Automated Detection and Identification of Blue and Fin Whale Foraging Calls by Combining Pattern Recognition and Machine Learning Techniques

Ho Chun Huang
Department of Oceanography
Naval Postgraduate School
Monterey, CA, USA

Ming Jer Huang
Coast Guard Radar Project
Mercuries Data Systems Ltd.
Taipei, Taiwan, ROC

John Joseph
Department of Oceanography
Naval Postgraduate School
Monterey, CA, USA

Tetyana Margolina
Department of Oceanography
Naval Postgraduate School
Monterey, CA, USA

Abstract*— a novel approach has been developed for detecting and classifying foraging calls of two mysticete species in passive acoustic recordings. This automated detector/classifier applies a computer-vision based technique, a pattern recognition method, to detect the foraging calls and remove ambient noise effects. The detected calls were then classified as blue whale D-calls [1] or fin whale 40-Hz calls [2] using a logistic regression classifier, a machine learning technique. The detector/classifier has been trained using the 2015 Detection, Classification, Localization and Density Estimation (DCLDE 2015, Scripps Institution of Oceanography UCSD [3]) low-frequency annotated set of passive acoustic data, collected in the Southern California Bight, and its out-of-sample performance was estimated by using a cross-validation technique. The DCLDE 2015 scoring tool was used to estimate the detector/classifier performance in a standardized way. The pattern recognition algorithm's out-of-sample performance was scored as 96.68% recall with 92.03 % precision. The machine learning algorithm's out-of-sample prediction accuracy was 95.20%. The result indicated the potential of this detector/classifier on real-time passive acoustic marine mammal monitoring and bioacoustics signal processing.

Keywords—whale; pattern recognition; machine learning; foraging call

I. INTRODUCTION

Fin whales produce a variety of low-frequency sounds. The 20-Hz pulse, based on average frequency of maximum level in early Atlantic recordings, is the most often reported fin whale sound worldwide e.g., in the Pacific, the Atlantic and the Southern Oceans [2], [4], [5], [6], [7], [8], [9], [10]. However, only males have been found to produce 20-Hz pulses and it is possibly a reproductive strategy [11]. Another 40-Hz calls, is produced by fin whales apparently in feeding contexts without gender exception [2]. This call is more variable in character and only classified manually so far, but should be more adequate for population estimation.

Blue whales have also been well-documented for four low-frequency sounds. The best-described pulsed A-calls and tonal B-calls have well-defined frequency content, long duration (15-20 sec) and repetitive sequences, thus using a matched filter or spectrogram correlation method has proven effective for detection and identification of these calls. However, like fin whale 20-Hz pulses, those calls are only produced by males. Blue whales also produce highly variable D-calls related to feeding behavior [12] without gender exception. Detectors for this call type are still being developed. Currently this call is also classified manually. The fourth type of blue whale sound is a highly variable amplitude-modulated and frequency-modulated call [13], [14] but its behavioral significance is unknown.

The foraging calls of both blue and fin whales present a challenge for conventional detection techniques because of their higher variability and lower stereotype as compared to other types of mysticete vocalizations with perhaps the exception of humpback songs. The foraging calls are statistically separable, especially when considered within a context of long trains of vocalizations, since the D-calls of blue whales typically have a distinctly broader bandwidth and longer duration than the 40-Hz fin whale calls. It is however not a trivial task to discriminate between the two call types in real data because of their similar downsweeping over a wide frequency range, as well as possible overlapping in call duration. The blue and fin whales are closely related species and their foraging calls have similar characteristics (Fig. 1). Furthermore, ambiguity of identifying the calls is caused not only by features inherent to a specific call type but also by effects of background noise distorting the real calls as recorded. The approach developed here applies two steps to spectral images of acoustic data. At first, a computer-vision based technique is used to detect the foraging calls and de-noise the images. The detected calls are then identified as blue whale D-calls or fin whale 40-Hz calls using a logistic regression classifier.

978-1-5090-1537-5/16/$31.00 ©2016 IEEE

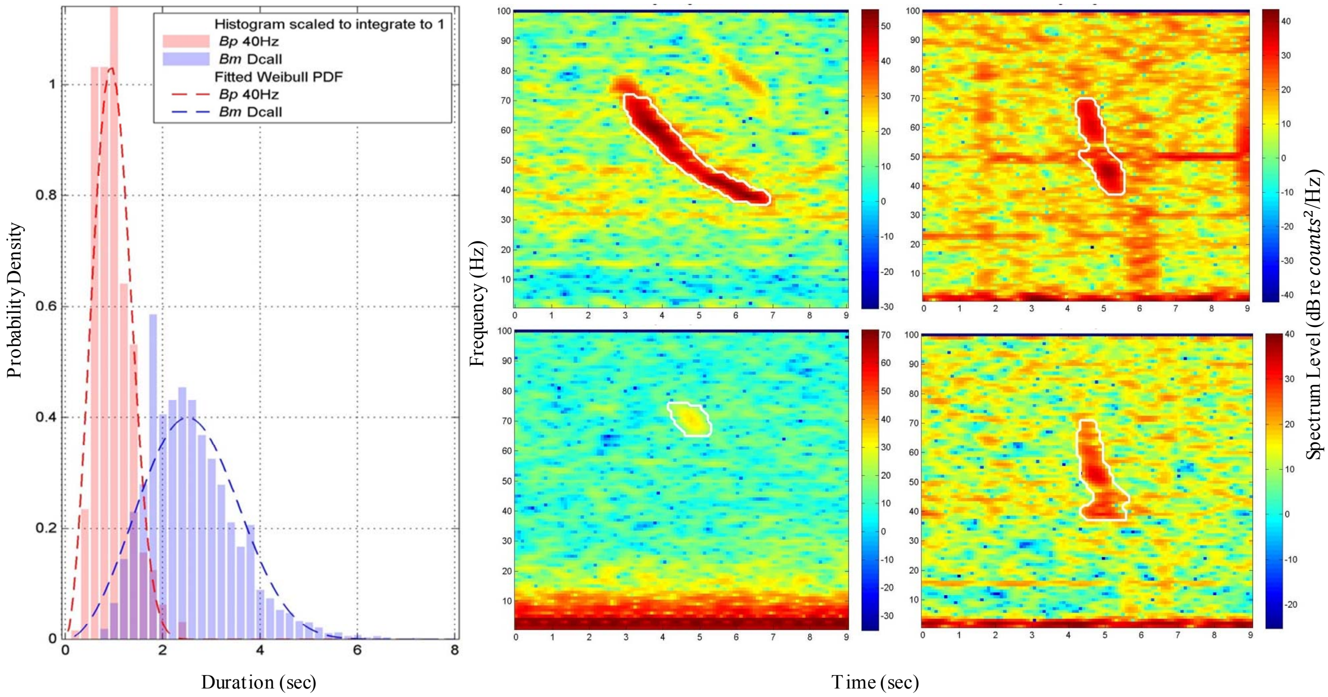

Fig. 1. *Left: The distribution of fin whale 40-Hz calls and blue whale D-calls duration (DCLDE 2015 dataset). Middle: The spectrogram of a typical D-call (top) and a typical 40-Hz call (bottom). Right: The spectrogram of a short D-call (top) and a similar 40-Hz call (bottom).*

II. DATA

A. DATA

The Detection, Classification, Localization and Density Estimation (DCLDE) 2015 dataset contains separate sets for mysticetes and odontocetes, and consists of data from multiple deployments of High-frequency Acoustic Recording Packages, HARPs [15], deployed in the Southern California Bight between 2009-2013 (Fig. 2). The annotated mysticete data were selected to cover all seasons from three locations and have been decimated to 1 and 1.6 kHz bandwidth [3].

The traditional protocol of visual scanning was applied using the custom MATLAB-based program Triton [15] to generate Long-Term Spectral Averages (LTSAs), with 5s temporal and 1Hz frequency resolution. Analysts manually scan the LTSA for the presence of foraging calls and further examine the call within a shorter time scale spectrogram with higher temporal resolution and specific frequency band. This research used the annotations to locate foraging calls to simulate the examining part of the manual protocol.

According to the annotations, there were 4,504 D-calls and 320 40-Hz calls in the mysticetes dataset. For each call, the annotations provide six lines of information: project, site, species-abbreviation, start-time, end-time, and call name. A MATLAB routine used start-time and end-time to locate calls and generate power spectral images with 9-second duration and 0-99Hz frequency band. Since there were two sampling rates, the number of points used to form each Fast Fourier Transform was chosen equal to each sampling rate in order to obtain consistent frequency and temporal resolution. By using 91% overlap, each image was represented by a 100 by 101 pixel matrix with 0.9s and 1Hz resolutions; the intensity of each pixel was color-coded. Additional information is also saved for each image for further evaluation.

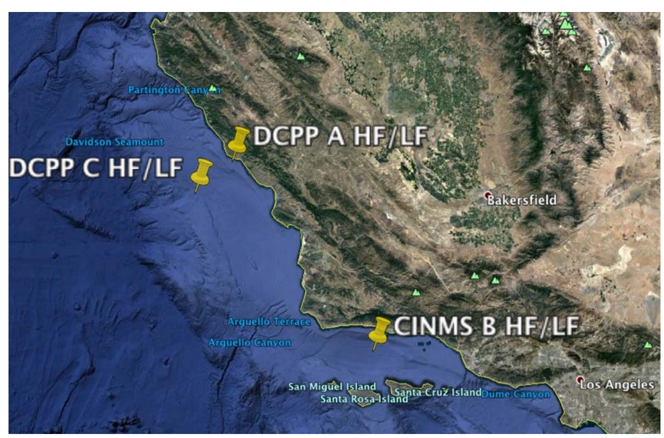

Fig. 2. *DCLDE 2015 data are provided from 3 different locations offshore Southern California as shown in yellow pins. CINMS-B site HARP recorded at 200 kHz sampling rate at 600m depth. DCPP-A site HARP recorded at 320 kHz sampling rate at 65m depth. DCPP-C site HARP recorded at 200 kHz sampling rate at 1000m depth..*

In order to evaluate the machine learning algorithm's performance when making predictions on new datasets it has not been trained on, the whole dataset was partitioned into three subsets, training (60%), testing (20%), and validation (20%), randomly so that each subset would still contain features of different sites and seasons for cross-validation, a model assessment technique. The machine learning algorithms are trained by the training subset and tuned by the testing subset. The out-of-sample performance is estimated by the validation subset.

III. METHOD

A. *PATTERN RECOGNITION*

The pattern recognition component applied pre-defined rules as well as a series of morphological image processes to the spectral images of both types of foraging calls. Calls are not distinguished in this step but classified later by machine learning algorithms. The fundamental rules were knowledge-based, translating expert knowledge of a trained analyst into algorithms, and the supplemental rules were data-based, as observed from DCLDE dataset. The optimum goal of pattern recognition was to filter all noise out and preserve a precise contour of every call's spectral image. The basic hypothesis was that the machine learning should be able to accurately classify two different calls if the inputs were almost exact contours with limited noise.

The first rule was to limit the frequency between 25 to 90 Hz. This frequency range was based on an expert's experience and the statistics of the training subset. In the training subset, only 1.73% (50/2895) calls were slightly above 90Hz and no call was below 20Hz. The low limit was designed to filter out fin whale 20Hz pulses which were often present at the same time as foraging calls. The upper limit was designed to filter out all the other sound sources. Though the contours of those 50 calls were not preserved exactly, the overall performance of both pattern recognition and machine learning algorithms was not affected by this imperfection. There was always a tradeoff between preserving as much frequency content as possible and de-noising, and the result indicated that this threshold was adequate. After applying this rule, each image was represented by a 71 by 101 pixel matrix.

The second rule kept only 4% pixels with highest intensity. This simulated the methods of changing brightness and contrast in the Triton display. By changing brightness and contrast, analysts were able to enhance the strength of the call so they can distinguish between calls and ambient noise. The assumption was that the call should be stronger than ambient noise in order to be found. Loud broad and narrow band noise would also be filtered out here by using the semi-equalization process and a de-noise function. The semi-equalization process, similar to the equalization functions in Triton display, only affects constant noise but not the whole image. Unlike the Triton equalization function, which is only able to mitigate the influence of narrowband noise, the semi-equalization process could mitigate both narrow and broadband noise or even remove the whole noisy column (broadband) or row (tonal). Another supplemental de-noise function could remove specific random noise. The 4% criteria is based on the amount of remaining pixels after applying the semi-equalization process and the de-noise function. The number of pixels needed to be limited so only calls could be preserved but not the noise. This 4% threshold was generated by evaluating both training and testing subsets, and could be adjusted depending on signal and noise strength.

The third rule described the size of foraging calls. Pixels next to each other are grouped by this rule. Groups with less than 6 pixels are filtered out. A 3-by-3 matrix is then used to dilate and smooth the remaining groups. This matrix can also be considered as another threshold since different matrices would dilate a group in different ways and might connect the group to other, unrelated groups. The pixels are grouped again and groups with less than 45 pixels are filtered out. The number of dilations can also be considered as another threshold.

The fourth rule connected broken parts of a call. A call can be separated by broad-band or narrow-band noise or just have two parts due to random discontinuity in sound. The thresholds of this rule were angle and length of bridges needed to connect the separated pieces. Since the slopes of foraging calls and the width or height of noise vary, it was difficult to determine an adequate bridge that was able to connect disrupted signal parts but not noise. By evaluating the training and testing subsets, three bridges were used in this rule. All three bridges had the same length of 5 pixels and used 20, 45, and 70 degrees angle, respectively, to follow the downsweep.

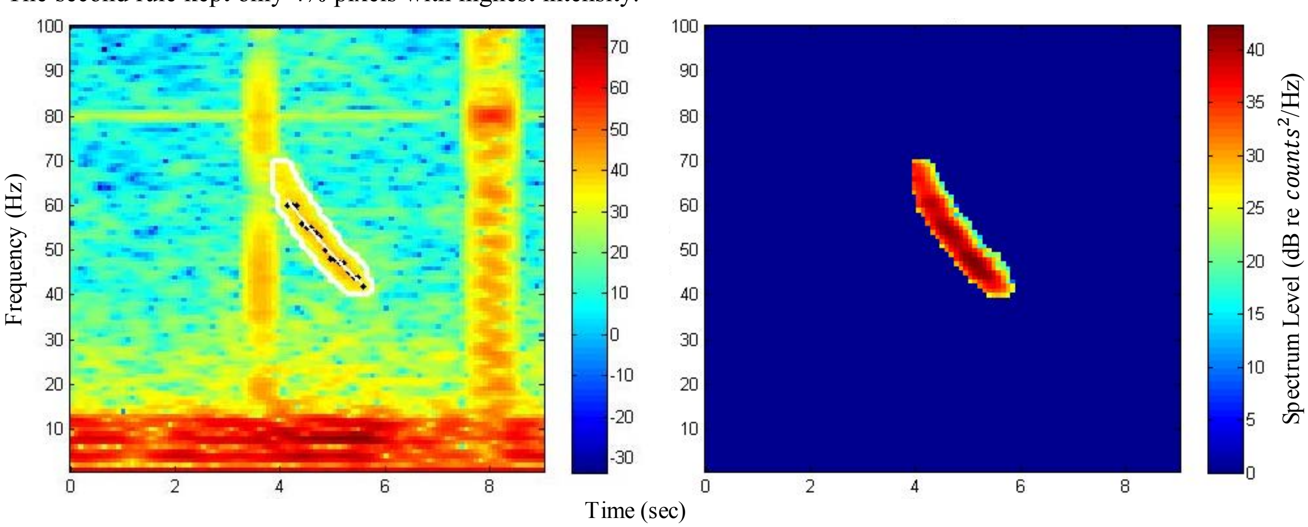

Fig. 3. *Left: The original image contains a lot of ambient noise. The white thin line connecting black spots, the highest intensity points of each temporal column, represents the ridge of the call. Right: After the pattern recognition step, the call is presented as a clear contour but not the whole image.*

The last rule was based on the critical feature of the baleen whale foraging calls: frequency downsweep. Any sound source without this feature would not be considered as a baleen whale foraging call. The highest intensity points of each temporal column in a group were selected to represent the ridge of this contour. The frequency downsweep feature requires the slope of this ridge to be negative. There were additional thresholds in this rule. The frequency bandwidth must be larger than 9 Hz and the duration must be longer than 0.5 second. The ratio of frequency and duration must be larger than 4 Hz/sec. Considering some D-calls with long duration but narrow frequency bandwidth, the ratio thresholds decreased to 3 Hz/sec if the duration was longer than 2.5 second. In case of increased duration by connecting to a regular narrow band noise, a de-noise loop would remove that noise when the rule was first implemented. The products of pattern recognition algorithm would be clear contours without ambient noise (Fig. 3).

B. MACHINE LEARNING

The machine learning technique is a series of computer algorithms which facilitate the learning processes and become an intelligent system. Its ability of learning from data is crucial in this study since there was limited analysis in the past to describe the difference between the two foraging calls. This study used the Logistic Regression method to produce a hypothesis which can predict the probability of correctly classifying blue and fin whale foraging calls. The initial hypothesis would have significant error; however, the logistic regression used the gradient descent technique to minimize the in-sample error. Provided with the training subset, the algorithm learns the difference between the two foraging calls and create a prediction function. It needed a sufficient number of iterations and an appropriate iteration step to determine the optimal in-sample error. The training subset was used multiple times to adjust the algorithms. Once the prediction function was sufficiently accurate to classify two calls of the training subset, the testing subset would then be used to test the prediction function. The testing subset could also be used multiple times to estimate the performance of prediction function in different situations. Finally, the validation subset was inputted to evaluate the out-of-sample performance of the prediction function. The validation subset was only used once to measure performance without the algorithm be further modified. This cross-validation method prevents overfitting and demonstrates that the prediction function can fit different situations.

IV. RESULTS AND DISCUSSION

A. RESULTS

The DCLDE 2015 scoring tool was used to estimate the algorithm's performance in a standardized way. A reliable detector must have high recall (detected ground-truth calls / total ground-truth calls) and precision (correct detections / total detections) values, meaning it would provide a minimum number of false detections and missed calls. There were 20 thresholds applied in the pattern recognition algorithm with 4 of having a the most sensitivity. To optimize both recall and precision, thresholds were determined by observing the in-sample results of different combinations (Fig. 4). The final major thresholds were 4.0%-pixel with 5-pixel-bridge and one-dilation at 25-90 Hz frequency band. This combination balanced precision and recall values and weighted the recall more than precision. Due to some calls being close to other calls, an image could contain more than one real call but one of them could be partitioned into validation subset which was the out-of-sample data (Fig. 5). When the dataset was partitioned, its annotation file was partitioned too. Once the algorithm detected the out-of-sample calls during the training procedure, those "correct detections" would be determined as false detections due to the scoring tool inability to pair the detection information to validation annotation. The precision value could be higher if the dataset was not separated into three subsets.

B. DISCUSSION

The out-of-sample performance, 96.68% recall with 95.20% prediction accuracy, is the core result of this research (Table I). Even the most well-trained analysts cannot have uniform performance. The performance of an individual analyst is also very difficult to be quantified. Moreover, the development of recording technique provides numerous data which are unpractical to be manually processed. The advantages of the algorithm are reproducibility, known performance, and cost-efficiency. It can accomplish two missions: post data processing and real-time monitoring.

TABLE I.

Performance	Data	
	In-sample Performance Training and Testing(3860 calls)	***Out-of-sample Performance Validation(964 calls)***
Precision	0.9256	0.9203
Recall	0.9710	0.9668
Truth Coverage Overall percentage	83.35	82.68
Detection Coverage Overall Percentage	85.20	84.41
Empirical expected fragmentation	1.0774	1.0655
Prediction Accuracy Percentage	97.12	95.20

The blue whale calls throughout the world are geographically distinct so that different "acoustic populations" have been suggested as a means to distinguish among whales [16], [17], [18]. The study of whale population and whale behavior requires adequate data to be analyzed. The obstacle of

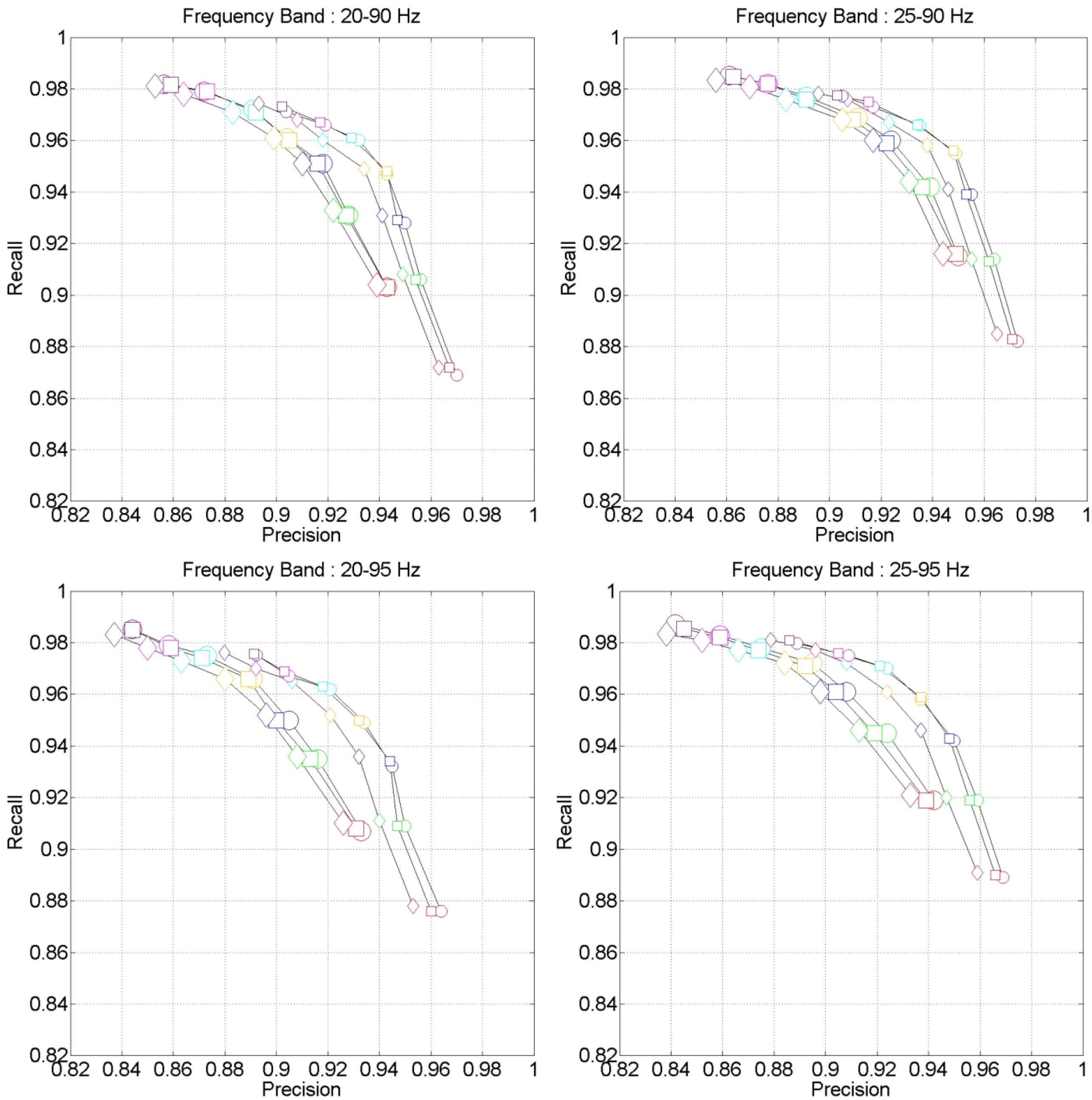

Fig. 4. *Based on four frequency bands, symbol's size represents dilation times: small for one-time and large for two-times. Symbol's shape represents bridge length: circle for 5-pixel, square for 7-pixel, and diamond for 9-pixel. Symbol's color represents the percentage of highest intensity pixel: red, green, blue, yellow, light blue, pink, and purple for 2.0, 2.4, 2.8, 3.2, 3.6, 4.0, and 4.4, respectively.*

analyzing blue and fin whale foraging calls is limited annotated data. Although this algorithm is not perfect and still needs the analysts to confirm its products, it will facilitate many related research efforts since manually processing of thousands of images is much more efficient than processing years of raw data. For the post data processing purpose, it is more appropriate to use a higher percentage of pixel since this will increase the recall value and false detections can be filtered out manually.

The development in real-time passive acoustic monitoring of baleen whale occurrence and distribution from autonomous platforms requires progress on both hardware and software systems [19]. Current software does not have the ability of detecting and classifying baleen whale foraging calls. This algorithm could enhance the existing software. Although

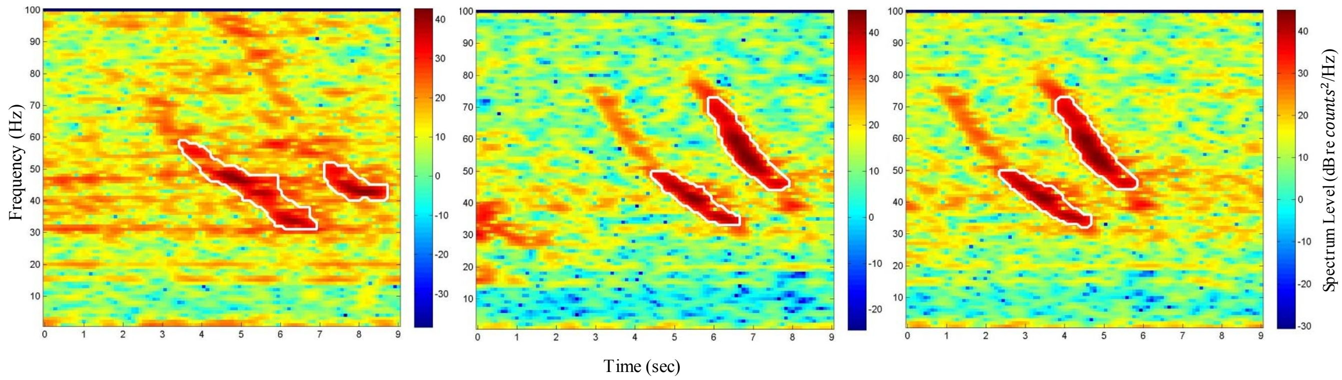

Fig. 5. *Left: Two real calls were correctly detected (white outline) but the right one was partitioned into the validation subset which made it be determined as a false detection by the scoring tool. Middle and Right: Two real calls were correctly detected twice which made the second one be determined as a fragmentation by the scoring tool.*

acoustic monitoring methods have become more important, visual surveying is still a trustworthy approach for marine mammal population estimation and their behavior analyzation. With the help of real-time acoustic monitoring, visual observation will be much more accurate. For the real-time monitoring purpose, equally weighting the recall and precision value is recommended thus 3.2% pixel will be a better threshold.

Machine learning was used independently, without the pattern recognition, in the very beginning of this research. Although the in-sample prediction accuracy was very high, the out-of-sample accuracy was less 50% because of overfitting. This study demonstrated the potential of the pattern recognition technique which can be combined with other classifiers to enhance their capability.

Direct application of the detector/identifier to large datasets may be computationally inefficient. For example, a one-year recording will require generating and processing of 3,504,000 images, many of which contain no feeding calls. To optimize a process, we will supplement the detector with a pre-screening module to scan LTSAs for sections of data with potential signals of interest. Results of this work in progress will be described in a separate paper.

ACKNOWLEDGMENT

This research was supported by the US Navy's Living Marine Resources (LMR) Program.

REFERENCES

[1] P. O. Thompson, "Marine Biological Sounds West of San Clemente Island: Diurnal Distributions and Effects of Ambient Noise During July 1963. San Diego,", U.S. Navy Electronics Laboratory, 1965.

[2] W. A. Watkins, "Activities and underwater sounds of fin whales," Deep Sea Research Part B. Oceanographic Literature Review, vol. 29, no. 12, p. 789, Jan. 1982.

[3] Scripps Institution of Oceanography UCSD, "Dataset documentation: 2015 DCLDE workshop," 2015. [Online]. Available: http://www.cetus.ucsd.edu/dclde/datasetDocumentation.html. Accessed: Apr. 1, 2016.

[4] P. L. Edds, "Characteristics of finback Balaenoptera Physalus vocalizations in the ST. Lawrence Estuary," Bioacoustics, vol. 1, no. 2-3, pp. 131–149, Jan. 1988.

[5] P. O. Thompson, "20-Hz pulses and other vocalizations of fin whales, Balaenoptera physalus, in the gulf of California, Mexico," The Journal of the Acoustical Society of America, vol. 88, no. S1, p. S5, 1990.

[6] W. Watkins et al., "Seasonality and distribution of whale calls in the north pacific," Oceanography, vol. 13, no. 1, pp. 62–67, 2000.

[7] C. W. Clark, J. F. Borsani, and G. Notarbartolo-Di-sciara, "Vocal activity of fin whales, Balaenoptera Physalus, in the Ligurian sea," Marine Mammal Science, vol. 18, no. 1, pp. 286–295, Jan. 2002.

[8] S. L. Nieukirk, K. M. Stafford, D. K. Mellinger, R. P. Dziak, and C. G. Fox, "Low-frequency whale and seismic airgun sounds recorded in the mid-atlantic ocean," The Journal of the Acoustical Society of America, vol. 115, no. 4, p. 1832, 2004.

[9] A. Širović, J. A. Hildebrand, S. M. Wiggins, M. A. McDonald, S. E. Moore, and D. Thiele, "Seasonality of blue and fin whale calls and the influence of sea ice in the western Antarctic peninsula," Deep Sea Research Part II: Topical Studies in Oceanography, vol. 51, no. 17-19, pp. 2327–2344, Aug. 2004.

[10] M. Castellote, C. W. Clark, and M. O. Lammers, "Fin whale (Balaenoptera physalus) population identity in the western Mediterranean sea," Marine Mammal Science, vol. 28, no. 2, pp. 325–344, Jul. 2011.

[11] D. A. Croll et al., "Bioacoustics: Only male fin whales sing loud songs," Nature, vol. 417, no. 6891, pp. 809–809, Jun. 2002.

[12] E. M. Oleson, S. M. Wiggins, and J. A. Hildebrand, "Temporal separation of blue whale call types on a southern California feeding ground," Animal Behaviour, vol. 74, no. 4, pp. 881–894, Oct. 2007.(a)

[13] A. M. Thode, G. L. D'Spain, and W. A. Kuperman, "Matched-field processing, geoacoustic inversion, and source signature recovery of blue whale vocalizations," The Journal of the Acoustical Society of America, vol. 107, no. 3, p. 1286, 2000.

[14] E. Oleson, J. Calambokidis, W. Burgess, M. McDonald, C. LeDuc, and J. Hildebrand, "Behavioral context of call production by eastern north pacific blue whales," Marine Ecology Progress Series, vol. 330, pp. 269–284, Jan. 2007.(b)

[15] S. M. Wiggins, E. M. Oleson, M. A. McDonald, and J. A. Hildebrand, "Blue whale (Balaenoptera musculus) Diel call patterns offshore of southern California," Aquatic Mammals, vol. 31, no. 2, pp. 161–168, Sep. 2005.

[16] P. O. Thompson, L. T. Findley, O. Vidal, and W. C. Cummings, "Underwater sounds of blue whales, Balaenoptera Musculus, in the Gulf of California, Mexico," Marine Mammal Science, vol. 12, no. 2, pp. 288–293, Apr. 1996.

[17] D. K. Mellinger and C. W. Clark, "Methods for automatic detection of mysticete sounds," Marine and Freshwater Behaviour and Physiology, vol. 29, no. 1-4, pp. 163–181, Jan. 1997.

[18] K. M. Stafford, S. L. Nieukirk, and C. G. Fox, "An acoustic link between blue whales in the eastern tropical Pacific and the northeast Pacific," Marine Mammal Science, vol. 15, no. 4, pp. 1258–1268, Oct. 1999.

[19] M. F. Baumgartner et al., "Real-time reporting of baleen whale passive acoustic detections from ocean gliders," The Journal of the Acoustical Society of America, vol. 134, no. 3, p. 1814, 2013.