Robust North Atlantic Right Whale Detection using Deep Learning Models for Denoising

William Vickers,[1] Ben Milner,[1, a] Denise Risch,[2] and Robert Lee[3]
[1] *School of Computing Sciences, University of East Anglia, Norwich, Norfolk, UK*
[2] *Scottish Association for Marine Science, Oban, UK*
[3] *Gardline Geosurvey Limited, Great Yarmouth, UK*

(Dated: 18 May 2021)

This paper is part of a special issue on Machine Learning in Acoustics

This paper proposes a robust system for detecting North Atlantic right whales by using deep learning methods to denoise noisy recordings. Passive acoustic recordings of right whale vocalisations are subject to noise contamination from many sources such as shipping and offshore activities. When such data is applied to uncompensated classifiers, their accuracy falls substantially. To build robustness into the detection process, two separate approaches that have proved successful for image denoising are considered. Specifically a denoising convolutional neural network (DNCNN) and a denoising autoencoder (DAE), each of which is applied to spectrogram representations of the noisy audio signal, are developed. Performance is improved further by matching the classifier training to include the vestigial signal that remains in clean estimates after the denoising process. Evaluations are performed first by adding white, tanker, trawler and shot noises at SNRs from -10dB to +5dB to clean recordings to simulate noisy conditions. Experiments show that denoising gives substantial improvements to accuracy and particularly when using the vestigial-trained classifier. A final test applies the proposed methods to previously unseen noisy right whale recordings and finds that denoising is able to improve performance over the baseline clean trained model in this new noise environment.
 [https://doi.org(DOI number)]

[XYZ] Pages: 1–14

I. INTRODUCTION

The aim of this work is to develop robust methods of detecting marine mammals from passive acoustic monitoring (PAM) devices in challenging environments. Being able to reliably detect marine mammals 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. This is of particular interest, as they are one of the world's most endangered marine mammal species and at risk of extinction with as few as 350 individuals remaining (Pace III *et al.*, 2017). Entanglement in fishing gear and ship strike are the most common lethal causes in North Atlantic right whales (Corkeron *et al.*, 2018; Davies and Brillant, 2019) and offshore industries, such as oil and gas exploration and offshore construction, pose additional risks (Leiter *et al.*, 2019). Detecting the animals' presence before they are in close proximity to large vessels or enter a mitigation zone can both protect animals and avoid costly shutdowns of offshore operations. Traditional methods for detecting marine mammals at sea use human observers on-board ships, but more recently long-term archival recorders (Davis *et al.*, 2014), gliders (Baumgartner *et al.*, 2013) and autonomous surface vehicles (ASVs) have been used as they offer a cheaper solution that can operate in zero visibility conditions, and in the case of autonomous vehicles, can provide real- or near real-time data (Verfuss *et al.*, 2019). Human experts may listen to these audio recordings and use spectrogram analysis to identify occurrences but this is time consuming and expensive. Automating the detection process and providing a robust solution to detecting North Atlantic right whales is the aim of this work.

A number of machine learning techniques have been applied to cetacean detection from audio data. Vector quantisation and dynamic time warping have been effective in detecting blue and fin whales from frequency contours extracted from spectrograms (Mouy *et al.*, 2009). Hidden Markov models (HMMs) have also been used to recognise low frequency whale sounds using spectrogram features (Mellinger and Clark, 2000). Comparisons have also been made between between artificial neural networks (ANNs) and spectrogram correlation for right whale detection (Mellinger, 2004). With the advent of deep learning, a number of approaches have been ap-

[a] b.milner@uea.ac.uk

plied to cetacean detection. Representing audio recordings as a two-dimensional spectrogram has led to convolution neural network (CNN) approaches for detection (Ibrahim *et al.*, 2018). A recent study comparing various time-series classification and deep learning methods to right whale detection found that those using CNNs had highest accuracy (Vickers *et al.*, 2019a). Further studies have also reported on the success of using CNNs for right whale detection when compared to other classification models such as recurrent neural networks (Shiu *et al.*, 2020; Smirnov, 2013; Vickers *et al.*, 2019b).

An important consideration when developing automatic detectors is the likelihood that right whale recordings will be corrupted by noise from various sources at differing signal-to-noise ratios (SNRs), depending on the distance of the right whale and noise source from the receiving hydrophone. Noise presents a challenge to most classification problems, from speech recognition to image identification (Liu *et al.*, 2020; Seltzer *et al.*, 2013), and consequently many different compensation techniques have been proposed. These can broadly be categorised into those that attempt to match the underlying model to the characteristics of the noisy input data and those that remove noise before classification (Lu *et al.*, 2013; Nazaré *et al.*, 2018).

In this work we consider both of these strategies for improving robustness within the framework of right whale detection and we also show that combining them gives further improvement. Our classifier is based on previous work which established that transforming the audio into a spectrogram representation, to consider it as an image, and inputting this into a CNN outperformed a range of other machine learning approaches (Vickers *et al.*, 2019b). To improve the robustness of this model, we consider both augmented training methods and denoising of the spectrograms prior to classification. Two different approaches for denoising spectrogram representations of the audio signal are considered and compared. These are the denoising autoencoder (DAE) and the denoising CNN (DNCNN) (Gondara, 2016; Grais and Plumbley, 2017; Zhang *et al.*, 2017). These are chosen as they have both been shown to be highly successful when applied to image denoising and yet represent very different architectures. The DAE builds on the structure of autoencoders and comprises a series of encoding layers leading to a compressed bottleneck representation of the input data followed by a series of decoding layers to return the input to its original size. Within a DAE, the learning aims to map noisy input features into a clean output representation (Gondara, 2016; Grais and Plumbley, 2017). DAEs have been highly successful in tasks such as image denoising and audio separation (Gondara, 2016; Grais and Plumbley, 2017) and consequently are a good candidate to apply to spectrogram denoising of right whale recordings. The DNCNN exploits and combines some of the most effective architectures that have been proposed for image recognition and denoising. This includes using deep architectures that are effective at increasing the learning capacity and flexibility of the model (He *et al.*, 2016; Krizhevsky *et al.*, 2012; Simonyan and Zisserman, 2015). To improve the learning of such deep models, residual learning frameworks have been shown to be more effective than attempting to learn a direct mapping (He *et al.*, 2016; Zhang *et al.*, 2017). Batch normalisation is also commonly applied and through the scaling and shifting applied at each layer, any internal covariate shift can be mitigated (Glorot and Bengio, 2010; Ioffe and Szegedy, 2015). Based on these factors, the approach taken for spectrogram denoising is based on a DNCNN framework that employs residual learning (Zhang *et al.*, 2017). Our evaluation of the two denoising methods found them to leave remains of the contaminating noise and to introduce artefacts into the denoised spectrograms. To better match these denoised signals to the classifier, we create training augmented data that contains this signal, which we term vestigial noise, and retrain the classifier.

The remainder of the paper is organised as follows. Section II gives a brief introduction to right whale calls and typical sources of marine noise that contaminate recordings. Section III introduces the CNN-based right whale detector and explains how data augmentation can be applied to make the model more generic. The denoising CNN (DNCNN) is explained in Section IV and how it is applied to spectrogram enhancement. Section V discusses the framework for using denoising autoencoders (DAEs) to denoise spectrograms. The experimental set-up is explained in Section VI and results and analysis from a set of evaluations presented in Section VII.

II. ACOUSTIC CHARACTERISTICS OF RIGHT WHALES IN MARINE ENVIRONMENTS

Right whales are one of the world's most endangered marine mammals and are at risk of extinction with as few as 350 individuals remaining (Pace III *et al.*, 2017). Right whales emit a range of vocalisations with common sounds being upcall tones and gunshot sounds and it is these two call types that we focus on in this work (Clark, 1983). Upcalls begin with a frequency of approximately 60Hz that rises to around 250Hz and typically last for about one second, although these calls are not always consistent and vary in duration and frequency (Pylypenko, 2015). Upcalls most likely play a role as a social contact call between individuals and are produced by both sexes and different age classes and are therefore most commonly used for passive acoustic detection of the species (Clark *et al.*, 2007; Parks *et al.*, 2011). The gunshot sounds are very different to upcalls and are characterised as an impulsive broadband-like signal, primarily produced during mating (Parks *et al.*, 2012). Due to their different behavioural function their relative frequency in different right whale habitats can vary (Van Parijs *et al.*, 2009). Figure 1 shows example spectrograms of an upcall and gunshot vocalisation. Both of these vocalisation types can be difficult to hear in noisy conditions, and to visualise in spectrograms, as low frequency regions are often masked by marine noise such as from passing ships, drilling and piling activities, seismic exploration or inter-

ference from other marine mammals, such as humpback whale song (Gillespie, 2004). In many cases, anthropogenic and environmental noises overlap in frequency with right whale calls, which makes detection difficult.

To represent different anthropogenic conditions in this work, we consider four types of noise as typical contaminants of right whale recordings, namely tanker noise, trawler noise, shot noise (representative of sounds produced by activities such as piling and seismic exploration) and white noise. These noise types are described in Section VI A and example spectrograms shown in Figure 4. For evaluation, noises are added artificially to the whale recordings at signal-to-noise ratios (SNRs) from -10dB to +5dB to simulate recordings at different ranges from the receiving hydrophone.

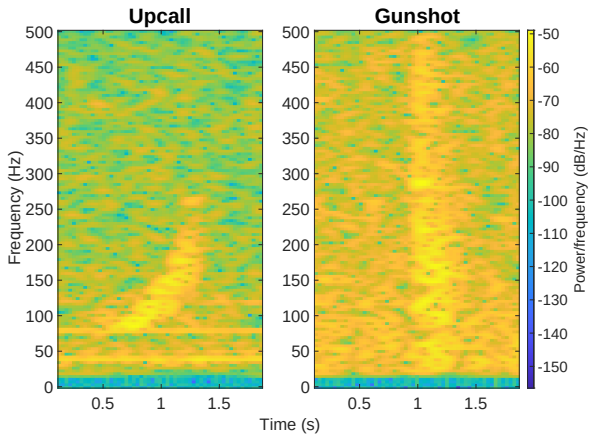

FIG. 1. *Two example spectrograms showing a right whale upcall (left) and a gunshot (right). Upcalls are characterised as a tone starting at around 60Hz and ending around 250Hz, with a duration of one second. Gunshots have less structure and are characterised as bursts of broadband noise.*

III. CNN RIGHT WHALE CLASSIFIER

The CNN right whale classifier is based on our earlier work that investigated a range of deep learning techniques (Vickers *et al.*, 2019a). Specifically, we compared time-series classification, RNNs and CNNs and found highest accuracy was achieved with an architecture that first extracted spectrogram features from the audio and input those into a CNN-based classifier.

A. Feature extraction

The requirement of feature extraction is to transform an input audio signal into a representation that is more effective for identifying whale sounds. Although many different approaches to audio feature extraction have been developed (for example MFCCs, PLP, filterbank (Milner, 2002)) we chose a straightforward power spectral-based representation. Our reasoning is that we wish to allow the subsequent CNN to learn the most appropriate feature representation and not to make any assumptions beforehand such as introducing a non-linear frequency scaling.

Feature extraction uses a sliding window to convert short-duration frames of the input audio signal into a sequence of log power spectral vectors, $\mathbf{x}_t$. Specifically, an N-point frame of time-domain samples is extracted from the audio, a Hamming window applied and a Fourier transform computed. The upper $N/2$ frequency points are discarded and the remaining complex points transformed to power and then logged. Analysis windows are advanced by S samples to compute each new spectral vector. Normalisation is applied to the elements of the log spectral vectors such that they are in the range 0 to 1. Spectral vectors are grouped in two-second blocks and used to create the spectrogram feature that is input into the CNN. Preliminary testing established that best performance is achieved with N=256 and S=32 which at a sampling frequency of f_s=1kHz, corresponds to a frame width of 256ms and a frame advance of 32ms, resulting in spectrogram features, $\mathbf{X}$, of size 55×128.

B. CNN classifier

The classifier, $\mathcal{C}(\mathbf{X})$, developed for this work comprises first a CNN encoder that maps input spectrogram features, $\mathbf{X}$, into a new space and contains convolutional layers that are each followed by max pooling layers. This outputs into a series of dense layers that perform classification. Preliminary testing established that best performance is attained using three convolutional layers and two dense layers (Vickers *et al.*, 2019a). Each convolutional layer uses 3×3 filter kernels with 32, 64, 128 filters in each subsequent layer. The max pooling layers use a pool size of 2×2 and have rectified linear unit (ReLU) non-linear activation functions applied to their outputs (Nair and Hinton, 2010). 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 respectively, with a ReLU activation function. The final dense layer uses a softmax activation function to output the probability of each class. Training used an Adam optimiser with a learning rate of 0.001 and categorical cross-entropy as the loss function (Kingma and Ba, 2014). Training took place over 200 epochs and was repeated 10 times for each test. The model that achieved highest validation accuracy was used for testing and reported accuracies were calculated as an average over all 10 tests. In the subsequent tests that are reported in Section VII, the standard deviation across 10 repetitions was very small (less than 0.5%) and no outliers were observed within each test.

For the whale data used in this work, three classes are defined - {*upcall, gunshot, no whale*} - and the classifier is trained using target labels for each training data sample. A baseline classifier is trained using noise-free

training data, see Section VI A for details on the data. The evaluations also consider variants of the classifier that have been trained using data augmentation methods. Specifically, the noise-free training data is contaminated with noise and the classifier re-trained. This is used to create classifiers that are trained on the specific noise condition under test or more generic classifiers that are trained across a range of noise conditions. These are defined in Section VI B.

IV. DENOISING CONVOLUTIONAL NEURAL NETWORK

This section explains how denoising convolutional neural networks (DNCNNs) are applied to enhance spectrogram representations of noisy audio. The DNCNN's architecture is explained first and then adjustments to the classifier to maximise accuracy.

A. DNCNN architecture

Residual learning has been shown highly effective for image denoising in scenarios such as additive white noise to deblocking the distortion introduced by JPEG compression (Zhang *et al.*, 2017). This success is based on the assumption that for an image contaminated by noise, it is easier to learn a mapping to the noise (i.e. residual) than it is to learn a mapping to the clean image. We make the same assumption for the problem of denoising spectrograms that have been extracted from audio contaminated by noise. Specifically, learning the mapping to the noise spectrogram (or residual) is easier than learning the mapping to the clean spectrogram. For spectrogram features extracted from noisy audio, the additivity of clean audio and noise is not necessarily linear and depends on whether spectral amplitudes are linear or have been log transformed. We consider both scenarios.

Considering first spectrogram features that are extracted from noisy audio as described in Section III A without the log operation being applied to their amplitudes. The noisy spectrogram, $\mathbf{Y}$, can be assumed equal the sum of the clean and noise spectrograms (ignoring cross-spectral terms), $\mathbf{X}$ and $\mathbf{D}$, as

$$\mathbf{Y} = \mathbf{X} + \mathbf{D} \tag{1}$$

When this is reformulated into a residual learning framework, rather finding a direct mapping from the noisy spectrogram to the clean spectrogram, i.e. $\mathcal{F}(\mathbf{Y}) = \mathbf{X}$, a residual mapping, $\mathcal{R}_{LIN}(\mathbf{Y}) = \mathbf{D}$, is instead learnt. This makes a prediction of the noise spectrogram (i.e. residual) and when subtracted from the noisy spectrogram gives an estimate of the clean spectrogram, $\widehat{\mathbf{X}}$, as

$$\widehat{\mathbf{X}} = \mathbf{Y} - \mathcal{R}_{LIN}(\mathbf{Y}) \tag{2}$$

The alternative spectrogram feature is represented by log spectral amplitudes, which is common practice for audio processing applications. In this case the noisy log spectrogram, $\log(\mathbf{Y})$, is expressed as,

$$\log(\mathbf{Y}) = \log(\mathbf{X} + \mathbf{D}) \tag{3}$$

The residual $(\log(\mathbf{Y}) - \log(\mathbf{X}))$ in this representation is obtained by expanding the log operation in Eqn. (3) to

$$\begin{aligned} \log(\mathbf{Y}) &= \log\left(\mathbf{X}\left(1 + \frac{\mathbf{D}}{\mathbf{X}}\right)\right) \\ &= \log(\mathbf{X}) + \log\left(1 + \frac{\mathbf{D}}{\mathbf{X}}\right) \end{aligned} \tag{4}$$

and so the residual mapping, $\mathcal{R}_{LOG}(\mathbf{Y})$, is

$$\mathcal{R}_{LOG}(\mathbf{Y}) = \log(\mathbf{Y}) - \log(\mathbf{X}) = \log\left(1 + \frac{\mathbf{D}}{\mathbf{X}}\right) \tag{5}$$

This residual is significantly different to that using linear spectral amplitudes and no longer comprises just a noise component. Instead, it is a combination of the noise and clean spectrogram components.

With these two formulations for the residual, two slightly different architectures for denoising the spectrogram features are required and shown in Figure 2. Both ultimately provide estimates of the clean log spectrum for the CNN classifier described in Section III B. To perform the residual mapping we base our architecture on the approach developed for image denoising and use a model with 17 convolutional layers (Zhang *et al.*, 2017). The first layer has 64 filters and outputs these into a ReLU activation function (Nair and Hinton, 2010). The next 15 convolutional layers also use 64 filters but now incorporate batch normalisation before outputting into a ReLU activation function (Glorot and Bengio, 2010; Ioffe and Szegedy, 2015). The final layer excludes the batch normalisation and ReLU operations and outputs a prediction of the residual, or noise, spectrogram. No pooling layers are used, so deeper models have a wider receptive field. With 17 layers, this corresponds to a receptive field of 35×35. For the spectrogram features this equates to a receptive field of 1.27 seconds and bandwidth of 137Hz which is broadly the duration of a whale vocalisation and the frequency range of an upcall.

The DNCNN is trained using pairs of spectrogram features with a noise-free version forming the training target and a noisy version as the input. Noisy spectrograms for training are produced by adding the desired noise type at the required SNR to the noise-free time-domain signals and extracting spectrogram features. Mean squared error is used as the loss function between the clean and predicted spectrogram features, along with the Adam optimiser. Training was performed over 50 epochs.

The classifier, $\mathcal{C}()$, in Section III B requires as its input an estimate of the clean log spectrum. For the DNCNN formulation that takes as input log spectrogram features, $\mathcal{R}_{LOG}()$, the residual output is subtracted from the log noisy spectrogram to give the clean log spectrogram estimate, $\widehat{\log(\mathbf{X})}$, that is input into the classifier,

$$\widehat{\log(\mathbf{X})} = \log(\mathbf{Y}) - \mathcal{R}_{LOG}(\log(\mathbf{Y})) \tag{6}$$

For the DNCNN using linear spectrogram features, $\mathcal{R}_{LIN}()$, the residual output is subtracted from the linear

noisy spectrogram to give the estimate of the clean linear spectrogram and this is then logged before being input into the classifier,

$$\log(\widehat{\mathbf{X}}) = \log\left(\mathbf{Y} - \mathcal{R}_{LIN}(\mathbf{Y})\right) \tag{7}$$

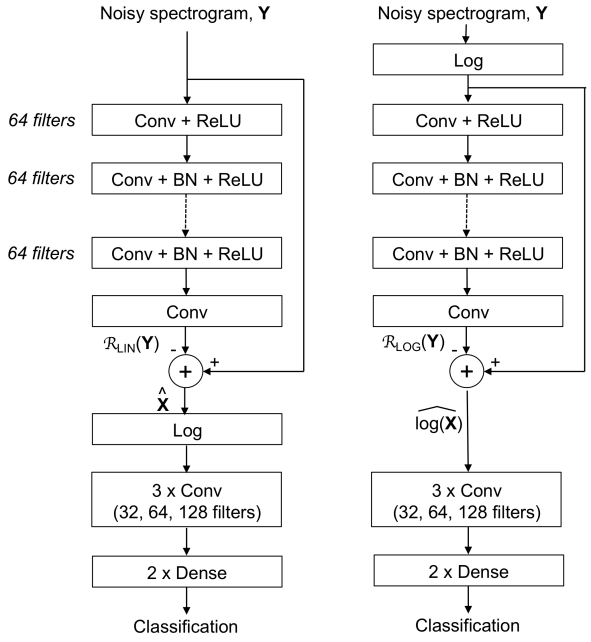

FIG. 2. *The left figure shows the denoising CNN (DNCNN) architecture applied to linear spectrogram amplitudes, and the right figure the DNCNN architecture applied to log spectrogram amplitudes. All CNN layers use filter sizes of* 3×3.

B. Classifier training

The aim of denoising is to estimate a noise-free spectrogram for input into the CNN classifier, $\mathcal{C}()$. The experimental evaluation in Section VII A shows that when this classifier is trained on noise-free data and then tested in noise-free conditions it achieves high accuracy. Expecting this clean-trained classifier to attain high accuracy when the input data has been denoised assumes that the denoising process is able to remove all of the noise and to introduce no artefacts of its own. However, it is likely that the denoising process will leave some noise and also introduce artefacts to the denoised spectrograms. We term these unwanted components as a vestigial spectrogram, $\mathbf{V}$ (we use the term 'vestigial' to avoid confusion with 'residual', which is used within the DNCNN). To address this mismatch a new classifier, $\mathcal{C}_{DNCNN}()$, is trained on data containing this vestigial error component resulting from the DNCNN denoising process. Training the CNN classifier, $\mathcal{C}()$, was explained in Section III B and used noise-free training data. To create the vestigial trained classifier, the set of noise-free training data is contaminated by noise and passed through the DNCNN to create a training set of denoised spectrogram features that will contain this vestigial signal. The new classifier, $\mathcal{C}_{DNCNN}()$, is then trained from this data. In this framework the training data can be contaminated by a single noise type and SNR, or from more generic data, depending on the test conditions. This is investigated in Section VII B.

V. DENOISING AUTOENCODER

The denoising autoencoder (DAE) approach to enhancing spectrogram representations of noisy audio is presented in this section. The DAE architecture is explained first before describing how the classifier is retrained to optimise performance.

A. DAE architecture

The denoising autoencoder, $\mathcal{A}(\mathbf{Y})$, takes noise-contaminated spectrogram features and predicts an estimate of the noise-free spectrogram that is input into the classifier, $\mathcal{C}()$. The DAE is illustrated in Figure 3 and first encodes the input spectrogram features using three convolutional layers each with 32 filters. Each layer is followed by a ReLU activation function and a max pooling layer to compress the output (Nair and Hinton, 2010). The max pooling layers use a pool size of 2×2. The compressed representation of the spectrogram is then decoded back to its original size using three more convolutional layers which also use 32 filters. Each convolutional layer is followed by a ReLU activation function and an upsampling operation that expands the size of the image by a factor of two in each dimension.

Similar to the DNCNN, the DAE is trained with pairs of spectrogram features with a noise-free version as the training target and a noisy version as the input. Binary cross-entropy is used as the loss function between the clean and predicted spectrogram features, along with the Adam optimiser. Training was performed over 100 epochs. A series of preliminary tests established that using three convolutional layers each for the encoder and decoder, with 32 filters, gave best performance.

B. Classifier training

Two methods for training the classifier used with the DAE are considered, as was done with the DNCNN in Section IV B. One option is to apply the DAE denoised spectrogram features directly into the CNN classifier trained using clean audio, $\mathcal{C}()$. However, as was considered for the DNCNN, it is likely that the predicted spectrogram features will contain a vestigial signal that retains some of the original noise and contains artefacts from the DAE denoising process. Consequently, a second classifier is trained, $\mathcal{C}_{DAE}()$, using data created by passing noisy training data through the DAE to create a new set of spectrogram features that comprise the clean signal and a vestigial component from the DAE denoising.

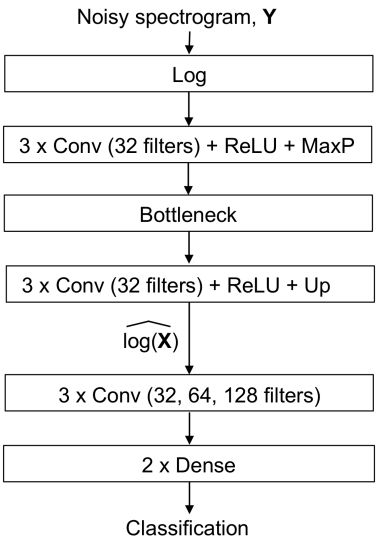

FIG. 3. *Denoising autoencoder (DAE) architecture for denoising log spectrogram features and inputting these into the CNN classifier. All CNN layers use filter sizes of* 3×3.

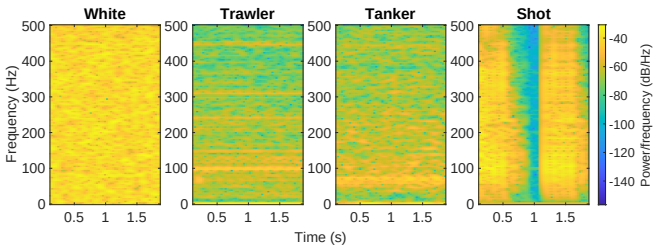

FIG. 4. *Spectrograms showing two-second examples of white noise, trawler noise, tanker noise and shot noise that are used in the evaluations in Section VII.*

VI. EXPERIMENTAL SETUP

This section introduces the whale data and noise conditions that the evaluation is based upon. The different system configurations under evaluation are then defined in terms of their method of denoising and training.

A. Datasets

The right whale recordings used for evaluating the classifiers and denoising methods were taken from the DCLDE 2013 workshop and were collected in the Gerry E. Studds Stellwagen Bank National Marine Sanctuary from the Massachusetts Bay area of the north-eastern coast of the US[1]. These were collected using marine autonomous recording units (MARUs) deployed in arrays of between 6 and 10 devices. For this dataset, the output of just one channel is taken, converted to 16 bits per sample and sampled at 2kHz. The audio recordings are subsequently arranged as two-second segments that either contain a right whale sound or do not, and have been annotated by human experts using data from all channels to maximise accuracy. Two different whale vocalisations are heard, upcalls and gunshots, which gives a three-class classification problem - {*upcall, gunshot, no whale*}. The recordings are relatively noise-free, as spectrograms in Figure 1 show, but they do contain some low amplitude noise. For the purposes of the evaluation in this work, we consider them as 'clean' and subsequently add noise to simulate noisy audio. The DCLDE 2013 workshop protocol supplied four days of recordings for use in training and a further three days for testing. However, labels for the test data are not available so we divided the four days supplied for training into non-overlapping training, validation and test sets, using a split of 70:15:15, which gives sizes of 2,784, 600 and 600, respectively. Each set contains an equal proportion of segments from each of the three classes and samples are taken randomly from the original corpus. Given the low frequency of right whale calls the audio was downsampled to 1 kHz, as previous work showed this introduces no loss in accuracy (Vickers *et al.*, 2019a). Our code for extracting the training, validation and test sets from the data available in the DCLDE 2013 workshop is available from GitHub[2].

Four noise types are considered for the evaluation - tanker noise, trawler noise, shot noise and white noise. Spectrogram examples of each of these noise types are shown in Figure 4. Tanker and trawler noises are chosen as shipping is a common source of marine noise that introduces horizontal bands in the spectrograms arising from harmonics of rotating machinery within the ship as well as low frequency noise. These noises were obtained from data that had been collected by the NOAA Northeast Fisheries Science Center from a passive acoustic monitoring project in the Stellwagen Bank National Marine Sanctuary. Shot noise is representative of sounds produced by activities such as piling and seismic exploration and is characterised by vertical structure in the spectrogram. The shot noise examples were taken from the 'gun' samples in the NOISEX-92 database (Varga and Steeneken, 1993). This noise is impulsive but was arranged so that each two-second recording contained at least one example of the shot noise. Finally, white noise is included as a more general noise type that affects all time and frequency regions within the spectrogram, and this was generated artificially. To create the noisy audio segments, noise is added to every two-second recording (upcall, gunshot and no whale) in the time-domain (waveform-domain) at SNRs of +5dB, 0dB, -5dB and -10dB. This set of SNRs is chosen to cover a range of reception conditions that represent signals received from right whales at both close and long range distances. For recordings that contain a whale vocalisation, the noise samples are scaled such that when added to the whale recording, their subsequent power achieves the target SNR. To create the noisy 'no whale' recordings, two-second segments with no whale vocalisation present are extracted from the original recordings at a time 5 seconds after an upcall or gunshot has occurred. To these

'no whale' segments, noise samples are added and scaled so that they have the same noise power as that in the preceding segment which contained a whale sound. This ensures that the actual power of the noise remains consistent across each pair of whale and 'no whale' examples. The procedure is illustrated in Figure 5. It should be noted that within a specific SNR and noise type, the noise examples that are added to the whale/no-whale examples are not duplicated, so each two-second segment is contaminated with unique noise examples. Further, there is no sharing of noise examples used across the training, validation and test sets.

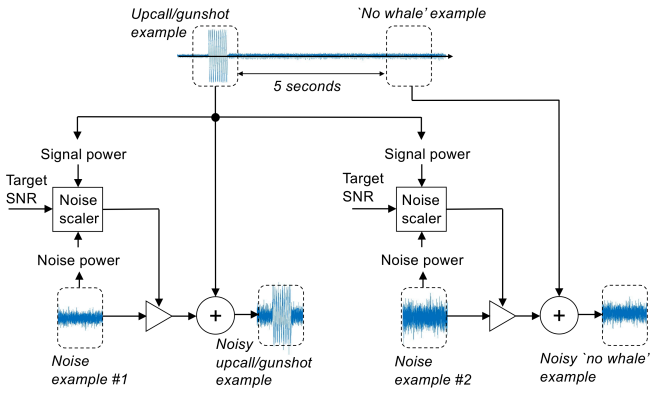

FIG. 5. *Method of adding noise to the two-second whale and 'no whale' segments to create noisy examples at the target SNR. 'No whale' examples are extracted 5 seconds after a whale vocalisation to give consistency in terms of the power of the noise examples that are added.*

A final, unseen, test condition is introduced for the final stage of evaluation and uses whale recordings collected from a different marine environment. These recordings are naturally more noisy and were taken from the Cornell NRW Buoys data, recorded in the area of Cape Cod (Spaulding *et al.*, 2009). For testing, a total of 2,142 two-second recordings are used, with 1,071 containing an upcall and a further 1,071 having no whale sound. This dataset has no gunshot examples.

B. System configurations

The aim of the experiments presented in Section VII is to investigate how the proposed denoising methods perform in different noise conditions and to explore the effectiveness of the various configurations. Specifically, the experiments aim to:

- Examine the effectiveness of augmenting clean training data with noisy examples when testing in noisy conditions
- Compare classification accuracy when using training data augmentation against the explicit denoising methods of the DAE and DNCNN
- Determine, when using denoising, whether the classifier is best trained on clean or vestigial data
- Consider how classification accuracy is affected when the noise condition in testing is unseen in training

To address these aims, a number of different classifier and denoising configurations are investigated. These are shown in Table I and divided into three sets:

1. Those using data augmentation for classifier training with no explicit denoising
2. Those applying DAE denoising
3. Those applying DNCNN denoising

For each method in Table I, the columns show the denoising method (i.e. none, DAE or DNCNN), the training data used for denoising (if applied) and the training data used to create the classifier. The final column shows the mean classification accuracy, measured across all noise types and SNRs, and summarises the results in Section VII.

From Table I, the first four configurations use no explicit denoising and instead differ in how the classifier is trained with regard to the test condition. Method CLEAN is the baseline classifier and trained on only clean (noise-free) training data. The classifiers used in configuration MATCH are trained on data that matches the specific noise type and SNR that is subsequently used in testing. This requires a set of 16 matched models that are used individually in each specific noise condition. The GENERIC classifier is trained on data contaminated with all four noises types at all four SNRs. This gives the most generic model for classification. The UNSEEN classifier is similar, however the specific noise type under test is excluded from the training data so that the test noise condition is unseen during classifier training.

The next four methods in Table I all use the DAE for denoising prior to classification. The naming convention for these methods follows the structure DAE-<*denoising training data*>-<*classifier training data*>. Method DAE-MATCH-CLEAN uses the DAE autoencoder that is trained on data matched to the specific noise test condition and the classifier is trained on clean data (i.e. method CLEAN). The denoising in method DAE-MATCH-VES is identical but the classifier is now trained on the vestigial data (Section IV B). Method DAE-GENERIC-VES uses a DAE trained across all four noise types and four SNRs and uses a vestigial-trained classifier. Finally, method DAE-UNSEEN-VES is similar except the DAE is trained on all noise types with the exception of the specific noise under test, i.e. on three noise types at the four SNRs.

The four final denoising methods in Table I use the DNCNN and have naming conventions as DNCNN-<*denoising training data*>-<*classifier training data*>.

These four methods follow the same structure as those shown for the DAE.

VII. EXPERIMENTAL RESULTS

The aim of these experiments is to explore the effectiveness of the various denoising methods under different noise type and SNR conditions. The first experiment uses no explicit denoising and examines the effect that augmenting clean training data with varying quantities of noise data, both seen and unseen in relation to the test conditions, has on accuracy. The second set of tests compares the effectiveness of the DAE and DNCNN denoising methods and examines whether higher accuracy is attained when using clean-trained or vestigial-trained classifiers. The third set of tests again compares the DAE and DNCNN denoising methods but now examines accuracy when the noise in testing is unseen during model training. A fourth experiment moves testing to the unseen Cape Cod dataset and compares classification accuracy using augmentation and denoising. A final section compares processing times for the denoising methods.

A. Augmented model training

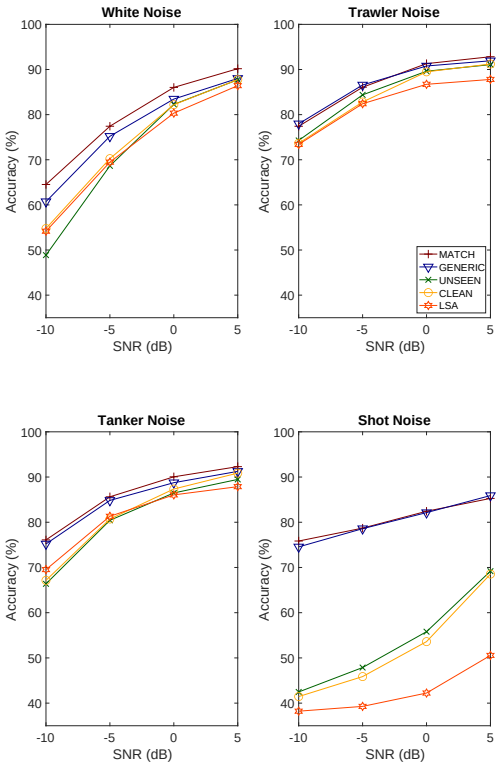

FIG. 6. *Right whale classification accuracies in the four different noise types at SNRs from -10dB to +5dB. The models are trained using different augmentation strategies with no explicit denoising, with the exception of the LSA method.*

This first set of tests does not use any denoising and instead examines the accuracy of the CNN classifier introduced in Section III using different training data augmentation scenarios. The evaluation is performed across all four noise types and SNRs with classification accuracies shown in Figure 6. Each noise condition is evaluated using four different classification models - trained on clean data (CLEAN), trained on data matched to the specific test condition (MATCH), trained on all four noise types and SNRs (GENERIC) and trained on three noise conditions excluding the noise type under test (UNSEEN). To benchmark the effectiveness of these methods against an existing method of noise reduction, a log spectral amplitude (LSA) estimator was also evaluated, given its success in denoising audio signals (Cohen, 2002). Using the implementation in (Loizou, 2013), the noisy examples were denoised and the resulting time-domain samples then input into the CNN of Section III for spectrogram extraction and classification. Classification accuracies are shown as method LSA in Figure 6.

In noise-free conditions the CLEAN system attains an accuracy of 94.1% but falls as SNRs reduce and in general has lowest performance. Testing using the matched model (MATCH) removes the mismatch between training and test conditions and improves accuracy substantially. However, this does require the model to be trained under the same noise conditions as seen in testing. Augmenting the training data to contain all noise types and SNRs (GENERIC) gives accuracy close to MATCH and occasionally attains higher performance which we attribute to the broad coverage of the training data. However, removing from the training data the noisy examples corresponding to the noise type under test, to give method UNSEEN, reduces accuracy considerably to be comparable with the CLEAN model. For the LSA method of denoising, performance is similar to that obtained using the CLEAN model, although in shot noise the performance is substantially worse. Examining spectrograms of the LSA denoised signals shows the noise to have been suppressed to a certain extent, but to now also contain short duration artefacts. We believe these lead to confusion with whale vocalisations in the classifier, particularly with upcalls, hence the inability of LSA to improve accuracy beyond the CLEAN model.

B. Denoising autoencoder and denoising CNN performance

The second set of experiments compares the performance of the denoising autoencoder (DAE) and denoising CNN (DNCNN), described in Sections IV and V. These tests also examine how best to train the classifier, on either clean data or vestigial data. Specifically, classification accuracy is measured across all four noise types and SNRs using both the DAE and the DNCNN denoising methods trained on data matched to the specific noise type and SNR under test. Methods DAE-MATCH-CLEAN and DNCNN-MATCH-CLEAN output their denoised spectrogram features into a CNN classifier trained on clean data, while methods DAE-MATCH-VES

Name	Denoising method	Denoising training data	Classifier training data	Mean accuracy
CLEAN	None	NA	Clean data	72.98%
MATCH	None	NA	Specific noise type and SNR under test	83.26%
GENERIC	None	NA	All noise types at all SNRs	82.24%
UNSEEN	None	NA	All noise types at all SNRs except the noise under test	72.81%
DAE-MATCH-CLEAN	DAE	Noise type and SNR under test	Clean data	82.80%
DAE-MATCH-VES	DAE	Noise type and SNR under test	Vestigial noisy data	85.18%
DAE-GENERIC-VES	DAE	All noise types at all SNRs	Vestigial noisy data	83.52%
DAE-UNSEEN-VES	DAE	All noise types at all SNRs except the noise type under test	Vestigial noisy data	73.45%
DNCNN-MATCH-CLEAN	DNCNN	Noise type and SNR under test	Clean data	79.57%
DNCNN-MATCH-VES	DNCNN	Noise type and SNR under test	Vestigial noisy data	84.71%
DNCNN-GENERIC-VES	DNCNN	All noise types at all SNRs	Vestigial noisy data	81.45%
DNCNN-UNSEEN-VES	DNCNN	All noise types at all SNRs except the noise type under test	Vestigial noisy data	72.85%

TABLE I. Definitions of the various system configurations under evaluation in terms of the denoising method and its training data, and the classifier training data. The first four methods use no explicit denoising, while the remaining methods use various configurations of either the denoising autoencoder (DAE) or the denoising CNN (DNCNN). The final column shows the mean classification accuracy of each method, taken across all noise types and SNRs from Section VII.

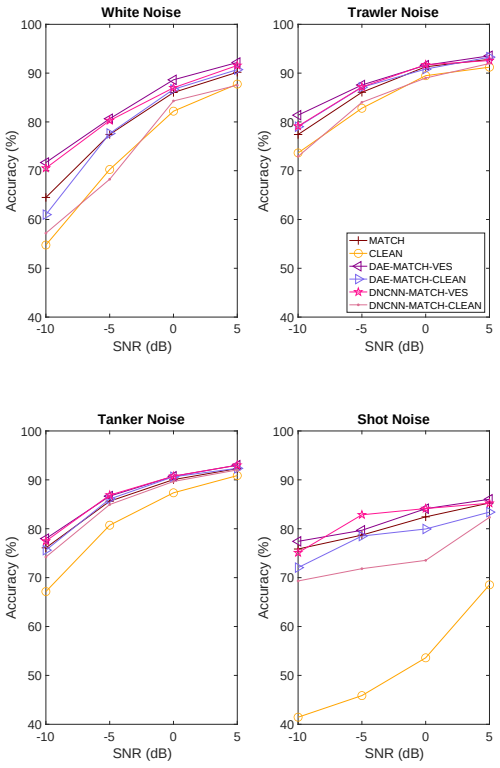

FIG. 7. *Right whale detection accuracies when applying the denoising autoencoder (DAE) and denoising CNN (DNCNN) to the four noise types at SNRs from -10dB to +5dB.*

and DNCNN-MATCH-VES output into a CNN classifier trained on vestigial data after noise removal. Table I shows specific configuration details on these systems. For comparison, the performance of the clean trained CNN model (CLEAN) and matched CNN models (MATCH) are included with classification accuracies shown in Figure 7. As a preliminary test for the DNCNN, we compared the accuracy of the system using log spectrogram denoising, $\mathcal{R}_{LOG}()$, with that using linear spectrogram denoising, $\mathcal{R}_{LIN}()$, introduced in Section IV A and shown in Figure 2. This established that using log spectrogram features for denoising achieved higher classification accuracy (for example a 3% increase in white noise at an SNR of 0dB), which we attribute to the better conditioned spectral values the log provides, making learning the residual function more effective. For clarity, we now report only DNCNN results using the log spectrogram feature as input.

Figure 7 shows that the two denoising methods using the vestigial trained classifier (DAE-MATCH-VES and DNCNN-MATCH-VES) attain best performance and their accuracy is almost equal in all noise conditions. When these two denoising approaches are applied to the clean-trained classifier their performance reduces. This suggests that the denoising methods are not able to remove the contaminating noise completely. However, classifying the output spectrograms using a classifier trained on the vestigial noise is able to recover performance. The results also suggest that the DAE is better able to remove noise and minimise distortion as its mean performance using the clean-trained classifier is higher than the DNCNN with the clean classifier as shown in Table I.

To illustrate the denoising ability of the DAE and DNCNN, the top row of Figure 8 shows a single upcall example that has been contaminated by each of the four noise types at an SNR of -5dB. For comparison, the original noise-free upcall is shown in Figure 1. The bottom two rows show spectrograms resulting from denoising with the DAE and DNCNN, and all spectrograms are shown using the same amplitude scale. These show that slightly more vestigial components remains after the DNCNN which may explain its lower performance compared to the DAE in Table I.

As a final investigation, the confusions between the three classes (*upcall(U)*, *gunshot(G)* and *no whale(NW)*) are examined across the four noise types. Tables II and III show confusions in white noise and shot noise at an SNR of 0dB with no denoising (i.e. CLEAN). Confusions in tanker and trawler noises were very similar to those in white noise and so are not shown. In white noise, gunshots are classified more accurately than upcalls, while in

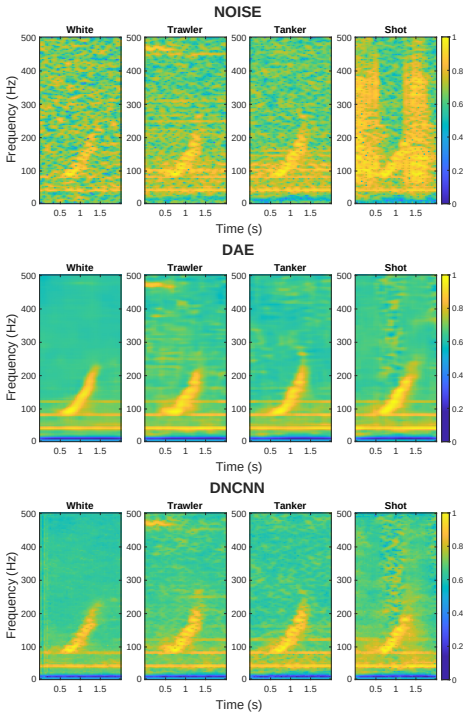

FIG. 8. *First row shows spectrograms of a single upcall (as displayed in Figure 1) that has been contaminated by white, trawler, tanker and shot noises at an SNR of -5dB. The second and third rows show the corresponding denoised spectrograms as produced by the DAE and DNCNN methods. The colourbar shows an amplitude range of 0 to 1 as these spectrograms are output from the denoising methods that are themselves trained on spectrograms with normalised energies, as discussed in Section III A.*

shot noise, upcalls are classified more accurately. This we attribute to the shot noise having more similar characteristics to gunshot vocalisations and so introducing more confusion. Tables IV and V show confusion matrices for the same two scenarios but now with denoising applied (specifically DAE-MATCH-CLEAN). In white noise, the primary effect of denoising is to reduce the percentage of no whale instances that are misclassified as either upcalls or gunshots, which represents a reduction in false alarms. This also happens when denoising in shot noise, but in addition, denoising also reduces the large number of gunshots that were misclassified as upcalls and are now classified correctly.

C. Denoising in unseen noise conditions

The previous tests were carried out where the denoising method was matched to the noise condition under test. In this section, the denoising training is no longer matched to the noise condition under test and instead is trained on different noise type and SNR conditions. Specifically, two scenarios are considered. First, where the denoiser is trained on all four noises and four

TABLE II. Confusion matrix for no denoising in white noise at 0dB SNR.

	U	G	NW
U	76%	6%	18%
G	1%	89%	10%
NW	9%	9%	82%

TABLE III. Confusion matrix for no denoising in shot noise at 0dB SNR.

	U	G	NW
U	58%	3%	39%
G	33%	51%	16%
NW	37%	21%	52%

TABLE IV. Confusion matrix for DAE denoising in white noise at 0dB SNR.

	U	G	NW
U	75%	2%	23%
G	0%	89%	11%
NW	4%	1%	95%

TABLE V. Confusion matrix for DAE denoising in shot noise at 0dB SNR.

	U	G	NW
U	84%	0%	16%
G	0%	81%	19%
NW	24%	1%	75%

SNRs (DAE-GENERIC-VES and DNCNN-GENERIC-VES) and secondly where training is on the three noise types that are not under test, which gives an unseen test condition (DAE-UNSEEN-VES and DNCNN-UNSEEN-VES). Given its superior performance in the previous section, all tests use the classifier trained on vestigial data rather than the clean-trained model. For comparison, results with no denoising are also shown and include the clean-trained model (CLEAN), the generic model (GENERIC) and unseen model (UNSEEN), as defined in Table I, with results shown in Figure 9. Methods that include training across all noise types and SNRs (GENERIC, DAE-GENERIC-VES and DNCNN-GENERIC-VES) achieve highest accuracies across all test conditions. This is attributed to the models having been trained on noise data that has similar characteristics to the specific test condition, whether it be in the denoising process (DAE-GENERIC-VES and DNCNN-GENERIC-VES) or in the classification stage (GENERIC). Moving to the unseen noise situations, where training does not include examples of the specific noise type under test, this leads to a reduction in accuracy for all systems (UNSEEN, DAE-UNSEEN-VES and DNCNN-UNSEEN-VES). Testing in white noise and shot noise, accuracy falls substantially below that of the equivalent systems trained on all noise types (i.e. the GENERIC systems), while for tanker and trawler noises the reduction in performance is much less. This we attribute to the similarity between tanker and trawler noises which allows the methods to learn at least some characteristics of the unseen noise and thereby perform better than the clean-trained model.

D. Performance in new unseen conditions

The evaluation of denoising methods in the previous sections used simulated noisy conditions by mixing

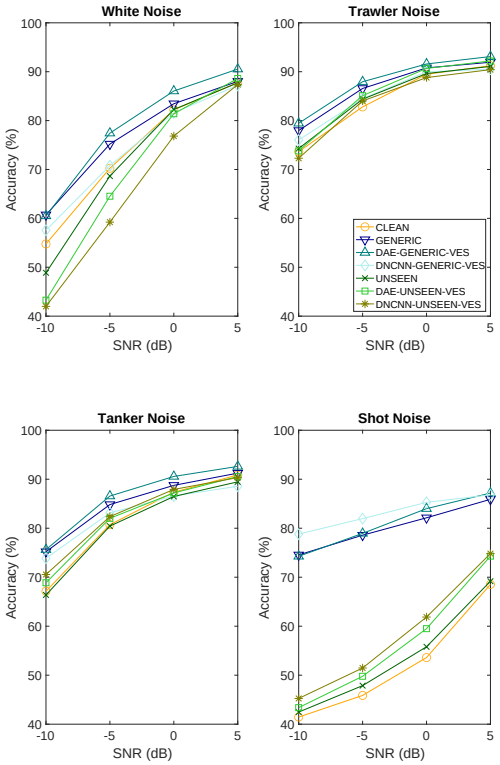

FIG. 9. *Right whale detection accuracies using denoising autoencoders and denoising CNNs in the four noises at SNRs from -10dB to +5dB. Results are shown with the denoising methods trained generically on all noise types or on noises not used in testing.*

clean audio with different noise types at varying SNRs. This is well suited for controlled evaluations of performance. We now consider an alternative scenario where the performance of whale detection on real noisy data is investigated. For this evaluation, data is taken from the Cape Cod corpus which was described in Section VI A and collected from a marine environment different from the Stellwagen corpus. Spectrogram analysis and listening to recordings has revealed them to contain significant amounts of different noise types which therefore represent a genuine unseen condition. To illustrate the recordings from Cape Cod, Figure 11a shows twelve example spectrograms of upcalls, arranged as a 2×6 grid. This shows that continuous broadband noise is present in most recordings as well as shorter duration impulses and some tonal noise, depending on the particular example.

Based on the evaluation in Section VII C on unseen noise conditions, the performance on the Cape Cod data is now evaluated using the CLEAN, GENERIC, DAE-GENERIC-VES and DNCNN-GENERIC-VES configurations. Instead of measuring classification accuracy, as has been done previously, these tests consider the task of whale detection (i.e. detecting whether a whale is present or not in a recording). For a practical whale detection system, knowing its precision and recall performance is more useful than classification accuracy. Consequently, we evaluate using these metrics with results shown for the four systems as precision-recall curves in Figure 10. It should be noted that the number of 'whale' and 'no whale' examples are equal. The DAE-GENERIC-VES, DNCNN-GENERIC-VES and GENERIC systems have similar precision-recall profiles. These all outperform the CLEAN system, particularly at higher levels of recall, where their precision is substantially better. This is investigated further in Figures 11b and 11c which show denoised spectrograms from the DAE and DNCNN systems that correspond to the noisy examples in Figure 11a. Both denoising methods are effective at removing much of the noise present in the original spectrograms of Figure 11a, and both do leave a vestigial signal. This reinforces the benefit of using a classifier trained on the vestigial signal rather than on clean data.

To compare with the DCLDE data, we also measured classification accuracy for the four methods and found that DAE-GENERIC-VES was best with 84.3%, followed by DNCNN-GENERIC-VES at 83.8%. The GENERIC system attained 81.7% and CLEAN 79.5%. From Table I, the DAE-GENERIC-VES method also outperforms DNCNN-GENERIC-VES, and both improve over the CLEAN model.

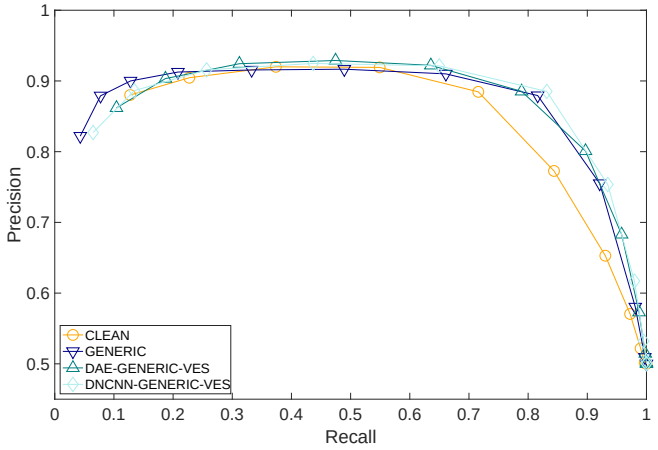

FIG. 10. *Precision-recall curves for the CLEAN, GENERIC, DAE-GENERIC-VES and DNCNN-GENERIC-VES models that are trained on Stellwagen data and tested on unseen recordings from Cape Cod.*

E. Classification processing times

An important consideration when deploying a practical right whale detection system is the processing time required to make a decision. This is examined by measuring the time taken from receiving a two-second block of audio to making a classification decision, which includes computing the spectrogram, denoising (where applied) and classification. Times were computed by averaging across the entire test set of recordings. The tests were performed on an Intel Quad Core i7 2.8GHz CPU which is a more realistic test than using a GPU, as was used in

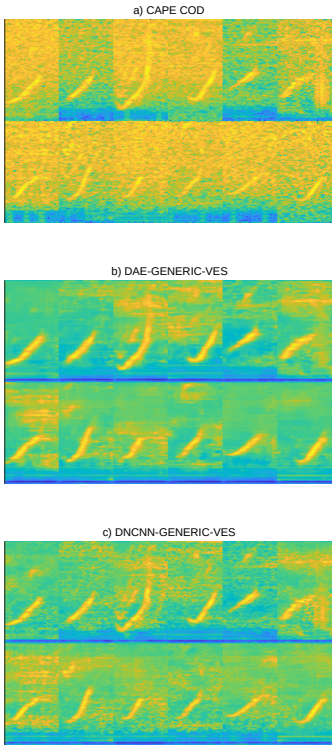

FIG. 11. *a) Spectrograms of twelve example upcalls taken from Cape Cod recordings, b) the resulting spectograms of the same upcalls after DAE-GENERIC-VES denoising, c) the resulting spectrograms after DNCNN-GENERIC-VES denoising.*

training. Three systems were evaluated: CLEAN, DAE-GENERIC-VES and DNCNN-GENERIC-VES, with the total time taken to process each two-second block broken down into the spectrogram extraction, denoising and classification times and shown in Table VI. This shows that all methods can process a two-second recording well within real-time constraints. The slowest method was the DNCNN-GENERIC-VES, where the majority of processing is taken by the denoising CNN although this is still capable of operating at 35-times real-time. The DAE-GENERIC-VES method of denoising was substantially faster, primarily due to the DAE denoising method operating eight times faster than the DNCNN denoising, which we attribute to it having fewer layers. Spectrogram extraction is the fastest of all stages, requiring just 0.72ms. In a practical deployment, these very fast classification times would allow a single CPU to process multiple channels of hydrophone array data simultaneously in real-time, 205 channels for the DAE and 35 channels for the DNCNN, ignoring multiplexing overheads.

VIII. CONCLUSION

This work has considered the problem of developing a robust system to detect right whales in differing noise conditions. Having the ability to deploy such an automated system, whether it be on buoys, ASVs or gliders,

Method	Spectrogram	Denoising	Classification	Total (ms)
CLEAN	0.72	-	2.63	3.35
DAE	0.72	6.40	2.63	9.75
DNCNN	0.72	53.02	2.63	56.37

TABLE VI. Mean processing times (in ms) for the spectrogram extraction, denoising and classification operations for the CLEAN, DAE and DNCNN methods when applied to a two-second audio recording.

that can achieve high levels of detection in real-time is vital to the long term future of right whales (Baumgartner *et al.*, 2020). Without applying any noise compensation, the CNN classifier was affected adversely in all four noise types with performance reducing from 94% in clean conditions to as low as 42% in severe conditions. Using augmentation to provide the classifier with data more representative of the noises under test improved performance. The DAE and DNCNN methods of denoising were to be able to improve accuracy further and this was more effective when the classifier was retrained on data containing the vestigial signal left after the denoising process. This indicates that denoising is not fully effective, and observations of spectrograms confirmed that artefacts do remain. However, matching the classifier to these artefacts through augmented vestigial training increases performance considerably. When compared to the conventional LSA method of denoising, this was found not to be able to increase accuracy above the clean-trained model and we attribute this in part to the method introducing artefacts that are confusable with whale vocalisations. This supports the benefit of applying denoising in the spectrogram domain.

When denoising was applied to unseen noise conditions, the accuracy of both methods reduced but remained higher than with no denoising. The evaluation also showed that if the unseen noise had similar attributes to the training data, such as training on one kind of shipping noise and testing on another, then higher performance was achieved. This suggests that if a representative collection of noises is used within training then good accuracy across a wide range of conditions is achievable. Furthermore, only a single model is required in this situation rather than a set of models matched to each noise condition. When the denoising methods were applied to a completely unseen noise condition the DAE improved performance substantially over the baseline classifier. The DNCNN performed almost as well and this was a trend observed throughout the evaluations - see Table I for a summary. Analysis of spectrograms produced from both denoising methods showed that clean predictions from the DNCNN were slightly more noisy while the DAE was able to remove the majority of noise as well as introduce fewer artefacts. This is beneficial not only for improving automated classification but also for human annotators, where having access to denoised spectrograms makes identification of whale sounds easier.

Measurement of processing times revealed the DAE to operate at 205 times real-time compared to 35 times real-time for the DNCNN. The faster operation and higher classification accuracy achieved by the DAE suggest this is a better choice for denoising within the framework of robust detection of right whales.

ACKNOWLEDGMENTS

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

[1]https://soi.st-andrews.ac.uk/static/soi/dclde2013/documents/WorkshopDataset2013.pdf

[2]https://github.com/williamvickerss/RightWhale_Jasa2021

Baumgartner, M., Fratantoni, D., Hurst, T., Brown, M., Cole, T., Van Parijs, S., and Johnson, M. (**2013**). "Real-time reporting of baleen whale passive acoustic detections from ocean gliders," The Journal of the Acoustical Society of America **134**(3), 1814–1823.

Baumgartner, M. F., Bonnell, J., Corkeron, P. J., Van Parijs, S. M., Hotchkin, C., Hodges, B. A., Bort Thornton, J., Mensi, B. L., and Bruner, S. M. (**2020**). "Slocum gliders provide accurate near real-time estimates of baleen whale presence from human-reviewed passive acoustic detection information," Frontiers in Marine Science **7**, 100.

Clark, C., Gillespie, D., Nowacek, D., and Parks, S. (**2007**). *Listening to their world: acoustics for monitoring and protecting right whales in an urbanized ocean* (In: Kraus SD, Rolland RM (eds) The urban whale: North Atlantic right whales at the crossroads. Harvard University Press, Cambridge, MA), pp. 333–357.

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

Cohen, I. (**2002**). "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," IEEE Signal Processing Letters **9**(4), 113–116.

Corkeron, P., Hamilton, P., Bannister, J., Best, P., Charlton, C., Groch, K., Findlay, K., Rowntree, V., Vermeulen, E., and Pace III, R. (**2018**). "The recovery of north atlantic right whales, Eubalaena glacialis, has been constrained by human-caused mortality," Royal Society open science **5**(11), 180892.

Davies, K., and Brillant, S. (**2019**). "Mass human-caused mortality spurs federal action to protect endangered north atlantic right whales in Canada," Marine Policy **104**, 157–162.

Davis, G., Baumgartner, M., Bonnell, J., Bell, J., Berchok, C., Thornton, J., Brault, S., Buchanan, G., Charif, R., Cholewiak, D., and Clark, C. (**2014**). "Long-term passive acoustic recordings track the changing distribution of North Atlantic right whales (Eubalaena glacialis) from 2004 to 2014," Scientific reports **7**(1), 1–12.

Gillespie, D. (**2004**). "Detection and classification of right whale calls using an 'edge' detector operating on a smoothed spectrogram," Canadian Acoustics **32**(2).

Glorot, X., and Bengio, Y. (**2010**). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Y. W. Teh and M. Titterington, JMLR Workshop and Conference Proceedings, Chia Laguna Resort, Sardinia, Italy, Vol. 9, pp. 249–256.

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

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

He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Ibrahim, A. K., Zhuang, H., Erdol, N., and Muhmed Ali, A. (**2018**). "Detection of North Atlantic right whales with a hybrid system of CNN and dictionary learning," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1210–1213.

Ioffe, S., and Szegedy, C. (**2015**). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Lille, France, Vol. 37, pp. 448–456.

Kingma, D., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," International Conference on Learning Representations .

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (**2012**). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Curran Associates, Inc., Vol. 25, pp. 1097–1105.

Leiter, S., Stone, K., Thompson, J., Accardo, C., Wikgren, B., Zani, M., Cole, T., Kenney, R., Mayo, C., and Kraus, S. (**2019**). "North Atlantic right whale Eubalaena glacialis occurrence in offshore wind energy areas near Massachusetts and Rhode Island," Endangered Species Research **34**, 45–59.

Liu, F., Song, Q., and Jin, G. (**2020**). "The classification and denoising of image noise based on deep neural networks," Applied Intelligence 1–14.

Loizou, P. (**2013**). *Speech Enhancement: Theory and Practice* (CRC Press, Inc.).

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (**2013**). "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, Vol. 2013, pp. 436–440.

Mellinger, D. K. (**2004**). "A comparison of methods for detecting right whale calls," Canadian Acoustics **32**(2), 55–65.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," The Journal of the Acoustical Society of America **107**(6), 3518–3529.

Milner, B. (**2002**). "A comparison of front-end configurations for robust speech recognition," in *ICASSP*, pp. 797–800.

Mouy, X., Bahoura, M., and Simard, Y. (**2009**). "Automatic recognition of fin and blue whale calls for real-time monitoring in the st. lawrence," The Journal of the Acoustical Society of America **126**, 2918–28.

Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, p. 807–814.

Nazaré, T., De Barros Paranhos da Costa, G., Contato, W., and Ponti, M. (**2018**). *Deep Convolutional Neural Networks and Noisy Images*, 416–424 (Springer).

Pace III, R., Corkeron, P., and Kraus, S. (**2017**). "State–space mark–recapture estimates reveal a recent decline in abundance of North Atlantic right whales," Ecology and Evolution **7**(21), 8730–8741.

Parks, S., Hotchkin, C., Cortopassi, K., and Clark, C. (**2012**). "Characteristics of gunshot sound displays by North Atlantic right whales in the Bay of Fundy," The Journal of the Acoustical Society of America **131**(4), 3173–9.

Parks, S., Searby, A., Célérier, A., Johnson, M., Nowacek, D., and Tyack, P. (**2011**). "Sound production behavior of individual North Atlantic right whales: implications for passive acoustic monitoring," Endangered Species Research **15**(1), 63–76.

Pylypenko, K. (**2015**). "Right whale detection using artificial neural network and principal component analysis," in *2015 IEEE 35th International Conference on Electronics and Nanotechnology (ELNANO)*, pp. 370–373.

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

Shiu, Y., Palmer, K., Roch, M., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D., and Klinck, H. (**2020**). "Deep neural networks for automated detection of ma-

rine mammal species," Scientific Reports **10**, 607.
Simonyan, K., and Zisserman, A. (**2015**). "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*.
Smirnov, E. (**2013**). "North atlantic right whale call detection with convolutional neural networks," in *Proc. Int. Conf. on Machine Learning, Atlanta, USA*, Citeseer, pp. 78–79.
Spaulding, E., Robbins, M., Calupca, T., Clark, C. W., Tremblay, C., Waack, A., Warde, A., Kemp, J., and Newhall, K. (**2009**). "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, Vol. 6, p. 010001.
Van Parijs, S., Clark, C., Sousa-Lima, R., Parks, S., Rankin, S., Risch, D., and Van Opzeeland, I. (**2009**). "Management and research applications of real-time and archival passive acoustic sensors over varying temporal and spatial scales," Marine Ecology Progress Series **395**, 21–36.
Varga, A., and Steeneken, H. (**1993**). "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication **12**(3), 247–251.
Verfuss, U. K., Aniceto, A. S., Harris, D. V., Gillespie, D., Fielding, S., Jiménez, G., Johnston, P., Sinclair, R. R., Sivertsen, A., Solbø, S. A. *et al.* (**2019**). "A review of unmanned vehicles for the detection and monitoring of marine fauna," Marine Pollution Bulletin **140**, 17–29.
Vickers, W., Milner, B., Lee, R., and Lines, J. (**2019**a). "A comparison of machine learning methods for detecting right whales from autonomous surface vehicles," in *27th European Signal Processing Conference, EUSIPCO*, pp. 1–5.
Vickers, W., Milner, B., Lines, J., and Lee, R. (**2019**b). "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*.
Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (**2017**). "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," IEEE Transactions on Image Processing **26**(7), 3142–3155.