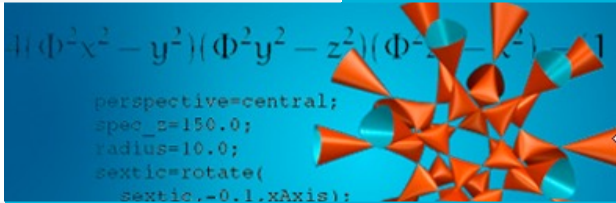
Philipps Universität Marburg

Deep Learning for Recognizing Bat Species and Bat Behavior in Audio Recordings

Markus Vogelbacher, Hicham Bellafkir, Jannis Gottwald, Daniel Schneider, Markus Mühling, and Bernd Freisleben

University of Marburg, Germany

vogelbacher@informatik.uni-marburg.de



Motivation

- Bats (*Chiroptera*) are excellent indicators for ecosystem health
- Bats emit different sounds to orient themselves and to communicate
- Monitoring bat populations is a very tedious task
- Automated methods are required
- Behavior analysis offers further insights

Audio Processing

- Time expansion
  → expands time domain by a factor of 10
  → reduces frequencies by a factor of 10

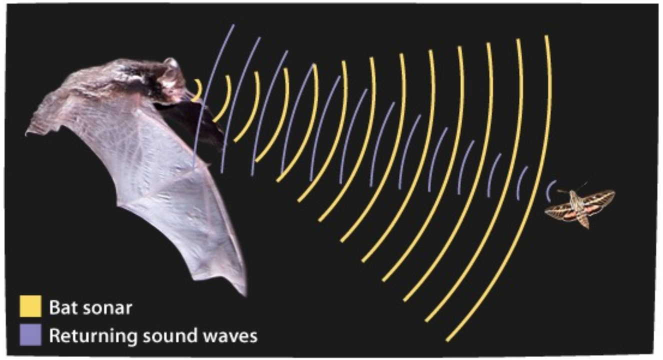
- Different ways to input audio to machine learning models
  - Raw audio
  - Linear spectrograms
  - Mel-scaled log spectrograms
  - Learnable filters (e.g., LEAF [1])

Bat Calls

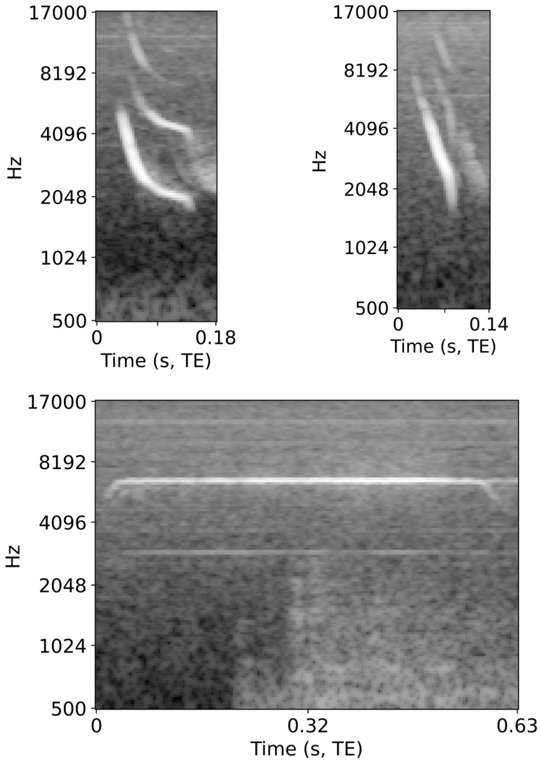
- Three main classes of bat calls
  - Echolocation calls for orientation
  - Feeding buzzes for hunting prey
  - Social calls for communicating with conspecifics
- Echolocation calls are widely used to determine the corresponding species

Echolocation Calls

- Bats emit and receive ultrasonic sounds to orient themselves
- Usually short pulses separated by longer periods of time

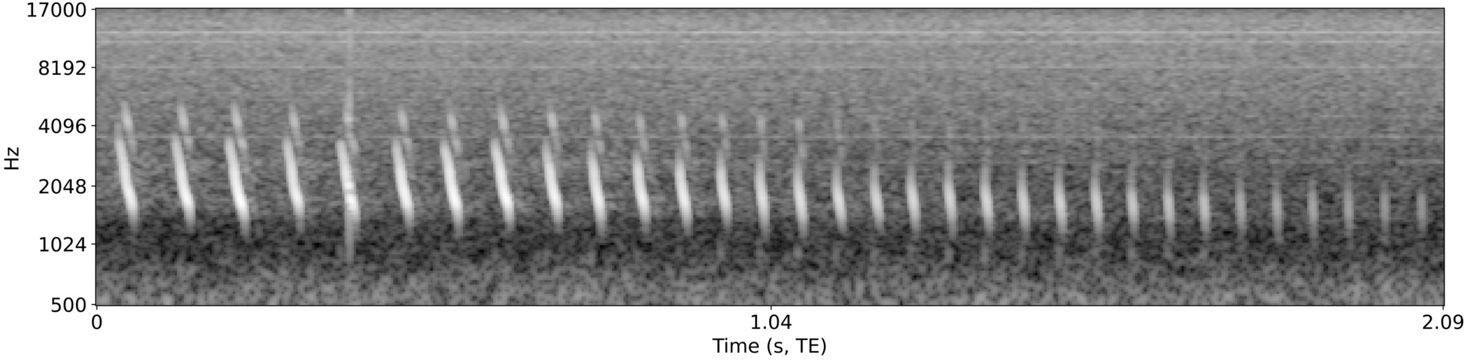


© Arizona Board of Regents / ASU Ask A Biologist. [Link]



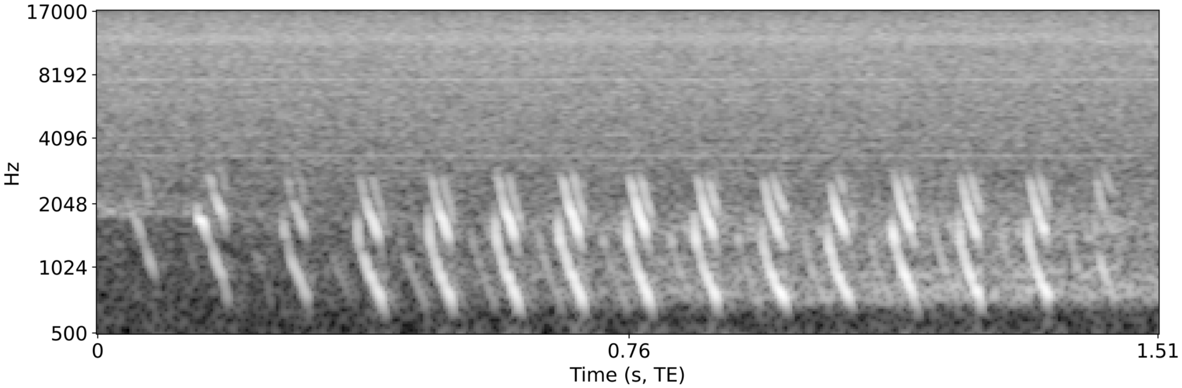
Feeding Buzzes

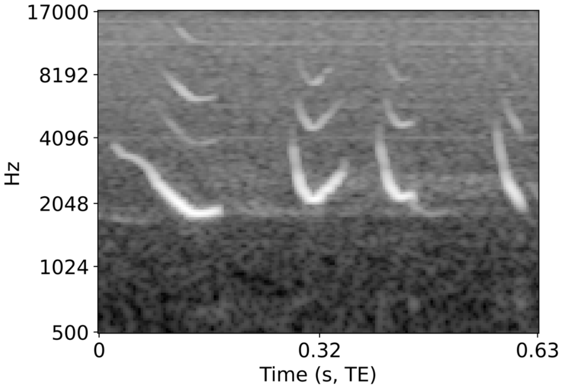
- Used to precisely locate prey while hunting
- Fast and accelerating sequence of ultrasonic calls
- Followed by an attempt to capture the target



Social Calls

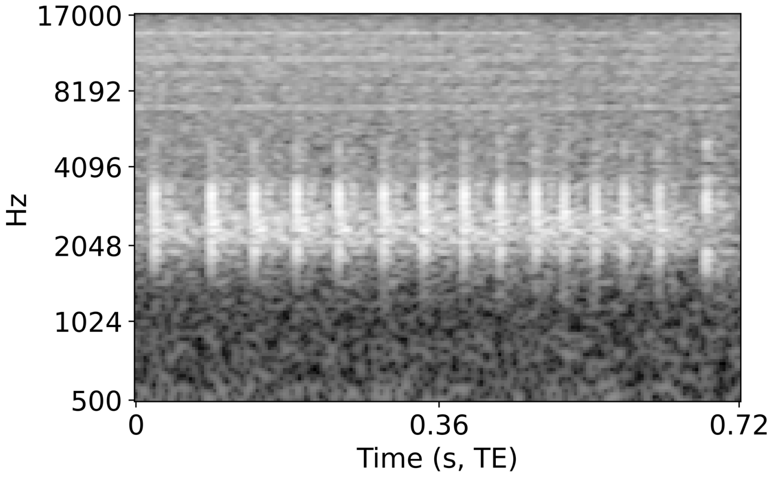
- Often audible for humans
- Great variety due to a wide range of applications
- More complex than other call types





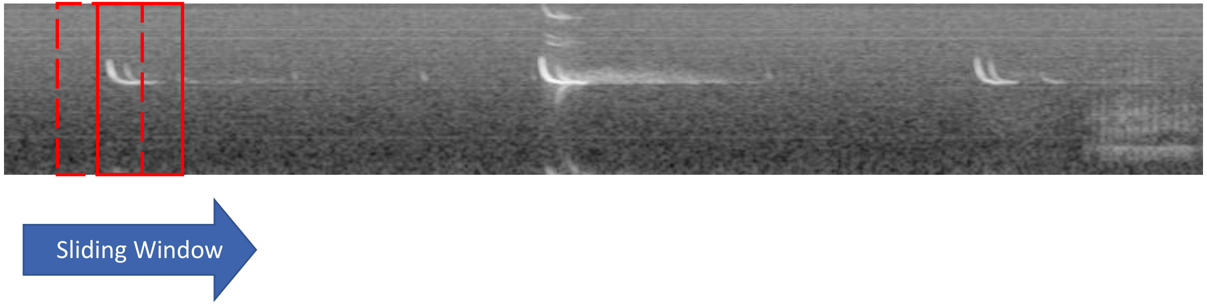
Challenges

- Different call lengths
- Similar calls
- Noises (e.g., crickets)



Related Work

- State-of-the-art approaches use deep learning (e.g., [2,3,4])
- Usually sliding windows and classifying the corresponding content [2,3]
- Not well suited for different call lengths



Our Approach

- Object detection in spectrograms to capture the boundaries of each call precisely
- 3 classes (i.e., behaviors)
  - Echolocation Call
  - Feeding Buzz
  - Social Call
- Species recognition with 19 species living in Europe and Northern Africa

Pre-processing

- Time-expand (TE) all audio recordings by a factor of 10
- Resample all audio recordings to 96 kHz
- According to the Shannon-Nyquist sampling theorem, frequencies up to 48 kHz (TE) can be captured
- Mel-spectrograms are used as a visual representation

Spectrograms

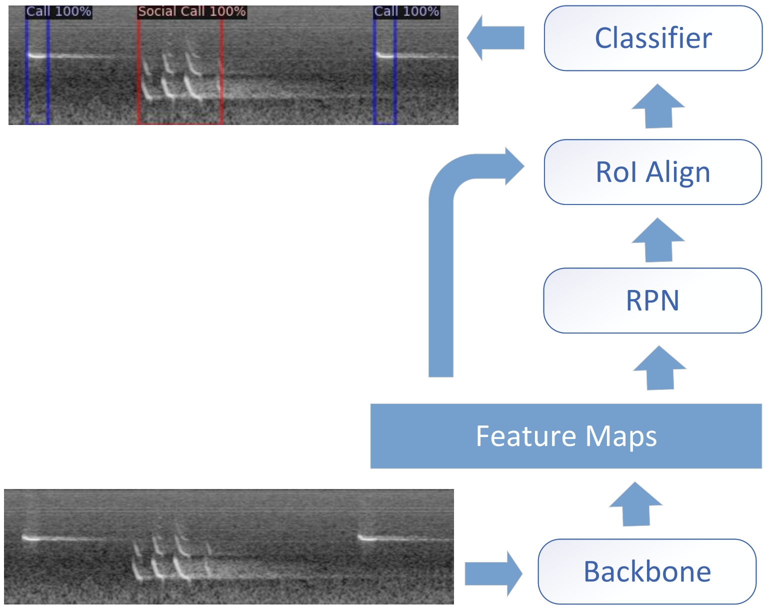
- Generate Mel-spectrograms as our input
  - 128 Mel bins
  - Window size of 23 ms (TE) and an overlap of 84.5%
  - 500 Hz to 17 kHz (TE) are considered

- Resulting in spectrograms of 2777 x 128 px for a 1s (10s TE) audio snippet

Architecture

- Faster-RCNN approach with different backbones
  - ResNet-50 + FPN
  - ResNeXt-101 + FPN
  - VitDet-Base [5]

FPN: Feature Pyramid Network



Experiments

Two data sets with hold-out test sets:

- Tierstimmenarchiv[1] (TSA)
  - Recorded on tape and digitized
  - 30,798 bounding boxes
  - Species and behavior annotations

Tierstimmenarchiv

- Audio exploratories
  - Passively recorded with AudioMoth[2] devices
  - 4,259 bounding boxes
  - Only behavior annotations

[1] https://tierstimmenarchiv.de
[2] https://www.openacousticdevices.info

Results

- Bat behavior recognition on TSA data set
- Average Precision @ IoU=0.5

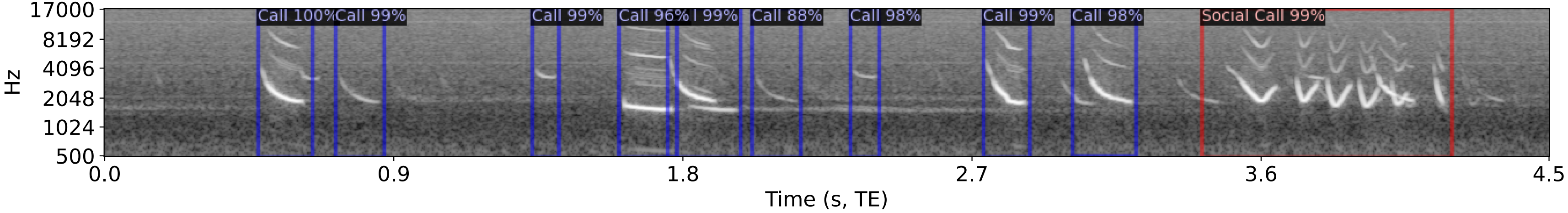
Method	Echo Call	Feeding Buzz	Social Call	Mean (mAP)
ResNet-50	0.975	0.953	0.952	0.960
ResNeXt-101	0.978	0.923	0.946	0.949
ViTDet-Base	**0.984**	**0.983**	**0.956**	**0.974**

Results

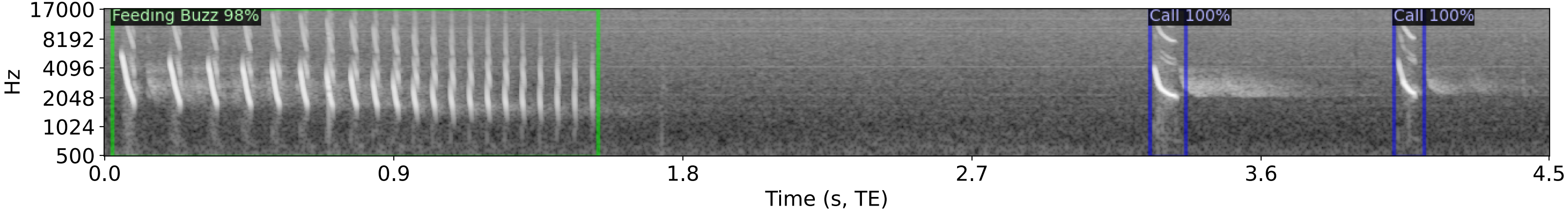
- Bat behavior recognition on Audio exploratories data set
- No evaluation of feeding buzzes due to lack of instances
- Average Precision @ IoU=0.5

Method	Echo Call	Social Call	Mean (mAP)
ResNet-50	**0.958**	0.913	0.936
ResNeXt-101	0.949	0.909	0.929
ViTDet-Base	0.957	**0.934**	**0.946**

Results



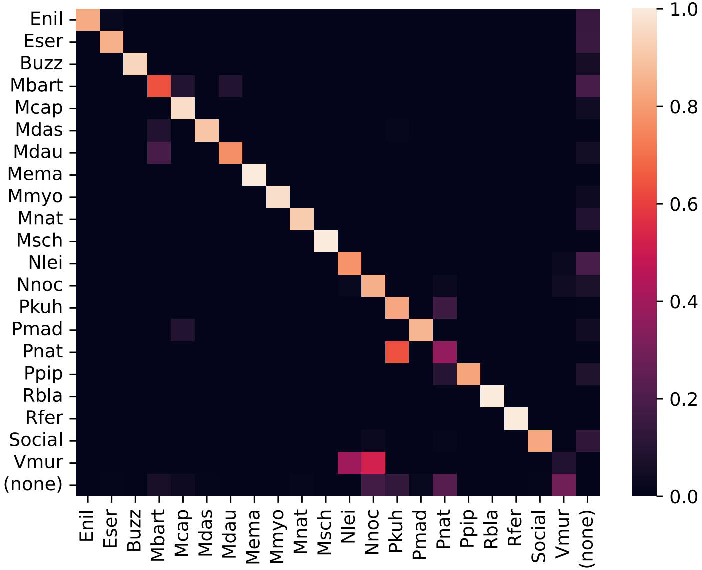
Call 100%
Call 99%
Call 99%
Call 96%
l 99%
Call 88%
Call 98%
Call 99%
Call 98%
Social Call 99%
17000
8192
4096
2048
1024
500
Hz
0.0
0.9
1.8
2.7
3.6
4.5
Time (s, TE)



Feeding Buzz 98%
Call 100%
Call 100%
17000
8192
4096
2048
1024
500
Hz
0.0
0.9
1.8
2.7
3.6
4.5
Time (s, TE)

Results

- Bat species recognition on TSA data set



- Mean Average Precision @ IoU=0.5

Method	mAP
ResNet-50	0.803
ResNeXt-101	0.824
VitDet-Base	**0.862**

Results

- Comparison to a state-of-the-art sliding window method
- TSA data set

Method	Echo Call Detection (AP)	Species Recognition (mAP@IoU=50%)
[3]	0.722	0.806
VitDet-Base	**0.988**	**0.862**

Conclusion

- Bat call recognition with object detection
  - Precise detection of call boundaries
  - Improves classification performance based on echolocation calls
- First approach to automated bat behavior recognition
- State-of-the art bat species recognition performance
- Up to 97.4% mAP in behavior recognition and up to 86.2% mAP in species recognition

Future Work

- Use smaller architecture to facilitate execution on edge devices
- Take all call types into account for species recognition
- Classify social calls into subclasses
- Use self-supervised approaches

References

[1] N. Zeghidour et al., “LEAF: A Learnable Frontend for Audio Classification”, Int. Conf. on Learning Representations (ICLR), 2021.

[2] O. Mac Aodha et al., “Bat Detective - Deep Learning Tools for Bat Acoustic Signal Detection”, PLOS Computational Biology, 2018.

[3] H. Bellafkir et al., “Bat Echolocation Call Detection and Species Recognition by Transformers with Self-Attention”, Int. Conf. on Intelligent Systems and Pattern Recognition (ISPR), 2022.

[4] I. Zualkernan et al., “A Tiny CNN Architecture for Identifying Bat Species from Echolocation Calls”, Int. Conf. on Artificial Intelligence for Good (AI4G), 2020.

[5] Y. Li et al., “Exploring plain vision transformer backbones for object detection”, Eur. Conf. on Computer Vision (ECCV), 2022.