Development of a Classifier for the Acoustic Identification of Delphinid Species in the Northwest Atlantic Ocean

Submitted to:

Naval Facilities Engineering Command Atlantic under
HDR Environmental, Operations and Construction, Inc.
Contract No. N62470-10-D-3011, Task Order 03

HDR

Prepared By:

Bio·Waves Incorporated

Julie N. Oswald
Bio-Waves, Inc.
364 2nd Street, Ste. #3
Encinitas, CA 92024
Julie.oswald@bio-waves.net
Phone: (760) 452-2575
Fax: (760) 652-4878

27 February, 2014

Suggested Citation:

Oswald, J.N. 2013. *Development of a Classifier for the Acoustic Identification of Delphinid Species in the Northwest Atlantic Ocean. Final Report*. Prepared for Naval Facilities Engineering Command Atlantic, Norfolk, Virginia, under HDR Environmental, Operations and Construction, Inc. Norfolk, Virginia Contract No. CON005-4394-009, Subproject 164744, Task Order 003, Agreement # 105067. Prepared by Bio-Waves, Inc., Encinitas, California.

Photo: Atlantic spotted dolphins (*Stenella frontalis*) taken by Heather Foley, Duke University. Photo taken under NOAA Permit No. 16185.

Executive Summary

Two random-forest classifiers were developed for the identification of whistles produced by five species of delphinids (bottlenose dolphin, *Tursiops truncatus*; Atlantic spotted dolphin, *Stenella frontalis*; striped dolphin, *S. coeruleoalba*; short-beaked common dolphin, *Delphinus delphis*; short-finned pilot whale, *Globicephala macrorhynchus*) recorded in the northwestern Atlantic Ocean. Acoustic data were provided by the National Marine Fisheries Service's (NMFS) Northeast and Southeast Fisheries Science Centers (NEFSC and SEFSC) and Duke University. One classifier was trained and tested with whistles detected and extracted with manual based methods using the bio-acoustic analysis software ROCCA (Real-time Odontocete Call Classification Algorithm). The other classifier was trained and tested with whistles detected and extracted using the fully automated Whistle and Moan Detector (WMD). Both ROCCA and the WMD are integrated as modules in the acoustic analysis software platform, PAMGuard (www.pamguard.org). Two classification approaches were tested: a single-stage random-forest approach, where whistles were classified directly to species, and a two-stage random-forest approach, where whistles were first classified into species groups (i.e., "large dolphin" or "small dolphin") in stage 1 and then classified again, to species within those groups, in stage 2. The two-stage approach produced more accurate results when the classifier was trained and tested using manually detected/extracted whistles and when the classifier was trained/tested using automatically detected/extracted whistles. Individual whistles within an acoustic encounter were classified as 'small dolphins' (short-beaked common dolphins, striped dolphins) or large dolphins (bottlenose dolphins, Atlantic spotted dolphins, short-finned pilot whales) in stage 1 of the manual classifier, and as short-finned pilot whales or dolphins (short-beaked common, striped, Atlantic spotted, bottlenose) in stage 1 of the automated classifier. Both classifiers were used to identify individual whistles to species and then to identify encounters (i.e., groups of whistles produced during an acoustic encounter) based on the combined classification results for all of the whistles in each encounter. Overall correct classification scores for the manual classifier were 78 percent (standard deviation [sd] = 1.2 percent) for individual whistles and 86 percent (sd = 2.5 percent) for encounters. For the automated classifier, correct classification scores were 80 percent (sd = 1.9 percent) for whistles and 91 percent (sd = 2.4 percent) for encounters. Both classifiers have been incorporated into PAMGuard's ROCCA module, and will be made available to users via PAMGuard's website (www.pamguard.org) in the next PAMGuard software update. Recommendations for future research and further development of ROCCA include testing the Atlantic classifiers using recordings containing whistles from visually validated species, as well as those with varying signal to noise ratios and noise environments. ROCCA should also be tested in real time during shipboard towed-array surveys. Several recommendations are provided for future classifier development, including: ground-truthing ROCCA and adding species to the Atlantic classifier. Both of these tasks could be accomplished using acoustic recordings of delphinids collected during summer 2013 surveys conducted by NMFS-NEFSC and SEFSC. In addition, collaborations among research groups are recommended to explore methods for increasing classification success and creating multi-species classifiers for tonal signals produced by baleen whales. Finally, we recommend using a classifier developed specifically for the northwest Atlantic Ocean to process archival towed-array and seafloor-mounted recorder data to examine questions related to occurrence and behavioral responses of animals to naval activities.

This page intentionally left blank.

TABLE OF CONTENTS

TABLES

FIGURES

APPENDICES

Acronyms and Abbreviations

.csv	Comma Separated Values text file extension
.wav	Windows Waves audio file extension
dB	decibel
GUI	graphical user interface
Hz	hertz
PAM	passive acoustic monitoring
ROCCA	Real-time Odontocete Call Classification Algorithm
kHz	kilohertz
kHz/sec	kilohertz per second
NEFSC	Northeast Fisheries Science Center
nmi	nautical miles
sd	standard deviation
sec	second(s)
SEFSC	Southeast Fisheries Science Center
U.S.	United States
WMD	Whistle and Moan Detector

This page intentionally left blank.

1. INTRODUCTION

In recent decades, passive-acoustic monitoring (PAM) has been adopted as an effective method for obtaining information about the occurrence, distribution, and behavior of marine mammals (Mellinger and Barlow 2003). The extensive use and adoption of PAM for detecting and monitoring marine mammals has generated huge volumes of data. In order for the data generated from PAM to be effectively used, they need to be efficiently analyzed and accurately interpreted. This, in turn, requires acoustic analysis software that is comprehensive with respect to species that occur in a given geographic region, and that also allow for reliable automated detection and classification of vocalizations.

Species identification from acoustic recordings of marine mammal vocalizations can be challenging due to the high variability in many of the characteristics of sounds that can be easily measured or extracted from spectrograms, both within species and among species. The development of classifiers for marine mammal vocalizations is a rapidly advancing area of research. Early work on delphinid whistle classifiers focused on time-frequency characteristics measured from spectrograms and classification algorithms such as discriminant-function analysis and classification-tree analysis (e.g., Steiner 1981, Fristrup and Watkins 1993, Wang et al. 1995, Matthews et al. 1999, Rendell et al. 1999, Oswald et al. 2003). More recently, other classification algorithms such as Gaussian mixture models (Roch et al. 2007), Hidden Markov models (Brown and Smaragdis 2009) and random forests (Oswald et al. 2013) have been used, with varying degrees of success.

ROCCA (Real-time Odontocete Call Classification Algorithm) is one of a few classifiers that are readily available for the general marine mammal research, conservation and management community (Oswald et al. 2013). At present, ROCCA is available as a module within the program PAMGuard. PAMGuard is an open-source software platform that is freely available to the public to record, process, and analyze bio-acoustic data (www.pamguard.org; Gillespie et al. 2008). Currently, ROCCA contains a random-forest classifier that was developed for whistles from eight different species of delphinids occurring in the tropical Pacific Ocean (Oswald et al. 2013). Correct classification scores for all species included in this classifier are significantly greater than the 12.5 percent expected by chance alone, and range from a low of 35 percent (for short- and long-beaked common dolphin, *Delphinus* spp., whistles) to a high of 90 percent (for false killer whale, *Pseudorca crassidens,* whistles; Oswald et al. 2013).

Although ROCCA is a useful tool for the classification of whistles from delphinids occurring in the tropical Pacific Ocean, geographic variation in whistle characteristics for delphinid species limits its application in other geographic areas. For example, May-Collado and Wartzok (2008) examined whistles produced by bottlenose dolphins (*Tursiops truncatus*) in six regions in the western North Atlantic and one in the eastern North Atlantic and found that both frequency and duration parameters varied significantly between regions. Similarly, Ansmann et al. (2007) determined that short-beaked common dolphins (*Delphinus delphis*) in the English Channel produced whistles that were higher in frequency than those in the Celtic Sea. Geographic variation has also been reported for other delphinid species, such as spinner dolphin (*Stenella longirostris;* Bazua-Duran and Au 2004), Indo-Pacific bottlenose dolphin (*Tursiops aduncus;* Morisaka et al. 2005), and Atlantic spotted dolphin (*Stenella frontalis;* Baron et al. 2008). Given that whistle structure has been demonstrated to vary both between and within ocean basins, classification algorithms will be more effective when specifically trained for the locations and populations for which they will be applied.

In order to analyze the enormous volumes of data recorded during PAM efforts, classifiers also need to be efficient and user-friendly. Classification of a sound first requires that the sound is detected in a recording and then a set of features is measured from the detected sound. In the case of whistles and other frequency modulated sounds, feature measurement generally requires the extraction of time-frequency contours from spectrograms. Detection, contour extraction and feature measurement can require significant human effort and expertise. In order to reduce these requirements, the entire process should be automated to the greatest extent possible. Although automated methods can reduce significantly the time required to detect, extract and measure potential whistle contours, they are typically less accurate than manual detection methods in which an expert analyst makes the detection decision and processes the detected contours. False detections (i.e., detections of sounds other than whistles), inaccurate contour extractions, and fragmentation of whistles (i.e., where a single continuous whistle is incorrectly labeled as several shorter, separate whistles), can result in biases and inaccuracies in the outputs of automated detectors. Because of these errors, the values obtained from whistle contours extracted using manual methods can be very different than those extracted using automated methods (**Figure 1**). Therefore, the choice of method used to generate training data for classifiers is crucial. It is likely that classifiers that are intended to be used on whistles detected and extracted automatically will perform better if trained using auto-detector output data, and classifiers to be used on whistles detected and extracted manually will perform better if trained using manually extracted whistles.

Although fully automated methods require less time and user interaction, environmental, electrical (i.e. self) and other sources of noise in some datasets makes it difficult or impossible to fully automate the process. Therefore, it is important to have classifiers trained using whistles that have been manually detected, extracted and measured for these noise conditions. In this effort, we developed ROCCA classifiers for whistles produced by delphinids in the western North Atlantic Ocean. Two different classifiers were developed—one using whistles detected and extracted using manual methods and a second using whistles detected and extracted using an automated detector.

a)

b)

c)

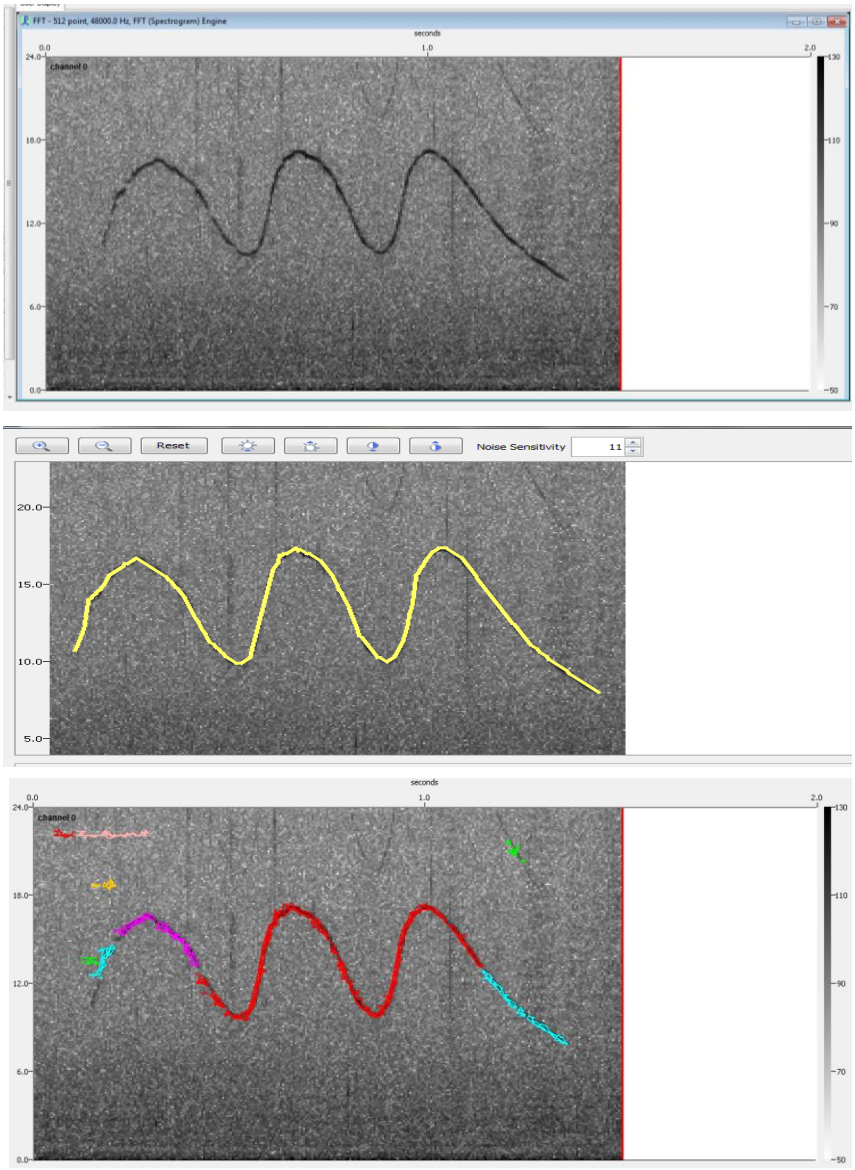

Figure 1. Example spectrograms of: (a) a dolphin whistle (without a contour outline); (b) contour manually traced and extracted using ROCCA; (c) traced and extracted automatically using the Whistle and Moan Detector (WMD) in PAMGuard, where different colors represent different individual whistles. In this example, WMD has labeled the sample whistle as four separate whistles, and false detections are shown in yellow, pink, and green.

2. METHODS

2.1 Data

Acoustic recordings of delphinid encounters were made during ship-based visual and acoustic line-transect surveys conducted by the Southeast Fisheries Science Center (SEFSC) and the Northeast Fisheries Science Center (NEFSC) of the National Marine Fisheries Service, and Duke University. The surveys took place off the Atlantic coast of the United States between central Florida and Georges Bank (in the Gulf of Maine) (**Figure 2**). The NEFSC and SEFSC surveys were several months in duration and covered large areas of the U.S. Atlantic coast. During these surveys, a team of experienced marine mammal observers searched for cetaceans using 25 × 150 binoculars, hand-held binoculars, and the naked eye. The Duke University surveys consisted of multiple one-day trips out of Onslow Bay, North Carolina and Cape Hatteras, North Carolina (Hodge 2011). For sightings during all cruises, species identification and group-size estimations were recorded. In addition, a hydrophone array was towed behind each of the research vessels during daylight hours. Acoustic signals from the arrays were monitored by acoustic technicians. Signals were monitored aurally using stereo headphones, and monitored visually from real-time scrolling spectrograms using Ishmael (Mellinger 2000) and PAMGuard (Gillespie et al. 2008) software. The frequency response characteristics of the arrays and recording equipment used during these surveys are provided in **Appendix A**.

Duke University researchers also provided acoustic data recorded with DTAGs (Digital Acoustic Recording Tags, Johnson and Tyack 2003) attached to short-finned pilot whales. Recordings from DTAGs were used only if the tagged animal was part of a single-species school of short-finned pilot whales and if there were no other species that whistle sighted within 3 nmi. Frequency response characteristics for the DTAG hydrophones are given in **Appendix A.**

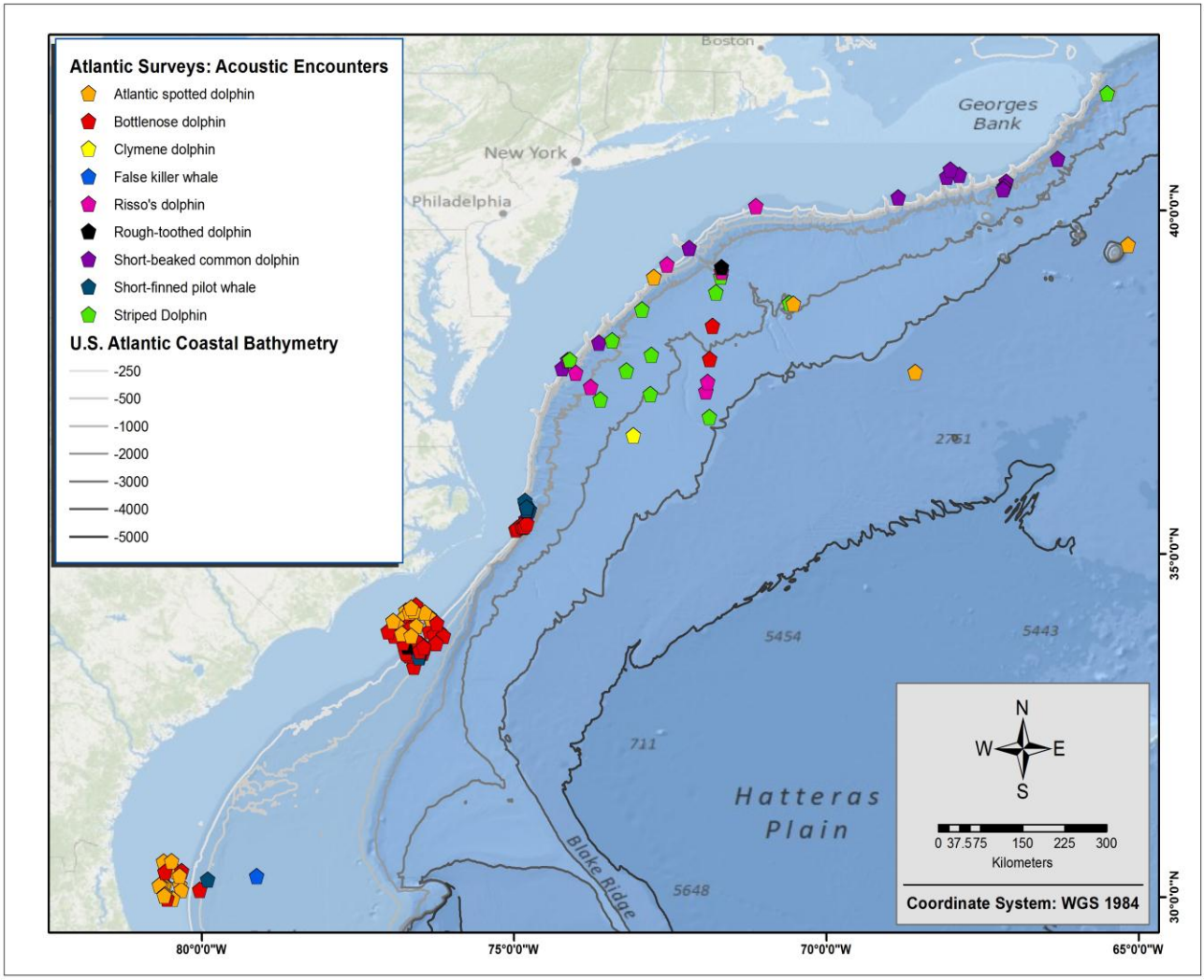

Figure 2. Western North Atlantic study area and location of recordings available for this work. Each species is represented by a different color.

2.2 Whistle Measurement

Only whistles produced by single-species delphinid schools that had visual confirmation of species identity were included in the analysis. It is possible that some recordings that were labeled as 'single species' may contain whistles produced by other species in the area. Acoustic localization was not performed to obtain exact locations of the dolphins being recorded; however the visual observers did note the latitude and longitude of each school periodically throughout the sighting. To reduce the risk of measuring whistles produced by other species in the area, encounters were only analyzed if the school was greater than 3nmi from sightings of any other whistling species. Based on work done by Rankin et al. (2008) we assumed that whistles produced by a school that was greater than 3nmi from the school being recorded would not be detected. The distance to the acoustic detection was calculated between the previous sighting (or multiple sightings if they were close together) and the position of the school in

question at the beginning of the recording. The distance was also calculated between the next sighting (or multiple sightings if they were close together) and the position of the school in question at the end of the recording. Acoustic recordings from all acoustic encounters that met the above criteria for including in the analysis were processed twice, once using manual methods for whistle detection and contour extraction and a second time using automated methods for whistle detection and contour extraction.

2.2.1 Manual Detection and Contour Extraction

Recordings from each acoustic encounter included in the analysis were examined aurally and visually using Raven Pro: Interactive Sound Analysis Software (Version 1.3; Cornell Bioacoustics Research Program 2008). Start times were noted for all whistles with a signal-to-noise ratio of 6dB or greater. Overlapping whistles were included only if each contour could be traced unambiguously. Selected whistles were saved as individual audio files in the ".wav" format for archival purposes. A maximum of 50 whistles was selected per encounter to avoid over-sampling of groups or individuals.

Time-frequency contours were extracted from spectrograms using the ROCCA module in PAMGuard. First, the start and end points of the whistle were manually selected by the operator by windowing it using with a pointing device on the computer. Next, ROCCA automatically extracted the whistle contour by stepping through the spectrogram one time slice at a time and searching for the peak frequency within a user-defined frequency band centered around the peak frequency in the previous time slice (see Oswald et al. 2007 and Barkley et al. 2011 for details). Once a contour was extracted, it was displayed on the spectrogram and the accuracy of the extraction could be adjusted by applying high-pass and/or low-pass filters, adjusting ROCCA's sensitivity to noise in the recording, and/or using the cursor to manually drag contour points to the correct location on the spectrogram (Oswald et al. 2013).

2.2.2 Automated Detection and Contour Extraction

Automated whistle detection and contour extraction was performed using the Whistle and Moan Detector (WMD) module within PAMGuard. The WMD automatically detects and extracts whistle contours by searching for spectral peaks within a user-specified frequency band. In order to be considered a true whistle detection, the spectral peak needs to occur within certain user-defined parameters relating to its amplitude and frequency in relation to other candidate spectral peaks detected the time-slices directly before and after the peak in question. For each acoustic encounter, parameters within the WMD module were adjusted manually via a GUI within the WMD to maximize accuracy of contour extraction and minimize false positives (**Figure 3,** see help files within PAMGuard for details on changing WMD parameters).

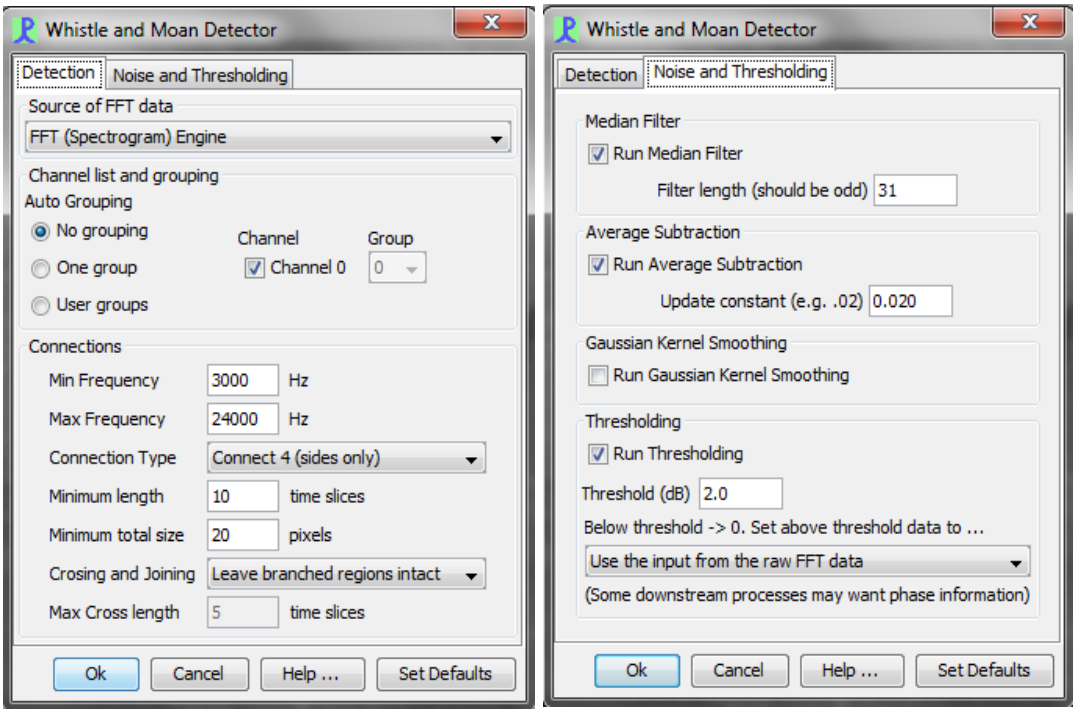

Figure 3. Parameters that can be set in PAMGuard's whistle and moan detector.

2.2.3 Feature Measurement

After whistle contours had been extracted they were saved as comma-separated text (.csv) files and then used as inputs to ROCCA for feature measurements. Fifty variables from each contour were automatically measured, including: duration, frequencies (e.g., minimum, maximum, beginning, ending, and at various points along the whistle), slopes, and variables describing shape of the whistles (see **Appendix B** and Barkley et al. 2011 for a complete list and description of variables measured).

2.2.4 Random-Forest Analysis

Classification algorithms were developed using random-forest statistical classification methods. A random forest is a collection of decision trees grown using binary partitioning of the data. Each binary partition of the data is based on the value of one feature (or in this case, a whistle variable; Breiman 2001). The goal for each split is to divide the data into two nodes, each as homogeneous as possible (i.e., containing whistles from the smallest number of species possible). Randomness is introduced into the tree-growing process by examining a random subsample of all of the features at each node. The feature that produces the most homogeneous split is chosen at each partition. When whistle features are run through a random forest, each of the trees in the forest produces a species classification. Each tree can be considered 1 'vote' for a given species classification. Votes are then tallied over all trees and the whistle classification is based on the species with the most 'votes'. In addition to classifying individual whistles, entire acoustic encounters were classified based on the number of tree classifications for each species, summed over all of the whistles that were analyzed for that encounter.

The number of tree classifications for the predicted species was also used as a measure of the certainty of the classification. It was assumed that if a greater percentage of trees classified the whistle as a

particular species, then that classification had a higher degree of certainty. Based on this assumption, a 'strong whistle threshold' was defined. If the percentage of trees that classified the whistle as a particular species was greater than this strong whistle threshold, the whistle was considered strongly classified, or simply 'strong' (Oswald et al. 2011). If the percentage of trees that classified the whistle as a particular species did not exceed the strong whistle threshold, then the classification was considered unreliable and the whistle was labeled as 'ambiguous.' Higher strong whistle thresholds generally resulted in higher correct classification scores, but also resulted in more whistles being labeled as ambiguous because fewer classifications will meet the strong whistle threshold. If all of the whistles within a single encounter were labeled as ambiguous, then that encounter was also classified as ambiguous and could not be classified. Strong whistle thresholds were chosen to maximize correct classification scores while minimizing the number of encounters that were labeled as ambiguous. In this project, strong whistle thresholds that allowed at least 90 percent of encounters to be classified were chosen.

To test the random-forest models, each dataset was randomly subsampled so that there was a maximum of 50 contours per encounter for the manual dataset, and 100 contours per encounter for the auto-detector dataset. Because of the higher overall sample size resulting from fragmented whistles, it was possible to use a greater number of contours per encounter for the auto-detector dataset (i.e., single whistles that were broken into multiple fragments) and false detections. Next, these datasets were subsampled so that each contained an equal number of contours per species. This avoided any one species dominating the data and skewing the results. The subsampled dataset was randomly divided in two, with whistles from the same encounter kept together in the same dataset. One dataset was used to train the model, while the other was used to test the model. The datasets were then swapped so that each was used both as a training and a testing set. This procedure was repeated 10 times in order to produce means and standard deviations for the confusion matrices.

Two random-forest models were trained, one using manually detected whistles with contours extracted by ROCCA and a second using whistles both automatically detected and extracted using the WMD. For each model, two different random-forest models were explored. In the first model, whistle contours were classified directly to species. The second model used two-stages, in which whistle contours were first classified to a broad species-group (such as 'small-sized delphinid' or 'medium-sized delphinid,') in stage one and then classified to species within that species-group in stage two. For each two-stage model, several different sets of species-groups were explored for stage one. To prevent whistle measurements from any one species from dominating the data and skewing the classification results, training datasets first were subsampled to give an equal number of whistles per species-group in stage one, and subsampled again to give an equal number of whistles per species in stage two. The manual and automated classifiers that produced the most accurate results were used for ROCCA module in PAMGuard.

2.3 Variable Importance

One of the outputs of a random-forest analysis is an estimate of variable importance and provides a relative measure of the degree to which each variable contributes to the random forest model predictions. This measure of variable importance is uses the Gini Index, which is a measure of the 'purity' of each node in a classification tree (Breiman et al. 1984). In our case, purity refers to the number of species represented in a node. Smaller Gini Index values represent increases in purity.

Splitting variables are chosen at each node so that the resulting subsets minimize the combined Gini Index as possible (Oh et al. 2003). To evaluate variable importance, decreases in the Gini Index from one node to the next are summed for each variable over all of the trees. This sum is known as the 'Gini Importance'. Variables with higher Gini Importance values contribute more to the random-forest model predictions than do those with lower Gini Importance values. In this project, Gini importance values were averaged over the 10 random-forest runs described previously (in **Section 2.2.4**) to evaluate which features were most important to the classification models.

This page intentionally left blank.

3. RESULTS

3.1 Acoustic Recordings

Acoustic recordings of single-species schools that met the criteria for analysis were available for nine delphinid species. The numbers of acoustic encounters and, the numbers of whistle contours detected manually and automatically for each species are compiled in **Table 1**. For the short-finned pilot whale, 6 of the 15 encounters were recorded using DTAGs. In general, the number of contours much greater for the auto-detector (n = 5,027) than for the manual method (n = 3,525) because the auto-detector fragmented some whistles, causing those whistles to be counted more than once. The auto-detector also produced false detections that were used as whistle contours in this project. Only species with at least four encounters and 200 manually detected whistle contours were included in the analysis. This was the minimum amount of data that we considered to be adequate for reliable training and testing of classifiers. Because of these strict criteria, data from only the following five species were used: short-beaked common dolphin, striped dolphin (*Stenella coeruleoalba*), Atlantic spotted dolphin, bottlenose dolphin, and short-finned pilot whale (*Globicephala macrorhynchus*).

Table 1. Numbers of acoustic encounters per species and total numbers of whistle contours for each species detected using ROCCA (Manually Detected) and using PAMGuard's WMD (Auto-detected).

Species	Encounters	Whistle Contours	
		Manually Detected	Auto-detected
Bottlenose dolphin	74	1,632	1,719
Atlantic spotted dolphin	45	706	988
Striped dolphin	12	293	648
Short-finned pilot whale	15	259	749
Short-beaked common dolphin	9	249	475
Risso's dolphin	8	119	99
Clymene dolphin	2	99	64
Rough-toothed dolphin	3	98	109
False killer whale	2	70	176
Total	**170**	**3,525**	**5,027**

3.2 Manual Classifier

3.2.1 Single-Stage Classifier

When whistles were classified with the single stage classifier to species, a mean of 60 percent (standard deviation [sd] = 1.1 percent) of whistles and 66 percent (sd = 1.5 percent) of encounters were classified correctly (strong whistle threshold = 40 percent). Confusion matrices for both individual whistles and overall encounters are provided in **Table 2**. Patterns in classifications were similar between the two,

with the highest percentage of correct classification scores occurring for short-finned pilot whales (76 percent and 88 percent for whistles and encounters, respectively) and the lowest percentage of correct classification scores attributed to striped dolphins (39 percent and 43 percent for whistles and encounters, respectively). Striped and short-beaked common dolphin whistles were most often misclassified as each other, bottlenose dolphin whistles were most often misclassified as Atlantic spotted dolphins, and Atlantic spotted dolphin whistles were most often misclassified as either short-finned pilot whales or bottlenose dolphins. Short-finned pilot whale classification errors were spread relatively evenly across the other four species.

Table 2. Confusion matrices for the single-stage classifier trained using manually detected and extracted whistles. The percent of whistles correctly classified for each species is in bold, with standard deviations in parentheses. A) Individual whistles. Overall, 60 percent (sd = 1.1 percent) of whistles were correctly classified when the strong whistle threshold was 40 percent. Sample size (n) is the number of whistles that were strongly classified. B) Encounters. Overall, 65.9 percent (sd = 1.5 percent) of encounters were correctly classified when the strong whistle threshold was 40 percent. Sample size (*n*) is the number of encounters that could be classified based on strong whistles alone.

A)

Actual species	Percent classified as					
	Short-beaked common dolphin	Short-finned pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**49.2 (0.8)**	0.2 (0.4)	33.0 (2.4)	5.3 (1.3)	12.3 (1.9)	202 (5)
Short-finned pilot whale	3.7 (0.7)	**76.1 (1.4)**	6.4 (1.3)	9.0 (1.6)	4.6 (1.2)	210 (5)
Striped dolphin	35.6 (4.0)	4.1 (0.9)	**39.0 (4.2)**	8.7 (1.1)	12.8 (1.8)	192 (5)
Atlantic spotted dolphin	0.8 (0.8)	11.5 (1.4)	3.1 (0.3)	**69.8 (3.8)**	14.4 (3.1)	201 (6)
Bottlenose dolphin	5.0 (1.9)	4.5 (0.8)	5.5 (1.4)	19.3 (3.7)	**66.0 (4.1)**	203 (5)

B)

Actual species	% Classified as					
	Short-beaked common dolphin	Short-finned pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**44.0 (0)**	0 (0)	19.8 (6.9)	22.0 (0)	13.2 (6.9)	9 (0)
Short-finned pilot whale	6.0 (0)	**88.0 (0)**	0 (0)	5.4 (1.9)	0.6 (1.9)	16 (0)
Striped dolphin	27.5 (8.6)	0.8 (2.5)	**43.4 (6.6)**	16.1 (2.8)	12.4 (6.1)	12 (0)
Atlantic spotted dolphin	0 (0)	7.1 (2.8)	0.9 (1.4)	**81.7 (4.3)**	10.5 (4.4)	38 (2)
Bottlenose dolphin	3.9 (1.9)	3.2 (1.6)	3.4 (1.6)	16.9 (5.0)	**72.5 (4.7)**	59.1 (2.1)

3.2.2 Two-Stage Classifier

Based on the confusion matrices for the single-stage classifier, several two-stage classifiers were tested, each with different combined species-groups in stage 1. Examples of species-groups that were tested include:

1. Short-finned pilot whales versus dolphins (short-beaked common, striped, Atlantic spotted, bottlenose)
2. Small dolphins (short-beaked common and striped) versus large dolphins (Atlantic spotted, bottlenose and short-finned pilot whales)
3. Small dolphins versus medium dolphins (Atlantic spotted and bottlenose) versus short-finned pilot whales

The small dolphins versus large dolphins combination produced the highest correct classification scores. When contours were classified to small versus large dolphins in stage 1, and then to species in stage 2, average overall correct classification scores were 78 percent (sd = 1.2 percent) for individual whistles and 86 percent (sd = 2.5 percent) for encounters (strong whistle threshold = 50 percent). Confusion matrices for these classifiers are given in **Table 3**.

When compared to the single-stage results, correct classification scores for whistles in the two-stage classifier were higher for every species. These differences were statistically significant (Fisher's exact test, $\alpha = 0.05$) for every species except for bottlenose dolphin and Atlantic spotted dolphin (**Table 4**). For encounters, correct classification scores for the two-stage classifier were significantly greater (Fisher's exact test, $\alpha = 0.05$) than the single-stage classifier for striped and short-beaked common dolphins.

Table 3. Confusion matrices for the two-stage classifier trained using manually detected and extracted whistles. The percent of whistles correctly classified for each species is in bold, with standard deviations in parentheses. A) Confusion matrix for individual whistles. Overall, 78 percent (sd = 1.2 percent) of whistles were correctly classified when the strong whistle threshold was 50 percent. Sample size (n) is the number of whistles that were strongly classified. B) Confusion matrix for overall encounters. Overall, 86 percent (sd = 2.5 percent) of encounters were correctly classified when the strong whistle threshold was 50 percent. Sample size (*n*) is the number of encounters that could be classified based on strong whistles alone.

A)

Actual species	% Classified as					
	Short-beaked common dolphin	Short-finned pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**85.5 (2.5)**	0.9 (0.3)	0 (0)	3.5 (0.5)	9.9 (2.4)	244 (2)
Short-finned pilot whale	2.6 (0.8)	**86.4 (1.6)**	11.2 (1.3)	0 (0)	0 (0)	181 (0)
Striped dolphin	2.6 (1.2)	3.8 (0.6)	**77.8 (1.7)**	7.6 (1.1)	8.3 (2.2)	279 (2)
Atlantic spotted dolphin	1.3 (0.8)	3.8 (1.1)	11.2 (2.3)	**77.6 (3.7)**	6.2 (2.3)	166 (4)
Bottlenose dolphin	5.6 (1.9)	3.7 (1.8)	18.5 (2.5)	8.3 (2.9)	**63.8 (2.6)**	164 (4.0)

B)

Actual species	% Classified as					
	Common dolphin	Pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**84.6 (5.7)**	0 (0)	0 (0)	11 (0)	4.4 (5.7)	9 (0)
Pilot whale	0 (0)	**94.6 (1.9)**	5.4 (1.9)	0 (0)	0 (0)	16 (0)
Striped dolphin	0 (0)	0 (0)	**91.1 (2.8)**	8.0 (0)	0.8 (2.5)	12 (0)
Atlantic spotted dolphin	0.3 (0.9)	2.5 (2.6)	4.8 (3.1)	**90 (6.6)**	2.5 (2.6)	36.7 (1.6)
Bottlenose dolphin	5.4 (2.2)	2.7 (1.6)	14.7 (4.2)	7.2 (2.9)	**70 (4.3)**	56.5 (2.4)

Table 4. Percentages of whistles and encounters correctly classified (with standard deviation in parentheses) for single-stage and two-stage classifiers trained using manually detected and extracted whistles and using whistles detected and extracted automatically. P-values are for Fisher's exact test comparing single-stage correct classification scores to two-stage correct classification scores for each species and dataset. Significant differences are shown with an asterisk.

Species	% correct classification - manual						% correct classification - automated					
	whistles			encounters			whistles			encounters		
	single-stage	two-stage	p	single-stage	two-stage	p	single-stage	two-stage	p	single-stage	two-stage	p
Short-beaked common dolphin	49.2 (0.8)	85.5 (2.5)	<0.0001*	44.0 (0)	84.6 (5.7)	<0.0001*	37.0 (0.8)	85.6 (3.4)	<0.0001*	25.8 (1.7)	95.2 (6.2)	0.007*
Short-finned pilot whale	76.1 (1.4)	86.4 (1.6)	0.01*	88.0 (0)	94.6 (1.9)	1	81.3 (1.9)	83.6 (0.8)	0.383	94.0 (2.8)	95.2 (2.5)	1
Striped dolphin	39.0 (4.2)	77.8 (1.7)	<0.0001*	43.4 (6.6)	91.1 (2.8)	<0.0001*	49.3 (1.6)	71.8 (3)	<0.0001*	55.0 (0)	87.9 (7.8)	<0.0001*
Atlantic spotted dolphin	69.8 (3.8)	77.6 (3.7)	0.098	81.7 (4.3)	90 (6.6)	0.516	85.0 (1.5)	77.2 (1.6)	0.037*	93.8 (2.8)	89.7 (5.8)	0.668
Bottlenose dolphin	66.0 (4.1)	63.8 (2.6)	0.74	72.5 (4.7)	70 (4.3)	0.838	88.6 (1.8)	84.4 (2.9)	0.58	88.7 (2.8)	88.9 (2.2)	1

3.3 Automated Classifier

3.3.1 Single-Stage Classifier

Using the single stage classifier, when whistles were classified directly to species, overall means of 68 percent (sd = 0.7 percent) of whistles and 71 percent (sd = 0.8 percent) of encounters were correctly classified (strong whistle threshold = 45 percent). Confusion matrices for both individual whistles and overall encounters are provided in **Table 5**. Correct classification scores for Bottlenose dolphin, Atlantic spotted dolphin, and short-finned pilot whale all were greater than 80 percent for whistles and close to, or greater than, 90 percent for encounters. Correct classification scores were lowest for short-beaked common dolphins (whistles: 37 percent, sd = 0.8 percent; encounters: 26 percent, sd = 1.7 percent). Short-beaked common dolphin whistles were most often misclassified as bottlenose, striped, or Atlantic spotted dolphin.

Table 5. Confusion matrices for the single-stage classifier trained using automatically detected and extracted whistles. The percentages of whistles correctly classified for each species is presented in bold, with standard deviations in parentheses. A) Confusion matrix for individual whistles. Overall, 68.2 percent (sd = 0.7 percent) of whistles were correctly classified when the strong whistle threshold was 45 percent. Sample size (n) is the number of contours that were strongly classified. B) Confusion matrix for overall encounters. Overall, 71.5 percent (sd = 0.8 percent) of encounters were correctly classified when the strong whistle threshold was 45 percent. Sample size (*n*) is the number of encounters that could be classified based on strong whistles alone.

A)

Actual species	% Classified as					
	Common dolphin	Pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	n
Short-beaked common dolphin	**37.0 (0.8)**	4.2 (0.6)	19.5 (1.1)	14.5 (1.8)	24.5 (1.2)	299 (7)
Pilot whale	1.6 (0.7)	**81.3 (1.9)**	3.9 (0.9)	8.5 (0.5)	4.6 (0.9)	343 (8)
Striped dolphin	13.5 (1.4)	12.6 (0.9)	**49.3 (1.6)**	10.1 (0.9)	14.5 (1.3)	294 (6)
Atlantic spotted dolphin	1.2 (0.6)	5.9 (1.1)	2.2 (0.6)	**85 (1.5)**	6.0 (0.9)	311 (9)
Bottlenose dolphin	2.8 (0.9)	1.8 (0.8)	1.4 (0.7)	5.1 (1.4)	**88.6 (1.8)**	328 (14)

B)

Actual species	% Classified as					
	Common dolphin	Pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	n
Short-beaked common dolphin	**25.8 (1.7)**	0 (0)	10.4 (5.5)	20.7 (6.5)	44.0 (7.1)	8 (0.4)
Pilot whale	0 (0)	**94.0 (2.8)**	0 (0)	6.0 (2.8)	0 (0)	17 (0)
Striped dolphin	0 (0)	13.5 (4.7)	**55.0 (0)**	14.4 (4.6)	17.1 (6.6)	11 (0)
Atlantic spotted dolphin	0 (0)	2.0 (1.8)	0.9 (1.4)	**93.8 (2.8)**	3.6 (2.4)	35 (2)
Bottlenose dolphin	0.6 (0.9)	0 (0)	0.6 (0.9)	10.1 (3.5)	**88.7 (2.8)**	57 (1)

3.3.2 Two-Stage Classifier

Results of the species-groups tested in stage 1 of the two-stage classifier were similar to those described for the manual classifier (see **Section 3.2.2**). The combination that produced the highest correct classification scores for the auto-detector data was short-finned pilot whales versus dolphins (short-beaked common, striped, Atlantic spotted, bottlenose). When test data were run through this two-stage classifier, mean overall correct classification scores were 80 percent (sd = 1.9 percent) for individual whistles and 91 percent (sd = 2.4 percent) for encounters (strong whistle threshold = 45 percent). For this classifier, correct classification scores for whistles were above 70 percent for all species and close to 85 percent for short-beaked common dolphins, bottlenose dolphins, and short-finned pilot whales. Correct classification scores for encounters were close to or above 90 percent for every species (**Table 6**).

The two-stage classifier resulted in statistically significant increases in correct classification of whistles for every species except short-finned pilot whales and bottlenose dolphins (Fisher's exact test, $\alpha = 0.05$; **Table 4**). For overall encounters, correct classification scores increased significantly (Fisher's exact test, $\alpha = 0.05$) for short-beaked common dolphins and striped dolphins (**Table 4**). For short-beaked common dolphins, mean correct classification scores increased from 37.0 percent (sd = 0.8 percent) to 85.6 percent (sd = 3.4 percent) for whistles and from 25.8 percent (sd = 1.7 percent) to 95.2 percent (sd = 6.2 percent) for encounters when using the two-stage classifier. For striped dolphins, correct classification scores increased from 49.3 percent (sd = 1.6 percent) to 71.8 percent (sd = 3.0 percent) for whistles and from 55.0 percent (sd = 0 percent) to 87.9 percent (sd = 7.8 percent) for encounters when using the two-stage classifier.

Table 6. Confusion matrices for the two-stage classifier trained using automatically detected and extracted whistles. The percent of whistles correctly classified for each species is in bold, with standard deviations in parentheses. A) Confusion matrix for individual whistles. Overall, 80.5 percent (sd = 1.9 percent) of whistles were correctly classified when the strong whistle threshold was 45 percent. Sample size (n) is the number of whistles that were strongly classified. B) Confusion matrix for overall encounters. Overall, 91.4 percent (sd = 2.5 percent) of encounters were correctly classified when the strong whistle threshold was 45 percent. Sample size (*n*) is the number of encounters that could be classified based only on strong whistles.

A)

Actual species	% Classified as					
	Short-beaked common dolphin	Short-finned pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**85.6 (3.4)**	14.4 (3.4)	0 (0)	0 (0)	0 (0)	188 (0)
Short-finned pilot whale	2.0 (0.5)	**83.6 (0.8)**	2.9 (0.6)	7.4 (0.7)	4.0 (0)	695 (8)
Striped dolphin	1.4 (0.8)	22.5 (2.8)	**71.8 (3.0)**	1.7 (1.2)	3.1 (1.4)	177 (3)
Atlantic spotted dolphin	0.8 (0.6)	18.8 (1.3)	1.0 (0.7)	**77.2 (1.6)**	2.1 (1.4)	176 (3)
Bottlenose dolphin	2.0 (0.9)	9.6 (2.0)	1.1 (0.7)	2.7 (1.2)	**84.4 (2.9)**	160 (5)

B)

Actual species	% Classified as					
	Short-beaked common dolphin	Short-finned pilot whale	Striped dolphin	Atlantic spotted dolphin	Bottlenose dolphin	*n*
Short-beaked common dolphin	**95.2 (6.2)**	5.2 (6.7)	0 (0)	0 (0)	0 (0)	8 (0)
Short-finned pilot whale	0 (0)	**95.2 (2.5)**	0 (0)	4.8 (2.5)	0 (0)	17 (0)
Striped dolphin	0 (0)	12.1 (7.8)	**87.9 (7.8)**	0 (0)	0 (0)	11 (0.4)
Atlantic spotted dolphin	0 (0)	8.9 (5.3)	0 (0)	**89.7 (5.8)**	1.2 (1.5)	32 (2)
Bottlenose dolphin	0.8 (1.4)	5.6 (2.5)	0.4 (0.8)	4.1 (2.0)	**88.9 (2.2)**	49 (3)

3.4 Whistle Measurements

3.4.1 Manual Measurements

Descriptive statistics for features that were important in the classifiers based on Gini Importance values are presented in **Tables 7-9**. The Gini Importance values indicated that features that characterize the slope and shape of whistles were most important in the single-stage classifier (**Table 10**). The range for the slope of whistles was large when compared among species (**Table 9**). For example, mean positive slope ranged from 101.5 kilohertz per second (kHz/s) (sd = 44.5 Hz/sec) for Atlantic spotted dolphins to 35.9 kHz/s (sd = 37.3 Hz/s) for short-beaked common dolphins. Features describing slope were similar for short-beaked common dolphins and striped dolphins, and for Atlantic spotted dolphins and bottlenose dolphins.

For the two-stage classifier, a different set of variables was most important in each stage. The most important variables in stage 1 (small dolphins versus large dolphins) were similar to the variables that were important in the single-stage classifier (**Tables 10, 11**). In stage 2, the small dolphins (short-beaked common and striped) were classified based on a mix of shape, slope, and frequency variables (**Table 11**). Striped dolphins produced whistles that were shorter and steeper with fewer inflection points than short-beaked common dolphins (**Tables 8, 9**). The most important features in the large dolphin classifier (Atlantic spotted dolphins, bottlenose dolphins, and short-finned pilot whales) were frequency variables (**Table 11**). Short-finned pilot whales produced the lowest frequency whistles and bottlenose dolphins produced the highest frequency whistles (**Table 7**).

Table 7. Descriptive statistics (mean, standard deviation, minimum, maximum) of frequency variables (in kHz) for manually detected and measured whistles. Because of the large number of features, only those features most important to the classifiers (based on the Gini Importance Index) are included in this table. See Appendix B for a description of variables.

Species	n		Max	Min	Beg	End	Mean	Standard deviation	Median	Center	1/4	1/2	Range	COFM
Short-beaked common dolphin	249	mean	14.4	9.4	12.1	11.9	11.5	1.4	11.3	11.9	11.2	11.5	5	0.8
		sd	3.2	2.0	3.7	2.9	1.9	0.9	2.1	2.1	2.4	2.5	3.4	0.8
		min	7.1	5.2	5.6	5.2	6.4	0.1	6.0	6.7	5.8	6.0	0.2	0.0
		max	37.7	16.4	37.7	21.6	18.1	5.0	18.2	23.8	19.5	20.2	27.7	7.0
Short-finned pilot whale	256	mean	8.3	4.9	6.3	6.7	6.6	0.9	6.6	6.6	6.6	6.7	3.4	0.6
		sd	3.7	2.7	3.2	3.8	2.9	0.8	3.1	2.9	3.0	3.2	2.9	0.7
		min	2.4	1.1	1.1	1.4	1.6	0.03	1.5	2.1	1.7	1.5	0.2	0.0
		max	20.1	15.7	17.2	20.0	16.3	3.9	17.4	17.9	17.1	19.3	14.8	7.8
Striped dolphin	293	mean	14.6	8.7	10.7	11.5	11.3	1.7	11.2	11.6	11.1	11.6	5.8	0.9
		sd	3.7	2.3	3.8	3.6	2.5	1.1	2.6	2.6	3.1	3.1	3.4	0.9
		min	2.7	1.2	1.2	2.7	2.1	0.1	2.2	1.9	1.6	2.2	0.2	0.0
		max	23.8	19.3	22.9	23.8	20.8	5.3	21.0	20.3	21.8	21.4	15	6.3
Atlantic spotted dolphin	706	mean	13.9	8.2	9.9	12.5	10.4	1.6	10.2	11.1	9.6	10.3	5.7	1.0
		sd	3.9	2.4	3.3	4.1	2.6	1.1	2.6	2.7	2.8	2.9	3.6	1.4
		min	3.0	2.8	2.8	2.8	3.0	0.0	3.0	3.0	3.0	3.0	0	0.0
		max	27.6	20.6	27.4	25.3	23.5	7.0	23.8	22.8	24.4	24.3	20.1	12.5
Bottlenose dolphin	1632	mean	16.8	8.5	11.9	12.0	12.5	2.4	12.4	12.7	12.8	12.7	8.3	2.0
		sd	4.4	2.9	4.6	4.9	3.2	1.4	3.5	3.0	4.1	4.1	4.4	2.2
		min	3.2	2.4	2.6	3.2	2.8	0.0	2.8	2.8	2.4	2.6	0	0.0
		max	38.4	26.4	30.0	38.4	28.4	10.6	31.9	27.9	34.4	37.9	31.5	18.2

Table 8. Descriptive statistics (mean, standard deviation, minimum, maximum) for features describing shape for manually detected and measured whistles. Duration and delta features (time between inflection points) are given in seconds. Because of the large number of features, only those features most important to the classifiers (based on the Gini Importance Index) are included in this table. See Appendix B for a description of variables.

Species	n		Duration	Max delta	Min delta	Mean delta	# inflections/ duration	# up-down	# down-flat	# flat-down
Short-beaked common dolphin	249	mean	0.8	0.4	0.2	0.3	4.7	1.0	4.7	4.7
		sd	0.5	0.5	0.2	0.3	4.9	2.8	7.6	7.6
		min	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	2.0	1.8	1.2	1.2	28.3	42.0	42.0	43.0
Short-finned pilot whale	256	mean	0.5	0.3	0.1	0.2	5.2	7.0	6.1	6.2
		sd	0.3	0.3	0.1	0.2	5.3	14.4	6.6	6.4
		min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	3.1	3.1	0.6	1.6	37.0	137.0	41.0	37.0
Striped dolphin	293	mean	0.7	0.3	0.1	0.2	4.3	0.7	4.7	4.8
		sd	0.3	0.3	0.2	0.2	6.9	0.7	7.3	7.3
		min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	2.0	1.5	1.3	1.3	111.1	4.0	35.0	35.0
Atlantic spotted dolphin	706	mean	0.3	0.2	0.1	0.1	12.0	4.4	3.2	2.9
		sd	0.3	0.3	0.1	0.2	12.9	9.6	3.9	3.9
		min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	2.4	2.2	0.7	1.1	137.9	87.0	29.0	28.0
Bottlenose dolphin	1632	mean	0.7	0.5	0.1	0.3	4.5	11.0	12.9	13.1
		sd	0.5	0.6	0.1	0.3	5.5	21.7	12.5	12.5
		min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	3.7	3.6	1.0	1.8	100.0	129.0	79.0	76.0

Table 9. Descriptive statistics (mean, standard deviation, minimum, maximum) for features describing slope (in kHz/sec) for manually detected and measured whistles. Because of the large number of features, only those features most important to the classifiers (based on the Gini Importance Index) are included in this table. See Appendix B for a description of variables.

Species	n		Mean slope	Absolute slope	Positive slope	Negative slope	Percent upswept	Percent downswept	Percent flat
Short-beaked common dolphin	249	mean	1.2	12.2	35.9	-38.1	45.5	36.7	17.8
		sd	9.6	9.2	37.3	38.9	20.4	18.5	11.2
		min	-26.9	0.6	0	-178.9	0.0	0.0	0.0
		max	57.1	59.1	281.2	0	100.0	100.0	32.5
Short-finned pilot whale	256	mean	1.9	22.4	71.2	-66.9	38.6	35.8	25.5
		sd	12.3	21.1	35.8	40.3	8.9	7.8	7.4
		min	-58.3	0.3	0	-234.4	0.0	0.0	0.0
		max	100.0	100	183.6	0	100.0	100.0	33.2
Striped dolphin	293	mean	1.5	15.1	48.9	-43.2	42.2	39.3	18.5
		sd	11.4	12.1	36.5	42.5	17.8	18.2	11.6
		min	-68.4	0.4	0	-272.6	0.0	0.0	0.0
		max	71.4	112.3	164.7	0	100.0	100.0	32.7
Atlantic spotted dolphin	706	mean	14.5	49.3	101.5	-92.7	46.1	33.3	20.6
		sd	31.6	39.7	44.5	53.8	13.3	10.7	8.1
		min	-125.0	0	0	-390.6	0.0	0.0	0.0
		max	242.8	281.2	406.2	0	100.0	100.0	33.3
Bottlenose dolphin	1632	mean	1.2	40.7	92.1	-90.5	39.3	38.4	22.3
		sd	17.9	30.1	44.2	48.5	9.4	8.4	7.1
		min	-79.6	0	0	-1031.2	8.0	0.0	0.0
		max	255.7	264.2	502.4	0	100.0	83.9	33.3

Table 10. Ten features most important in the single-stage classifier trained using manually detected and extracted whistles. See Appendix B for a description of each feature.

Feature	Gini Importance value
positive slope	104.1
# flat-down	93.7
absolute slope	86.1
# down-flat	78.1
negative slope	76.5
# inflection points/duration	76.5
duration	75.8
% flat	72.5
center frequency	70
1/4 frequency	69.2

Table 11. Ten features most important in each classifier in the two-stage classifier trained using manually detected and extracted whistles. See Appendix B for a description of each feature.

Stage 1		**Stage 2**			
Small dolphins vs. large dolphins		**Short-beaked common dolphins vs. Striped dolphins**		**Atlantic spotted dolphins vs. bottlenose dolphins vs. short-finned pilot whales**	
Feature	Gini Importance value	Feature	Gini Importance value	Feature	Gini Importance value
positive slope	64.9	duration	11.7	center frequency	34.9
absolute slope	54.8	positive slope	11.4	mean frequency	26.9
negative slope	53.7	# inflection points/duration	10.4	minimum frequency	24
% flat	32.4	beginning frequency	9.9	maximum frequency	23.8
duration	23.5	absolute slope	8.8	median frequency	22.5
mean slope	15.8	1/2 frequency	8.7	1/4 frequency	21.3
# up-down	11.2	mean frequency	8.4	# inflection points/duration	19.3
# down-flat	11.1	median frequency	8.1	1/2 frequency	19
% downswept	11.1	center frequency	7.3	# flat-down	19
# inflection points/duration	10.7	1/4 frequency	6.9	absolute slope	18.6

3.4.2 Automated Measurements

Descriptive statistics for features that were important in the classifiers based on Gini Importance values are presented in **Tables 12–14**. Based on Gini Importance values, duration was the most important feature in both the single-stage and the two-stage automated classifiers (**Tables 15 and 16**). Atlantic spotted dolphins and short-finned pilot whales produced the shortest whistles (mean = 0.3 seconds (sec), sd = 0.1 sec and 0.3 sec, sd = 0.2 sec, respectively; **Table 13**), and short-beaked common dolphins produced the longest whistles (mean = 0.6 sec, sd = 0.4 sec). For the single-stage classifier, features describing slope were also important (**Table 15**).

For the two-stage classifier, features describing frequency were most important (after duration) in stage 1 (short-finned pilot whales versus dolphins, **Table 16**). All frequency features were lowest for pilot whale contours (**Table 12**) and highest for bottlenose and short-beaked common dolphins. Atlantic spotted and striped dolphin contours were generally in the middle of the frequency range relative to the other species and their frequency features were similar to each other. In stage 2, features describing slope were the most important for separating the four smaller dolphin species (short-beaked common, striped, Atlantic spotted, and bottlenose; **Table 16**). Slopes of whistle contours were steepest for Atlantic spotted dolphin (mean absolute value = 147 kHz/s, sd = 68 kHz/s; **Table 14**) and were predominantly positive (mean slope = 9.6 kHz/s, sd = 20.1 kHz/s). Bottlenose dolphins were the only species with contours containing predominantly negative slopes (mean slope = -1.4 kHz/s, sd = 16.1 kHz/s).

Table 12. Descriptive statistics (mean, standard deviation, minimum, maximum) for frequency variables (in kHz) for automatically detected and measured whistles. Because of the large number of features, only those features most important to the classifiers (based on the Gini Importance Index) are included in this table. See Appendix B for a description of variables.

Species	n		Max	Min	Beg	End	Mean	Standard deviation	Median	Center	1/4	1/2	Range	COFM
Short-beaked common dolphin	475	mean	16.5	11.6	13.6	14.6	13.9	1.4	13.8	14.0	13.6	13.7	4.9	1.0
		sd	6.1	5.3	6.0	6.0	5.4	0.9	5.5	5.5	5.6	5.6	3.2	1.4
		min	2.6	0.9	0.9	1.1	1.5	0.1	1.3	1.8	0.9	1.5	0.2	0.0
		max	44.8	39.9	44.1	43.7	42.9	7.9	43.7	41.5	43.7	44.2	28.5	17.3
Short-finned pilot whale	749	mean	10.9	7.6	9.1	9.2	9.2	0.9	9.2	9.2	9.3	9.2	3.3	0.4
		sd	6.1	5.1	5.8	5.8	5.5	0.8	5.6	5.5	5.7	5.6	2.8	0.5
		min	1.3	0.9	0.9	0.9	0.9	0.03	0.9	1.1	0.9	0.9	0.2	0.0
		max	39.9	35.8	38.2	39.6	38.5	4.6	38.3	37.9	37.7	38.1	19.7	4.1
Striped dolphin	648	mean	13.8	9.0	11.2	11.9	11.1	1.4	11.0	11.4	10.9	10.9	4.8	0.9
		sd	4.3	3.5	4.2	4.4	3.7	0.8	3.7	3.6	3.9	3.8	2.9	1.4
		min	2.1	0.9	1.9	1.7	1.9	0.03	1.9	1.9	0.9	1.9	0.2	0.0
		max	42.2	32.4	42.2	41.6	36.3	4.8	35.3	37.3	39.0	35.2	17.1	19.7
Atlantic spotted dolphin	988	mean	14.5	8.2	10.3	12.8	10.9	1.7	10.7	11.3	9.9	10.6	6.3	1.2
		sd	4.8	3.4	4.4	5.0	3.8	1.0	3.8	3.8	3.8	3.9	3.4	1.2
		min	2.2	0.9	0.9	0.9	1.5	0.1	1.7	1.6	1.5	1.3	0.6	0.0
		max	44.6	42.2	43.1	44.4	42.7	8.6	42.7	42.7	42.7	42.7	22.7	12.0
Bottlenose dolphin	1719	mean	17.4	9.3	13.7	13.0	13.1	2.3	13.0	13.3	13.4	13.0	8.1	1.5
		sd	4.8	3.5	5.4	5.3	3.8	1.2	3.9	3.7	4.4	4.3	3.9	2.3
		min	6.0	3.4	3.7	4.3	5.1	0.2	5.2	5.2	3.9	4.7	0.7	0.1
		max	44.8	33.0	44.8	39.6	36.2	9.6	35.6	36.7	39.2	44.4	32.2	55.5

Table 13. Descriptive statistics (mean, standard deviation, minimum, maximum) for features describing shape for automatically detected and measured whistles. Duration and delta features (time between inflection points) are given in seconds. Because of the large number of features, only those features most important to the classifiers (based on the Gini Importance Index) are included in this table. See Appendix B for a description of variables.

Species	n		Duration	Max delta	Min delta	Mean delta	# Inflections/ duration	# up-down	# down-flat	# flat-down
Short-beaked common dolphin	475	mean	0.6	0.5	0.1	0.3	24.1	3.9	5.8	5.9
		sd	0.4	0.4	0.2	0.3	25.1	3.3	7.4	7.4
		min	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	3.1	2.3	1.2	1.3	114.1	19.0	44.0	44.0
Short-finned pilot whale	749	mean	0.3	0.1	0.03	0.1	27.2	2.2	4.4	4.8
		sd	0.2	0.1	0.05	0.1	24.8	2.3	5.7	5.7
		min	0.05	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	1.2	1.0	0.4	0.7	145.2	14.0	45.0	46.0
Striped dolphin	648	mean	0.5	0.4	0.1	0.2	19.3	3.0	4.8	4.8
		sd	0.3	0.3	0.2	0.2	17.7	3.2	6.9	7.0
		min	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	2.7	2.6	1.1	1.5	88.7	33.0	57.0	58.0
Atlantic spotted dolphin	988	mean	0.3	0.2	0.04	0.1	51.1	5.9	3.1	3.1
		sd	0.1	0.2	0.1	0.1	28.0	4.4	3.1	3.0
		min	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	1.4	1.3	0.4	0.8	166.7	32.0	23.0	20.0
Bottlenose dolphin	1719	mean	0.5	0.4	0.1	0.2	18.3	3.4	13.7	14.0
		sd	0.3	0.3	0.1	0.2	15.5	3.5	12.1	12.0
		min	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		max	4.3	4.3	0.9	1.5	105.5	35.0	109.0	105.0

Table 14. Descriptive statistics (mean, standard deviation, minimum, maximum) for features describing slope (in kHz/sec) for automatically detected and measured whistles. Because of the large number of features, only those features most important to the classifiers (based on the Gini Importance Index) are included in this table. See Appendix B for a description of variables.

Species	n		Mean slope	Absolute slope	Positive slope	Negative slope	Percent upswept	Percent downswept	Percent flat
Short-beaked common dolphin	475	mean	2.6	39.4	95.8	-98.5	42.2	37.4	20.3
		sd	13.1	50.8	133.3	123.1	15.0	14.0	8.2
		min	-46.2	1.3	0	-1229.6	5.3	3.2	0.0
		max	5.5	385.1	1143.1	0	96.8	94.7	32.1
Short-finned pilot whale	749	mean	1.1	35.1	98.5	-93.1	37.7	39.0	23.4
		sd	21.3	38.2	82.7	74.7	14.5	15.5	7.8
		min	-15.9	0.6	0	-690.5	0.0	0.0	0.0
		max	20.6	385.2	928.8	0	100.0	100.0	33.1
Striped dolphin	648	mean	1.9	31.8	86.5	-93.3	41.6	39.4	19.0
		sd	12.2	37.9	93.4	116	17.0	17.4	10.0
		min	-65.8	0.8	4.3	-1343.7	0.0	0.0	0.0
		max	62.1	339.6	633.7	0	100.0	100.0	32.7
Atlantic spotted dolphin	988	mean	9.6	86.6	146.7	-155.6	46.6	35.5	17.9
		sd	20.1	56.8	67.7	82.6	11.3	10.7	6.5
		min	-76.9	6.8	62.5	-864.6	8.2	10.8	0.0
		max	84.8	399.9	724.3	0	83.1	83.6	32.0
Bottlenose dolphin	1719	mean	-1.4	40.5	123.4	-113.9	36.8	38.8	24.3
		sd	16.1	37.4	121.9	87.9	9.7	9.2	4.7
		min	-64.8	3.7	0	-1348.8	5.4	5.7	5.0
		max	64.0	541.2	2495.2	0	88.7	89.3	32.5

Table 15. Ten features most important in the single-stage classifier trained using automatically detected and extracted whistles. See Appendix B for a description of each feature.

Feature	Gini Importance value
duration	137.4
absolute slope	80.1
negative slope	70.8
% flat	70.5
positive slope	66.7
# inflection points/duration	61
center frequency	60.8
maximum frequency	57.6
mean frequency	56.3
maximum delta	52.6

Table 16. Ten features most important in each classifier in the two-stage classifier trained using automatically detected and extracted whistles. See Appendix B for a description of each feature.

Stage 1		**Stage 2**	
Short-finned pilot whales vs. dolphins		**Short-beaked common dolphins vs. striped dolphins vs. Atlantic spotted dolphins vs. bottlenose dolphins**	
Feature	Gini Importance value	Feature	Gini Importance value
duration	63	duration	93.9
maximum frequency	45.8	% flat	64.1
coefficient of frequency modulation	45.5	negative slope	63.4
center frequency	40.7	absolute slope	61.7
mean frequency	32.8	positive slope	56.9
standard deviation of the frequency	29	# inflection points/duration	56.9
maximum delta	28.1	# down-flat	52.9
frequency range	26.9	# flat-down	50.9
mean delta	21.6	max delta	35.7
minimum frequency	21.4	1/4 frequency	35.5

3.5 PAMGuard ROCCA

Both the manual classifier and the automated classifier have been incorporated into the ROCCA module in PAMGuard. The updates to ROCCA will be available for users at www.pamguard.org in the next PAMGuard update. Until that time, users can obtain the update directly from Bio-Waves, Inc. (www.bio-waves.com). With these new ROCCA updates, users can now choose to run the automated Atlantic classifier, the manual Atlantic classifier, or the manual tropical Pacific classifier when starting ROCCA. Whistles are detected, extracted and measured using the methods described in **Sections 2.2.1–2.2.3**. A User's Manual describing the set-up and use of both the manual and the automated classifiers is available, and detailed help files are contained within the software (Oswald and Oswald 2013).

4. DISCUSSION

Two new classifiers were trained for whistles produced by delphinid species in the northwestern Atlantic Ocean. One classifier was trained and tested using whistles detected and extracted manually and the second was trained and tested using whistles detected and extracted automatically (using PAMGuards WMD module). These are the only classifiers that we are aware of that currently are available to the general public and scientific community specifically for classifying whistles of delphinid species occurring in the western Atlantic ocean. These classifiers will be valuable tools in the analysis of acoustic data, both in real time and for post-processing. Correct classification scores were excellent for both classifiers, with 78 percent and 80 percent of whistles, and 86 percent and 91 percent of schools being correctly classified using the manual classifier and the automated classifier, respectively. These results compare very favorably with multi-species classifiers trained for other species-groups and locations. For example, the original ROCCA classifier is a single-stage random-forest classifier that was trained using whistles recorded from eight delphinid species in the tropical Pacific Ocean (Oswald et al. 2013). Overall correct classification scores for this classifier were only 43 percent for whistles and 60 percent for encounters (Oswald et al. 2013). Other researchers have used multivariate discriminant function analysis to classify whistles produced by five species in the western North Atlantic, with an overall correct classification score of 70 percent (Steiner, 1981). Roch et al. (2007) used cepstral feature vectors and Gaussian mixture models to classify whistles and clicks produced by four species recorded in the Southern California Bight and the Gulf of California. Correct classification scores in this study ranged from 67 percent to 75 percent, depending on how their training and test data were partitioned.

We believe the main reason for the high correct classification scores that were obtained in this project were due to the use of a two-stage classifier. For both the manual and automated data, the two-stage classifier was able to classify whistles and encounters more accurately than the single-stage classifier (**Table 4**). This is because different features were important for separating certain species or species groups. Using two classifiers instead of only one, allows different feature-sets to be exploited more effectively. For example, in the single-stage manual classifier, short-beaked common dolphins were most often misclassified as striped dolphins (and vice versa), and bottlenose dolphins were most often misclassified as Atlantic spotted dolphins (and vice versa; **Table 2**). It is likely that this is because of the similar slope and shape features for whistles from those two species-groups. When species with similar slope and shape whistle features were combined into two groups and the whistles were first classified into either the small dolphin or large dolphin class, correct classification scores increased dramatically. Stage two contained two different classifiers that used different sets of features in each. The most important features in the small dolphin classifier (striped versus short-beaked common dolphins) were a mixture of shape, slope, and frequency features (**Table 11**). Although the differences in those features did not seem large for common dolphins and striped dolphins, they were sufficient to create a classifier that could distinguish between those two species with almost 100 percent accuracy. A different set of features was most important in the large dolphin (bottlenose, Atlantic spotted dolphin and short-finned pilot whale) classifier. The ten most important features in this classifier primarily consisted of frequency variables (**Table 11**). The whistles produced by pilot whales were much lower in frequency than those produced by Atlantic spotted and bottlenose dolphins and as a result, no pilot whale whistles were misclassified as Atlantic spotted or bottlenose dolphins, and very few Atlantic spotted or bottlenose whistles were misclassified as pilot whales.

The two-stage approach also worked well for the automated data. In the single-stage automated classifier, duration and contour slope features were most important for separating species (**Table 15**). In

this classifier, classification errors for all species were distributed relatively evenly among all species (**Table 5**). Major sources of these errors are likely due to the similar durations of whistles for pilot whales and Atlantic spotted dolphins, and similar slope characteristics for Atlantic spotted and bottlenose dolphins and striped and short-beaked common dolphins (**Tables 13 and 14**). In addition, frequency characteristics were similar for short-beaked common and bottlenose dolphins, as well as for Atlantic spotted and striped dolphins (**Table 14**).

The two-stage classifier had higher correct classification scores because in stage one, pilot whales were effectively separated from the other species based mainly on duration and frequency characteristics (**Table 16**). Although whistles from the four smaller species had frequency characteristics that were similar, pilot whale whistles were consistently lower in frequency than the whistles of all other species. Pilot whale whistles also had shorter durations than whistles from all other species with the exception of Atlantic spotted dolphin whistles (**Tables 12 and 13**). Separating pilot whales from the other species based on frequency in stage one removed some of the sources of error. These sources of error included slope characteristics (similar between pilot whales and striped dolphins and pilot whales and short-beaked common dolphins) and duration (similar between pilot whales and Atlantic spotted dolphins) (**Tables 13 and 14**). In stage two, frequency variables were not as important as they were in stage one because these features were so similar among the small dolphins. Instead, duration and slope variables were used to separate the small dolphin species (**Table 16**). Because pilot whales were not included in stage 2, Atlantic spotted dolphins could be separated based on duration. Contours of the remaining three species had similar durations, and so slope variables were important for their identification. Bottlenose dolphins had the most distinctive slope characteristics of the remaining three species (**Table 14**). Most of the classification errors in the two-stage classifier were common, striped, and bottlenose dolphins being misclassified as pilot whales (**Table 6**), which suggests that frequency variables should be refined further to allow greater separation between pilot whales and smaller dolphins.

The accurate performance of the automated classifier in relation to the manual classifier was unexpected. For each encounter, WMD settings were optimized for whistle detection and contour extraction; however, as with any automated system, inaccurate extractions and false detections were unavoidable. In addition, the WMD sometimes fragmented whistles, resulting in a single whistle incorrectly labeled as multiple individual whistles (**Figure 1**). Because the goal of this work was to create a fully automated classifier, it was decided that inaccurately extracted whistles, false detections, or fragmented whistles would not be removed. Only by using the entire output of the WMD to train a classifier is it possible to create a classification system that is fully automated from start to finish. An advantage to using the entire output of the WMD is that it provided a larger sample size for training and testing the classifier. The larger sample size likely allowed the classifier to capture a greater amount of the variability in the dataset and consequently resulted in higher correct classification scores. One disadvantage to the fully automated classifier is that its performance is affected, to a greater extent, by noise and other sounds in the recordings than the performance of the manual classifier is. However, this is true of most automated acoustic detection and classification systems. The manual classifier requires the user to detect whistles so there are no (or very few) false detections caused by noise. In addition, the user has the option to adjust the extracted contour to make it more accurate. Low signal-to-noise ratios in the recordings can cause the automated detector to produce false detections or inaccurate whistle extractions. In addition, noise can mask portions of whistles, causing them to be fragmented. Adjusting the noise and thresholding settings in the WMD (**Figure 3**) can reduce the effects of noise in the recordings, but it is not always possible to remove it completely. The amount and type of noise in recordings will vary by location, recording equipment, and time. The recordings used to train this classifier were made using towed hydrophone arrays, which have very different self and ambient noise

characteristics than recordings made using other devices such as seafloor-mounted hydrophones and acoustic recorders. It is important to test the classifier on each new dataset to evaluate whether the automated classifier is appropriate for that dataset. In some cases, interference from noise may be too great, and it may be necessary to use the manual classifier.

While both the manual and automated classifiers produced good results when evaluated using the test data, it is important to treat these results with caution, especially when using the classifiers to analyze novel data. Differences in recording platforms and noise environments may affect the performance of automated detectors as well as whistle measurements made using manual methods. For example, ROCCA's classifiers were trained with data collected using an array of hydrophones towed near the surface of the water. Whistles recorded at depth (ex. using seafloor mounted autonomous recorders) may have different characteristics due to propagation effects and attenuation. In addition, animals may produce whistles with different characteristics in response to the presence of a research vessel towing a hydrophone array. If this is the case, then the classifiers may perform differently when used on recordings made using less obtrusive platforms such as autonomous recorders or sonobuoys. It is not possible to assess the performance of classifiers without ground-truthing them using visually validated acoustic recordings. If possible, it is important to ground-truth the classifiers using visually validated acoustic data before each new analysis. This will provide error rates specific to the recordings in question and will allow a more accurate assessment of the results.

Although ROCCA does not classify with 100 percent accuracy, we believe that the correct classification scores of ~ 85-90 percent that we achieved are high enough to provide reliable information about the identity of whistles produced by many species of odontocete that are monitored with passive acoustics. These results can be used to provide new and important information on occurrence, distribution, and even the behaviors of dolphins and pilot whales that would not be possible (or would be extremely difficult or expensive) to obtain otherwise. Nevertheless, additional effort is needed to continue to improve classification success. Some of the features (ex., duration, beginning and ending frequencies, maximum frequency) currently included in the classifier could be unreliable, depending on the SNR of the recording. Alternative features that may be more robust to noise may increase the accuracy and reliability of classifiers. Additional variables such as cepstral features (Roch et al. 2007) or statistical measures of signals such as those used by Fristrup and Watkins (1993) in the program ACOUSTAT should be explored.

Finally, it is important to note that the classifiers presented here include only five whistling species that occur in the northwest Atlantic Ocean. At least seven additional species that occur in this region are known to produce whistles (pantropical spotted dolphin, *Stenella attenuata;* rough-toothed dolphin, *Steno bredanensis*; Clymene dolphin, *Stenella clymene*; long-finned pilot whale, *Globicephala melas;* false killer whale; Risso's dolphin, *Grampus griseus*; Atlantic white-sided dolphin, *Lagenorhynchus acutus*) (Palka 2012, Waring et al. 2012). These species are not yet included in ROCCA because recordings of single-species schools either were unavailable or there were not enough acoustic encounters to reliably train the classifier. This lack of data for some species is due to the fact that these species were either rarely encountered during the NEFSC, SEFSC and Duke surveys or, if they were encountered, they produced few or no whistles near the survey vessel. Unfortunately, If recordings containing any of these species are analyzed using ROCCA, they will be misclassified as one of the five species included in the classifier. This type of error will result in an incorrect picture of the occurrence of species in recordings analyzed using ROCCA. Given the fact that whistles produced by these species were rare during the surveys, it is expected that the magnitude of this error will be low for at least some species. Some species may change their acoustic behavior (i.e. stop whistling) in response to the

presence of ships. A lack of recordings does not necessarily reflect low abundance for those species. Therefore, we believe that it is important to add all missing species to the classifier. The ability to identify the full complement of species in the northwest Atlantic would allow for a more complete understanding of the occurrence and distribution of whistling species in the northwest Atlantic Ocean.

5. CONCLUSIONS

This project resulted in the successful development of delphinid whistle classifiers that are user-friendly, accurate and freely available for download, making them useful tools for the analysis of acoustic data collected along the Atlantic coast of the United States. Classifiers were trained using whistles detected both manually and automatically, providing alternative methods for analyzing data with a wide variety of signal-to-noise characteristics. The ability to identify delphinid species based on whistles will allow a deeper understanding of the distribution, occurrence, and vocal behaviors of these important, federally protected, living marine resources.

This page intentionally left blank.

6. RECOMMENDATIONS FOR FUTURE RESEARCH

Although results from the manual and automated classifier processes as developed and applied here were good, there are several important tasks that should be undertaken to make ROCCA more complete, accurate, precise, and dependable. Perhaps the highest priority is testing the Atlantic classifiers using visually validated recordings containing different noise levels and characteristics. This is particularly important for the automated classifier, and for recordings made from different platforms, such as seafloor recorders. Testing classifiers on recordings made using seafloor recorders is challenging because seafloor recordings rarely have associated visual observations of marine mammals, which are often necessary to validate species acoustic identifications. In cases when acoustic identification is uncertain, surface observations concurrent with seafloor-mounted acoustic recordings and localizations should be conducted in order to confirm species identity.

When possible, classifiers should be tested in real time to assess their capabilities for field-work applications. The classifiers in this study were trained using data analyzed during post-processing, which allowed more time for ensuring accuracy and testing of WMD parameters. In a real time setting, there are often many activities occurring simultaneously and conditions are constantly changing. In the field, detection-classification software needs to be robust, easy to use, and time-efficient in order to be truly useful. ROCCA should be tested in real time scenarios in order to evaluate its robustness, accuracy, and user-friendliness, as well as the frequency with which WMD settings need to be adjusted. Based on such at-sea testing and evaluation, improvements can be made to make ROCCA a more effective field tool.

ROCCA's Atlantic classifier currently contains five delphinid species recorded off the East Coast of the United States. Recordings are currently available for five additional species (i.e., not currently included in ROCCA): pantropical spotted dolphin, rough-toothed dolphin, Clymene dolphin, false killer whale and Risso's dolphin. Unfortunately, sample sizes are not yet large enough for training and testing detectors and classifiers for these species. In order to be able to correctly identify all of the whistles on recordings made in the western North Atlantic Ocean, it is necessary to add these species, and any others that are missing, to the classifier. During the summer of 2013, the SEFSC conducted a concurrent visual and acoustic survey off the East Coast of the United States. Because this survey occurred as work on the Atlantic classifier was wrapping up, recordings made during the survey were not included in our dataset. It is recommended that the data recorded during SEFSC's 2013 survey be analyzed and existing data from other researchers be obtained to increase sample sizes and allow additional species to be included in the Atlantic classifiers.

While ROCCA's classifiers have been trained using tonal whistles produced by delphinids, these are not the only marine mammal species that produce tonal signals. Many baleen whale species, including right (*Eubalaena glacialis*), fin (*Balaenoptera physalus*), blue (*Balaenoptera musculus),* and humpback whales (*Megaptera novaeangliae*), also produce tonal signals (Payne and McVay 1971, Mattila et al. 1987, Thompson et al. 1992, Matthews et al. 2001, Mellinger and Clark 2003). These signals can easily be detected and extracted using either the WMD or ROCCA's manual methods. ROCCA's random-forest classifiers have proven to work well with variable high-frequency tonal sounds and are likely to also work well with low-frequency tonal sounds such as those produced by baleen whales. Fristrup and Watkins (1993) showed that the same feature vectors can be used to classify delphinids, baleen whales and pinnipeds. Their detection/classification system, AcouSTAT correctly classified 85 percent of sounds produced by 53 marine mammal species. However, this classifier has not been optimized or tested specifically on data collected in the northwest Atlantic Ocean. A collaboration with the developers of

AcouSTAT and/or other researchers working on the detection and classification of sounds produced by baleen whales is recommended to develop a multi-species baleen whale classifier specifically for the northwest Atlantic Ocean and other areas of high naval activity.

Although the western Atlantic contains areas that are important areas for PAM (especially related to United States (U.S.) Navy activities), there are many other locations for which classifiers should be created. Previous studies have shown geographic variation in characteristics of whistles (e.g., Bazua-Duran and Au 2004, Morisaka et al. 2005, Ansmann et al. 2007, Baron et al. 2008, May-Collado and Wartzok 2008), and so it is important to develop classifiers for the specific locations in which they will be used. Areas such as the Gulf of Mexico, the Gulf of Alaska, the temperate Pacific, the waters surrounding the Mariana Islands, and the Caribbean Sea are all being monitored using passive acoustics and are important areas for naval exercises. A standardized method for detection and classification of delphinids would allow results to be compared among locations. It is recommended that detectors and classifiers be developed using data specific to each of the locations mentioned above. It is also recommended that these detection/classification systems be based on common software systems so that operators can easily work in multiple areas without having to learn new methods for different study areas.

Methods for efficiently analyzing acoustic data and examining questions related to occurrence, distribution, and behavior of animals are greatly needed. For example, it is crucial to be able to accurately and efficiently identify species in order to evaluate potential responses to naval and other human activities. Lumping species into a 'delphinid' group instead of examining individual species may result in the reactions of one species being masked by the opposite reactions of another species. A large amount of archival towed-array and seafloor-mounted recorder data exist for the northwestern Atlantic Ocean. It is recommended that a detector/classifier developed specifically for the northwestern Atlantic Ocean be used to process these data and examine questions related to the occurrence of animals and responses to naval activities such as sonar exercises, explosions and ship traffic.

7. ACKNOWLEDGEMENTS

We would like to thank Lynne Hodge and Andrew Read (Duke University), Melissa Soldevilla and Lance Garrison (SEFSC), and Danielle Cholewiak and Sofie Van Parijs (NEFSC) for generously providing data for this project. We are grateful to Susan Baron, Kait Frasier, John Hildebrand, Tony Martinez, and Jesse Wicker for their efforts in data collection. Thank you to Shannon Coates, Talia Dominello, Kerry Dunleavy, and Cory Hom-Weaver (Bio-Waves, Inc.) for many hours spent measuring whistles and aiding with data analysis and to Michael Oswald for computer programming. Tom Norris, Michael Oswald, Elizabeth Ferguson, and Robyn Walker provided helpful comments on this manuscript. We thank the Bureau of Ocean Energy and Management for providing funding for SEFSC surveys and to Naval Facilities Engineering Command, Atlantic (NAVFAC LANT) for providing funding for this analysis. We are grateful for the logistic support and advice from Joel Bell and Anurag Kumar (NAVFAC LANT) and from Dan Engelhaupt and Carter Esch (HDR).

This page intentionally left blank.

8. LITERATURE CITED

Ansmann, I.C., J.C. Goold, P.G.H. Evans, M. Simmonds and G.K. Simon. (2007). Variation in the whistle characteristics of short-beaked common dolphins, *Delphinus delphis*, at two locations around the British Isles. *J. Mar. Biol. Assoc. U.K.* 87, 19-26.

Barkley, Y., J.N. Oswald, J.V. Carretta, S. Rankin, A. Rudd, and M.O. Lammers. (2011). Comparison of real-time and post-cruise acoustic species identification of dolphin whistles using ROCCA (Real-time Odontocete Call Classification Algorithm). NOAA Technical Memorandum NOAA-TM-NMFS-SWFSC-473. National Marine Fisheries Service, La Jolla, CA. 29 pp.

Baron, S.C., A. Martinez, L.P. Garrison, and E.O. Keith. (2008). Differences in acoustic signals from delphinids in the western North Atlantic and northern Gulf of Mexico. *Mar. Mamm. Sci.* 24, 42-56.

Bazua-Duran, M.C., and W.W.L. Au. (2004). Geographic variations in the whistles of spinner dolphins (*Stenella longirostris*) of the main Hawaiian Islands. *J. Acoust. Soc. Am*. 116, 3757-3769.

Bioacoustics Research Program. (2008). Raven Pro: Interactive Sound Analysis Software (Version 1.3) [Computer software]. The Cornell Lab of Ornithology, Ithaca, NY. Available from http://www.birds.cornell.edu/raven.

Breiman, L. (2001). Random forests. *Machine Learn.* 45, 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. (1984). *Classification and Regression Trees.* Wadsworth International Group, Belmont, CA.

Brown, J.C., and P. Smaragdis. (2009). Hidden Markov and Guassian mixture models for automatic call classification. *J. Acoust. Soc. Am*. 125, EL221-EL224.

Gillespie, D., J. Gordon, R. McHugh, D. McLaren, D.K. Mellinger, P. Redmond, A. Thode, P. Trinder, and D. Xiao. (2008). PAMGUARD: Semiautomated, open-source software for real-time acoustic detection and localization of cetaceans. *Proceed. Instit. Acoust.* 30, Part 5. 9 pp.

Hodge, L.E.W. (2011). *Monitoring Marine Mammals in Onslow Bay, North Carolina, Using Passive Acoustics.* PhD Dissertaion, Department of the Environment, Duke University, 197pp.

Johnson, M.P., and P.L. Tyack. 2003. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. *IEEE J. Oc. Eng.* 28, 3-12.

Matthews, J.N., S. Brown, D. Gillespie, M. Johnson, R. McLanaghan, A. Moscrop, D. Nowacek, R. Leaper, T. Lewis, and P. Tyack. 2001. Vocalisation rates of the North Atlantic right whale (*Eubalaena glacialis*). *J. Cet. Res. Manage.* 3, 271-282.

Matthews, J.N., L.E. Rendell, J.C.D. Gordon, and D.W. MacDonald. (1999). A review of frequency and time parameters of cetacean tonal calls. *Bioacoustics* 10, 47-71.

Mattila, D.K., L.N. Guinee and C.A. Mayo. 1987. Humpback whale songs on a North Atlantic feeding ground. *J. Mamm.* 68, 880-883.

May-Collado, L.J., and D. Wartzok. (2008). A comparison of bottlenose dolphin whistles in the Atlantic Ocean: factors promoting whistle variation. *J. Mamm.* 89, 1229-1240.

Mellinger, D.K. 2001. Ishmael 1.0 User's Guide. NOAA Technical Memorandum OAR PMEL-

120. Available from NOAA/PMEL/OERD, 2115 SE OSU Drive, Newport, OR, 97365-5258, http://www.pmel.noaa.gov/pubs/PDF/mell2434/mell2434.pdf.

Mellinger, D.K., and J. Barlow. (2003). Future directions for marine mammal acoustic surveys: stock assessment and habitat use. Workshop held in La Jolla, CA, 20-22 November 2002. NOAA/PMEL Contribution No. 2557. NOAA/PMEL, Seattle, WA. 45 pp.

Mellinger, D.K. and C.W. Clark. 2003. Blue whale (*Balaenoptera musculus*) sounds from the North Atlantic. *J. Acoust. Soc. Am.* 114, 1108-1119.

Morisaka, T., M. Shinohara, F. Nakahara, and T. Akamatsu. (2005). Geographic variation in the whistles among three Indo-Pacific bottlenose dolphin (*Tursiops aduncus*) populations in Japan. *Fish. Sci.* 71, 568-576.

Oh, J., M. Laubach, and A. Luczak. 2003. Estimating neuronal variable importance with random forest. Proceedings of the IEEE Bioengineering Conference, pp 33-34.

Oswald, J.N., and M. Oswald. 2013. ROCCA (Real-time Odontocete Call Classification Algorithm) User's Manual. Submitted to HDR Environmental, Operations and Construction, Inc. Norfolk, VA under Contract No. CON005-4394-009, Subproject 164744, Task Order 003, Agreement # 105067. Prepared by Bio-Waves, Inc., Encinitas, CA.

Oswald, J.N., J. Barlow, and T.F. Norris. (2003). Acoustic identification of nine delphinid species in the eastern tropical Pacific Ocean. *Mar. Mamm. Sci.* 19, 20-37.

Oswald, J.N., S. Rankin, J. Barlow, and M.O. Lammers. (2007). A tool for real-time acoustic species identification of delphinid whistles. *J. Acoust. Soc. Am*. 122, 587-595.

Oswald, J.N., J.V. Carretta, M. Oswald, S. Rankin and W.W.L. Au. (2011). Seeing the species through the trees: Using random forest classification trees to identify species-specific whistle types*. J. Acoust. Soc. Am. 129*, 2639.

Oswald, J.N., S. Rankin, J. Barlow, M. Oswald and M.O. Lammers. (2013). Real-time Call Classification Algorithm (ROCCA): software for species identification of delphinid whistles. *In*: O. Adam, and F. Samaran (eds). Detection, Classification and Localization of Marine Mammals using Passive Acoustics, 2003-2013: 10 years of International Research. DIRAC NGO, Paris, France, pp. 245-266.

Palka, D. 2012. Cetacean abundance estimates in U.S. Northwestern Atlantic Ocean waters. Northeast Fisheries Science Center Reference Document 12-29. 37pp.

Payne, R.S., and S. McVay. 1971. Songs of humpback whales. *Science* 173, 585-597.

Rankin, S., J.N. Oswald, and J. Barlow. 2008. Acoustic behavior of dolphins in the Pacific Ocean: implications for using passive acoustic methods for population studies. *Can. Acoust.* 36, 88-92.

Rendell, L.E., J.N. Matthews, A. Gill, J.C.D. Gordon, and D.W. MacDonald. (1999). Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation. *J. Zool.* 249, 403-410.

Roch, M.A., M.S. Soldevilla, J.C. Burtenshaw, E.E. Henderson, and J.A. Hildebrand. (2007). Gaussian mixture model classification of odontocetes in the southern California Bight and the Gulf of California. *J. Acoust. Soc. Am. 121*, 1737-1748.

Steiner, W.W. (1981). Species-specific differences in pure tonal whistle vocalization of five western North Atlantic dolphin species. *Behav. Ecol. Sociobiol.* 9, 241-246.

Thompson, P.O., L.T. Findley and O. Vidal. 1993. 20-Hz pulses and other vocalizations of fin whales, *Balaenoptera physalus*, in the Gulf of California, Mexico.

Wang, D., B. Würsig and W. Evans. (1995). Comparisons of whistles among seven odontocete species. *In*: R.A. Kastelein, J.A. Thomas, and P.E. Nachtigall (eds). Sensory Systems of Aquatic Mammals. De Spil Publishers, Woerden, Netherlands, pp. 299-323.

Waring, G.T., E. Josephson, K. Maze-Foley, and P.E. Rosel. 2012. U.S. Atlantic and Gulf of Mexico marine mammal stock assessments – 2011. NOAA Technical Memorandum NMFS-NE-221. 319 pp.

APPENDIX A:

CHARACTERISTICS OF THE HYDROPHONE ARRAYS AND RECORDING SYSTEMS USED BY THE SOUTHEAST FISHERIES SCIENCE CENTER AND THE NORTHEAST FISHERIES SCIENCE CENTER OF THE NATIONAL MARINE FISHERIES SERVICE, AND DUKE UNIVERSITY

Appendix A:
Characteristics of the hydrophone arrays and recording systems used by the Southeast Fisheries Science Center and the Northeast Fisheries Science Center of the National Marine Fisheries Service, and Duke University.

Source	Cruise	Year	Number of Hydrophones	Array Frequency Response	Recording Medium	Sampling Rate
SEFSC	AMAPPS	2011	6	+/- 3 dB from 1 to 180 kHz except for a 3 dB peak at 150 kHz	Computer hard drive	192 kHz
SEFSC	GU-02-01	2002	5	+/- 3dB from 2 to 15 kHz, except for a 4dB peak at 35 kHz	Digital Audio Tape	48 kHz
SEFSC	GU-02-01	2002	2	+/- 1 dB from 1 Hz to 15 kHz and +/- 2 dB from 15 to 25 kHz.	Digital Audio Tape	48 kHz
SEFSC	GU-06-03	2006	2	+/- 1 dB from 1 Hz to 15 kHz and +/- 2 dB from 15 to 25 kHz.	Digital Audio Tape	4 kHz
NEFSC	AMAPPS	2011	2-3	+/- 1.5 dB up to 15 kHz and +/- 2 dB up to 25 kHz	Computer hard drive	192 kHz
Duke	One-day surveys	2007-2010	4	+/- 3dB from 2 kHz to 100 kHz	Computer hard drive	192 kHz
Duke	DTAG	2008, 2011, 2012	1	+/- 3dB from 400 Hz to 96 kHz	DTAG	96 – 192 kHz

Key: AMAPPS = Atlantic Marine Assessment Program for Protected Species; dB = decibel(s); Duke = Duke University; GU = NOAA Ship *Gordon Gunter*; Hz = Hertz; kHz = kilohertz; NEFSC = Northeast Fisheries Science Center; SEFSC = Southeast Fisheries Science Center

This page intentionally left blank.

APPENDIX B:

VARIABLES MEASURED BY ROCCA

Appendix B:
Variables measured by ROCCA.

Variable	Explanation
Begsweep	slope of the beginning sweep (1 = positive, -1 = negative, 0 = zero)
Begup	binary variable: 1 = beginning slope is positive, 0 = beginning slope is negative
Begdwn	binary variable: 1 = beginning slope is negative, 0 = beginning slope is positive
Endsweep	slope of the end sweep (1 = positive, -1 = negative, = 0 zero)
Endup	binary variable: 1 = ending slope is positive, 0 = ending slope is negative
Enddwn	binary variable: 1 = ending slope is negative, 0 = ending slope is positive
Beg	beginning frequency (Hertz [Hz])
End	ending frequency (Hz)
Min	minimum frequency (Hz)
Dur	duration (seconds)
Range	maximum frequency - minimum frequency (Hz)
Max	maximum frequency (Hz)
mean freq	mean frequency (Hz)
median freq	median frequency (Hz)
std freq	standard deviation of the frequency (Hz)
Spread	difference between the 75th and the 25th percentiles of the frequency
quart freq	frequency at one-quarter of the duration (Hz)
half freq	frequency at one-half of the duration (Hz)
Threequart	frequency at three-quarters of the duration (Hz)
Centerfreq	(minimum frequency + (maximum frequency-minimum frequency))/2
rel bw	relative bandwidth: (maximum frequency - minimum frequency)/center frequency
Maxmin	maximum frequency/minimum frequency
Begend	beginning frequency/end frequency
Cofm	coefficient of frequency modulation (COFM): take 20 frequency measurements equally spaced in time, then subtract each frequency value from the one before it. COFM is the sum of the absolute values of these differences, all divided by 10,000
tot step	number of steps (10 percent or greater increase or decrease in frequency over two contour points)
tot inflect	number of inflection points (changes from positive to negative or negative to positive slope)
max delta	maximum time between inflection points
min delta	minimum time between inflection points
maxmin delta	maximum delta/minimum delta
mean delta	mean time between inflection points
std delta	standard deviation of the time between inflection points
median delta	median of the time between inflection points

Variable	Explanation
mean slope	overall mean slope
mean pos slope	mean positive slope
mean neg slope	mean negative slope
mean absslope	mean absolute value of the slope
Posneg	mean positive slope/mean negative slope
perc up	percent of the whistle that has a positive slope
perc dwn	percent of the whistle that has a negative slope
perc flt	percent of the whistle that has zero slope
up dwn	number of inflection points that go from positive slope to negative slope
dwn up	number of inflection points that go from negative slope to positive slope
up flt	number of times the slope changes from positive to zero
dwn flt	number of times the slope changes from negative to zero
flt dwn	number of times the slope changes from zero to negative
flt up	number of times the slope changes from zero to positive
step up	number of steps that have increasing frequency
step dwn	number of steps that have decreasing frequency
step.dur	number of steps/duration
inflect.dur	number of inflection points/duration