Mastor\_detecter: A noise tolerant detection, classification, and review system for archival acoustic data

Daniel Woodrich1

NOAA NMFS AFSC MML Acoustics division

1JISAO, University of Washington. University of Washington  
3737 Brooklyn Ave NE  
Seattle, WA 98195

Table of Contents

[Glossary 2](#_Toc7614854)

[Introduction 3](#_Toc7614855)

[Methods 5](#_Toc7614856)

[Construction and design 5](#_Toc7614857)

[Building and evaluating autodetectors 6](#_Toc7614858)

[Upsweep 6](#_Toc7614859)

[Gunshot 7](#_Toc7614860)

[Holdout data experiment: 7](#_Toc7614861)

[Results 8](#_Toc7614862)

[Performance on labelled data 8](#_Toc7614863)

[Upsweep 10](#_Toc7614864)

[Gunshot 11](#_Toc7614865)

[Holdout data experiment 15](#_Toc7614866)

[Upsweep 15](#_Toc7614867)

[Gunshot 15](#_Toc7614868)

[Discussion 16](#_Toc7614869)

[Performance on labelled data 16](#_Toc7614870)

[Upsweep 16](#_Toc7614871)

[Gunshot 16](#_Toc7614872)

[Holdout data experiment 17](#_Toc7614873)

[Upsweep 17](#_Toc7614874)

[Gunshot 17](#_Toc7614875)

[Future directions 18](#_Toc7614876)

[References 19](#_Toc7614877)

[Supplemental 21](#_Toc7614878)

[1. Raven and custom algorithm 21](#_Toc7614879)

[2. Pre-whitening 23](#_Toc7614880)

[3. Feature extraction 23](#_Toc7614881)

[Hough slope 24](#_Toc7614882)

[Island slope 24](#_Toc7614883)

[Switches 24](#_Toc7614884)

[4. Detector parameters 28](#_Toc7614885)

[5. Adaptive\_compare 29](#_Toc7614886)

# Glossary

|  |  |
| --- | --- |
| Adaptive compare (A\_c) | Filter that uses probability values to compare, and then combine or eliminate detections within a certain time and probability threshold. Helps reduce MBs |
| AUC | Area under the ROC Curve. Represents the likelihood of correct classification along probability thresholds |
| AURAL | Autonomous Underwater Recorders for Acoustic Listening. Manufactured by Multi-Électronique in Rimouski Canada. |
| 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 |
| Custom algorithm | User designed algorithm that uses the time and frequency components of MDs among other parameters and rules to filter likely FPs. |
| Data segments (DS) | Sections of high graded time waveform data that are used to build and analyze performance of a detector |
| 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. |
| 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 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 I seem to be using it the same way as flightcallr paper. |
| Island | Isolated region of 1s (shape presence) spectrogram representation of sound after binarizing image. |
| LFDCS |  |
| Masking | Effect of higher signal noise obscuring desired signal when overlaid on same time and frequency. |
| Mastor\_detecter | Name of described detection and classification system. |
| MD Run | A sequence of minidetections that may represent an autodetection when assessed with the custom algorithm. |
| Meantime | Average time of a time by frequency box. = (start time + end time)/2 |
| Minidetection (MD) | Output from the BLED suites. |
| 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. |
| Pitch tracker (PT) | Part of the detector that refers to the combined output of the Raven BLED suites and custom algorithm |
| Random Forest (RF) | Bagged decision tree based machine learning approach. Robust to overfitting, light computing requirements, excellent performance, and doesn’t require data standardization |
| Rho | The length of the perpendicular line to the normal drawn from the origin Along with Theta used as a factor in an alternative expression of a line that avoids infinite values. |
| Shape presence | Refers to the area expressed as 1s in a binarized spectrogram |
| Signal of interest (SOI) | Signal in time wave that is the target for the detector. |
| Theta | The angle of the perpendicular line to the normal drawn from the origin. Along with Rho used as a factor in an alternative expression of a line that avoids infinite values. |
| 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. |
| True positive rate (TPR) | Also known as sensitivity. Defined as TP/(TP+FN) |

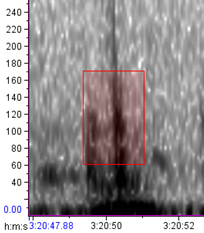
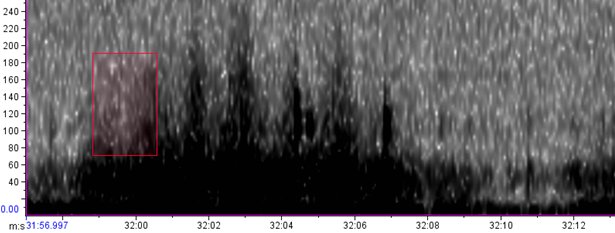
# Introduction

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 (Basset, Thomson, Dahl, & Polagye 2014), is a common source of interference for low frequency acoustic analysis (Strasberg & Taylor 1979, Mceachern & Lauchle 1995). Mooring self-noise is intermittent and highly polymorphic, and can have high amplitude due to receiver proximity to the source signal which can have a masking effect on the SOI or corrupt background noise calculation. In current environments where flow noise is audible as turbulent “gusts” that extend upwards to 200 Hz, pitch tracking detectors can persistently mistake the FM boundary of the gust as desired signal and additionally mistake a variety of noise produced by mechanical disturbance of the instrument (chain rattling, squeeks, bumps, etc.) as desired signal resulting in high FPR (Figure 1). Amplitude detectors will trigger at any form of mooring self-noise that exceeds their threshold, similarly resulting in increased FPR.

**Figure 1**. Examples of flow noise **(A)** and disturbance noise **(B)** triggering FPs for the upsweep PT.

**B.**

**A.**



AURAL recorders in the arctic region suffer from a high degree of self-noise in their low frequency range, making implementation of autodetectors challenging due to increased FPR in analyses of SOI in that range. Popular autodetectors implemented in the LFDCS (Baumgartner & Mussoline 2011) and Ishmael (Mellinger 2002) are designed to be lean computationally and don’t offer learning schemes to distinguish mooring self-noise from SOI on a region to region basis. High FPR hurts the implementation of an autodetector as it either increases the time of an analyst to verify detections or reduces TPR due to adjusting the TPR threshold to optimize the TPR/FPR tradeoff for time efficient analysis.

Mastor\_detecter was created as a solution to the high FPR issue encountered with existing autodetectors on the high self-noise recordings from arctic AURALs. While the idea to model true positives from a band limited energy detector against negatives or null data is not new (flightcallr, Mellinger 2004), Mastor\_detecter is to our knowledge the first detector that combines a pitch tracker and machine learning classifier to filter out reoccurring sources of false positives. This ability to reduce the FPR of analyzed data 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 the features available in the labelled dataset, 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 than arctic AURALs.

Mastor\_detecter has similar use limitations to existing pitch tracking detectors- it can be used to identify stereotyped, distinct calls that ideally feature FM, but is vulnerable to interference that will mask or obscure the shape of calls. This vulnerability is most prominent in the Raven BLEDs, where there is a built in background noise calculation that cannot discriminate against self-noise resulting in masking from lowered sensitivity.

At present there are no built in variables to represent species presence or temporal 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, I 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 for your SOI must be coded in R language. Due to its implementation in R Mastor\_detecter will never be an executable program, but rather a flexible infrastructure that can be customized to various call types and easily reviewed in Raven 1.5. 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.

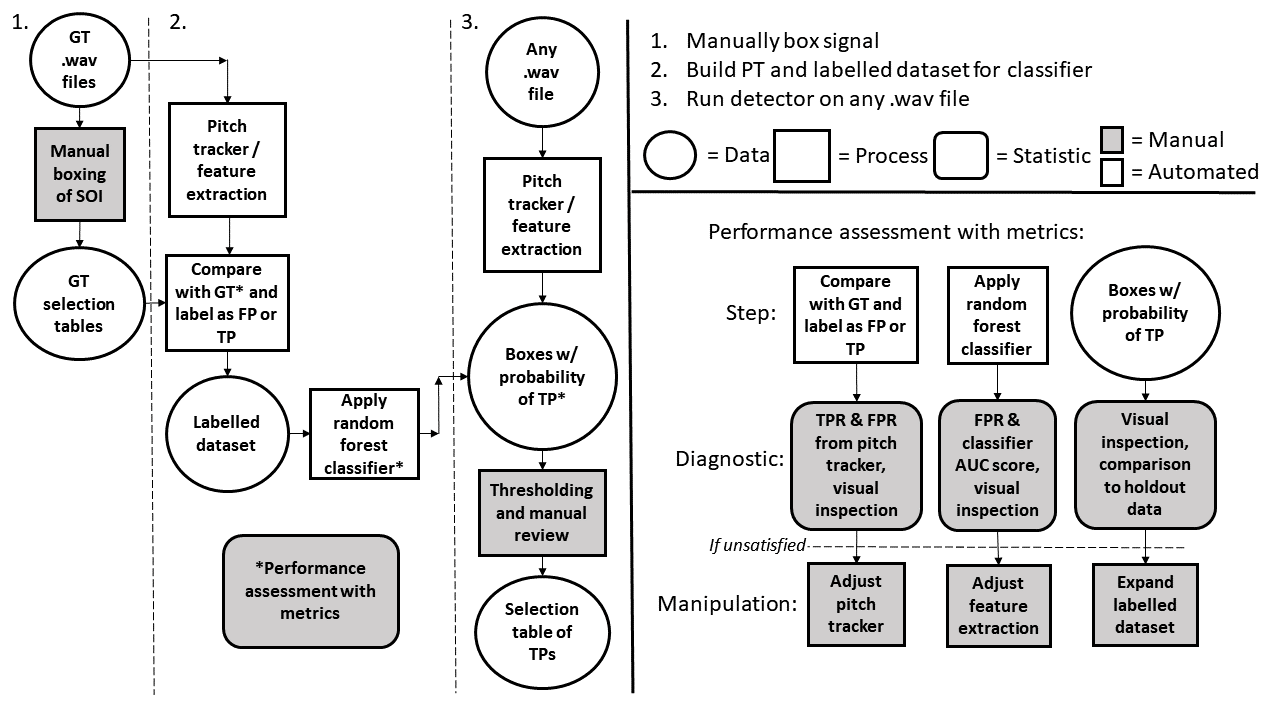
I have successfully applied Mastor\_detecter to right whale upcalls and gunshots and demonstrate results for these applications, as well as provide guidance for its implementation to any distinct, stereotyped signal of interest. With Mastor\_detecter still being in an unrefined state this report is not meant to serve as a user guide, but an explanation of the underlying processes.

# Methods

## Construction and design

I 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 the R package ‘Rraven’ which allowed us to call this algorithm from API, I decided to continue developing around Raven and the R language (Araya-Salas 2017). 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 must 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 initial 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 Ross & Allen (2014) was implemented to help weed out FPs. Image analysis was later introduced to improve the classification performance of the RF models (Supplemental 3: Feature extraction). Parallelization is implemented while running the pitch tracking algorithm, extracting features, and building the RF models, and is highly recommended to improve run time. The infrastructure for these parallelized sections was developed with R package ‘foreach’ which is in theory portable to Azure to enable cloud computing (Microsoft & Weston 2017).

Much of the infrastructure was reworked for compatibility with simultaneous multi-call detection in a single detector run. 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, binary detection for each call of interest 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. The capability for multiclass detection was scaled back as our timeline prioritized effective single species detection during development, but the looping structure is largely intact for this capability.

**Figure 2.** Generalized workflow of Mastor\_detecter. 

## Building and evaluating autodetectors

### Upsweep

An upsweep detector was developed using the Mastor\_detecter workflow (Figure 1). Data segments from 12 different AURAL deployments that had high presence of RW were chosen to build the pitch tracker and train the RF models (Table 1). Eight of these DS were hand-boxed to evaluate the performance of the PT, and the rest (n=5) were machine assisted for ease of analysis once satisfactory performance was seen in the hand-boxed DS. 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 performance metrics on TPR and FPR.

Once satisfactory results were generated (TABLE#) from the newly designed pitch tracker, I 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 I had built a pitch tracker and generated the labelled data, the detector can use the features and labels from these data to build a classification scheme. The detector does this by randomly partitioning the data into training (75%) and test (25%) sets (when running the detector on unlabeled data, the test set is replaced in the structure with the entire unlabeled dataset). The classes in the training set are balanced by downsampling the larger class to the size of the smaller, the smaller typically the positive class, to correct for the propensity of the Breiman random forest algorithm to deprioritize the precision of a rare class (Chen, Liaw, and Breiman 2004). This reduced training set is compared to the test set to generate random forest models according to the CV specified which each determine probabilities for the test set, which are then averaged. To test the performance of the complete detector, I trained models recursively on the labelled set (TABLE #) and evaluated results using TPR, FPR, graphically, and by visual inspection of the selection tables in Raven. Manipulations performed here to improve the model included adding features new features to the labelled data and adding post hoc filters to clean up issues with multiple boxes being assigned to single calls (Adaptive\_compare).

I used the following set of parameters for the upcall detector in the analysis presented: Decimate = ‘y’, decimationFactor = 16, whiten = ‘n’, TPRthresh = 0.95, grpsize = 3, allowedZeros = 4, detskip = 4, groupInt = 0.45, Maxdur = 3.5, Mindur = 0.2, timediffself = 1.25, probdist = 0.2, ImgTresh = 65 (Supplemental: Detector parameters)

### Gunshot

The same process to build and evaluate the detector 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 labelled data and analyzing performance. Fewer DS (n=4) were hand-boxed due to the apparent success of the PT to identify signal, and the difficulty associated with hand-boxing the frequent and often muddled gunshot signal that densely occupied each DS.

Given the large structural difference between upsweeps and gunshots, I constructed a new BLED suite and custom algorithm to detect these sounds. The algorithm looked for ‘stacked’ MDs (MDs which have an identical or very close start time) to identify broadband signal, and evaluated runs based on the frequency range of the broadband sound and possible propagation ‘downsweep’ effect. I incorporated an extra variable ‘timesepGS’ for the gunshot algorithm, which is an arbitrary value that serves as a coefficient to help shape the function that compares the time difference of MDs as frequency rank decreases. To preserve the ability to isolate double gunshot (FIGURE#), I adjusted the Adaptive\_compare time threshold to 0.15 while keeping the probability threshold at 2.

I used the following set of parameters for the gunshot detector in the analysis presented: Decimate = ‘y’, decimationFactor = 8, whiten = ‘n’, TPRthresh = 0.95, grpsize = 4, allowedZeros = 2, detskip = 7, groupInt = 0.35, Maxdur = 3.5, Mindur = 0, timesepGS = 1.2, timediffself = 0.15, probdist = 0.2, ImgTresh = 90 (Supplemental: Detector parameters)

## Holdout data experiment:

Once I found detector performance to be sufficient as measured by the ability of the classifier to effectively generalize to the DS provided, I applied the detector to holdout DS to test the generalization of the classifier with a limited labelled dataset against a larger data population. In this experiment, a labelled set was used to generate a classifier that was tested against holdout DS, which was then incorporated into the labelled set itself until all DS were included. Since the effect of growing the labelled dataset on overall performance was obscured by the amount of variation in performance for each individual DS added I conducted 30 trials with randomly selected DS to obtain average classifier performance at each iteration of comparison to holdout data.

**Table 1.** Ground truthing effort for gunshot and upcalls. ‘Machine assisted = n’ means that data was boxed manually, and ‘Machine assisted = pt’ 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 ‘pt’ 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 | Randomly selected |
| Upcall | M2 | 194 | 53.61 | 5.88 | 06/22/2015 | 08/09/2015 | n | n |
| Upcall | M4 | 179 | 100.00 | 10.08 | 12/02/2014 | 09/20/2015 | n | n |
| Upcall | BS3 | 217 | 100.00 | 12.10 | 08/11/2012 | 09/12/2013 | n | n |
| Upcall | M4 | 304 | 100.00 | 16.85 | 09/18/2013 | 10/04/2014 | n | n |
| Upcall | M2 | 325 | 53.85 | 9.83 | 05/17/2016 | 09/08/2016 | n | n |
| Upcall | M2 | 62 | 100.00 | 3.44 | 10/01/2015 | 12/24/2015 | n | n |
| Upcall | BS3 | 443 | 36.12 | 9.00 | 10/20/2014 | 06/26/2015 | n | n |
| Upcall | BS1 | 158 | 100.00 | 9.06 | 09/26/2016 | 11/25/2016 | pt | y |
| Upcall | M2 | 58 | 100.00 | 3.15 | 07/03/2013 | 07/18/2013 | pt | y |
| Upcall | M5 | 13 | 100.00 | 0.75 | 09/29/2016 | 12/08/2016 | pt | y |
| Upcall | M4 | 540 | 21.48 | 6.56 | 08/15/2016 | 08/26/2016 | pt | y |
| Upcall | BS2 | 86 | 100.00 | 4.71 | 07/21/2015 | 09/25/2015 | pt | y |
| Upcall |  |  |  | 91.42 |  |  |  |  |
| Gunshot | BS3 | 1452 | 3.10 | 2.50 | 11/06/2015 | 11/14/2015 | n | n |
| Gunshot | M2 | 455 | 10.11 | 2.65 | 06/19/2012 | 07/20/2012 | n | n |
| Gunshot | BS3 | 583 | 12.18 | 3.96 | 11/05/2014 | 12/04/2014 | n | n |
| Gunshot | M4 | 289 | 30.45 | 4.79 | 08/15/2014 | 09/11/2014 | n | n |
| Gunshot | BS3 | 3764 | 6.64 | 13.77 | 08/11/2012 | 08/26/2012 | pt | n |
| Gunshot | BS3 | 3764 | 1.17 | 2.38 | 10/07/2012 | 10/11/2012 | pt | n |
| Gunshot | BS3 | 583 | 10.29 | 3.40 | 12/21/2014 | 12/27/2014 | pt | y |
| Gunshot | BS2 | 391 | 23.02 | 3.98 | 10/07/2015 | 10/09/2015 | pt | y |
| Gunshot | BS3 | 576 | 16.15 | 5.35 | 12/03/2016 | 07/01/2017 | pt | y |
| Gunshot | M2 | 1383 | 4.19 | 3.27 | 09/20/2012 | 09/23/2012 | pt | y |
| Gunshot | M4 | 166 | 44.58 | 4.15 | 04/02/2015 | 08/30/2015 | pt | y |
| Gunshot |  |  |  | 50.19 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

# Results

## Performance on labelled data

**Table 2.** Performance of the upsweep and gunshot detectors 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 that were not registered from calculating the accuracy of the PT.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Upsweep DS: Hand boxed | PT TPR | RF TPR | Total TPR | PT FPR | Total FPR | MB % | OB % | AUC |
| BS15\_AU\_02a | 0.98 | 0.98 | 0.96 | 0.82 | 0.55 | 0.37 | 0.00 | 0.97 |
| BS14\_AU\_04 | 0.94 | 0.96 | 0.90 | 0.91 | 0.71 | 0.48 | 0.00 | 0.97 |
| AW12\_AU\_BS3 | 0.94 | 0.97 | 0.91 | 0.85 | 0.61 | 0.00 | 0.00 | 0.97 |
| BS13\_AU\_04 | 0.92 | 0.94 | 0.86 | 0.88 | 0.58 | 0.07 | 0.07 | 0.96 |
| BS16\_AU\_02a | 0.96 | 0.97 | 0.93 | 0.85 | 0.58 | 0.00 | 0.00 | 0.97 |
| BS15\_AU\_02b | 0.89 | 0.92 | 0.82 | 0.94 | 0.71 | 0.00 | 0.00 | 0.94 |
| AW14\_AU\_BS3 | 0.93 | 0.96 | 0.90 | 0.90 | 0.66 | 0.91 | 0.00 | 0.97 |
| Hand boxed all | 0.93 | 0.95 | 0.89 | 0.88 | 0.62 | 0.22 | 0.02 |  |
| Upsweep DS: machine assisted |  |  |  |  |  |  |  |  |
| AL16\_AU\_BS1 |  | 0.94 | ≤0.94 | 0.93 | 0.72 |  |  | 0.96 |
| BS13\_AU\_02a |  | 0.96 | ≤0.96 | 0.76 | 0.52 |  |  | 0.96 |
| BS16\_AU\_05 |  | 0.93 | ≤0.93 | 0.92 | 0.73 |  |  | 0.94 |
| BS15\_AU\_04 |  | 0.94 | ≤0.94 | 0.83 | 0.56 |  |  | 0.95 |
| AW14\_AU\_BS2 |  | 0.87 | ≤0.89 | 0.86 | 0.60 |  |  | 0.93 |
| Machine assisted all |  | 0.93 | ≤0.93 | 0.88 | 0.63 |  |  |  |
| Upsweep all |  |  |  | 0.88 | 0.62 |  |  | 0.96 |
| Gunshot DS:  Hand boxed | PT TPR | RF TPR | Total TPR | PT FPR | Total FPR | MB % | OB % | AUC |
| AW15\_AU\_BS3 | 0.97 | 0.97 | 0.93 | 0.67 | 0.35 | 6.67 | 0.43 | 0.95 |
| AW14\_AU\_BS3 1 | 0.93 | 0.97 | 0.91 | 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 | 0.94 | 0.89 | 0.81 | 0.54 | 19.14 | 0.94 | 0.93 |
| Hand boxed all | 0.95 | 0.97 | 0.91 | 0.68 | 0.39 | 13.25 | 0.51 |  |
| Gunshot DS: Machine assisted |  |  |  |  |  |  |  |  |
| AW15\_AU\_BS2 |  | 0.88 | ≤0.88 | 0.86 | 0.43 |  |  | 0.95 |
| AW12\_AU\_BS3 1 |  | 0.97 | ≤0.97 | 0.68 | 0.42 |  |  | 0.95 |
| BS14\_AU\_04 |  | 0.89 | ≤0.89 | 0.79 | 0.47 |  |  | 0.92 |
| AW14\_AU\_BS3 2 |  | 0.90 | ≤0.90 | 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 |  | 0.94 | ≤0.94 | 0.70 | 0.35 |  |  |  |
| Gunshot all |  |  |  | 0.70 | 0.37 |  |  | 0.95 |

**Table 3.** Performance of the upsweep and gunshot detectors on high graded, hand boxed ground truth DS with the Adaptive\_compare filter. The filter reduced the MB% for these DS on both detectors and only slightly raised the OB%.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Upsweep DS: Hand boxed | PT TPR | RF TPR | Total TPR | PT FPR | Total FPR | MB % | OB % | AUC |
| BS15\_AU\_02a | 0.98 | 0.98 | 0.96 | 0.82 | 0.55 | 0.37 | 0.00 | 0.97 |
| BS14\_AU\_04 | 0.94 | 0.96 | 0.90 | 0.91 | 0.71 | 0.00 | 0.00 | 0.97 |
| AW12\_AU\_BS3 | 0.94 | 0.96 | 0.91 | 0.85 | 0.61 | 0.00 | 0.00 | 0.97 |
| BS13\_AU\_04 | 0.92 | 0.94 | 0.87 | 0.88 | 0.58 | 0.00 | 0.14 | 0.96 |
| BS16\_AU\_02a | 0.96 | 0.97 | 0.93 | 0.85 | 0.58 | 0.00 | 0.00 | 0.97 |
| BS15\_AU\_02b | 0.89 | 0.92 | 0.82 | 0.94 | 0.71 | 0.00 | 0.00 | 0.94 |
| AW14\_AU\_BS3 | 0.93 | 0.97 | 0.90 | 0.90 | 0.66 | 0.18 | 0.00 | 0.97 |
| Hand boxed all | 0.93 | 0.96 | 0.89 | 0.88 | 0.63 | 0.05 | 0.05 |  |
| Gunshot DS:  Hand boxed | PT TPR | RF TPR | Total TPR | PT FPR | Total FPR | MB % | OB % | AUC |
| AW15\_AU\_BS3 | 0.97 | 0.96 | 0.93 | 0.67 | 0.34 | 1.22 | 0.46 | 0.95 |
| AW14\_AU\_BS3 1 | 0.93 | 0.97 | 0.91 | 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 | 0.94 | 0.89 | 0.81 | 0.53 | 11.55 | 1.02 | 0.93 |
| Hand boxed all | 0.95 | 0.97 | 0.91 | 0.70 | 0.39 | 5.58 | 0.59 |  |

### Upsweep

The seven DS that had a corresponding hand boxed GT tables were pooled and analyzed to assess accuracy of the PT and of the entire detector. The PT had an accuracy of 0.93 (n=7) for these DS. These DS had an overall TPR of 0.89 (n=7), and an RF classifier TPR of 0.95 (n=7) (Table 2). As the hand boxes did not necessarily line up with the PT generated boxes as with the machine assisted DS, MB% and OB% were calculated as measures of box quality compared to by-hand boxes. MB% for upsweeps was 0.22% (n=7), meaning 0.22% of TPs were multiboxes (redundant relative to the GT), and this was reduced by a factor of ~4 to 0.05% (n=7) with the addition of the Adaptive\_compare filter (Supplemental 5: Adaptive\_compare). The OB% was 0.02% (n=7), meaning 0.22% of the TPs were overboxes (contained multiple true detections relative to the GT), and this increased by a factor of ~2 to 0.05% (n=7) with the addition of the Adaptive\_compare filter (Table 3).

The remaining machine assisted DS were not able to have 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 component. The machine assisted DS had an average RF classifier TPR of 0.93 (n=5) giving these DS an upper bound of 0.95 (n=5) for overall TPR. At a PT TPR of 0.93 such as seen in the hand boxed DS, this would represent an average total TPR of 0.86 for the machine assisted DS (Table 2).

Statistics that did not involve the number of FNs from the PT were able to be compared across hand and machine assisted DS. The FPR after implementing the PT is 0.88 (n=12), and this is reduced to 0.62 (n=12) with the RF classifier. There were a wide range of differences in FPR between DS due to the prevalence of mooring self-noise, which can be visualized in FIG#. The FPR can also be interpreted as a ratio: for every 3 TPs in the final detector output at an FPR of 0.62, expect 4-5 FPs on average for these DS (Table 2).

The total AUC score across all DS is 0.96 (n=12) (Table 2). Various features were effective in reducing the Gini coefficient for the random forest classifier, but the most important features all represented variations on measurements of slope: MedSlope Hough (V56), BestTheta Hough (V50), BestSlope Hough (V51), MedTheta Hough (V55), Meanslope (V69) and freqrange (freqrange) (Figure 3).

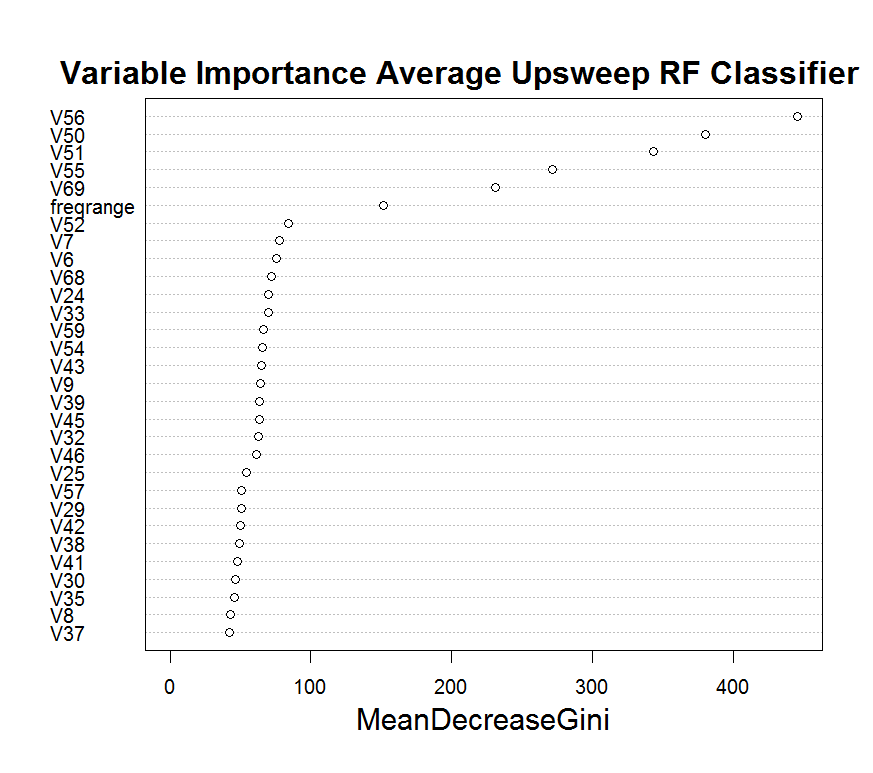
### Gunshot

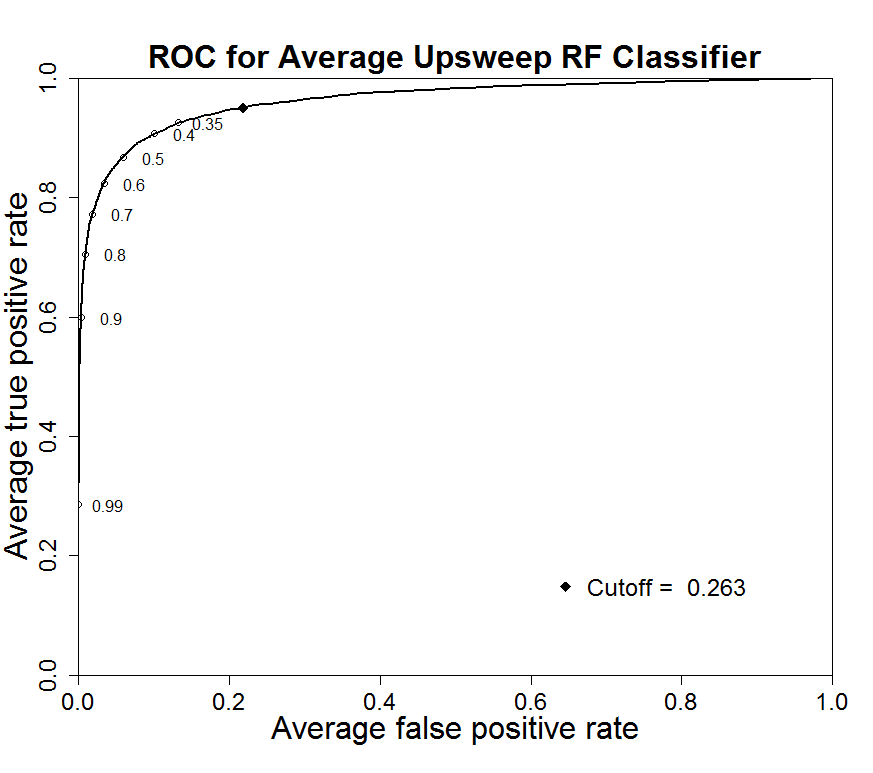
The PT has an accuracy of 0.95 (n=4) for the hand-boxed DS. These DS had an overall TPR of 0.92 (n=4), and an RF classifier TPR of 0.97 (n=4). MB% occurred at a rate of 13.25 (n=4), with higher incidence at BS12\_AU\_02a and BS13\_AU\_04, and OB% occurs at a very low rate throughout. MB% was reduced by a factor of ~2.5 by the Adaptive\_compare filter to 5.58% (n=4), and the OB% was increased fractionally by this filter.

The machine assisted DS had an average RF classifier TPR of 0.95 (n=7) giving these DS an upper bound of 0.95 for overall TPR. At a PT TPR of 0.95 such as seen in the hand boxed DS, this would represent an average total TPR of 0.92 for the machine assisted DS.

The FPR after implementing the PT is 0.7 (n=11), and this is later reduced to 0.37 (n=11) with the RF classifier. Differences in FPR between DS can be visualized in FIG#. 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 for these DS.

The total AUC score across all DS is 0.95 (n=11). A variety of features were effective in reducing the Gini coefficient for the random forest classifier. The most important features by this measure were freqrange (freqrange), MedSlope Hough (V56), SwitchesY max (V67), autoc se (V8), SwitchesY mean (V65) and Total entropy (V26).

**Figure 3. (A,C)** ROC curve of the random forest classifier for gunshot **(A)** and upsweep **(C)** classifier. Probability cutoff values along the curve are labelled. The diagonal line represents the performance of a random classifier. **(B,D)** Mean decrease in Gini coefficient for features included in the RF classifier for gunshot **(B)** and upsweep **(D)**. More informative variables correspond to a larger decrease in Gini coefficient. Feature names and description correspond to V# listed in table x.

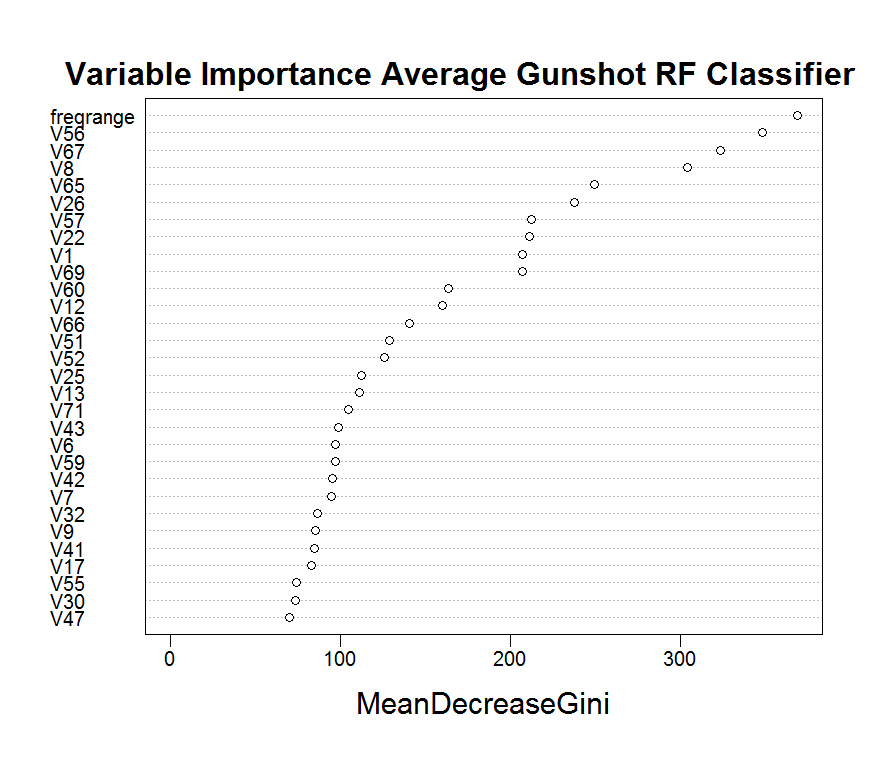


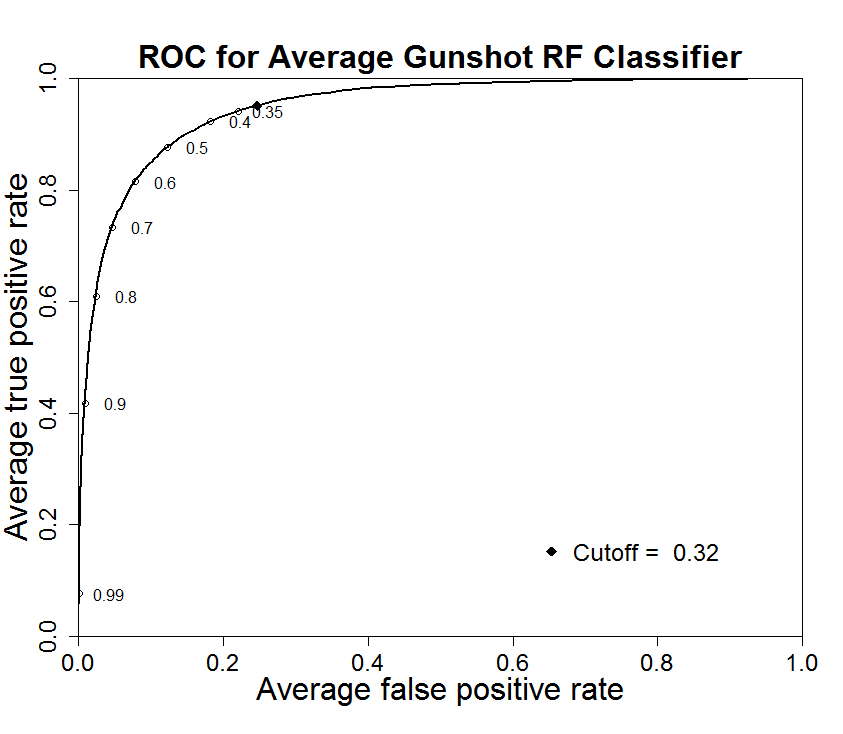
**B.**

**A.**

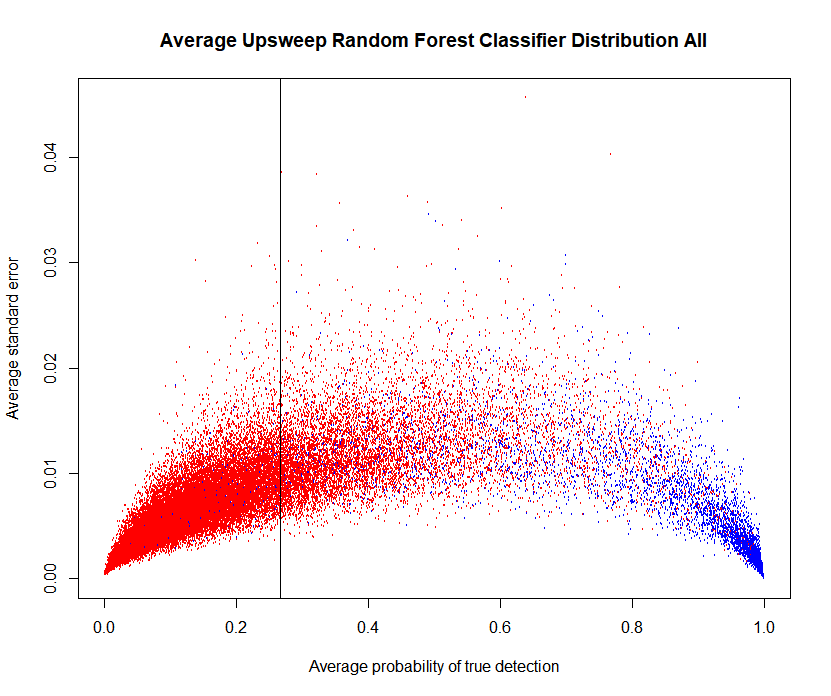
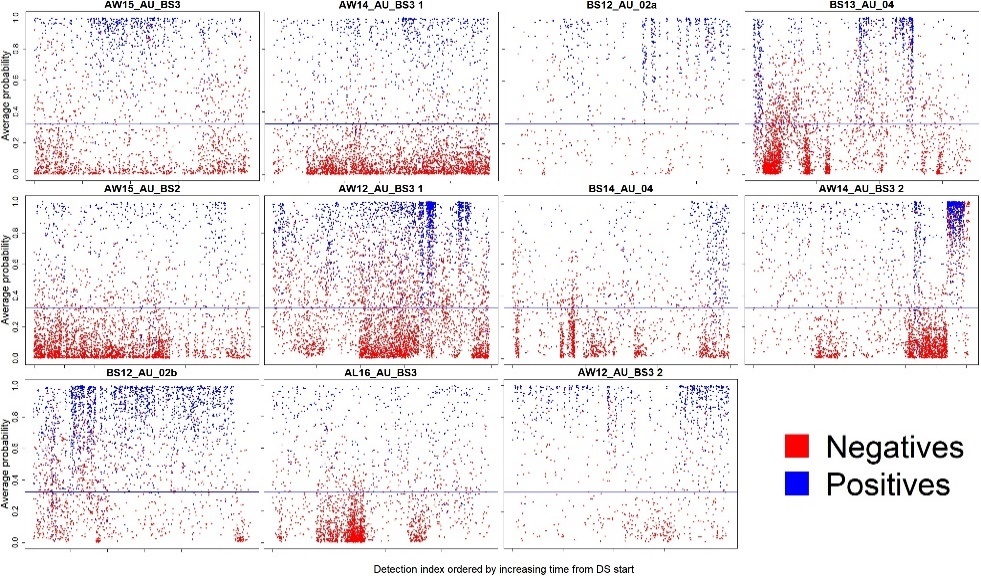
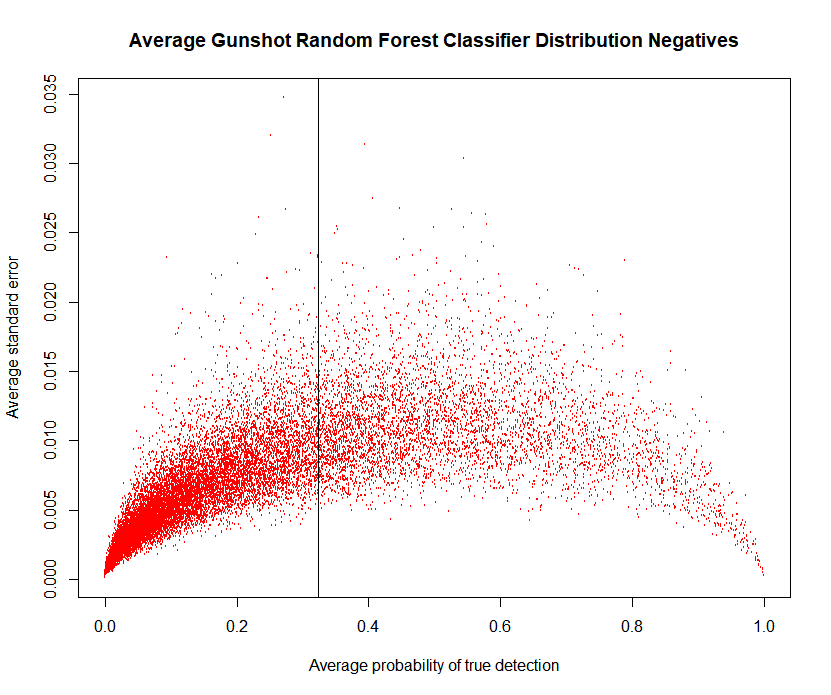
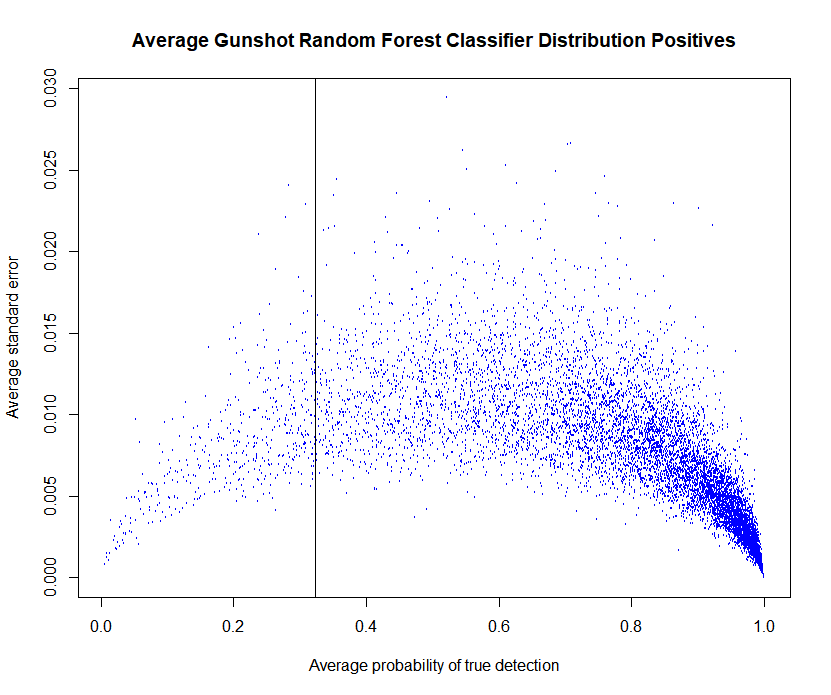
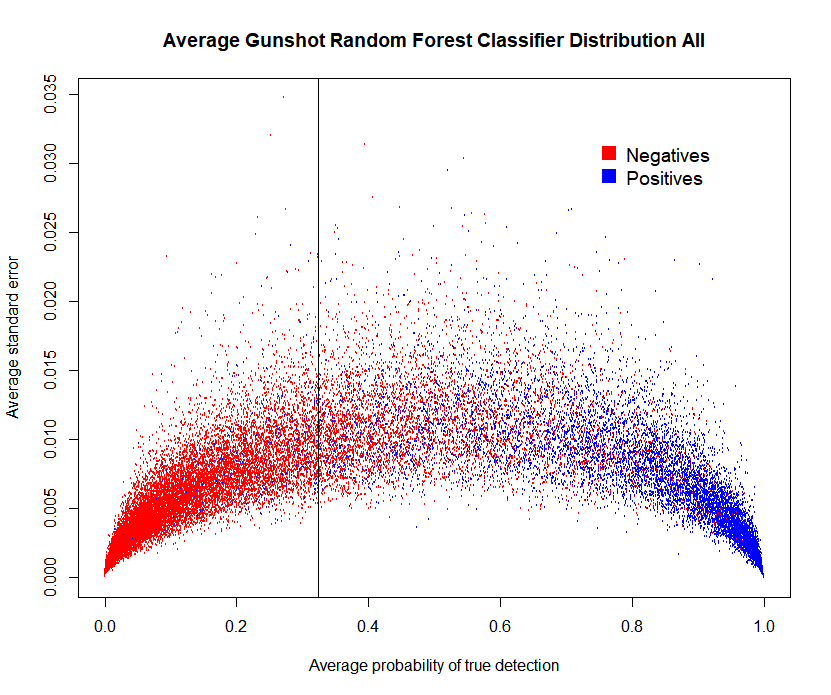
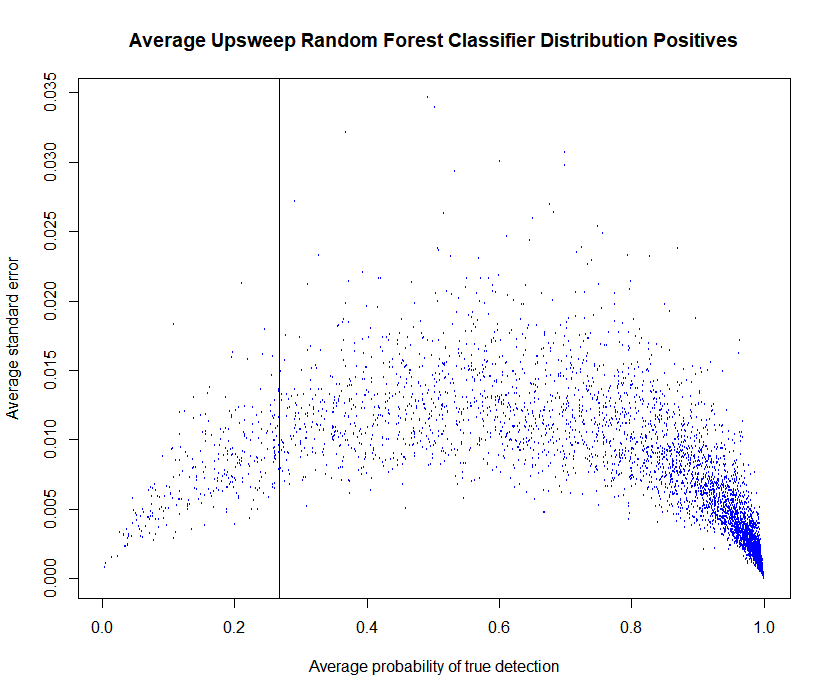
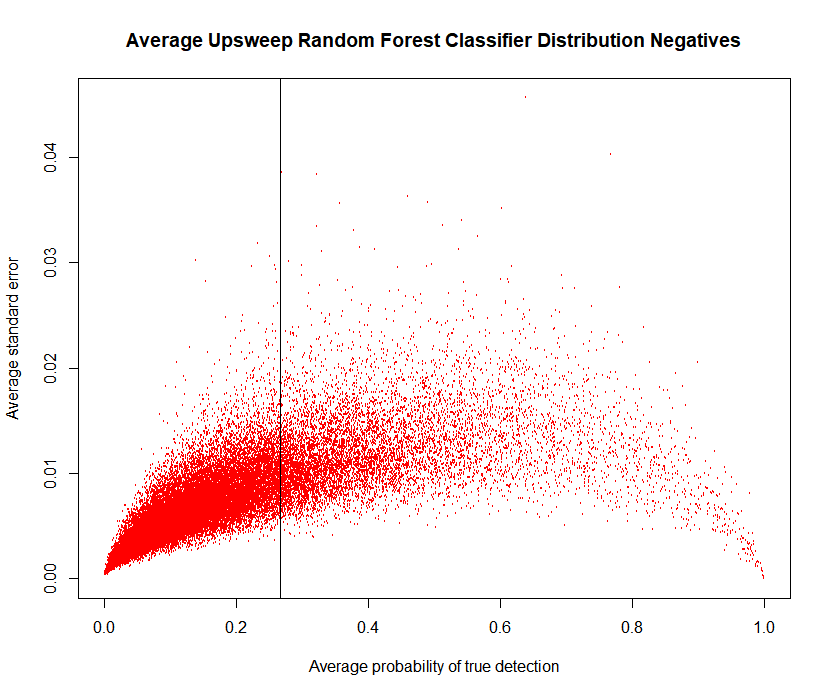
**D.**

**C.**





**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. Autodetections and cutoff value shown for upsweep **(A)** and gunshot detector **(B)**



**Threshold: x = .27**

**Gunshot DS**

**Upsweep DS**

**B.**

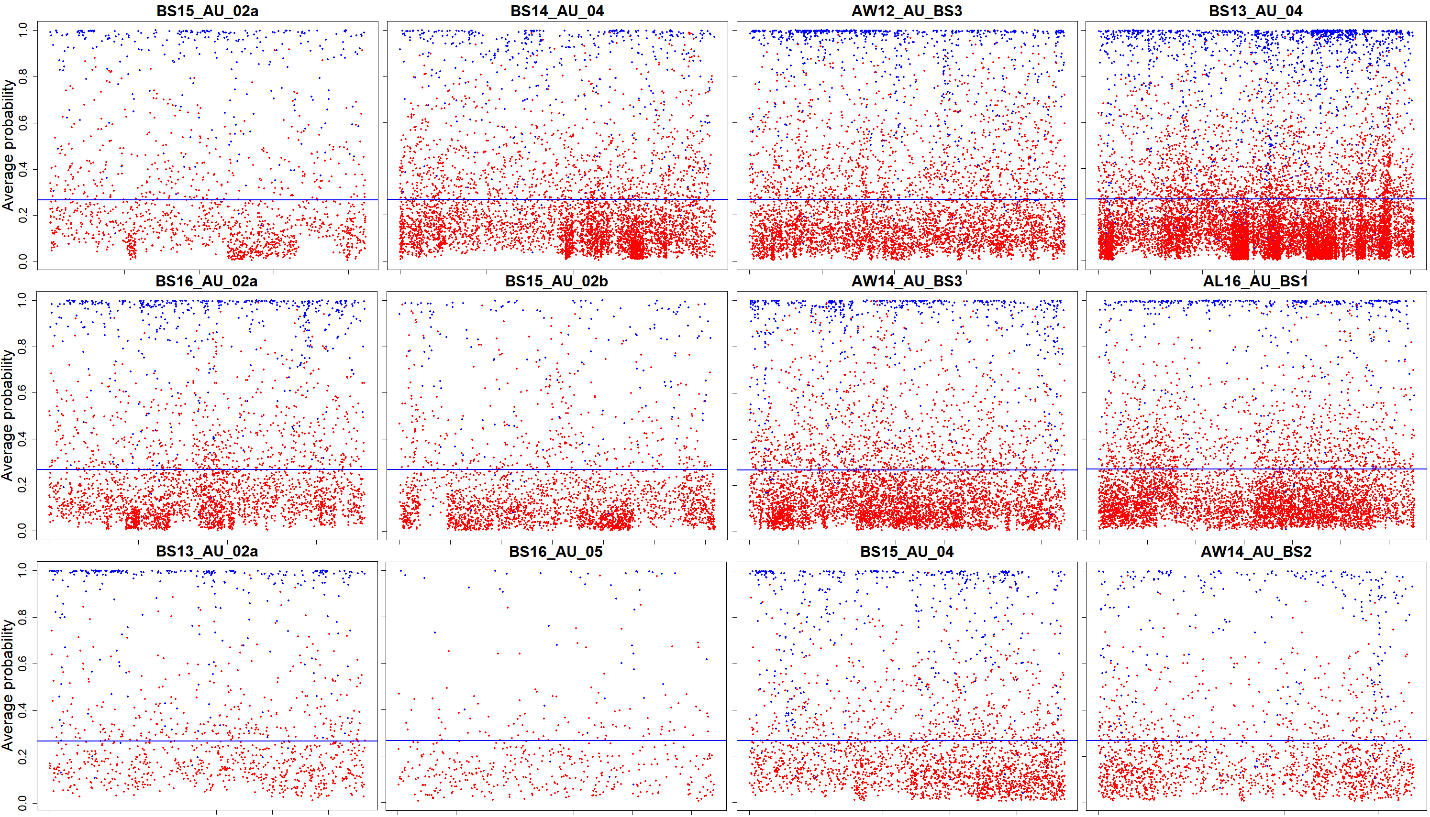
**A.**

**Threshold: x = .32**

**Figure x.** Probability over index arranged consecutively for all DS included in the upsweep **(A)** and gunshot **(B)** detectors 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 for each detector.

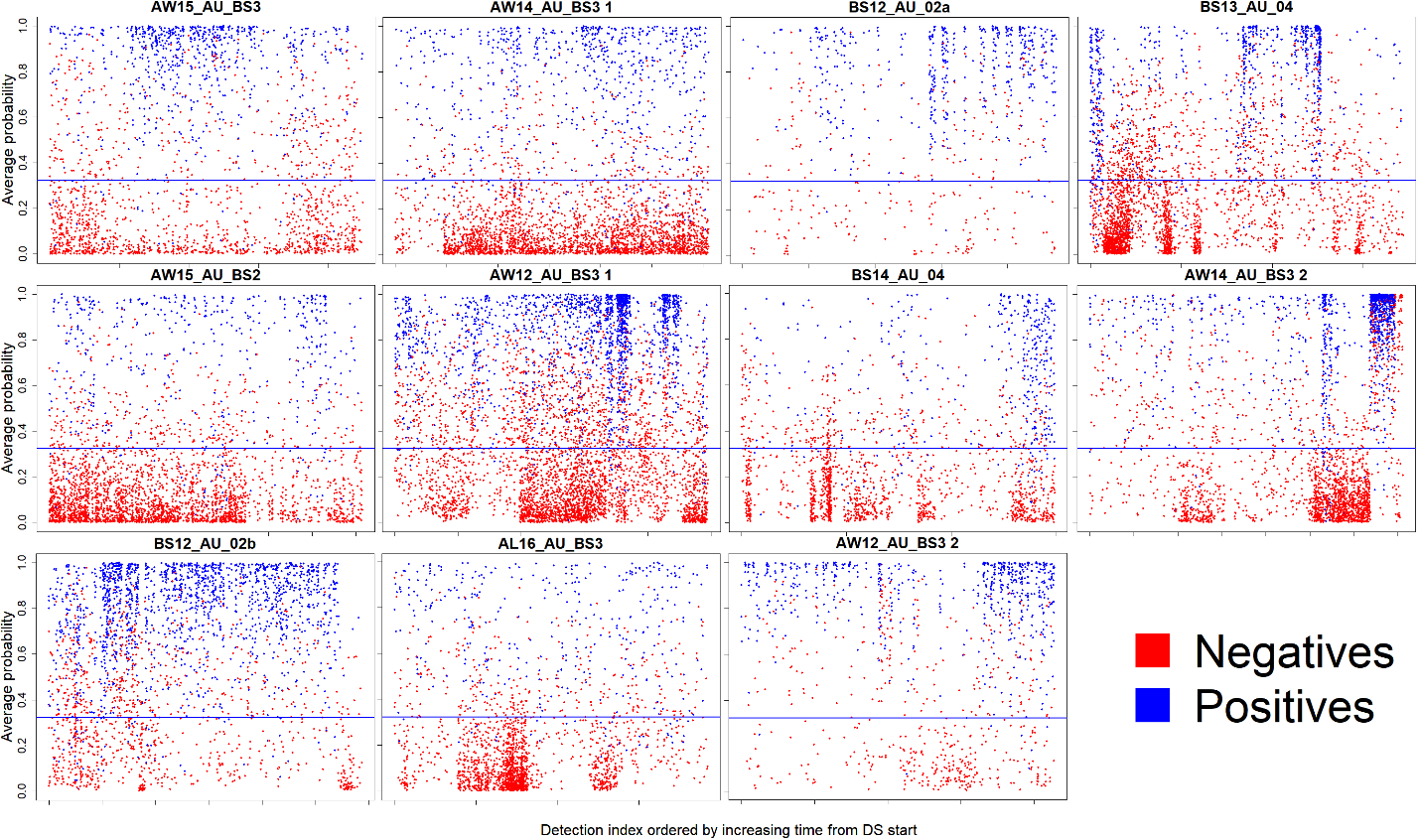
**Upsweep DS**

**A.**



**Gunshot DS**

**B.**



Autodetection index (in consecutive order from start of DS, arbitrary)

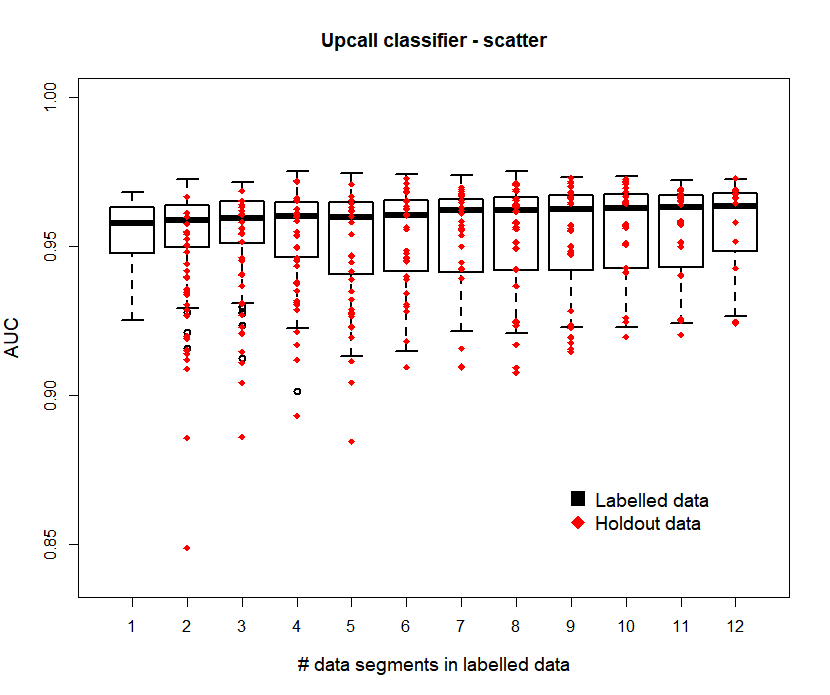
## Holdout data experiment

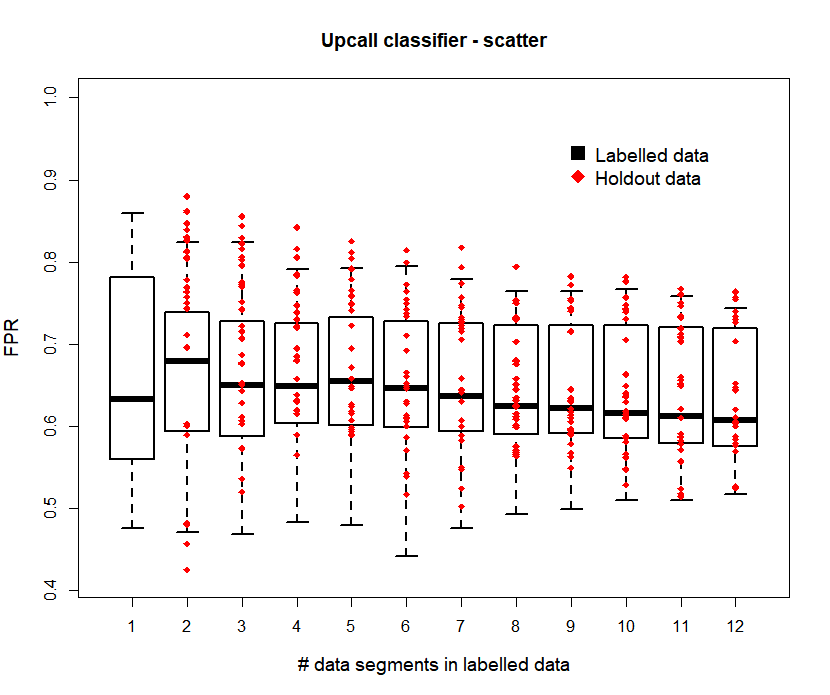
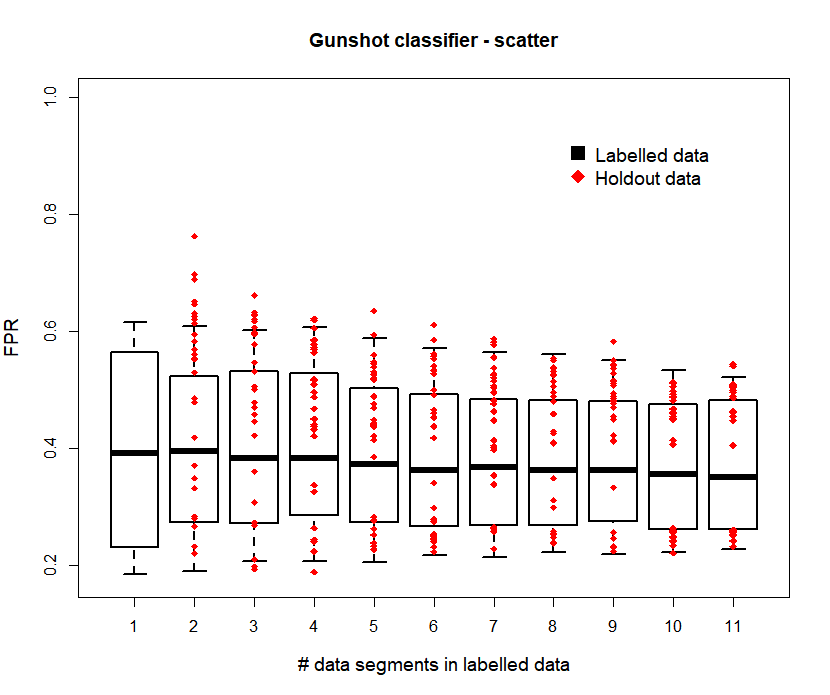
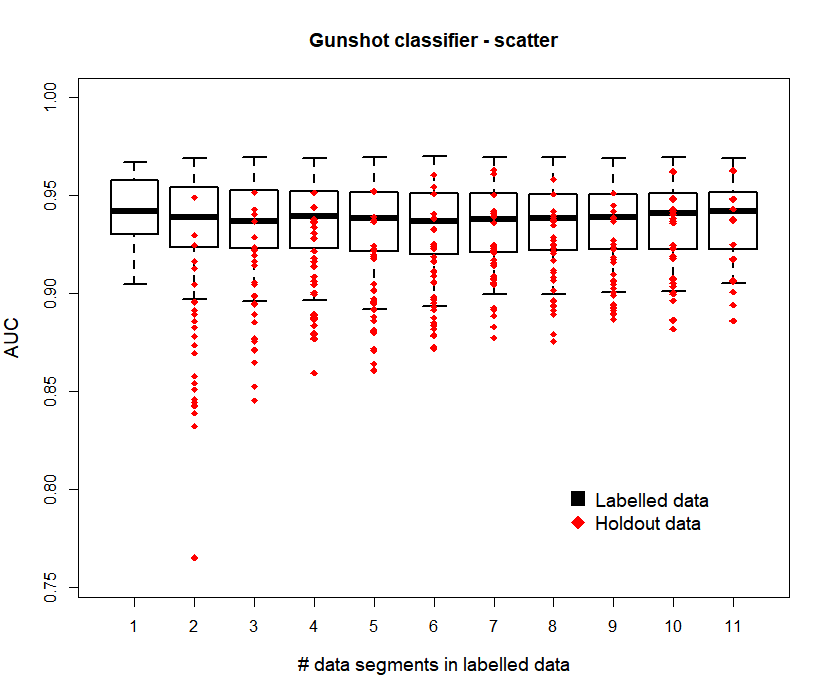
### Upsweep

The results are presented graphically for this experiment in (FIG#). The upsweep classifier appears to have a gradual learning curve throughout, and seems to achieve neutrality with the AUC scores and FPR of the labeled data at n = 11 & n = 12.

### 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). While large improvement are seen in both FPR and AUC with the addition of new DS to the labelled data at n = 2 & n = 3, performance on holdout data steadily increases but does not appear to achieve neutrality with the labelled set.

**Figure x.** Comparison of classifier performance on the FPR **(A,C)** and AUC **(B,D)** score DS in the labelled and holdout sets for gunshot **(A,B)** and upsweep **(C,D)**. A total of 30 trials were performed for each experiment on each call type classifier. **(A,B)** **(C,D)**



**D.**

**C.**

**D.**

**A.**

**B.**

# Discussion

## Performance on labelled data

### Upsweep

The upsweep detector demonstrates good TPR for the PT and the classifier. The FPR for the PT alone would likely be prohibitively high for analysis (0.88: ~1 TP for every 9 FPs), but this is reduced to a feasible amount by the effective RF classifier. The effectiveness of the RF classifier is likely inflated due to the less discriminatory tuning of the PT- the inclusion of more options that have less subtle distinctions allow the classifier to correctly categorize ‘easier’ choices. The high PT FPR and position of the cutoff value above the ROC inflection point (Habibzadeh, Habibzadeh, & Yadollahie 2016) suggest that the upsweep detector is more weighed towards maximizing TPR than optimal at its current parameters. Increased discrimination from the PT or lowering of the TPR threshold could balance the performance towards lower FPR while lowering TPR. The decision of how much to tune the PT and where to set the TPR threshold is dependent on the needs of the analysis around the accuracy and manual review time tradeoff.

The upsweep classifier shows excellent discrimination among the highest quality signal. The ROC curve shows that at a probability cutoff of 0.9, 60% of upsweep TPs would be included along with a fractional % of FPs, which would allow for an extraction of high quality upsweeps from mooring self-noise if useful for an analysis. This also could allow for effective lower TPR analyses such as daily upsweep presence. However, high TPR analyses involving calling rate or behavior would likely suffer from more FPs than a more evenly discriminating classifier such as for the gunshots.

Although the AUC score appears excellent for the upsweep classifier, it may be inflated by the less discriminatory PT, and also shows a notable lack of diversity in the most effective features for reduction Gini coefficient. This suggests possible weakness to the classifier, given that slope (or sweep-rate) is known to be a highly variable characteristic of NARW upcalls (McDonald & Moore 2002): the classifier could be weak if applied to bout of upcalls where slope was outside the range of what was considered acceptable by the classifier (FIG#).

### Gunshot

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 gunshots which often places these sounds outside of the typical mooring self-noise frequency range, which typically tops out under 500 Hz. 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. However, including this feature does mean that the classifier will discriminate against lower frequency gunshot calls and suffer worse performance relative to that positive class. 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 accurate PT will give the classifier a higher proportion of difficult choices than a weaker one, reducing the probability of correct classification for each detection.

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 robust to call plasticity. 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 such as daily presence.

MB% is a larger issue with gunshots given that the pitch tracker is optimized to be sensitive to double gunshot. The Adaptive\_compare filter was effective in reducing the incidence of MB% while keeping the OB% nearly constant, increasing the average quality of each box as a representation of a single signal. Depending on the need of the analysis to differentiate closely grouped signal, optimization for MB% and OB% is possible by manipulation of the Adaptive\_compare filter, the Raven BLED parameters, or the custom algorithm.

## Holdout data experiment

One of the most important assumptions to ensure good performance from the detector is that the labelled dataset is an adequate sample to represent the population of your SOI. The classifier treats an unlabeled dataset the same way it treats the test data for the performance evaluation on labelled data: quite literally, as the internal structure is reused to apply a model generated from randomly selected training data not to the randomly selected test data, but to the entire unlabeled dataset. Because of this, the performance metrics shouldn’t be expected to match perfectly to any individual dataset given that individual datasets have specific attributes that aren’t consistent over the entire population, which is ideally represented in your labelled data sample. This experiment was designed to visual the effect on performance of a general addition of data to the labelled dataset, and shouldn’t be interpreted as a guarantee of performance on any given DS.

### Upsweep

The AUC score and FPR of the holdout data appeared to achieve neutral performance to the labeled data, suggesting that the number (n = 12) and variety (FIG# gt data status) of DS were sufficient to provide consistent results on unseen data. Given the low sample size of DS, it is possible that DS of heteromorphic data within the high graded upcall population were coincidentally not represented in the labelled dataset and would be misclassified once the detector were applied. Lacking a reliable estimate on the polymorphy of gunshot and mooring self-noise signal in our entire data population, it is not feasible to estimate the likelihood of this possibility. From an analysis perspective, the labeled seemed to be a faithful representation of the variety seen in the high graded upcall population.

### Gunshot

As the AUC score and FPR of the holdout data did not appear to achieve neutral performance to the labelled data, the addition of more DS to the labelled set may improve the RF classifier if perceived performance is insufficient. Fewer data hours were included in the gunshot DS than in the upcall DS: while this was a response to the greater number of calls per unit of time seen in gunshot DS, it still could limit the variety of the data given, as call plasticity is strongly influenced by propagation and attenuation factors which depend on the physical properties of the water, distance of the calling animal (as well as associated changes in bathymetry due to change in distance), and orientation of the calling animal all of which I expect to change more over a larger period of time. The morphology of gunshot calls is known to be strongly influenced by these propagation effects (Crance, Berchok, & Keating 2017), and one of the most informative features of the classifier “SwitchesY max” is a measurement of propagation line occurrence, indicating that propagation effects certainly influence classifier performance. A larger temporal scale of data may be required to account for polymorphism due to propogation.

## Future directions

Mastor\_detecter is a flexible solution to large scale analysis of datasets that feature consistent intermittent noise from self-noise or the environment. A sizable library of wave files containing SOI is needed to build the labelled dataset, so Mastor\_detecter is not well suited to exceptionally rare signal. It is likely not well suited for cursory analysis of a dataset that does not correspond to the geographic and seasonal range of the labelled dataset, as building a labelled dataset that is soundscape specific would likely be necessary to account for the local differences in noise composition and possibly ‘album effect’ issues due to differences in propagation environment and recorder specification (Roch, Stinner-Sloan, Baumann-Pickering & Wiggins 2015). Mastor\_detecter is likely best suited for organizations that are seeking long term solutions for machine assisted analysis of data collected from a consistent region using consistent equipment, and should be effective within those constraints.

When thinking of how to apply Mastor\_detecter to a given problem, it is essential to think of

1. What can I reasonably expect of the detector, given the known constraints?
2. Does the data provided, and the way I label these data, fit my expectation for the detector?

Often, this will require thinking about the features that allow your detector to work, and using your knowledge of the system to think down the road for issues that could hamper the effectiveness of the detector.

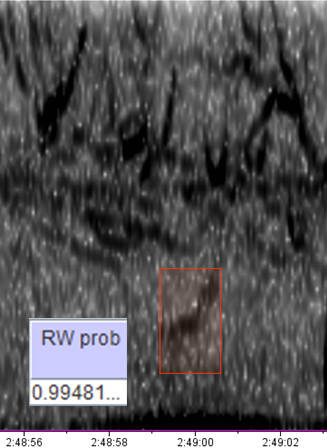
For instance: the upsweep detector seems to rely heavily on slope related features for the brunt of its classification. Because of this, and given that there is no established criteria for distinguishing upsweeps between species, it is not reasonable to expect the detector to be able to distinguish upsweeps from bowhead and humpback from right whale upcalls. It is possible that when presented with upsweeps that fit the existing criteria to a tee, but were marked as negatives, it would be able to dig deeper down the feature list to find ones that are more effective but considering that to a human eye these signal are indistinguishable out of context this hope is far-fetched. When designing an upcall or upsweep detector, you may wish to include data from dense bowhead song, as this this can be a significant source of FP due to the frequency modulation and sharing tonal characteristics. However, the presence of bowhead upsweeps in the data complicates this, leaving you with a couple choices:

1. Ground truth data so that bowhead upsweeps are not marked as positive detections. Pros: technically correct (helpful for data management and organization perspectives), may allow the detector to “dig deeper” in the feature list to discriminate right whale and bowhead upcalls. Cons: unlikely to be able to do so, positives and negatives belonging to the upsweep feature class will likely lower probability of correctly identifying upcalls outside of bowhead song.
2. Ground truth data so that bowhead upsweeps are not marked as positive detections and add bowhead presence proxy variables to classifier. Pros: may give the detector more power to discriminate these bowhead upsweeps. Cons: these variables will apply a flat probability penalty to upsweeps during bowhead song, which would lower the likelihood of identifying a positive upcall detection within bowhead song. (Worth mentioning that a human analyst would almost certainly treat upcalls seen within bowhead song more skeptically, however, they are also ideally considering variables such as patterning and bowhead song form that are not able to be considered by the detector)
3. Ground truth data so that bowhead upsweeps are marked as positive detections. Pros: will maintain good performance at identifying upsweeps throughout the data. Cons: no longer explicitly identifies upcalls, labelled dataset now contains detections that aren’t upcalls being marked as positive detections (making your detector an upsweep detector instead of a upcall detector)

Systems that have complicated biotic soundscapes will have to consider situations like this when deciding how to design and apply detectors, as well as interpret their outputs. The golden rule may be to consider “how would a human analyst do that, and am I ok with a computer trying to do it that way to?”. Providing context related variables that human analysts rely on such as information on location, call patterning (timing), and season to machine learning based classifiers is an area of ongoing research (Roch 2019) and is a direction that could be explored in the future.

Given the high computational requirements of the holdout data experiment, I recommend not to attempt to replicate the holdout data 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 to the labelled set 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.

**Figure x.** Upsweep in bowhead data with high probability score from upsweep detector



# References

Araya-Salas. (2017), Rraven: connecting R and Raven bioacoustic software. R package version 1.0.2.

Araya-Salas, M. and Smith-Vidaurre, G. (2017), warbleR: an r package to streamline analysis of animal acoustic signals*. Methods Ecol Evol. 8*, 184-191.

Bassett, C., Thomson, J., Dahl, P., & Polagye, B. (2014). Flow-noise and turbulence in two tidal channels. *The Journal of the Acoustical Society of America,* *135*(4), 1764-74.

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.

Chen et al. 2004. Using Random Forest to Learn Imbalanced Data. *UC Berkeley technical report No.666*.

Crance, J., Berchok, C., Wright, D., Brewer, A., & Woodrich, D. (2019). Song production by the North Pacific right whale, *Eubalaena japonica*. Manuscript in review.

Duda, R., & Hart, P. (1972). Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM,* *15*(1), 11-15.

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: The case of tests with continuous results. *Biochemia Medica,* *26*(3), 297-307.

Ligges, U., Krey, S., Mersmann, O., and Schnackenberg, S. (2018). tuneR: Analysis of Music and Speech.

McDonald, M.A., and S.E. Moore. 2002. Calls recorded from North Pacific right whales (*Eubalaena japonica*) in the eastern Bering Sea. *J. Cetacean Res. Manage. 4*: 261-266.

Mceachern, J., & Lauchle, G. (1995). Flow‐induced noise on a bluff body. *The Journal of the Acoustical Society of America,* *97*(2), 947-953.

Mellinger, D. (2002). *Ishmael : 1.0 user's guide ; Ishmael : Integrated system for holistic multi-channel acoustic exploration and localization* (NOAA technical memorandum OAR PMEL ; no. 120). Seattle, Wash.: NOAA, Pacific Marine Environmental Laboratory.

Microsoft and Steve Weston (2017). foreach: Provides Foreach Looping Construct for R. R package version 1.4.4.

Roch, M. (2019) Using context to improve marine mammal classification. *ONR* *Marine Mammal & Biology Program Review, 2019*. 40-41

Roch, M., Stinner-Sloan, J., Baumann-Pickering, S., & Wiggins, S. (2015). Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks. *The Journal of the Acoustical Society of America,* *137*(1), 22-9.

Ross, & Allen. (2014). Random Forest for improved analysis efficiency in passive acoustic monitoring. *Ecological Informatics,* *21*, 34-39.

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.

Sueur J., Aubin T., Simonis C. (2008). Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics, 18*: 213-226

# Supplemental

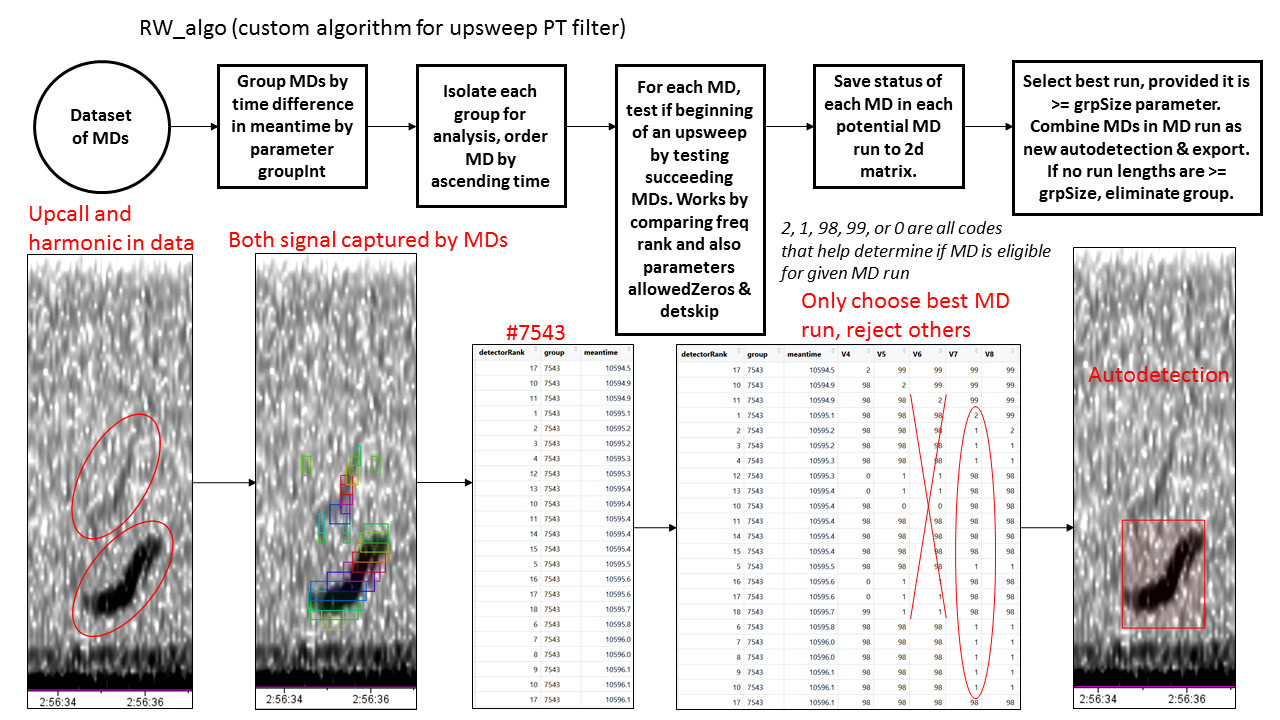
## Raven and custom algorithm

**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.

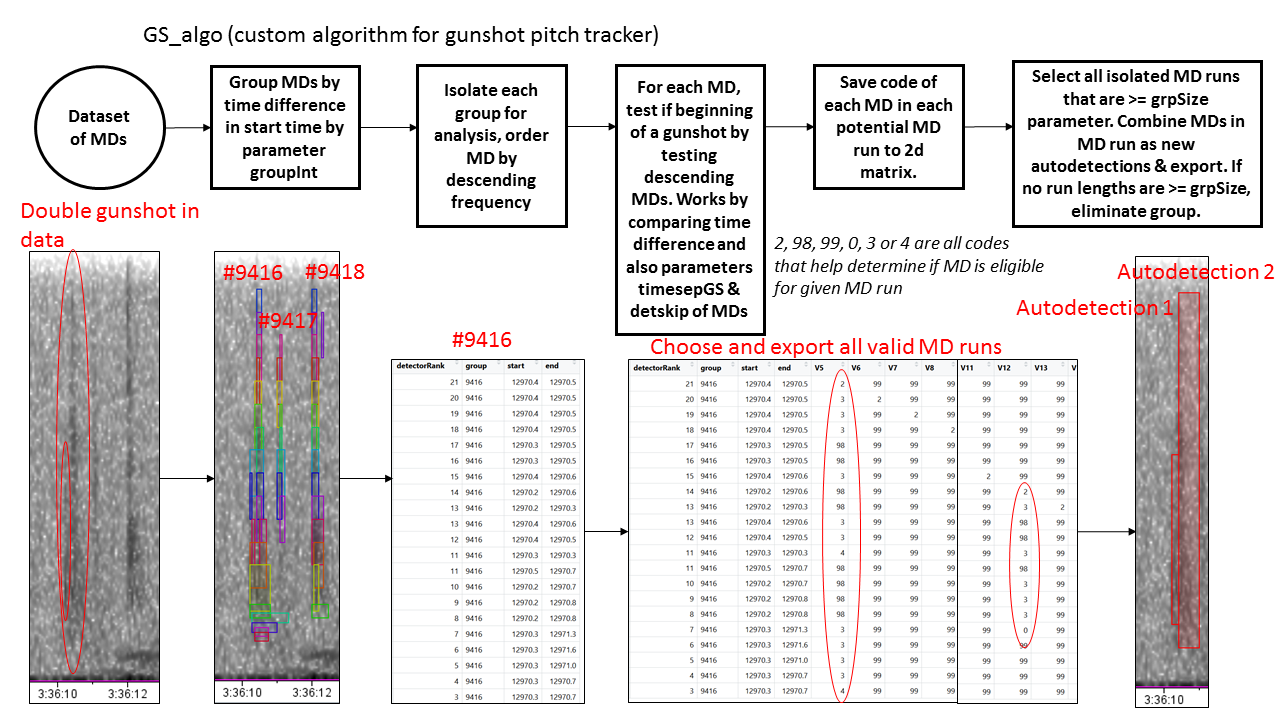
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Suite ID | Low freq (Hz) | High freq (Hz) | Min dur (s) | Max dur (s) | Min sep (s) | Min % occupancy | SNR Threshold (dB) | Block Size (s) | Hop Size (s) | Percentile |
| P11a | 60 | 80 | 0.25 | 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. (A)** Description of algorithm for upsweep pitch tracker, and an example when applied to a MD group that features a harmonic. This algorithm was designed to be exclusive of multiple candidate SOI in the same group, due to the higher inter-call interval typically seen in upcalls and the ability of harmonics to resemble calls themselves and be incorrectly counted as a separate call. **(B)** Description of algorithm for gunshot pitch tracker, and an example when applied to a MD group that features a double gunshot. This algorithm was designed to be inclusive of multiple SOI being present within a group, due to the prevalence and recent interest in NPRW patterns containing double gunshot (Crance, Berchok, Wright, Brewer, & Woodrich 2019).

**A.**



**B.**

## 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 labelled 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

The random forest algorithm is tolerant to many considered features in a large dataset, so features from many sources were included to take advantage of this property. We initially started out with the features provided by specprop() in the ‘seewave’ R package (Sueur, Aubin, & Simonis 2008), expanded to features not included in from specan() in the ‘warbleR’ R package (Araya-Salas & Smith-Vidaurre 2017) (Ligges, Krey, Mersmann & Schnackenberg 2018) until we believed we had included the most commonly used measurements on a wave form or FFT. To further increase performance, we added custom built features from image analysis enabled by the R package ‘imager’. We built 40 of the 75 current features by hand using different measurements of the image. Each feature could be elaborated on, but below are examples of a few highly ranked to reduce Gini coefficient for the random forest.

### Hough slope

MedSlope Hough (V56), BestTheta Hough (V50), BestSlope Hough (V51), MedTheta Hough (V55) are the top ranking features for upcall classifier and all are variations on the “Hough slope” measurement. MedSlope Hough is also one of the highest ranking features for the gunshot classifier. The Hough transform (Duda & Hart 1972) is commonly used in machine learning as a means of normalizing complex image data into the outline of shapes present in an image. The Hough transform is robust to noise in images, allowing it to make clean outlines in its normal application. We took advantage of this property to apply it as an approximation of slope in our noisy spectrogram data. The implementation of the Hough line algorithm in R exports image data in theta/rho form, these measurements as well as the transformation into slope/intercept form were all included as features.

### Island slope

An alternative measure of slope was used to calculate Meanslope (V69). This calculates an average slope from the important islands in the image. The slope here is not calculated with the Hough transform, but with the frequency difference of the average of each islands leftmost and rightmost point in time.

### Switches

SwitchesY max (V67) and SwitchesY mean (V65) both scored highly for the gunshot classifier. These features are both statistics on the number of switches between shape presence and absence that occur on six x axis quantiles (the Y referring the line being drawn on the fixed x value quantile). For gunshots, this likely is informative due to propagation lines generating a high amount of switches on one or more quantiles.

**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 mean | Spectrogram | V61 | mean number of times a line along y axis quantiles switches from shape presence to absence and vice versa |
| SwitchesX se | 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 mean | Spectrogram | V65 | mean number of times a line along x axis quantiles switches from shape presence to absence and vice versa |
| SwitchesY se | 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 |

## Detector parameters

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 performed by hand and iterative testing.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Parameter | Type | | Example value | Description |
| CV | Classifier | | 60 | How many times to generate RF models on random training data from the labelled set. |
| TPRthresh | Classifier | | 0.95 | Determines that cutoff that this % of TP will be included from the labelled set. |
| numFeatures | Classifier | | 75 | Number of features given to the classifier. Change when adding features to preserve correct indexing |
| fileCombine size | Signal manipulation | | 280 | Number of .wav files to combine into larger sections to improve speed of BLED suites. |
| fileCombine size2ndIt | Signal manipulation | | 12 | Number of combined .wav files to combine again (done twice due to limitation with SoX and windows getting mad that too many files were ‘open’ ) |
| Decimate | Signal manipulation | | y | Controls if data is decimated |
| decimation Factor | Signal manipulation | | 8 | Factor to decimate data. Performed iteratively by prime factors. Sampling rate should be divisible by factor. |
| Whiten | Signal manipulation | | n | Sets correct pathing if using whitened data- does not automatically whiten data |
| FO | Signal manipulation | | 100 | Filter order. Parameter for whitening |
| LMS | Signal manipulation | | 0.1 | Least Sean Squared. Parameter for whitening. |
| Filtype | Signal manipulation | | Bband | Type of filter used to whiten signal |
| spStart | PT | | 21 | File number in BLED folder to start BLED suite |
| spEnd | PT | | 42 | File number in BLED folder to start BLED suite |
| grpSize | PT | | 4 | Minimum number of eligible MDs in a MD run to constitute a positive detection |
| allowedZeros | PT | | 2 | Maximum number of consecutive ineligible MDs in a MD run to end the MD run. |
| detskip | PT | | 7 | Maximum amount of skips in rank for a MD to be considered as part of the MD run |
| groupInt | PT | 0.35 | | Maximum time difference for MDs to be considered within the same group. |
| Maxdur | PT | 3.5 | | Maximum duration for a detection |
| Mindur | PT | 0 | | Minimum duration for a detection |
| timesepGS | PT, only gunshot | 1.2 | | Coefficient that helps fit ideal curve of tolerance for MD start time difference in a MD run |
| timediffself | A\_c | 0.15 | | Maximum time difference for detections to be compared by Adaptive\_compare filter |
| probdist | A\_c | 0.2 | | Max difference in detection probability to be average by Adaptive\_compare filter, otherwise lower % detection is just eliminated |
| ImgThresh | Image analysis | 90 | | Noise threshold at which to binarize the spectrogram image. Lower values return more 1s, higher values return more 0s. |
|  |  | |  |  |

## Adaptive\_compare

Adaptive\_compare is a filter that helps lessen the effect of MB. It is particularly helpful on the gunshot detector which is sensitive to picking out multiple MD runs per group (which it does to isolate double gunshot when present). This filter uses the probability of a detection in the decision to combine boxes or only select one of the boxes, so can only be used after the RF classifier has been applied.

**Figure x.** Explanation and example of Adaptive\_compare applied to gunshot autodetections

