Deep learning detection and classification of baleen whale vocalizations using a novel data representation

Mark Thomas[1,2,*], Bruce Martin[2], Katie Kowarski[2], Briand Gaudet[2], and Stan Matwin[1]

[1]*Dalhousie University, Faculty of Computer Science* [2]*JASCO Applied Sciences*

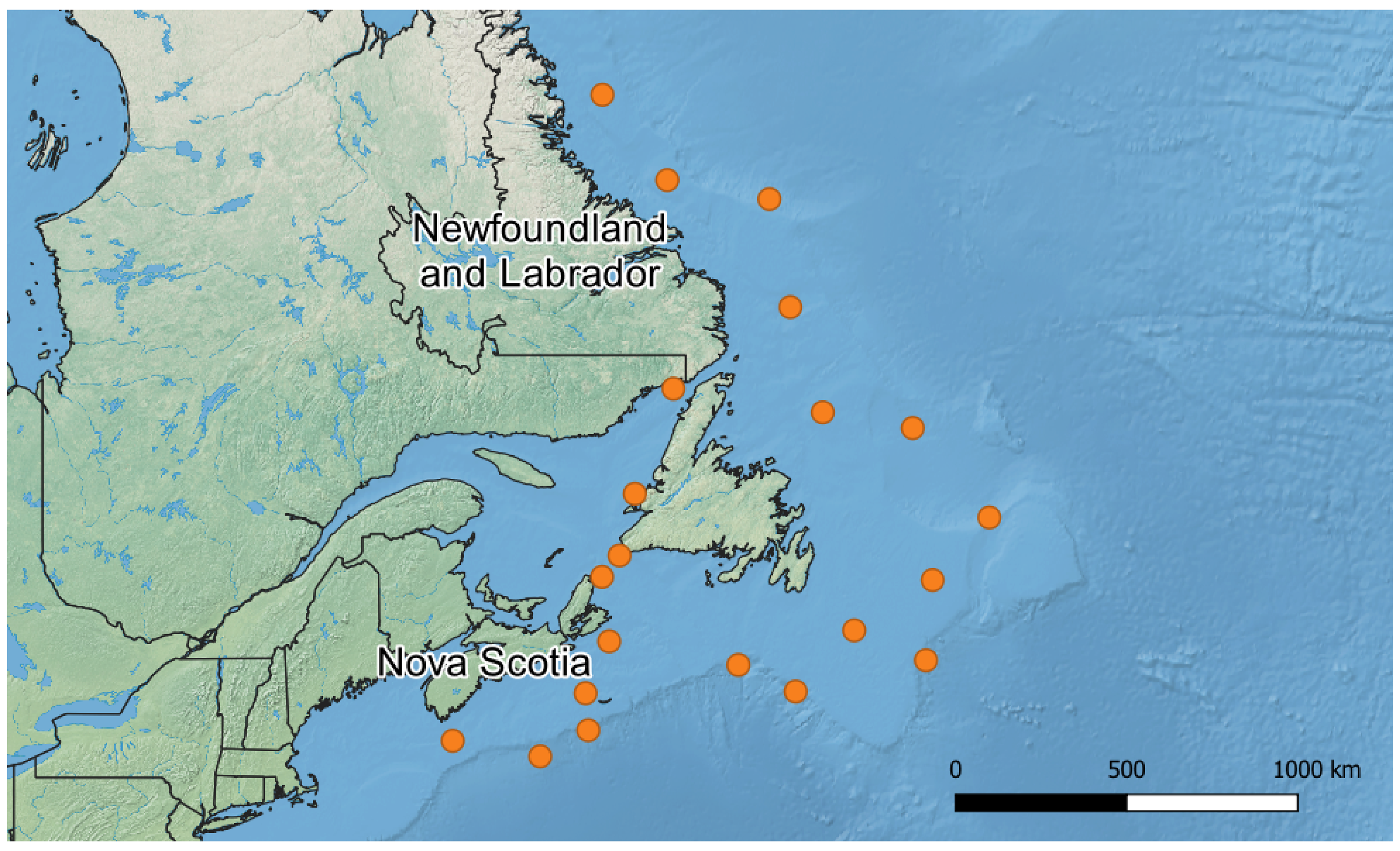
[*]`mark.thomas@dal.ca`

Introduction

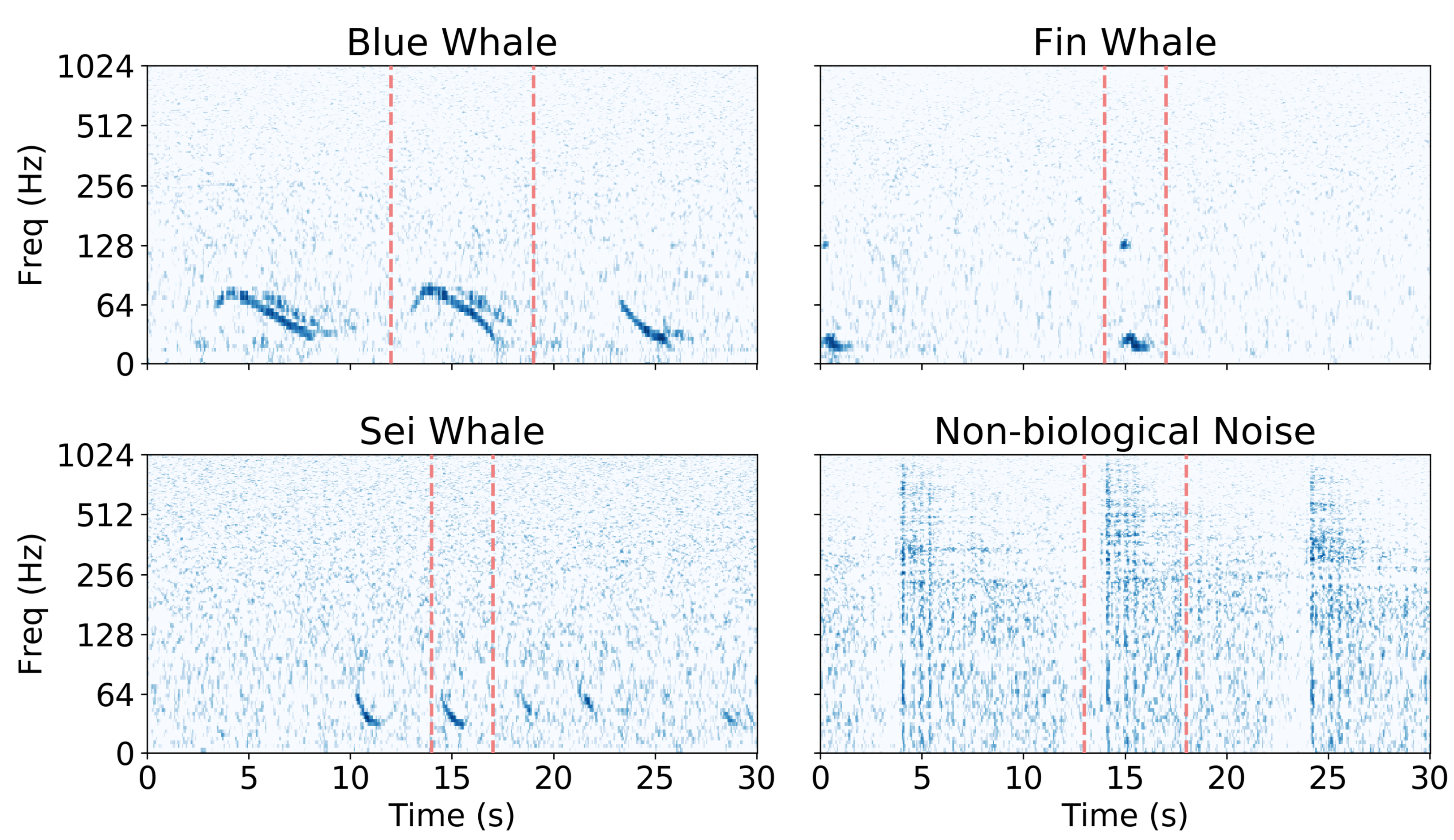
- Marine biologists use acoustic data collected through Passive Acoustic Monitoring (PAM) to determine presence, abundance, behaviour and migratory patterns of marine life, especially marine mammals
- Collections of acoustic recordings obtained through PAM are very large, making complete human analysis infeasible
- Can we use deep learning to detect and classify marine mammal vocalizations in acoustic recordings?

Acoustic Recordings and Training Data

- The acoustic recordings were collected by JASCO Applied Sciences during the summer and fall months of 2015 and 2016 in the areas surrounding the Scotian Shelf



- The recordings were analyzed by marine biologists producing annotations pertaining to marine mammal vocalizations and other acoustic sources labelled as "non-biological"
- We focus on identifying three species of baleen whales with similar call types (blue, fin, and sei whales) against non-biological and ambient sources
- We use spectrograms of the acoustic recordings containing each annotation and treat this problem as an image-classification task



Source	Training	Validation	Test
Blue Whale	2692 (6.23%)	601 (6.49%)	574 (6.20%)
Fin Whale	15118 (35.01%)	3244 (35.06%)	3272 (35.36%)
Sei Whale	1701 (3.94%)	332 (3.59%)	383 (4.14%)
Non-biological	2078 (4.81%)	449 (4.85%)	398 (4.30%)
Ambient	21589 (50.00%)	4626 (50.00%)	4627 (50.00%)

Stacked and Interpolated Spectrograms

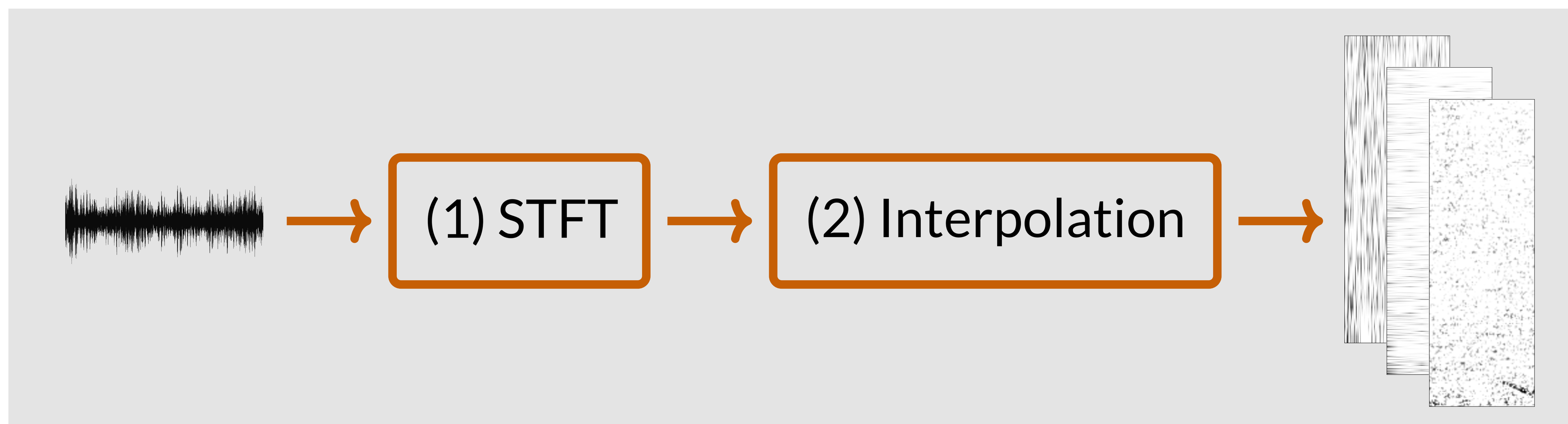
- Experts in marine biology use multiple spectrograms with different resolutions when analyzing acoustic recordings
- How can we exploit the strategy used by marine biologists without simply training multiple classifiers?
  - Generate k spectrograms using multiple sets of parameters to the Short-time Fourier Transform

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[m-n]e^{-j\omega m} \tag{1}$$

  - Interpolate the original spectrograms over a pre-defined resolution

$$\omega = \omega_i + \frac{\omega_{i+1} - \omega_i}{n_{i+1} - n_i}(n - n_i) \tag{2}$$

  - Stack the interpolated spectrograms to form a k-channel tensor

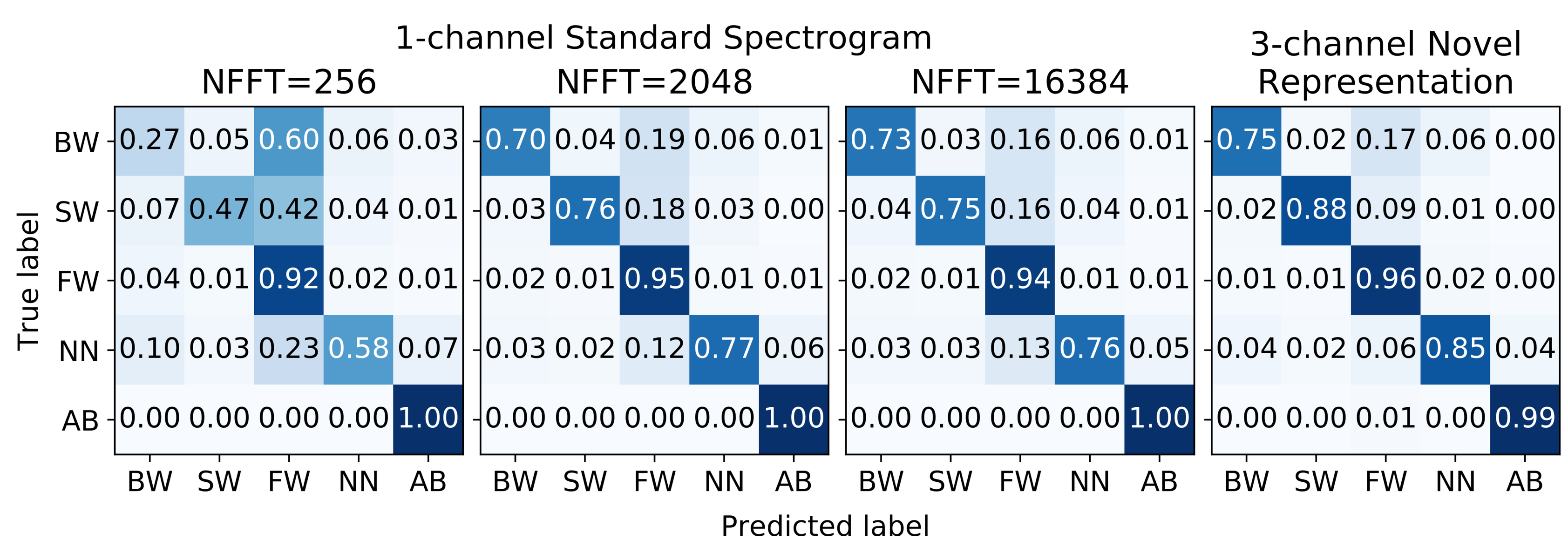


Neural Network Architecture and Training Details

- We train a commonly used deep Convolutional Neural Network (CNN) known as ResNet-50 [1]
- A cross-entropy loss function was optimized using Stochastic Gradient Descent (SGD) with momentum
- Other training parameters: batch size=128, learning rate=0.001 with exponential decay ($\lambda = 0.01$) every 30 epochs

Experimental Results

	1-channel Standard Spectrogram			3-channel Novel
	NFFT=256	NFFT=2048	NFFT=16384	Representation
Accuracy	0.88512	0.94326	0.94196	**0.95331**
Precision	0.71979	0.86621	0.85686	**0.89265**
Recall	0.64634	0.83627	0.83814	**0.88409**
F-1 Score	0.67394	0.85003	0.84697	**0.88735**



References and Acknowledgements

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770--778, 2016.

Collaboration between researchers at JASCO Applied Sciences and Dalhousie University was made possible through an NSERC Engage Grant. The acoustic recordings were collected by JASCO Applied Sciences as part of the Environmental Studies Research Fund (ESRF) program.