Fisheries and Oceans Canada | Pêches et Océans Canada
Ecosystems and Oceans Science | Sciences des écosystèmes et des océans

Canadian Science Advisory Secretariat (CSAS)

Research Document 2023/050

Newfoundland and Labrador Region

A preliminary review of the efficacy of several acoustic autodetection algorithms to identify North Atlantic right whale calls, and recommendations for next steps to further assess and optimize these algorithms

Jack W. Lawson[1]

[1]Fisheries and Oceans Canada
Northwest Atlantic Fisheries Centre
80 E. White Hills Rd., P.O. Box 5667
St. John's NL A1C 5X1

July 2023

Canada

Foreword

This series documents the scientific basis for the evaluation of aquatic resources and ecosystems in Canada. As such, it addresses the issues of the day in the time frames required and the documents it contains are not intended as definitive statements on the subjects addressed but rather as progress reports on ongoing investigations.

Published by:

Fisheries and Oceans Canada
Canadian Science Advisory Secretariat
200 Kent Street
Ottawa ON K1A 0E6

http://www.dfo-mpo.gc.ca/csas-sccs/
csas-sccs@dfo-mpo.gc.ca

ISSN 1919-5044
ISBN 978-0-660-49012-0 Cat. No. Fs70-5/2023-050E-PDF

Correct citation for this publication:

Lawson, J. 2023. A preliminary review of the efficacy of several acoustic autodetection algorithms to identify North Atlantic right whale calls, and recommendations for next steps to further assess and optimize these algorithms. DFO Can. Sci. Advis. Sec. Res. Doc. 2023/050. iv + 17 p.

Aussi disponible en français :

Lawson, J. 2023. Examen préliminaire de l'efficacité de plusieurs algorithmes de détection acoustique automatique des vocalises de baleine noire de l'Atlantique Nord, et recommandations par rapport aux prochaines étapes relatives à l'évaluation et à l'optimisation de ces algorithmes. Secr. can. des avis sci. du MPO. Doc. de rech. 2023/050. iv + 20 p.

TABLE OF CONTENTS

ABSTRACT iv
INTRODUCTION 1
METHODS 1
AUTOMATED ACOUSTIC NARW UPCALL DETECTORS 1
TERMINOLOGY TO DESCRIBE AND COMPARE DETECTORS 1
RESULTS 2
2010 COMPARISON OF EARLY MACHINE LEARNING NARW UPCALL DETECTORS 2
2013 COMPARISON OF NARW UPCALL AUTOMATED DETECTORS 3
BAUMGARTNER'S LFDCS DETECTOR 3
JASCO'S SPECTROPLOTTER DETECTOR 4
2017 COMPARISON OF LFDCS AND JASCO NARW UPCALL DETECTORS 4
DISCUSSION 5
DCS PERFORM "ACCEPTABLY" WITH NARW UPCALL DATA 5
UPCALL CONTEXT IS AN IMPORTANT FACTOR IN DCS PERFORMANCE 6
GOOD UPCALL SAMPLES AND MANUALLY-VALIDATED DATASETS ARE CRITICAL TO TEST DCS 6
FUTURE RESEARCH 7
DCS COMPARISON/OPTIMIZATION – MANUAL VALIDATION 7
DCS COMPARISON/OPTIMIZATION – DETECTOR PROCESSING 7
DCS COMPARISON/OPTIMIZATION – INTERNATIONAL WORKSHOP 8
CONCLUSIONS 8
ACKNOWLEDGMENTS 9
REFERENCES CITED 9
TABLES AND FIGURES 11

ABSTRACT

Automated detection and classification of the vocalizations of North Atlantic right whale (NARW) and other marine mammals is a highly desirable goal for researchers and managers seeking to monitor areas for whale presence as the basis to implement mitigation measures. Such automated acoustic processing is particularly important for real-time monitoring approaches where there are large-scale acoustic data inputs.

All of the Detection and Classification Systems (DCSs) used by Fisheries and Oceans Canada (DFO) are expected to perform similarly well, given the metric (e.g., hours with calls/day) used to present NARW occurrence time-series. Previously, this was demonstrated by comparing performances of a variety of detectors during studies in 2004, 2013, and 2017. Spectroplotter (a commercial programme) and Low-Frequency Detection and Classification System (LFDCS), which are the two systems that have been used to analyse acoustic data in Newfoundland and Labrador (NL) and Maritimes regions, perform well; although in one small study the LFDCS detected more actual NARW upcalls than Spectroplotter, but also generated more false positives.

DCS performance is influenced by multiple factors, including the ambient noise levels relative to the characteristics of the NARW upcalls, the location of the hydrophone, the characteristics of the recorder instrumentation, software settings and thresholds, and other contextual features, such as the presence of other species. The next generation of DCSs will incorporate context into their logic (e.g., presence of other marine mammals or abiotic sound sources and signal-to-noise ratio [SNR]).

Algorithm comparisons are less crucial in the historic NARW analyses as the metrics in which the present detection results are presented at a large enough scale (“has there been NARW detected at this recorder location today?”) that slight differences in algorithm performance would be subsumed in the amalgamation and summation process.

At smaller spatial and temporal sampling scales, differences in algorithm performance become more apparent. Thorough testing of the different DCSs being used in Atlantic Canada would require a series of manually validated acoustic datasets from a representative set of locations, time frames, seasons, and recording hardware. Such a DCS comparison would be a useful activity but would require agreed upon performance metrices and thresholds for the DCS.

INTRODUCTION

Passive acoustic monitoring (PAM) provides a powerful tool to detect and identify marine mammals underwater, and has been used in many studies (Baumgartner et al. 2018; Mellinger et al. 2007a; Van Parijs et al. 2009; Verfuß et al. 2007). Unlike visual survey methods, acoustic monitoring can collect data continuously, in remote locations, and in light and weather conditions that would limit visual detection. However, PAM can produce a great deal of data and these data are a complex mixture of the sounds of target species of interest, other species, anthropogenic activities, and environmental processes. Therefore, there is a need for a method to analyse these data quickly.

Fortunately, many marine species produce sounds that are unique. For instance, North Atlantic right whales (*Eubalaena glacialis*)(NARW) generate a distinctive vocal repertoire, including a diagnostic 50-300 Hz upsweeping contact call known as the “upcall” (Figure 1a)(Mellinger et al. 2007b). To process large amounts of acoustic recording data to detect species-specific sounds, an automated DCS requires significantly less time than manual searching by a trained expert, assuming the DCS is acceptably accurate. In particular, accurate automated detection and classification of the vocalizations of NARWs and other marine mammals is a highly desirable goal for researchers and managers seeking to monitor areas for whale presence as the basis to implement mitigation measures.

A variety of DCS approaches have been developed for marine mammal sounds, including low-frequency detectors for the NARW (see Davis et al. 2017)(Table 1). Such detectors are particularly relevant for DFO’s monitoring and mitigation efforts for this species in Atlantic Canada as a variety of detection/classification algorithms are used currently; however, these detectors seem to differ in their efficacy and accuracy, so it is worth comparing them to ensure that DFO uses a system with the best accuracy and analytical speed.

In addition to reviewing previous DCSs, I provide recommendations for the Department as to how we could further compare these detectors and how to better allocate resources for optimizing these systems.

METHODS

AUTOMATED ACOUSTIC NARW UPCALL DETECTORS

In recent years, a variety of automated NARW detector-classifiers have been developed (Baumgartner and Mussoline 2011; Dugan et al. 2010a; Gillespie 2004; Mellinger 2004; Mouy et al. 2009; Simard and Roy 2008; Urazghildiiev and Clark 2006), including combinations of detectors (Dugan et al. 2010b). Many of these detectors have been targeted to tonal sounds, such as NARW upcalls, rather than broadband signals such as NARW “gunshot” sounds which they also produce. Some DCSs classify signals based on direct measures of features such as signal frequency and duration (e.g., JASCO’s algorithm [Spectroplotter] [Figure 1b]), whereas others classify signals based on measures derived from basic signal features (e.g., the LFDCS algorithm; Baumgartner and Mussoline 2011).

TERMINOLOGY TO DESCRIBE AND COMPARE DETECTORS

Terminology and metrics are useful when quantifying differences in NARW call detector performance and manual validation (the process whereby a highly-trained analyst reviews acoustic recordings aurally and visually to classify signals they contain):

True positives: calls classified as NARW calls that are actually NARW calls as determined by manual validation.

False positives: calls classified as NARW calls that are not NARW calls as determined by manual validation.

True negatives: calls classified as not NARW calls that are actually not NARW calls as determined by manual validation.

False negatives: calls classified as not NARW calls that are actually NARW calls as determined by manual validation.

Recall: (also known as sensitivity): ability of a classification algorithm (detector) to identify all actual NARW calls.

Precision: ability of a classification algorithm to return only correct NARW call detections.

While recall expresses the ability to detect all NARW calls in a dataset, precision expresses the proportion of the calls the detection algorithm correctly classifies as “right whale” that actually were NARW:

$$\text{Recall} = \frac{\text{True Upcalls}}{\text{True Upcalls + False Negatives}} \qquad \text{Precision} = \frac{\text{True Upcalls}}{\text{True Upcalls + False Positives}}$$

F1 score: a single metric that combines recall and precision using the harmonic mean. The F1 score is a better measure to obtain a balance between precision and recall, when there is an imbalanced class distribution. An imbalanced class distribution is a scenario where the number of incidents of one call type (such as NARW) is much lower than those belonging to other call types (such as humpback whales, Megaptera novaeangliae).

$$F1 = 2 \times \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

Receiver operating characteristic (ROC) curve: plots the true positive rate versus the false positive rate as a function of the classification algorithm’s threshold for correctly classifying a NARW call; the ROC curve displays how the recall versus precision relationship changes as the threshold for identifying a true NARW call is changed. Altering the threshold can achieve an acceptable precision versus recall balance.

Area under the curve (AUC): metric to calculate the overall performance of a classification algorithm based on the area under the ROC curve.

RESULTS

2010 COMPARISON OF EARLY MACHINE LEARNING NARW UPCALL DETECTORS

Dugan et al. (2010b) compared three early computer-based approaches for recognizing NARW upcalls. The performance of two newer approaches (machine learning algorithms based either on artificial neural networks [NET] or classification and regression tree classifiers [CART]) were compared with an earlier system that employed a multi-stage feature vector testing (FVT)

approach. A large sample of underwater noise and NARW upcall events recorded from Cape Cod Bay and the Great South Channel was used. Of the three classifiers, CART had the highest precision of 86% with the same false positive rates as the NET algorithm. The FVT algorithm was not as good at classifying NARW upcalls as the newer methods (lower recall) but had very low false positive rates.

2013 COMPARISON OF NARW UPCALL AUTOMATED DETECTORS

Gillespie (2018 pers. comm.) provided unpublished summary information from the 6th International Workshop on Detection, Classification, Localization, and Density Estimation (DCLDE) of Marine Mammals using Passive Acoustics in St. Andrews, Scotland in 2013 ("Comparison of Right Whale Detector Results - DCLDE6"). Meeting participants were provided with an acoustic dataset containing a number of NARW and other species' calls which had been manually validated by an experienced analyst beforehand. Subsequently, this test dataset was found to be flawed as it was reviewed by only a single validator and contained undetected NARW calls (false negatives).

The seven DCSs tested included multiple Cornell detectors, multiple detectors by Gillespie, the Mouy (i.e., early JASCO; Mouy et al. 2009) kernel-based detector, and the continuous region analysis (CRA) detector (a neural network algorithm). Baumgartner's LFDCS detector was not tested at that meeting, nor was the latest Simard algorithm.

Overall, NARW "gunshots" were a poor candidate for automatic detection by the systems tested; there had been no large improvement in performance of upcall detectors since previous testing, all detectors had similar false positive rates for a given efficiency (Figure 2); and detector performance was heavily dependent on the goal of its application. Some detectors had better recall at very low false positive rates, while others had better recall at higher false positive rates (Figure 2). Performance of these detectors was influenced by several factors:

- SNR, which is related to ambient noise levels relative to the characteristics of the target calls
- SNR is also dependent on the location of the hydrophone, and the characteristics of recorder instrumentation
- Other contextual features, such as the presence of other calling species (particularly humpback whales in the NW Atlantic study area)

Meeting participants concluded that while the DCSs produced slightly differing detection and classification results, they all proved to achieve about 70% recall (Figure 2).

BAUMGARTNER'S LFDCS DETECTOR

The LFDCS is a sound pitch contour-based detector (Baumgartner and Mussoline 2011) that is used to search for the species-specific calls in PAM data from fixed acoustic recorders and in customized autodetection software/hardware packages used in mobile gliders (Baumgartner et al. 2013). The LFDCS characterizes temporal variation of dominant call frequencies via pitch-tracking, and classifies NARW calls based on attributes of the resulting pitch tracks using quadratic discriminant function analysis (Baumgartner and Mussoline 2011). The software is somewhat "generalized" in that the underlying species libraries and settings on which its performance is based can be adjusted by the user. DFO Maritimes has transitioned to using Baumgartner's LFDCS algorithms to determine presence of NARW in their acoustic recorder data; for instance, the Maritimes' 2006–14 datasets have been analysed using this algorithm and included in Davis et al. (2017), as were 18 other groups' including NOAA's.

Baumgartner and Mussoline (2011) examined the LFDCS detector accuracy by using it to analyse acoustic recordings collected in the Gulf of Maine during spring of 2006 and 2007. They found that the LFDCS algorithm was able to compensate for persistent narrowband and transient broadband noise in the recordings, and its accuracy was similar to that of a human analyst. That is, variability in differences between the DCS and an analyst was similar to that between independent analysts, and temporal variability in call rates was similar among the LFDCS and several analysts.

Similarly, Davis et al. (2017) determined that the rate of missed upcall detections using LFDCS was low (25%), and while this rate depended on the characteristics of individual deployments, such as ambient and anthropogenic background noise at the site, the resulting detections still provided a satisfactory indication of the broad-scale distribution of NARW.

JASCO'S SPECTROPLOTTER DETECTOR

In the past, DFO Maritimes and DFO NL Regions used JASCO's multispecies detector (Spectroplotter) to determine the presence of NARW upcalls in their acoustic datasets. This algorithm is a proprietary kernel-based detector developed by JASCO Applied Sciences detailed below. As described in a number of JASCO reports, Spectroplotter's NARW multiple upcall exemplars were sampled from Cornell's fixed platform dataset, which was deployed off the coast of Massachusetts in 2000, 2008, and 2009. For comparison, the NARW upcall described in the LFDCS call library is based on 254 upcall exemplars from the Southwestern and Central Gulf of Maine in 2005 and 2009 (Martin et al. 2014). Since these exemplars were obtained in similar locations and dates, JASCO concluded that any differences in detector performance should not be attributed to differences in the underlying call libraries.

JASCO (Delarue et al. 2018; Martin et al. 2014) outlined work carried out on NARW call detection for some of the same datasets on which DFO has conducted their LFDCS analysis. JASCO ran Spectroplotter through a large dataset and determined it performed with low precision due to interference by humpback tonal moans; this implies that the precision of this system will vary seasonally with the local abundance of humpbacks. JASCO then had to rely on manual validation of some datasets; for NARW they manually reviewed one min of every 11-min sound file (corresponding to three mins per hr, or 5% of the recorded data)(see page 76 in Delarue et al. 2018). From the report: *To ensure an accurate representation of right whale occurrence, we performed additional manual review of data recorded where and when right whale presence was expected, based on the current knowledge of the species' seasonal distribution* (Delarue et al. 2018). JASCO, recently re-tuned Spectroplotter to increase the probability of NARW upcall detection and reduce the number of false positives, but the degree of improvement is not yet published.

2017 COMPARISON OF LFDCS AND JASCO NARW UPCALL DETECTORS

Although not a comprehensive comparison of DCSs, Moore (2017) undertook a test of the LFDCS and Spectroplotter DCS with NARW upcall data from Roseway and Emerald Basin. Acoustic recordings collected by both Slocum gliders and fixed PAM systems in the summer and fall of 2015, and a subsample of 7% of all recordings were validated manually by two experienced acousticians. The manual analysis results were compared to the two automated detector results at three temporal scales.

Spectroplotter uses a binary spectrogram of time frequency bins above an empirical threshold, rather than a smoothed spectrogram (subtraction of a long-duration mean) which is used in the LFDCS (Baumgartner and Mussoline 2011; Martin et al. 2014). While Spectroplotter processes the spectrogram for time-frequency objects (or events) by selecting for contiguous temporal bins

that are above a threshold amplitude level, the LFDCS identifies the beginning of a tonal sound above an empirical threshold (hereafter pitchtrack) and employs forward pitch-tracking and backward pitch-tracking to formalize a pitchtrack (Baumgartner and Mussoline 2011; Martin et al. 2014). Attributes selected for in Spectroplotter include time (date), duration (s), minimum and maximum frequency (Hz), sweep rate (slope), tonality, and bandwidth (Hz) (Figure 1b) (Martin et al. 2014). In comparison, the attributes selected for in the LFDCS include an average frequency (Hz), frequency variation, time variation, and slope (Baumgartner and Mussoline 2011). The attributes selected for in both detectors appear similar; however, while the kernel-based detector extracts call attributes directly from the spectrogram, the contour-based detector first estimates a pitchtrack using the time variation of the fundamental frequency, from which attributes of the sound are extracted (Baumgartner and Mussoline 2011).

For this small dataset, Moore found that the LFDCS detected more true NARW upcalls (i.e., true positives) at all time scales relative to the Spectroplotter detector, but also generated a higher number of false positives. There was no significant difference in percentage of true positive detections in Emerald Basin between the two DCSs on per recording file (termed snippet), hourly, and daily sampling scales (Table 2; Figure 3a) (Moore 2017). In Roseway Basin, the percentage of true positive NARW detections on hourly and daily scales between detectors was not different, whereas the LFDCS found more true positive detections on a per snippet basis (Table 3; Figure 3b). Moore concluded that the Spectroplotter and LFDCS results were comparable on hourly and daily scales.

Both DCSs experienced reduced accuracy in Emerald relative to Roseway Basin (Tables 2 and 3), but there appeared to be a trade-off between maximizing detections and minimizing false negatives. While the LFDCS detected between 9 and 17% more days with true upcalls, Spectroplotter minimized false detections by approximately 10%. An analysis of a dataset which yields a high number of false negatives, such as the one produced by LFDCS, will require more intensive validation, whereas a dataset with less detections may underestimate NARW acoustic presence. If the objective is to detect presence or absence of NARW, using a detector that minimizes false detections (such as the Spectroplotter) would be the most efficient. If the objective is to observe seasonal habitat use, using a detector that maximizes true upcall classifications (such as the LFDCS) would be more efficient.

Since this comparison employed data collected over a short time period, it did not address call detection issues over the broader scale of varying background noise levels and how those impact detections; one of the comparisons also didn't compare individual detections but rather examined false negative/false positives rates on a per snippet basis. The latter metric is different than false alarm versus missed call rate when you use individual calls and detections, but nonetheless did provide some information on how the two detectors compare to each other on the same manually validated dataset on a per file basis.

DISCUSSION

DCS PERFORM "ACCEPTABLY" WITH NARW UPCALL DATA

Based on several studies, most of the DCSs described in this paper perform acceptably in that they process acoustic datasets much more quickly than a manual validator, and their true NARW classifications and error rates are acceptable – particularly if the questions they address are broad ("have NARW been detected near the recorder today?"). Earlier work by Mellinger and Clark (2000) showed that the performance of spectrogram correlation (such as is used by LFDCS and Spectroplotter) compared favourably to three other methods that had been used for automatic call recognition (matched filters, neural networks, and hidden Markov models). With

further development, (Mellinger 2004), again compared upcall detection by spectrogram correlation and a neural network, and this time the neural network performed better (Figure 5). As expected, performance of all methods generally improved with increasing SNR.

Subsequent testing of a variety of newer DCSs in 2013 and 2017 further determined that these DCSs can perform with a degree of precision and recall that is sufficient to answer questions based on presence and absence of calling NARW (e.g., Figure 2). Moore (2017) concluded that the LFDCS detected more true NARW upcalls at all time scales relative to Spectroplotter, but also generated a higher number of false positives; the only DCS difference in this study was Spectroplotter's higher F1 score in the per snippet scale of Roseway Basin data (Tables 2 and 3, Figure 3).

Depending on the thresholds set (such as for the level of acceptable false negative or positive detections of NARW upcalls), and the amount of manual validation needed following the DCS processing, the LFDCS and Spectroplotter DCS can be used to automate the analyses of copious acoustic recordings, and thereby answer questions related to NARW presence.

UPCALL CONTEXT IS AN IMPORTANT FACTOR IN DCS PERFORMANCE

There are a variety of contextual factors which have important impacts on DCS upcall classification performance. The most important factors are the relative strength and clarity of an upcall record (SNR, Mellinger 2004), and whether there are similar-sounding species (e.g., humpback whales) calling nearby. Another factor may be calling rate. Moore (2017) found that both the LFDCS and Spectroplotter detectors detected 23–31% fewer true upcalls in Emerald than Roseway Basin (Tables 2 and 3). In this case, Moore thought that NARW in Roseway Basin might have called at a higher rate and thereby offered an increased opportunity to detect at least one call within a snippet, hour, or day sample period.

Marine animal behaviour, including their acoustic behaviour, varies between seasons and geographical areas, driven by varying life history parameters and experiences (Van Parijs et al. 2009). For example, analyses of NARW call types across the three habitat areas display differences in the calling behaviour of these whales in the spring versus the summer habitat areas (Figure 4) (Van Parijs et al. 2009). It also suggests that other call types, such as gunshots, may be more effective to detect NARW presence in some areas.

When using acoustic-based approaches to implement management and mitigation approaches for NARW it is therefore important that we account for how their call repertoire varies seasonally or geographically, and how this and other factors outside the mechanics of the DCS will influence the ability of the DCS to detect and classify NARW sounds.

GOOD UPCALL SAMPLES AND MANUALLY-VALIDATED DATASETS ARE CRITICAL TO TEST DCS

PAM is most useful within the context of the acoustic behavioural ecology of NARW and applied in a regionally and seasonally appropriate context. In order to improve PAM, more information is needed on NARW individual, group, and population sound usage and sources of variability in vocalization.

The LFDCS and Spectroplotter DCS rely on underlying libraries of upcall exemplars to seek matches in the analysed datasets. Ideally these sample calls will have been collected with good signal-to-noise conditions, and for areas and seasons that match as closely as possible the area of the analysed datasets.

Differences in these particular DCSs will not impact the ability to make general conclusions about NARW presence in Atlantic Canada, although we did not evaluate the Simard DCS.

Based on differences in the approaches employed in the acoustic studies presented herein, we cannot straightforwardly compare recording sites, but trends in minimum presence observed from each study will provide useful information. A plan for a more comprehensive detector comparison, including the algorithm used in the Gulf of St. Lawrence, would be relevant for the topic of limitations and strengths of different DCS technology and could be highlighted as a step that should be taken.

FUTURE RESEARCH

Recognising that detection and classification of marine mammals using PAM has become a critical tool for DFO, there is a need for a zonal/national consultation on the way forward for dealing with a number of issues with PAM data in terms of processing speed and classification accuracy:

- Develop similar-performing detection/classification tools for multispecies considerations in different environments (optimization)
- Recommendations for how to advance the detector development process into a national approach for DFO
- Sanction guidelines for best practice for acoustic analysis and reporting
- Address optimal data format and large-scale acoustic data storage issues

DCS COMPARISON/OPTIMIZATION – MANUAL VALIDATION

A robust detector comparison/optimization would require a significant amount of work, including challenging the DCS with multiple datasets collected using different recording systems, with varying background noise levels, and from different geographic areas, to validate calls. These datasets would be manually validated by at least two experienced acousticians[1]. For this DCS comparison we recommend using at least five 10-hour recorded underwater samples:

1. digital recording with many NARW calls,
2. digital recording with few NARW calls,
3. digital recording with many NARW calls and a high level of ambient noise,
4. digital recording with few NARW calls and a high level of ambient noise, and
5. digital recording with a similar ratio of NARW and humpback whale calls.

Prior to the start of the manual validation process, an analysis process would be established which would include rules for annotating targets as definite NARW upcalls; while many cases are clear, there are instances where it is unsure if the call is actually a NARW or something else and analysts can differ in how they categorize this. There should be scientific consensus on the steps in the validation process.

DCS COMPARISON/OPTIMIZATION – DETECTOR PROCESSING

As for the manual validation process, clear steps must be undertaken for DCS processing of the trial sound samples. For instance, the point selected for the setting of each algorithm on the

[1] These recommendations could be used to write a Statement of Work for manual validation of these datasets. To examine acoustic analyst bias there should be at least two trained analysts to process a subset of the recordings and compare their results.

recall versus precision curve must be documented; each DCS algorithm will have to meet a precise point on that curve (target), depending on the objective of either not missing any target species call (at the cost of accepting higher rates of false positives), or limiting the false positive rate. All algorithms would need to meet a given level of accuracy, and in this case, we should consider this to be less than a 10% false positive rate.

Further, an optimal DCS test or comparison will be based on files with manually-validated time-tagged NARW contact call detections, and summaries of hours with NARW contact calls present (i.e., after elimination of confounding identical humpback contact call-like sounds); comparing these will provide evidence of detector efficacy.

With advances in machine learning capabilities and processing power of individual computers, it is worth reconsidering neural network DCSs that previously performed as well or better than spectrogram-matching systems in some trials (Figure 5; Mellinger 2004).

DCS COMPARISON/OPTIMIZATION – INTERNATIONAL WORKSHOP

A workshop, perhaps through collaboration with Meridian, to review output from this DCS comparison and optimization process would improve our knowledge on this topic. Relevant work proposed by Meridian over a longer period could contribute to a more complete detector comparison study (such as including other types of NARW sounds, and broadening the DCS to additional species as Baumgartner (Baumgartner and Mussoline 2011) has done with the LFDCS. In addition, a workshop could be a forum for further improvements in the DCS approach for DFO by strengthening working relationships with collaborators.

CONCLUSIONS

- Automated detection and classification of the vocalizations of NARW and other marine mammals is a highly desirable goal for researchers and managers seeking to monitor areas for whale presence as the basis for implementing management measures.
- Such automated acoustic processing is particularly important for real-time monitoring approaches where there are large-scale acoustic data inputs.
- A DCS assessment meeting in 2013 demonstrated similar performances of seven detectors (~70% correct), and this was influenced by several factors:
    - SNR, which is related to ambient noise levels relative to the characteristics of the target calls.
    - SNR is also dependent on the location of the hydrophone, and the characteristics of the recorder instrumentation.
    - Other contextual features, such as the presence of other species (particularly humpback whales in the NW Atlantic study area).
- Spectroplotter and LFDCS (the algorithms that have been used to analyze acoustic recordings in Maritime and NL Regions) perform well, although in one study the LFDCS detected more true NARW upcalls relative to Spectroplotter, but also generated more false positives.
- All of the detectors used are expected to perform similarly well, given the metric (e.g., hours with calls/day) used to present the NARW occurrence time-series; several DCSs demonstrated similar performance by the DCLDE 2013 meeting and several 2017 studies. If false positive rates and missed call rates are presented in PAM Research Documents, then these accuracy measures will give an idea of how the detectors compare without doing a direct comparison.

- Thorough testing of different DCSs would require a series of manually validated acoustic datasets from a representative set of locations, time frames, seasons, and recording hardware. Such a DCS comparison, in particular comparing performance of the LFDCS, JASCO (Spectroplotter), and Simard detectors, plus potentially others, would be a useful activity but require agreed performance metrics and thresholds for the DCS.
- Algorithm comparisons are less crucial in historic NARW analyses because the metrics in which the current detection results are presented are at a large enough temporal scale ("has there been a NARW detected at this recorder location today"?) that slight differences in algorithm performance would be subsumed in the detection data amalgamation and summation process. With the same data collected in similar locations and contexts, there should be less concern about the influence of these underlying factors. At smaller spatial and temporal sampling scales, differences in algorithm performance become more significant.
- The next generation of some U.S. DCSs will incorporate context into their logic (e.g., presence of other marine mammals or abiotic sound sources and SNR). For instance, currently the performance of some DCSs varies by location and season, and adjusting precision or including context-specific detection thresholds may correct for this.

ACKNOWLEDGMENTS

I thank Drs. H. Moors-Murphy (DFO), Y. Simard (DFO), and S. Van Parijis (NOAA) for providing important feedback on an outline for this Research Document. Dr. Van Parijis obtained the 2013 unpublished DCS comparison meeting results from Dr. D. Gillespie.

REFERENCES CITED

Baumgartner, M.F., and Mussoline, S.E. 2011. A generalized baleen whale call detection and classification system. J. Acoust. Soc. Am. 129(5): 2889–2902.

Baumgartner, M.F., Fratantoni, D.M., Hurst, T.P., Brown, M.W., Cole, T.V.N., Van Parijs, S.M., and Johnson, M. 2013. Real-time reporting of baleen whale passive acoustic detections from ocean gliders. J. Acoust. Soc. Am. 134(3): 1814–1823.

Baumgartner, M.F., Stafford, K.M., and Latha, G. 2018. Near Real-Time Underwater Passive Acoustic Monitoring of Natural and Anthropogenic Sounds. In Observing the Oceans in Real Time. Edited by R. Venkatesan, A. Tandon, E.A. D'Asaro, and M.A. Atmanand. Springer International Publishing, Cham. 203–226.

Davis, G.E., Baumgartner, M.F., Bonnell, J.M., Bell, J., Berchok, C., Bort Thornton, J., Brault, S., Buchanan, G., Charif, R.A., Cholewiak, D., Clark, C.W., Corkeron, P., Delarue, J., Dudzinski, K., Hatch, L., Hildebrand, J., Hodge, L., Klinck, H., Kraus, S., Martin, B., Mellinger, D.K., Moors-Murphy, H., Nieukirk, S., Nowacek, D.P., Parks, S., Read, A.J., Rice, A.N., Risch, D., Širović, A., Soldevilla, M., Stafford, K., Stanistreet, J.E., Summers, E., Todd, S., Warde, A., and Van Parijs, S.M. 2017. Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (*Eubalaena glacialis*) from 2004 to 2014. Sci. Rep. 7(1): 13460.

Delarue, J., Kowarski, K.A., Maxner, E.E., MacDonnell, J.T., and Martin, S.B. 2018. Acoustic Monitoring Along Canada's East Coast: August 2015 to July 2017. Document Number 01279. Environmental Studies Research Funds Report Number 215, Version 1.0. Tech. Rep. by JASCO Applied Sciences for Environmental Studies Research Fund, Dartmouth, NS, Canada. 120 pp + appendices.

Dugan, P.J., Rice, A.N., Urazghildiiev, I.R., and Clark, C.W. 2010a. North Atlantic Right Whale acoustic signal processing: Part I. Comparison of machine learning recognition algorithms. In 2010 IEEE Long Island Systems, Applications and Technology Conference, Farmingdale, NY. 1–6.

Dugan, P.J., Rice, A.N., Urazghildiiev, I.R., and Clark, C.W. 2010b. North Atlantic right whale acoustic signal processing: Part II. improved decision architecture for auto-detection using multi-classifier combination methodology. In 2010 IEEE Long Island Systems, Applications and Technology Conference. Farmingdale, NY, USA. 1–6.

Gillespie, D. 2004. Detection and classification of right whale calls using an 'edge' detector operating on a smooth spectrogram. Can. Acoust. 32(2): 39–47.

Martin, B., Kowarski, K., Mouy, X., and Moors-Murphy, H. 2014. Recording and identification of marine mammal vocalizations on the Scotian Shelf and slope. In Oceans 2014 Conference, St. John's, NL. 1–6.

Mellinger, D. 2004. A comparison of methods for detecting right whale calls. Can. Acoust. 32(2): 55–65.

Mellinger, D.K., and Clark, C.W. 2000. Recognizing transient low-frequency whale sounds by spectrogram correlation. J. Acoust. Soc. Am. 107(6): 3518–3529.

Mellinger, D.K., Nieukirk, S.L., Matsumoto, H., Heimlich, S.L., Dziak, R.P., Haxel, J., Fowler, M., Meinig, C., and Miller, H.V. 2007a. Seasonal occurrence of North Atlantic right whale (*Eubalaena glacialis*) vocalizations at two sites on the Scotian Shelf. Mar. Mamm. Sci. 23(4): 856–867.

Mellinger, D.K., Stafford, K.M., Moore, S.E., Dziak, R.P., and Matsumoto, H. 2007b. An Overview of Fixed Passive Acoustic Observation Methods for Cetaceans. Oceanogr. 20(4): 36–45.

Moore, D. 2017. Is Emerald Basin, Scotian Shelf, Canada, a North Atlantic right whale (*Eubalaena glacialis*) feeding habitat? Honours Thesis, Bachelor of Science in Marine Biology. Biology Dept., Dalhousie University. viii + 55 p.

Mouy, X., Bahoura, M., and Simard, Y. 2009. Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence. J. Acoust. Soc. Am. 126(6): 2918–2928.

Simard, Y., and Roy, N. 2008. Detection and localization of blue and fin whales from large-aperture autonomous hydrophone arrays: A case study from the St. Lawrence estuary. Can. Acoust. 36(1): 104–110.

Urazghildiiev, I.R., and Clark, C.W. 2006. Acoustic detection of North Atlantic right whale contact calls using the generalized likelihood ratio test. J. Acoust. Soc. Am. 120(4): 1956–1963.

Van Parijs, S.M., Clark, C.W., Sousa-Lima, R.S., Parks, S.E., Rankin, S., Risch, D., and Van Opzeeland, I.C. 2009. Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales. Mar. Ecol. Prog. Ser. 395: 21–36.

Verfuß, U.K., Honnef, C.G., Meding, A., Dähne, M., Mundry, R., and Benke, H. 2007. Geographical and seasonal variation of harbour porpoise (*Phocoena phocoena*) presence in the German Baltic Sea revealed by passive acoustic monitoring. J. Mar. Biol. Assoc. UK. 87(1): 165–176.

TABLES AND FIGURES

Table 1. Operating features of detection and classification algorithms considered in this document.

Detector	Target Species	Function	Notes
Gillespie DCS	NARW	Spectrogram feature comparison	First, the spectrogram is smoothed by convolving it with a Gaussian kernel and the 'outlines' of sounds are extracted using an edge detection algorithm. Second, parameters are used in a classification function in order to determine which sounds are from NARW
Cornell DCS	NARW	A variety, but one based on a generalized likelihood ratio	Detector of polynomial-phase signals with unknown amplitude and polynomial coefficients observed in the presence of locally stationary Gaussian noise. The closed form representation for a minimal sufficient statistic is derived and a realizable detection scheme is developed.
Mouy (JASCO) DCS	NARW + others	Spectrogram feature comparison	Multispecies DCS is based on underlying mammal sounds library
CRA detector	NARW	Neural network detector	Despite a relatively low level of false positive NARW upcall detections, CRA demonstrated the lowest precision of all detectors
LFDCS	NARW, Fin, Sei	Sound pitch contour-based detector with an underlying call library	Low frequency detection and classification system with adjustable settings and editable underlying call libraries
LSTM	NARW	-	DFO-BIO OPP colleague working with Wright on a new NARW call detection algorithm
Dugan et al. DCS	NARW	Classification and regression tree classifiers (CART)	The CART had higher true positive rate, and matched the NET algorithm for false positive rate
Dugan et al. DCS	NARW	Artificial neural networks (NET)	-

Table 2. Number of false negative, false positive, true positive, and true negative upcall detections identified per snippet, per hour, and per day for the LFDCS and Spectroplotter detectors in acoustic recordings from Emerald Basin. Percentage of true detections (number of true detections of all manually validated true calls), percentage of false detections (number of false detections out of all manually validated units without calls), recall, precision, and F1 score are also displayed (adapted from Moore 2017).

-	False Negative	False Positive	True Positive	True Negative	% True	% False	Recall	Precision	F1 Score
Per Snippet									
LFDCS	19	410	49	5,522	72%	7%	0.72	0.11	0.19
Spectroplotter	29	278	39	5,654	57 %	5%	0.57	0.12	0.20
Per Hour									
LFDCS	30	67	30	2,873	50%	2%	0.50	0.31	0.38
Spectroplotter	38	36	22	2,904	37%	1%	0.37	0.38	0.37
Per Day									
LFDCS	10	27	20	10	67%	28%	0.67	0.43	0.52
Spectroplotter	15	17	15	15	50%	18%	0.50	0.47	0.48

Table 3. Number of false negative, false positive, true positive, and true negative upcall detections identified per snippet, per hour, and per day for the LFDCS and Spectroplotter detectors in acoustic recordings from Roseway Basin. Percentage of true detections (number of true detections of all manually validated true calls), percentage of false detections (number of false detections out of all manually validated units without calls), recall, precision, and F1 score are also displayed (adapted from Moore 2017).

-	False Negative	False Positive	True Positive	True Negative	% True	% False	Recall	Precision	F1 Score
Per Snippet									
LFDCS	74	1,690	415	3,821	85%	31%	0.85	0.20	0.32
Spectroplotter	255	94	234	5,417	48%	2%	0.48	0.71	0.57
Per Hour									
LFDCS	167	157	218	2,459	57%	6%	0.57	0.58	0.57
Spectroplotter	193	75	192	2,541	50%	3%	0.50	0.72	0.59
Per Day									
LFDCS	7	23	65	30	90%	43%	0.90	0.74	0.81
Spectroplotter	14	18	58	35	81%	34%	0.81	0.76	0.78

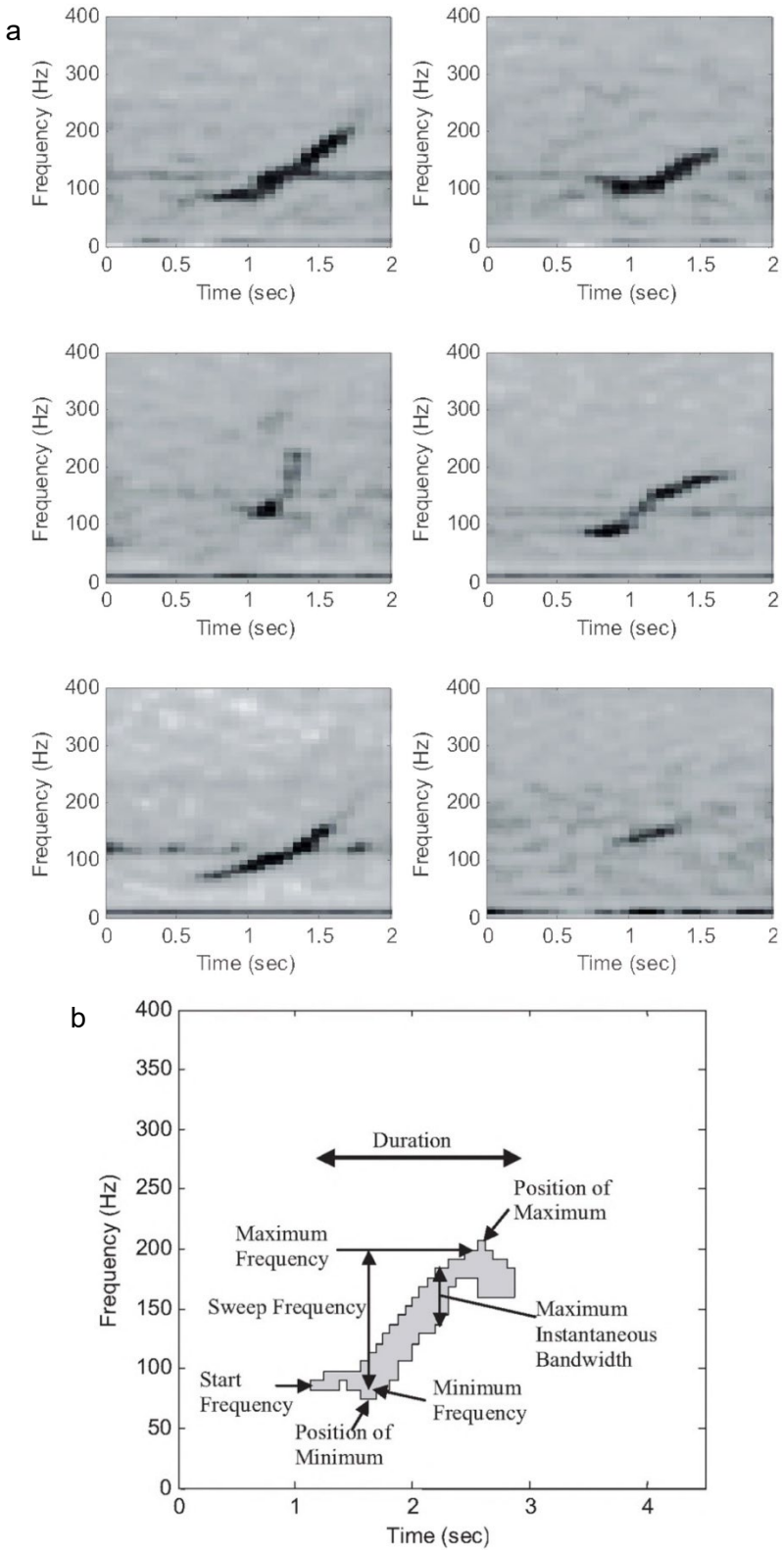

Figure 1(a). Spectrograms for upcall vocalization of NARW (adapted from Gillespie 2004), and 1(b). sample parameters measured by several NARW DCS discussed in this paper.

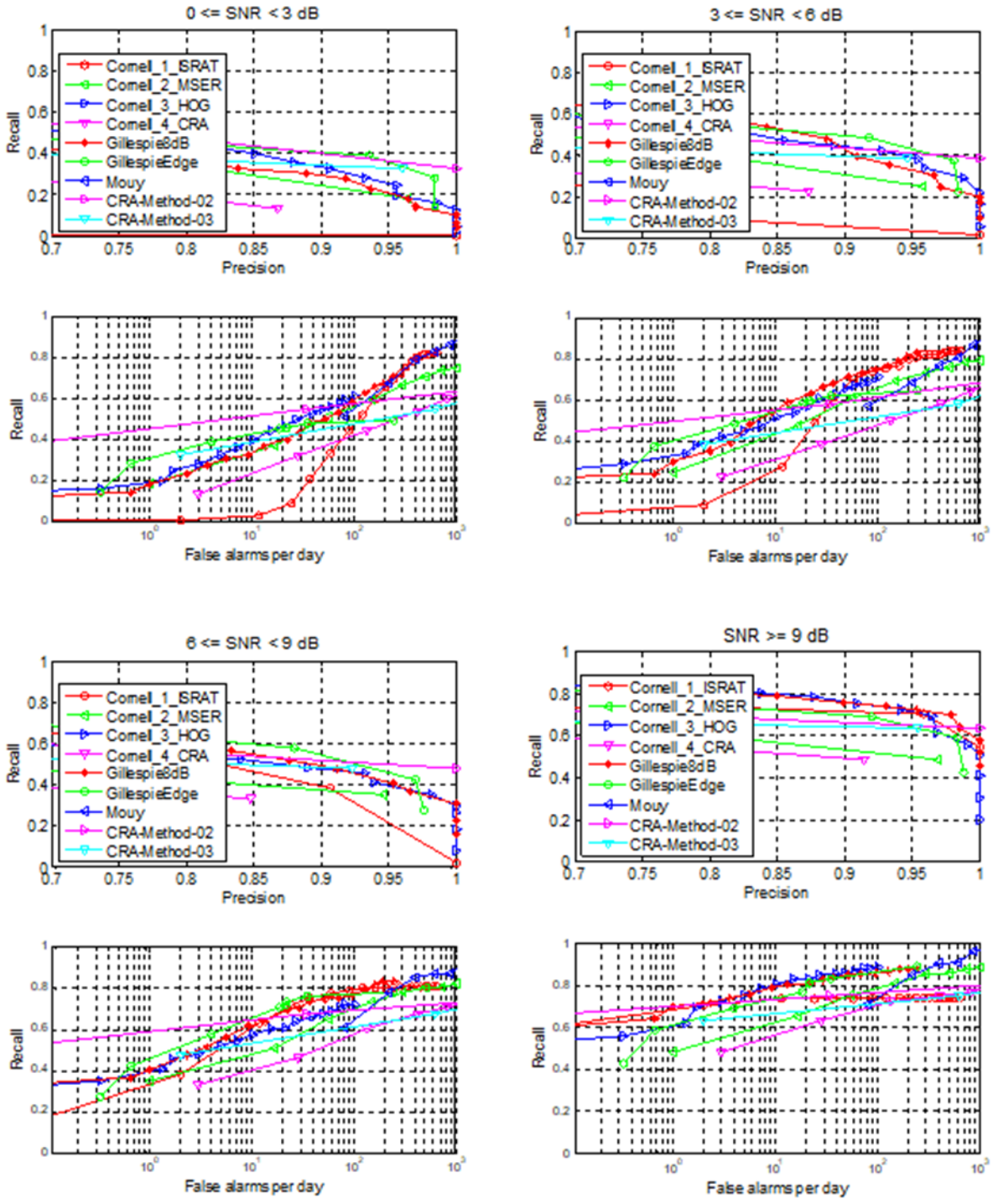

Figure 2. Relative performance of automated detectors in four different noise regimes. Detectors by Cornell, Gillespie, Mouy (JASCO), and CRA. Images courtesy of D. Gillespie.

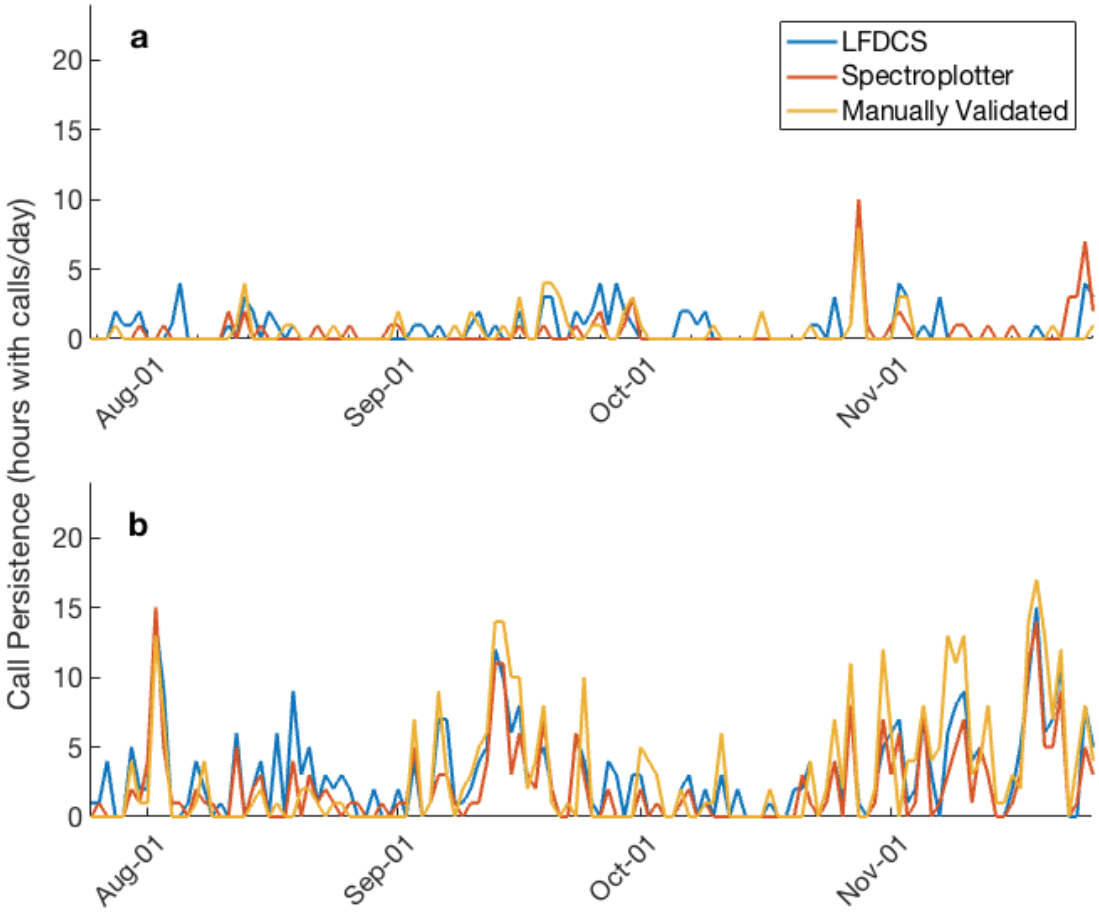

Figure 3. Plots of daily call persistence in (a) Emerald Basin and (b) Roseway Basin. LFDCS unvalidated auto-detections in blue, Spectroplotter unvalidated auto-detections in orange, and manually validated results in yellow (from Moore 2017).

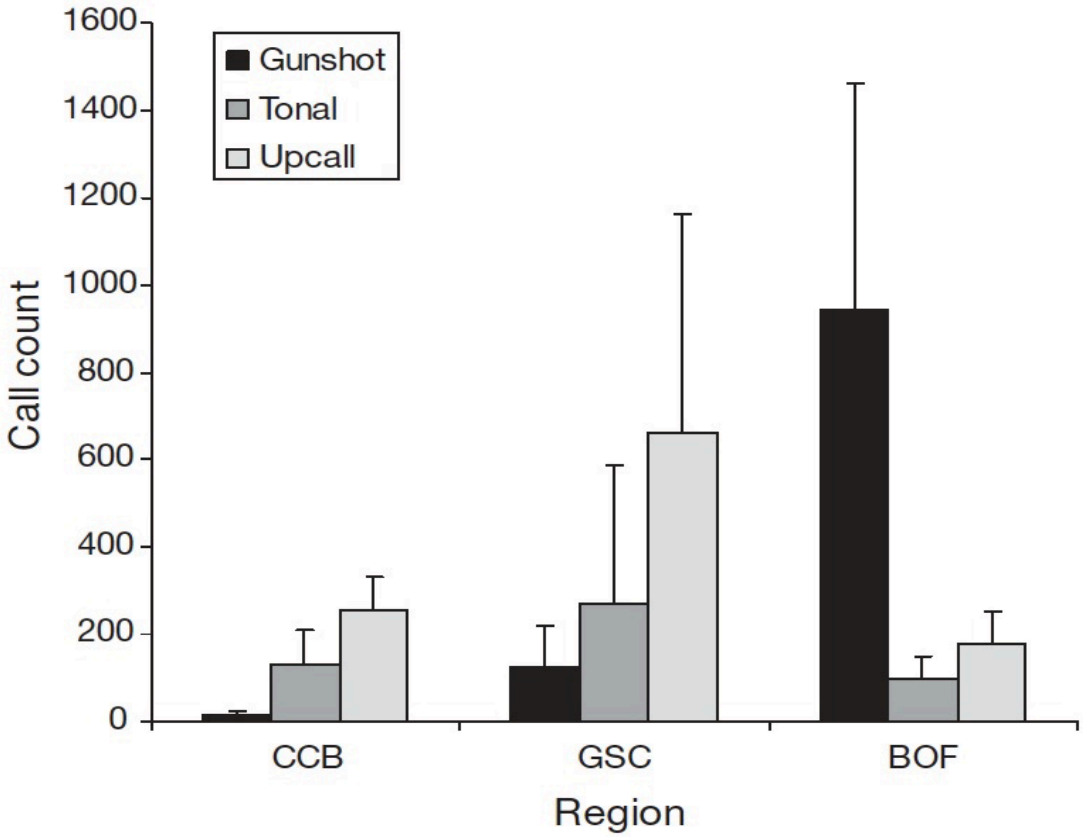

Figure 4. Archival PAM recorders deployed in three NARW habitats {Cape Cod Bay (CCB), Great South Channel (GSC), and Bay of Fundy (BOF)} demonstrated differences in the median (±SE) daily NARW call counts for three call types (Figure 2 from Van Parijs et al. 2009).

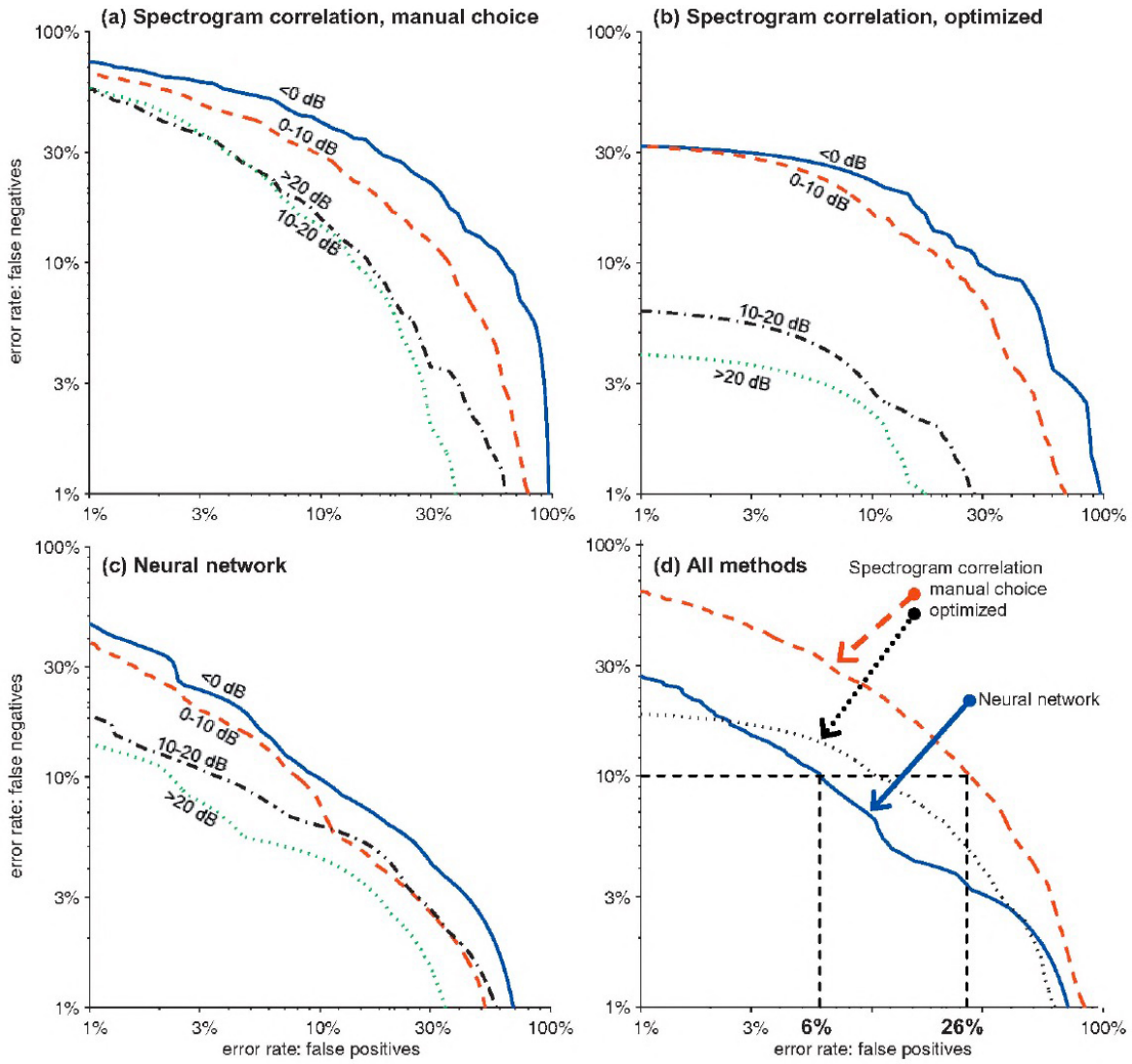

Figure 5. Performance curves of the three DCS methods (spectrogram correlation and neural network) compared in Mellinger (2004); the lower area under the curve, the better-performing the detector. The labels on the curves in (a) to (c) are the signal-to-noise ratio of the upcalls used for that curve. In (d) the DCS are compared with data from all signal-to-noise ratio upcalls combined (Figure 4 from Mellinger 2004).