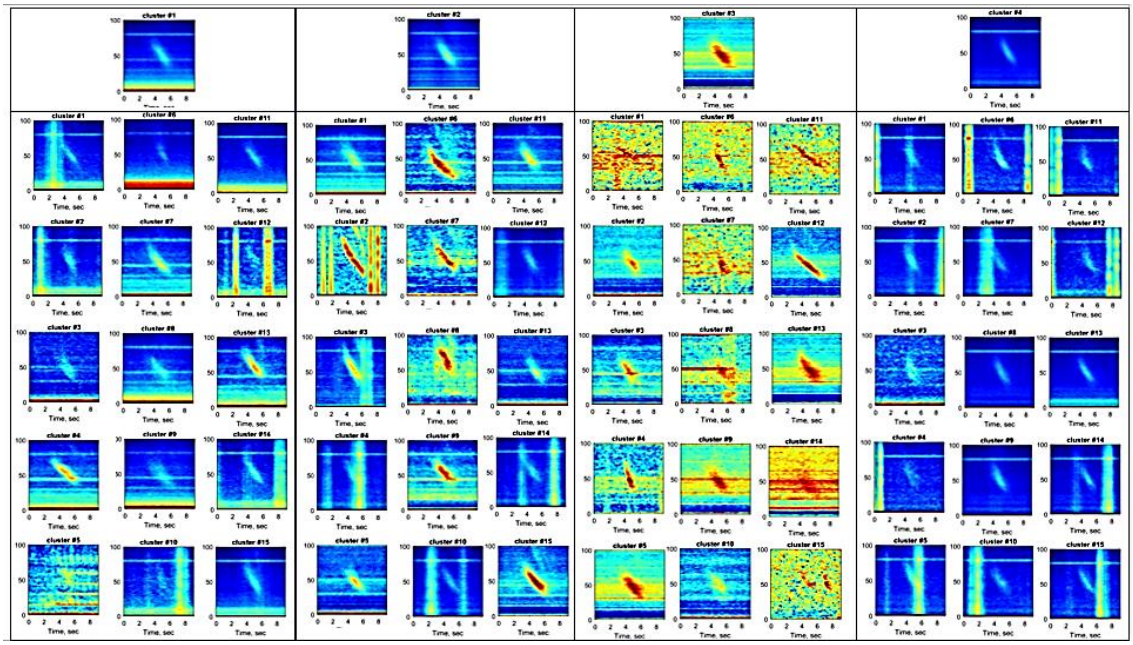
APPENDIX. CLUSTER ANALYSIS

A. BLUE WHALES: IN-SAMPLE CLUSTER ANALYSIS

1. Lowest Performing Run

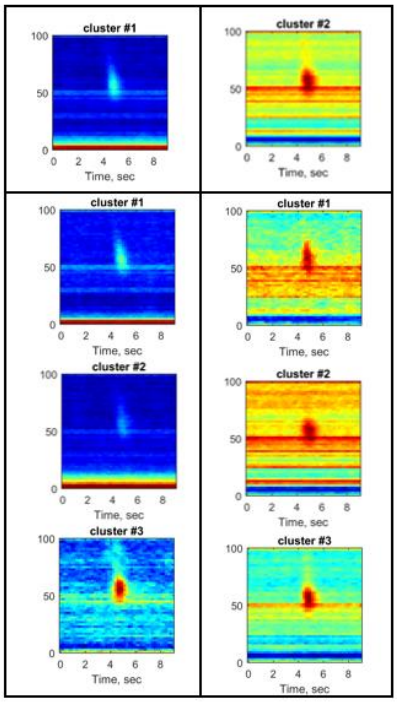


2. Highest Performing Run

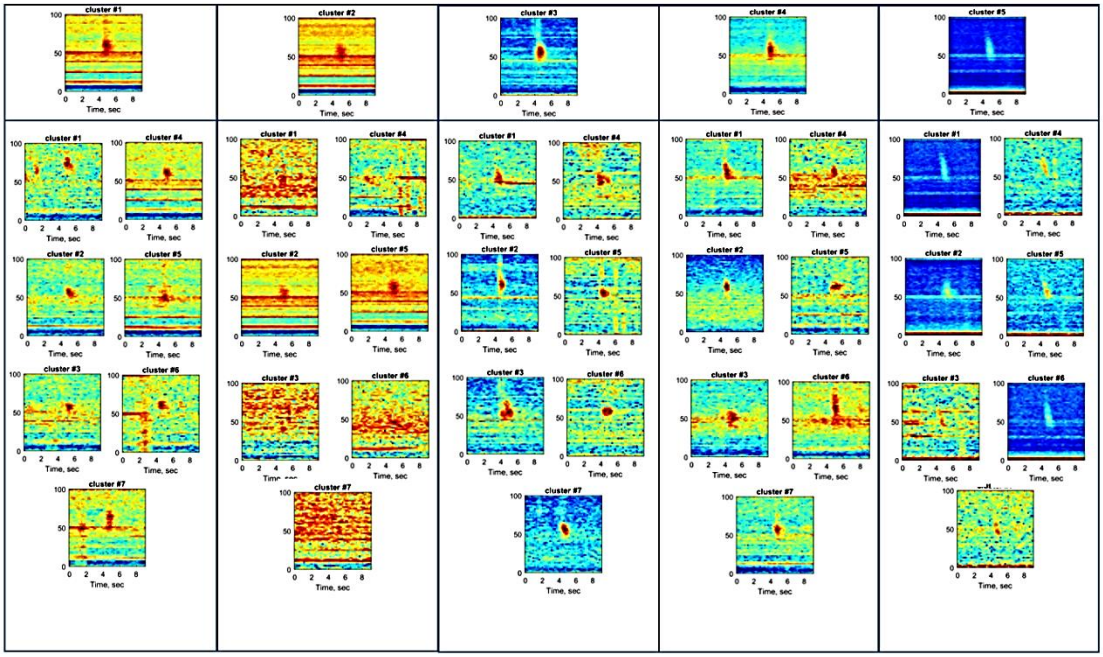


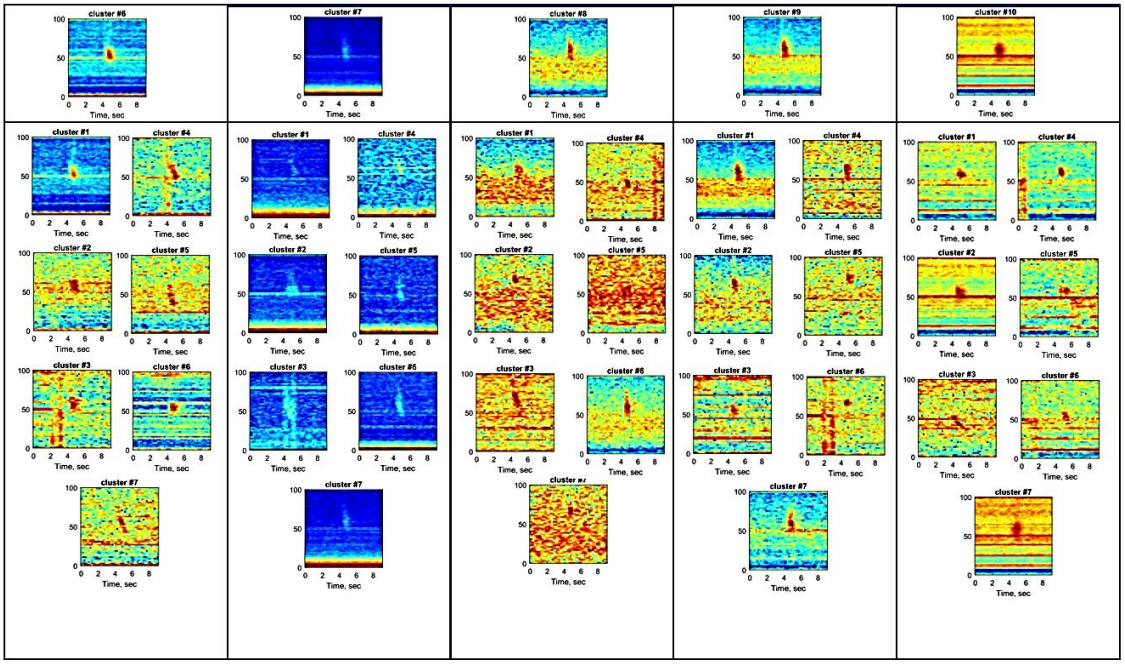
B. FIN WHALES: IN-SAMPLE PERFORMANCE

1. Lowest Performing Run



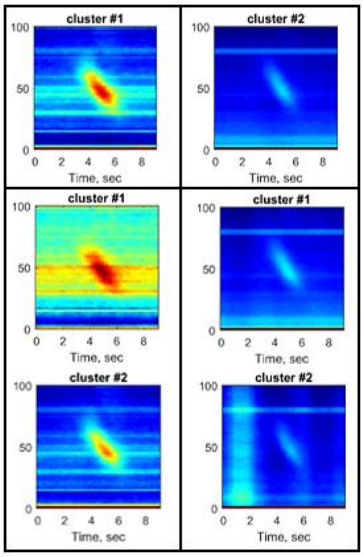
2. Highest Performing Run



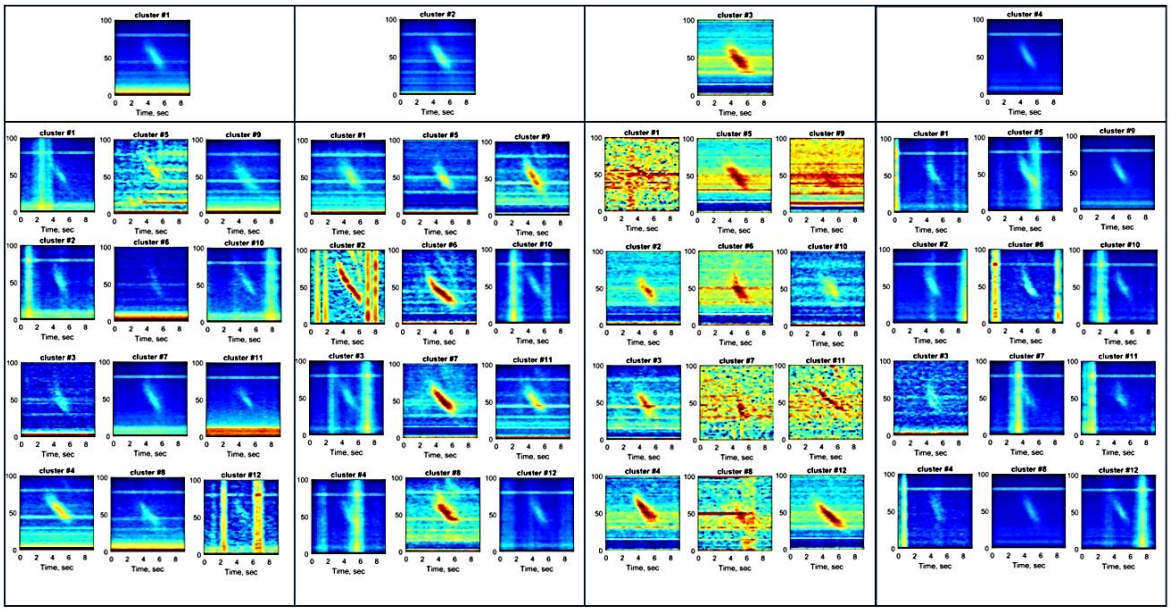

cluster #6
cluster #7
cluster #8
cluster #9
cluster #10
cluster #1
cluster #4
cluster #2
cluster #5
cluster #3
cluster #6
cluster #7
Time, sec

C. BLUE WHALES: OUT-OF-SAMPLE PERFORMANCE

1. Lowest Performing

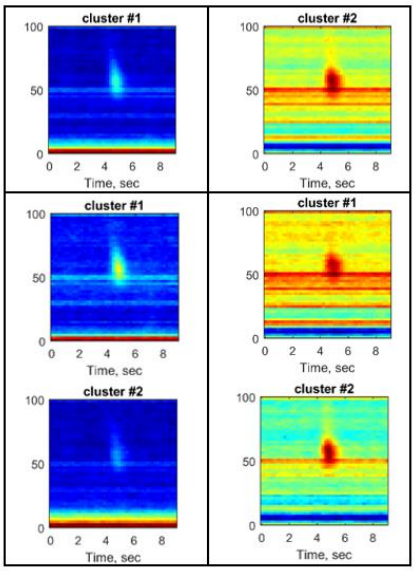


2. Highest Performing

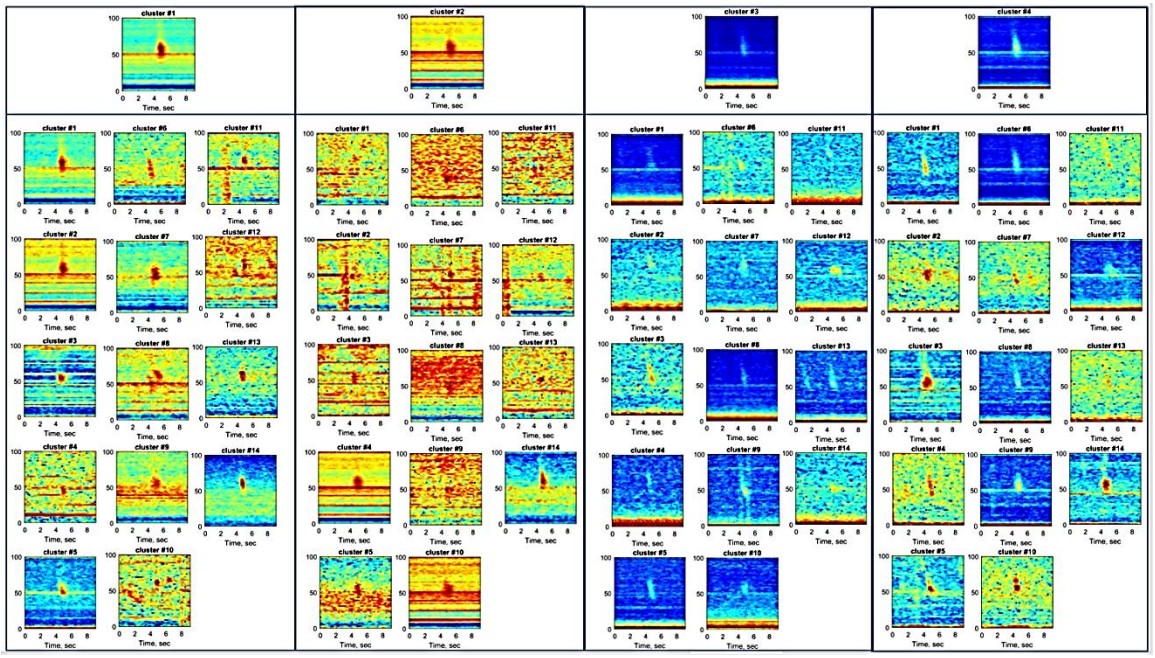


D. FIN WHALES: OUT-OF-SAMPLE PERFORMANCE

1. Lowest Performing

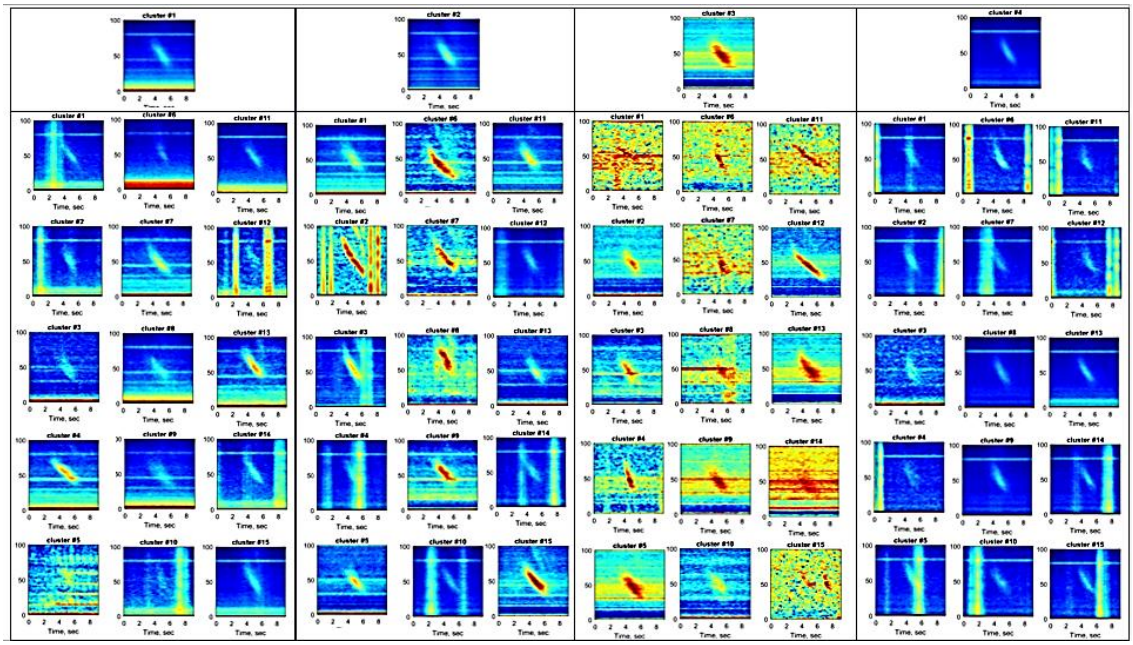


2. Highest Performing

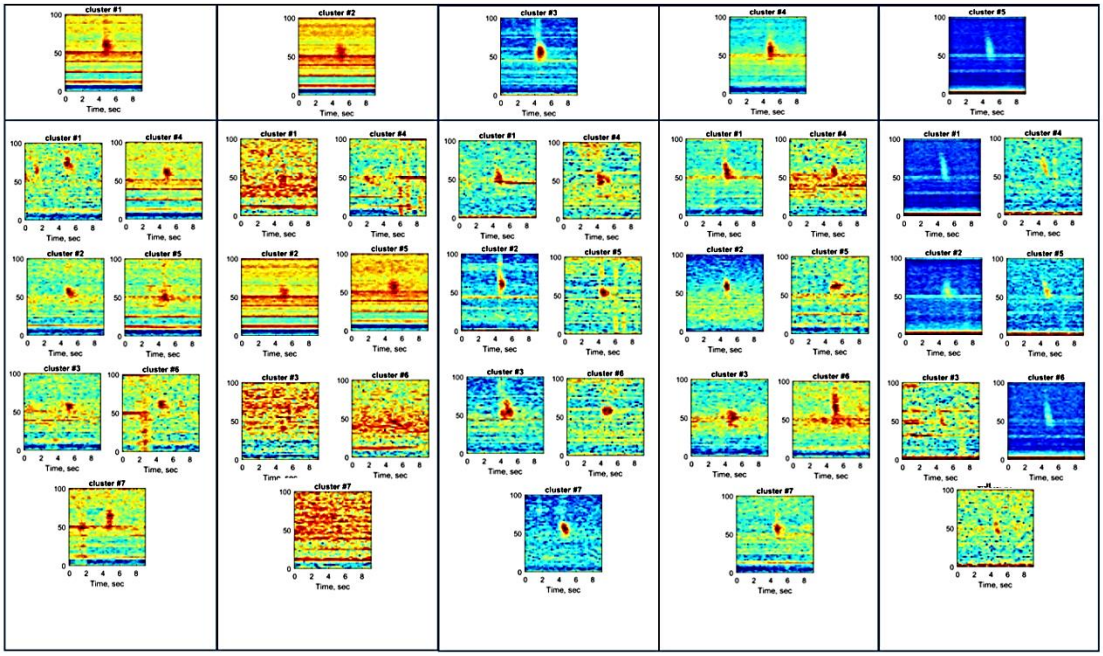


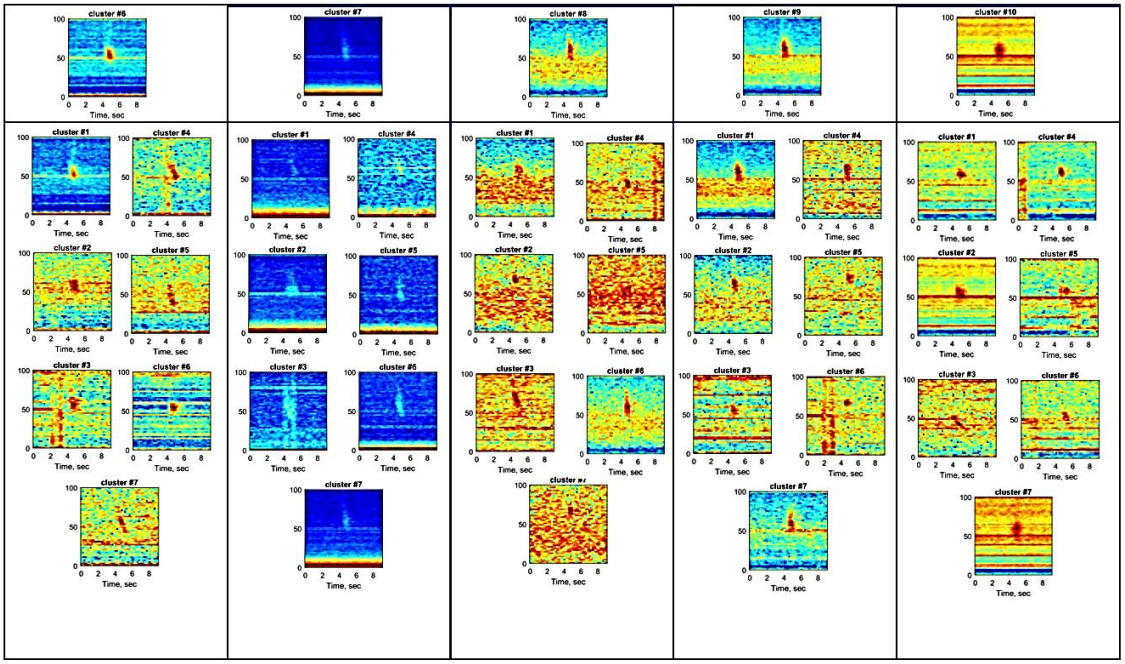
E. OUT-OF-SAMPLE PERFORMANCE: RUN 14

1. Blue Whales



2. Fin Whales




cluster #6
cluster #7
cluster #8
cluster #9
cluster #10
cluster #1
cluster #4
cluster #2
cluster #5
cluster #3
cluster #6
cluster #7
Time, sec

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Arthur, D., and S. Vassilvitskii, 2007: K-means++: the advantages of careful seeding. *Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, New York, NY, USA, 1027–1035, http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf

Aurenhammer, F., 1991: Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, **23**(3), 345–405.

Bahoura, M., 2009: Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes. *Computers in Biology and Medicine*, **39**, 824–843, doi:10.1016/j.compbiomed.2009.06.011.

Bahoura, M., Y. Simard, 2012: Serial combination of multiple classifiers for automatic blue whale calls recognition. *Expert Systems with Applications*, **39**, 9986–9993, doi:10.1016/j.eswa.2012.01.156.

Barlow, J., 1995: The abundance of cetaceans in California waters. Part I: Ship surveys in summer and fall of 1991. *Fish Bull*, **93**, 1−14.

Barlow J, K. Forney, 2007: Abundance and population density of cetaceans in the California Current ecosystem. *Fish Bull*, **105**, 509−526.

Becker EA, Forney KA, Ferguson MC, Foley DG, Smith RC, Barlow J, Redfern JV (2010) Comparing California Current cetacean−habitat models developed using in situ and remotely sensed sea surface temperature data. *Mar Ecol Prog Ser*, **413**, 163−183.

Bendersky, Eli. “Memory layout of multi-dimensional arrays” September 26, 2015. Accessed May 3, 2018. https://eli.thegreenplace.net/2015/memory-layout-of-multi-dimensional-arrays/#id1

Berman-Kowalewski, M., Gulland FMD, S. Wilkin, J. Calambokidis, B. Mate, et al. 2010: Association between blue whale (balaenoptera musculus) mortality and ship strikes along the California coast. Aquat Mam, 36, 59–66.

Briscoe, G. and T. Caelli, 1996: A Compendium of Machine Learning, **1**, 6.

Burtenshaw J.C., E.M. Oleson, J.A. Hildebrand, M.A. McDonald, R.K. Andrew, B.M. Howe, and J.A. Mercer, 2004: Acoustic and satellite remote sensing of blue whale seasonality and habitat in the Northeast Pacific. *Deep-Sea Res II*, **51**, 967−986.

Carretta, J. V., 1995: Report of 1993- marine mammal aerial survey conducted with the U.S. Navy outer sea test range off Southern California.

Carretta, J.V., M. Muto, S. Wilkin, J. Greenman, K. Wilkinson, et al., 2015: Sources of human-related injury and mortality for U.S. Pacific West Coast marine mammal stock assessments, 2009–2013. doi: 10.7289/V5/TM-SWFSC-548.

Castellote, M., C. W. Clark, and M. O. Lammers, 2012: Acoustic and behavioural changes by fin whales (Balaenoptera physalus) in response to shipping and airgun noise. *Biological Conservation,* **147**, 115–122, doi:10.1016/j.biocon.2011.12.021.

Croll, D. A., J. Gedamke, A. Acevedo, C. W. Clark, J. Urban, S. Flores, and B. Tershy, 2002: Bioacoustics only male fin whales sing loud songs. *Nature*, **417**, 809, doi: 10.1038/417809a.

Davis, J., and M. Goadrich, 2006: The relationship between precision-recall and ROC curves. *23rd International Conference on Machine Learning*, Pittsburgh, PA, International Machine Learning Society, 233-240, doi:10.1145/1143844.1143874.

Dohl, T. P., K. Norris, R. Guess, J. Bryant and M. Honig, 1980: Part I1 of summary of marine mammals and seabird survey of the Southern California Bight area, 1975-78. Cetacea of the Southern California Bight. Final Report to the Bureau of Land Management, NTIS Report No. PB81248189, 414.

Fiedler PC, Reilly SB, Hewitt RP, Demer D and others,1998: Blue whale habitat and prey in the California Channel Islands. *Deep-Sea Res II,* **45**, 1781−1801.

Flach, P.A., 2012: *Machine Learning South Asia Edition: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, 14 pp.

Forney K.A., J. Barlow, 1998: Seasonal patterns in the abundance and distribution of California cetaceans, 1991– 1992. *Mar Mamm Sci,* **14**, 460−489

Forney K.A., J. Barlow, and J.V. Carretta, 1995: The abundance of cetaceans in California waters. Part II: Aerial surveys in winter and spring of 1991 and 1992. *Fish Bull,* **93**, 15−26.

NOAA, 2012: Shipping lanes to be adjusted to protect endangered whales along California coast. Accessed 16 May 2018, https://sanctuaries.noaa.gov/news/press/2012/pr122712.html

Huang, H. C., 2016: Detection and classification of baleen whale foraging calls combining pattern recognition and machine learning techniques. MS thesis, Naval Postgraduate School, Monterey, CA.

Jain, A.K., 2008: Data Clustering: 50 Years Beyond K-Means. This paper is based on the King-Sun Fu Prize lecture delivered at the 19th International Conference on Pattern Recognition (ICPR).

Jefferson, T. A., M. Smultea, and C. Bacon, 2014: Southern California Bight marine mammal density and abundance from aerial surveys, 2008-2013. *Journal of Marine Animals and Their Ecology: Oceanographic Environmental Research Society*, **7 (2)**, 14-30.

Joseph, J. and T. Margolina, 2015: Quantifying response in vocal behavior of fin whales to local shipping in the Southern California. *The Journal of the Acoustical Society of America*, **137**, 2395. doi: 10.1121/1.4920721.

Kohavi, R, Foster Provost, 1998: Glossary of terms. *Machine Learning*, **30**, 271–274.

Lloyd, S.P., 1982: Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**, 129–137.

MacQueen, J. B., 1967: Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, 281–297.

MathWorks, 2018: k-means-2018a. Retrieved 16 January 2018 from https://www.mathworks.com/help/stats/kmeans.html#responsive_offcanvas.

Madhusudhana, S. K., M. A. Roch, E. M. Oleson, M. S. Soldevilla, and J. A. Hildebrand, 2009: Blue whale B and D call classification using a frequency domain based robust contour extractor. *OCEANS 2009 - EUROPE*, Bremen, 2009, 1–7. doi: 10.1109/OCEANSE.2009.5278220.

Margolina T., J. E. Joseph, and M. J. Huang, 2015. Object-oriented rule-based classifier for blue and fin whales. *7th International DCLDE Workshop*, La Jolla, CA, American Acoustical Society.

Margolina, T., 2010: High frequency automatic recording package data summary report PS05, August 4, 2008–January 6, 2009, NPS Project Report NPS-OC-10-003, 40 pp.

Melcón M.L., A.J. Cummins, S.M. Kerosky, L.K. Roche, S.M. Wiggins, J.A. Hildebrand, 2012: Blue whales respond to anthropogenic noise. *PLoS ONE* 7: e32681

Mellinger, D., and C. Clark, 1994: Methods for automatic detection of mysticete calls. *The Journal of the Acoustical Society of America*, **96**(5), 3317–3317. doi:10.1121/1.410749.

McDonald, M.A., J.A. Hildebrand, and S.M. Wiggins, 2006: Increase in deep ocean ambient noise in the Northeast Pacific west of San Nicolas Island, California. *Journal of the Acoustical Society of America*, **120**(2), 711–717.

National Research Council 2003. Ocean Noise and Marine Mammals National Academies Press, Washington, DC.

National Research Council 2005. Marine Mammal Populations and Ocean Noise: Determining When Noise Causes Biologically Significant Effects National Academy Press, Washington, DC.

Oleson E.M., 2005: Calling behavior of blue and fin whales off California. PhD thesis, University of California San Diego, La Jolla, CA.

Oleson, E. M., J. A. Hildebrand, J. Calambokidis, G. Schorr, and E. Falcone, 2007a: 2006 progress report on acoustic and visual monitoring for Cetaceans along the outer Washington coast, NPS Project Report NPS-OC-07-003, 30 pp.

Oleson, E. M., S. M. Wiggins, and J. A. Hildebrand, 2007c: Temporal separation of blue whale call types on a southern California feeding ground. *Animal Behaviour*, **74**, 881–894, doi:10.1016/j.anbehav.2007.01.022.

Pace, R.M. III, T.V.N. Cole, A.G. Henry, 2014: Incremental fishing gear modifications fail to significantly reduce large whale serious injury rates. *Endang Species Res* **26**, 115−126.

Rockwood, R. C., J. Calambokidis, and J. Jahncke, 2017: High mortality of blue, humpback and fin whales from modeling of vessel collisions on the U.S. West Coast suggests population impacts and insufficient protection. *PLoS ONE*, 12(8), e0183052. http://doi.org/10.1371/journal.pone.0183052

Schaathun, Hans Georg, 2012: *Machine Learning in Image Steganalysis*. Hoboken, NJ: Wiley, 290–293 pp.

Samaran, F., Guinet, C., Adam, O., Motsch, J., & Cansi, Y., 2010: Source level estimation of two blue whale subspecies in southwestern Indian Ocean. *The Journal of the Acoustical Society of America*, **127**(6), 3800. doi:10.1121/1.3409479

Scripps Institution of Oceanography, 2015: Dataset retrieval for the 2015 DCLDE workshop. Accessed 2 October 2016. [Available online at http://www.cetus.ucsd.edu/dclde/dataset.html.]

Smith P.E., and R.W. Eppley, 1982: Primary production and the anchovy population in the Southern California Bight: comparison of time-series. *Limnol Oceanogr*, **27**, 1−17.

Singh K., M. Dimple, N. Sharma, 2011: Evolving limitations in K-means algorithm in data mining and their removal. *International Journal of Computational Engineering & Management*, **12**, 105-109.

Širović, A., J. A. Hildebrand, S. M. Wiggins, M. A. McDonald, S. E. Moore, and D. Thiele, 2004: Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. *Deep-Sea Research Part II*, **51**, 2327–2344, doi:10.1016/j.dsr2.2004.08.005.

Širovic´ A., L. Williams, S. Kerosky, S. Wiggins, J. Hildebrand, 2013: Temporal separation of two fin whale call types across the eastern North Pacific. *Mar. Biol.*, **160**, 47−57, doi:10.1007/s00227-012-2061-z.

Širovic´ A., A. Rice, Chou E, S. Wiggins, J. Hildebrand, A. Marie, 2015: Seven years of blue and fin whale call abundance in the Southern California Bight. *Endangered Species Research*, **28**, 61–76, doi: 10.3354/esr00676.

Thompson P., L. Findley, W. Cummings, 1996: Underwater sounds of blue whales, Balaenoptera musculus, in the Gulf of California, Mexico. *Mar. Mamm Sci*, **12**, 288−293.

Wiggins, S.M., 2010: Engineering Tools for Studying Marine Mammals. In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2009 Symposium*. Washington, DC, The National Academies Press, 13-21, doi: 10.17226/12821.

Wiggins, S. M., E. M. Oleson, M. A. McDonald, and J. A. Hildebrand, 2005: Blue whale (Balaenoptera musculus) diel call patterns offshore of Southern California. *Aquatic Mammals*, **31**, 161–168, doi:10.1578/AM.31.2.2005.161.

Wiggins, S. M., and J. A. Hildebrand, 2007: High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring. *Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies*, 551–557, doi:10.1109/UT.2007.370760.

Wiggins, S. M., and J. A. Hildebrand, 2018: Gulf of Alaska fin whale calling behavior studied with acoustic tracking. MPL Technical Memorandum #622.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California