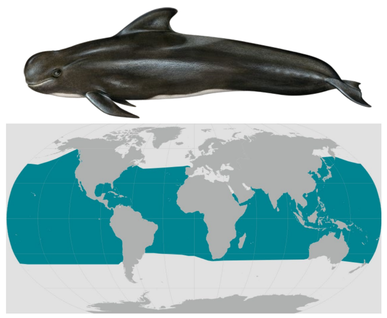
Detecting and Classifying Short-finned Pilot Whale Acoustics with Deep Learning

Virginia Pan, Dr. Nicola Quick, Dr. Douglas Nowacek
Pratt School of Engineering and Duke Marine Laboratory

OD44A-3476

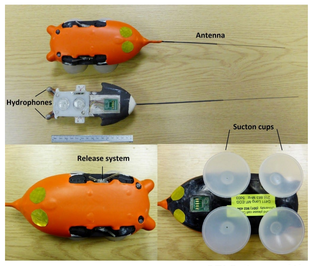
Short-finned Pilot Whales and DTAGs



Short-finned pilot whale (top) [3]
Short-finned pilot whale distribution (bottom) [3]

Globicephala macrorhynchus
- Highly social
- Squid and fish diet
- Hunt at +1,000 ft
- Use sound for navigation, hunting and communication
- Often involved in mass standings
- "Cheetahs of the deep sea"

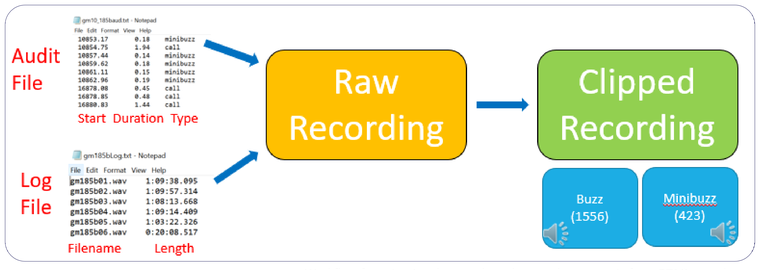
Digital Acoustic Recording Tag (DTAG)
- Records audio, pitch, roll, heading, and depth
- Can study social interactions, foraging, diving and ecology
- Nowacek Lab annually tags short-finned pilot whales off Cape Hatteras



DTAG diagram [1]

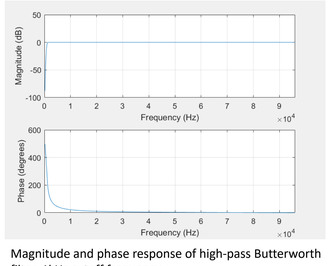
Manual review of DTAG audio is time consuming and requires a trained ear!

Audio Pre-processing



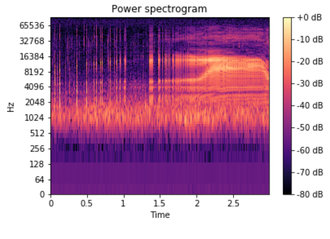
Workflow for extracting buzz and minibuzz audio segments from DTAG recordings

- Focused on buzz and minibuzz vocalizations
- Data from 2008, 2010, and 2011 (25 individuals total)
- Log file: chronological list of audio segments for an individual
- Audit file: Manually reviewed audio for whale sounds
- Extracted audio clips
- High-pass Butterworth Filter with cutoff of 1kHz
- STFT of filtered audio clips served as input to classification network

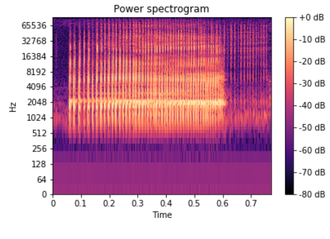


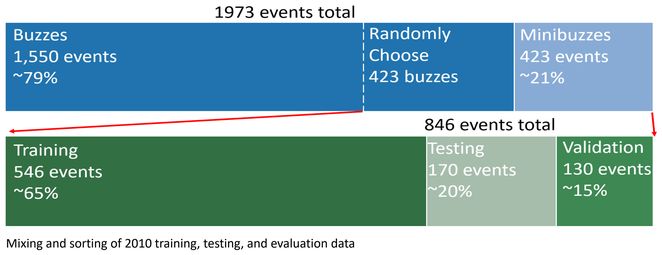
Magnitude and phase response of high-pass Butterworth filter, 1kHz cutoff frequency

Buzz STFT



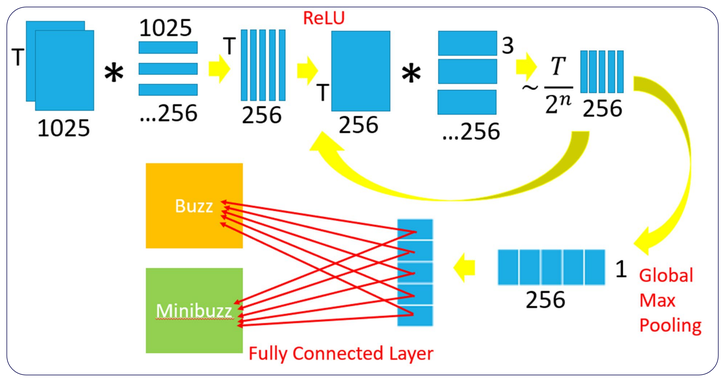
Minibuzz STFT





Mixing and sorting of 2010 training, testing, and evaluation data

Classification with Deep Learning



Network Structure for Classifying Variable Length Audio

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 3559, 1, 256)      262400
re_lu_1 (ReLU)               (None, 3559, 1, 256)      0
conv2d_2 (Conv2D)            (None, 1780, 1, 256)      196608
re_lu_2 (ReLU)               (None, 1780, 1, 256)      0
conv2d_3 (Conv2D)            (None, 890, 1, 256)       196608
re_lu_3 (ReLU)               (None, 890, 1, 256)       0
conv2d_4 (Conv2D)            (None, 445, 1, 256)       196608
re_lu_4 (ReLU)               (None, 445, 1, 256)       0
global_max_pooling2d_1 (Glob (None, 256)               0
dense_1 (Dense)              (None, 2)                 512
=================================================================
Total params: 852,736
Trainable params: 852,736
Non-trainable params: 0
```

Convolutional Neural Network Model Summary

Deep Learning Network:
- Based on [2]
- Identify presence of 256 unique sounds
- Group sounds into sequences
- Global max pooling layer checks for sound sequence in whole audio clip (handled different length audio segments)
- Fully connected layer
- AWS: S3 storage, Sagemaker: Jupyter Notebook, Python 3 backed by Keras

Classification Results

```
[INFO] training w/ generator...
Epoch 1/5
 - 103s - loss: 0.6696 - acc: 0.7418 - val_loss: 0.6291 - val_acc: 0.9154
Epoch 2/5
 - 94s - loss: 0.5574 - acc: 0.9011 - val_loss: 0.4374 - val_acc: 0.9385
Epoch 3/5
 - 93s - loss: 0.3821 - acc: 0.9267 - val_loss: 0.2493 - val_acc: 0.9692
Epoch 4/5
 - 94s - loss: 0.2963 - acc: 0.9396 - val_loss: 0.1698 - val_acc: 0.9769
Epoch 5/5
 - 94s - loss: 0.2499 - acc: 0.9487 - val_loss: 0.1231 - val_acc: 0.9769
ran fit_generator
```

Training: 2010 data
Testing: 2010 data
✓ **97.69% accuracy**

2010: 170 minibuzz and buzz events

```
13/13 [==============================] - 14s 1s/step
[0.14795524455033815, 0.9763313623575064]
['loss', 'acc']
```

Figure 5: Detection performance metrics

2008: 227 minibuzz and buzz events

```
13/13 [==============================] - 16s 1s/step
[0.27890387210784806, 0.9053254494300256]
['loss', 'acc']
```

2011: 575 minibuzz and buzz events

```
13/13 [==============================] - 16s 1s/step
[2.2900024331532993, 0.5325443916595899]
['loss', 'acc']
```

Training: 2010 data
Testing: 2008 & 2011 data
x **Poor accuracy**

```
Epoch 1/5
 - 118s - loss: 0.6587 - acc: 0.7546 - val_loss: 0.5866 - val_acc: 0.8000
Epoch 2/5
 - 117s - loss: 0.5060 - acc: 0.8901 - val_loss: 0.4182 - val_acc: 0.9154
Epoch 3/5
 - 117s - loss: 0.3859 - acc: 0.8919 - val_loss: 0.3530 - val_acc: 0.9308
Epoch 4/5
 - 117s - loss: 0.3050 - acc: 0.9139 - val_loss: 0.3460 - val_acc: 0.8923
Epoch 5/5
 - 117s - loss: 0.3268 - acc: 0.8956 - val_loss: 0.3716 - val_acc: 0.9077
ran fit_generator

13/13 [==============================] - 49s 4s/step
[0.2015399462901629, 0.9763313623575064]
['loss', 'acc']
```

Training: 2008, 2010, 2011 (equal mix)
Testing: 2008, 2010, 2011 (equal mix)
✓ **97.63% accuracy**

References

[1] DTAG: A Digital Acoustic Recording Tag. (2017, May 12). Retrieved December 1, 2019, from https://www.whoi.edu/website/marine-mammal-behavior-lab/dtag

[2] Hertel, Lars & Phan, Huy & Mertins, Alfred. (2016). Classifying Variable-Length Audio Files with All-Convolutional Networks and Masked Global Pooling.

[3] NOAA. (n.d.). Short-Finned Pilot Whale. Retrieved December 1, 2019, from https://www.fisheries.noaa.gov/species/short-finned-pilot-whale

[4] Auditok Library: https://github.com/amsehili/auditok

[5] B. Mcfee, C. Raffel, D. Liang, D. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," Proceedings of the 14th Python in Science Conference, 2015.

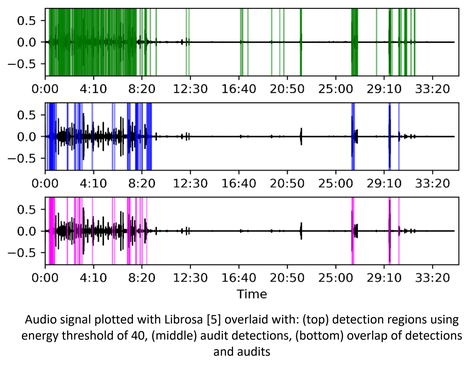
Detection in Time Domain

- High-pass Butterworth Filter with cutoff of 1kHz
- Auditok library [4] for event detection
- Parameters for peak detection:
  - Minimum signal length: 0.1s
  - Maximum signal length: 20s
  - Maximum length of silence within an audio segment: 0.5s
  - Energy threshold: 30-50, increments of 2

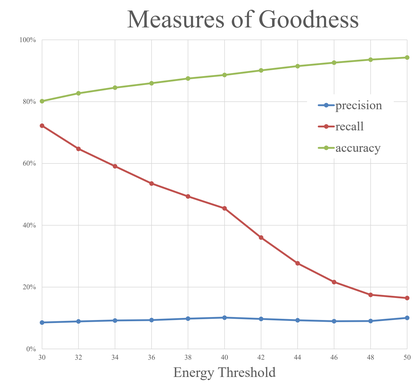
$$energy = 10 * \log_{10} \frac{\|signal\|_2^2}{\text{length}(signal)}$$

- Detected events strictly non-overlapping
- Tested with shorter (35 minute) DTAG audio segment with an abundance of sound events

Detection Results



Audio signal plotted with Librosa [5] overlaid with: (top) detection regions using energy threshold of 40, (middle) audit detections, (bottom) overlap of detections and audits



- Lower energy threshold:
  - Captured more audit events
  - Increased recall
  - Similar precision
  - Lowered accuracy

Detection Performance Metrics

$$precision = \frac{total\ overlap\ time}{total\ detected\ time} \qquad recall = \frac{total\ overlap\ time}{total\ audit\ time}$$

$$accuracy = \frac{(total\ overlap\ time) + (total\ nonoverlap\ time)}{total\ time}$$

→ *Aim to maximize recall (prevent missed events)*
→ *Can identify false positives in classification stage*

Successes
- Ignoring silent periods
- 72% recall rate at e = 30
- Clear relationship to parameters, future tuning has the potential to further improve performance

Shortcomings
- Acted as a peak detector, but many audit events do not occur at peak sound levels
- Detector breaks up continuous sounds (should be processed as single event)
- Detector does not allow for overlapping events

Future Work

- Deep Learning methods for simultaneous detection and classification
  - Purely feed-forward CNN on a continuously sliding window of the most recent audio spectrogram data
  - Can use classification scheme described here to continuously classify sound chunks using a sliding window
  - Current classification scheme already has some duration invariance built in
- Retrain the network
  - Incorporate more training data from new short-finned pilot whales
  - Train network with distribution of buzzes and minibuzzes proportional to life
  - Train network on other short-finned pilot whale sounds
- Investigate biological trends in data
  - Acoustic behavior in relation to time of day, depth, and movement