IMPROVING THE ROBUSTNESS OF RIGHT WHALE DETECTION IN NOISY CONDITIONS USING DENOISING AUTOENCODERS AND AUGMENTED TRAINING

W. Vickers, B. Milner

School of Computing Sciences
University of East Anglia
Norwich, UK
{w.vickers, b.milner}@uea.ac.uk

R. Lee

Gardline Environmental
Gardline Geosurvey Limited
Great Yarmouth, UK
robert.lee@gardline.com

ABSTRACT

The aim of this paper is to examine denoising autoencoders (DAEs) for improving the detection of right whales recorded in harsh marine environments. Passive acoustic recordings are taken from autonomous surface vehicles (ASVs) and are subject to noise from sources such as shipping and offshore construction. To mitigate the noise we apply DAEs and consider how best to train the classifier by augmenting clean training data with examples contaminated by noise. Evaluations find that the DAE improves detection accuracy and is particularly effective when the classifier is trained on data that has itself been denoised rather than using a clean model. Further, testing on unseen noises is also effective particularly for noises that exhibit similar character to noises seen in training.

Index Terms— cetacean detection, autonomous surface vehicles, CNN, autoencoder, right whale

1. INTRODUCTION

The aim of this work is to develop robust methods of detecting marine mammals from autonomous surface vehicles (ASVs). This is important for population monitoring and for mitigation as many species are endangered and protected by environmental laws. Specifically, we consider the challenge of detecting North Atlantic right whales (Eubalaena glacialis) in situations where they may be approaching potentially harmful and noisy offshore activities. Detecting their presence before they enter a mitigation zone both protects the animal and avoids costly shutdowns of sub-sea operations. For the right whale in particular, these are one of the most endangered marine mammals and at risk of extinction with as few as 350 individuals remaining [1]. Traditional methods for detecting marine mammals use human observers on-board ships, but more recently ASVs have been used as they offer a cheaper solution that can operate in zero visibility conditions [2].

Many machine learning techniques have been applied to cetacean detection. Vector quantisation and dynamic time warping have been effective in detecting blue and fin whales from frequency contours extracted from spectrograms [3]. Hidden Markov models (HMMs) have also been used to recognise low frequency whale sounds using spectrogram features [4]. Comparisons have also been made between between artificial neural networks (ANNs) and spectrogram correlation for right whale detection [5]. A study of various time-series classification and deep learning approaches to right whale detection found convolutional neural networks (CNNs) to give highest accuracy [6]. Further studies have also found success in using CNNs for right whale detection when compared to other classification models such as recurrent neural networks [7, 8].

Right whale recordings are susceptible to corruption from various noise sources at different signal-to-noise ratios (SNRs) depending on the distances of the right whale and noise source from the hydrophone. Noise presents a challenge to classification problems ranging from speech recognition to image identification [9, 10]. Many methods have been proposed to combat noise and can be broadly categorised into those that match the underlying model to the operating environment and those that remove noise before classification [11, 12]. We consider both and investigate augmentation strategies to make training closer to test conditions. Given the changing nature of operating conditions, it is not always possible to match the training conditions and so we examine performance not only for matched models (same noise type and SNR) but also for models trained on a range of noises that both include and exclude samples from the test condition. As a second approach we apply denoising autoencoders to remove noise before classification as they perform well on other audio tasks [13, 12, 14, 15].

The remainder of the paper is organised as follows. Section 2 gives a brief introduction to right whale calls and typical sources of marine noise. Section 3 introduces the baseline CNN right whale detector, while the implementation of denoising autoencoders is explained in Section 4. Experimental results are presented in Section 5 that evaluate right whale detection accuracy in varying noise types and SNRs.

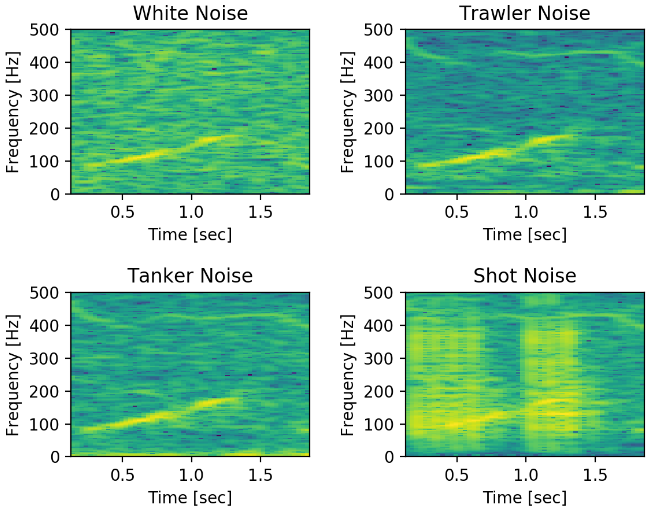

Fig. 1. *Spectrograms of right whale upsweep contaminated with white, trawler, tanker and shot noise at -5dB.*

2. ACOUSTIC CHARACTERISTICS OF RIGHT WHALES IN MARINE ENVIRONMENTS

Right whales emit a range of vocalisations and in this work we focus on their up-sweep tones [16]. These sweep from approximately 60Hz to 250Hz typically lasting for one second, although these calls are not always consistent and vary in duration and frequency [17]. Calls can be difficult to hear, and visualise in spectrograms, as low frequency bands are often masked by sounds from passing ships, drilling and piling activities, seismic exploration or interference from other marine mammals [18]. In many cases the noises overlap in frequency with the right whale calls and make detection difficult.

We consider four types of noise as typical contaminants of right whale recordings, namely tanker noise, trawler noise, shot noise and white noise. Figure 1 shows an example right whale upsweep contaminated by each of these noises at an SNR of -5dB. Tanker and trawler noises are chosen for investigation as they represent common types of marine noise and are similar in character, creating horizontal bands of energy in spectrograms from their engine and propeller. Shot noise represents noises that arise from piling and seismic exploration and has an impulsive character that introduces vertical bands of energy in the spectrogram. White noise is a more generic noise type and is chosen as it affects all time-frequency regions of the spectrogram.

3. BASELINE CNN DETECTOR

The baseline CNN right whale detector is based on earlier work that compared a range of deep learning techniques [6]. Spectrogram features are extracted from the input audio using an N=256-point sliding window with a frame slide of S=32 samples, with normalisation applied to transform amplitudes into the range 0 to 1. A CNN encoder, M_S, maps input spectrogram features into a new space using three convolutional layers, each followed by a max pooling layer. This outputs into a network comprising two dense layers for classification, C [6]. Each convolutional layer uses 3×3 filter kernels with 32, 64, 128 filters on subsequent layers. Max pooling layers use a pool size of 2×2 and have ReLU activation functions applied to their outputs. At the edges of the input, zero-padding is applied to convolutional layers to maintain the size of the output. After the last max pooling layer a dropout of 0.5 is applied. The two dense layers use 200 and 50 nodes with a ReLU activation function. The final dense layer has a sigmoid activation function and outputs the probability of a right whale being present. For training, an Adam optimiser is used with a learning rate of 0.001 and binary cross-entropy as the loss function [19]. Training used 200 epochs and was repeated 10 times for each test. The model with highest validation accuracy was used for testing and accuracies calculated as an average over all 10 tests.

4. DENOISING AUTOENCODER

The denoising autoencoder takes as input the spectrogram features and outputs into the classifier, C, for a detection result. The DAE function, M_A, encodes input features using 3 convolutional layers, with each using a 3×3 filter kernel with 32 filters. A max pooling layer is applied after each convolution layer to compress the feature further. Each max pooling layer uses a pool size of 2×2 and has a ReLU activation function on the output. Previous testing found that 32 filters per layer and a network depth of 3 achieved highest accuracy across the range of SNRs. The DAE uses non-corrupted samples as validation for training. Binary cross-entropy produces a loss between the output and non-corrupted samples with Adam being utilised as the optimiser. Training was carried out over 100 epochs for all tests involving the DAE.

With the DAE, we consider two methods of training the final classifier. The first approach uses noise-free training data to create a clean-trained classifier, with the assumption that the DAE is able to remove the noise. The second approach trains the classifier on noisy training data that has been passed through the DAE. This data is more likely to match the denoised test data and contains residual noise that the DAE was either unable to remove or that the DAE introduced.

5. EXPERIMENTAL RESULTS

The purpose of the experiments is to explore data augmentation for classifier training and denoising autoencoders on right whale detection across both seen and unseen noise types. The first experiment considers the baseline CNN method of detection and examines the effect of augmenting clean training data with varying quantities of noise data, both matched and unmatched to the test conditions. The second set of tests explores the effectiveness of the denoising autoencoder in noisy conditions. Finally, a third experiment compares data augmentation with the denoising autoencoders in both seen and unseen noise conditions.

Right whale recordings were obtained from the Cornell NRW Buoys data, recorded in the Cape Cod region [20]. Audio recordings are arranged as two-second segments that either contain a right whale sound or do not, as checked by human experts. Given the low frequency of right whale calls the audio was downsampled to 1 kHz, as previous work established that this introduces no loss in accuracy [6]. Recordings are divided into non-overlapping training, validation and test sets, using a split of 70:15:15, which gives sizes of 10,000, 2,142 and 2,142. All sets contain equal proportions of segments with and without right whales and samples are taken randomly from the original corpus.

Four noise types are considered for the experiments - tanker noise, trawler noise, shot noise and white noise, as shown in Figure 1. Noisy data is created by adding noise samples to the whale recordings at SNRs from +5dB down to -10dB. These cover a range of reception conditions that represent signals received from right whales at both close and long range distances.

5.1. Augmented model training

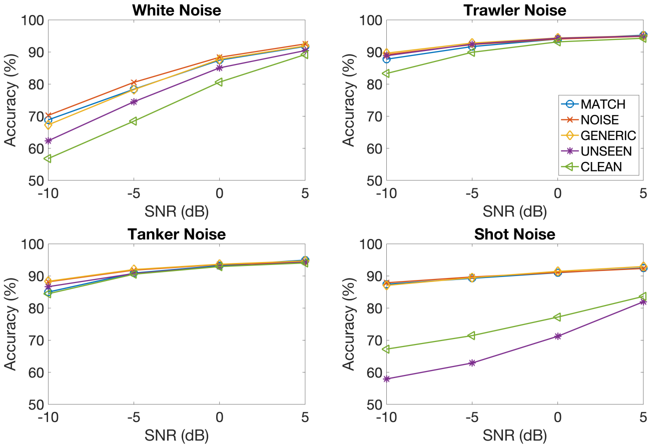

Fig. 2. *Right whale detection accuracies with training data augmentation in four noises at SNRs from -10dB to +5dB.*

The first tests use the baseline CNN method of Section 3 and examine the effect that augmenting the training data with different noise types has on accuracy. Tests are performed on the four noises with detection accuracy shown in Figure 2 across all SNRs. Each noise condition is evaluated against five different models - trained on only clean data (CLEAN), matched to the specific test condition (MATCH), specific to the noise type but across all SNRs (NOISE), trained on all four noises at all SNRs (GENERIC) and trained on the three noises that exclude the noise type under test (UNSEEN).

The clean-trained model (CLEAN) generally performs worst due to the mismatch between the training and test conditions, while the matched model (MATCH) performs well. The GENERIC model, trained on data from all conditions, also performs very well. Removing the test noise type from the training data, to give an UNSEEN test condition has little effect on the tanker and trawler noise conditions but reduces performance on white and shot noise. This we attribute to tanker and trawler noises being similar and so the UNSEEN model has been exposed to a similar noise type. For the white and shot noises, these have more unique spectral properties and in the UNSEEN case the models have not learnt any of their character, hence the larger reduction in performance.

5.2. Denoising autoencoder

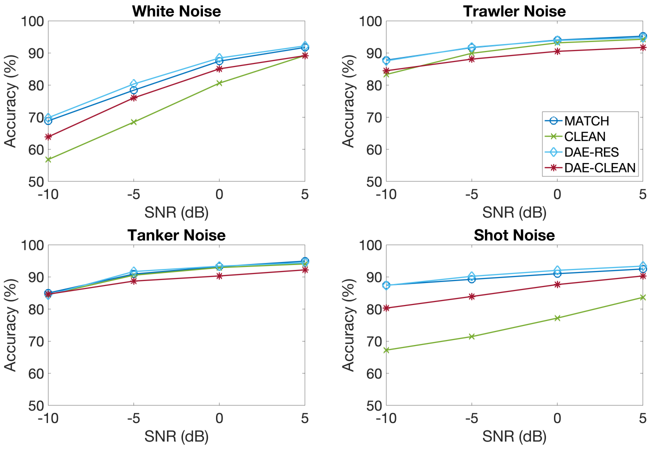

Fig. 3. *Right whale detection accuracies for denoising autoencoders in four noises at SNRs from -10dB to +5dB.*

The second set of experiments uses the denoising autoencoder of Section 4 and tests on the same set of four noise types and four SNRs. In method DAE-CLEAN, the DAE is trained on data matched to the test condition with its output applied to a clean-trained classification model. In method DAE-RES, the same matched DAE is used but the classification model is now trained on the residual signal after noise removal. For comparison, both clean trained CNN models (CLEAN) and matched CNN models (MATCH) are included.

Figure 3 shows detection accuracies and reveals that applying the output of the DAE to the residual-trained classifier is better than applying it to the clean-trained model. This we attribute to the DAE not being able to remove entirely all noise and so applying its output to a model trained on a similar signal performs better than a clean trained model. Compared to using the clean model with no DAE, accuracy in tanker and trawler noises for DAE-CLEAN is worse but is improved when applied to DAE-RES. Performance is again different

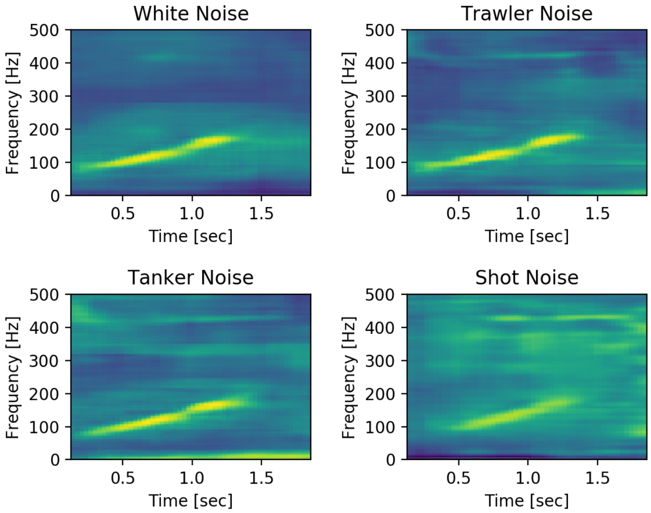

Fig. 4. *Spectrograms of DAE enhanced right whale upsweeps contaminated with white, trawler, tanker and shot noises at -5dB.*

for the white and shot noises, where performance with the clean model is much worse than for tanker and trawler noises. However, applying the DAE before classifying with the clean model (DAE-CLEAN) improves accuracy and moving to the residual model (DAE-RES) improves performance even further and is comparable to matched models (MATCH).

To show the denoising ability of the DAE, Figure 4 shows denoised spectrograms of the signals shown in Figure 1. Upsweeps contaminated by tanker, trawler and white noises are recovered well, with small amounts of the original noise remaining as can be seen in tanker and trawler noises. Shot noise has also been largely removed but the DAE has introduced distortion and is less effective at recovering the upsweep regions masked by specific shots.

5.3. Combined denoising autoencoder and augmentation

The final tests investigate the DAE in both seen and unseen noisy test conditions with results shown in Figure 5. DAEs are trained on both seen (DAE-RES-GENERIC) and unseen (DAE-RES-UNSEEN) noises across all SNRs with the classifier trained on residual signals, as Section 5.2 showed this to be better than using a clean-trained classifier. For comparison, three non-DAE models are evaluated, namely one trained on noise free data (CLEAN), one on all noises and SNRs (GENERIC) and one on the three noises that are not under test (UNSEEN).

The models trained on all noise types and SNRs (GENERIC and DAE-RES-GENERIC), achieve either highest or close to highest detection accuracies across all conditions. Excluding the noise type under test from the training data (UNSEEN and DAE-RES-UNSEEN) has little effect on accuracy for tanker and trawler noises but reduces performance for white noise and shot noise. In fact, for shot noise the unseen noise trained models perform worse than the clean training model (CLEAN), although performance is improved when shot noise is included in the model training, i.e with GENERIC and DAE-RES-GENERIC.

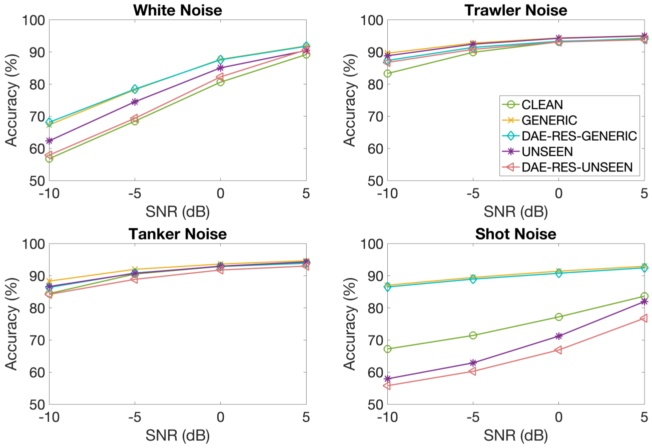

Fig. 5. *Right whale detection accuracies for denoising autoencoders with training data augmentation in four noises at SNRs from -10dB to +5dB.*

6. CONCLUSION

The experiments have shown that augmenting training data with noisy samples improves right whale detection and in many conditions this model outperforms ones trained under matched conditions. Furthermore, only a single model is then required rather than a set of matched models. In the more realistic situation where the noise under test is not included in the model training data, performance still improves in all cases with the exception of when the noise has very different character to the other noises, as was the case for shot noise. That said, detection accuracy for shot noise can be improved by applying a DAE to denoise the signal prior to classification. In fact, the DAE improved accuracy across all noise conditions. Applying the DAE output to classification models trained on residual signals gave higher accuracy than using clean trained models as the denoising process is unable to create an entirely clean representation.

7. ACKNOWLEDGEMENTS

We acknowledge the support of the Next Generation Unmanned Systems Science (NEXUSS) Centre for Doctoral Training, Gardline Geosurvey Limited and NVIDIA.

8. REFERENCES

[1] Scott D. Kraus, Moira W. Brown, Hal Caswell, Christopher W. Clark, Masami Fujiwara, Philip K. Hamilton, Robert D. Kenney, Amy R. Knowlton, Scott Landry, Charles A. Mayo, William A. McLellan, Michael J. Moore, Douglas P. Nowacek, D. Ann Pabst, Andrew J. Read, and Rosalind M. Rolland, "North Atlantic Right Whales in Crisis," *Science*, vol. 309, no. 5734, pp. 561–562, July 2005.

[2] Ursula K Verfuss, Ana Sofia Aniceto, Danielle V Harris, Douglas Gillespie, Sophie Fielding, Guillermo Jiménez, Phil Johnston, Rachael R Sinclair, Agnar Sivertsen, Stian A Solbø, et al., "A review of unmanned vehicles for the detection and monitoring of marine fauna," *Marine Pollution Bulletin*, vol. 140, pp. 17–29, 2019.

[3] Xavier Mouy, Mohammed Bahoura, and Yvan Simard, "Automatic recognition of fin and blue whale calls for real-time monitoring in the st. lawrence," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2918–28, 12 2009.

[4] David K. Mellinger and Christopher W. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3518–3529, May 2000.

[5] David K. Mellinger, "A comparison of methods for detecting right whale calls," *Canadian Acoustics*, vol. 32, no. 2, pp. 55–65, June 2004.

[6] W. Vickers, B. Milner, J. Lines, and R. Lee, "A comparison of machine learning methods for detecting right whales from autonomous surface vehicles," in *EUSIPCO*, 2019.

[7] Evgeny Smirnov, "North atlantic right whale call detection with convolutional neural networks," in *Proc. Int. Conf. on Machine Learning, Atlanta, USA*. Citeseer, 2013, pp. 78–79.

[8] W. Vickers, B. Milner, J. Lines, and R. Lee, "Detecting right whales from autonomous surface vehicles using RNNs and CNNs," in *EUSIPCO - Satellite Workshop: Signal Processing, Computer Vision and Deep Learning for Autonomous Systems*, 2019.

[9] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7398–7402.

[10] Fan Liu, Qingzeng Song, and Guanghao Jin, "The classification and denoising of image noise based on deep neural networks," *Applied Intelligence*, pp. 1–14, 2020.

[11] Tiago Nazaré, Gabriel De Barros Paranhos da Costa, Welinton Contato, and Moacir Ponti, *Deep Convolutional Neural Networks and Noisy Images*, pp. 416–424, 01 2018.

[12] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, vol. 2013, pp. 436–440.

[13] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, 2016, pp. 241–246.

[14] Emad M Grais and Mark D Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *2017 IEEE global conference on signal and information processing (GlobalSIP)*. IEEE, 2017, pp. 1265–1269.

[15] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.," *Journal of machine learning research*, vol. 11, no. 12, 2010.

[16] Colin W Clark, "Acoustic communication and behavior of the southern right whale (eubalaena australis)," *Communication and behavior of whales*, pp. 163–198, 1983.

[17] K. Pylypenko, "Right whale detection using artificial neural network and principal component analysis," Apr. 2015, pp. 370–373.

[18] Douglas Gillespie, "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," *Canadian Acoustics*, vol. 32, no. 2, June 2004.

[19] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[20] Eric Spaulding, Matt Robbins, Thomas Calupca, Christopher W Clark, Christopher Tremblay, Amanda Waack, Ann Warde, John Kemp, and Kris Newhall, "An autonomous, near-real-time buoy system for automatic detection of north atlantic right whale calls," in *Proceedings of Meetings on Acoustics 157ASA*. Acoustical Society of America, 2009, vol. 6, p. 010001.