PAMAI : Passive Acoustic Monitoring using AI

Arthur **ZUCKER**[1,2,3], Clément **APAVOU**[1,2,3], Masato **HASHIZUME**[2,3], Nobuhito **OHTE**[2], Anja M. **COLLAZO**[2,3], Christian E. **VINCENOT**[2,3]
[1]Sorbonne Science University's engineering School, Department of Applied mathematics and computer science, Paris, France
[2]Biosphere Informatics Laboratory, Graduate School of Informatics, Kyoto University Department of Social Informatics, Japan
[3]Island Bat Research Group, Kyoto, Japan, www.batresearch.net
This project was supported by a research grant to C.E. Vincenot by Pantagonia Inc. and the Tides foundation (Ref. 1908-59458)

1. Introduction

Flying foxes are globally **threatened** phytophagous bats [1]. which mainly feed on fruits and flowers thus providing **seed dispersal** and pollination services important for **forest regeneration**, especially in insular ecosystems. However, below a certain population density threshold, flying foxes stop competing for fruiting trees and tend to stay on the same tree under the crown of which feces and dispersed seeds have a very low probability of germinating. It is important to keep the number of individuals **above a critical threshold** necessary to provide ecosystem services. Hence the need to monitor populations of flying foxes, which is also necessary to support ecological studies and conservation programs. Passive acoustic monitoring in natural environments, albeit challenging, might offer an adequate solution to this problem. In this conference, we propose **PAMAI**, a Passive Acoustic Monitoring pipeline based on Artificial Intelligence, which aims at facilitating monitoring and **replacing invasive capture-mark-recapture** and by allowing density population estimation based on the identification of speakers.

2. Research Goals

Assessing activity patterns and monitoring population dynamics in flying foxes can be challenging, especially in species living in densely forested mountains or forming only small colonies. This research aims at providing a new AI-based Passive Acoustic Monitoring (PAM) system to automatically detect flying fox vocalizations and extract ecological information, such as behavior (e.g. aggression, mating attempt, etc), sex and possibly identity of the emitter. Ultimately, this shall support bat conservation.

Solitary Ryukyu flying fox *Pteropus dasymallus* hanging in a tree. Credits: C.E. Vincenot (IBRG)

5. Conclusion & Discussions

We proposed a novel approach to Passive Acoustic Monitoring (PAM) in wildlife surveys using state of the art deep learning methods applied to audio processing. The proposed pipeline yielded **promising results** for the audio segmentation task, which reinforces the potential uses of the pipeline on other animals. Moreover, we introduced possible ideas and improvements about **how to tackle both the speaker identification** and the audio classification task, using feature embedding. Future development will include the test of other AI systems for audio classification (gender / context and speaker), and transfer learning of face recognition algorithms to voice recognition.

7. References

[1] C.E. Vincenot, F.B.V. Florens, and T. Kingston. Can we protect island flying foxes? *Science*, 355:1368–1370, 2017.

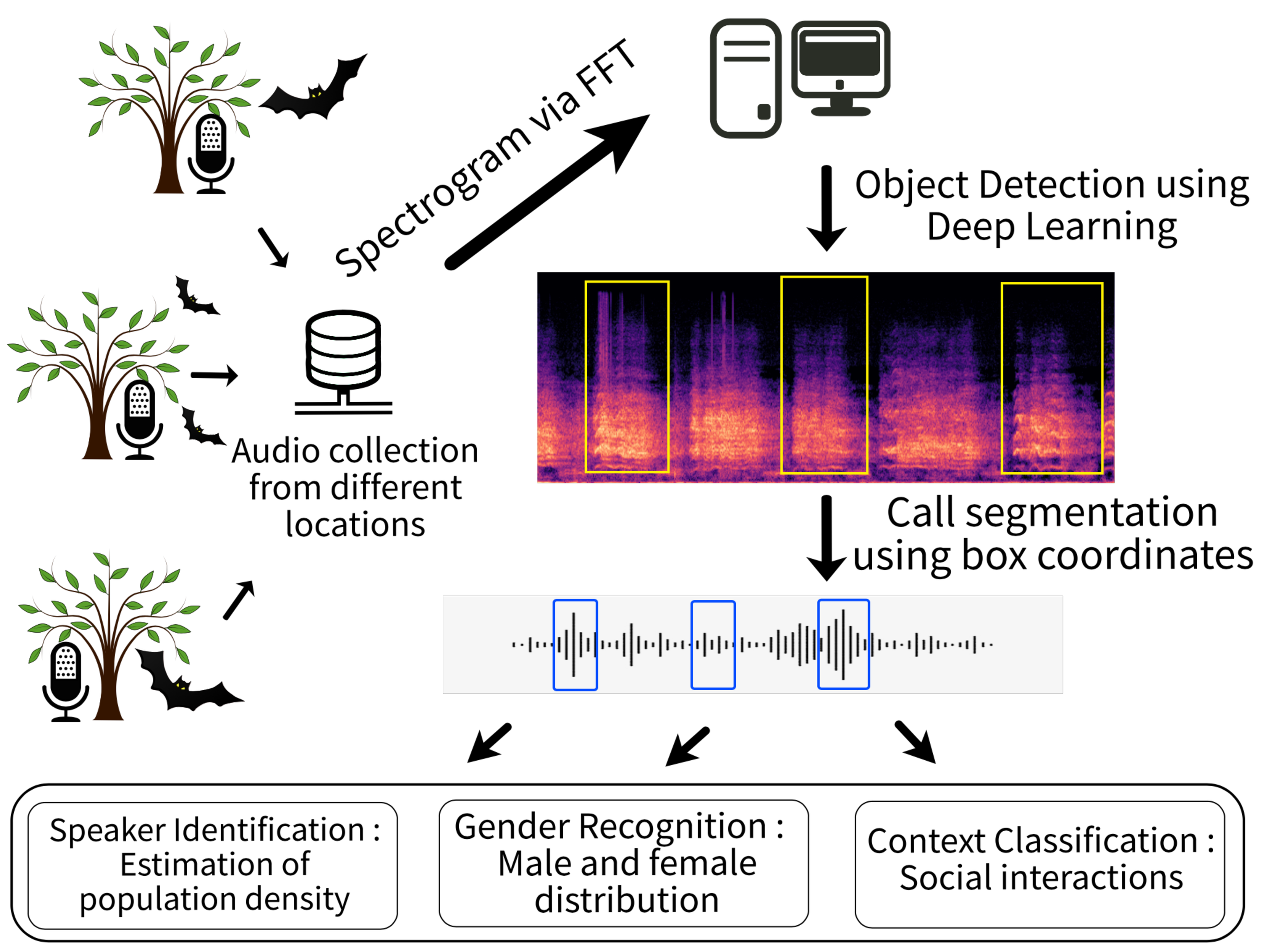
[2] Gregory R. Koch. Siamese neural networks for one-shot image recognition. 2015.

[3] Yosef Prat, Mor Taub, Ester Pratt, and Yossi Yovel. An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Scientific data*, 4:170143, 10 2017.

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020.

3. Materials & Methods

PAMAI's pipeline can be decomposed in the following steps :



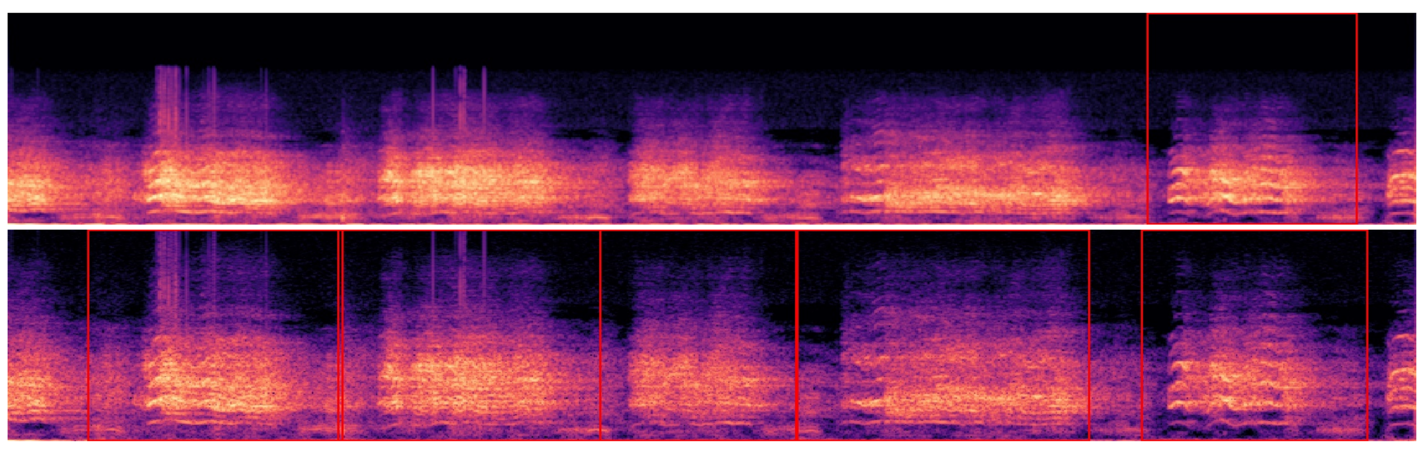
1. Gathering : natural recordings from different locations are gathered and retrieved.
2. Call detection : each audio file is converted to a spectrogram using Fast Fourier Transforms. A trained real-time object detection deep neural network then detects the calls on the spectrogram, outputting pixel coordinates that are then converted back to audio time stamps.
3a. Speaker recognition : using a siamese neural network [2], each individual is tentatively identified as belonging to the database or not. A new individual is added without the need of computation. Hence the relative number of individuals and their movements can be potentially inferred from the location of the recordings.
3b. Gender classification : the gender of the emitter is recognized using a vanilla Multi Layer Perceptron (MLP) and MFCC (Mel Frequency Cepstral Coefficients).
3c. Context classification : the context of emission can also be classified using deep learning networks based on a known ethogram (incl. aggression, mating attempt, etc).
In order to train our neural networks, we used two different datasets of audio recordings :

- *Pteropus dasymallus* dataset : composed of 3448 hand labeled audio files. Annotated with the start and the end of the call, the gender (male, female, juvenile, unknown) and the context of the call.
- *Rousettus aegyptiacus* : a publicly available dataset containing 220 000 label audio files (Prat et al. 2017 [3]). Annotated with the start and the end of the call, the gender, the context of the call and the call emitter id.

We used Google Colab's online GPU services to train our deep learning architectures.

4. Results

We trained YOLOv4 [4] on a sample of the *Rousettus aegyptiacus* dataset, obtaining a mean average precision with IoU threshold of 0.50 (mAP/50) of 93.2%. The IoU metric represents the Intersection over Union of the areas of the ground truth and the predicted bounding box. We then used transfer learning to evaluate qualitatively the method on the other dataset. This did not require any retraining, but rather some image processing.



On the left, the result of our trained network on a natural environment recording with (bottom figure) and without (top figure) transfer learning are shown. Transfer learning increases the accuracy of call detection.
Our gender classification MLP reached a 78% accuracy, which is encouraging given the simplicity of the model used.
Finally, we trained siamese neural networks on spectrograms and images MFCC to attempt the speaker recognition task, which did not yield the expected results: the 7-way evaluation stagnated at 38%. We deduced that our approach was mistaken, and set out to explore different solutions, taking inspiration from face recognition. We are currently working on a solution.

	Call detection	Speaker recognition	Gender classification
Method	Object detection	Siamese Neural Network	Multi Layer Perceptron
Database	*Rousettus aegyptiacus*	*Rousettus aegyptiacus*	*Pteropus dasymallus*
Features	Spectrograms	Spectrograms and MFCC images	MFCC coefficients
Result	mAP/50 of 93.2%	7-way evaluation of 38%	Accuracy of 78%