Automatic annotation of spectrograms – Comparison of traditional energy detectors and ResNet deep learning to analyse 20 Hz fin whale calls

Shaula Garibbo[1], Philippe Blondel[1], Gary Heald[2,3], Ross Heyburn[4], Alan Hunter[5], and Duncan Williams[2]

[1]Department of Physics, University of Bath, Claverton Down, Bath, BA2 7AY, UK
[2]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, Scotland, EH14 4AS, UK
[3]Dstl: Defence Science and Technology Laboratory, Salisbury, Wiltshire, SP4 0JQ, UK
[4]AWE Blacknest, Aldermaston, Reading, RG7 4RS, UK
[5]Department of Electrical Engineering, University of Bath, Claverton Down, Bath, BA2 7AY, UK

Shaula Garibbo, 3.23 Department of Physics, University of Bath, Claverton Down, Bath, BA2 7AY, UK. Email: sg2340.bath.ac.uk

Abstract: *Passive underwater acoustic measurement systems produce very large amounts of data, which need to be analysed to detect sources of noise (e.g. ships, marine life, natural physical processes). Supervised/semi-supervised machine learning applications rely on annotated datasets for training. In this study, the annotated dataset comes from manual picking and the aim is that machine learning will produce automated detections fast and repeatably which are in agreement with the analyst's annotations. We consider data from two different ocean observatories (namely, Lofoten-Vesterålen (LoVe) in Norway and the Ascension Island station of the Comprehensive Nuclear-Test-Ban Treaty network), and three sampling rates (32 or 64 kHz at LoVe, 250 Hz at Ascension Island). We look at how the annotation of data, spectrogram parameters (such as window length and frequency resolution), and signal-to-noise in the training data affect performance. As well as examining whether or not the signals of interest are detected, accuracy in determining the start and end times of the signals is also considered.*

Keywords: *deep learning, detector, fin whale*

1. INTRODUCTION

The global distribution of fin whales mean that their 20 Hz call is a common feature of the world's ocean soundscape [1], found *inter alia* at observatories such as Lofoten-Vesterålen (LoVe, Norway) and Ascension Island (Comprehensive Nuclear-Test-Ban Treaty Organisation, CTBTO). Fin whale calls are 1 s long, sinusoid-like signals centred at 20 Hz but varying between 13 Hz and 29 Hz worldwide [1, 2]. They are important and can be used to determine marine mammal abundance with seasons and other environmental factors [1]. Examples of using automatic detectors to identify 20 Hz fin whale calls are numerous [3, 4, 5], the most common based on energy summation over the frequency band of interest (like the 'fin index') [6], correlation [7], and more recently, machine and deep learning [8]. Here we focus on a pre-existing deep learning detector [9], assess performance using different parameters, and compare it with an energy summation detector and a correlation detector. A map showing where the acoustic data was collected is included in Figure 1.

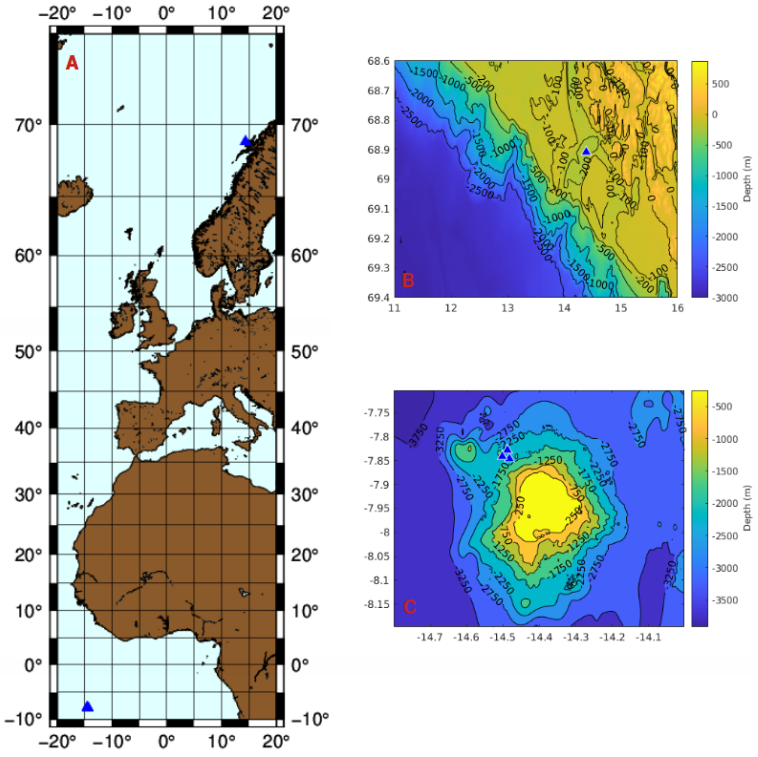

Figure 1: Left: A wide area map showing the locations of hydrophones (blue triangles) offshore Ascension Island (South Atlantic) and northern Norway (North Atlantic). Right, upper: Bathymetry at LoVe; the single hydrophone is 15 km offshore Lofoten, 255 m deep and 0.5 m above the seafloor. Right, lower: Bathymetry surrounding Ascension Island with northern hydrophone triplet. The hydrophones on average are 845 m deep and located in the SOFAR channel [10]. Bathymetry is mapped using Matlab's Mapping Toolbox [11] and GEBCO bathymetry data [12].

2. DEEP LEARNING DETECTOR

The deep learning detector is based on a pre-existing Ketos ResNet detector initially trained on data only from LoVe (details in Garibbo *et al.* 2021 [9]). Additional data from Ascension Island was added into the training data - resulting in 3570 training annotations, and 837 vali-

dation annotations. These annotations were then used with a range of parameters (like window length and time resolution) to generate different spectrograms to train the network, as well as different training approaches - specifically, if the data was augmented in training or not, and how the noise annotations were selected. These are addressed in the following subsections.

2.1. SPECTROGRAM WINDOW LENGTHS

Although the 20 Hz fin whale call is typically 1 s long, the signal can often appear longer in a spectrogram due to effects of propagation-like reverberation. Fin whale call lengths (including propagation effects), annotated by an analyst, ranged between 1.0 s to 3.4 s, so window lengths of 1, 2, and 3 seconds were tested to compare their effect on the performance criteria of accuracy, precision, recall, and timing - as displayed in Table 1. Accuracy, precision, and recall are calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN}$$

where TP: True Positive detection (the positive detector output coincides with the analyst's annotation of a fin whale call), FP: False Positive (the positive detector output coincides with the analyst's annotation of noise), TN: True Negative (the negative detector output coincides with the analyst's annotation of noise), FN: False Negative (the negative detector output coincides with the analyst's annotation of a fin whale call).

2.2. SPECTROGRAM TIME RESOLUTION

Time resolutions of the spectrograms fed into the Ketos deep learning network were varied between 0.1 s and 1.0 s to see what effect the finer details of the spectrogram had on the Ketos detector's performance. The results are displayed in Table 1.

Window Length (s)	Time resolution (s)	Augmentation (s)	No. of detections	No. of annotations	Accuracy (%)	Precision (%)	Recall (%)	Start time difference (s)	End time difference (s)
1	0.1	1	86	59	82.39	65.56	100.00	3.30 ± 2.98	2.92 ± 2.85
1	0.25	1	64	59	81.38	71.05	91.53	5.18 ± 4.85	3.42 ± 3.39
1	0.5	1	58	59	75.82	61.70	98.31	3.11 ± 2.85	5.44 ± 5.31
1	1	1	71	59	78.31	62.11	100.00	3.11 ± 2.84	4.70 ± 4.58
2	0.1	2	63	59	79.61	65.91	98.31	4.30 ± 3.95	5.40 ± 5.41
2	0.25	2	66	59	80.92	67.86	96.61	3.19 ± 2.93	1.65 ± 1.66
2	0.5	2	63	59	79.61	65.91	98.31	2.72 ± 2.50	1.93 ± 1.93
2	1	2	45	59	74.82	62.77	100.00	4.86 ± 4.68	3.39 ± 3.36
3	0.1	3	40	59	79.20	69.41	100.00	5.64 ± 4.53	17.62 ± 16.84
3	0.25	3	57	59	80.74	73.24	88.14	3.78 ± 3.58	4.58 ± 4.58
3	0.5	3	55	59	80.74	72.00	91.53	4.66 ± 4.30	1.42 ± 1.43
3	1	3	45	59	79.03	72.60	89.83	4.59 ± 4.16	5.47 ± 5.50

Table 1: Summary of key spectrogram parameters and associated Ketos detector performance on one 10-minute segment of LoVe data. The start and end time differences refer to the RMS mean difference between Ketos detection and analyst's annotation ± one standard deviation. All spectrograms were magnitude spectrograms between 10 and 30 Hz at 250 Hz sampling using a Hamming window. No data augmentation was implemented.

2.3. DATA AUGMENTATION

Data augmentation is a commonly utilised process in machine and deep learning [13, 15, 16]. There are various approaches [15], but their common aim is to increase the volume of training data for deep and machine learning models. Here we use a standard form of data augmentation available within the Ketos Python module [14], which involves lengthening the analyst's annotation and then shifting the position of the start and end times of the annotation within the segment that is fed to the network [14]. At least 50% of an annotation was included in the spectrogram segment fed into the network. The performance of the Ketos detector which was trained on augmented data is shown in Table 2.

2.4. NOISE SELECTION

The performance shown in Table 1 is based on randomly selected data using the Ketos inbuilt function for generating random background noise selections [14]. This function assumes that all audio present between the annotations of the signal of interest are noise. Although this visually appeared to be the case in many of the annotations, swapping out the files that the noise was randomly sampled from for files that contained no fin whale calls visible to an analyst had a significant effect - see Table 2. This may suggest that the detector was more sensitive to the boundaries of the fin whale calls than a human analyst.

Window Length (s)	Time resolution (s)	Augmentation (s)	No. of detections	No. of annotations	Accuracy (%)	Precision (%)	Recall (%)	Start time difference (s)	End time difference (s)
1	0.1	0.5	67	70	95.00	89.47	97.14	2.20 ± 2.22	3.13 ± 2.97
1	0.25	0.5	66	70	90.91	80.00	97.14	4.15 ± 2.95	2.95 ± 2.79
1	0.5	0.5	67	70	80.00	60.55	94.29	0.38 ± 2.32	2.94 ± 2.72
1	1	0.5	54	70	66.12	34.81	90.00	0.47 ± 2.19	4.75 ± 4.07

Table 2: Summary of the key spectrogram parameters and associated Ketos detector performance on one 10 minute segment of data from LoVe. The start and end time differences referring to the RMS mean difference between the Ketos detection and the analyst's annotation ± one standard deviation. Note: all spectrograms were computed as magnitude spectrograms between 10 Hz and 30 Hz at a sampling frequency of 250 Hz using a Hamming window.

3. COMPARISON

The Ketos detector was compared with two other automatic detectors - the energy summation and correlation detector capabilities in analytical acoustic software Ishmael [17]. All the data were bandpass-filtered and normalised between 13 and 30 Hz, with a specified minimum detection duration of 1 s and maximum of 10 s. The detection threshold in Ishmael is dictated by the detection function, which is a time-series function of arbitrary amplitude that is defined by the parameters set by the user - parameters of influence could be signal duration or bandwidth. These parameters also vary with the type of detector [17]. For the LoVe data, the detection threshold was set at the unitless value of 0.3, and for the Ascension Island data the detection threshold was set at 500. This difference is quite stark and the reasons are to be investigated in further work. The correlation detector was based on an average of five fin whale calls (duration and bandwidth) present in a range of files - but this also needed adjusting to better reflect the calls seen at LoVe and at Ascension Island in terms of frequency range and call duration. The detection threshold for this detector was set at 0.3 (unitless) as well. The performance of each

detector (including the best performing Ketos-based detector) can be seen in Table 3.

4. CONCLUSION

Detector	Accuracy (%)	Precision(%)	Recall (%)	Δs (s)	Δe (s)
Energy Sum	93.44	93.33	93.33	1.53 ± 1.21	2.12 ± 2.10
Correlation	71.83	68.25	68.25	2.31 ± 2.25	2.19 ± 2.37
Ketos	95.00	89.47	97.4	2.20 ± 2.22	3.13 ± 2.97

Table 3: Comparison of two traditional approaches to automatic fin whale detection and Ketos on LoVe and H10 data.

Tables 1 and 2 show the performance of the Ketos deep learning detector varies considerably with simple modifications to key spectrogram parameters. The timings of detections and how they compare to annotations may also be important to consider depending on the application; for example in Table 1 a 3 s window at 0.25 s time resolution yielded good accuracy, precision, and recall measurements, but the detections could be up to 9 seconds out from the analyst's.

Table 3 compares Ketos performance against two other detectors, and it appears that the correlation detector does not perform as well. This is not too surprising as fin whale 20 Hz calls, although very similar, are not identical in frequency bandwidth in different geographic regions and are known to change over time [3].

Although the energy summation detector is more computationally efficient than the deep learning detector, the threshold requires monitoring - especially when changing datasets. Although this could be argued as a minor point when considering the time and computational price of building and training a deep learning detector, future work can focus on the development of Receiver Operator Characteristic (ROC) curves to fully understand and compare performance. The results here have also been obtained using very small data segments; using a longer data set would be necessary for a more complete representation of the detectors' performance.

5. ACKNOWLEDGEMENTS

This research is supported by the UK Engineering and Physical Sciences Research Council (EPSRC), as part of industrial Cooperative Award in Science and Technology (iCASE) project #2279119, supported by Defence Science & Technology Laboratory (Dstl) and AWE Forensic Seismology.

REFERENCES

[1] Morano, J.L., Salisbury, D.P., Rice, A.N., Conklin, K.L., Falk, K.L., and Clark, C.W.: "Seasonal and geographical patterns of fin whale song in the western North Atlantic Ocean", *J. Acoust. Soc. Am.* **132(2)**, 1207-1212 **(2012)**.

[2] Watkins, W.A., Tyack, P., Moore, K.E., and Bird, J.E.: "The 20-Hz signals of finback whales (*Balaenoptera physalus*)", *J. Acoust. Soc. Am.* **82(6)**, 1901-1912 **(1987)**.

[3] Helble, T.A., Guazzo, R.A., Alongi, G.C., Martin, C.R., Martin, S.W., and Henderson, E.E.: "Fin Whale Song Patterns Shift Over Time in the Central North Pacific" *Frontiers in Marine Science* **7**, 1-16 **(2020)**.

[4] Pereira, A., Harris, D., Tyack, P., and Matias, L.: "Lloyd's mirror effect in fin whale calls and its use to infer the depth of vocalizing animals", **POMA 27(1)**, 070002 **(2016)**.

[5] Mellinger, D. K., and Clark, C. W.: Recognizing transient low-frequency whale sounds by spectrogram correlation", *J. Acoust. Soc. Am.* **107(6)**, 3518-3529 **(2000)**.

[6] Haver, S.M., Klinck, H., Nieukirk, S.L., Matsumoto, H., Dziak, R.P., and Miksis-Olds, J.L.: "The not-so-silent world: Measuring Arctic, Equatorial, and Antarctic soundscapes in the Atlantic Ocean", *Deep-Sea Res. I*, **122**, 95-104 **(2017)**.

[7] Širović, A., Rice, A., Chou, E., Hildebrand, J.A., Wiggins, S.M., and Roch, M.A.: "Seven years of blue and fin whale call abundance in the Southern California Bight", *Endangered Species Research* **28(1)**, 61-76 **(2015)**.

[8] Zhong, M., Torterotot, M., Branch, T.A., Stafford, K.M., Royer, J.-Y., Dodhia, R., and Lavista Ferres, J.: "Detecting, classifying, and counting blue whale calls with Siamese neural networks", *J. Acoust. Soc. Am.* **149(5)**, 3086-3094 **(2021)**.

[9] Garibbo, S., Blondel, P., Heald, G., Heyburn, R., Hunter, A., and Williams, D.: "Characterising and detecting fin whale calls using deep learning at the Lofoten-Vesterålen Observatory, Norway", **POMA 44(1)**, 070021 **(2021)**.

[10] Metz, D., Watts, A.B., Grevemeyer, I., Rodgers, M., and Paulatto, M.: "Ultra-long-range hydroacoustic observations of submarine volcanic activity at Monowai, Kermadec Arc", *Geophysical Research Letters* **43(4)**, 1529-1536 **(2016)**.

[11] MATLAB: "Mapping Toolbox" at *https://uk.mathworks.com/help/map/*, Accessed 24-Mar-2023.

[12] GEBCO Bathymetric Compilation Group 2020 (2020). The GEBCO 2020 Grid - a continuous terrain model of the global oceans and land. BODC/NOC, NERC, UK. DOI:10.5285/a29c5465-b138-234d-e053-6c86abc040b9.

[13] O.S. Kirsebom, F. Fabio, Y. Simard, N. Roy, S. Matwin, and S. Giard: "Performance of a deep neural network at detecting N. Atlantic right whale upcalls", *J. Acoust. Soc. Am.* **147**, 2637-2646 **(2020)**.

[14] Meridian: "Welcome to Ketos's documentation!", *https://docs.meridian.cs.dal.ca/ketos/* Accessed on 07/05/2021, **(2020)**.

[15] Padovese, B., Frazao, F., Kirsebom, O.S., and Matwin, S.: "Data augmentation for the classification of N. Atlantic right whales upcalls", *J. Acoust. Soc. Am.* **149(9)**, 2520-2530 **(2021)**.

[16] Goldwater, M., Zitterbart, D.P., Wright, D., and Bonnel, J.: "Machine-learning-based simultaneous detection and ranging of impulsive baleen whale vocalizations using a single hydrophone", *J. Acoust. Soc. Am.* **153(2)**, 1094-1107 **(2023)**.

[17] Mellinger D.K.: "Ishmael 1.0 User Guide", *NOAA Tech. Memo. OAR PMEL-120*, **(2001)**.