INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

ORCA-SLANG: An Automatic Multi-Stage Semi-Supervised Deep Learning Framework for Large-Scale Killer Whale Call Type Identification

Christian Bergler[1], *Manuel Schmitt*[1], *Andreas Maier*[1], *Helena Symonds*[2], *Paul Spong*[2], *Steven R. Ness*[3], *George Tzanetakis*[3], *Elmar Nöth*[1]

[1]Friedrich-Alexander-University Erlangen-Nuremberg, Pattern Recognition Lab, Erlangen, Germany
[2]OrcaLab, Alert Bay, British Columbia, Canada
[3]University of Victoria, Victoria, British Columbia, Canada

{christian.bergler, elmar.noeth}@fau.de

Abstract

Identification of animal-specific vocalization patterns is an imperative requirement to decode animal communication. In bioacoustics, passive acoustic recording setups are increasingly deployed to acquire large-scale datasets. Previous knowledge about established animal-specific call types is usually present due to historically conducted research. However, time- and human-resource constraints, combined with a lack of available machine-based approaches, only allow manual analysis of comparatively small data corpora and strongly distort the actual data representation and information value. Such data limitations cause restrictions in terms of identifying existing population-, group-, and individual-specific call types, subcategories, as well as unseen vocalization patterns. Thus, machine learning forms the basis for animal-specific call type recognition, to facilitate more profound insights into communication. The current study is the first fusing task-specific neural networks to develop a fully automated, multi-stage, deep-learning-based framework, entitled ORCA-SLANG, performing semi-supervised call type identification in one of the largest animal-specific bioacoustic archives – the Orchive. Orca/noise segmentation, denoising, and subsequent feature learning provide robust representations for semi-supervised clustering/classification. This results in a machine-annotated call type data repository containing 235,369 unique calls.

Index Terms: Killer Whale, Deep Learning, Call Type

1. Introduction

Large-scale animal-specific call type identification is fundamental for robust recognition of statistically significant recurring call patterns, in order to infer semantic and syntactic communication structures. Such language-like elements represent an essential prerequisite for animal language decoding. Therefor large acoustic datasets are imperative, wherefore bioacoustics makes extensive use of passive acoustic monitoring [1, 2]. One of the largest bioacoustic data repositories – the Orchive [3, 4, 5] – was acquired over 25 years (1985 – 2010) via a network of 6 hydrophones in northern British Columbia. The Orchive comprises approximately 20,000 h of underwater recordings, split into ≈45 min audio tapes. Since 50 years, killer whales (*Orcinus Orca*) have been studied in the coastal areas of the northeastern Pacific Ocean [3, 6, 7, 8, 9, 10]. Orca vocalizations are divided into three categories [8, 11]: (1) *Echolocation Clicks* – short pulsed events used for object localization and orientation [8, 11], (2) *Whistles* – non-pulsed single narrow band tones, between 1.5 – 18 kHz, with no or few harmonics, mainly utilized in close-range interactions lasting 50 ms to 12 s [8, 11], and (3) *Pulsed Calls* – most abundant and intensively studied vocal activity, split into discrete, variable, and aberrant calls, comprising a primary energy of 1 – 6 kHz up to 30 kHz or higher, together with a pulse repetition rate between 250 – 2,000 Hz and duration of 50 ms to 10 s [8, 11]. Discrete pulsed calls, also known as *Call Types*, are repetitive, stereotyped, and distinguishable vocalizations with large inter- and intra-class variability, possessing a wide diversity of diverse tonal properties [8, 11]. Figure 1 visualizes some excerpts of call types reported in [6, 8], describing the acoustic behavior of orcas.

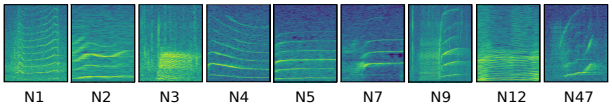

Figure 1: *Killer whale call type categories reported in [6]*

Large-scale, data-driven, and machine-based orca call type identification is imperative to gain deeper insights into killer whale communication. A large data corpus of call types will allow new studies on dialects, animal-ID, unknown vocalization patterns, and other interesting scientific questions. However, this scenario results in two major challenges: (1) massive, noise-heavy data volumes require robust, task-specific machine learning approaches, to prepare, enhance, and process the audio material before conducting call type identification, and (2) lack of sufficiently-large, human-labeled, and representative call type data, combined with missing information about intra call type variations, as well as inter call type differences. The current study introduces a fully automated, multi-stage, semi-supervised, deep-learning-based orca call type identification pipeline, named ORCA-SLANG, processing the entire Orchive, designed for large-scale recognition of already known call types besides the detection of possible sub-call patterns and/or unlabeled vocalization categories. It combines the following individual steps (see Figure 2): (1) orca sound type versus background noise segmentation, (2) signal enhancement by denoising pre-segmented orca vocalizations, (3) deep feature learning on these signals, and (4) hybrid semi-supervised call type identification, including (4.1) unsupervised clustering to detect and group known call types, new sub-call categories, as well as unseen orca vocalization patterns, and (4.2) supervised, multi-class, call type classification with respect to the cluster hypothesis, utilizing a small human-labeled call type dataset, to detect clusters of known call types by measuring cluster purity in order to build the machine-annotated data foundation for k-Nearest Neighbor (k-NN) [17] classification of remaining calls, while confirming feature and clustering quality.

http://dx.doi.org/10.21437/Interspeech.2021-616

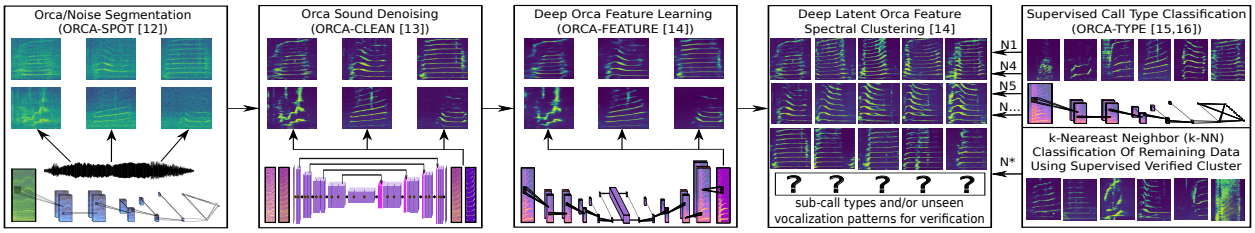

Figure 2: *ORCA-SLANG: multi-stage, semi-supervised, deep-learning-based, and large-scale killer whale call type recognition fusing: (1) orca/noise segmentation [12], (2) denoising [13], (3) deep feature learning [14], (4) call type clustering/classification [14, 15, 16]*

2. Related Work

Passive acoustic monitoring, mainly caused by decreasing costs for recording equipment [1, 2], allows bioacoustics access to very large data archives [3]. In recent years, machine learning has gradually been applied to a variety of bioacoustic research topics. In particular, time- and resource-critical, cumbersome, and labor-intensive procedures, concerning a majority of such fundamental bioacoustic signal identification scenarios, need to be automated. Next to machine-learning-based bioacoustic segmentation/detection [3, 18, 19, 20, 21] between various animal-specific vocalizations and environmental noise, several animal species [22, 23, 24] and/or call-type [3, 20, 25, 21] classification studies, have been conducted. Besides pure supervised scenarios, animal-specific semi-supervised [26, 22, 27, 23, 24] and unsupervised [28, 29, 30, 31] concepts for species and/or call type identification were additionally explored. In comparison, less work has been conducted on bioacoustic signal enhancement (denoising) [32, 33, 34], primarily caused due to a lack of cleaned/denoised ground truth samples. To the best of the authors' knowledge, this is the first study presenting a bioacoustic pipeline for large-scale, fully automated, orca call type identification by fusing previous task-specific approaches: (1) orca/noise segmentation [12], (2) denoising [13], (3) deep feature learning [14], and (4) semi-supervised call type identification clustering/classification [14, 15, 16].

3. Methodology

ORCA-SPOT – In our previous work, a ResNet18-based segmentation model, entitled ORCA-SPOT [12] (see Figure 2), was designed to robustly distinguish between orca sounds and background noise within noisy underwater recordings. ORCA-SPOT was trained/tested in a supervised manner, based on 17,104 orca and 44,323 noise excerpts, transformed into 256×128 (256 frequency bins, covering 500 Hz to 10 kHz and 128 time frames) pre-processed, augmented, and 0/1-dB-normalized power-spectrograms. Besides an accuracy of 95.0 % on unseen pre-extracted test samples, it obtained a precision of 93.2 % on 238 randomly chosen unseen Orchive tapes (191.5 h), and 95.2 % Area-Under-The-ROC-Curve (AUC) on 9 fully annotated unseen Orchive tapes [12].

ORCA-CLEAN – Besides pure environmental noise (filtered by ORCA-SPOT), there also exist complex overlaying noises within the orca vocalizations. In order to separate such superimposed noise characteristics, while simultaneously counteracting the lack of clean ground truth samples, a U-Net-based deep denoising network was invented, named ORCA-CLEAN [13] (see Figure 2), to enhance noisy orca vocalizations. The model incorporates a hybrid training algorithm, providing a pool of additive noise variants (synthetic and real-world), together with machine-generated binary (orca/noise) mask alternatives, acting as network attention mechanism. ORCA-CLEAN was trained on 256×128-large spectral input/output pairs, where the input was always noisier than the output [13]: (1) randomly chosen additive noise variants to further distort the original noisy sample (input) compared to the original file (output), and (2) noisy original (input) versus random selected machine-generated orca/noise mask alternatives (output).

ORCA-FEATURE – In order to learn and derive robust feature embeddings from distinct orca sound types, a ResNet18-based convolutional undercomplete autoencoder was trained, entitled ORCA-FEATURE (see Figure 2), initially presented in [14, 16]. Similar to ORCA-SPOT and ORCA-CLEAN, ORCA-FEATURE covers the same frequency range and also uses a 256×128-large input spectrogram [14, 16]. Whereas the data clustered in [14], incorporated a small, Orchive-based, human-selected, and noisy orca call type corpus, comprising comparatively distinct and clearly distinguishable orca call patterns, large-scale and machine-based orca segmentations of the entire Orchive are significantly less differentiable due to strong call/noise variations and higher noise levels. Most of the spectral content observed in underwater orca vocalizations belongs to envrionmental noise. Consequently, major parts of inferred features often characterize background noise, rather than focusing on the actual spectral orca call structure, which in turn strongly affects subsequent unsupervised clustering approaches. Therefore, the denoised output of ORCA-CLEAN was directly utilized as input for ORCA-FEATURE. Moreover, experimental investigations resulted in two architectural modifications compared to the version reported in [14, 16]: (1) dropout layer after each residual layer in the encoder/decoder (compression/decompression) path, to avoid occasionally observed overfitting during training, and (2) a modified bottleneck layer. In [14], the bottleneck layer contained two convolutional layers (1×1, stride 1), to compress/decompress the 512×16×8-large encoder output into 4×16×8 (1×512 latent features) and back to the 512×16×8 decoder input. However, in this study, the 512×16×8 encoder output was compressed to 512 features by a convolutional layer (5×5, stride 1) and a subsequent max-pooling (16×8). The 512×16×8 decoder input was generated via max-unpooling (16×8) combined with a transposed convolutional layer (5×5, stride 1), to decompress the 512 latent features. A comparatively large kernel size of 5×5 was chosen to keep contextual information in each of the 512 feature maps, encoded by subsequent max-pooling into a final 1×512-large latent feature vector, representing the clustering input.

ORCA-TYPE – Supervised multi-class orca call type classification is very difficult due to a lack of human-labeled data, combined with a very limited extent of covered, known call types. In our previous work [15, 16], a ResNet18-based 12-class deep neural network, named ORCA-TYPE (see Figure 2), was introduced. ORCA-TYPE was trained on 256×128-large

samples of the Call Type Data Corpus (CTDC), a small human-labeled data archive, consisting of two catalogs with 514 audio excerpts in total, distributed across 9 orca call types, echolocations, whistles, and noise (12 classes) [16]. ORCA-TYPE achieved the following results on the unseen CTDC test set: (1) supervised training on the original CTDC dataset (mean 12-class test accuracy 87 %) [15], (2) semi-supervised training by conducting deep representation learning on numerous machine-segmented orca sounds, combined with downstream supervised training/fine-tuning, utilizing the original CTDC data (mean 12-class test accuracy 94 %, best model 96 %) [16], and (3) supervised training on the denoised CTDC dataset by applying ORCA-CLEAN as additional data preprocessing/enhancement step [13] (mean 12-class test accuracy 95 %, best model 98 %). In this study, the best ORCA-TYPE model, trained on denoised signals (case 3), was utilized to verify unsupervised results.

Spectral clustering and k-NN – Bottleneck features derived by ORCA-FEATURE, provide the input for unsupervised orca call type identification. Spectral clustering [35] was applied using a radial basis function in order to build the affinity matrix. Furthermore, k-NN was performed to assign remaining unclustered data samples (1×512 features) to the respective call categories.

4. Data Material and Preprocessing

Orchive – Seg90 (OSEG90) – Significant call type analysis requires large-scale data material. Thus, ORCA-SPOT was utilized to segment the entire Orchive in about 8 days first, by performing a sliding window approach (window-size = 2 s, step-size = 0.5 s, confidence > 0.90). Only ORCA-SPOT predictions above this confidence level were considered as valid vocalization frames. Given this constraint, ORCA-SPOT identified 2,521,078 orca segments of various durations. The resulting data repository, called Orchive-Seg90 (OSEG90), including $\approx$ 2,992.02 h of machine-detected orca vocalizations, leading to an average duration per segment of $\approx$ 4.2 s, which in turn reduced the entire Orchive to $\approx$ 15 %. However, ORCA-SPOT, ORCA-CLEAN, ORCA-FEATURE, and ORCA-TYPE require a fixed, uniformly large temporal context. According to [11], pulsed calls exhibit durations spanning from less than 50 ms to $>$10 s, although the majority is between 0.5 and 1.5 s. As in our previous studies [12, 13, 14, 15, 16], a temporal context of 1.28 s was chosen, together with the cross-module FFT-settings (window-size = 4,096, hop-size = 441, sampling rate = 44,1 kHz) resulted in 128 time frames. An orca signal detection algorithm was invented in [13], to guarantee (1) equal temporal context, (2) isolated calls, and (3) reduction of superfluous preceding, succeeding, and intermediate noises, besides considering the average orca vocalization duration of pulsed calls, reported in Ford et al. [11]. Specifically, the orca detection algorithm models the spectral intensities of a preprocessed power spectrogram as a function of time, by summing up spectral intensities for each time step via a sliding window approach (window-size = 1.28 s, step-size = 100 ms) [13]. The window containing the global function maximum was extracted, resulting in an F$\times$128 spectrogram, where F is the number of frequency bins. Potential zero-padding was never required due to the minimum segmentation window-size of 2 s.

Automated Orca Feature Corpus (AOFC) – Deep orca feature learning (training ORCA-FEATURE) was conducted on a machine-annotated (ORCA-SPOT, window-size = 2 s, step-size = 0.5 s, confidence > 0.9999) dataset – the Automated Orca Feature Corpus (AOFC). The entire data archive includes 1,969 representative machine-labeled excerpts (training – 1,378 samples, 70.0 %, validation – 293 samples, 14.9 %, test – 298 samples, 15.1 %) from various Orchive tapes spread over all years, summing up to a total duration of $\approx$ 1.65 h, resulting in an average sample duration of $\approx$ 3.0 s. Similarly to OSEG90, 1.28s-large sections were obtained during training by means of the previously described orca detection algorithm (see also [13]).

5. Experiments

The experimental setup of ORCA-SLANG (see Figure 2) can be described by a 4-step sequentially ordered procedure, to semi-supervised identify known/unknown orca call types within the entire Orchive: (1) ORCA-SPOT to segment 20,000 h of Orchive recordings, resulting in the OSEG90 repository, (2) embedding ORCA-CLEAN as enhancement step for deep feature learning and call type classification, (3) train ORCA-FEATURE (batch-size = 4, Adam optimizer – learning rate = 10^{-5}, $\beta_1 = 0.5$, $\beta_2 = 0.999$, Mean Squared Error (MSE) loss), on the AOFC corpus to learn compact/representative bottleneck features, while using identical data preprocessing steps and network parameters as presented in [14], except the fact that data augmentation was disabled, and (4) unsupervised spectral call type clustering as well as supervised classification, applying the ORCA-TYPE model of [13], besides k-NN, all together inside a hybrid identification scenario. Step 4, described in detail hereafter, was repeated 5 times, where the inter-trial call type duplicates were removed, to determine the actual number of unique and additional detected known call types across all 5 trials. Step 4, involves the following single steps: spectral clustering on a random selection of 20,000 ORCA-SPOT segments (confidence > 0.9999), taken from the 2,521,078 samples of the OSEG90 archive with 200 clusters (1 %). Afterwards, ORCA-TYPE was applied to classify the content of all clusters in order to find clusters which belong to one of the 9 trained call types of ORCA-TYPE (data enlargement). All elements of a cluster were assigned to one of the nine call type categories, if more than 70 % (cluster purity) of the cluster was classified as a certain call type. Only clusters fulfilling this purity constraint were call-type-specifically grouped, summarized, and consequently built the machine-annotated data reference of call type prototypes for subsequent k-NN classification (k = 5) of all remaining 2,501,078 data excerpts. Clusters which did not satisfy the purity criterion of > 0.7 were mapped to a rejection class. The hybrid approach between spectral clustering and downstream k-NN classification is significantly more efficient than clustering the entire OSEG90 archive in terms of computational overhead, data storage, manageability of cluster number, and performance. Finally, ORCA-TYPE was again applied, but now to the k-NN-extended call type-specific data repositories, to evaluate purity, prove clustering quality, and verify general approach applicability.

6. Results and Discussion

ORCA-SLANG, in combination with the 2,521,078-large OSEG90 data corpus, enables bioacoustics to gain completely new and deeper insights into orca communication by providing identification results on: (1) known call types, (2) potential sub-call types, (3) and unseen vocalization patterns. Table 1 presents call types plus respective findings while applying ORCA-SLANG (see Figure 2). Considering the cluster purity of > 0.7 ($C_{p>0.7}$), across all 5 clustering trials (20,000 random samples, 200 cluster each), 6 out of 9 call types, known to ORCA-TYPE, were identified, summing up to 7,637 samples.

Table 1: *Call type findings of 6 known categories, across all 5 trials, k-NN-based call assignments, verified by ORCA-TYPE (accuracy), and amount of unique identified call types*

Types \ Mode	$C_{p>0.7}$	k-NN	ORCA-TYPE		Unique
	samples	samples	samples	accuracy[%]	samples
N1	742	64,124	60,281	94.0	19,280
N3	208	9,273	7,023	75.7	8,484
N4	4,854	424,727	387,406	91.2	145,760
N5	233	13,826	10,809	78.2	8,861
N9	1,324	96,661	84,366	87.3	41,990
N47	276	17,070	13,725	80.4	10,994
$\sum$	**7,637**	**625,681**	**563,610**	**90.0**	**235,369**

Subsequent k-NN- and trial-based call type assignment of the remaining 2,501,078 excerpts lead to 625,681 total findings, whereas 563,610 were correctly classified as either one of the 6 call types, resulting in an overall multi-class classification accuracy of 90.0 %. This in turn is a solid proof concerning the quality of learned/derived features, initial clustering, as well as general feasibility of the entire semi-supervised ORCA-SLANG pipeline. The frequency distribution of call type specific observations also coincides with the reported distribution of frequent and less occurring vocalization types in [3]. Removing these 6 call types from the entire OSEG90 archive, leads to a change in the prior probabilities, and thus also restructures the high-dimensional feature space, whereby other, rare call types will receive more weight which in turn allows better clustering. Figure 3 visualizes denoised call types for each of the 6 call categories, semi-supervised identified by ORCA-SLANG.

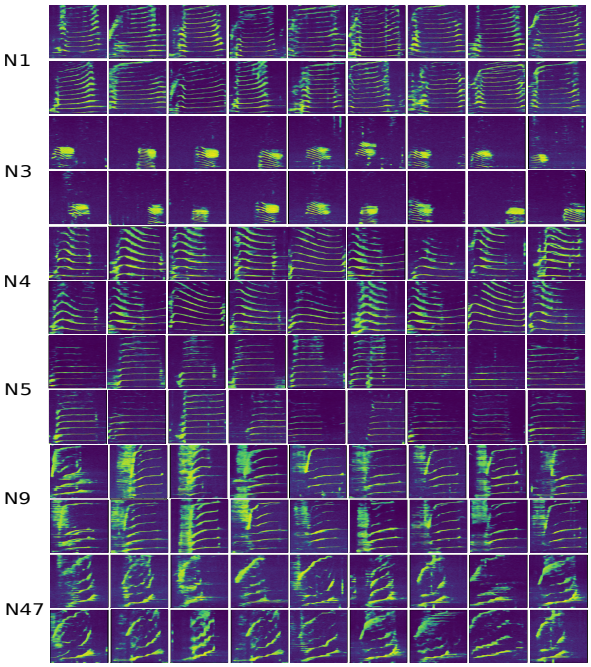

Figure 3: *Large-scale known orca call type [6] identification*

Besides results on already known call types, ORCA-SLANG also provides the possibility to identify potential unseen subtypes. As an example, the 145,760 unique N4 call types, listed in Table 1, were fully-unsupervised clustered. Figure 4 visualizes examples of various potential sub-call types, each of them taken from a distinct cluster, showing the strong spectral intra call type variety.

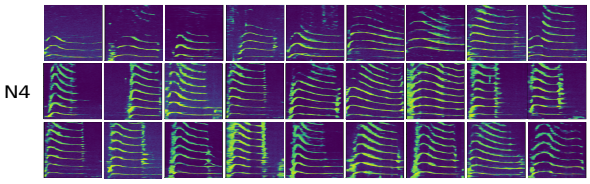

Figure 4: *Potential sub-call type hypothesis of N4*

In addition, very interesting observations were made in terms of unseen call types. Figure 5 shows one example of such a call type, apparently also containing various sub-structures, identified by ORCA-SLANG, primarily occurring in a strongly concentrated manner across 7 tapes in the year 2000. In this context, the handwritten transcripts of the Orcalab (OrcaLab-Books [5]) were consulted. The OrcaLab-Books were written by different domain experts and scientists to document additional information such as communication and behavioral patterns, while passively recording. Figure 5 visualizes exemplarily two OrcaLab-Book pages of the corresponding Orchive tapes, 15A and 15B from the year 2000, out of which the respective call excerpts originate. According to the handwritten documentations, these kind of vocal activities were produced by transients, one of the three orca populations present in the coastal areas of the northeastern Pacific Ocean. However, while most of the Orchive is based on resident orca vocalization (N-calls, S-calls [6]), ORCA-SLANG detected previously unlabeled, rare transient vocalizations in significantly large quantities.

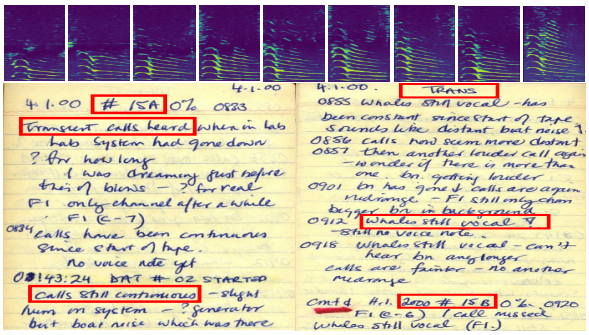

Figure 5: *Transient killer whale call type recognition*

7. Conclusion and Future Work

In this study we present ORCA-SLANG, a multi-stage, deep-learning framework, unifying (1) orca/noise segmentation [12], (2) denoising [13], (3) deep feature learning [14], and (4) call type identification [14, 15, 16] in a machine-based, semi-supervised manner, to identify known and unknown killer whale call types within 20,000 h of underwater recordings. ORCA-SLANG allows biologists to access large-scale pre-processed data archives with respect to known orca call patterns, sub-call types, and unseen vocal activities, enabling totally new insights into population-, group-, and individual-related acoustic analysis, which in turn is the foundation for a revised/updated version of the existing call type catalog [6]. In our future studies we will utilize known/unknown call type findings to identify semantic/syntactic language-like patterns within the entire Orchive, which will then be correlated with the OrcaLab-Book [4, 5]. Moreover, ORCA-SLANG with all individual components will be generalized to be animal independent. The source code of ORCA-SLANG will be publicly available under [36].

8. References

[1] E. Browning, R. Gibb, P. Glover-Kapfer, and K. E. Jones, "Passive acoustic monitoring in ecology and conservation," WWF-UK, Tech. Rep., 10 2017.

[2] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones, "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring," *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 169–185, 2019.

[3] S. R. Ness, "The Orchive : A system for semi-automatic annotation and analysis of a large collection of bioacoustic recordings," Ph.D. dissertation, Department of Computer Science, University of Victoria, British Columbia, Canada, 2013.

[4] ORCALAB, "Orcalab - A whale research station on Hanson Island," http://orcalab.org (April 2021).

[5] S. R. Ness, "Orchive," http://orchive.cs.uvic.ca/ (April 2021).

[6] J. K. B. Ford, "A catalogue of underwater calls produced by killer whales (Orcinus orca) in British Columbia," *Canadian Data Report of Fisheries and Aquatic Science*, no. 633, p. 165, 1987.

[7] M. A. Bigg, P. F. Olesiuk, G. M. Ellis, J. K. B. Ford, and K. C. Balcomb, "Organization and genealogy of resident killer whales (Orcinus orca) in the coastal waters of British Columbia and Washington State," *International Whaling Commission*, pp. 383–405, 1990.

[8] J. K. B. Ford, "Vocal traditions among resident killer whales (Orcinus orca) in coastal waters of British Columbia," *Canadian Journal of Zoology*, vol. 69, pp. 1454–1483, June 1991.

[9] J. Ford, G. Ellis, and K. Balcomb, *Killer whales: The natural history and genealogy of Orcinus orca in British Columbia and Washington*. UBC Press, 2000.

[10] J. Towers, G. Sutton, T. Shaw, M. Malleson, D. Matkin, B. Gisborne, J. Forde, D. Ellifrit, G. Ellis, J. Ford *et al.*, "Photo-identification catalogue, population status, and distribution of bigg's killer whales known from coastal waters of british columbia, canada," *Fisheries and Oceans Canada, Pacific Biological Station, Nanaimo, BC*, 2019.

[11] J. K. B. Ford, "Acoustic behaviour of resident killer whales (Orcinus orca) off Vancouver Island, British Columbia," *Canadian Journal of Zoology*, vol. 67, pp. 727–745, January 1989.

[12] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer, and A. Maier, "Orca-spot: An automatic killer whale sound detection toolkit using deep learning," *Scientific Reports*, vol. 9, 12 2019.

[13] C. Bergler, M. Schmitt, A. Maier, S. Smeele, V. Barth, and E. Nöth, "ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication," in *Proc. Interspeech 2020*, 2020, pp. 1136–1140.

[14] C. Bergler, M. Schmitt, R. X. Cheng, A. Maier, V. Barth, and E. Nöth, "Deep Learning for Orca Call Type Identification – A Fully Unsupervised Approach," in *Proc. Interspeech 2019*, 2019, pp. 3357–3361.

[15] H. Schröter, E. Nöth, A. Maier, R. Cheng, V. Barth, and C. Bergler, "Segmentation, Classification, and Visualization of Orca Calls Using Deep Learning," in *International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*. IEEE, May 2019, pp. 8231–8235.

[16] C. Bergler, M. Schmitt, R. X. Cheng, H. Schröter, A. Maier, V. Barth, M. Weber, and E. Nöth, "Deep Representation Learning for Orca Call Type Classification," in *Proc. Text, Speech, and Dialogue 2019*, vol. 11697 LNAI. Springer, 2019, pp. 274–286.

[17] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, pp. 21–27, 1967.

[18] N. Priyadarshani, S. Marsland, and I. Castro, "Automated birdsong recognition in complex acoustic environments: a review," *Journal of Avian Biology*, vol. 49, no. 5, pp. jav–01 447, 2018.

[19] Y. Shiu, K. Palmer, M. 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, p. 607, 01 2020.

[20] P. Bermant, M. Bronstein, R. Wood, S. Gero, and D. Gruber, "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," *Scientific Reports*, vol. 9, pp. 1–10, 08 2019.

[21] D.-H. Jung, N. Y. Kim, S. H. Moon, C. Jhin, H.-J. Kim, J.-S. Yang, H. S. Kim, T. S. Lee, J. Y. Lee, and S. H. Park, "Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering," *Animals*, vol. 11, no. 2, 2021.

[22] M. Zhong, J. LeBien, M. Campos-Cerqueira, R. Dodhia, J. M. Lavista Ferres, J. P.Velev, and T. M. Aide, "Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling," *Applied Acoustics*, vol. 166, September 2020.

[23] L. Zhang, D. Wang, C. Bao, Y. Wang, and K. Xu, "Large-scale whale-call classification by transfer learning on multi-scale waveforms and time-frequency features," *Applied Sciences*, vol. 9, p. 1020, 03 2019.

[24] M. Thomas, B. Martin, K. Kowarski, and B. Gaudet, *Marine Mammal Species Classification Using Convolutional Neural Networks and a Novel Acoustic Representation*, ser. LNCS. Springer, 4 2020, vol. 11908, pp. 290–305.

[25] S. Madhusudhana, Y. Shiu, H. Klinck, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, A. Širović, and M. A. Roch, "Temporal context improves automatic recognition of call sequences in soundscape data," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2442–2442, 2020.

[26] J. Salamon, J. P. Bello, A. Farnsworth, M. Robbins, S. Keen, H. Klinck, and S. Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PLOS ONE*, vol. 11, pp. 1–26, November 2016.

[27] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ-the Journal of Life and Environmental Sciences*, vol. 488, p. 24, July 2014.

[28] J. Brown, A. Hodgins-Davis, and P. Miller, "Classification of vocalizations of killer whales using dynamic time warping," *JASA Express Letters*, vol. 119, no. 3, pp. 617–628, March 2006.

[29] G. Picot, O. Adam, M. Bergounioux, H. Glotin, and F.-X. Mayer, "Automatic prosodic clustering of humpback whales song," in *New Trends for Environmental Monitoring Using Passive Systems*, November 2008.

[30] P. Rickwood and A. Taylor, "Methods for automatically analyzing humpback song units," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1763–1772, January 2008.

[31] T. Sainburg, M. Thielk, and T. Q. Gentner, "Latent space visualization, characterization, and generation of diverse vocal communication signals," *bioRxiv*, 2020. [Online]. Available: https://www.biorxiv.org/content/early/2020/01/28/870311

[32] R. Sinha and P. Rajan, "A deep autoencoder approach to bird call enhancement," in *2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS)*, 2018, pp. 22–26.

[33] J. Castro and E. Meneses, "Parallelization of a denoising algorithm for tonal bioacoustic signals using openacc directives," in *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018, pp. 1–8.

[34] N. Hassan and D. Ramli, "A comparative study of blind source separation for bioacoustics sounds based on fastica, pca and nmf," *Procedia Computer Science*, vol. 126, pp. 363–372, 01 2018.

[35] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[36] C. Bergler, "Open Source Github-Repository Christian Bergler," https://github.com/ChristianBergler (2021).