Beluga whale acoustic signal classification using deep learning neural network models

Ming Zhong,¹ Manuel Castellote,²,a) Rahul Dodhia,¹ Juan Lavista Ferres,¹ Mandy Keogh,³ and Arial Brewer²

¹AI for Good Research Lab, Microsoft, Redmond, Washington 98052, USA
²Alaska Fisheries Science Center—National Oceanic and Atmospheric Administration (NOAA) Fisheries and Joint Institute for the Study of the Atmosphere and Ocean (JISAO), University of Washington, Seattle, Washington 98195, USA
³Alaska Department of Fish and Game, Juneau, Alaska 99802, USA

ABSTRACT:
Over a decade after the Cook Inlet beluga (Delphinapterus leucas) was listed as endangered in 2008, the population has shown no sign of recovery. Lack of ecological knowledge limits the understanding of, and ability to manage, potential threats impeding recovery of this declining population. National Oceanic and Atmospheric Administration Fisheries, in partnership with the Alaska Department of Fish and Game, initiated a passive acoustics monitoring program in 2017 to investigate beluga seasonal occurrence by deploying a series of passive acoustic moorings. Data have been processed with semi-automated tonal detectors followed by time intensive manual validation. To reduce this labor intensive and time-consuming process, in addition to increasing the accuracy of classification results, the authors constructed an ensembled deep learning convolutional neural network model to classify beluga detections as true or false. Using a 0.5 threshold, the final model achieves 96.57% precision and 92.26% recall on testing dataset. This methodology proves to be successful at classifying beluga signals, and the framework can be easily generalized to other acoustic classification problems.


I. INTRODUCTION

A. Background

The beluga whale (Delphinapterus leucas), also known as the white whale, is a highly social species that lives in large groups. Like other odontocetes, this species relies heavily on sound. They produce impulsive signals to find and capture prey to feed (Au, 1993), and also produce tonal, impulsive, and combined signals to communicate (Sjare and Smith, 1986). In the U.S., there are five separate populations of beluga whales, all within Alaska. Of those five, the Cook Inlet population is the smallest and has declined by about 75% since 1979. Subsistence hunting contributed to the initial population decrease, but this practice was regulated in 1999, with the last hunt occurring in 2005 (National Marine Fisheries Service, 2008). Still, the Cook Inlet beluga whale population has yet to recover. In 2008, this population was listed as endangered under the Endangered Species Act with hopes that the population would begin to recover. Now, more than a decade later they continue to decline with a current population estimate of 279 whales (Shelden and Wade, 2019).

In order to better assess the seasonal distribution of this population for management purposes, as well as increase our understanding of the effects of noise on beluga whales, there is critical need for accurate and reliable tools that allow the description of beluga whales’ presence. Historically, this information has been collected from annual aerial surveys and vessel and shore-based photo identification surveys that use unique markings to identify individuals (e.g., Shelden et al., 2015). More recently, scientists have begun using passive acoustic monitoring to listen for beluga whale signals and describe their long-term spatial and temporal occurrence and habitat use by maintaining a network of underwater acoustic recording moorings (Castellote et al., 2020). However, the acoustic detector performance is degraded when background noise is high due to strong current conditions (including self-noise such as flow noise), commercial shipping, and other diverse anthropogenic noise sources (Castellote et al., 2018). These variables along with often high levels of background noise trigger false detections in the standard semi-automated processing of acoustic recordings, incurring tedious manual validation of automated classifications, which is labor and time intensive.

B. Related work

The classification of biological sounds using machine learning techniques, especially with deep learning models, has recently begun to attract the attention of researchers in the field of bioacoustics.

For example, a random forest model was implemented to improve the analytical efficiency in monitoring and detection of passerine nocturnal flight calls, in which spectrotemporal features are generated (Ross and Allen, 2013). A more popular approach is to conduct image classification
TABLE I. PAMGuard whistle and moan detector parameter settings used to detect beluga whale calls in sound recordings from passive acoustic moorings in Cook Inlet, AK, following Keating et al. (2015).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFT length</td>
<td>1024</td>
</tr>
<tr>
<td>FFT hop</td>
<td>512</td>
</tr>
<tr>
<td>Window type</td>
<td>Hann</td>
</tr>
<tr>
<td>Sample rate</td>
<td>24 kHz</td>
</tr>
<tr>
<td>Time resolution</td>
<td>42.67 ms</td>
</tr>
<tr>
<td>Frequency resolution</td>
<td>23.44 Hz</td>
</tr>
<tr>
<td>Time step size</td>
<td>21.33 ms</td>
</tr>
<tr>
<td>Median Filter length</td>
<td>61</td>
</tr>
<tr>
<td>Average Subtraction</td>
<td>0.02</td>
</tr>
<tr>
<td>Gaussian Kernel Smoothing</td>
<td>on</td>
</tr>
<tr>
<td>Threshold</td>
<td>8.0 dB</td>
</tr>
<tr>
<td>Minimum Hz</td>
<td>1000</td>
</tr>
<tr>
<td>Maximum Hz</td>
<td>12000</td>
</tr>
<tr>
<td>Type</td>
<td>connect 8 (sides and diagonals)</td>
</tr>
<tr>
<td>Minimum length</td>
<td>5 time slices</td>
</tr>
<tr>
<td>Minimum total size</td>
<td>50 pixels</td>
</tr>
<tr>
<td>Crossing and joining</td>
<td>Re-link across joins</td>
</tr>
<tr>
<td>Max cross length</td>
<td>5 time slices</td>
</tr>
</tbody>
</table>

TABLE II. Eight datasets processed for model training, validation and testing. These datasets are collected from seven locations in Alaska, covering both summer and winter seasons.

<table>
<thead>
<tr>
<th>Mooring location</th>
<th>Recording Period</th>
<th>Area Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cairn Point</td>
<td>5/2018–9/2018</td>
<td>Knik Arm, highly disturbed zone, port turning basin, continuous dredging operations, nearby military base</td>
</tr>
<tr>
<td>Point Woronzof</td>
<td>5/2018–9/2018</td>
<td>Moderate disturbance, access area to Knik Arm</td>
</tr>
<tr>
<td>Little Susitna River</td>
<td>5/2018–9/2018</td>
<td>Moderate disturbance, important foraging grounds</td>
</tr>
<tr>
<td>Chuitna River</td>
<td>5/2018–9/2018</td>
<td>Highly disturbed zone, oil and gas operation area.</td>
</tr>
<tr>
<td>Kalgin Island</td>
<td>10/2017–4/2018</td>
<td>Moderate disturbance, exposed to shipping traffic, winter concentration area</td>
</tr>
</tbody>
</table>
whale calls). In relatively quiet areas (i.e., areas with low levels of anthropogenic noise disturbance and low currents), the whistle and moan detector’s precision is close to 95%, while in areas of elevated anthropogenic disturbance and/or high currents (i.e., self-noise), the precision can be as low as 20%-60%.

**B. Classification models using convolutional neural networks (CNN)**

Using custom written scripts in Python 3.6, spectrograms were extracted from audio files (with NFFT = 256, non overlapping Hanning window). Each spectrogram, resized as 300 pixels by 300 pixels, was generated from a 2-s audio segment from the start time of each of the whistle and moan detections (see Fig. 2, for example). Since the intent of the machine learning model was to replicate the human validation process of the semi-automated detectors’ results described in Sec. II.A, only audio segments corresponding to PAMGuard detections were processed. The manually validated labeling of detections for each dataset was used as the ground truth to train and evaluate the binary classification model.

A total of eight datasets processed with the PAMGuard whistle and moan detector were used in this study (Table II), and a series of 2-s spectrograms have been generated from the raw audios. Based on the manually validated detections, we generated 89,000 spectrograms from true detections, and 146,000 spectrograms from false detections. Among all the available labeled data from true and false detections, they were split into training and testing portions, according to a 70/30 ratio; and then further split the training data into training and validation using the same ratio. As a result, there are 49% of the data for training, 21% for
validation, and the remaining 30% for out-of-sample testing. In order to train the model to better differentiate between the true beluga whale calls from various other types of acoustic signals (such as human voices, sonar signals, waves, and rain), we also generated 25,000 spectrograms from randomly chosen clips when there were no detections, and included them into the training and validation datasets. The spectrograms from true detections were assigned with label 1, and those from false detections and without detections were assigned with label 0. All the generated spectrograms along with their assigned labels were used as input for building a binary image classification model. The model is initially fit on the training dataset, and the validation dataset provides an unbiased evaluation of the model which tunes the model’s hyperparameters to avoid overfitting; finally, the testing dataset is used to evaluate how the model generalizes to an unseen dataset.

With this data, we trained four different CNNs to implement binary classification of beluga whale calls, and the final prediction is an ensemble of the weighted average of each individual model.

Among these models, we trained one CNN model from scratch, while the remaining three models were built upon pre-trained model weights known as transfer learning (Oquab et al., 2014). Transfer learning is a machine learning technique where a model trained on one task (or domain) is re-purposed on a second related task (or domain). For example, the model weights were initially trained on ImageNet (Deng et al., 2009) database with 1000 classes of objects, but their pre-trained weights can be leveraged by a different task or domain. This approach is effective because the images were trained on a large corpus of photographs and require the model to make predictions on a relatively large number of classes, in turn, requiring that the model efficiently learn to extract features from photographs in order to perform well on the problem.

1. **Build CNN using AlexNet architecture**

AlexNet (Krizhevsky et al., 2012) is a deep neural network which learns visual knowledge from images to classify a given image. Based on the original architecture of the AlexNet model, we made minor modifications by adding batch normalization layer after each max-pooling layer to solve the internal covariate shift problem (this refers to the change in the input distribution to each layer of the deep neural network, where even small changes might get amplified down the network) as it forces the input of every layer to have approximately the same distribution, and by adding dropout layer (with dropout rate $= 0.5$) after each fully connected layer to reduce interdependent learning amongst the neurons and thus to prevent over-fitting.

2. **Fine-tuning with pre-trained VGG16 model**

VGG16 is a convolutional neural network model proposed by (Simonyan and Zisserman, 2015). It makes improvements over AlexNet by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer, respectively) with multiple $3 	imes 3$ kernel-sized filters one after another.

Instead of training the entire model from scratch, we used transfer learning and fine-tuning hyperparameters. With transfer learning, the weights of the first 13 layers from pre-trained VGG16 model were retained, and we fine-tuned the hyperparameters of the last three layers, and added a fully connected layer, a dropout layer, and an output layer in the end.

3. **Fine-tuning with pre-trained ResNet model**

Deep residual networks (He et al., 2016) were arguably the most groundbreaking work in the computer vision/deep learning community in the last few years. ResNet makes it possible to train up to hundreds or even thousands of layers and still achieves compelling performance. Because of the compelling results, ResNet quickly became one of the most popular architectures in various computer vision tasks.

Without training the entire model from scratch, we used the pre-trained weights of ResNet, and fine-tuned the parameters by adding a fully connected layer, dropout layer and an output layer.

4. **Fine-tuning with pre-trained DenseNet model**

DenseNet (Densely Connected Convolutional Networks) (Huang et al., 2017) is one of the latest neural networks for visual object recognition. It is quite similar to ResNet but has some fundamental differences. In DenseNet, a layer receives all the outputs of previous layers and concatenates them in the depth dimension; while in ResNet, a layer only receives outputs from the previous second or third layers. Using a similar approach described in Secs. II and III, we used the
pre-trained weights of DenseNet, and fine-tuned the parameters by adding a fully connected layer, dropout layer and an output layer.

5. Ensemble modeling

Ensemble modeling is a process where multiple diverse models are created to predict an outcome in order to improve the overall performance, either by using many different modeling algorithms or using different training datasets. The motivation for using ensemble models is to reduce the generalization error of the prediction. As long as the base models are diverse and independent, ensemble methods help to minimize the prediction error and improve the stability of the predictions. To ensemble, we take the weighted average of each individual model’s predictions while their optimal weights are derived towards the optimization of the overall accuracy.

The developed model code for this study is open source and freely available at https://github.com/microsoft/belugasounds.

III. RESULTS

To evaluate and compare the performance of different models on the same testing dataset, we report four key metrics: precision, recall, accuracy and area under a curve (AUC). Precision measures the proportion of positive identifications that were actually correct [i.e., TP/(TP+FP)]; recall measures the proportion of actual positives that were identified correctly [i.e., TP/(TP+FN)]; and accuracy measures the overall percentage of correct classifications [i.e., (TP + TN)/(TP+FP+TN+FN)]. While precision, recall and accuracy are dependent on the choice of threshold score, average precision provides an aggregate measure of performance across all possible classification thresholds, and it is not affected by the class imbalance.

With a default neutral threshold score of 0.5 used for classification, the results show that each individual CNN model has relatively similar performance, while leveraging the pre-trained model with transfer learning yields better results, especially in recall (Table III). The ensemble model has a 1.5% increase in precision when compared to best individual model results, as well as the highest accuracy and AUC among all models. Testing on the out of sample dataset, the CNN models were able to satisfactorily classify most detections and achieve high accuracy, while the output probabilities also indicate the level of confidence by the model predictions.

While the ensemble model has high confidence scoring for most observations (that is, with probability score $>0.9$ for true positives, and $<=$0.1 for true negatives, see Fig. 3), in the real implementation, the classification threshold score can be tuned with flexibility, for example, from 0.5 to 0.8 or 0.9, to include fewer false positives and thus to get higher precision, or vice versa.

The amount of time required for an expert analyst to manually validate the detections triggered by the PAMGuard whistle and moan detector from the eight datasets included in this study, ranged from 56 to 112 h per dataset (on average 70 h). As a comparison, after the ensemble model has been trained (it is a one-time task, taking about 30 h for training four individual models and their ensembles), on average it only takes 4.5 h for the ensemble model to score each dataset with the same detections using Azure Deep Learning Virtual Machine NC24 (with 224 GB memory and 4 x K80 GPU).

IV. DISCUSSION

The main objective of this study was to test the performance of a machine learning approach to support the manual validation process of semi-automated detections of beluga whale acoustic signals. Compared to an expert analyst manually validating the detections used in this study, we calculate that scoring by our model could save 93% of that time. The CNN ensemble model presented here has high accuracy in differentiating Cook Inlet beluga whale sounds from various types of noises. This methodology could replace the labor-intensive manual validation process. In some cases, faint beluga whale calls were missed by the PAMguard detector or wrongly labeled by manual validation process suggesting the detection process could be improved; on the other hand, there are some cases where various types of signals that look similar to beluga whale calls may be wrongly classified by the model (see Fig. 4). To further improve the ability of the model to distinguish such false signals from true beluga whale calls, a simple solution is to collect more comprehensive data to be included in the training dataset. A manual inspection of these events revealed most are associated with tonal elements of distant ship noise, outboard engine noise, or ice noise from mechanical stress.

Our model was built based on eight datasets from seven locations within the Cook Inlet beluga whale critical habitat, and the majority of the datasets were collected during the ice-free season (May to September), while only one location

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>Accuracy</th>
<th>AUC</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN with AlexNet Architecture</td>
<td>0.9504</td>
<td>0.8550</td>
<td>0.9362</td>
<td>0.9783</td>
<td>0.9681</td>
</tr>
<tr>
<td>Fine-tuning with pre-trained VGG16</td>
<td>0.9383</td>
<td>0.8894</td>
<td>0.9437</td>
<td>0.9798</td>
<td>0.9700</td>
</tr>
<tr>
<td>Fine-tuning with pre-trained ResNet50</td>
<td>0.9422</td>
<td>0.9195</td>
<td>0.9544</td>
<td>0.9868</td>
<td>0.9796</td>
</tr>
<tr>
<td>Fine-tuning with pre-trained DenseNet</td>
<td>0.9503%</td>
<td>0.9248</td>
<td>0.9588</td>
<td>0.9898</td>
<td>0.9842</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.9656</td>
<td>0.9226</td>
<td>0.9633</td>
<td>0.9906</td>
<td>0.9859</td>
</tr>
</tbody>
</table>
included data for both ice-free and winter seasons. Considering the underwater acoustic environment may vary seasonally as well as spatially, it would be helpful to expand this study for both seasons in all selected locations. Refreshing the model with a more comprehensive and larger size training datasets will likely yield improved accuracy.

In building each CNN model, we used the default viridis color scales when generating spectrograms, although some previous studies have shown that the choice of color scales may impact model performance (Amiriparian et al., 2017). Even though the average frequency range of beluga whale calls in this dataset is from 2.0 to 5.9 kHz, we did not apply this restriction when plotting spectrograms to make our approach more generic. On the other hand, it will yield some false detections due to convolutional neural models’ translation invariant property (that is, the model cannot distinguish noises that have similar frequency modulation but are outside the normal frequency range of beluga whale calls).

Building deep neural network models typically requires a large amount of annotated data due to the multitude of parameters that need to be optimized. However, the application of transfer learning with pre-trained models makes it a viable approach when only small datasets are available for training purposes. While transfer learning leverages the learning from one task, which is generally trained on a large dataset, it does not require learning from scratch for the new task, which directly addresses the smart parameter initialization point for training neural networks. So, for situations where a large amount of data are being collected or archived, but manually validated subsets are small (e.g., only a subset of data are processed, typical of large underwater acoustic efforts), this method is still applicable. While each individual model performs well already, ensembling them with their respective optimal weights will further improve the prediction accuracy and make the final outcome more robust.

The manual validation of detections as false or true beluga whale acoustic signals is the most time-consuming, expensive, and subjective process of all the currently applied acoustic data analysis methods in this long-term study. Because the ground truth data are predefined by what the PAMGuard whistle and moan detector was able to identify as beluga signal, the annotated data used in this study may not be completely representative of the diversity of beluga calls and whistles present in the datasets. However, this approach ensures that, even if the methods to validate semi-automated detections are improved, the detection performance is equivalent and comparable to previous acoustic

FIG. 3. (Color online) Top left: histogram of predicted probabilities for true positive (i.e., manually validated as “with beluga whale call”) testing dataset consisting of 15 000 observations of manually validated detections with beluga whale signals. Top right: histogram of predicted probabilities for true negative (i.e., manually validated as “false beluga whale call”) testing dataset consisting of 30 000 observations of manually validated periods false beluga whale signals. Bottom left: receiver operating characteristic (ROC) curve. Bottom right: precision-recall curve.
monitoring efforts. Once the manual validation process is replaced by this initial machine learning framework, the next step is to add a detection process routine before the CNN ensemble-based classification, and conduct further comparative testing between this new procedure and the standard analysis procedure based on manually validated semi-automated detections. By doing this, an easy end-to-end processing methodology fully based on deep learning CNN models would be achieved, providing a more comprehensive, efficient, objective, and accurate approach to new efforts in the description of Cook Inlet beluga whale spatial and temporal distribution within the critical habitat.

Finally, the methodology and implementation framework presented in this study can be easily adopted by other similar bioacoustics applications, where target signals require manual validation. This study describes initial steps for placing deep learning CNN analysis, based on weighted output from a model ensemble, as the natural evolution of analysis methods for passive acoustic monitoring data.

ACKNOWLEDGMENTS

The authors would like to thank those who participated in this project for their expertise and support: Jennifer Keating McCullough, for providing insight and assistance with the PAMGuard whistle and moan detector; Benjamin X. Hou, for his support in the integration of machine learning into the IT system of the Marine Mammal Laboratory; the Marine Mammal Program of the Alaska Department of Fish and Game, for their partnership in the

FIG. 4. (Color online) Examples of spectrograms which have been validated as “False beluga whale calls” but were wrongly predicted as beluga whale signals by the model. Upper left—tonal component of noise generated by ice mechanical stress, upper right—outboard motor noise, lower left and right—large size ship noise.
passive acoustic monitoring efforts in Cook Inlet; the NMFS Alaska Regional Office, for their logistics support in Anchorage. This work was supported by AI for Earth grants at Microsoft. Our appreciation goes to Dan Morris for connecting different parties for fruitful discussions and helpful online materials. Passive acoustic data collection was funded by the NOAA Fisheries Species Recovery Grants to States Program.


