Pattern Recognition Lab

FAU Friedrich-Alexander-Universität Technische Fakultät

icassp 2022 Singapore

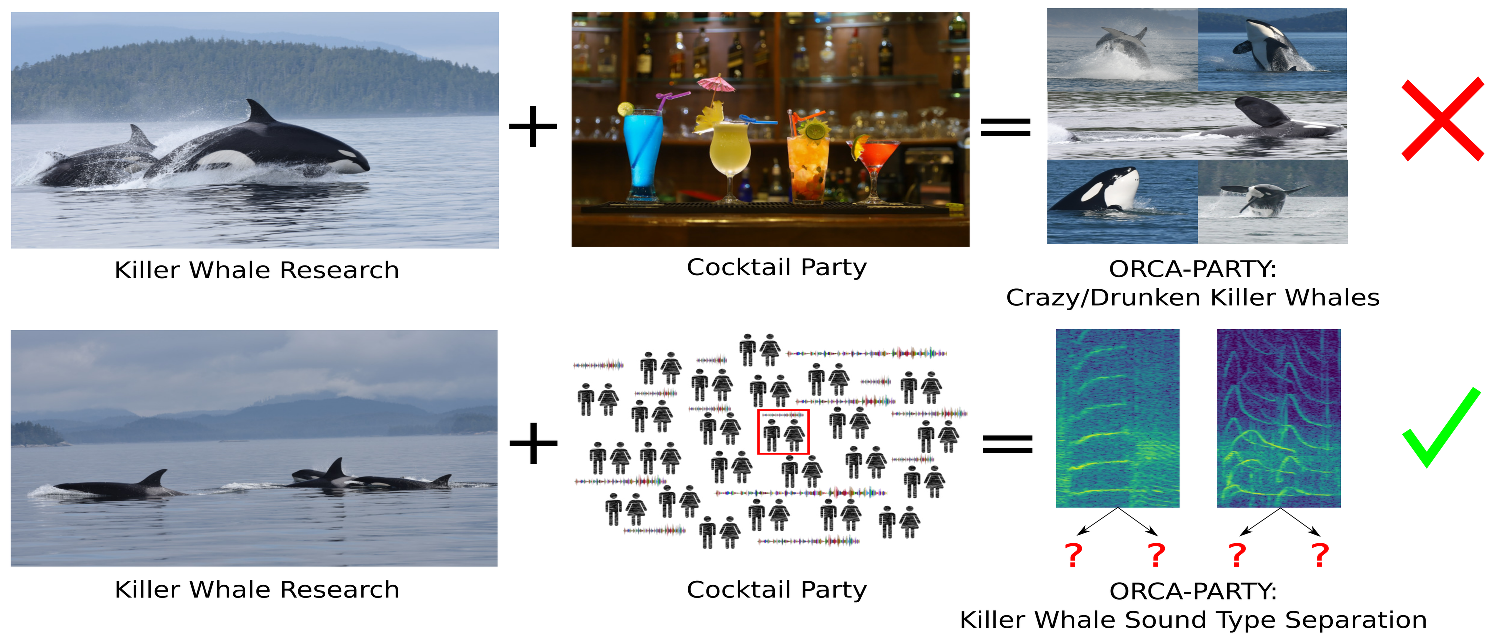
AUD-36.5

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

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

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

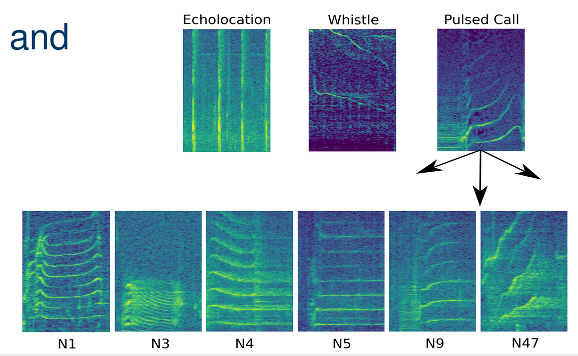
INTRODUCTION



Introduction – Killer Whale [1, 2, 3, 4, 5]

The largest member of the dolphin family – the Killer Whale (*Orcinus Orca*) – lives in stable (family-based) social units of several individuals, and produces three types of vocalization:

- *Echolocation Clicks* – short pulses used for navigation and object localization
- *Whistles* – narrow-band signals primarily used within close-range interactions
- *(Discrete) Pulsed Calls* – most common, stereotyped and repetitive vocal activities with a wide diversity of distinctive tonal properties/categories *(Call Types)*



Source: Killer whale images taken from FIN-PRINT [3], Copyright Jared Towers & Gary J. Sutton, Other Images, Pixabay License – taken from https://pixabay.com/ – and recreated

MOTIVATION & CHALLENGES

Motivation – Killer Whale Call Type Classification

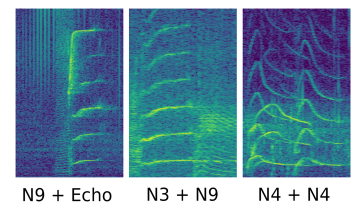
Killer whale call types indicate a wide diversity of distinctive categories with significant inter- and intra-class spectral variations [2]. Large-scale, data-driven, and machine-based orca call type identification is imperative to gain deeper insights into orca communication

→ Machine-based call type recognition is highly affected by overlapping call type structures!

Motivation – Killer Whale Sound Type Separation

In particular, longer acoustic regions of orca communication, containing a large number of vocalization events in consecutive short time intervals, are essential for communication analysis

→ High probability of overlapping category-specific call events!



Challenges:

- Robust machine learning pipeline to process massive and noise-heavy data repositories
- Limited knowledge about entire inter-/intra killer whale call type variations
  → combinatorial and spectral diversity
- 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

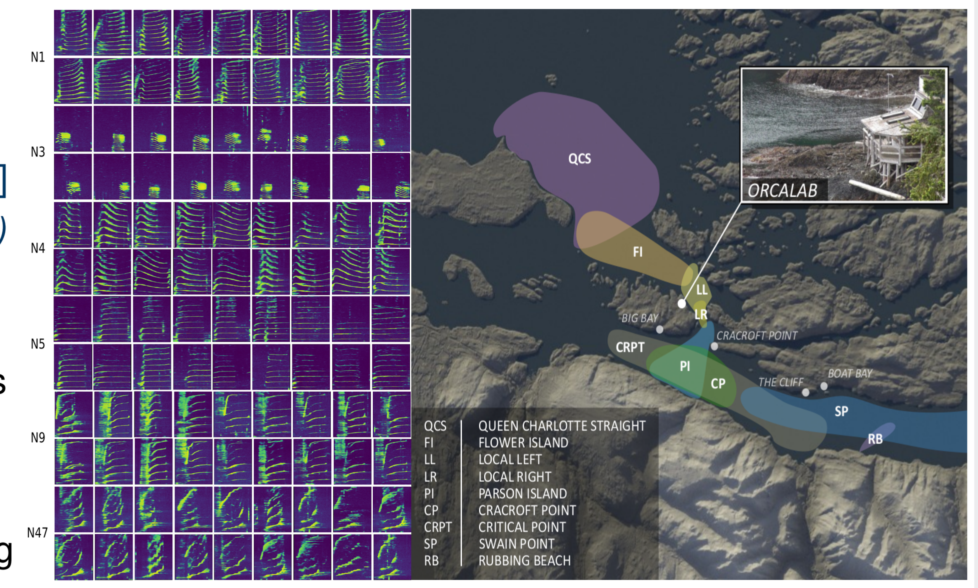
DATA MATERIAL

Killer Whale Sound Type Archive (KWSTA)

Large-scale, data-driven, and machine-generated orca sound type repository, consisting of three sub-archives, which are the result of applying machine (deep) learning algorithms [2, 4] to the Orchive (≈20,000 h underwater recordings)

- *ORCA-SLANG Call Type Data Corpus (OSDC)*
  → 235,369 orca samples, split across 6 call types
- *Echolocation Repository (ELRP)*
  → 9,382 echo events, identified via ORCA-TYPE [4]
- *ORCA-SLANG Unknown Signal Repository (OSUR)*
  → 2,101 excerpts of either so far unseen/unknown orca sounds or background noise

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



Call Type Data Corpus (CTDC)

Human-annotated dataset including 514 non-overlapping orca call type events, unequally split and categorized into 12 distinct classes [4, 6, 7]

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

Source: Images taken from ORCA-SPOT [1], ORCA-SLANG [2], and from the DeepAL 2017–2019 expedition image collection (copyright Anthro-Media)

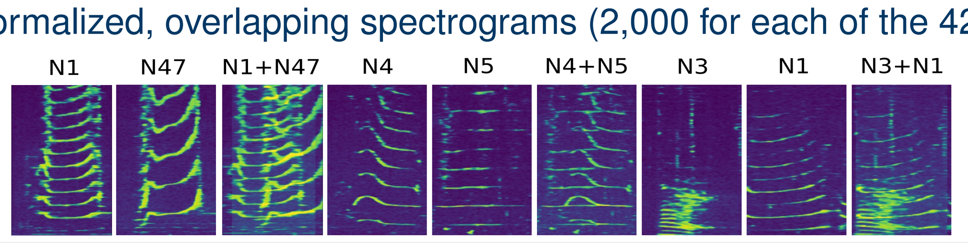
METHODOLOGY – DATA PROCESSING

Multi-Stage Data Preprocessing Procedure [1, 6]

- Conversion to mono, resampling to 44.1 kHz, and STFT (window/step ≈ 100 ms/10 ms) to build a F×T (Frequency×Time) decibel-converted power-spectrogram
- Orca detection algorithm [6], to return 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) → 256×128 0/1-dB-normalized spectrogram

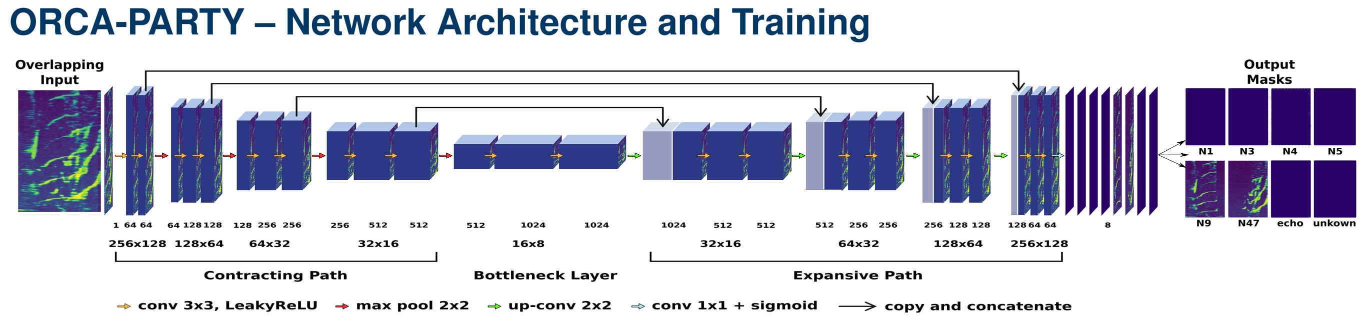
Multi-Stage Overlapping Data Generation Procedure

- 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 $\delta \in [0.64s, 1.28s]$
- Randomly sub-sampling a temporal context of 1.28 s (T = 128) and 0/1-min/max-normalization
- *ORCA-PARTY Overlapping Dataset (OPOD)* → 84,000 256×128-large, 0/1-min/max-normalized, overlapping spectrograms (2,000 for each of the 42 combinations)



METHODOLOGY – NETWORK & EXPERIMENTS

ORCA-PARTY – Network Architecture and Training



- Network Input: 256×128-large, 0/1-min/max-normalized, overlapping signals from the *OPOD* – Train: 58,800 – 70 %, Dev: 12,600 – 15 %, Test: 12,600 – 15 %
- Network Output: 8 category-specific activated segmentation masks

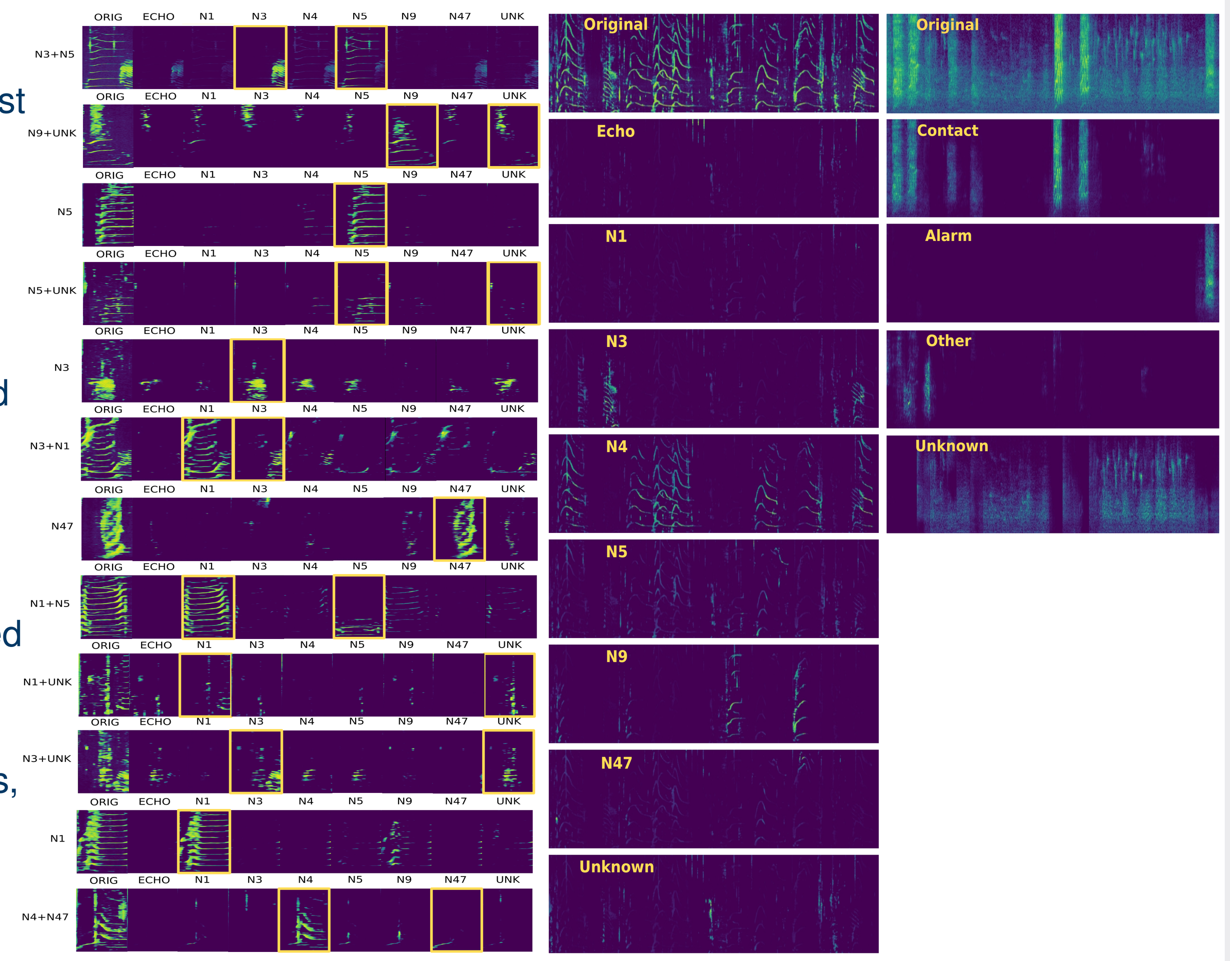
Experiments

- Visual inspection and classification of the network output masks from the unseen *OPOD* test set (ignore the “unknown” class → 8,400 out of 12,600 test events)
- *ORCA-TYPE* [4] was trained on the denoised (ORCA-CLEAN [6]) *CTDC* mask-specific data, with & 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-by-frame over pre-segmented/-denoised excerpts $\Psi \in [10.0s, 30.0s]$ of the unlabeled *DLFD* (*O-WP* vs. *O-BL*) → Classification hypotheses!
- Model transfer to a bird species, named Monk parakeets *(Myiopsitta monachus)*

RESULTS & DISCUSSION

Results

- Visualizations from the unseen *OPOD* test set (overlapping input spectrogram vs. class-based separation output)
- *O-WP* Classification accuracy ≈86.0 % (8,400 overlapping *OPOD* test samples)
- *O-BL* vs. *O-WP* average classification accuracy ≈96.0 % vs. ≈94.5 % (dev) and ≈94.5 % vs. ≈93.0 % (test) with respect to the non-overlapping *CTDC* data
- Classification hypotheses on the entire DLFD archive amount to 39,569 (*O-BL*) vs. 51,684 (*O-WP*) orca events distributed across 7 categories (increase of ≈30 %)
- *ORCA-PARTY*, trained on 3,000 (noisy) overlapping monk parakeet spectrograms, derived from 3,251 human-annotated events across 4 classes (contact, alarm, other call, different songbirds/noise)



REFERENCES

[1] 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.

[2] C. Bergler, M. Schmitt, A. Maier, H. Symonds, P. Spong, S. R. Ness, G. Tzanetakis, and 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.

[3] C. Bergler, A. Gebhard, J. Towers, L. Butyrev, G. Sutton, T. Shaw, A. Maier, and E. Nöth, “FIN-PRINT A Fully-Automated Multi-Stage Deep-Learning-Based Framework for the Individual Recognition of Killer Whales,” *Scientific Reports*, vol. 11, p. 23480, 12 2021.

[4] 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, pp. 274–286, Springer, 2019.

[5] 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.

[6] 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.

[7] 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)*, pp. 8231–8235, IEEE, May 2019.

ACKNOWLEDGEMENTS

DFG Deutsche Forschungsgemeinschaft

Thanks to the German Research Council (DFG) for funding and “The Paul G. Allen Frontier’s 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 orca-specific data archive, and Simeon Smeele (Max-Planck-Institute), for giving us permission to the Monk parakeet data.

CONTACT

Christian Bergler, M. Eng.

Pattern Recognition Lab
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
☎ +49 9131 85 27872
✉ christian.bergler@fau.de
https://lme.tf.fau.de/person/bergler/
https://github.com/ChristianBergler