ORCA-WHISPER: An Automatic Killer Whale Sound Type Generation Toolkit Using Deep Learning

Christian Bergler¹, Alexander Barnhill¹, Dominik Perrin¹, Manuel Schmitt¹, Andreas Maier¹, Elmar Nöth ¹

¹Friedrich-Alexander-University Erlangen-Nürnberg, Pattern Recognition Lab, Erlangen, Germany
{christian.bergler, elmar.noeth}@fau.de

Abstract

Even today, the current understanding and interpretation of animal-specific vocalization paradigms is largely based on historical and manual data analysis considering comparatively small data corpora, primarily because of time- and human-resource limitations, next to the scarcity of available species-related machine-learning techniques. Partial human-based data inspections neither represent the overall real-world vocal repertoire, nor the variations within intra- and inter-animal-specific call type portfolios, typically resulting only in small collections of category-specific ground truth data. Modern machine (deep) learning concepts are an essential requirement to identify statistically significant animal-related vocalization patterns within massive bioacoustic data archives. However, the applicability of pure supervised training approaches is challenging, due to limited call-specific ground truth data, combined with strong class-imbalances between individual call type events. The current study is the first presenting a deep bioacoustic signal generation framework, entitled ORCA-WHISPER, a Generative Adversarial Network (GAN), trained on low-resource killer whale (Orcinus Orca) call type data. Besides audiovisual inspection, supervised call type classification, and model transferability, the auspicious quality of generated fake vocalizations was further demonstrated by visualizing, representing, and enhancing the real-world orca signal data manifold. Moreover, previous orca/noise segmentation results were outperformed by integrating fake signals to the original data partition.

Index Terms: Killer Whale, Deep Learning, Call Type, Generative Adversarial Networks

Machine- and data-driven killer whale call type recognition is an imperative requirement to gain a more profound understanding of the actual communicative behaviour, providing the foundation for identification of reappearing linguistic patterns, which in turn, are an essential ingredient to decode animal communication [12]. To obtain an overall picture about the diversity of killer whale call types, existing knowledge, derived from historically conducted research [5, 6, 7, 14], has to be combined with modern, task-specific, and data-driven machine (deep) learning techniques, together with large-scale bioacoustic data volumes [8, 9, 10, 11, 12, 13]. However, besides missing information about the entire animal-specific call type repertoire, as well as the associated inter and intra call type variety, the scarcity of human-labeled, species-related, and task-specific ground truth killer whale call type data severely impedes straightforward supervised training of robust (deep) machine learning algorithms. In addition, a strong class imbalance within either small human-labeled, and known call type data collections [9], or large-scale machine-identified known and unseen call pattern hypotheses [12], significantly affects the setup of a balanced, robust, and generalizing training procedure. Similar to human speech, the use of certain orca vocalization patterns is very diverse, combined with population-, group-, and individual-specific unique acoustic paradigms, all together causing such disparities. Data augmentation is often applied to artificially enlarge low-resource training material, enhance data variation, and compensate category-specific data imbalances. However, traditional image-/audio-based data augmentation techniques provide only limited reasonable alternative data samples [17], present the risk of producing unrealistic examples [18], and proved to be challenging while creating data points from the “true distribution” [19]. To counteract all of the aforementioned challenges – (1) lack of human-labeled and representative killer whale call type data, (2) strong class imbalance between orca vocalization paradigms, and (3)
restrictions of traditional data augmentation techniques – Generative Adversarial Networks (GANs) were deployed to design a deep data-driven bioacoustic signal generation procedure. The current study presents ORCA-WHISPER, an automatic, machine-driven, deep learning-based killer whale signal generation framework, trained on low-resource vocalization type-specific data archives, performing robust, category-specific, and adversarial bioacoustic sound synthesis to represent and enhance the original real-world killer whale data manifold with a very high fidelity.

2. Related Work

In recent years, Generative Adversarial Networks (GANs) have gained remarkable and massive attention, especially in the research fields of computer vision [20, 21, 22, 23, 24]. However, extensive research has also been conducted within the auditory domain. Radford et al. [25] introduced Deep Convolutional GAN (DCGAN), acting as a core-concept for various image but also audio-based generative adversarial concepts. Donahue et al. [26] presented two GAN-based approaches for unsupervised raw-waveform audio synthesis, both of them using the underlying concept of DCGAN: (1) WaveGAN, using raw-waveform audio input to generate raw-waveform audio output, and (2) SpecGAN, operating on the time-frequency representation of a given raw audio waveform as network input (spectrogram), while also performing raw-waveform audio synthesis as model output. Marafioti et al. [27] invented a new and revised method for generative adversarial audio synthesis based on time-frequency (spectral) information, entitled TifGAN, inspired by DCGAN [25] and applying similar principles as introduced in WaveGAN/SpecGAN [26]. Especially in the domain of human speech, generative adversarial approaches are increasingly deployed for speech synthesis [28, 29], but also as data augmentation mechanisms to produce realistic speech samples, showing a very high fidelity [18, 30]. However, generative adversarial animal-specific audio synthesis is considerably less studied in bioacoustics. Pagliarini et al. [31] trained WaveGAN [26] on a huge dataset of canary syllables, while using a heavily extenuated latent feature space, whereas the generated syllable types were evaluated via a RNN-based classification system, next to a UMAP-based comparison between real-world and generated signals [31]. To the best of the authors knowledge, this is the first study addressing deep learning-based orca sound type generation, using a generative adversarial network, trained on low-resource killer whale call type data, in order to produce orca vocalization patterns with a high fidelity to compensate class imbalance and missing ground truth data.

3. Data Material

3.1. Call Type Data Corpus (CTDC)

The Call Type Data Corpus (CTDC) [9, 11, 12, 13], combines three stand-alone human-labeled sub-datasets, including monosignal killer whale sound types of various duration, sampled at 44.1 kHz: (1) OrcaLab Call Type Catalog ( CCS), consisting of 138 audio excerpts, spread across 7 killer whale call type categories (N1 – 33, N2 – 10, N4 – 21, N5 – 14, N7 – 18, N9 – 26, and N12 – 16), (2) Ness Call Type Catalog (CCN), including 286 audio samples, distributed among 6 call types (N1 – 36, N3 – 56, N4 – 60, N7 – 31, N9 – 70, and N47 – 33), and (3) Extension Catalog (EXT), containing 90 audio files, split into echos, whistles, and environmental noise (echo – 30, whistle – 30, noise – 30). Overall, the entire CTDC dataset comprises 514 audio samples, distributed across 12 classes [9, 11, 12, 13].

3.2. ORCA-WHISPER Low-Resource Dataset (OWLR)

The ORCA-WHISPER Low-Resource Dataset (OWLR) is a balanced low-resource data collection including 600 (= 15.0 min., duration per sample = 1.5 s) human-verified, high-quality, and representative killer whale call type audio samples, spread across 6 categories (N1, N3, N4, N5, N9, N47 – see Figure 1) with a total of 100 mono-signal orca excerpts per category, each with a sampling rate of 44.1 kHz. The OWLR archive was derived from the ORCA-SLANG Call Type Data Corpus (OSDC), the result of a multi-stage, large-scale, machine-driven deep learning framework, entitled ORCA-SLANG [12], processing the entire Archive (20,000 h) [14, 15, 16] via the following sequentially-ordered processing steps: (1) orca/noise segmentation [8], (2) orca signal denoising [11], (3) deep orca feature learning [10], (4) semi-supervised killer whale call type identification, applying clustering [10] and classification [9, 32].

The OSDC dataset consists of 235,369 (=393.2 h, average duration per sample ≈ 6.0 s) audio excerpts, spread across 6 distinct killer whale call type categories (see Figure 1) [12, 13]. The call type specific data partitions are distributed as follows [12]: N1 – 19,280, N3 – 8,484, N4 – 145,760, N5 – 8,861, N9 – 41,990, N47 – 10,994. However, in real-world situations, such large species-specific call type data sources usually do not exist, either due to the size of the original and available raw data material, or because of missing animal-independent machine and data-driven (deep) learning frameworks. Nevertheless, in order to facilitate a broad usage of the proposed framework, it is absolutely necessary to simulate the existing real-world data scarcity of annotated ground truth data. Therefore, ORCATYPE [9], a ResNet18-based multi-class Convolutional Neural Network (CNN), designed for supervised orca sound type classification and trained on the CTDC dataset, was applied to machine-select the top-200 data samples from each call type data pool of the OSDC repository showing the highest classification (posterior) probability. For this purpose, a fixed temporal context of 1.5 s was previously extracted for all OSDC audio data excerpts of varying duration, applying the orca intensity detection algorithm presented in [11]. In a last step, the top-200 machine-identified excerpts per killer whale call type were manually verified via audiovisual inspection, whereas only the top-100, category-specific, high-quality, and representative 1.5 s-long data samples were chosen, resulting in the final, balanced, and low-resource OWLR repository, completely independent and without any overlap to the CTDC data pool.

4. Methodology

4.1. Data Preprocessing

In order to facilitate robust network training, a multi-level data preprocessing scheme was applied, consisting of the following sequentially-ordered individual steps: (1) audio conversion to mono-signal (averaging) and resampling to 44.1 kHz, (2) Discrete Gabor Transform (DGT) with a Gaussian window of 1,024 samples (≈23 ms) and hop-size of 260 samples (≈5 ms), resulting in a 513 (window-size/2 + 1) (=frequency bins) × 256 (=time bins) spectral representation, (3) linear call type specific frequency compression (nearest neighbor, f_max = 0 kHz, f_min = 10 kHz, f_max(N1,N3,N4,N5,N9,N47) = 12 kHz) to 512 frequency bins, and (4) spectral normalization, including 0/1-min/max-normalization (file-by-file), followed by clipping of x ≤ e⁻¹⁰, subsequent log-transformation, and -1/4-1-conversion, resulting in the definite -1/4-1-normalized 512×256-large spectral data preprocessing output of real-world orca call type samples.
4.2. ORCA-WHISPER

Generative Adversarial Networks (GANs), originally proposed by Goodfellow et al. [33], utilize two neural networks competing against each other within an adversarial supervised training procedure, whereas optimization is done in a joint and iterative manner: (1) Generator – randomly sampling information from a known distribution (e.g. \( \mathcal{N}(0, 1) \)), to produce artificial data samples being as close as possible to a given real-world distribution, which is determined by the training data material (2) Discriminator – differentiating if the input data samples belong to the real-world or generated fake signal distribution, by performing pure classification. Optimization of both networks – Generator and Discriminator – continues until an equilibrium is reached, a network state in which the discriminator is no longer able to discern between signals from the real and fake distribution. For a more detailed and general information about Generative Adversarial Networks see [33]. ORCA-WHISPER, visualized in Figure 2, is a Generative Adversarial Network using the concepts of TiFGAN [27] due to reported performance improvements compared to WaveGAN/SpecGAN [26]. The original TiFGAN-architecture [27] was extended by two additional layers in the Generator, as well as in the Discriminator to increase both, time and frequency resolution. However, in order to counteract the resulting increase in model complexity, the original feature map dimensions were reduced by a scaling factor of 8, which did not affect model performance (audio-visual inspection), but significantly decreased training times and GPU-based hardware requirements (larger Nvidia P100, \( \approx 33h \) versus smaller Nvidia 2070 Super, \( \approx 20h \)). For each of the 6 different orca call types (see OWLR) an individual ORCA-WHISPER model was trained. A single cross-call type network is not effective or recommended due to the demonstrated strong intra- and inter killer whale call type variations [12]. The associated data complexity, together with the low-resource data material in OWLR, severely interferes with model convergence, performance, and generalization. In addition, a cross-call type approach significantly increases the probability of observing mixtures of call type specific patterns, causing fake signals which cannot be assigned to any category. ORCA-WHISPER samples from a \( \mathcal{N}(0, 1) \)-distribution, leading to a 100x1-large network input for the Generator, whereas the Discriminator part returns in a 1x1 output (see Figure 2). The model uses an Adam optimizer with an initial learning rate of \( 10^{-3} \), \( \beta_1 = 0.5 \), and \( \beta_2 = 0.9 \), besides the WGAN-GP loss function [34] (penalty coefficient \( \lambda = 10 \)) originally used with TiFGAN [27], batch-size of 8, and training iterations \( \min/\max = 10,000/20,000 \). Therefore, each training iteration processes 8x512x256 Generator samples, next to 8x512x256 real-world signals, whereas the weight-update of the Generator was performed only after every 5th-iteration, while freezing weights of the Discriminator and vice-versa. Model training was terminated based on an audio-visual analysis of the 512x256 Generator results, along with monitoring the Generator/Discriminator loss. The final audio conversion (1.5 s-long fake samples) of the -1/+1-converted (see TanH activation in Figure 2) 512x256-large Generator output was realized via: (1) inverse spectral normalization, (2) linear spectral decompression, and (3) inverse DGT (see section 4.1).

5. Experiments

The entire experimental setup consists of five major experiments. First, a stand-alone ORCA-WHISPER model was trained for each call type within the OWLR dataset (6 models in total). In a first instance, the quality of the generated call type specific fake signals was evaluated via audiovisual inspection. In a second experiment, a total of 125 1.5 s-long fake signals were generated from each model, resulting in a balanced call type fake data pool of 750 samples (\( \approx 19 \) min.). In addition, 25 samples of unseen, representative, and real-world killer whale call type exceptions were added, resulting in a total of 150 real samples. The quality of these 900 files was assessed via multi-class classification applying ORCA-TYPE [9] (ResNet18-based multi-class CNN), trained on the CTDC dataset, however, only on the existing 9 different call types, while ignoring echolocation clicks, whistles, and noise. The accuracy was used as a quality criterion for the 9-class problem. In a third experiment, the 1x512 latent feature vector of ORCA-TYPE [9] was extracted for all 900 data files followed by a dimensionality reduction using PCA (1x512 \( \rightarrow 1x50 \)) and t-SNE (1x50 \( \rightarrow 1x2 \)), together with various visualizations of the 2D-compressed representations, including: (1) 25 real-world signals per call type, (2) 25 fake signals per call category, (3) 25/25 real/fake signals per vocalization type, and (4) 25/125 real/fake signals per killer whale call type. In a fourth experiment, a total of 30,000 killer whale fake signals (5,000 per call type) were produced, using all 6 previously trained models. All signals were added as part of a data augmentation procedure to the training set (validation/test set remained the same as in [8]) of ORCA-SPOT [8], a ResNet18-based binary-class CNN for orca/noise segmentation, to further improve performance on the original unseen test material. In a fifth and last experiment, ORCA-WHISPER was applied to a bird species, namely Monk parakeets (Myiopsitta monachus), by modeling 3 different call type categories – contact, alarm, and other calls – in order to prove species-specific model transferability and performance.

6. Results and Discussion

Figure 3 visualizes different 512x256 spectrogram excerpts (9 real vs. 9 fake, \( f_{min} = 0 \) KHz, \( f_{max} = 10 \) KHz, duration=1.5 s), produced by the generator for each of the 6 distinct ORCA-WHISPER models, trained on low-resource call type specific
data pools, including only 100 real-world samples. Besides a very promising spectral and acoustic quality (audiovisual examples available under [35]), combined with a well-defined discrimination between individual call types, significant intra-call type variations are represented despite the low-resource training data, next to a precise reproduction of the presented spectral call type structures (see Figure 3). Moreover, the individual GAN-specific Generator models represent varying intensity characteristics, both, for the orca signal components (faint vs. strong) as well as the superimposed environmental background noise, which further favours the overall data variation (see Figure 3).

Figure 3: Real vs. fake killer whale call type spectrograms

In addition to the audiovisual analysis, ORCA-TYPE [9] was applied to perform orca call type classification on a total of 750 generated equally distributed fake killer whale call type signals and 150 unseen real-world excerpts. ORCA-TYPE [9] achieved a mean test accuracy (10 runs, while ignoring the 3-best and 3-worst models) of 92.5% based on the 750 category-specific fake signals, 97.8% on the 150 real-world samples, and 93.3% regarding all 900 files (best model 95.0%), all together demonstrating the auspicious quality of the fake signals. Further assessments regarding the generated fake call type quality were performed by analyzing the latent feature (PCA/t-SNE) space extracted via ORCA-TYPE [9] and visualized in Figure 4. Figure 4.1 and Figure 4.2 show a very similar intra- and inter-call type distribution, proving an accurate representation and coverage of the original real-world data manifold. In addition, the joint and merged fake vs. real feature space (balanced number of files – Figure 4.3, all samples – Figure 4.4) also represented, extended, and enhanced the real-world call type data manifold. Moreover, the improved results of ORCA-SPOT [8] are another confirmation of the promising quality of the generated fake call type data samples. In comparison to the best model in [8], the following test set metrics were achieved through additional fake training material, outperforming the original performance of ORCA-SPOT [8]: Accuracy (old/new) = 94.97/96.10 %, Precision = 92.28/96.94 %, True-Positive-Rate = 93.77/91.93 %, F1-Score = 93.00/94.37 %, False-Positive-Rate = 4.36/1.60 %, and Area-Under-The-ROC-Curve (test set vs. 9-unseen Archive recordings – see [8]) = 98.28/98.64 % vs. 95.23/95.24 %.

Figure 4: Visualization real vs. fake call type feature space

Finally, species-specific model transferability and performance turned out to be productive, while considering the 512×256 audiovisual results (see Figure 5 – 6 real vs. 6 fake, f_{min}=0kHz, f_{max}=10kHz, duration=0.25-0.40s) of the generated Monk parakeet call types (examples available here [35]), all together trained on just 100 excerpts per category and similar conditions.

Figure 5: Real vs. fake Monk parakeet call type spectrograms

7. Conclusion and Future Work

This study presents ORCA-WHISPER, a fully-automated machine-driven killer whale call type generation framework, based on low-resource deep adversarial learning. Within a comprehensive and significant experimental setup – audiovisual inspection, supervised classification, data manifold and feature space analysis, data augmentation, and model transferability – the auspicious quality of the generated fake signals was clearly demonstrated. In future work, ORCA-WHISPER will be further improved and used for data augmentation, counteracting data imbalance, and generation of additional ground truth data from low-resource archives. ORCA-WHISPER will be adapted to operate on animal-independent data, whereas source code and audiovisual examples, will be publicly available under [35].

8. Acknowledgements

The authors would like to acknowledge the German Research Council (DFG) for funding the project “Deep Animal Linguistic Analysis (DALA) – Decoding animal communication using a hybrid approach between bioacoustics and machine learning” (project number MA–4898/18–1) and “The Paul G. Allen Frontiers Group” for their initial grant. Thanks also to Helena Symonds & Paul Spong (OrcaLab), Steven Ness (formerly UVIC), for giving us permission to use the raw data archive from the orcalab.org and Simeon Smeele (Max-Planck-Institute), for giving us permission to the Monk parakeet data.
9. References


