Robust Automated Cetacean Identification Using Underwater Acoustics

Nathan Staffa, Thomas Teisberg

Stanford
Electrical Engineering

Abstract

Thousands of hours of underwater acoustic data has been recorded by researchers hoping to learn more about marine mammal populations. One of the main existing challenges in the field is the detection and classification of sounds made by different species. Being able to automatically detect and classify whale species from underwater hydrophones would greatly improve our ability to monitor and track whale populations.

We approach this problem by training a convolutional neural network to classify 1 second audio clips as either a particular whale species or as not containing whale sounds. Our approach achieves 94.4% correct detection of whales and greater than 99.3% correct classification of detected whale sounds.

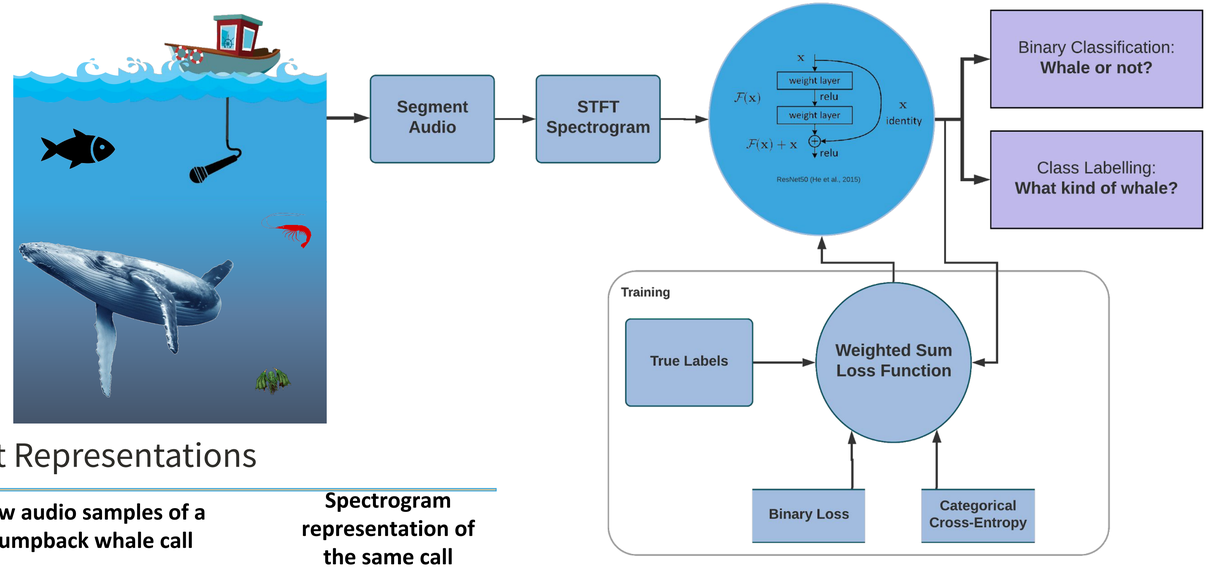

Figure 3: System overview: Underwater recordings are classified using a deep neural network. A weighted loss function between categorical cross-entropy and binary (whale/no whale) classification is used during training.

Results

For our final model, we used mel-frequency binning, our modified ResNet50 architecture, and the weighted loss function. The confusion matrix for this network evaluated on our test set is shown below.

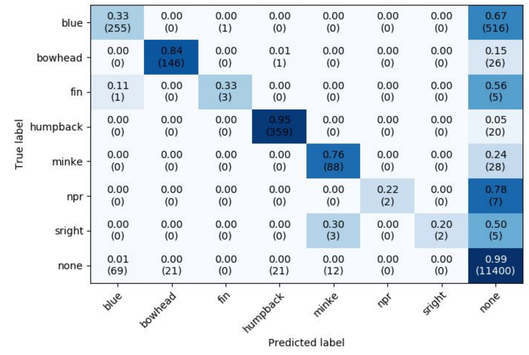

Figure 4: Normalized confusion matrix for our best-performing model evaluated on our test dataset.

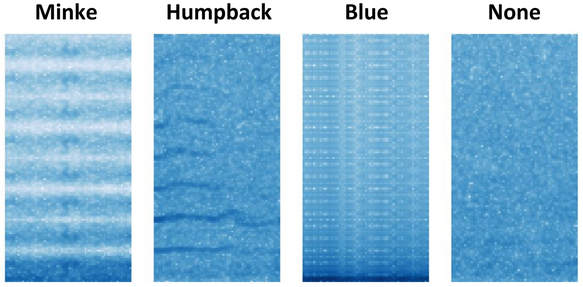

Figure 1: Examples of whale vocalizations (log-compressed spectrograms)

Dataset

We use the MobySound collection of annotated underwater recordings. The MobySound database contains recordings of different aquatic species made on a variety of recording devices in different locations. We pre-process the data by:

- Extracting random 1 second clips, resampling them to 4 kHz, and sorting them by label.
- Randomly mixing background (i.e. no whale) clips from other recorders into every clip
- Randomly shuffling and splitting the clips into train/validation/test sets with an 80/10/10 split

Blue	8246 (6.3%)	**Bowhead**	1760 (1.3%)
Fin	121 (<0.1%)	**Humpback**	3144 (2.4%)
Minke	1127 (0.8%)	**N. Pacific Right**	38 (<0.1%)
Southern Right	48 (<0.1%)	**None**	114,908 (88.8%)

Table 1: Dataset class distribution (before train/validation/test split)

Input Representations

Raw audio samples of a humpback whale call

Spectrogram representation of the same call

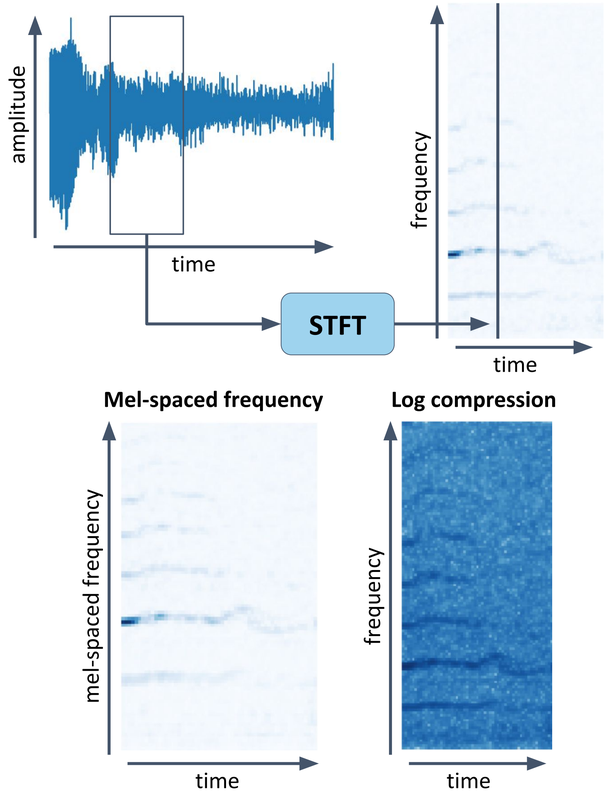

Figure 2: Top: Diagram of how a spectrogram representation is made from raw audio. Bottom: Examples of other representations commonly used.

Model Architectures

We modified the input layers of a number of network architectures that achieve high performance on ImageNet to accept our input. Because our input type is significantly different, we retrain each network from scratch.

Base Architecture	Validation Set Binary Accuracy	Validation Set Categorical Accuracy
2 Layer Convolutional	94.0%	94.0%
VGG16	88.5%	88.5%
ResNet50	94.5%	94.5%
InceptionResnetV2	94.4%	94.4%

Table 2: Validation results from training modified versions of several common architectures.

Weighted Loss Function

In order to encourage correct detection of whales, we use a custom loss function that weights correct "binary" detection of the whale (i.e. whale/no whale) more than correct categorical classification.

$$\text{Loss}_o = -\sum_{c=0}^{7} y_{o,c} \log(\hat{p_{o,c}}) - w_b \left[y_{\text{binary},o} \log(\hat{p}_{\text{binary},o}) + (1 - y_{\text{binary},o}) \log(1 - \hat{p}_{\text{binary},o}) \right]$$

After experimentation, we selected $w_b = 2$ as the best trade-off between categorical and binary accuracy.

Future Work

- There's a huge amount of unlabelled underwater acoustic data. We'd like to be able to leverage this model to explore the unlabelled data and perhaps to assist a human in labelling more of the recordings.
- One of the challenges in this field is generalizing a classifier between different recording setups. Our data noise mixing process is designed to help with this, but there's currently insufficient labelled data to fully test this.
- Some work has found success with trainable audio representations. We think there's strong potential for trainable audio representations that might be able to find better representations specifically for animal sounds (as opposed to human perception).
- Lots of other things apart from whales make noise underwater. Machine learning in underwater acoustics could be leveraged for many other applications, ranging from tracking other marine mammals to detecting illegal fishing.

Acknowledgements

Thanks to Cindy Jiang for her input on our project and the MobySound team for assembling the recordings used in our dataset. Please see our paper for full references.