North Atlantic Right Whale Acoustic Signal Processing:

Part II. Improved Decision Architecture for Auto-Detection Using Multi-Classifier Combination Methodology

Peter J. Dugan*, Aaron N. Rice, Ildar R. Urazghildiiev and Christopher W. Clark
Bioacoustics Research Program, Cornell Laboratory of Ornithology,
Cornell University
Ithaca, NY 14850 USA
*e-mail: pjd78@cornell.edu

Abstract—**Autonomous signal detection of the North Atlantic right whale (NRW), *Eubalaena glacialis*, is becoming an important factor in monitoring and conservation for this highly endangered species. Both online and offline systems exist to help study and protect animals within this population. In both cases auto-detection of species-specific calls plays a vital role in localizing individual animal by searching time-frequency passive acoustic data. This research presents an experimental system, referred to as the NRW-CRITIC, for automatic detection of the NRW contact call. In general, the CRITIC uses a combinatorial classifier approach to integrate a series of existing machine learning algorithms; each designed specifically for NRW contact call identification. The proposed configuration consists of several recognition methods running in parallel; these include linear discriminant analysis, artificial neural network (NET) and classification regression tree (CART). This paper presents the details for the NRW-CRITIC and discusses the approach used to combine multiple independent decisions into a single result. A side-by-side performance comparison, between the CRITIC and a well-known method, the feature vector testing (FVT), is summarized. Performance metrics are evaluated based on a large database of acoustic recordings consisting of over 58,000 NRW contact calls from various locations, including two critical habitats, Great South Channel and Cape Cod Bay. Results indicate the FVT algorithm yields a74.7% detection probability with an error rate of 4.35%. In comparison the CRITIC, operating at similar information level yields a 78.02% detection probability with a 3.25% error rate, exceeding the performance of the FVT. Performance was also measured using data from a multi-channel acoustic array located in Massachusetts Bay. A side-by-side comparison of array presence is discussed for two separate days. Results show that with the FVT and CRITIC operating at 0% error for array presence, the FVT method had 18,769 and 24,469 false positives for the Massachusetts Bay datasets respectively. With the same 0% error condition the CRITIC provided successful detection with significantly lower number of false positive rates: 1,072 and 2,324 calls, respectively. Future extensions of this experimental work are also discussed.**

Keywords--Right Whale; Acoustic Monitoring; Automated Detection; Multi-Classifier

I. INTRODUCTION

The North Atlantic right whale (NRW, *Eubalaena glacialis*) is a critically endangered marine mammal. Passive acoustic monitoring (PAM) for NRW sounds is increasingly a significant component for detection and conservation management [1, 2]. Right whale species produce a frequency-modulated vocalization that increases in frequency (referred to as a contact call or "up-call") that functions as a contact call between individuals [1, 3]. Because the up-call is a common part of their vocal repertoire, it is used as the primary basis for acoustically detecting the occurrence of right whales [1, 2, 4, 5].

Underwater PAM combined with automatic detection technology is becoming a popular method for detecting marine mammal sounds. A typical tradeoff in most software allows the user to control detection thresholds: too low of a threshold results in high numbers of false alarms; too high of a threshold results in missed NRW calls. In either case, and despite recent advances in automatic detection, false alarm rates with some of the best detection algorithms to date still provide relatively high false positive rates, causing operators to spend significant time confirming reports and inspecting data. Changing ambient noise levels and call variability are the two major causes of error in the automatic detection software [6, 7], resulting in high numbers of false alarms.

Some of the earliest automatic detection approaches for NRW calls used single-stage algorithms. Edge detection [8] and time-frequency convolution [9] are good at finding NRW calls, but provide relatively high levels of false positive errors [7]. In an effort to reduce false positives, later approaches added on additional stages, including feature extraction and classification [7, 10]. Multi-stage classifiers generally perform significantly better than previously used single stage approaches for the NRW problem, though they still result in a significant number of false alarms [7]. Similarly, alternate classifier technologies have been used for NRW detection [6], and provide some increase in detection performance with small, but significant, increases in error rates. However, such methods [6] only use a single classifier for determining output results.

In recent years, there has been increased focus within the pattern recognition community on combining classifiers to improve overall system performance at the cost of higher system complexity. Voting methods and consensus theory play an important role in combining multi-expert systems [11].

978-1-4244-5550-8/10/$26.00 ©2010 IEEE

These social mechanisms were discussed in terms of artificial intelligence by [12]. Early democratic human cultures, used voting (a form of group decision theory) to establish order and make decisions; recent cultures use opinion polls as a tool to evaluate consensus within groups. According to [11] there is significant interest in moving away from systems that have complex classifiers. One consideration is to use a series of less complex classifiers, combining the results into a single decision for improved recognition performance, though this approach is not unanimously accepted [13]. There are at least three reasons for using multiple classifiers: these are statistical behavior, computational behavior and representational behavior [11]. Statistical behavior diminishes the effect of picking one classifier that may provide worse predictions than the others. Computational behavior protects against classifiers that train to local extremes, in some cases the optimal learning point may never be attained as a result of random weight initialization. Some problems have requirements that exceed the operational behavior, or representation, of the classifier; for example, a linear classifier used on a problem that requires a non-linear decision surface. The fundamental idea is to use multiple classifiers to help average out the fore mentioned effects, providing more stable performance. Combined classifiers have been successfully used in systems that handle large amounts of online data that continuously update in the presence of variable inputs. Early work by [14, 15] incorporated techniques called "bagging" and "boosting" for training an ensemble of classifiers whereby decisions are made using a voting methodology. Since both of these configurations exploit variability of the training set, optimal classifiers are selected based on a group vote. Others approaches, such as the Random Forest method [16], extend the concept of voting to randomizing the feature space as well as the training data for regression tree classifiers. Early computer memory constraints required alternate methods whereby multi-expert systems were trained using samples from an online database by the Hedge(β) algorithm [15]. Further work from [17] uses a multi-classifier system, which is trained using a series of several short data samples taken over many data sets, called pasting small votes; promoting final decisions through voting rules.

To our present knowledge, the whale bioacoustic scientific community has not employed combinational classifier architectures for automated detection. The new approach presented here, NRW-CRITIC, integrates three different NRW classifiers [6, 7]. The focus for improving pattern recognition primarily involves reducing false positive rates, while maintaining sufficient detection accuracy. For this study, accuracy is measured using traditional approaches, including assignment rate, false positive rates and missed detection rates. However, from the human operator's perspective, these measures are not alone sufficient indicators of NRW presence, therefore detection accuracy is considered in the context of multiple detections of the same call within a multi-channel recording array of sensors. The array system allows for any one of a plurality of sensors to detect an animal. However, inspecting several channels can consume significant operator time and resources; therefore, reducing false alarms is important. The final objective is to reduce the amount of operator time required to analyze detector results, without missing the identification, or observation that a NRW call occurred, meaning that a whale was present. The goal of this study is to present an experimental, new technology for automatically detecting NRW vocalizations by combining the results of independent classifiers in order to develop a single, more accurate answer for improved pattern recognition.

II. METHODS

A. System Operation, Input Data Collection

This study proposes an experimental system, as shown in Fig. 1, designed to recognize NRW calls using multiple classifiers. Acoustic data are recorded in the field, and post-processed in the laboratory using a multi-stage detection system [6]. Input data, x, consists of 11 features as documented in [6]. Two sources of acoustic data are used for this report. The first set is from a series of recordings in Cape Cod Bay (CCB), Great South Channel (GSC) and Brunswick, GA [data are summarized in 6]. This dataset contains over 58,000 NRW contact call events and more than 70,000 noise samples. Classifiers used herein were trained using portions of this database; details are discussed in [6]. A second series of acoustic recordings are taken from a 19-unit hydrophone array deployed in Massachusetts Bay (Mass-Bay). For the focus of this study, data from two arbitrary days of recordings are used, 25 February, 2008 (Mass-Bay-20080225), and 31 March, 2008 (Mass-Bay-20080331). The classifiers used in the NRW-CRITIC were completely blind to any data measured in Mass-Bay recordings; thereby this is a completely independent set isolated from training, testing and validating samples.

B. Classifiers

This study proposes an experimental system shown in Fig. 1 called the NRW-CRITIC. Input samples, x, are presented to a cascade network of multiple algorithms, each providing an independent class label. Classifier algorithms consist of feature vector testing (FVT), classification regression tree (CART) and the artificial neural net (NET). The FVT algorithm uses a linear discriminant analysis described in [7]. The CART and NET are modeled after [18]. Both the NET and CART topologies are considered instable classifiers [11], and designed to offer variability in decision making. The CRITIC was configured using a total of nine classifiers, three FVT's, three NET's and three CART algorithms. To give equal representation of the algorithms, the FVT had three instances of the same training weights, the NET and CART each contribute another three classifiers, trained on data from Cape Cod Bay, Great South Channel and a Generic set, comprised of a series of random samples drawn from a database of over 58,000 NRW calls and 70,000 noise samples. Training for each expert was done with replacement, using roughly 10% of the data set [6]. The NET and FVT have distance-based outputs, and classifiers were trained to provide result such that noise signals were represented by 0 and NRW calls by 1. As mentioned by [11, 18]; labels can be accommodated by confidence scores, quality measures, or probability values. The FVT algorithm did not provide such measures, only a class label. To keep consistent, the CART and NET did not provide quality either, only class labels. For the FVT and NET, samples with the closest distance to either extreme were labeled accordingly. The FVT

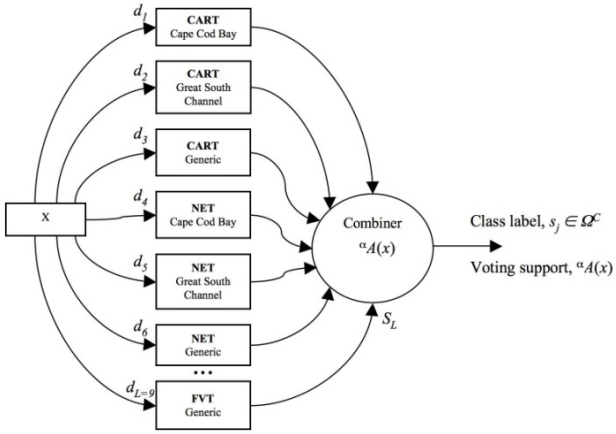

Figure 1. System architecture for the NRW-CRITIC. The input vector *x*, defined by the feature set in [6] is used as input to *L*=9 classifiers. Output from classifiers combined using a decision algorithm capable of providing a vote along with a support measure defined by the α level.

applies penalty functions for values beyond either extreme (0, 1) [7]; therefore, data are guaranteed to be in the interval. The NET however, used a linear activation, practically 90% of the data existed on the interval between (-1, 1) and a threshold was selected heuristically to map output decisions to the (0, 1) interval. The CART is inherently non-distance-based, providing a discrete answer for an output label. Together, all classifiers provide a label for each input. Therefore, for input vector *x* an output may look like [0 0 1 0 1 1 0 0 1], indicating that classifier 3, 5, 6 and 9 each identified the event as a NRW call. Although noise samples were present in the training set, they did not provide a separate class vector. Instead, the single observed NRW recognition was considered the lowest value to support a contact-call detection.

C. Combining Classifiers, the NRW-CRITIC approach

Building systems that use multiple classifiers is becoming more common [19, 20]. These systems use multiple inputs to make a single decision. In a multi-expert system, there are two modes of operation, fusion and combination [21]. Classifier fusion combines data in order to reduce system error and improve decision accuracy. Classifier selection aims at selecting the optimal classifier to make decisions in a partitioned decision space.

The NRW-CRITIC operates using a fusion approach (Fig. 1). The notation used to develop the cascaded architecture is borrowed in part from [11], which assumes a series of *L* classifiers arranged in a parallel format, with combined output vector for each class D^C. Each classifier, d_i, is connected such that $d_i \in D$, $i = 1, \ldots, L$, or $D = [d_1, d_2, \ldots, d_L]$ or with a *C* dimensional output $D^C = [d_{i,1}, d_{i,2}, \ldots, d_{i,c}]$. The output of each classifier, per class, is labeled s_i such that $s_i \in \Omega^C$ and the value of d_i ranges from $[0,1]^C$ to support class *c*. Information or sample, *x,* is combined for each class such that,

$$A(x) = \max_{j=1}^{c} \frac{1}{L} \sum_{i=1}^{L} d_{i,j} \quad (1)$$

Equation (1) provides a measure per class, where $A(x) \in \mathbb{R}^n \rightarrow [0,1]$, is a real number ranging from 0 to 1 on a closed interval. Only considering two classes simplifies our result, such that any non-signal is considered noise. Also, tie-breaker votes do not have to be computed as in [11]. The combined output statistic $A(x)$ has discrete levels. For this study, the combiner is modeled using a crisp set notation [22],

$${}^{\alpha}A = \{x \in X \mid A(x) \geq \alpha\}, \quad (2)$$

where α will be referred to as the *alpha-cut* value and *x* as the input sample. The various α levels are computed such that $\alpha \in \mathbb{R}^n \rightarrow [0,1]$. This means, if α=0.5, then at least 50% of classifiers in D^C are required to make a decision in favor of class label Ω^C. Similarly α=1.0 means that all classifiers in D^C must agree on the same label. A graphical representation for different α levels is show in Fig. 2, which illustrates the overall classifier space, represented by the outer ring. This space coincides with $\alpha > 0$, where objects have at least one classifier in D^C observing the true label. The smallest ring is the case with α=1, which encloses the collection of all objects, whereby D^C agrees unanimously. By modeling each true identification as a vote, the smaller inner ring coincides with the highest level of consensus by all voters [11], providing support for the given decision.

D. Performance Analysis

Performance is measured for NRW by using assignment rates and error rates. As pointed out in [23], detection probability p_d, often referred to as true positives, is the ratio of total NRW calls detected n_d to the total true calls available n_t. Errors occur in two ways: false alarms or false positives (Type I error), and missed detections or false negatives (Type II error). False positive rate is the ratio of the total non-calls detected n_{fa} divided by the total non-calls present n_{nse}. For this study, we will consider a false positive rate, e_{far}, which we

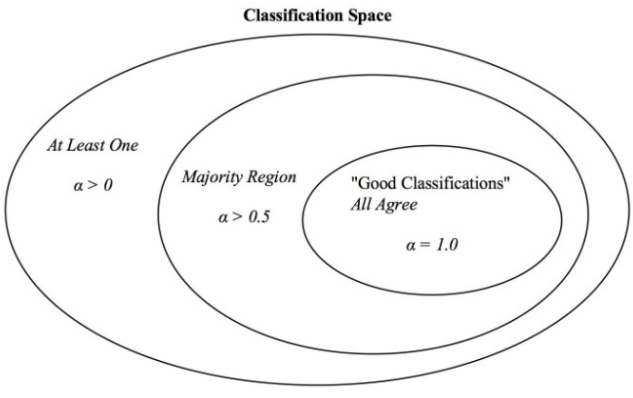

Figure 2. Representative classification space, where the outside region is the entire classifier space, sub regions represent different alpha cuts. The outside region, *At Least One,* represents any one classifier determining a result. The next region inward, *Majority Region,* corresponds at least fifty percent of the classifiers determining the same label. The inside ring, *All Agree,* corresponds to all classifiers making the same prediction, this is labeled as the "Good Classifications" region following [11].

define as the total non-calls detected n_{fa} divided by the total true calls n_t. The second error, probability of missed detections, e_{md}, is the total NRW calls missed n_{md} divided by the total true NRW calls, n_t, available. To compare the operating point of the FVT to the CRITIC, interpolated assignment probabilities and error rates are computed for intermediate α values. The following relationship is used:

$$e_{f_{ar}}^{(i)} = \left(e_{f_{ar}}^{(i-1)} - e_{f_{ar}}^{(i+1)}\right)\left(\frac{p_d^{(i)} - p_d^{(i+1)}}{p_d^{(i-1)} - p_d^{(i+1)}}\right) + e_{f_{ar}}^{(i+1)}, \quad (3)$$

where $e_{f_{ar}}^{(i)}$ and $p_d^{(i)}$ are the i^{th} false alarm rate and percent detection probability, respectively, that correspond to a given α level.

To put the NRW-CRITIC performance into a "real-world" data analysis context, this study considers the auto-detection performance using acoustic data recorded from two days within a 19 unit PAM array in Massachusetts Bay, where the units are separated by 5 km from one another. Therefore, over a surveillance period, the total calls measured by all sensors and confirmed by operators is n_t. The total number of calls sensed by the auto detection algorithm is n_d. Because NRW up-calls propagate many kilometers and can be recorded by multiple units in the array [1, 2], this means that many signals are redundant. Taking this into account, recording units can serve as a built-in mechanism to confirm true NRW signals, reducing the work load by allowing the operator to discount confirmed up-calls measured on multiple hydrophones. Hourly presence of NRW within or near the Massachusetts Bay recording array is a relevant mitigation-related metric for evaluating their occurrence, so the performance of both algorithms was evaluated in the context of detecting the number of whale calls per hour. For each hour, a total array score, a_t, is determined as the number of sensors that correctly detected a NRW call, as confirmed by the operator. Therefore if 2 out of 19 sensors report correct NRW calls for the given hour, $a_t = 2$. Likewise, the detection score a_d is the total number of sensors that correctly auto detect an NRW call without operator intervention. The hourly errors for the array, a_e, is equal to 1 for each hour the auto detection algorithm failed to detect at least one NRW call on any sensor, providing one was present. Array false alarm count, a_{fac}, is the total number of false alarms recorded across all channels per hour and represents the operator workload where the operator must inspect all the results, $a_t + a_{fac}$, to confirm detections.

III. RESULTS

The FVT and NRW-CRITIC were run on the database of 58,624 NRW calls, and results p_d, and e_{far} are summarized in Tables I and II, respectively. Several α values $\{\alpha = 0.0, 0.6, 1.0\}$ for the CRITIC were computed. These are indicated in Tables I, and II. Identical input samples were used for the FVT and CRITIC. These were generated as in (Dugan et al., 2010). The operating point for the CRITIC is also computed and corresponds to the FVT algorithm using the intermediate computation (Eq. 3). For $p_d^{(i)} = 74.70$, and using p_d and e_{far} values which correspond to $\alpha = \{0.6, 0.7\}$ in Tables I and II,

TABLE I. PROBABILITY OF DETECTION OR FVT AND NRW-CRITIC, FOR THE NRW CRITIC VARIOUS α LEVELS ARE SHOWN. $\alpha = 0.0$ MEANS ALL CLASSIFIERS ARE COUNTED, $\alpha = 0.6$, ONLY DETECTIONS WITH AT LEAST 60% OR MORE OF THE CLASSIFIERS THAT AGREED, $\alpha = 1.0$ UNANIMOUS VOTE, ALL CLASSIFIERS AGREE.

Data set	Total Signals	NRW FVT	NRW CRITIC		
			$\alpha = 0.0$	$\alpha = 0.6$	$\alpha = 1.0$
CCB00	15672	83.17	85.93	76.51	67.96
CCB02	1812	78.81	80.63	69.65	60.65
CCB03	120	83.33	89.17	80.00	70.00
CCB04	3056	82.33	86.09	73.69	65.02
CCB05	1791	83.08	94.03	87.88	81.30
CCB06	10438	82.96	92.36	87.79	81.60
CCB09	2215	83.79	89.03	84.70	78.10
CCB15	8304	69.50	85.36	77.20	69.79
CCB17	1693	64.62	79.80	72.18	62.61
ESI06	287	83.97	84.67	80.14	74.91
GSC00	791	50.32	84.45	67.26	52.59
GSC01	1239	49.15	85.31	66.67	51.65
GSC02	3000	61.40	84.53	71.77	60.23
GSC03	8002	58.45	88.17	74.33	61.52
GSC04	204	64.22	70.59	46.57	34.31
Total Calls	**58624**	**43791**	**51091**	**45598**	**40423**
Percent Calls	**100**	**74.70**	**87.15**	**77.78**	**68.95**

respectively, the interpolated CRITIC error is $e_{f_{ar}}^{(i)} = 2.87$ percent (data for $\alpha = 0.7$, not shown in Tables I and II, p_d=72.27, e_{far}=2.57,).

Results for the multi-channel Mass-Bay data were generated using the FVT and CRITIC. Settings for both of these approaches were identical to the results used in measuring p_d, and e_{far}. Values for hourly total calls, at, automatic detection score, a_d, hourly errors, a_e, and false alarm count, a_{fac}, were computed for the Mass-Bay array data (Table III). For the CRITIC, $\alpha = 0.6$ provided zero errors for a_e and a total automatic detection score $n_d = 425$ for Mass.Bay-20080331, $n_d = 63$ for Mass.Bay-20080225. The false alarm count $a_{fac} = 1{,}072$ for Mass.Bay-20080331, $a_{fac} = 2, 324$ for Mass.Bay-20080225. In comparison, the FVT algorithm also provided zero array errors a_e, and a total detection $n_d = 501$ for Mass.Bay-20080331, $n_d = 71$ for Mass.Bay-20080225. The FVT false alarm counts were significantly higher, a_{fac}=18,769 and a_{fac}= 24,469 for Mass.Bay-20080331 and Mass.Bay-20080225, respectively. For hourly presence, the FVT algorithm used the same settings as shown in Tables I and II, the CRITIC was set at $\alpha = 0.6$. This result was determined by running the CRITIC over several α levels and finding the value that minimized the error and false positive rates; which is the classical tradeoff between assignment rate (p_d) and errors (e_{md}, e_{far}).

IV. DISCUSSION

Comparing the results in Tables I and II, the FVT performed as follows. Percent detection probability, $p_d = 74.7$ and percent false alarm rate, $e_{far} = 4.35$. The CRITIC used a range of information levels. Using $\alpha = 0.0$ resulted in $p_d = 87.7$ and $e_{far} = 9.93$. Using $\alpha = 0.6$ resulted in $p_d = 77.78$ and $e_{far} = 3.25$. With all nine classifiers observing an NRW contact call

TABLE II. FALSE ALARM RATES FOR FVT AND NRW-CRITIC, FOR THE NRW CRITIC VARIOUS ALPHA CUT (α) LEVELS ARE SHOWN. $\alpha = 0.0$ MEANS ALL CLASSIFIERS ARE COUNTED. AT $\alpha = 0.5$, ONLY DETECTIONS WITH AT LEAST 50% OF THE CLASSIFIERS THAT AGREED. FOR $\alpha = 1.0$ ONLY THOSE SIGNALS ARE COUNTED WHERE ALL CLASSIFIERS ARE CONSIDERED TO BE THE SAME MATCH.

Data set	Total Signals	NRW FVT	NRW CRITIC		
			$\alpha = 0.0$	$\alpha = 0.6$	$\alpha = 1.0$
CCB00	15672	1327	1689	595	405
CCB02	1812	150	179	38	21
CCB03	120	10	11	2	2
CCB04	3056	160	137	29	21
CCB05	1791	20	87	25	20
CCB06	10438	128	555	245	180
CCB09	2215	200	240	118	87
CCB15	8304	305	1071	428	320
CCB17	1693	25	142	41	30
ESI06	287	26	86	17	9
GSC00	791	3	73	12	4
GSC01	1239	8	68	17	10
GSC02	3000	62	524	116	88
GSC03	8002	100	929	215	125
GSC04	204	25	31	7	2
Total Calls	**58624**	**2549**	**5822**	**1905**	**1324**
Percent Calls	**100**	**4.35**	**9.93**	**3.25**	**2.26**

($\alpha = 1.0$), the CRITIC provided percentages, $p_d = 69.8$ and $e_{far} = 2.26$. The $\alpha = 0.0$ and $\alpha = 1.0$ values can be viewed as the operating extremes, as represented schematically in Fig. 2. The outer-most classification region represents at least one classifier providing a decision *(dictator voting)*, the inner most coincides with all classifiers agreeing, or *(unanimous vote)*. It should be pointed out, that for the CRITIC, there is a steady, simultaneous decrease in detection rates and error rates from α=0.0 to α=1.0. The FVT has similar assignment rates between $0.6 < \alpha < 0.7$ compared to the CRITIC. Assuming the CRITIC provides piece-wise linear behavior and using the FVT detection percentage, the CRITIC operating at an alpha value between $0.6 < \alpha < 0.7$, would yield an equivalent error rate of $e_{far} = 2.87$ for $p_d = 74.7$.

For the array presence test, the FVT and CRITIC had operated with zero error, Table III. However, the FVT had significantly higher false positives than the CRITIC. For the two observed days, the FVT had a combined total false positive rate of 43,238 reports, the CRITIC had only 3,396 reports. This is an overall reduction by a factor of twelve, using the CRITIC over the FVT method.

The rationale for using multiple classifiers for the NRW problem has several points. First, this work is best described by the fusion type, whereby multiple classifiers are used to provide an enhanced decision space. However, this approach could be expanded to also include combinational requirements. Furthermore, the reason for using multiple classifiers for the NRW problem is based on several factors. First, evolving new classifiers as data becomes available is a practical consideration; second, limited resources and parallel collaboration efforts are ideal for an architecture that allows for additional classifiers to be added to the network instead of retraining new ones. And lastly, providing an architecture to make decisions will allow more efficient integration into online systems for future work.

TABLE III. NRW DETECTION PERFORMANCE OF FVT AND NRW-CRITIC ON MASSACHUSETTS BAY ACOUSTIC DATA.

	Mass.Bay-20080225		Mass.Bay-20080331	
	NRW FVT	NRW-CRITIC	NRW FVT	NRW-CRITIC
Number true calls (n_t)	71	71	502	502
Number found by detector (n_d)	71	63	501	425
Total channels with hourly presence (truth) (a_t)	31	31	60	60
Total channels with hourly presence (found by detector) (a_d)	31	29	60	57
Number errors (hourly presence) (a_e)	0	0	0	0
Total number false alarms (a_{fac})	24,469	2,324	18,769	1,072

The fundamental goal for automated classifier algorithms is to reduce the workload of human operators while maintaining minimal error conditions. In the Massachusetts Bay system, multiple sensors are used to meet the requirement of detecting an animal within a given hour. For the CRITIC, an operating point must exist that minimizes errors while maintaining suitable detection. These results raise at least two interesting questions for future work: 1) Is a combined architecture necessary for reducing false positives for the NRW problem? Rather, could a combined architecture be built into a single classifier that performs as well, or better than the NRW-CRITIC? This is an interesting question, especially if we expand the FVT algorithm to include confidence measures or quality scores as part of the output. Another rationale for using multiple classifiers is to combine several outputs using data provided by an online system. This is an especially intriguing possibility since the acoustic environment is ever changing as a result of variability in ambient noise, climate and shipping to name a few. 2) Can a multiple classifier system provide an online training process to optimize for these various effectors? More specifically, instead of using a fusion architecture, as presented here, would a better solution be a combinational expert system designed to select individual classifiers that find specific patterns within the general population of incoming acoustic objects? For example, build classifiers to identify specific bioacoustic behaviors (e.g., NRW contact calls, humpback song syllables) that cluster into isolated regions in the animal's acoustic feature space. Therefore, one could train a suite of classifiers, enabling and disabling different parameters (e.g., through weighting different features or geographic data sets, etc.), to meet current optimal application-specific conditions.

Although we consider the results presented here experimental in nature, they are promising. Future extensions

of using multi-classifier systems for NRW acoustic detection will expand by optimizing for online methodologies and incorporate more dynamic combinational methods.

ACKNOWLEDGEMENT

Research was supported through funding provided by the National Oceanographic Partnership Program, Excelerate Energy, Suez Energy, and the Massachusetts Division of Marine Fisheries.

We would like to thank I. Biedron, T. Calupca, W. Kroska, and C. Tremblay for collecting field data; M. Fowler, C. McCarthy, J. Morano, C. Muirhead, A. Murray, D. Ponirakis, E. Rowland, and A. Warde for providing verified libraries of right whale sounds; and S. Dedrick, C. Diamond, B. Estabrook, J. Evans-Wilent, B. Howard, C. McCarthy, J. Morano, C. Muirhead, A. Murray, D. Nelson, M. Pitzrick, D. Ponirakis, B. Roberts, E. Rowland, D. Salisbury, K. Conklin, J. Tielens, A. Warde and K. Wurzell for analysis of the Massachusetts Bay acoustic data. Also like to thank R. Charif, E. Spaulding and A. Warde for thoughtful input.

REFERENCES

[1] C. W. Clark, D. Gillespie, D. P. Nowacek, and S. E. Parks, "Listening to their world: acoustics for monitoring and protecting right whales in an urbanized ocean," in *The Urban Whale: North Atlantic Right Whales at the Crossroads*, S. D. Kraus and R. M. Rolland, Eds. Cambridge, MA: Harvard University Press, 2007, pp. 333-357.

[2] C. W. Clark, M. W. Brown, and P. J. Corkeron, "Visual and acoustic surveys for North Atlantic right whales, *Eubalaena glacialis*, in Cape Cod Bay, Massachusetts, 2001-2005: management implications," *Mar. Mamm. Sci.*, In press.

[3] S. E. Parks and C. W. Clark, "Acoustic communication: social sounds and the potential impacts of noise," in *The Urban Whale: North Atlantic Right Whales at the Crossroads*, S. D. Kraus and R. M. Rolland, Eds. Cambridge, MA: Harvard University Press, 2007, pp. 310-332.

[4] M. A. MacDonald and S. E. Moore, "Calls recorded from North Pacific right whales (*Eubalaena japonica*) in the eastern Bering Sea," *J. Cetac. Res. Manage.*, vol. 4, pp. 261-266, 2002.

[5] D. K. Mellinger, S. L. Nieukirk, H. Matsumoto, S. L. Heimlich, R. P. Dziak, J. Haxel, M. Fowler, C. Meinig, and H. V. Miller, "Seasonal occurrence of North Atlantic right whale (*Eubalaena glacialis*) vocalizations at two sites on the Scotian Shelf," *Mar. Mamm. Sci.*, vol. 23, pp. 856-867, 2007.

[6] P. J. Dugan, A. N. Rice, I. R. Urazghildiiev, and C. W. Clark, "North Atlantic right whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms," *Proc. IEEE: LISAT 2010*, 2010.

[7] I. R. Urazghildiiev, C. W. Clark, T. P. Krein, and S. E. Parks, "Detection and recognition of North Atlantic right whale contact calls in the presence of ambient noise," *IEEE J. Ocean. Eng.*, vol. 34, pp. 358-368, 2009.

[8] D. Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Can. Acoust.*, vol. 32, pp. 39-47, 2004.

[9] D. K. Mellinger and C. W. Clark, "A method for filtering bioacoustic transients by spectrogram image convolution," *Proceedings of the IEEE: OCEANS '93*, vol. 3, pp. 122-127, 1993.

[10] D. K. Mellinger, "A comparison of methods for detecting right whale calls," *Can. Acoust.*, vol. 32, pp. 55-65, 2004.

[11] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

[12] C. Berenstein, L. N. Kanal, and D. Lavine, "Consensus rules," in *Uncertainty in Artificial Intelligence*, L. N. Kanal and J. F. Lemmer, Eds. New York: Elsevier Science Publishers, 1986, pp. 27-32.

[13] T. K. Ho, "Multiple classifier combination: lessons and next steps," in *Hybrid Methods in Pattern Recognition*, H. Bunke and A. Kandel, Eds. Singapore: World Scientific Publishing Co., 2002, pp. 171-198.

[14] L. Breiman, "Bagging predictors," *Mach. Learning*, vol. 24, pp. 123-140, 1996.

[15] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119-139, 1997.

[16] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, pp. 5-32, 2001.

[17] L. Breiman, "Pasting small votes for classification in large databases and on-line," *Mach. Learning*, vol. 36, pp. 85-103, 1999.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd ed.* New York: John Wiley and Sons, Inc., 2001.

[19] C. Pilcher and A. Khotanzad, "Nonlinear classifier combination for a maritime target recognition task," *Proc. IEEE: Radar Conference*, pp. 1-5, 2009.

[20] P. J. Dugan, L. K. Lewis, R. D. Paradis, and D. A. Tillotson, "Cognitive Arbitration System," U. S. Patent 7,340,443, March 4, 2008.

[21] A. Altmann, M. Rosen-Zvi, M. Prosperi, E. Aharoni, H. Neuvirth, E. Schülter, J. Büch, D. Struck, Y. Peres, F. Incardona, A. Sönnerborg, R. Kaiser, M. Zazzi, and T. Lengauer, "Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy," *PLoS ONE*, vol. 3, pp. e3470, 2008.

[22] G. J. Klir and M. J. Wierman, *Uncertainty-Based Information: Elements of Generalized Information Theory*. Heidelberg: Physica-Verlag, 1998.

[23] R. J. Urick, *Principles of Underwater Sound for Engineers*. New York: McGraw-Hill Book Company, 1967.