Name ideas: RMDCS (raven mediated detection and classification system)

LFDCS: 1. Not optimized to reduce FPR, optimized to go fast

2. only on mac

3. interfaced through terminal, which makes it hard to look under the hood and tweak actual program

Mastor\_detecter

1. Optimized to reduce FPR, much slower computation
2. On PC, could probably port to mac
3. R script is interpretable and customizable by any researcher familiar with the language.

Autodetection tools are popular in passive acoustics to streamline manual analysis. Detector performance is commonly compared via TPR and FPR, but factors such as run time, ease of setup and use, and platform compatibility are also important considerations to performance. The LFDCS is a premier analysis tool in the field. It was designed to be compatible with near real-time detection on wave gliders so it is very computationally efficient. The trade off is that the learning method of the LFDCS (quadratic discriminant function analysis) only utilizes four features that do not contain information that allows for the discrimination of noise that can resemble positives for the target species. This can result in high FPR, particularly with environments and recorders that experience regular instrument noise, which is the case with AURAL recorders in the arctic. High FPR results in longer analysis time and more expensive analysis.

Despite not being optimized for this tradeoff from the perspective of acoustics labs working on archival data, the LFDCS is often used for this purpose. The LFCDS is also a Mac only implementation and interfaced through the command line, so ubiquity and customization for analysts is limited.

We submit an alternative approach for machine assisted analysis of low frequency sounds, optimized for efficient analysis time on archival data. This approach relies on a representative library of both the positives and negatives of a pitch tracker on ground truth reference data. It extracts measurements of each known and putative call to build and compare with random forest models. As you supply positive and negative detections, the models not only learn the identifying features of true positives, but also that of consistent types of false positives. In this way it ‘learns from its mistakes’. This is a flexible architecture that has been successfully applied to right whale upcalls and gunshots, as well as being able to discriminate from a variety of consistent sources of false positives due to noise. It is designed to be compatible with any stereotyped, distinct call type, and resilient in a variety of acoustic environments.

The detector is associated with Raven Pro 1.5 as part of its core functionality. The detector uses Raven Pro 1.5 Band Limited Energy Detectors (BLEDs) through API, and selection table outputs are formatted to be reviewed in Raven. The use of Raven software, as well as an implementation of the R language, make Mastor\_detecter comfortable to interface with and customize for scientists in the field.

Contents

[Glossary 2](#_Toc6480221)

[Introduction: 4](#_Toc6480222)

[Methods: 4](#_Toc6480223)

[Construction and design: 4](#_Toc6480224)

[Building autodetectors: 6](#_Toc6480225)

[Results 7](#_Toc6480226)

[Discussion 7](#_Toc6480227)

[Performance and recommendations 7](#_Toc6480228)

[Future directions 7](#_Toc6480229)

[References 7](#_Toc6480230)

[Supplemental 7](#_Toc6480231)

[1. Raven and custom algorithm 7](#_Toc6480232)

[2. Feature extraction 9](#_Toc6480233)

[3. Random forest models and ground truth considerations 13](#_Toc6480234)

[4. Adaptive compare 13](#_Toc6480235)

# Glossary

|  |  |
| --- | --- |
| Pitch tracker (PT) | Part of the detector that refers to the combined output of the Raven BLED suites and custom algorithm |
| Band Limited Energy Detector (BLED) | Implemented in Raven |
| Band Limited Energy Detector (BLED) suites | Group of BLEDs that is ran to produce MDs with high resolution time and frequency components |
| Minidetection (MD) | Output from the BLED suites. |
| Custom algorithm | User designed algorithm that uses the time and frequency components of MDs among other parameters and rules to filter likely FPs. |
| AURAL |  |
| LFDCS |  |
| Mastor\_detecter | Name of described detection and classification system. |
| True positive rate (TPR) | Also known as sensitivity. Defined as TP/(TP+FN) |
| False positive rate (FPR) | Precision of false positives. Defined as FP/(TP+FP). Note for myself: FPR may be a misnomer, seems to be calculated differently for technical definition, but we seem to be using it the same way as flightcallr paper. |
| True positive (TP) | Defined as an autodetection that bounds the meantime of a manually annotated detection and/or has its own meantime bounded by a manually annotated detection. |
| False positive (FP) | Defined as an autodetection that does not bound the meantime of a manually annotated detection and does not have its own meantime bounded by a manually annotated detection. |
| False negative (FN) | Defined as an manually annotated detection that does not bound the meantime of an autodetection and does not have its own meantime bounded by an autodetection. |
| Meantime | Average time of a time by frequency box. = (start time + end time)/2 |
| Masking | Effect of higher signal noise obscuring desired signal when overlaid on same time and frequency. Note to self put in another section: The Raven BLEDs and feature extraction both respond poorly to masking. For the Raven BLEDs, recorder self-noise can interfere with calculations of background noise, and for the image analysis component of feature extraction, thresholding will only choose the loudest portion of an autodetection to measure and will miss attributes of more faint call. It is possible to do feature extraction with variable amplitude thresholding, where different ranges of amplitude for a given autodetection each had features extracted, but this would increase run time of script by roughly 2 \* # of amplitude ranges. |
| Island | Isolated region of 1s (shape presence) spectrogram representation of sound after binarizing image. |
| Shape presence | Refers to the 1s in a binarized spectrogram |
| Random Forest (RF) | Bagged decision tree based machine learning approach. Robust to overfitting, light computing requirements, excellent performance, and doesn’t require data standardization |
| Multibox (MB) | A redundant TP- it shares a GT box with one or more autodetection. Contributes to inflated #TP, and can be lower quality TP as it only represents part of a TP. Only applicable for hand boxed DS |
| Overbox (OB) | A TP which encompasses more than one GT box. Contributes to deflated #TP, and is a lower quality TP as it has features of multiple SOI instead of 1. |
| AUC | Area under the ROC Curve. Represents the likelihood of correct classification along probability thresholds |
| Data segments (DS) | Sections of high graded time waveform data that are used to build and analyze performance of a detector |

# Introduction:

Mooring self-noise, defined as the noise created from disturbance of the instrument by water turbulence as well as the direct result of turbulence on the transducer, is a common issue among acoustic studies. Mooring self-noise is highly intermittent and hard to characterize. Mooring self-noise can have a high amplitude due to receiver proximity, in which case it becomes a source of interference for determination of background noise levels and has a masking effect on short-duration low frequency SOI. Mooring self-noise can strongly effect the viability of signal detection for an analysis. In regimes where mooring noise is the highest amplitude signal, amplitude detectors will be overwhelmed by false positives (FPR), and in regimes where mooring noise is audible as turbulent “gusts”, pitch tracking detectors will persistenly identify their desired slope along their FM boundary and also suffer from high FPR (FIG#).

AURAL recorders in the arctic region can suffer from a high degree of self-noise, making implementation of autodetectors challenging due to increased FPR. Popular autodetectors such as the LFDCS (which uses QDFA on 4 features: ) and Ishmael (which uses spectrogram correlation) are not resilient against false positives produced by intermittent noise as they don’t have the necessary features and learning system to distinguish mooring self-noise from positive calls. High FPR hurts the implementation of an autodetector as it either 1. increases the time of an analyst to verify detections or 2. reduces TPR due to having to set a lower threshold for time efficient analysis.

Mastor\_detecter was created as a solution to the high FPR issue encountered with existing autodetectors on high self-noise recordings. While the idea to model true positives from a band limited energy detector against null data is not new (flightcallr, Mellinger 2004), mastor\_detector is the first detector that models the true and false positives of a pitch tracker result to identify and weed out consistent sources of false positives. Reducing the FPR of outputs enables higher TPR analysis for use cases like boxing or behavior studies, or for faster analysis at a lower set TPR for use cases like seasonal presence. As the models are spontaneously generated from a set of features in the ground truth library, this method should be applicable to other regions that experience high levels of self-noise on their recorders even if the composition and prevalence of the noise is different.

Mastor\_detecter has similar use limitations to existing pitch tracking detectors- it can be used to identify stereotyped, distinct calls that preferably feature FM, but is vulnerable to anything that will mask or obscure the shape of calls. It cannot use species presence or patterning of calls as a criteria for detection, so manual verification is required in cases where the call type in question may be attributable to multiple species (ie NPRW upcalls and bowhead, humpback, bearded seal upsweeps). For this reason, we refer to the detector that was designed for right whale upcalls as an upsweep detector, since there are multiple species in our system that will produce sounds that are indistinguishable from high probability upcalls.

Mastor\_detecter is easy to use, but hard to set up. BLEDs must be created manually in Raven, and the custom algorithm must be coded in R language. Mastor\_detector will never be an executable program, but rather a flexible infrastructure that can be customized to various call types. Due to high runtimes, it is not meant to be applicable to real time or near real time detection, but serves as an efficient method for reducing the analyst time needed for detecting or boxing desired calls from archival acoustic data.

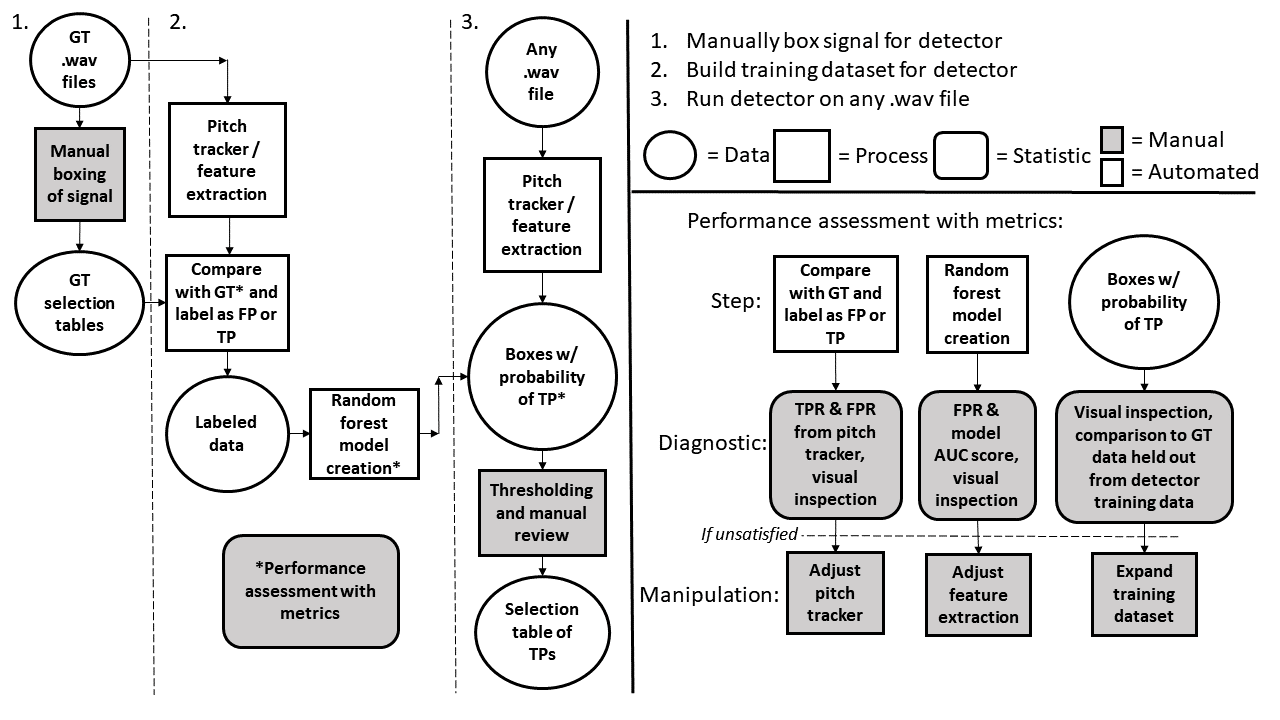
We have successfully applied Mastor\_detector to right whale upcalls and gunshots and demonstrate results for these applications, as well as provide guidance for its implementation to any distinct, stereotyped signal of interest.

# Methods:

## Construction and design:

We identified the Raven BLEDs as a good starting place to assess the viability of applying energy detectors to our data, given the comfortability of the Raven interface. Upon finding an R implementation to call this algorithm from API, we decided to continue developing around Raven and the R language. It was determined that in environments where the amplitude of mooring noise was greater than that of individual calls, wide band limited energy detectors worked poorly for pulling out faint calls, whereas narrowband BLEDs were more likely to hit pieces of faint calls (hereafter referred to as minidetections (MDs)). With this knowledge, we pursued using a ‘suite’ of narrowband BLEDs throughout the frequency range of the desired call. Prewhitening was built in as an option, although it has to be manually initiated Raven and did not have a strong effect on performance (see Supplemental: 2. Whitening). Using the suite of detectors allowed for an additional ‘pitch-tracking’ filter, by comparing the frequency and time of each MD to assess FM of the signal and filter out some unlikely detections. Using only this pitch tracker method garnered too many FPs to make analysis feasible, so a machine learning solution using RF models inspired by the detector design in (flightcallr) was implemented to help weed out FPs. Image analysis was later introduced to improve the classification performance of the RF models. Parallelization is implemented while running the pitch tracking algorithm, extracting features, and building the models, and is highly recommended to improve run time. The infrastructure for these parallelized sections in R package ‘foreach’ is in theory portable to Azure given that Microsoft developed this package for compatibility with Azure (<https://cran.microsoft.com/web/packages/foreach/foreach.pdf>).

Much of the infrastructure was reworked for compatibility with multiple species detection in a single run of the detector. Included in this idea was the implementation of a more computationally efficient multiclass RF classifier, which seemed to have poor results compared to binary classification. For future attempts at multiclass detection, iterative binary detection and *post hoc* comparison is a safe approach if using RF models as a classifier. Use of different models besides RF may demonstrate better results at multiclass detection. This capability was scaled back as our timeline prioritized effective single species detection during development, but the looping structure is largely intact for this capability.

**Figure 1.** Generalized workflow of mastor\_detecter. 

## Building autodetectors:

### Upsweep

An upsweep detector was developed using the mastor\_detecter workflow (Figure 1). Data segments from seven different AURAL deployments that had high presence of RW were chosen to build the pitch tracker and train the RF models (Table 1). These data were decimated by a factor of 16, and combined into single sound files corresponding to mooring year. Upcalls in the initial datasets were boxed by hand in Raven, and later using the output of the pitch tracker (Table 1). A suite of Raven BLEDs were designed by hand within a conservative frequency range (table 2) to best pick up upcall MDs. A custom algorithm was designed in R to identify likely groups of MDs as putative detections and filter out likely FPs (TABLE#). The Raven BLEDs and algorithm were tested by iteratively running the section of the script “runRavenGT” while changing structure and parameters of both. Performance was evaluated using visual examination of generated selection tables in Raven 1.5 and using autogenerated performance metrics on TPR and FPR.

Once satisfactory results were generated (TABLE#) from the newly designed pitch tracker, we built a labelled dataset using the section of the script “runProcessGT”. This section of the script associates the generated selection tables with your ground truth selection tables to assign TP or FP labels for each box, and extracts features from each to build the training dataset that will inform the random forest models for the detector.

Once we had built a pitch tracker and generated the labelled data, the detector is functional to be applied to any .wav file. To test the performance of the final detector, we trained models recursively on the labelled set to test performance (TABLE #) and evaluated results by visual inspection in Raven. Manipulations performed here to improve the model included adding features new features to the labeled data and adding post hoc filters to clean up issues with multiple boxes being assigned to single calls (adaptive\_compare). For upcalls, a detection combination time threshold of 1.2 seconds and a probability threshold of 0.2 was found to show good adaptive\_compare performance.

### Gunshot

The same process was applied for gunshot detection. For this detector the data were decimated by a factor of 8. Data were originally selected from moorings that had high incidence of gunshot (Table 1), and a total of 11 DS were used for building the labeled data and analyzing performance. We incorporated an extra variable ‘timesepGS’ for the gunshot algorithm, which is an arbitrary value that allows the algorithm to draw a line that serves to shape the exponential function that determines the frequency dynamic time comparison of MDs. To preserve the ability to isolate double gunshot (FIGURE#), we adjusted the adaptive\_compare time threshold to 0.15 while keeping the probability threshold at 2.

## Unlabeled data experiment:

Once detector performance on the labeled data was found to be sufficient, we applied the detector to new data to test the generalization of the random forest classifier. To test the effect of increasing the number of DS in the labelled set on model performance we constructed a learning curve for AUC score and FPR. The labelled set was built iteratively with randomly selected DSs, comparing the detector performance in each iteration to the both the labelled and unlabeled DSs until all data were included in the labelled set. The effect of adding DSs on performance was obscured by the amount of variation in performance for each individual DS, so to account for this effect we ran the experiment 30 times to decrease individual differences in DS.

**Table 1.** Ground truthing effort for gunshot and upcalls. ‘Machine assisted = n’ means that data was boxed manually, and ‘Machine assisted = r’ means that boxing was performed by annotating the output of the pitch tracker. This switch in protocol was made to increase analysis speed and convenience, initiated once 1. The pitch tracker was performing to satisfaction on a variety of data 2. More data was needed to train the model, which doesn’t consider FN. Because of this, total TPR is not comparable between ‘n’ and ‘r’ machine assisted data, although AUC and FPR are still valid comparisons.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Call type | Location | # high graded files | % analyzed | Data hours | Analysis start | Analysis end | Machine assisted | How data was  selected |
| Gunshot | BS3 | 1452 | 3.10 | 2.50 | 11/06/2015 | 11/14/2015 | n | Hand |
| Gunshot | M2 | 455 | 10.11 | 2.65 | 06/19/2012 | 07/20/2012 | n | Hand |
| Gunshot | BS3 | 583 | 12.18 | 3.96 | 11/05/2014 | 12/04/2014 | n | Hand |
| Gunshot | M4 | 289 | 30.45 | 4.79 | 08/15/2014 | 09/11/2014 | n | Hand |
| Gunshot | BS3 | 3764 | 6.64 | 13.77 | 08/11/2012 | 08/26/2012 | r | Hand |
| Gunshot | BS3 | 3764 | 1.17 | 2.38 | 10/07/2012 | 10/11/2012 | r | Hand |
| Gunshot | BS3 | 583 | 10.29 | 3.40 | 12/21/2014 | 12/27/2014 | r | Random |
| Gunshot | BS2 | 391 | 23.02 | 3.98 | 10/07/2015 | 10/09/2015 | r | Random |
| Gunshot | BS3 | 576 | 16.15 | 5.35 | 12/03/2016 | 07/01/2017 | r | Random |
| Gunshot | M2 | 1383 | 4.19 | 3.27 | 09/20/2012 | 09/23/2012 | r | Random |
| Gunshot | M4 | 166 | 44.58 | 4.15 | 04/02/2015 | 08/30/2015 | r | Random |
| Upcall | M2 | 194 | 53.61 | 5.88 | 06/22/2015 | 08/09/2015 | n | Hand |
| Upcall | M4 | 179 | 100.00 | 10.08 | 12/02/2014 | 09/20/2015 | n | Hand |
| Upcall | BS3 | 217 | 100.00 | 12.10 | 08/11/2012 | 09/12/2013 | n | Hand |
| Upcall | M4 | 304 | 100.00 | 16.85 | 09/18/2013 | 10/04/2014 | n | Hand |
| Upcall | M2 | 325 | 53.85 | 9.83 | 05/17/2016 | 09/08/2016 | n | Hand |
| Upcall | M2 | 62 | 100.00 | 3.44 | 10/01/2015 | 12/24/2015 | n | Hand |
| Upcall | BS3 | 443 | 36.12 | 9.00 | 10/20/2014 | 06/26/2015 | n | Hand |
| Upcall | BS1 | 158 | 100.00 | 9.06 | 09/26/2016 | 11/25/2016 | r | Random |
| Upcall | M2 | 58 | 100.00 | 3.15 | 07/03/2013 | 07/18/2013 | r | Random |
| Upcall | M5 | 13 | 100.00 | 0.75 | 09/29/2016 | 12/08/2016 | r | Random |
| Upcall | M4 | 540 | 21.48 | 6.56 | 08/15/2016 | 08/26/2016 | r | Random |
| Upcall | BS2 | 86 | 100.00 | 4.71 | 07/21/2015 | 09/25/2015 | r | Random |
|  |  |  |  |  |  |  |  |  |

# Results

## Performance on labeled data

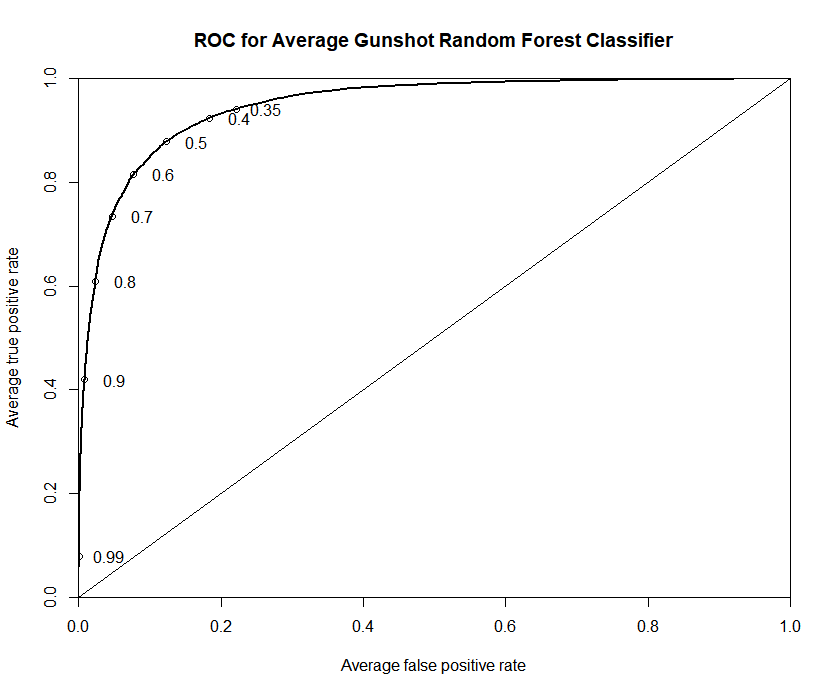
### Gunshot

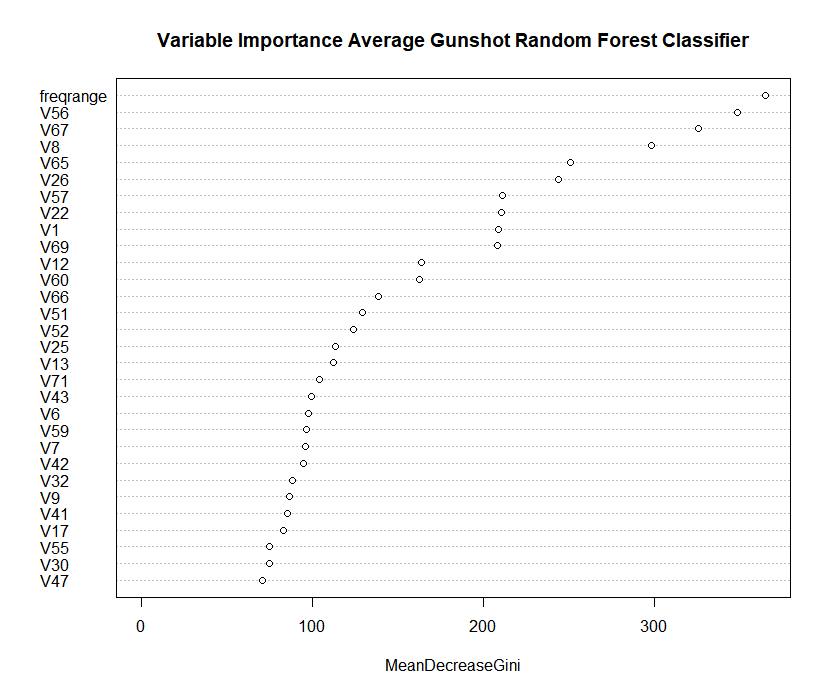
The four DS that had a corresponding hand boxed GT set were able to be compared with each other to assess accuracy of the PT and the entire detector. The PT has an accuracy of 0.95 (n=4, ±?) for these moorings. These DS had an overall TPR of 0.92 (n=4, ±?), and an RF classifier TPR of 0.97 (n=4, ±?). As the hand boxes did not necessarily line up with the PT generated boxes as with the machine assisted DS, MB% and OB% could be calculated as measures of box quality compared to by-hand boxes. MB% occurs at a rate of 13.25 (n=4, ±?), with higher incidence at BS12\_AU\_02a and BS13\_AU\_04, and OB% occurs at a very low rate throughout.

The remaining moorings were machine assisted which did not allow for a total TPR estimate due to a missing component of FNs but did give an upper bound for total TPR based on the FNs generated from the RF classifier. The machine assisted moorings had an average RF classifier accuracy of 0.95 (n=7, ±?) giving these DSs an upper bound of 0.95 for overall TPR.

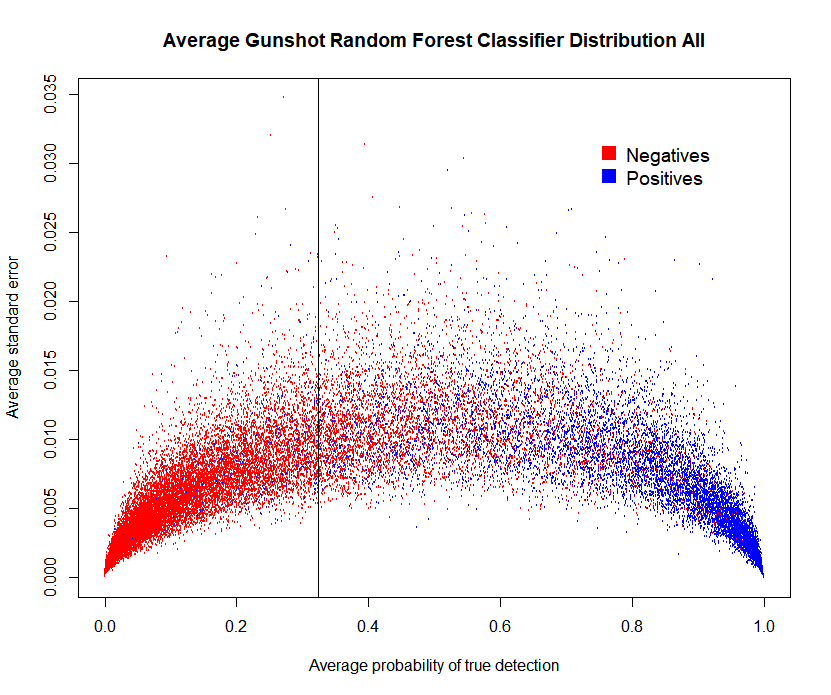
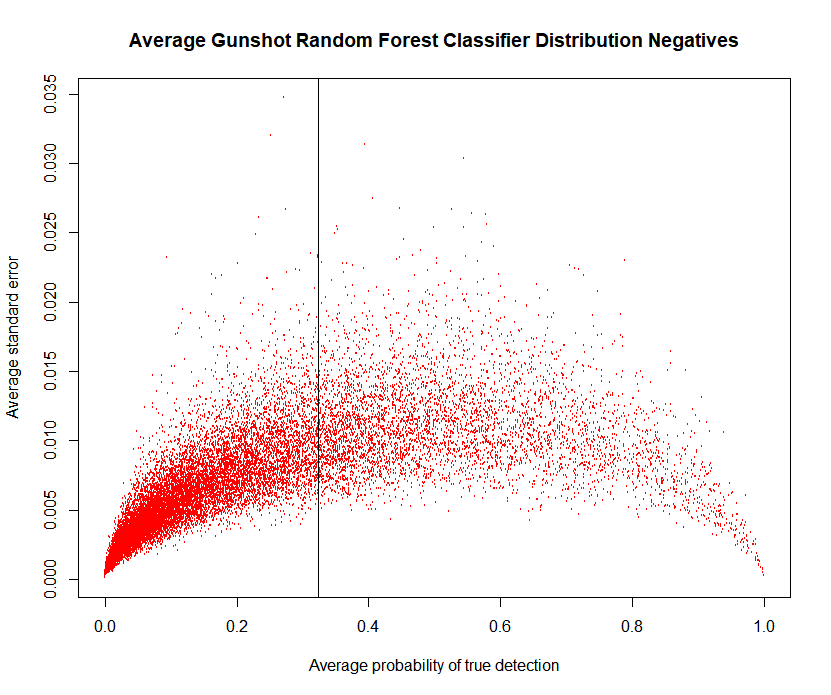
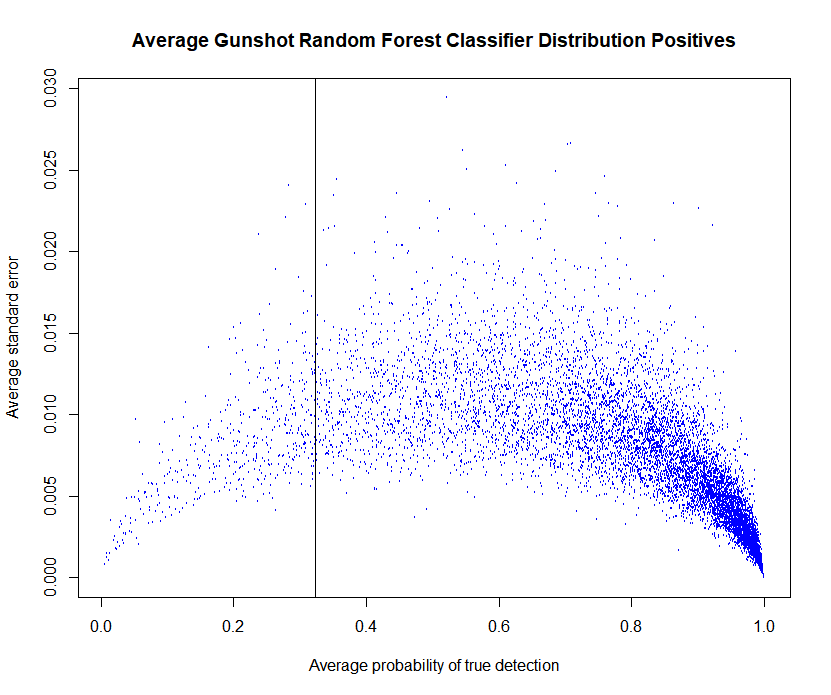
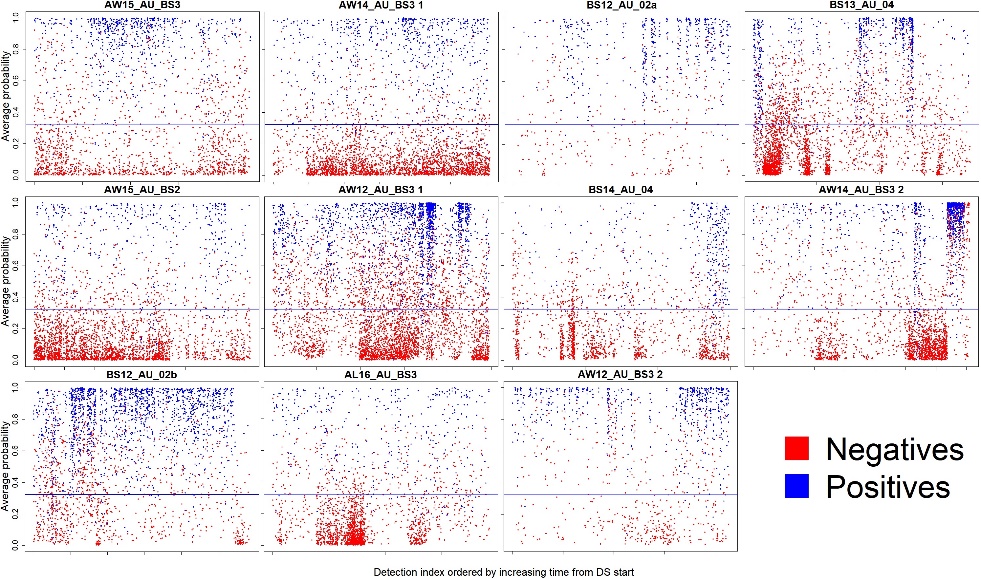
Statistics that did not involve the number of FNs from the PT were able to be compared across hand and machine assisted DS. The PT FPR is 0.7 (n=11, ±?), and the FPR is reduced to 0.37 (n=11, ±?) with the RF classifier. The FPR can also be interpreted as a ratio: for every 20 TPs in the final detector output at an FPR of 0.37, expect 7-8 FPs on average. The total AUC score across all DS is 0.95 (n=11, ±?).

A variety of features are effective in reducing the Gini coefficient for the random forest classifier. The most important features by this measure were freqrange, MedSlope Hough (V56), SwitchesY max (V67), autoc se (V8).

**Figure x.** ROC curve of the random forest classifier for gunshots. Different thresholds along the curve are labeled. The diagonal line represents the performance of a random classifier. 

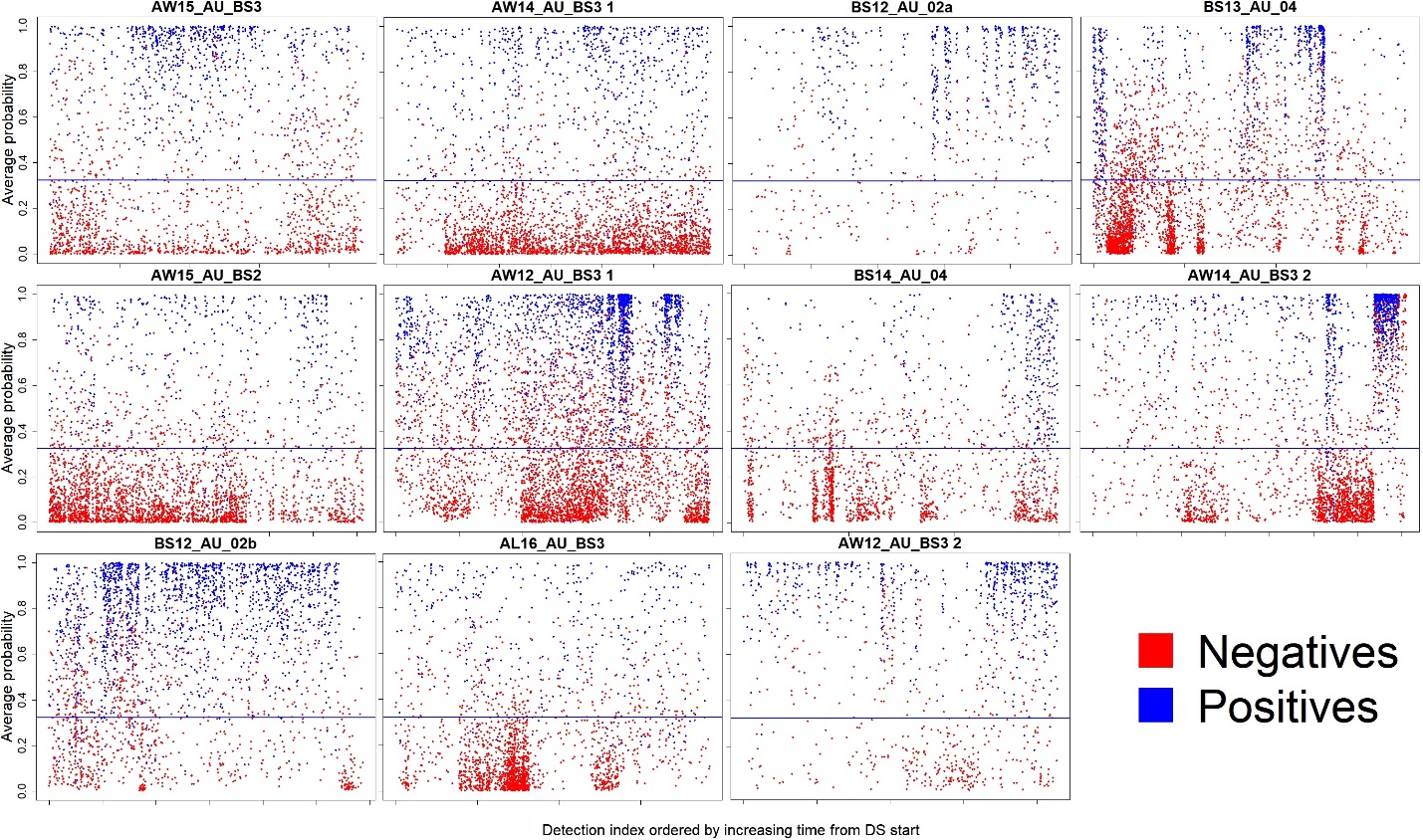
**Figure x.** Mean decrease in Gini coefficient for features included in the random forest classifier. More informative variables correspond to a larger decrease in Gini coefficient. Feature names and description correspond to V# listed in table x. 

**Figure x.** Average probability and average probability standard error for each autodetection in labelled set after applying random forest classifier with x60 cross validation. Vertical line shows the probability cutoff, where autodetections with probability >= cutoff are kept and probability <= cutoff are discarded. This cutoff is set such that 95% of TPs are kept. High and low probability autodetections show lower variance in probability standard error, while those in the midrange see higher variance in probability standard error.



**Threshold: x = .32**

**Figure x**. Probability over index arranged by time for all DS included in the gunshot detector labelled data. Horizontal line shows the probability cutoff, where autodetections with probability >= cutoff are kept and probability <= cutoff are discarded. This cutoff is set such that 95% of TPs are kept.



Detection index (in consecutive order from start of DS)

**Table x.** Performance of gunshot detector on high graded, ground truth data segments without the adaptive\_compare filter. Total TPR was not calculated for DS where the ground truth protocol was machine assisted as there are an unknown component of FNs from the PT.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DS:  Hand boxed | PT TPR | RF TPR | Total TPR | PT FPR | Total FPR | MB % | OB % | AUC |
| AW15\_AU\_BS3 | 0.97 | 0.97 | 0.94 | 0.67 | 0.35 | 6.67 | 0.43 | 0.95 |
| AW14\_AU\_BS3 1 | 0.94 | 0.98 | 0.92 | 0.68 | 0.35 | 10.60 | 0.49 | 0.94 |
| BS12\_AU\_02a | 0.96 | 0.99 | 0.95 | 0.41 | 0.26 | 19.32 | 0.00 | 0.91 |
| BS13\_AU\_04 | 0.94 | 0.95 | 0.90 | 0.81 | 0.54 | 19.14 | 0.94 | 0.93 |
| Hand boxed all | 0.95 | 0.97 | 0.92 | 0.70 | 0.40 | 13.25 | 0.51 |  |
| DS: Machine assisted |  |  |  |  |  |  |  |  |
| AW15\_AU\_BS2 |  | 0.89 | <0.89 | 0.86 | 0.43 |  |  | 0.95 |
| AW12\_AU\_BS3 1 |  | 0.97 | <0.97 | 0.68 | 0.42 |  |  | 0.95 |
| BS14\_AU\_04 |  | 0.90 | <0.90 | 0.79 | 0.47 |  |  | 0.92 |
| AW14\_AU\_BS3 2 |  | 0.91 | <0.91 | 0.79 | 0.25 |  |  | 0.97 |
| BS12\_AU\_02b |  | 0.96 | <0.96 | 0.39 | 0.23 |  |  | 0.92 |
| AL16\_AU\_BS3 |  | 0.91 | <0.91 | 0.84 | 0.46 |  |  | 0.95 |
| AW12\_AU\_BS3 2 |  | 0.99 | <0.99 | 0.45 | 0.26 |  |  | 0.94 |
| Machine assisted all |  | 0.95 | <0.95 | 0.70 | 0.35 |  |  |  |
| All |  |  |  | 0.70 | 0.37 |  |  | 0.95 |

**Table x.** Performance of gunshot detector on high graded, hand boxed ground truth DS with the adaptive\_compare filter. The filter reduced the MB% for these moorings and incrementally raised the OB%. It also had the effect of slightly reducing overall FPR.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DS:  Hand boxed | PT TPR | RF TPR | Total TPR | PT FPR | Total FPR | MB % | OB % | AUC |
| AW15\_AU\_BS3 | 0.97 | 0.97 | 0.94 | 0.67 | 0.34 | 1.22 | 0.46 | 0.95 |
| AW14\_AU\_BS3 1 | 0.94 | 0.98 | 0.92 | 0.68 | 0.34 | 4.56 | 0.62 | 0.94 |
| BS12\_AU\_02a | 0.96 | 0.99 | 0.95 | 0.41 | 0.24 | 5.00 | 0.00 | 0.91 |
| BS13\_AU\_04 | 0.94 | 0.95 | 0.90 | 0.81 | 0.53 | 11.55 | 1.02 | 0.93 |
| Hand boxed all | 0.95 | 0.97 | 0.92 | 0.70 | 0.39 | 5.58 | 0.59 |  |

## Unlabeled data experiment

### Gunshot

The results are presented graphically for this experiment in (FIG#). While the selection for holdout data was random, by coincidence AW12\_AU\_BS3 2 was the holdout data for the last iteration in 9/30 trials, which is a statistical outlier from the average of the remaining 10 eligible DS that were present in the last iteration an average of 2.1 times in the experiment, with a max of 3 (n=10, ±?) (3+2+1+3+3+2+2+1+2+2 / 10). It does not appear that the classifier performance on the holdout data ever matched the performance on the labelled data for these 11 DSs.

**Figure x.** Comparison of gunshot classifier performance on the FPR and AUC score DS in the labeled and holdout sets. A total of 30 trials were performed for the experiment. While large improvement are seen in both FPR and AUC with the addition of new DS to the labelled data at n = 2 & 3, performance on holdout data steadily increases but does not appear to achieve neutrality with the labeled set.

### 

The

# Discussion

## Performance on labelled data

### Gunshot

The gunshot detector shows good performance for the PT and the RF classifier. The PT for gunshots is stronger than that of the upcalls, while the classifier shows weaker performance (TAB#,TAB#). The stronger PT can be explained by the higher frequency range of the top end of the gunshots which often places these sounds outside of the typical mooring self-noise frequency range. This idea is supported by the frequency range being the most effective feature to reduce Gini coefficient, suggesting the height of gunshot detection often had a powerful effect on distinguishing a given call from mooring self-noise. The poorer performance for the gunshot classifier (AUC: TAB#) is at least in some part contributed to by the great performance of the PT- a more discriminatory PT will give the classifier better choices to choose from, reducing the probability of correct classification. While the AUC score of the classifier is worse than for upsweeps, the features that best reduce the Gini coefficient are varied in what they measure and likely more robust to call plasticity, whereas the most informative features in the upsweep classifier are all variations on slope measurements and may be weaker if applied to upcalls with a non-standard slope. Visual analysis of the ROC curve (FIG#) and probability distribution (FIGU#) shows that the classifier sees consistent good separation at both low and high probability but does not show the extreme separation at high probabilities like the upsweep detector. This means that some gunshot FPs should be expected even in analyses that require low TPR threshold.

## Unlabeled data experiment

Given the high computational requirements of this experiment, we recommend not to attempt to replicate this experiment as a step in detector construction, and suggest that the best method for evaluating detectors is visual inspection of generated selection tables on new data, providing data as needed in situations where the model is underperforming. Adding data from sections of data that have particularly high FPR or low TPR (unless in the case of masking) to the training set will make the model more sensitive to similar cases.

## Future directions

# References

Baumgartner, M., & Mussoline, S. (2011). A generalized baleen whale call detection and classification system. *The Journal of the Acoustical Society of America,* *129*(5), 2889-902.

Strasberg, M., & Taylor, D. (1979). Nonacoustic noise interference in measurements of infrasonic ambient noise. *Journal of the Acoustical Society of America,66*(5), 1487-1493.

# Supplemental

## Raven and custom algorithm

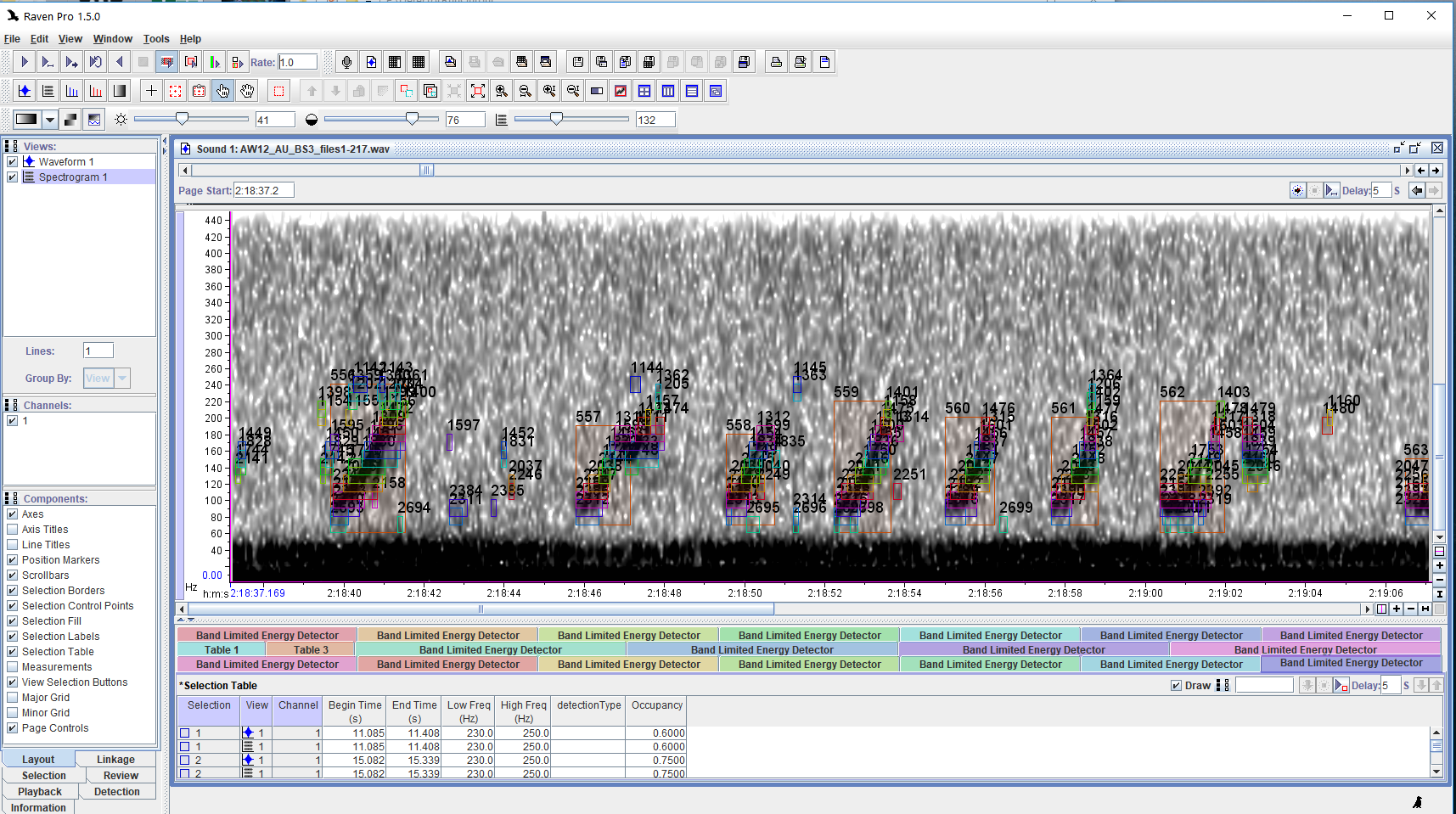
**Table x.** Parameters for suite of Raven BLEDs configured for upsweeps (P11). For the upsweep detector the parameters stay constant through the frequency range. See Raven Pro 1.4 users manual “Configuring Band Limited Energy Detectors” page 274 for in depth description of these parameters.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suite ID | Low freq (Hz) | High freq (Hz) | Min dur (s) | Max dur (s) | Min sep (s) | Min % occupancy | SNR Threshold (dB) | Block Size (s) | Hop Size (s) | Percentile |
| P11a | 60 | 80 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11b | 70 | 90 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11c | 80 | 100 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11d | 90 | 110 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11e | 100 | 120 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11f | 110 | 130 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11g | 120 | 140 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11h | 130 | 150 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11i | 140 | 160 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11j | 150 | 170 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11k | 160 | 180 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11l | 170 | 190 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11m | 180 | 200 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11n | 190 | 210 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11o | 200 | 220 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11p | 210 | 230 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11q | 220 | 240 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
| P11r | 230 | 250 | 0.25 | 4 | 0.25 | 30 | 3.5 | 5 | 0.25 | 60 |
|  |  |  |  |  |  |  |  |  |  |  |

**Table 3.** Parameters for suite of Raven BLEDs configured for gunshots (Pg2). This detector has changing parameters over the frequency range to better account for the effects of propagation on perceived call duration. See Raven Pro 1.4 users manual “Configuring Band Limited Energy Detectors” page 274 for in depth description of these parameters.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suite ID | Low freq (Hz) | High freq (Hz) | Min dur (s) | Max dur (s) | Min sep (s) | Min % occupancy | SNR Threshold (dB) | Block Size (s) | Hop Size (s) | Percentile |
| Pg2a | 750 | 850 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2b | 700 | 800 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2c | 650 | 750 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2d | 600 | 700 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2e | 550 | 650 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2f | 500 | 600 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2g | 450 | 550 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2h | 400 | 500 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2i | 350 | 450 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2j | 300 | 400 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2k | 250 | 350 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2l | 200 | 300 | 0 | 0.625 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2m | 150 | 250 | 0.125 | 1 | 0.125 | 50 | 3.5 | 1 | 0.25 | 65 |
| Pg2n | 135 | 165 | 0.125 | 1.75 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2o | 125 | 145 | 0.125 | 1.75 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2p | 115 | 135 | 0.125 | 1.75 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2q | 105 | 125 | 0.125 | 2.5 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2r | 95 | 115 | 0.125 | 3 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2s | 85 | 105 | 0.125 | 3 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2t | 75 | 95 | 0.125 | 3 | 0.5 | 20 | 4 | 6 | 1 | 65 |
| Pg2u | 65 | 85 | 0.125 | 3 | 0.5 | 20 | 4 | 6 | 1 | 65 |

**Figure x.** Screenshot of Raven Pro 1.5 on ground truth data showing MDs from Raven suite P11, superimposed with positive detections from the upsweep custom algorithm (larger orange boxes) that compare these MDs within the script. At detection #557 at 2:18:44, there is an example of a missed box- a positive detection that misses a part of the call. At present missed boxes still register as detections if they pass the criteria, and their prevalence is not quantified.



**Figure x.** Demonstration of pitch tracker. Superimpose two calls vertically with MDs- one for a negative and one for a positive detection (could also throw in a FP). Under both write the logic for the first step of the algorithm- “for every MD, compare all succeeding MDs for a viable upsweep within the given rules (determined by allowedZeros, detskip, etc)”. Show table comparing runs (and grpsize parameter). Show criteria for comparing runs, and the row with highest ranked run. Show that best run is chosen as PT detection, and show the box on the positive detection, and no box on the negative detection. Do this for RW and GS.

## Pre-whitening

Pre-whitening, or normalization, is a data transformation that is designed to reduce the effect of long term narrowband noise, commonly vessels. Given that the data used to create the labeled dataset is prescreened to not include vessel, and the strongest source of interference is composed of intermittent mooring self-noise, the application of this transformation is suspect for the purposes of constructing the detector.

If a detector is poorly performing when confronted with persistent narrowband noise, pre-whitening the data may be a good option. This must be done in Raven using the batch adaptive filter (see Raven Pro 1.4 users manual “Adaptive Filtering” page 148), and your whitening preference must be updated in the script parameters to enable proper pathing.

## Feature extraction

**Table x.** Features extracted for each autodetection.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature name | Type | Column name | Description (italics if from CRAN) |
| freq range | Simple | freq range | Top freq - bottom freq |
| Rugosity | Time wave | V1 | *The rugosity of a time wave* |
| Crest | Time wave | V2 | *Returns the crest factor and localizes the different crest(s)* |
| Temporal entopy | Hilbert amplitude envelope of time wave | V3 | *Compute the entropy of a temporal envelope.* |
| Shannon entropy | Hilbert amplitude envelope of time wave | V4 | *Shannon entropy of a frequency spectrum* |
| Roughness | Mean frequency spectrum of time wave | V5 | *The roughness or total curvature of a curve, i.e. of a time wave or of a spectrum* |
| autoc mean | Short term autocorrelation of time wav | V6 | Mean of short-term autocorrelation of time wave |
| autoc median | Short term autocorrelation of time wav | V7 | Median of short-term autocorrelation of time wave |
| autoc se | Short term autocorrelation of time wav | V8 | Standard error of short-term autocorrelation of time wave |
| dfreq mean | Dominant frequency of time wave | V9 | Mean of dominant frequency of time wave |
| dfreq se | Dominant frequency of time wave | V10 | Standard error of dominant frequency of time wave |
| specprop mean | Frequency spectrum of time wave | V11 | *Mean frequency* |
| specprop sd | Time wave | V12 | *Standard deviation of the mean* |
| specprop se | Time wave | V13 | *Standard error of the mean* |
| specprop median | Time wave | V14 | *Median* |
| specprop mode | Time wave | V15 | *Mode (dominant)* |
| specprop q25 | Time wave | V16 | *First quartile* |
| specprop q75 | Time wave | V17 | *Third quartile* |
| specprop IQR | Time wave | V18 | *Interquartile range* |
| specprop centroid | Time wave | V19 | *Centroid* |
| specprop skewness | Time wave | V20 | *Skewness* |
| specprop kurtosis | Time wave | V21 | *Kurtosis* |
| specprop sfm | Time wave | V22 | *Spectral flatness measure* |
| specprop sh | Time wave | V23 | *Spectral entropy* (possibly redundantwith V4) |
| specprop precision | Time wave | V24 | *Frequency precision* |
| Amp env median | Amplitude envelope of time wave | V25 | *Acoustic index based on the median of the amplitude envelope* |
| Total entropy | Time wave | V26 | *Total entropy of a time wave* |
| NULL | NULL | V27 | Removed feature. Occupied by values of 1 |
| Modinx | Dominant frequency of time wave | V28 | Cummulative change in dominant frequency, taken from warbleR |
| Startdom | Dominant frequency of time wave | V29 | Dominant frequency at initial time |
| Enddom | Dominant frequency of time wave | V30 | Dominant frequency at final time |
| Mindom | Dominant frequency of time wave | V31 | Lowest frequency dominant frequency |
| Maxdom | Dominant frequency of time wave | V32 | Highest frequency dominant frequency |
| Dfrange | Dominant frequency of time wave | V33 | Range of dominant frequency values |
| Dfslope | Dominant frequency of time wave | V34 | (Enddom-Startdom)/duration |
| Meanpeakf | Mean frequency spectrum of time wave | V35 | frd\_wrblr\_int.R . Unsure of purpose but has been informative |
| AreaX maxP | Spectrogram | V36 | Which quantile contains max of sum of shape presence (1s in binary matrix) along x axis quantiles |
| AreaX max | Spectrogram | V37 | Max of sum of shape presence (1s in binary matrix) within x axis quantiles |
| AreaX dom | Spectrogram | V38 | Max of sum of shape presence (1s in binary matrix) within x axis quantiles / total sum of shape presence along all quantiles |
| AreaX std | Spectrogram | V39 | standard error of sum of shape presence (1s in binary matrix) within x axis quantiles. Given value of 0 if NA |
| AreaY maxP | Spectrogram | V40 | Which quantile contains max of sum of shape presence (1s in binary matrix) within y axis quantiles |
| AreaY max | Spectrogram | V41 | Max of sum of shape presence (1s in binary matrix) within y axis quantiles |
| AreaY dom | Spectrogram | V42 | Max of sum of shape presence (1s in binary matrix) within y axis quantiles / total sum of shape presence along all quantiles |
| AreaY std | Spectrogram | V43 | standard error of sum of shape presence (1s in binary matrix) within y axis quantiles. Given value of 0 if NA |
| NULL | NULL | V44 | Removed feature. Occupied by values of 1 |
| AreaMax | Spectrogram | V45 | sum of shape presence (1s) in largest island |
| AreaMax Dom | Spectrogram | V46 | sum of shape presence (1s) in largest island/sum of all shape presence |
| AreaTop3 Dom | Spectrogram | V47 | sum of shape presence (1s) in top 3 largest islands/sum of all shape presence. Given value of 1 if NA |
| Num Shapes | Spectrogram | V48 | number of islands in image |
| BestRho Hough | Spectrogram | V49 | highest scoring Hough line rho |
| BestTheta Hough | Spectrogram | V50 | highest scoring Hough line theta |
| BestSlope Hough | Spectrogram | V51 | highest scoring Hough line slope |
| BestB Hough | Spectrogram | V52 | highest scoring Hough line B |
| NULL | NULL | V53 | no feature, likely typo. Occupied by values of 1 |
| MedRho Hough | Spectrogram | V54 | median scoring Hough line rho of all lines scoring .7 or higher to max score |
| MedTheta Hough | Spectrogram | V55 | highest scoring Hough line theta of all lines scoring .7 or higher to max score |
| MedSlope Hough | Spectrogram | V56 | highest scoring Hough line slope of all lines scoring .7 or higher to max score |
| MedB Hough | Spectrogram | V57 | highest scoring Hough line B of all lines scoring .7 or higher to max score |
| num Goodlines | Spectrogram | V58 | Number of lines scoring .7 or higher to max score |
| xavg | Spectrogram | V59 | x coordinate of shape presence centroid |
| yavg | Spectrogram | V60 | y coordinate of shape presence centroid |
| SwitchesX | Spectrogram | V61 | mean number of times a line along y axis quantiles switches from shape presence to absence and vice versa |
| SwitchesX mean | Spectrogram | V62 | standard error number of times a line along y axis quantiles switches from shape presence to absence and vice versa |
| SwitchesX max | Spectrogram | V63 | max number of times a line along y axis quantiles switches from shape presence to absence and vice versa |
| SwitchesX min | Spectrogram | V64 | minimum number of times a line along y axis quantiles switches from shape presence to absence and vice versa |
| SwitchesY | Spectrogram | V65 | mean number of times a line along x axis quantiles switches from shape presence to absence and vice versa |
| SwitchesY mean | Spectrogram | V66 | standard error number of times a line along x axis quantiles switches from shape presence to absence and vice versa |
| SwitchesY max | Spectrogram | V67 | max number of times a line along x axis quantiles switches from shape presence to absence and vice versa |
| SwitchesY min | Spectrogram | V68 | minimum number of times a line along x axis quantiles switches from shape presence to absence and vice versa |
| Meanslope | Spectrogram | V69 | average slope of all islands. Slope calculated by (average freq at start time of island - average freq at end time of island) / island duration |
| Varslope | Spectrogram | V70 | standard error of slope of all islands. Slope calculated by (average freq at start time of island - average freq at end time of island) / island duration |
| SumCent | Spectrogram | V71 | Sum of all island centroid distances to the slope line. |
| SumCent Abs | Spectrogram | V72 | Sum of absolute value of all island centroid distances to the slope line. |
| meanCent | Spectrogram | V73 | Mean of all island centroid distances to the slope line. |
| meanCent Abs | Spectrogram | V74 | Mean of absolute value of all island centroid distances to the slope line. |
| perconcave | Spectrogram | V75 | Percentage of shapes that are concave, as calculated by negative or positive distance from a shape centroid to its slope |

## Random forest models and ground truth considerations

## Adaptive compare

There are many parameters that allow you to control and fine tune the script (parameters, supplemental). There are too many parameters to optimize automatically, so optimization is best fine tuned by hand.

Building detectors for a new signal requires constructing the detector suite in Raven, as well as an algorithm to relate the MD to one another and filter out likely negatives based on time and frequency criteria. New detectors must be built into the script, but it is compartmentalized such that minimal effort is needed to add support for additional detectors.