DCLDE 2026 DATA

Authors:

Kaitlin Palmer, Ruth Joy, James Pilkington, Harald Yerk, Holger Klinck, Holly LeBlond, Jenn Waldichuck, Scott Viers, Val Viers, Lucy Quale, Jasper Kanes, Thomas Doniol-Valcroze, Lauren Laturnus, Malahat contributors

# Abstract

# Cultural Perspective?

# Introduction

Terminology- call, pulsed call, whistle, echolocation click, detection, detection algorithm,



**ONC Data**

This line of questioning brings up an important point to consider though that I hadn’t thought of! In the files with confirmed orca calls, I annotated every *orca  pulsed call*, but not every other sound. So there are echolocation clicks that weren’t annotated, and possibly some whistles (though I think maybe I annotated all the whistles too – I’d have to review to confirm). Also, the files from 2013 I *only* annotated orcas – they were added after the fact to increase the orca call class size and include other ecotypes. There may be other species in the 2013 files that I didn’t annotate.

**Preprocessing**

A mixed approach

# Methods

Each section should have deployment details, processing details, and annotation details

## Data Summary

Table with sample rate, Owner, Deployment, data start date, end date, number of annotations, KW annotations

Annotations Labels

Data for this project represent a large collaboration of groups and institutions and each dataset was processed in accordance with each groups project goals. Post processing of the annotations was done to provide a uniform system for machine learning algorithms. However, users should consider details from each deployment carefully to determine whether they wish to do any additional post-processing. For example, multiple annotations from the Pilkington datasets may represent different harmonics of the same call. ONC data analysis similarly focused in impulsive calls from killer whales and thus each file may contain un-annotated whistles and or echolocation clicks.

To aid in the data cleaning process we provide the following columns

**SoundFile** – audio file from which the annotation was derived

**StartTime** – seconds into the SoundFile representing the beginning of the annotation

**EndTime** – Seconds into the SoundFile representing the end of the call annotation

**Low Frequency** – low frequency bound of the annotation, in Hz

**High Frequency** – high frequency bound of the annotation, in Hz

**UTC** – UTC of the StartTime

**ClassSpecies**- Character string with four options, Killer Whale (KW), Humback Whale (HW), Abiotic (AB), and Undetermined Biological sound (UndBio).

**KW**- bool indicating whether or not the annotate denotated that the annotation represented a killer whale call

**KW\_certain** – bool indicating whether or not the annotator was certain that the annotation was a KW. This is often represented by a question mark in the annotations. For ONC data, annotators listed all potential species that the thought the call could come from.

**Ecotype** – Ecotype of the killer whale annotation. SRKW- Southern Resident Killer Whale, BKW- Biggs killer whale, NRKW- Northern Resident Killer Whale, or OKW- Offshore Killer Whale

**Data Provider**- Character indicating the data provider

**Dep**- Character, deployment location

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | **Annotations** | | | |
| **Dataset Provider** | **Deployment Location** | **Lat\_Lon** | **Depth (m)** | **fs (khz)** | **All Annotations** | **Annotation Level** | **KW** | **Und. Bio** | **Abiotic** |
| **Orca Sound** |  |  |  |  |  | Call, pulsed only |  |  |  |
| **ONC** | Berkley Canyon | 48.426 -126.174 | 40 | 64 | 14120 | Call, pulsed only |  |  |  |
| **DFO\_** **Pilkington** | West Vancouver Island |  | 114 | 16.384 | 114056 | Detection |  |  |  |
| **DFO\_Pilkington** | Northern Mainland British Colombia |  | 35 | 16 | 44818 | Detection |  |  |  |
| **DFO\_Yerk** | Carmanah Point |  | 55 | 192 |  | Detection |  |  |  |
| **DFO\_Yerk** | Swanson Channel |  | 72 | 256 |  | Detection |  |  |  |
| **DFO\_Yerk** | Strait of Georgia North |  | 193 | 256 |  | Detection |  |  |  |
| **DFO\_Yerk** | Strait of Georgia South |  | 245 | 256 |  | Detection |  |  |  |
| **JASCO/Malahat** |  |  |  |  | 36936 | Call |  |  |  |
| **SIMRES** |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

This dataset is based on roughly 116 days of duty-cycled audio data recorded with a SM2M autonomous recorder deployed on the northern mainland coast of British Columbia. The recording effort focused on the winter months, and the batteries depleted earlier than anticipated. The dataset consists of the full deployment’s duty-cycled recordings converted to FLAC (for ease of sharing), as well as a table of the associated automated detections that were identified to species and sound type by human analysts. This dataset was not initially intended to be used in the training of detectors/classifiers.

## Ocean Networks Canada

**Deployment**

Acoustic data were collected using an Ocean Sonics SC2 hydrophone deployed at 48.426, -126.174, 70m from the Barkley Canyon Upper Slope platform of ONC’s North-East Pacific Time-series Underwater Networked Experiments (NEPTUNE) observatory. The hydrophone collected data from 16 May 2013 to 13 Jan 2015. It was mounted 1 m above the sea floor and sampled continuously at 64 kHz. The hydrophone model used applies an anti-aliasing filter during data collection and digitization, yielding a 32kHz bandwidth with reduced apparent sound intensities above 25.6kHz. Data collection was near-continuous, with some gaps due to maintenance and outages. Most data were archived in near-real-time. Data from some time periods have been split into high- and low-pass filtered files with overlapping bandwidths due to collection during times of concern for national security; low-pass filtered data were archived in near-real-time and high-pass filtered data that did not contain classified signals were archived after review by regional navies.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Location** | **Instrument** | **Date Start** | **Date Stop** | **SRKW** | **Biggs** | **NRKW** | **Offshore** | **HW** |
| Barkley Canyon Upper Slope | Ocean Sonics SC2 | 2013-05-16 | 2015-06-13 | 129 | 315 | 0 | 418 | 2946 |

**Processing**

These annotations were produced in three separate efforts with different effort protocols, using JASCO Applied Science’s PAMLab software, and reviewed for accuracy, signal diversity, and completeness using Raven Pro.

**Annotation**

Pulsed Calls

The first effort produced one annotation per species per file for every other file of the first 4 days of each month in 2014. Annotations were produced using a logarithmic spectrogram display with different spectral settings in 4 different bands, enabling multi-species identification across the full bandwidth in a single pass. All visible signals were considered for annotation, and no signal-to-noise ratio threshold was used. As a result, files within the supplied file list that do not contain an annotation may be assumed to have no marine mammal sounds, though it is worth noting that these files have not been reviewed and it is possible some sounds were missed.   
  
The second effort added approximately 2 to 20 annotations per file to files containing Pacific white-sided dolphin pulsed calls, and an annotation for every orca call in the 2014 data set, as these classes were not large enough for training an effective classifier.   
  
The third effort added one annotation per orca call for every file flagged as containing orcas in the previously-made weak-label annotations for the full 2013 to 2015 Barkley Canyon deployment period of Ocean Sonics hydrophone icListen HF 1251, as orcas were uncommon during the annotated time periods.

All files originally identified as containing marine mammal signals have been manually reviewed a second time using species-specific spectrogram settings in Raven Pro to fix erroneous or misclassified annotations. During the review process, some humpback whale and blue whale annotations were added or replaced to improve the diversity of signals in the data set. Delphinid click annotations, which were not originally annotated to species due to insufficient bandwidth to capture the full click structure, were updated to differentiate between Risso’s dolphin, Pacific white-sided dolphin, and other delphinids clicks based on their literature descriptions and comparison to broader bandwidth data containing clicks from these species recently collected from the same location.

## Department of Fisheries and Oceans Canada

Two groups within DFO provided datasets to the challenge, the Yerk and Pilkington labs. Data processing methods were consistent across projects within each lab but varied slightly between each lab.

### Pilkington

**Deployment**

The Pilkingon lab provided data from two deployment locations. As with all DFO deployments, exact locations are not provided.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Location** | **Instrument** | **Date Start** | **Date Stop** | **SRKW** | **Biggs** | **NRKW** | **Offshore** | **HW** |
| West coast of  Vancouver Island. | AURAL-M2 | 2011-02-15 | 2012-05-25 | 48 | 5336 | 4558 | 258 | 95861 |
| mainland coast of British Columbia. | SM2M | 2013-10-10 | 2014-02-04 | 0 | 2309 | 3501 | 947 | 26058 |

**Processing**

The raw audio recordings (WAV) were post-processed using the Whistle and Moan Detector in PAMGuard version 1.12.08 (Gillespie et al. 2013). The detector was user- configured with a high-pass filter of 800Hz to limit the number of humpback whale detections and lessen the manual validation burden. The SNR detection threshold was set to 6dB. All detections in the first two seconds of each file were excluded because the detection algorithm produces several false detections within this period. The automated detections from the detector were stored in an Excel spreadsheet

**Annotation**

All detections **including whistles and pulsed calls** were aurally and visually reviewed using PAMGuard and identified to species (for biotic) and sound type (for abiotic). Where applicable and as time allowed, detections were also acoustically identified to intra-specific groupings including ecotype. Note that files may contain more identifiable calls than the annotations indicate. These manual reviews were conducted by trained and experienced analysts. A small portion of annotations contain question marks for both the KW class and the Ecotype class. As these calls were validated to ecotype we assume that KW class is certain and the question mark is removed.

Note that individual detections may be separate components of the same discrete call (ie harmonics or sidebands), thus, not every detection represents a unique vocalization. The PAMGuard Whistle and Moan detector detects individual contours, so all individual harmonics within a call would constitute separate detections if they meet the detector’s criteria (this happens quite frequently). Also, the settings of the detector mean that independent tones (like from multiple individuals) that cross or overlap in frequency and time may be detected as a single detection.

Note that the original WAV files were used in post-processing with the detector. These WAV files were converted to FLAC in the creation of dataset KkHK0R2F-NML1 for sharing. If FLAC are limiting for your purposes, the WAV files are also available if necessary – get in touch with the dataset contacts.

### Yerk

The Yerk lab provided data from four deployment locations, Carmanah, Swanson Channel, and two locations Strait of Georgia.

**Deployment**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Location** | **Instrument** | **Date Start** | **Date Stop** | **SRKW** | **Biggs** | **NRKW** | **Offshore** | **HW** |
| Carmanah Point | ST6249 | 2022-03-08 | 2022-06-29 | 1610 | 694 | 364 | 0 | 131 |
| Swanson Channel | AMAR 777 | 2021-11-13 | 2022-01-09 | 5630 | 25 | 0 | 0 | 1660 |
| Strait of Georgia North | AMAR 610 | 2021-09-05 | 2021-10-01 | 4184 | 593 | 0 | 0 | 0 |
|  | AMAR 617 | 2021-11-28 | 2021-11-30 | 0 | 324 | 0 | 0 | 42 |
| Strait of Georgia South | AMAR 607 | 2021-11-11 | 2021-11-18 | 159 | 191 | 0 | 0 | 221 |
|  | AMAR 779 | 2021-09-04 | 2021-10-01 | 2070 | 71 | 0 | 0 | 114 |

**Processing**

The raw audio recordings (WAV) were post-processed using the Whistle and Moan Detector in PAMGuard version 1.12.08 (Gillespie et al. 2013). The detector was user- configured with a high-pass filter of 2 KHz. The detector was also configured to use a SNR threshold of 8dB.

**Annotation**

All detections including whistles

## JASCO Malahat First Nation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Location** | **Instrument** | **Date Start** | **Date Stop** | **SRKW** | **Biggs** | **NRKW** | **Offshore** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Deployment**

**Processing**

**Annotation**

## OrcaSound

Orcasound has been collecting and annotating audio data as part of google summer of code challenges as well as

**Deployment**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Location** | **Instrument** | **Date Start** | **Date Stop** | **SRKW** | **Biggs** | **NRKW** | **Offshore** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Processing**

**Annotation**

## SIMRES

**Deployment**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Location** | **Instrument** | **Date Start** | **Date Stop** | **SRKW** | **Biggs** | **NRKW** | **Offshore** |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Processing**

**Annotation**