ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication

Christian Bergler[1], *Manuel Schmitt*[1], *Andreas Maier*[1], *Simeon Smeele*[2], *Volker Barth*[3], *Elmar Nöth*[1]

[1]Friedrich-Alexander-University Erlangen-Nuremberg, Pattern Recognition Lab, Erlangen, Germany
[2]Max Planck Institute of Animal Behavior, Cognitive and Cultural Ecology Lab &
Max Planck Institute for Evolutionary Anthropology, Department for Human Behavior, Ecology and Culture, Radolfzell & Leipzig, Germany
[3]Anthro-Media, Berlin, Germany

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

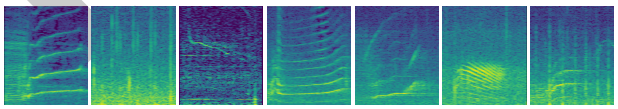
Abstract

In bioacoustics, passive acoustic monitoring of animals living in the wild, both on land and underwater, leads to large data archives characterized by a strong imbalance between recorded animal sounds and ambient noises. Bioacoustic datasets suffer extremely from such large noise-variety, caused by a multitude of external influences and changing environmental conditions over years. This leads to significant deficiencies/problems concerning the analysis and interpretation of animal vocalizations by biologists and machine-learning algorithms. To counteract such huge noise diversity, it is essential to develop a denoising procedure enabling automated, efficient, and robust data enhancement. However, a fundamental problem is the lack of clean/denoised ground-truth samples. The current work is the first presenting a fully-automated deep denoising approach for bioacoustics, not requiring any clean ground-truth, together with one of the largest data archives recorded on killer whales (*Orcinus Orca*) – the Orchive. Therefor, an approach, originally developed for image restoration, known as Noise2Noise (N2N), was transferred to the field of bioacoustics, and extended by using automatic machine-generated binary masks as additional network attention mechanism. Besides a significant cross-domain signal enhancement, our previous results regarding supervised orca/noise segmentation and orca call type identification were outperformed by applying ORCA-CLEAN as additional data preprocessing/enhancement step.

Index Terms: Killer Whale, Denoising, Call Type, Deep Learning, Orca

1. Introduction

Passive audiovisual monitoring techniques are often used to obtain more natural information and insights into the communication and behavior of different animal species [1, 2]. Thus, the natural animal habitat is affected as little as possible by external influences, which in turn significantly increase the chance of observing all natural and rarely occurring communication and behavioral patterns in sufficient quantity (observer's paradox) [3]. Large and noise-heavy data archives [1, 4] are the result of such passive observations. For more than 40 years, the killer whale *(Orcinus Orca)*, the largest member of the dolphin family, has been recorded and studied in the coastal areas of the northeastern Pacific Ocean [1, 5, 6, 7, 8, 9]. The Orchive [1, 4, 9], one of the largest bioacoustic data archives collected over 25 years (1985-2010) on a single animal species – the killer whale *(Orcinus Orca)* – contains about ≈20,000 h of underwater recordings captured via stationary hydrophones located in northern British Columbia (Hanson Island). A large part of these data is noise in various forms (underwater noise, boat noise, microphone artifacts, etc.), whereas also the noise characteristics shift over the years (e.g. increasing touristic boat traffic). In general such noise can be divided into two types: (1) pure noise which does not contain any killer whale communication, and (2) noise that overlays killer whale vocalization. Killer whales produce three different types of vocal activities [8]: (1) *Echolocation Clicks* (short pulses with variable duration utilized for navigation and localization [8]), (2) *Whistles* (narrow band tones between 1.5 and 18 kHz having no or less harmonic structures mostly used in close-range social interactions [8]), and (3) *Pulsed Calls* (repetitive, stereotyped and distinct tonal properties with a primary energy between 1-6 kHz showing sudden and patterned shifts in frequency [8]). Pulsed calls are separated into discrete, variable, and aberrant calls [8]. Figure 1 visualizes all three different types of orca sounds (echolocation clicks, whistles, and various discrete call types) within diverse noisy underwater conditions.



N09 Echo Whistle N01 N47 N03 N07+Echo

Figure 1: *Orca vocalizations within various noise conditions*

ORCA-SPOT [10] was utilized to counteract the above mentioned noise-type 1, by segmenting the entire ≈20,000 h of Orchive underwater recordings in order to separate large parts of pure and surrounding environmental noise from relevant orca signals. Nevertheless, the resulting machine-segmented killer whale extractions still contain various and complex overlaying noises (noise-type 2) (see Figure 1) negatively affecting human interpretation possibilities and machine-based feature learning. Therefore, it is imperative to develop an automatic and robust deep denoising procedure to remove superimposed and overlaying noise structures. However, this lead to a number of challenges and problems: (1) huge and complex noise variety, (2) large and unknown orca call type variety, and (3) no clean/denoised killer whale signals as ground-truth. In this study we present a fully-automated deep denoising approach not requiring any denoised/clean labeled ground-truth signals. The developed algorithm is based on the Noise2Noise [11] approach, originally designed for image restoration. We have transferred the idea of Noise2Noise [11] to the field of bioacoustics, by corrupting noisy original spectrograms via different additive noise variants. The distorted orca spectrograms, together with a deep denoising network, were utilized to reconstruct the original noisy orca vocalizations. Furthermore, we have extended the traditional Noise2Noise [11] approach by using machine-generated binary masks (1 = orca, 0 = noise) as clean ground-truth, acting as an additional attention mechanism for the network. In summary, our deep denoising network randomly selects from a collection consisting of various additive noise alternatives and different orca/noise mask manifestations.

2. Related Work

In recent years, deep learning has had a strong impact on image denoising [12, 13, 14], image restoration [11, 15, 16], speech enhancement [17, 18, 19, 20], and speech separation [21, 22, 23]. In bioacoustics, however, those fields of research in connection with machine learning, especially deep learning, are comparatively less explored. The work of Priyadarshani et al. [24] provides a literature review of various approaches using traditional noise reduction methods (wavelet denoising, various filtering techniques, spectral substraction, median clipping, etc.) for birdsongs. Brown et al. [25] examined the Minimum Mean Square Error Short Time Spectral Amplitude (MMSE STSA) algorithm to remove environmental noise from bird recordings. Hassan et al. [26] addressed the problem of Blind Source Separation (BSS) in mixed and noisy bioacoustic signals (frogs) utilizing various source separation approaches. Castro et al. [27] proposed a multi-stage denoising algorithm containing wavelet transformation, various signal processing techniques, and k-means clustering for West Indian manatees. Sinha et al. [28] illustrated a deep learning based approach for bird call enhancement using a deep denoising autoencoder resulting in a significant improvement of the subsequent DNN-based bird species classification. To the best of our knowledge, there is no study transferring the image-based Noise2Noise [11] concept to the field of bioacoustics, and additionally integrating an attention mechanism by using machine-generated spectral binary masks in various alternatives to denoise and enhance orca signals, in order to improve feature learning for our subsequent orca/noise segmentation [10] and orca call-type classification [29, 30].

3. Data Material

Automated Orca Extraction Corpus (AOEC) – In order to train our deep denoising network a machine-segmented data archive – the Automated Orca Extraction Corpus (AOEC) – was utilized. ORCA-SPOT [10] was used in combination with the Orchive [1, 4, 9] to identify killer whale vocalization events. In total, the AOEC dataset contains 31,151 machine-annotated killer whale segments having a variable duration of $d \in [1.28, 2.00]$ seconds. The AEOC archive includes machine-labeled [10] killer whale content from 11,907 randomly chosen $\approx$ 45-minute Orchive tapes, spread over a time period from 1985 to 2010. Overall the AEOC dataset comprises a total duration of $\approx$13.53 hours resulting in an average duration of $\approx$1.56 seconds per orca segment. AOEC has been split into a training (21,806 samples, 70.0 %), validation (4,673 samples, 15.0 %), and test set (4,672 samples, 15.0 %), utilized for training and evaluating ORCA-CLEAN. We have ensured that killer whale segments extracted from the same Orchive tape are only present in one of the three sets. Moreover, the AOEC dataset contains data from Orchive tapes not being part and/or in conflict with any of our other and previously used data corpora [10, 29, 30].

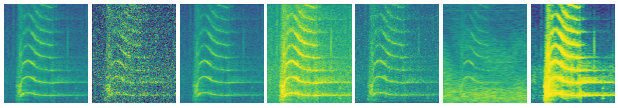
Call Type Data – In order to ensure comparability to our previous works [29, 30] we used the Call Type Data Corpus to verify our proposed deep denoising approach. As described in [29] and [30], the corpus consists of three subsets including (1) Orcalab Call Type Catalog (CCS), (2) Ness Call Type Catalog, and (3) Extension Catalog (EXT). The CCS data corpus contains 7 different call types summing up to a total amount of 138 audio samples distributed as follows [30]: 33 N01, 10 N02, 21 N04, 14 N05, 18 N07, 26 N09, and 16 N12. The CCN dataset includes 286 call type examples spread across 6 classes: 36 N01, 56 N03, 60 N04, 31 N07, 70 N09, and 33 N47. The EXT catalog comprises 90 additional audio samples split into 3 classes: 30 echolocation clicks, 30 whistles, and 30 noises. Consequently, the entire corpus consists of 514 audio files divided into 12 different classes [30]. Compared to our previous work in [29, 30], an identical data distribution was used (Training – 363 samples, 70.6 %, Validation – 72 samples, 14.0 %, Test – 79 samples, 15.4 %).

4. Methodology

Data Preprocessing – Compared to previous works [10, 29, 30] our multi-level preprocessing pipeline includes the following components: (1) creating a mono and resampled audio signal of 44.1 kHz, (2) STFT (fft-size = 4,096 ($\approx$100 ms), hop = 441 ($\approx$10 ms)) to generate a decibel-converted power-spectrogram, and (3) random intensity [-6dB, +3dB], pitch [0.5, 1.5], and time augmentation [0.5, 2.0] resulting in a 2,049$\times$T decibel-converted and augmented power-spectrogram, where T is the number of analyzed time frames. The machine-segmented AOEC archive consists of orca samples containing an unequal temporal context T (1.28-2.00 s), which in turn may include a varying number of orca vocalizations per segment. In order to provide ORCA-CLEAN during training with a uniformly large temporal context T per segment, under the constraint of extracting isolated and stand-alone killer whale sounds, an orca detection algorithm is introduced. This method acts as an additional preprocessing step and extracts isolated orca signals of a fixed temporal context T (in our case 1.28 s) from our various-sized machine-segmented [10] orca samples of the AEOC data corpus. In cases, where the time augmentation led to orca segments $<$ 1.28 s zero-padding was conducted. The orca detection algorithm takes as input a 2,049$\times$T decibel-converted and augmented power-spectrogram and returns a 2,049$\times$128-large version of it by conducting the following steps: (1) various spectral operations (e.g. maximum/median filtering, histogram equalization, global thresholding, self-potentiating, moving average, and morphological operations) to spot, highlight, and keep only high intensity regions while removing the rest, (2) summation of the 0/1-dB-normalized (min = 100 dB, ref = +20 dB [30]) remaining spectral intensities using a sliding window approach (window-size = 1.28 s, hop-size = 10 ms) representing the spectral intensity per window as a function of time, (3) peak-picking algorithm to identify the global spectral maximum of the respective intensity function, and (4) extraction of the 1.28 s-large window, representing the intensity maximum, from the 2.049$\times$T decibel-converted and augmented power-spectrogram. In summary, the data preprocessing pipeline provides a 2,049$\times$128 augmented dB-spectrogram.

Additive Noise Variants – Similar to the image-based Noise2Noise [11] approach, we use various distributions of synthetic noise in order to corrupt the noisy original spectral shapes of our preprocessed 2,049$\times$128-large killer whale spectrograms. In addition, we used real-world underwater noise (e.g. boat noise, water noise, and/or other animals) from other Orchive tapes, and histogram-equalization to redistribute the spectral intensities and thereby simulating different noise characteristics, as another variants to distort the original noisy signals. Before corrupting every 2,049$\times$128 augmented dB-spectrogram via one of the additive noise alternatives, each spectrogram is linear frequency compressed (nearest neighbor, 256 bins, fmin = 800 Hz, fmax = 10 kHz) resulting in a 256$\times$128 dB-spectrogram. In summary, our proposed deep denoising approach integrates the following additive noise variants, including parts of the artificial noise-distributions illustrated in Noise2Noise [11], as well as additional real-world underwater noises, and histogram-equalized modified spectrograms: (1) Gaussian distribution using a zero-mean and random standard deviation $\sigma \in [0, 25]$ for training and $\sigma = 12.5$

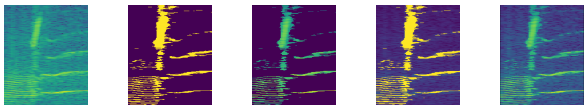
during testing, (2) Chi-squared distribution using a random factor $\chi \in [0, 30]$ for training and $\chi = 15.0$ during testing, (3) Poisson distribution using a random factor $\lambda \in [0, 30]$ for training and $\lambda = 15.0$ during testing, (4) Exponential distribution using a random factor $\epsilon \in [0.05, 0.15]$ for training and $\epsilon = 0.10$ during testing, (5) Real-world underwater noise using a random SNR $\in [-2, -8]$ dB for training and SNR $= -5.0$ dB during testing, and (6) Histogram-equalization to re-distribute intensity scalings. All mentioned random scalings are based on a uniform distribution. Finally, the 256×128 noise-corrupted spectrogram is normalized via a 0/1-dB-normalization (min = 100 dB, ref = +20 dB) [30]. Figure 2 exemplary visualizes all potential outputs of the 6 various additive noise alternatives.



N04 Gauss Chi2 Poisson Exp. Real Histo

Figure 2: *Noisy N04 orca call type – Additive noise variants*

Binary Mask Generation – By means of the different additive noise variations the entire spectrum is affected equally and thus a global denoising in all spectral ranges can be represented and learned by a network. However, compared to the traditional Noise2Noise [11] approach designed for image restoration, where all image regions are of equal interest, denoising leads to a differentiated attention within specific spectral areas. Whereas pure noise can be entirely removed, noisy orca vocalizations need to be cleaned in a way that as much noise as possible, but as few parts of the orca voicings as necessary, are eliminated. Therefore, in addition to the traditional Noise2Noise [11] idea, machine-generated binary masks (orca = 1, noise = 0) of killer whale sounds were utilized, acting as a network attention mechanism. Before automatically creating spectral binary masks, the 2,049×128 augmented dB-spectrogram is normalized by means of a 0/1-dB-normalization (min = 100 dB, ref = +20 dB) [30]. The result is used as an input for the binary mask generation including the following steps: (1) a maximum filter (kernel = 4×4) to spot and highlight local maxima in the 0/1-dB-normalized spectrum, (2) moving average (window-size = 25 bins, hop-size = 25 bins) calculating and substracting the spectral average per window, (3) global thresholding ($\delta = 25 \times 10^{-3}$) to set small intensity values to zero, (4) Otsu binarization [31] to binarize (orca = 1, noise = 0) the remaining spectral content, (5) morphological operation – erosion (kernel = 3×3), (6) median filter (kernel = 5×5), (7) morphological operation – erosion (kernel = 2×2), (8) median filter (kernel = 3×3), and (9) zeroing all spectral content below 800 Hz and above 10 kHz. Finally, the resulting 2,049×128-large 0/1-mask is linear frequency compressed (nearest neighbor, 256 bins, fmin = 800 Hz, fmax = 10 kHz) to a 256×128 binary spectrogram. The automatically generated binary masks, together with the corresponding original noisy spectrograms, enable different manifestations of potential denoised ground-truth samples (see Figure 3). There exists an integrated random selection between four distinct ground-truth possibilities: (1) the binary mask itself, (2) the binary mask multiplied with the original noisy spectrogram, (3) self-potentiating of the noisy original spectrogram using a random exponent $\xi \in [1.3\ ,\ 2.7]$ for training and 2.0 during testing, whereas all non-zero elements of the binary mask are set to one within the self-potentiated spectrogram, and (4) identical to alternative 3 while setting all non-zero elements of the binary mask to the original values of the input spectrogram. Figure 3 visualizes a noisy N07 call type, together with the illustrated orca/noise mask alternatives. The different versions of binary masks should ensure a balanced learning process, to counteract as many binary outliers as possible being a consequence of potential errors within the binarization process. Considering only the orca vocalizations: in alternative 1 and 3 the orca vocalizations are set to maximum intensity. The option 2 and 4 use the original spectral values of the orca calls. Thus, ORCA-CLEAN learns a trade-off between spectral highlighting of weak, but also strong orca vocalizations, compared to surrounding noisy regions, while maintaining the original orca-related spectral distribution as much as possible. Regarding the background noise: in variants 1 and 2 the surrounding noise is set to zero. Options 3 and 4 self-potentiate surrounding structures without removing it completely. The aim is to learn a trade-off between a complete removal of environmental noise, and a reduction of the surrounding noise, while preserving important residual orca structures. In general, the automatic binary mask generation provides a result that distinguishes between spectrally strong and weak signal regions, assuming that the emitted orca calls have stronger spectral intensities. If only the generated binary masks are used for training ORCA-CLEAN, it learns primarily to distinguish only between strong and weak spectral regions. However, not all regions with strong spectral intensities can always be assigned to orca vocalizations (e.g. stationary boat/engine-noises, microphone artifacts, buzzes, etc.). The enormous, and often very different noise characteristics, as well as short-term fluctuating noise, lead also to a certain error-proneness within the automatic creation process of the binary masks (eliminating valuable orca vocalization versus keeping too much noise). Consequently, using binary masks without the additive noise variants, result to sometimes undesirable misinterpretations by the network. Either due to possible arbitrary and self-contained insertions of orca-like vocal structures, especially when dealing with pure noise samples, or removing too much orca speech, particularly in case of faint and/or unseen sound structures.



N07 (1) (2) (3) (4)

Figure 3: *Noisy N07 orca call type – Various orca/noise masks*

Deep Denoising Network – ORCA-CLEAN is based on the U-Net architecture [32, 11] (see Figure 4). The network is trained via the previously mentioned 0/1-dB-normalized and frequency-compressed 256×128 spectrograms. The 256×128 spectral signal pairs (input/output) can be divided into two use cases: (1) the original file and an additive noise modified version, and (2) the original file and one of the corresponding mask alternatives. In the first case, the network input is the noise corrupted spectrogram, whereas the original file is the output. In the second case the network input/output consists of the original file and the respective mask. In both cases the network input spectrogram is more noisy than the ground truth (see Figure 4). During training, each input sample is randomly modified by one of 10 possible options (6 additive noise variants, 4 orca-/noise-mask options) to build the spectral signal pair. The network was implemented in PyTorch [33], using an Adam optimizer with an initial learning rate of 10^{-4}, $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and a batch size of 16. The Mean Squared Error (MSE) loss was used as loss function. For evaluation of the best model the validation loss was used as target metric. The individual convolutional layers (kernel = 3×3, stride = 1) are followed by a batch normalization and LeakyReLU ($\alpha = 0.1$). The last convolutional layer (kernel = 1×1, stride = 1) in the expansive path is just a plain convolution followed by a sigmoid. In the contracting path traditional max-pooling (kernel = 2×2, stride = 2) is used.

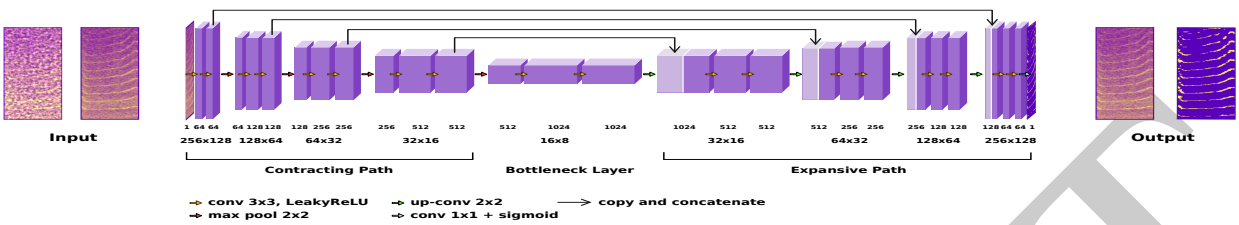


Figure 4: *ORCA-CLEAN – Deep denoising network architecture*

Transposed convolutions (kernel = 2×2, stride = 2) are utilized for upsampling within the expansive path. We used a learning rate decay of 0.5 after 4 epochs and stopped training after 10 epochs without any improvements on the validation loss.

5. Experiments

In a first experiment ORCA-CLEAN was trained and evaluated on the machine-segmented AOEC archive. Afterwards, the unseen Call Type Data corpus was used to verify and visualize the denoised output. In a second experiment we reproduced all our previous approaches regarding (1) orca/noise segmentation [10], and (2) orca call type identification [29, 30] under identical conditions, while using ORCA-CLEAN as an additional data preprocessing and enhancement step. The last experiment includes two steps: (1) back transformation of the complex denoised spectrum to the audio domain and subsequent spectral visualization, and (2) cross-domain generalization and transferability of ORCA-CLEAN utilizing bird and human speech data.

6. Results and Discussion

Visualization of denoised orca vocalizations – Figure 5 shows the 256×T denoised network output of multiple orca sounds in different noisy underwater conditions, being part of the unseen Call Type Data corpus. ORCA-CLEAN without any doubt substantiates: (1) significant signal denoising/enhancement, (2) spectral emphasis of orca vocalizations, (3) model robustness towards various vocalization types and environmental/ambient noises, and (4) model generalization regarding unseen orca data.

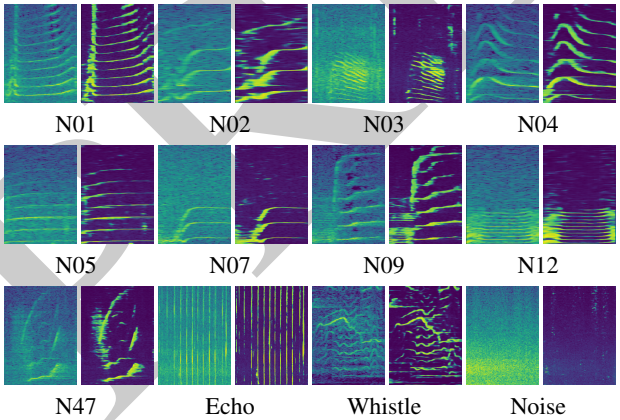


Figure 5: *ORCA-CLEAN on the unseen Call Type Data corpus*

Orca/Noise Segmentation – ORCA-SPOT [10], together with ORCA-CLEAN, was re-trained using the same data and parametric setup as in [10]. In comparison, the following old/new results were achieved, outperforming our previous best model in [10]: Accuracy = 94.97/96.04 %, Precision = 92.28/96.64 %, True-Positive-Rate = 93.77/92.10 %, False-Positive-Rate = 4.36/ 1.77 %, and Area-Under-The-ROC-Curve = 98.28/98.44 %.

Orca Call Type Identification – Our previous supervised orca call type classification results can be distinguished in (1) classification without pretraining [29], and (2) classification using representation learning [30]. In case 1 we achieved on the Call Type Data corpus a mean test accuracy (10 runs) of 87 % [29] and in case 2 of 94 % [30]. The best model in [30] reached 96 %. Together with ORCA-CLEAN, without the benefit of any pretraining and an identical parametric/data setup as in [29, 30], we outperformed both previous results by achieving a mean test accuracy (10 runs) of 95 % and best model accuracy of 98 %.
Model Generalization/Transferability – To show and prove transferability as well as model generalization even more, we evaluated ORCA-CLEAN, only trained on killer whale signals, on other bioacoustic data, using our own bird (monk parakeet – *Myiopsitta monachus*) corpora, and also switched to a different domain, utilizing noisy human speech [34]. We decompressed the 256×T frequency-compressed network output (see Figure 5), multiplied it with the original 2,049×T complex spectrum, and transformed it back to the audio domain. Due to lack of space we can only visualize one example for human speech, birds, and orcas (see Figure 6). More denoised orca, bird, and human speech signals can be viewed and listened to under [35].

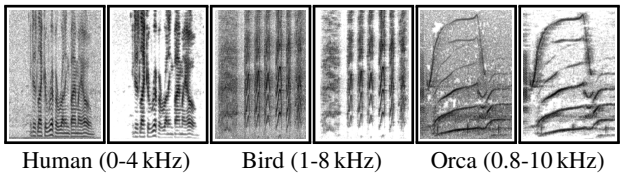


Figure 6: *Original versus denoised 2,049×T audio spectrum*

7. Conclusion and Future Work

In this work we present an orca-based deep denoising approach not requiring any clean ground truth, being a hybrid solution between the image-based Noise2Noise [11] idea and a network attention mechansim including machine-generated orca/noise masks. While using ORCA-CLEAN, we managed to outperform our previous works [10, 29, 30]. Furthermore, ORCA-CLEAN leads to an explicit visual/auditory cross-domain (orca, bird, human speech) signal enhancement. In our future studies we will integrate ORCA-CLEAN also into our fully unsupervised feature learning/clustering approach for call type identification (see [36]). Moreover, we will further investigate ORCA-CLEAN on other bioacoustic data/human speech. The source code of ORCA-CLEAN will be publicly available under [35].

8. Acknowledgements

The authors would like to thank Helena Symonds and Paul Spong from OrcaLab, and Steven Ness, formerly UVIC, for giving us permission to use the raw data and annotations from the orcalab.org, and the Paul G. Allen Frontiers Group for their initial grant for the pilot research. Moreover, the authors would like to thank Michael Weber for designing the U-Net image.

9. References

[1] S. 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, 3800 Finnerty Road, Victoria, British Columbia, Canada, V8P 5C2, 2013.

[2] J. Zhang, K. Huang, M. Cottman-Fields, A. Truskinger, P. Roe, S. Duan, X. Dong, M. Towsey, and J. Wimmer, "Managing and analysing big audio data for environmental monitoring," in *2013 IEEE 16th International Conference on Computational Science and Engineering*, Dec 2013, pp. 997–1004.

[3] G. Håkansson and J. Westander, *Communication in Humans and Other Animals*. John Benjamins Publishing Company, 2013.

[4] S. Ness, "Orchive," http://orchive.cs.uvic.ca/ (April 2020). [Online]. Available: http://orchive.cs.uvic.ca/

[5] 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, Jan. 1987.

[6] J. Towers, G. M. Ellis, and J. K. B. Ford, "Photo-identification catalogue and status of the northern resident killer whale population in 2014," Fisheries and Oceans Canada, Science Branch, Pacific Region, Pacific Biological Station, 3190 Hammond Bay Road, Nanaimo, British Columbia, Canada V9T 6N7, Tech. Rep. 3139, September 2015.

[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, January 1990.

[8] 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.

[9] ORCALAB, "Orcalab - a whale research station on Hanson Island," http://orcalab.org (April 2020). [Online]. Available: http://orcalab.org/

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

[11] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," in *ICML*, 2018.

[12] C. Tian, Y. Xu, L. Fei, and K. Yan, "Deep learning for image denoising: A survey," in *Genetic and Evolutionary Computing*. Singapore: Springer Singapore, 2019, pp. 563–572.

[13] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 341–349.

[14] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void - learning denoising from single noisy images," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2124–2132, 2018.

[15] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-cnn for image restoration," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 886–88 609, 2018.

[16] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," in *Advances in Neural Information Processing Systems*, 2018, pp. 1680–1689.

[17] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

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

[19] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Proc. Interspeech 2017*, 2017, pp. 3642–3646. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1428

[20] D. Hepsiba and J. Justin, "Role of deep neural network in speech enhancement: A review," in *Artificial Intelligence*, J. Hemanth, T. Silva, and A. Karunananda, Eds. Singapore: Springer Singapore, 2019, pp. 103–112.

[21] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[22] J. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016 - Proceedings*, vol. 2016-May, May 2016, pp. 31–35.

[23] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, 08 2017.

[24] N. Priyadarshani, S. Marsland, and I. Castro, "Automated birdsong recognition in complex acoustic environments: a review," *Journal of Avian Biology*, vol. 49, 01 2018.

[25] A. Brown, S. Garg, and J. Montgomery, "Automatic and efficient denoising of bioacoustics recordings using mmse stsa," *IEEE Access*, vol. PP, pp. 1–1, 12 2017.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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 *Text, Speech, and Dialogue, 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings*, vol. 11697 LNAI. Springer Verlag, 2019, pp. 274–286.

[31] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[32] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351. Springer, 2015, pp. 234–241.

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS 2017 Workshop*, October 2017.

[34] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and tts models," 2017. [Online]. Available: http://datashare.is.ed.ac.uk/handle/10283/2791

[35] C. Bergler, "Github repository." [Online]. Available: https://github.com/ChristianBergler

[36] 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 *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 3357–3361.