Detecting right whales from autonomous surface vehicles using RNNs and CNNs

W. Vickers
School of Computing Sciences
University of East Anglia
Norwich, UK
w.vickers@uea.ac.uk

B. Milner
School of Computing Sciences
University of East Anglia
Norwich, UK
b.milner@uea.ac.uk

J. Lines
School of Computing Sciences
University of East Anglia
Norwich, UK
j.lines@uea.ac.uk

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

Abstract—**This work is concerned with the problem of detecting right whales from autonomous surface vehicles (ASVs) and investigates the effectiveness of a range of deep learning methods. Given the success of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) across many different applications, these form the basis for comparison. From the input audio, power spectral features are extracted and consideration is given to how their time resolution and frequency resolution affects the detection accuracy and the number of points that need to be processed which is an important consideration within the limited processing power on an ASV. The effect of downsampling the input audio before feature extraction is also investigated. Tests establish that CNNs consistently give best performance on the detection task with accuracy of over 92% compared to around 90% with RNNs. Furthermore, tests measuring the processing time for detection found the CNN to be three times faster than the RNN.**

Index Terms—**Cetacean detection, CNNs, RNNs, machine learning, autonomous surface vehicles**

I. Introduction

This work investigates the performance of deep learning techniques on the detection of right whale up-sweep vocalisations. This investigation is aimed at providing an insight into the best detection methods for use on an autonomous surface vehicle (ASV) where processing power and communications are limited. Accurate detection of marine mammals is important for population monitoring and for mitigation as many species are endangered and protected by environmental laws. We specifically explore the detection of North Atlantic right whales (*Eubalaena glacialis*) as they are currently under threat of extinction. Detecting their presence before they enter a mitigation zone both protects the animal and avoids the shutdown of costly offshore operations.

As the human population grows so does the demand for commercial shipping. With this comes increased ocean sound, much of which has recently been under scrutiny for impacting the wellbeing of marine mammals. Ship sounds such as propellers and engine noise are often the source of loud low frequency tones within the ocean. These have the potential to not only interfere with marine mammal communication but also effect their physiological stress levels resulting in possible fatalities [1]. Military sonar testing has also been hypothesised as the cause of mass cetacean fatalities in Greece 1996, with the post mortem report concluding that injuries were consistent with acoustic or impulsive signals causing cardiovascular collapse, which is often associated with extreme stress [1]. With a number of studies providing strong evidence to suggest physiological harm to marine mammals through anthropogenic noise, it is necessary to create techniques to help mitigate the future risk to mammals.

Detection of cetaceans has traditionally been made by human observers on-board ships, but more recently ASVs have been used [2]. Using an ASV limits the detection to using just an acoustic signal, as opposed to visual with a human observer, but it provides a cheaper and more accessible alternative. ASVs typically employ passive acoustic monitoring (PAM) which processes acoustic signals from a hydrophone to determine if marine mammals are present. This presents a number of challenges that include performing audio analysis and detection with limited processing power whilst maximising detection accuracy.

A broad range of machine learning techniques have previously been applied to cetacean detection. For example, methods such as vector quantisation and dynamic time warping have been effective in detecting blue and fin whales from their frequency contours extracted from spectrograms [3]. Hidden Markov models (HMMs) have also been effective at recognising low frequency whale sounds using spectrogram features [4], [5]. Further research utilised artificial neural networks (ANNs) for right whale detection, comparing its effectiveness to that of spectrogram correlation, with the ANN giving a better performance in samples with low signal-to-noise ratio (SNR) [6]. Neural networks were further used for classifying clicks of Blainville's beaked whales, with a good performance recorded for correctly detecting beaked whale clicks [7]. Convolutional neural networks (CNNs) have been also applied to whale detection. For example, [8] explored using CNNs to detect whale sounds by creating spectral images represented as a series of mel-frequency cepstral vectors extracted from the input audio.

Given the widespread success of deep learning across a range of applications that include image classification, speech recognition and text classification [9]–[11], the aim of this work is to investigate their effectiveness for right whale detection. Specifically, CNN and recurrent neural network (RNN) architectures are developed and optimised for the task of right whale detection. Furthermore, we also investigate the effect on detection accuracy of the time and frequency resolution of the features extracted from the input audio. Finally, given that an application of this work is deployment on ASVs, processing times for the two architectures are also considered which is an important consideration in low power applications.

The remainder of the paper is organised as follows. Section II introduces the right whale and describes the characteristics of the sounds produced. Detection within the constraints of an ASV are discussed in Section III. Sections IV, V and VI explain the process of feature extraction and the CNN and RNN architectures developed. Finally, detection performance across a range of architectures and configurations are presented in Section VII.

II. Right whale characteristics

Cetaceans are a large and diverse group of marine mammals. They are split into two suborders, odontocetes (toothed whales) and mysticetes (baleen whales). Odontocetes have teeth and feed on fish whilst mysticetes have a comb like structure (baleen) which helps them to feed on large amounts of crustaceans and zooplankton at once. Right whales are part of the mysticeti suborder and are known to move seasonally to feed and give birth [12]. Communication between whales is achieved primarily through sound. Large amounts of water make sight extremely difficult however sound propagation over hundreds of kilometres is very common. Most cetaceans can vocalise in several ways with whistles, clicks and burst pulses being the most common [13]. These methods of vocalisation have been predominantly recorded for use in the tasks of communication, feeding and navigation.

With as few as 350 individuals remaining [14], our focus is on right whales. They currently have a high possibly of extinction due to human activity within areas where they migrate. Right whale calls have been well documented [15] and we focus on their most commonly documented sound, a tonal up-sweep from approximately 60Hz to 250Hz, typically lasting 1 second [16]. Tonal up-sweeps are believed to be used as contact calls and are produced by all ages of animal [17]. Examples of these tonal sounds are shown in Figure 1 which illustrates calls at different signal-to-noise ratios (SNRs) caused by marine noise. Calls, however, are not always consistent with one another and can often vary in duration, frequency range, by time of day, season and age of the animal [18]. Right whale vocalisation patterns are also extremely variable with periods of silence regularly spanning many hours [19]. Right whale calls can be difficult to hear or visualise in spectrograms as their calls can be hidden within background noise. Low frequency bands are often congested with anthropogenic sounds such as ship noise, drilling, piling, military sonar or explosives. Figure 1 shows a range of spectrograms with varying amounts of background noise.

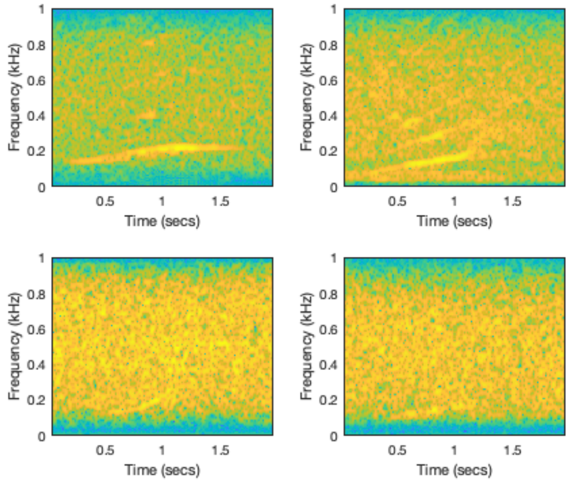

Fig. 1. *Example spectrograms showing up-sweep calls from right whales. Top Left: High SNR; Top Right: Medium SNR with clear harmonic structure; Bottom: Low SNR with low energy up-sweeps and high noise*

III. Detection from autonomous surface vehicles

Current methods of collecting cetacean data involve towing a hydrophone array from a ship and using trained observers to listen and watch the water for mammal activity. Visual surveys are often hindered by poor weather conditions (e.g. high seas, fog, presence of ice and darkness), uncertainty of species, and short surfacing intervals [20], causing this method to be extremely unpredictable for mammals that rarely surface. A combination of visual and acoustic monitoring from a ship will hypothetically yield the best result for detection however, ship time is expensive and often sporadically timed. Contrary to this, PAM-only systems can record continuously for several months at a time, without human interaction, giving a cheaper solution with a much higher likelihood of recording the animal of interest [2]. Furthermore, using an ASV with a PAM system will minimise local noise as no ship is needed for movement of the hydrophone.

For the task of mitigation monitoring a positive detection result needs to be communicated immediately so that mitigation measures can be put in place to protect the animal. This differs from, for example, population monitoring where data is stored on an ASV and then transferred and processed at a later time. Two potential ASV architectures can be considered for mitigation monitoring and can be termed ‘thick’ and ‘thin’. The ‘thick’ ASV samples the acoustic data from the hydrophone and inputs this into an on-board detection algorithm with positive detections transmitted for mitigation alert. The ‘thin’ ASV performs only the sampling on-board and transmits the data remotely for detection processing and mitigation alerts. Providing communication beyond a few miles, where a wireless modem could be employed, requires a satellite link. For the ‘thin’ ASV, the communication costs are generally prohibitive as a permanent satellite link is necessary. Furthermore, transmission would likely exceed the 2.4kbps limit for the Iridium network and thereby require a connection to the Inmarsat network which is substantially more expensive and has much higher power consumption (100 W, as opposed to 2.5 W). Based on these limitations of the 'thin' ASV architecture, we consider only the 'thick' ASV and explore how processing requirements can be minimised. To reduce false alarms (and the potentially large resulting costs) with the 'thick' ASV architecture, the segment of audio associated with a detection can be transmitted for a human to check, with the frequency of occurrence of this unlikely to be prohibitive in terms of transmission costs.

IV. Feature extraction

The purpose of feature extraction is to transform the input audio signal into a representation that is more effective for detecting whale sounds. Although many different methods of audio feature extraction have been developed (for example MFCCs, PLP, filterbank [21]) we chose to use a standard power spectral representation. Our reasoning is that we wish to allow the subsequent networks (CNN or RNN) to learn discriminative representations and not remove what could be useful information, such as may happen when using, for example, a mel-scaled filterbank.

Therefore, 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, Hamming windowed and a Fourier transform computed. The upper $N/2$ frequency points are discarded and the remaining points logged. Analysis windows are advanced by S samples to compute each new spectral vector. At a sampling frequency of f_s Hz, a total of $\frac{f_s-N+1}{S}$ spectral vectors are computed each second. This gives the total number of time-frequency points, L, that are produced each second as

$$L = \frac{f_s - N + 1}{S} \times \frac{N}{2} \tag{1}$$

Normalisation is applied to the elements of the power spectral vectors such that they are in the range 0 to 1.

V. CNN architectures for detection

Convolutional neural networks have been applied successfully to a number of tasks which include image classification, object detection, lip-reading and voice activity detection [9], [22]–[24]. By considering sequences of power spectral vectors as input images, the CNN can make a decision as to whether a whale sound is present or not.

The structure of the CNN takes the form of a number convolutional layers each followed by a max pooling layer, outputting into a final dense layer. In each convolutional layer a number of $M \times M$ filter kernels are convolved with the input and a ReLU non-linear activation function applied to their outputs. At the edges of the input, zero-padding is applied to maintain the size of the output. The final dense layer uses a sigmoid activation function and outputs a probability of whale detection. Section VII investigates the effect of different numbers of layers, numbers of filters and their sizes.

VI. RNN architectures for detection

The second deep learning architecture applied to whale detection is the recurrent neural network. RNNs are an effective architecture for temporal modelling and have been successful in a range of applications that include speech recognition and parsing [25], [26]. In the application of the RNN to whale detection, each feature vector is applied sequentially to the model, rather than the whole sequence of vectors as with the CNN.

The structure of the RNN comprises a sequence of recurrent layers, each with a number of nodes, followed, optionally, by dense layers. To avoid the diminishing gradients problem, each RNN layer is implemented using long short term memory (LSTM) cells [27]. All layers use the hyperbolic tangent activation function with the exception of the final layer which uses a sigmoid to output a probability of whale detection.

VII. Experimental results

The aim of these experiments is to explore the accuracy of the deep learning detection methods. Initially optimisation of the CNN and RNN is carried out by varying their individual architectures in order to reach a maximum accuracy for a fixed set of input features. After the highest achieving system for both CNN and RNN has been found we use these architectures to examine the effect of time and frequency resolutions in feature extraction and the effect of sampling frequency. This aims to identify whether reducing processing time through input features has an affect on detection accuracy and to discover which model performs best overall. All training was carried out over 100 epochs, using an RMSprop optimiser and a binary cross-entropy loss function.

Tests use a database of North Atlantic right whale up-calls that was obtained as part of the Marinexplore and Cornell University Whale Detection Challenge [1], where the audio is segmented into 2 second duration blocks. Each block is labelled as either containing a right whale or not, with annotations produced manually. A set of 10,934 audio blocks for are used for training, 1,122 for validation and 1,962 for testing. The training, validation and test sets are configured to contain equal numbers of blocks with and without right whales.

A. CNN optimisation

Initial testing with the CNN considered the effects of the number of layers, filter size and number of filters in each layer. Specifically, testing from 1 to 4 layers, each with 32, 64 or 128 filters of sizes of 3×3, 5×5 and 8×8 were performed. For consistency, the spectral vectors for all tests were extracted every 32ms using a 64ms window which gave a frequency resolution of 15.6Hz at the 2kHz sampling frequency that was used. The entire two second sequence of feature vectors formed the input image to the CNN. Highest detection accuracy of 91.6% was obtained with 3 layers with 32, 64 and 128 filters in each layer of size 3×3. This configuration forms the baseline architecture for further CNN tests.

[1] https://www.kaggle.com/c/whale-detection-challenge/data

B. RNN optimisation

Initial testing with the RNN considered the effects of the number of layers and nodes, investigating 1, 3 and 5 layers with 32, 64 or 128 nodes. Further tests added either 1 or 2 dense layers to the output of the RNN. For consistency, the spectral vectors for all tests were extracted using the same 32ms temporal resolution and 15.6Hz frequency resolution configuration as for the CNN tests. Stacking of the spectral vectors was also investigated with the input to the RNN comprising 1 to 4 stacked spectral vectors. Highest detection accuracy of 89.8% was found using 3 layers, all with 32 nodes, two dense layers, also with 32 nodes, and no frame stacking and this forms the baseline configuration for subsequent RNN tests. A bidirectional architecture was also tested, but gave no increase in accuracy.

C. Feature extraction and sampling frequency

These tests now compare the best performing CNN and RNN architectures under a range of different input features that vary according to their time and frequency resolutions and sampling frequency. Frame widths between 256ms and 16ms are considered first with a fixed 50% overlap of frames. This gives a time resolution, Δt, between 128ms and 8ms. In terms of the frequency resolution, Δf, this varies between 3.9Hz and 62.5Hz, depending on the window size and sampling frequency. The number of time-frequency points per second, L, for each configuration is computed using (1) and gives an indication of processing requirements. For each time resolution, Table I shows the resulting frequency resolution, number of time-frequency points per second and the detection accuracy for the CNN and RNN models, for sampling frequencies of 2kHz and 1kHz. We chose not to downsample further as initial testing showed a reduction in accuracy as this begins to loose spectral regions containing whale sounds.

For the CNN the highest accuracy for both sampling frequencies is found with the 64ms-7.8Hz time-frequency resolution, with 92.1% for 2kHz and 92.0% for 1kHz. The highest accuracy achieved by the RNN is 90.6% with a 1kHz sampling frequency using a 32ms-15.6Hz time-frequency resolution. This accuracy is reduced slightly to 90.4% with a 2kHz sampling frequency on a 16ms-31.3Hz resolution, however both are noticeably lower than the detection accuracy of the CNN. Considering the number of points (for both the CNN and RNN), and hence processing time, the 1kHz systems requires half the computations and gives almost equal performance to the 2kHz systems.

To compare the processing time for detection on the CNN and RNN, which is an important consideration for deployment on an ASV, the average time taken to process each 2 second audio block is measured. This is for detection only, not training, and tests are carried out on an Intel Core i7-870 CPU. For both the RNN and CNN the same input features of Δt=64ms and Δf=7.8Hz are used, both at the 1kHz sampling frequency. The average time per detection for the CNN is 3.05ms while for the RNN is 11.98ms. While these are much faster than real-time, in practice a much slower processor would be used on the ASV, making the faster detection time of the CNN more significant.

TABLE I

CNN AND RNN DETECTION ACCURACIES AND NUMBER OF POINTS PER SECOND FOR VARYING TIME AND FREQUENCY RESOLUTION FEATURES WITH 50% FRAME OVERLAP.

	Δt	128ms	64ms	32ms	16ms	8ms
2kHz	Δf	3.9Hz	7.8Hz	15.6Hz	31.3Hz	62.5Hz
2kHz	L	1489	1745	1873	1937	1969
2kHz	CNN	91.4%	**92.1%**	91.6%	90.2%	89.9%
2kHz	RNN	89.2%	89.9%	89.8%	**90.4%**	88.3%
1kHz	Δf	3.9Hz	7.8Hz	15.6Hz	31.3Hz	62.5 Hz
1kHz	L	745	873	937	969	985
1kHz	CNN	91.2%	**92.0%**	91.6%	90.6%	90.0%
1kHz	RNN	89.3%	90.5%	**90.6%**	89.8%	88.9%

Furthermore, we also compared our optimal CNN against a baseline ResNet architecture, pre-trained on the ImageNet dataset, with no additional layers [28]. This used the 64ms-7.8Hz acoustic feature sampled at 2kHz as input. This gave a detection accuracy of 91.4% which is slightly lower than the CNN trained on the same features. However, ResNet took 22ms to process each audio block which is seven times longer, which we attribute to its depth.

The tests in Table I were performed with 50% frame overlap which means that frequency resolution deteriorates as time resolution improves. In Table II we consider these independently by varying the frame slide, S, whilst using a fixed frame width. We specifically investigate two fixed widths to give both high and low frequency resolutions - Δf={3.9Hz, 15.6Hz}. The frame slide continues to be adjusted to give varying time resolutions, Δt, from 64ms to 8ms.The resulting CNN and RNN detection accuracies and number of time-frequency points per second are shown in Table II for the 2kHz and 1kHz sampling frequencies.

For both frequency resolutions and both sampling frequencies the time resolution has relatively little effect between 64ms and 16ms, with highest accuracy for both detection methods at 32ms. In terms of frequency resolution, the finer resolution gives higher accuracy for all CNN configurations. However, the maximum RNN accuracy is achieved using a frequency resolution of 15.6Hz, and where all configurations of the 1kHz sampling frequency achieve higher accuracy than their finer resolution counterparts. For example, highest performance of 92.5% is attained with a CNN using a 1kHz sampling frequency, 3.9Hz frequency resolution and 32ms time resolution which produces 2980 points. This could be reduced to 937 points (corresponding to a three times reduction in data) by using a wider frequency resolution but with a reduction in accuracy to 91.6%.

VIII. Conclusion

A range of deep learning methods have been applied to the detection of right whales, with the best performance, in terms of accuracy given by the CNN. An RNN also achieved high accuracies across a variety of input features however the accuracy of the CNN was consistently higher. Downsampling

TABLE II
CNN AND RNN DETECTION ACCURACIES AND NUMBER OF POINTS PER SECOND FOR VARYING TIME RESOLUTIONS AND FREQUENCY RESOLUTIONS OF 15.6HZ AND 3.9HZ.

	Δt	64ms	32ms	16ms	8ms
2kHz	Δf	15.6Hz	15.6Hz	15.6Hz	15.6Hz
2kHz	L	936.5	1873	3746	7492
2kHz	CNN	91.1%	**91.6%**	91.0%	90.0%
2kHz	RNN	89.1%	**89.8%**	88.7%	89.2%
2kHz	Δf	3.9Hz	3.9Hz	3.9Hz	-
2kHz	L	2978	5956	11912	-
2kHz	CNN	92.1%	**92.3%**	91.3%	-
2kHz	RNN	**90.2%**	89.8%	89.3%	-
1kHz	Δf	15.6Hz	15.6Hz	15.6Hz	15.6Hz
1kHz	L	468.5	937	1874	3748
1kHz	CNN	91.0%	**91.6%**	91.5%	91.0%
1kHz	RNN	90.4%	**90.6%**	90.5%	90.4%
1kHz	Δf	3.9Hz	3.9Hz	3.9Hz	3.9Hz
1kHz	L	1490	2980	5960	11920
1kHz	CNN	92.3%	**92.5%**	91.6%	91.0%
1kHz	RNN	**90.1%**	89.6%	89.6%	89.3%

the audio leaves accuracy almost unchanged with some tests showing better results. Downsampling however leads to a substantial reduction in processing time which is advantageous for use on ASVs. Considering time and frequency resolution the CNN results reveal that a wide time resolution of 32ms gives good accuracy whilst higher frequency resolutions are better, albeit at an increased processing cost. The RNN achieved its best accuracy using downsampled audio and also achieved higher accuracies from the wider 15.6Hz frequency resolution. Furthermore, the CNN was found to be able to perform detections three times faster than the RNN. This makes the CNN more suitable for deployment on an ASV, given both its higher detection accuracy and much smaller detection times.

ACKNOWLEDGMENT

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

REFERENCES

[1] T. M. Cox, T. Ragen, A. Read, E. Vos, R. W. Baird, K. Balcomb, J. Barlow, J. Caldwell, T. Cranford, and L. Crum, "Understanding the impacts of anthropogenic sound on beaked whales," SPACE AND NAVAL WARFARE SYSTEMS CENTER SAN DIEGO CA, Tech. Rep., 2006.

[2] U. K. Verfuss, A. S. Aniceto, D. V. Harris, D. Gillespie, S. Fielding, G. Jiménez, P. Johnston, R. R. Sinclair, A. Sivertsen, S. 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] X. Mouy, M. Bahoura, and Y. 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] D. K. Mellinger and C. 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] J. C. Brown and P. Smaragdis, "Hidden markov and gaussian mixture models for automatic call classification," *The Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. EL221–EL224, 2009.

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

[7] ——, "A neural network for classifying clicks of blainville's beaked whales (mesoplodon densirostris)," *Canadian Acoustics*, vol. 36, no. 1, pp. 55–59, 2008.

[8] E. 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.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[10] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8599–8603.

[11] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[12] P. O. Thomas and S. M. Taber, "Mother-infant interaction and behavioral development in southern right whales, eubalaena australis," *Behaviour*, pp. 42–60, 1984.

[13] C. Clark, "Acoustic communication and behavior of the southern right whale, eubalaena australis," *Communication and behavior of whales*, pp. 163–198, 01 1983.

[14] S. D. Kraus, M. W. Brown, H. Caswell, C. W. Clark, M. Fujiwara, P. K. Hamilton, R. D. Kenney, A. R. Knowlton, S. Landry, C. A. Mayo, W. A. McLellan, M. J. Moore, D. P. Nowacek, D. A. Pabst, A. J. Read, and R. M. Rolland, "North Atlantic Right Whales in Crisis," *Science*, vol. 309, no. 5734, pp. 561–562, Jul. 2005. [Online]. Available: http://science.sciencemag.org/content/309/5734/561

[15] C. W. Clark, "The acoustic repertoire of the southern right whale, a quantitative analysis," *Animal Behaviour*, vol. 30, no. 4, pp. 1060–1071, 1982.

[16] S. E. Mussoline, D. Risch, L. T. Hatch, M. T. Weinrich, D. N. Wiley, M. A. Thompson, P. J. Corkeron, and S. M. Van Parijs, "Seasonal and diel variation in north atlantic right whale up-calls: implications for management and conservation in the northwestern atlantic ocean," *Endangered Species Research*, vol. 17, no. 1, pp. 17–26, 2012.

[17] S. E. Parks, C. W. Clark, and P. L. Tyack, "Short-and long-term changes in right whale calling behavior: The potential effects of noise on acoustic communication," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3725–3731, 2007.

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

[19] J. N. Matthews, S. Brown, D. Gillespie, M. Johnson, R. McLanaghan, A. Moscrop, D. Nowacek, R. Leaper, T. Lewis, and P. Tyack, "Vocalisation rates of the North Atlantic right whale (Eubalaena glacialis)," p. 12, 2001.

[20] M. F. Baumgartner and S. E. Mussoline, "A generalized baleen whale call detection and classification system," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 2889–2902, 2011.

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

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

[23] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[24] T. le Cornu and B. Milner, "Voicing classification of visual speech using convolutional neural networks," *Proc. FAAVSP*, 2015.

[25] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[26] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," *Advances in Neural Information Processing Systems*, pp. 2773–2781, 2015.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.