Description of the dataset

A Workshop data set of Right Whale (Eubalaena glacialis) has kindly been provided by Sofie Van Parijs from the NOAA North East Fisheries Science Center and the Cornell Bioacoustics Research Program. Additional data have also been provided by the International Fund for animal Welfare.

The recordings were made with Cornell MARUs (http://www.birds.cornell.edu/brp/hardware/autonomous-recording-units) deployed off the coast of Massachusetts in 2000, 2008 and 2009. Recorders were deployed in arrays of 6 or 10 devices. For the workshop dataset a single channel has been stripped out and converted to 16 bit wav file format. The sample rate on all recordings is 2kHz. Recordings are arranged in folders by date and the start time of each recording is coded into the file name, e.g. NOPP6_EST_20090330_001500.wav is a recording which started at 15 minutes past midnight on 30 March. All recordings are 15 minutes long.

Seven days of right whale upsweep calls have been carefully annotated by Genevieve Davis, from Sofie's research group using XBat. When logging the calls, all 6 or 10 data channels were viewed. Often calls would be clearly visible and identified as right whale on one channel, but would have a much lower SNR on the logged channel. This means that the logs contain some very low SNR calls which would probably have been impossible to identify visually if only the single channel data had been viewed.

In all, the dataset contains the following:

Seven days of data containing annotated right whale upsweep calls.

Seven days of data containing annotated right whale gunshot calls.

Seven days of data containing (to the best of our knowledge) no right whale calls at all.

For development purposes, four days of data from each set and the corresponding log files (6916 logged calls) are being made available now. The final three days of data will be made available approximately one month before the workshop, but without the log files. Participants will be asked to submit results in a standard format for comparison with the undisclosed log files shortly before the workshop.

The log files, having been made by humans, are not of course perfect. While the operators have attempted to log every right whale sound it is possible that they have missed some and may have logged sounds which are not right whales. Generally though, you can assume that a sound logged as a right whale is a right whale and that a sound not logged as a right whale isn't a right whale with the following exceptions:

1. Some sounds have been named with an event .tag set to '?' indicating that they weren't sure. These events should probably be excluded from any analysis.

2. The period from 14:30 to 16:30 on 29 March 2009 when there was so much going on that it was impossible to reliably log calls.

For the IFAW 2000 data set, we are confident that there were no right whales present in any of the recordings. Aerial surveys found no animals in the vicinity of the pop-ups during the deployment period and human operators were unable to find any right whales in the data.

Channel Numbers

The channel numbering in the log files refers to the channel number in the 6 or 10 channel recordings generated from multiple MARU units. However, for the workshop, we have stripped out the relevant channel for each day. Channel numbering should therefore be ignored, i.e. assume that all channel numbers in the log files are channel 1 (or 0 if you program in Java or C).

Log file format

The logs have been provided in standard Xbat forma which is basically a single Matlab data structure describing both the sounds and the human made logs. There is one log file per day of data and times within the log files are relative to the start of each day. If you open a log file with Matlab the format is reasonably obvious. E.g. for 30 March, the log file structure is:

```
Log_NOPP6_20090330_RW_upcalls =
               type: 'log'
            version: 'PRE R5  (MATLAB 7.10)'
                 id: 1.3893e+015
               path: 'Y:\DATA TEMPLATES AND DETECTORS\DCL\Right
Whale Upcall Logs\'
               file: 'NOPP6_20090330_RW_upcalls.mat'
             author: 'Genevieve'
            created: 7.3513e+005
           modified: '26-Sep-2012 10:47:32'
              sound: [1x1 struct]
         annotation: [1x1 struct]
              event: [1x1663 struct]
      deleted_event: [1x17 struct]
    selected_events: [1x1 struct]
             length: 1663
            curr_id: 1681
            channel: []
               time: []
               freq: []
           duration: []
          bandwidth: []
        measurement: [1x1 struct]
         generation: []
            visible: 1
              color: [1 0 0.0500]
          linestyle: '-'
          linewidth: 2
              patch: 0
           event_id: 1
               open: 0
           readonly: 0
           autosave: 1
         autobackup: 1
              saved: 1
           userdata: []
```

The critical data element for this study is the event array which in this case contains 1663 annotated sounds, e.g.

```
Log_NOPP6_20090330_RW_upcalls.event(1180)  =
             id: 1191
           tags: {'?'}
         rating: []
          notes: {}
          score: []
        channel: 10
           time: [5.6652e+004 5.6654e+004]
           freq: [100.8772 238.3041]
       duration: 1.8852
      bandwidth: 137.4269
        samples: []
           rate: []
          level: 1
       children: []
         parent: []
         author: 'Genevieve'
        created: 7.3513e+005
       modified: 7.3513e+005
       userdata: []
      detection: [1x1 struct]
     annotation: [1x0 struct]
    measurement: [1x0 struct]
```

The most important information items for each event are the time, which is a two element array giving the start and end times of the event relative to the start of the day and the frequency which is the lower and upper bounds of the sound in Hz.

Any questions about the dataset should be addressed to dg50@st-andrews.ac.uk

.