Automated Classification of Vocalisations from Wild and Captive Seal Populations

William Butler[1], Harrison Smith[2], Wil Lloyd[1], Marios Impraimakis[1], Andrew Barnes[1], and Alan Hunter[1]

[1]University of Bath, Bath, BA2 7AY, United Kingdom
[2]Celtic Sea Power, Hayle, TR27 4DD, United Kingdom

Contact Author: William Butler, wmb34@bath.ac.uk
Keywords: *seal vocalisation, bioacoustics, deep learning.*

1. INTRODUCTION

Grey seals (*Halichoerus grypus*) and Harbour seals (*Phoca vitulina*) are key marine predators distributed across the North Atlantic, with significant populations in the U.K. where they are protected under several conservation frameworks ([1], [2], [3]). As the U.K. rapidly expands offshore infrastructure, particularly floating offshore wind farms, understanding how these developments interact with marine species such as grey seals is increasingly critical ([4], [5]). Passive acoustic monitoring (PAM) provides a non-invasive method to assess seal presence, behaviour, and potential disturbance responses, especially in offshore regions beyond visual monitoring range. Grey seal vocalisations are known to vary seasonally and contextually ([6], [7]), but classification has been limited by observational constraints and variation in wild populations [8]. While recent work has begun to map acoustic repertoires using spectrogram features [9], the behavioural significance of individual call types remains poorly understood.

Some notable work has been done on the automated detection and classification of marine mammal vocalizations. PAMGuard's Whistle and Moan Detector [10] is a widely used tool for the detection of odontocete vocalisations through noise reduction and thresholding techniques. More recently, several works have successfully utilised neural network approaches to classify marine mammals ([11], [12]), [13]), though focus has generally been on whales and other types of odontocetes rather than seals.

Here, we introduce a novel contribution to grey and harbour seal bioacoustics: we present and empirically validate two previously undocumented call types, termed SP4 and SP5 through expert-tagged recordings, and establish their consistency across multiple vocalising individuals. We then use these labels to train a spectrogram-based deep learning model, contributing a scalable pipeline for future automated classification. This work supports acoustic ecology efforts in the southwest UK, where dialectal variation may exist [14], and informs conservation planning amid expanding offshore energy development.

Class	Count	Description
S1	741	Multiple Harmonic Tonals
S2	520	Non-tonal Bursts
S3	363	Harmonic Pulses
SP4	134	Dyadic Pulse Train
SP5	170	“Zipper” Pattern Pulse Train

Table 1: Seal call class densities and descriptions

2. METHODOLOGY

The data used in this paper was provided by Celtic Sea Power (CSP). The data is comprised of 2-hour, uninterrupted recordings of captive and wild seals, taken at the Cornish Seal Sanctuary and in-situ at a location near to Godrevy island, U.K. using the HydroMoth [15] and AudioMoth [16] hydrophones with a sample rate of 96kHz. Near Godrevy, the audio was taken from a “Camera Frame” and a “Lobster Pot” situated around 30 meters apart and contains vocalisations of Grey Seals. Audio from the Cornish Seal Sanctuary was taken from the Common Seal Pool and Grey Seals Pool.

2.1. DATA LABELLING AND CLASS DESCRIPTIONS

Seal vocalizations were manually tagged by CSP employees with a start time and end time, checked by the marine biologist, and classified into one of five types of seal calls. The call types were labelled S1, S2, S3, SP4, and SP5, expanding upon Nowak’s classification of S1, S2, and S3 [8]. The class density of the calls is described in Table 1. with a total of 1928 seal calls tagged and classified. Calls that were not classified into one of these five categories were removed and are not included in the subsequent figures in this work, nor used in training any of the models described later.

The S1 and S2 classes were both categorised as described in Nowak’s paper. S1 calls are described as tonal, with multiple harmonic elements. S2 calls are repeated short bursts of vocalisation, lacking any distinct tonal elements. It is theorised that these may be mating calls [8].

A subset of tagged signals was provisionally classified as S3-type vocalisations, following the definitions established by Nowak who describes S3 calls as short, pulsed sounds with distinct harmonic structure and no observable air exhalation during production. In the Nowak dataset, S3 vocalisations predominantly occur in repetitive sequences (dyads or pulse trains) with inter-pulse intervals typically shorter than 1 second, often around 0.5 seconds. In contrast, the majority of sounds classified at Godrevy as S3 were singular pulses, occasionally occurring in pairs or triplets, or as isolated sounds with longer inter-call intervals (e.g., 2 seconds). Although a few examples displayed characteristics consistent with Nowak’s descriptions such as sequences of 4–5 evenly spaced pulses with clear structure, the overall pattern in our dataset is less rhythmically structured and lacks consistent evidence of paired airflow dynamics.

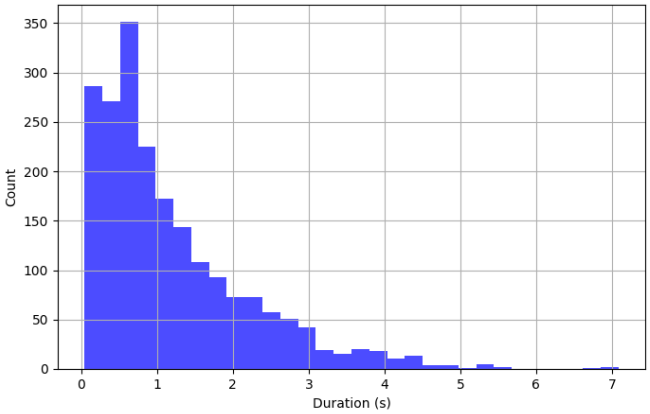


Figure 1: Histogram of call durations.

2.1.1. SP4 - RHYTHMIC DYADIC PATTERN

During the labelling process, a distinct acoustic signal—likely produced by a grey seal—was classified as "SP4", where "P" denotes *Provisional* categorisation. These calls typically last 1.5–2 seconds (ranging from 1–4+ seconds) and comprise 10–30 discrete, rhythmic elements. Each element features a steady tone with spectral peaks around 1 kHz, followed by a rapid downsweep to 250 Hz. The evenly spaced elements create a drumming or clapping-like rhythm, reminiscent of a seal clapping/striking its body as observed out of water. Closer analysis reveals a dyadic structure: elements occur in pairs with slightly shorter gaps within pairs than between them, producing a syncopated rhythm that occasionally shifts to even spacing mid-call.

2.1.2. SP5 - ZIPPER PATTERN PULSE TRAIN

Another distinct signal, labelled "SP5" was identified at Godrevy. Lasting ∼0.5 seconds (occasionally up to 1s), SP5 comprises 5–10 evenly spaced spectral peaks (0.5–1 kHz), and acoustically resembles a zipper or dragging a finger along metal coils, suggesting a tonal, percussive quality. SP5-type vocalisations are hypothesised to originate from a phocids (seals) as, based on the high intensity and propagation, they are unlikely caused by invertebrates or small fish which typically produce weaker, more broadband sounds. Known UK vocal fish (e.g. *Chelidonichthys* spp., *Sciaenidae*) do not match SP5's spectral or rhythmic features. Additionally, SP5 sometimes co-occurs or transitions into low-frequency "grunt" or "sniff" sounds typical of seals. These characteristics support a seal origin, though concurrent visual confirmation is recommended for species attribution.

2.2. DATA PRE-PROCESSING

Fig. 1 presents a histogram of call durations across all sites, showing that most last less than 3–4 seconds, with few exceeding 5 seconds and none over 10. Therefore, a 10-second time window was chosen for spectrogram generation.

For tagged seal calls, a random offset between 0 and (10 – call duration) seconds was applied before the labelled start time, and a 10-second audio clip was extracted. To estimate the noise floor at each site, 2–10 second gaps between tagged calls were used. Linear spectrograms from these segments were stacked horizontally and the median was computed for each frequency.

Spectrograms were generated using Python's Librosa library [17] with a Short-Time Fourier Transform (STFT) using an FFT window size of 2048, a Hann window, and a hop size of 1024, at a 96 kHz sample rate. Absolute FFT magnitudes were used, and the first and last frames were trimmed to avoid artifacts. This process yielded square spectrograms of shape (1025, 936), about 4× larger than ResNet's input (224×224), which will be introduced and used later in this paper. It is assumed there is acceptable information loss upon 4× downsampling. Finally, linear spectrograms were divided by the site-specific, frequency dependent noise floors and converted to dB scale (ref=1.0)

Several transformations were defined as part of the dataloaders for the fine tuning, including scaling, normalisation and resizing. ResNet expects a 3 channel input with values normalised to a mean of 0.485, 0.456 and 0.406 and standard deviation of 0.229, 0.224 and 0.225 for each RGB channel, respectively. The matrices generated from the STFT are 1 channel, so we take the mean of the ImageNet means (0.449) and standard deviations (0.226), as the matrices are duplicated across 3 channels to match the input dimensions expected by ResNet.

The standard deviation of the dB values, after noise reduction, is estimated by randomly sampling 20 seal and background noise spectrograms from each recording session at each location. The average standard deviation across all locations was found to be 6.6dB (2 s.f.). We assume a mean of 0 after the background noise subtraction in the dataset here and calculate appropriate minimum and maximum dB values of -13.11dB and 16.09dB for the following Min-Max scaling, to ensure the spectrograms approximate the mean and standard deviation of the data ResNet was initially trained on:

$$y = \text{clip}\left(\frac{x - \text{dB}_{\text{min}}}{\text{dB}_{\text{max}} - \text{dB}_{\text{min}}},\ 0,\ 1\right) \tag{1}$$

Finally, the data are normalised using the mean of 0.449 and standard deviation of 0.226 across all 3 channels and reshaped using linear interpolation to 224x224. Fig. 2 shows the original background subtracted spectrograms, the transformed inputs to the model and histograms of the pixel intensities after the transformations.

2.3. RESNET MODEL AND FINE-TUNING

A pretrained ResNet18 model from PyTorch [18], originally trained on ImageNet [19], was selected due to its broad success in computer vision tasks across domains [20]. The model was fine-tuned for a 5-class classification task corresponding to seal call classes S1, S2, S3, SP4, and SP5. The final fully connected layers of ResNet were replaced with the following sequence:

- Dropout (rate = 0.5)
- Fully Connected Layer (512 neurons)
- ReLU activation
- Dropout (rate = 0.5)
- Fully Connected Layer (5 neurons for class scores)

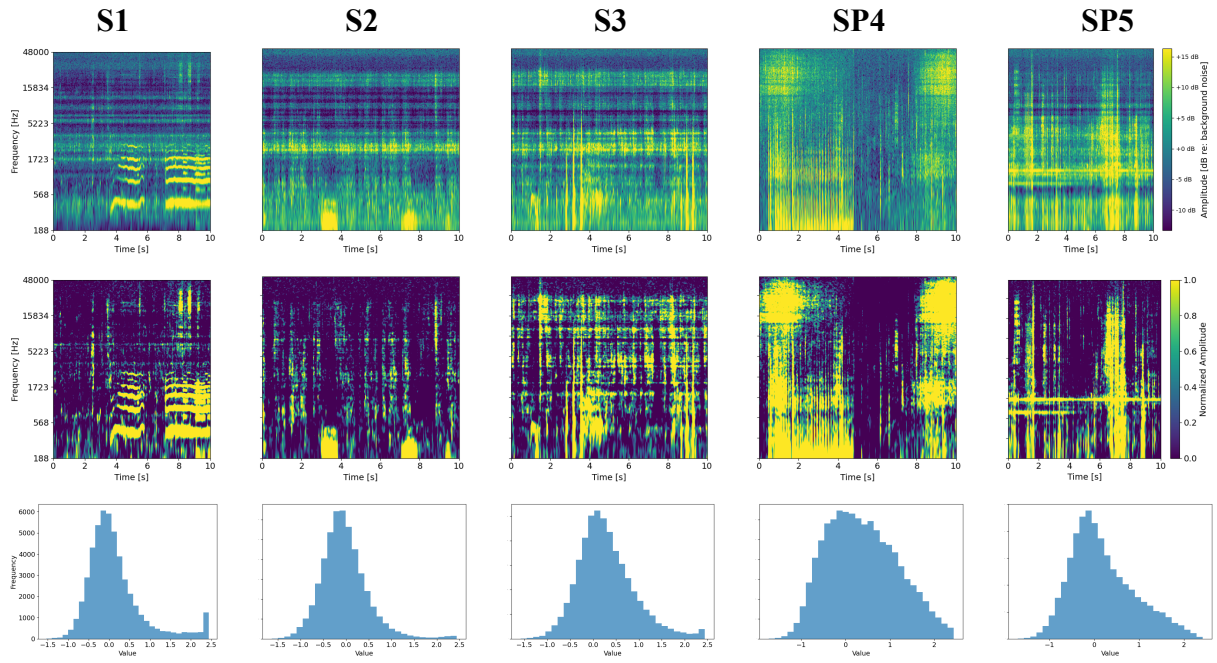


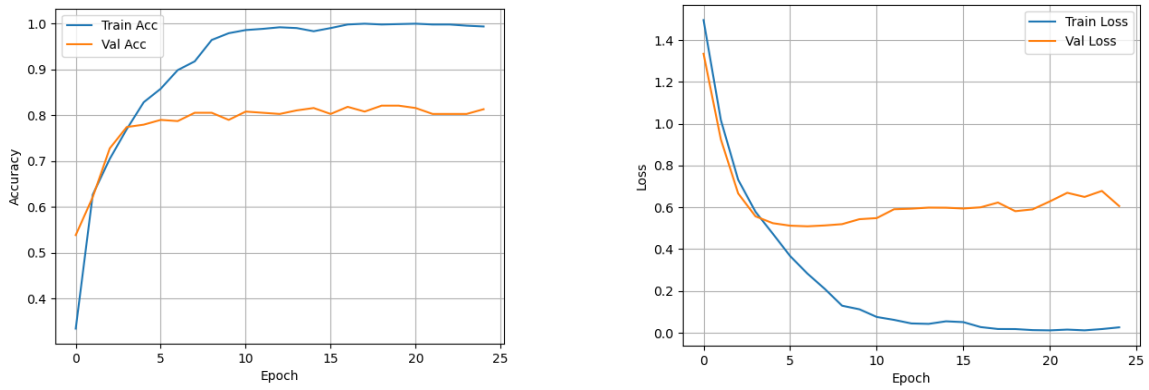
Figure 2: Examples of all five call types. Top row: Spectrograms after background subtraction. Middle row: Spectrograms after transformations. Bottom row: Histogram of intensities used as input to ResNet. Spectrograms are resampled to logarithmic frequency bins for visualization.

No output activation function was used, as PyTorch's CrossEntropyLoss implicitly applies softmax

The ResNet model was fine-tuned over 25 epochs. The data were split into training, evaluation and testing datasets with a ratio of 60:20:20, stratified according to the class densities. The CrossEntropyLoss function, weighted according to the class densities, was the criterion. Adam optimiser [21] was used with a learning rate of 1e-4. The model with the lowest validation loss, computed at each epoch, was selected. The initial layers were frozen to mitigate against overfitting, leaving only the last convolutional (Pytorch's "Layer 4") and fully connected layers trainable.

3. RESULTS

Fig. 3 present the model performance in terms of accuracy and loss over each epoch, and shows the model converging quickly before overfitting, despite freezing initial layers and high dropout rates (0.5) in the final layers. The final accuracy was 0.77. Fig. 4 presents the confusion matrix for the final model and reveals class S3 likely shares many similar features to class S1 and S2, evidenced by 42% of true S3 calls being misclassified as either S2 or S1. In contrast, only 4% of S3 calls are misclassified as SP4 and none classified as SP5. SP4 and SP5 are much more distinct with only 2 instances of SP4 misclassified (one each as S1 and S3) while all instances of SP5 were correctly identified. Fig 4. also highlights that S1 and S2 recall and precision are negatively impacted by the S3 class. It is clear from both Fig. 4 and Table 2 which show the model performance per-class that this model finds the S3 class significantly harder to model. The proposed SP4 and SP5 classes had the highest F-1 scores across all classes.



(a) Training and validation accuracy *(b) Training and validation loss*

Figure 3: Training curves for ResNet model.

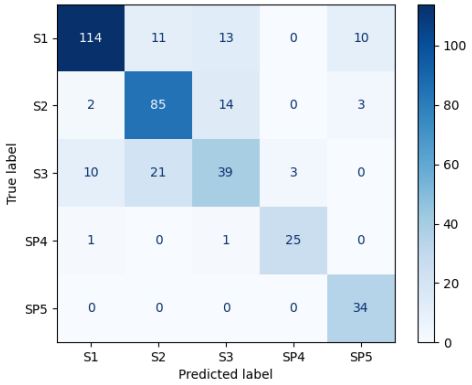


Figure 4: Test data confusion matrix

Class	Precision	Recall	F1-score	Support
S1	**0.90**	0.77	0.83	148
S2	0.73	0.82	0.77	104
S3	0.58	0.53	0.56	73
SP4	0.89	0.93	**0.91**	27
SP5	0.72	**1.00**	0.84	34
Accuracy		*0.77*		386
Macro avg	0.76	0.81	0.78	386
Weighted avg	0.78	0.77	0.77	386

Table 2: Classification report for fine-tuned ResNet

4. DISCUSSION AND CONCLUSION

As noted by Tadeo et al. [9], the S3 vocalisation category exhibits a high degree of variability, often cross-correlating with different vocalisation classes identified in previous studies. This variability complicates consistent classification and may reflect either ambiguity in our labelling or intrinsic heterogeneity within the S3 call type itself. It is therefore plausible that S3 represents a weaker clustering of vocal behaviour given it has the lowest F1-score of just 0.56 (next lowest is 0.77). Further work is needed to reassess the categorisation of these signals, particularly as we remain confident that the singular noise events originate from seals, despite their divergence from more stereotyped S3 structures. Given this variability, we acknowledge a lower level of confidence in the classification of S3 vocalisations during the labelling process. While some calls clearly match the acoustic profile of S3, others may represent contextually or functionally distinct call types. Additionally, observation conditions may have contributed to the ambiguity.

Despite these limitations, it is evident that the ResNet model performed significantly worse at classifying S3 vocalisations having both the lowest recall (0.58, next lowest 0.72) and precision (0.53, next lowest 0.77), compared to S1 and S2 as well as the proposed SP4 and SP5 classes. This supports the hypothesis that the S3 class represents a heterogeneous or poorly defined category, potentially encompassing multiple vocalisation types. In contrast, the higher classification performance observed for SP4 and SP5 suggests that these newly proposed categories

correspond to acoustically and structurally distinct vocalisation types that form well-separated clusters in the feature space. These findings underscore the need for further investigation into the acoustic properties and behavioural context of SP5 vocalisations. Additionally, future work should consider integrating supplementary classifications from existing literature to refine or subdivide the ambiguous S3 category and enhance the overall taxonomy of seal vocal behaviour.

5. ACKNOWLEDGEMENTS

The authors would like to thank Seiche Water Technology Group and the ART-AI CDT for funding and supporting William Butler's PhD.

REFERENCES

[1] D. Bowen, "Halichoerus grypus. the iucn red list of threatened species 2016: e. t9660a45226042," 2016.

[2] A. J. Hall and D. J. Russell, "Gray seal: Halichoerus grypus," in *Encyclopedia of marine mammals*. Elsevier, 2018, pp. 420–422.

[3] Joint Nature Conservation Committee, "Species interest features – special areas of conservation," n.d., accessed: 2025-05-29. [Online]. Available: https://sac.jncc.gov.uk/species/

[4] G. D. Hastie, P. Lepper, J. C. McKnight, R. Milne, D. J. Russell, and D. Thompson, "Acoustic risk balancing by marine mammals: anthropogenic noise can influence the foraging decisions by seals," *Journal of Applied Ecology*, vol. 58, no. 9, pp. 1854–1863, 2021.

[5] F. Chen, G. I. Shapiro, K. A. Bennett, S. N. Ingram, D. Thompson, C. Vincent, D. J. Russell, and C. B. Embling, "Shipping noise in a dynamic sea: a case study of grey seals in the celtic sea," *Marine Pollution Bulletin*, vol. 114, no. 1, pp. 372–383, 2017.

[6] S. Asselin, M. O. Hammill, and C. Barrette, "Underwater vocalizations of ice breeding grey seals," *Canadian Journal of Zoology*, vol. 71, no. 11, pp. 2211–2219, 1993.

[7] Y. P. Pozo Galván, M. Pérez Tadeo, M. Pommier, and J. O'Brien, "Static acoustic monitoring of harbour (phoca vitulina) and grey seals (halichoerus grypus) in the malin sea: A revolutionary approach in pinniped conservation," *Journal of Marine Science and Engineering*, vol. 12, no. 1, p. 118, 2024.

[8] L. J. Nowak, "Observations on mechanisms and phenomena underlying underwater and surface vocalisations of grey seals," *Bioacoustics*, vol. 30, no. 6, pp. 696–715, 2021.

[9] M. Pérez Tadeo, M. Gammell, and J. O'Brien, "First steps towards the automated detection of underwater vocalisations of grey seals (halichoerus grypus) in the blasket islands, southwest ireland," *Journal of Marine Science and Engineering*, vol. 11, no. 2, p. 351, 2023.

[10] D. Gillespie, M. Caillat, J. Gordon, and P. White, "Automatic detection and classification of odontocete whistles," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2427–2437, 2013.

[11] Y. Shiu, K. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Scientific reports*, vol. 10, no. 1, p. 607, 2020.

[12] M. Thomas, B. Martin, K. Kowarski, B. Gaudet, and S. Matwin, "Marine mammal species classification using convolutional neural networks and a novel acoustic representation," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2019, pp. 290–305.

[13] D. Duan, L.-g. Lü, Y. Jiang, Z. Liu, C. Yang, J. Guo, and X. Wang, "Real-time identification of marine mammal calls based on convolutional neural networks," *Applied Acoustics*, vol. 192, p. 108755, 2022.

[14] S. M. Van Parijs, P. J. Corkeron, J. Harvey, S. A. Hayes, D. K. Mellinger, P. A. Rouget, P. M. Thompson, M. Wahlberg, and K. M. Kovacs, "Patterns in the vocalizations of male harbor seals," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3403–3410, 2003.

[15] T. A. Lamont, L. Chapuis, B. Williams, S. Dines, T. Gridley, G. Frainer, J. Fearey, P. B. Maulana, M. E. Prasetya, J. Jompa *et al.*, "Hydromoth: Testing a prototype low-cost acoustic recorder for aquatic environments," *Remote Sensing in Ecology and Conservation*, vol. 8, no. 3, pp. 362–378, 2022.

[16] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, and A. Rogers, "Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment," *HardwareX*, vol. 6, p. e00073, 2019.

[17] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.

[18] 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.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] T. S. Prajwal and I. A K, "A comparative study of resnet-pretrained models for computernbsp;vision," in *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing*, ser. IC3-2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 419–425. [Online]. Available: https://doi.org/10.1145/3607947.3608042

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization. arxiv [preprint](2014)," *arXiv preprint arXiv:1412.6980*, vol. 5, 2014.