Friedrich-Alexander-Universität
Technische Fakultät

Pattern Recognition Lab

ORCA-PARTY: An Automtatic Killer Whale Sound Type Separation Toolkit Using Deep Learning

2022 IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2022, Singapore, May 22nd – 27th, 2022

Christian Bergler[1], Manuel Schmitt[1], Andreas Maier[1], Rachael Xi Cheng[2], Volker Barth[3], Elmar Nöth[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab, Erlangen, Germany
[2]Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany
[3]Anthro-Media, Berlin, Germany

May 13th, 2022

INTRODUCTION

ORCA-PARTY – What it is about?

Speech perception among killer whales...

+

=

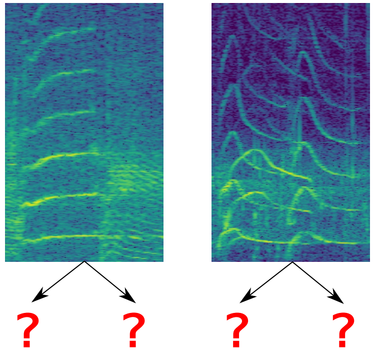
Killer Whale Research

Cocktail Party

ORCA-PARTY:
Crazy/Drunken Killer Whales

+

=



Killer Whale Research

Cocktail Party

ORCA-PARTY:
Killer Whale Sound Type Separation

- "Cocktail Party Problem", caused by multiple vocalizing killer whales → Overlapping call type structures

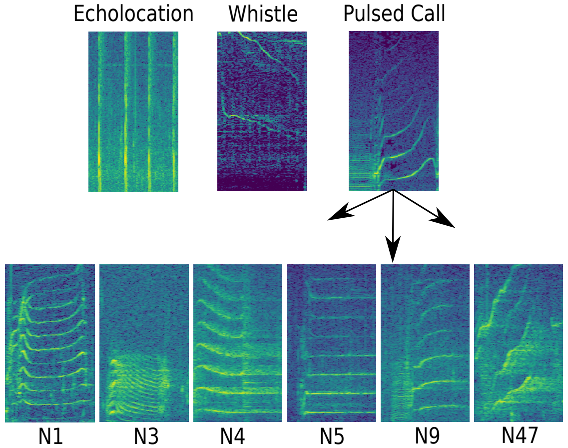
Source: Killer whale images taken from FIN-PRINT [2], Copyright Jared Towers & Gary J. Sutton, Other Images, Pixabay License – taken from https://pixabay.com/ – and recreated
Source: [2] Bergler et al., FIN-PRINT A Fully-Automated Multi-Stage Deep-Learning-Based Framework for the Individual Recognition of Killer Whales, Scientific Reports, 2022

The Killer Whale

...and the phenomenon of communication

- The Killer Whale (*Orcinus Orca*) is the largest member of the dolphin family [1] [2] [3]
- Lives in stable, family-based, and social groups of several individuals [1] [2] [4]
- Communicative behavior is based on three different types of vocalization paradigms [1] [3] [5]
  - Echolocation Clicks – Short pulses used for navigation and object localization
  - Whistles – Narrow-band signals primarily used within close-range interactions
  - Pulsed Calls – Most common type of vocalizations, subdivided into discrete, variable, and aberrant calls, showing distinct tonal properties
- Discrete Pulsed Calls (Call Types) are stereotyped and repetitive vocal activities, indicating a wide diversity of distinctive categories with significant inter- and intra-class spectral variations

Group of Killer Whales



Source: Killer whale images, Copyright Jared Towers & Gary J. Sutton, FIN-PRINT [2]
Source: [1] Bergler et al., ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning, Scientific Reports, 2019
Source: [2] Bergler et al., FIN-PRINT A Fully-Automated Multi-Stage Deep-Learning-Based Framework for the Individual Recognition of Killer Whales, Scientific Reports, 2022
Source: [3] Bergler et al., Deep Representation Learning for Orca Call Type Classification, Text, Speech, and Dialogue, 2019
Source: [4] Bergler et al., Deep Learning for Orca Call Type Identification – A Fully Unsupervised Approach, INTERSPEECH, 2019
Source: [5] Bergler et al., ORCA-SLANG: An Automatic Multi-Stage Semi-Supervised Deep Learning Framework for Large-Scale Killer Whale Call Type Identification, INTERSPEECH 2021

MOTIVATION & CHALLENGES

Killer Whale Sound Type Classification

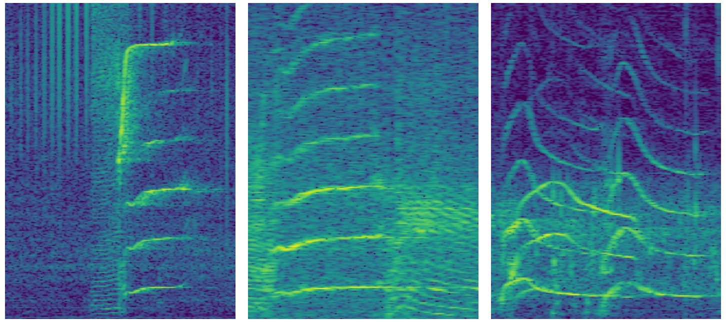
- Wide diversity of distinctive call type categories with significant inter- and intra-class spectral variations [5]
- Large-scale, data-driven, and machine-based orca call type identification is imperative to gain deeper insights into orca communication

→ Machine-based call type recognition [3] [4] [5] is substantially affected by overlapping call type structures!

Killer Whale Sound Type Separation

- Especially longer acoustic regions of orca communication, containing a large number of vocalization events in consecutive short time intervals
- Essential for communication analysis

→ High probability of overlapping call-specific events!



Source: [3] Bergler et al., Deep Representation Learning for Orca Call Type Classification, Text, Speech, and Dialogue, 2019
Source: [4] Bergler et al., Deep Learning for Orca Call Type Identification – A Fully Unsupervised Approach, INTERSPEECH, 2019
Source: [5] Bergler et al., ORCA-SLANG: An Automatic Multi-Stage Semi-Supervised Deep Learning Framework for Large-Scale Killer Whale Call Type Identification, INTERSPEECH 2021

Challenges

Killer Whale Sound Type Separation

- Robust machine learning pipeline to process massive and noise-heavy data repositories
- Limited knowledge about entire inter-/intra killer whale call type variations
- No ground truth data of overlapping call events and the associated individual components
- Huge call type-specific datasets are required to cover as much spectral variation as possible
- Single-channel acoustic events with no information about number of speakers, sound source location, speaker-specific data material, and various recording environments/setups.

Goal: Fully-automated machine (deep) learning-based orca sound type separation, independent of speaker-, sound source location-, and recording condition-specific knowledge, not requiring human-annotated overlapping ground truth data

DATA MATERIAL

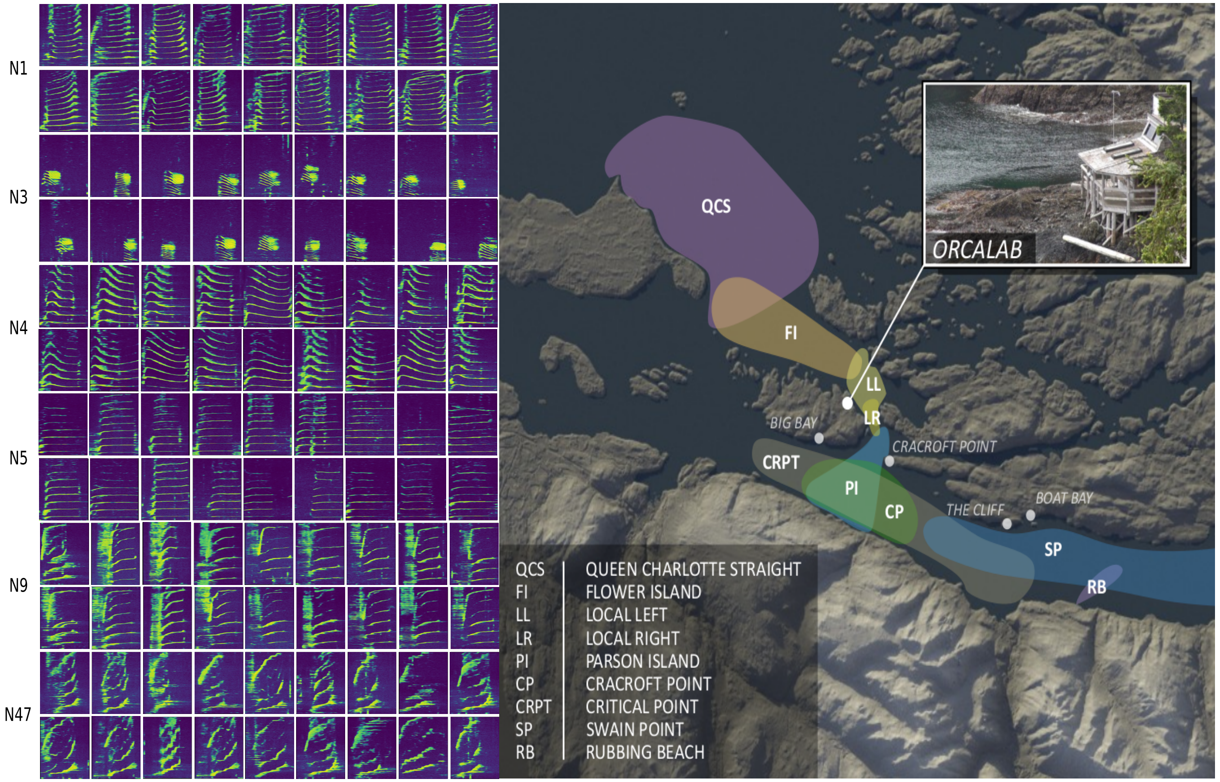
Data Archives

Killer Whale Sound Type Archive (KWSTA)

KWSTA consists of three sub-archives and is the result of applying machine (deep) learning algorithms (see ORCA-SLANG [5]) to one of the largest animal-specific data archives – The Orchive – including ≈20,000 h underwater recordings!

- ORCA-SLANG Call Type Data Corpus (OSDC)
  235,369 machine-identified orca samples, uneven distribute across 6 known call types
- Echolocation Repository (ELRP)
  9,382 echolocation events, machine-identified via ORCA-TYPE [3]
- ORCA-SLANG Unknown Signal Repository (OSUR)
  2,101 excerpts of either so far unseen/unknown orca sounds or background noise

The final KWSTA data repository includes 246,852 (≈398.1 h) unique orca events (mono, 44.1 kHz) with an average duration of ≈6.0 s



Source: Images taken from ORCA-SLANG [5], ORCA-SPOT [1]
Source: [1] Bergler et al., ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning, Scientific Reports, 2019
Source: [3] Bergler et al., Deep Representation Learning for Orca Call Type Classification, Text, Speech, and Dialogue, 2019
Source: [5] Bergler et al., ORCA-SLANG: An Automatic Multi-Stage Semi-Supervised Deep Learning Framework for Large-Scale Killer Whale Call Type Identification, INTERSPEECH 2021

Data Archives

Call Type Data Corpus (CTDC), DeepAL Fieldwork Data 17-19 (DLFD)

Call Type Data Corpus (CTDC)

Human-annotated dataset including 514 non-overlapping orca call type events, unequally split and categorized into 12 distinct classes [3] [6] [7] (9 killer whale call type categories, echolocation click, whistle, and noise)

DeepAL Fieldwork Data 2017/2018/2019 (DLFD)

Additional acoustic data collection via a 15-meter research trimaran during our fieldwork expedition along the coastal waters of northern British Columbia (2017–2019), resulting in ≈177.3 h (mono, 96 kHz) raw killer whale underwater recordings [1]

Source: Images taken from the DeepAL 2017–2019 expedition image collection (copyright Anthro-Media) and ORCA-SPOT [1]
Source: [1] Bergler et al., ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning, Scientific Reports, 2019
Source: [3] Bergler et al., Deep Representation Learning for Orca Call Type Classification, Text, Speech, and Dialogue, 2019
Source: [6] Bergler et al., ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication, INTERSPEECH 2020
Source: [7] Bergler et al., Segmentation, Classification, and Visualization of Orca Calls Using Deep Learning, ICASSP 2019

DATA PROCESSING

Data Preprocessing

...from Audio to a Spectral Representation

Multi-Stage Data Preprocessing Procedure [1] [6]

- Conversion to a single-channel audio file
- Resampling to 44.1 kHz
- Short-Time-Fourier-Transform (STFT) using a window-size = 4,096 samples ($\approx$100 ms) and hop-size = 441 samples ($\approx$10 ms) $\rightarrow$ Frequency$\times$Time (F$\times$T) power-spectrogram
- Decibel conversion of the F$\times$T power-spectrogram
- Orca Detection Algorithm [6] to extract a fixed temporal context of 1.28 s (T = 128)
- Linear frequency compression (nearest neighbor, fmin = 500 Hz, fmax = 10 kHz, F = 256)
- 0/1-dB-normalization (min = 100 dB, ref = +20 dB)

$\rightarrow$ Final Output: 256$\times$128 0/1-dB-normalized spectrogram

Source: [1] Bergler et al., ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning, Scientific Reports, 2019
Source: [6] Bergler et al., ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication, INTERSPEECH 2020

Data Generation

Machine-Generated Overlapping Killer Whale Vocalizations

Multi-Stage Data Generation Procedure

- Random selection of 37,101 samples from the KWSTA repository – 5,000 events per call type from the OSDC, 5,000 echolocation clicks of the ELRP, plus the entire OSUR data pool
- Spectral signal enhancement (denoising) by applying ORCA-CLEAN [6]
- Overlap a pair of spectrograms using a randomly chosen duration interval $\delta \in [0.64\,s, 1.28\,s]$
- Randomly sub-sampling a temporal context of 1.28 s (T = 128)
- 0/1-min/max-normalization of the 256×128-large overlapping spectrogram
- 2,000 overlapping spectral events for each of the 42 combinations (8 categories – 7 orca sound types plus a rejection class)

$\rightarrow$ Final Output: ORCA-PARTY Overlapping Dataset (OPOD), consisting of 84,000 256×128-large, 0/1-min/max-normalized, overlapping spectral representations

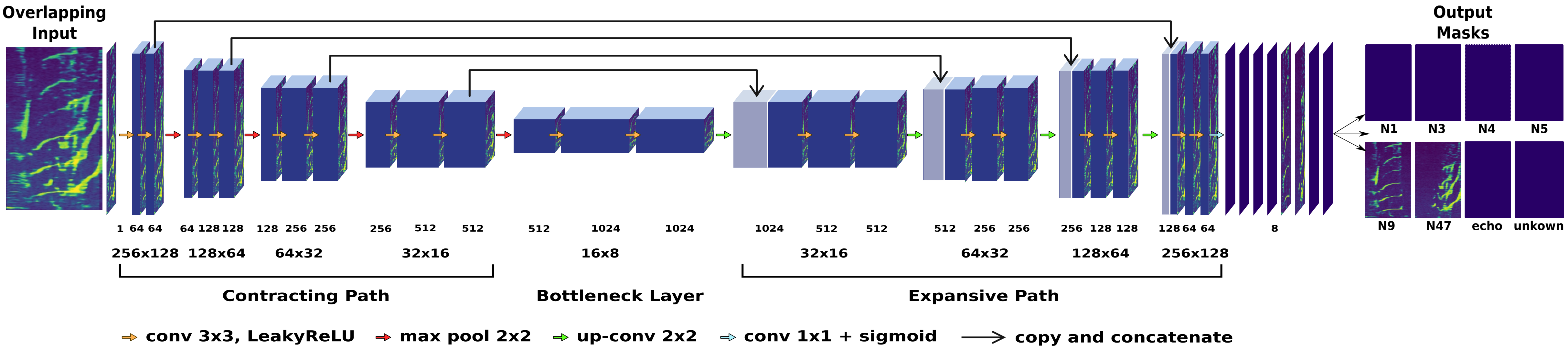
Source: [6] Bergler et al., ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication, INTERSPEECH 2020

METHODOLOGY

Network Architecture and Training

The Setup of ORCA-PARTY

ORCA-PARTY Architecture



- Network Input: 256×128-large, 0/1-min/max-normalized overlapping signals from the OPOD
- Network Output: 8 category-specific activated segmentation masks (7 orca sound types plus a rejection class)
- Data distribution: train – 58,800 (70 %), dev – 12,600 (15 %), test – 12,600 (15 %)

1st Experiment

Visual inspection and classification of the network output masks from the unseen OPOD test set, while ignoring the “unknown” class $\rightarrow$ 8,400 out of 12,600 test events

2nd Experiment

ORCA-TYPE [3] was trained on the denoised (ORCA-CLEAN [6]) human-labeled CTDC mask-specific data, with and without ORCA-PARTY (O-WP & O-BL) as additional data preprocessing step, evaluated on:

- Unseen non-overlapping CTDC test set
- Sliding window approach to iterate frame-wise over pre-segmented/-denoised excerpts $\Psi \in [10.0s, 30.0s]$ of the unlabeled *DLFD* $\rightarrow$ Classification hypotheses of O-WP vs. O-BL !

3rd Experiment

Model transfer to train and evaluate ORCA-PARTY on a bird species, named Monk parakeets *(Myiopsitta monachus)*, with a total of 3,000 bird call events across 4 categories (alarm, other, contact call & noise)

Source: [3] Bergler et al., Deep Representation Learning for Orca Call Type Classification, Text, Speech, and Dialogue, 2019
Source: [6] Bergler et al., ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication, INTERSPEECH 2020

RESULTS & DISCUSSION

Results

Visualization/Classification Overlapping OPOD Test & CTDC Data

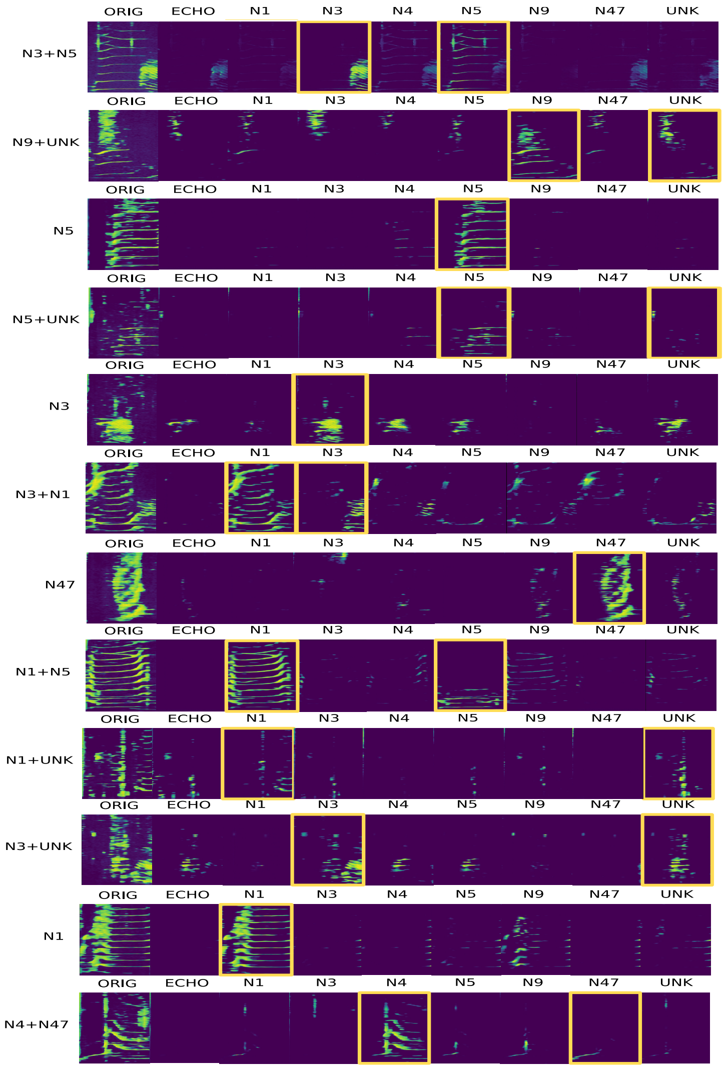
- Visualizations from the unseen OPOD test set, showing the original overlapping input spectrogram, compared to the class-based separation outputs

- Applying O-WP to the unseen overlapping 8,400 OPOD test samples (16,000 classification hypotheses) results in a multi-class accuracy of ≈86.0 %

- Applying O-BL as well as O-WP to the unseen non-overlapping CTDC dataset, an average classification accuracy of ≈96.0 % vs. ≈94.5 % (dev) and ≈94.5 % vs. ≈93.0 % (test) was achieved

  → O-WP almost reaches the upper classification boundary for non-overlapping signals, provided by O-BL!

ORCA-PARTY achieved auspicious results on overlapping data, besides robustly processing non-overlapping call type events!



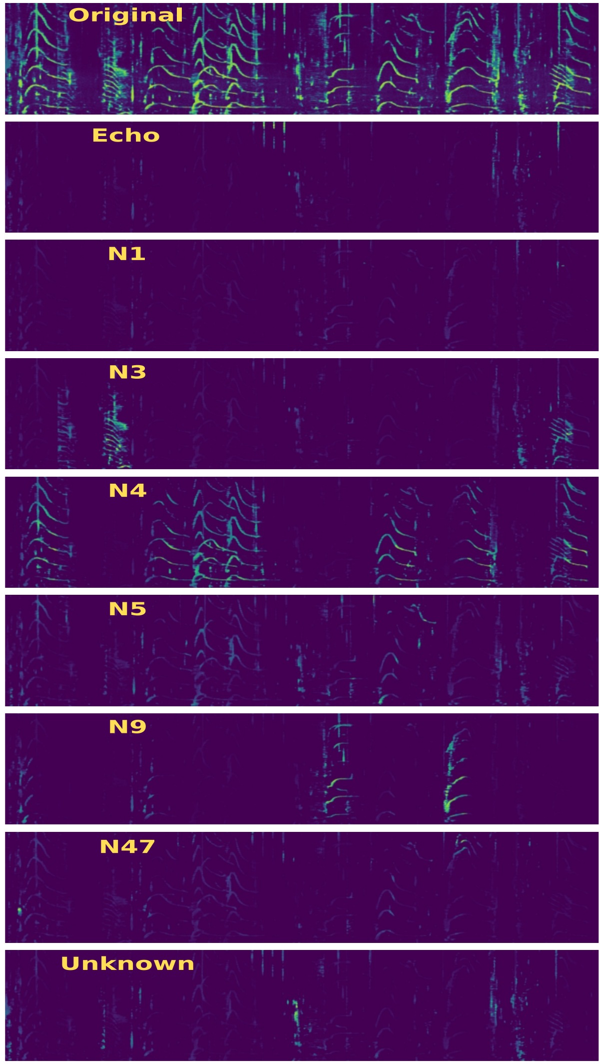
Results

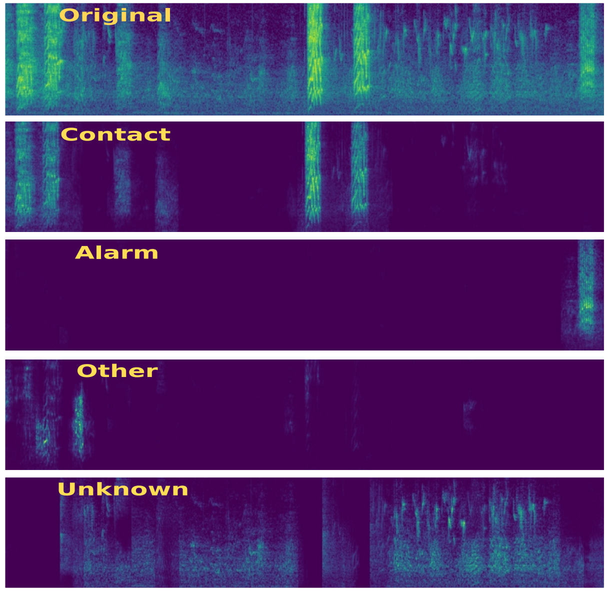
DeepAL Fieldwork Data & Monk Parakeets

- Applying O-BL vs. O-WP to frame-wise classify the entire DLFD archive results in the following overall amount of classification hypotheses:

  $\rightarrow$ 39,569 (O-BL) vs. 51,684 (O-WP) orca events distributed across 7 categories (increase of $\approx$30 %)

- ORCA-PARTY, trained on overlapping monk parakeet spectrograms, proved model transferability and achieved promising results even in noisy conditions





CONCLUSION & FUTURE WORK

Conclusion & Future Work

ORCA-PARTY – Wrap up and what's next?

Conclusion

ORCA-PARTY, is an automatic deep learning-based approach for orca sound type separation, not requiring any human-labeled overlapping ground truth data and is independent of speaker/-source information and various recording conditions.

- Additional data enhancement step
- Similar classification results were obtained for non-overlapping events
- Significant improvements were observed during the analysis of acoustic regions with high vocalization volumes, leading to $\approx$30 % more call identifications
- Promising initial results on various noisy bird calls

Future Work

- Future studies will evaluate performance on additional animal-related bioacoustic datasets
- Source code and audiovisual excerpts produced by ORCA-PARTY will be publicly available under [8]

Source: [8] Bergler Christian, Open Source GitHub-Repository

Thank you for your attention!

https://lme.tf.fau.de/person/bergler/

References

References I

[1] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer und A. Maier, „ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning“, *Scientific Reports*, Jg. 9, Dez. 2019. DOI: 10.1038/s41598-019-47335-w.

[2] C. Bergler, A. Gebhard, J. Towers, L. Butyrev, G. Sutton, T. Shaw, A. Maier und E. Nöth, „FIN-PRINT A Fully-Automated Multi-Stage Deep-Learning-Based Framework for the Individual Recognition of Killer Whales“, *Scientific Reports*, Jg. 11, S. 23 480, Dez. 2021. DOI: 10.1038/s41598-021-02506-6.

[3] C. Bergler, M. Schmitt, R. X. Cheng, H. Schröter, A. Maier, V. Barth, M. Weber und E. Nöth, „Deep Representation Learning for Orca Call Type Classification“, in *Proc. Text, Speech, and Dialogue 2019*, (Ljubljana), Bd. 11697 LNAI, Springer, 2019, S. 274–286. DOI: 10.1007/978-3-030-27947-9{_}23.

[4] C. Bergler, M. Schmitt, R. X. Cheng, A. Maier, V. Barth und E. Nöth, „Deep Learning for Orca Call Type Identification – A Fully Unsupervised Approach“, in *Proc. Interspeech*, (Graz), 2019. DOI: 10.21437/Interspeech.2019-1857.

[5] C. Bergler, M. Schmitt, A. Maier, H. Symonds, P. Spong, S. R. Ness, G. Tzanetakis und E. Nöth, „ORCA-SLANG: An Automatic Multi-Stage Semi-Supervised Deep Learning Framework for Large-Scale Killer Whale Call Type Identification“, in *Proc. Interspeech*, 2021. DOI: 10.21437/Interspeech.2021-616.

[6] C. Bergler, M. Schmitt, A. Maier, S. Smeele, V. Barth und E. Nöth, „ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication“, in *Proc. Interspeech*, 2020. DOI: 10.21437/Interspeech.2020-1316.

[7] H. Schröter, E. Nöth, A. Maier, R. Cheng, V. Barth und C. Bergler, „Segmentation, Classification, and Visualization of Orca Calls Using Deep Learning“, in *International Conference on Acoustics, Speech, and Signal Processing, Proceedings (ICASSP)*, IEEE, 2019, S. 8231–8235, ISBN: 978-1-4799-8131-1. DOI: 10.1109/ICASSP.2019.8683785.

[8] C. Bergler, *Open Source GitHub-Repository*, https://github.com/ChristianBergler.