Making Weak Supervision Interactive: Exploring Transfer from Sound Libraries to Passive Acoustic Monitoring Data

Novruz Mammadli[1,2] , **Rida Saghir**[1,2] , **Kanwar Ammar Ali**[1] , **Prathmesh Doddanawar**[1] , **Thiago S. Gouvêa**[1,2] , **Daniel Sonntag**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Oldenburg, Germany
[2]Carl von Ossietzky University of Oldenburg, Oldenburg, Germany
{first.last}@dfki.de,

Abstract

Passive Acoustic Monitoring (PAM), an increasingly popular method for wildlife monitoring, generates large volumes of data whose analysis depends on instance-level annotations that are costly to obtain. Archival sound collections provide weak labels that lack temporal localisation. In prior work, we demonstrated that Multiple Instance Learning (MIL) can extract approximate event locations from weakly labelled PAM data, suggesting it may be applied to sound collection data.

This demo operationalizes that approach within an interactive workflow that connects weakly annotated sound collections to downstream PAM deployment. The system supports configurable MIL-based localisation, lightweight interactive refinement, and transfer to an independent PAM dataset.

We carried out a preliminary evaluation with an actual sound library from a museum collection and a benchmark PAM dataset. Results confirm that weakly annotated sound collections can serve as a viable training signal for downstream PAM detection and illustrate differences between alternative MIL instantiations under real transfer conditions. (Video available at https://cst.dfki.de/projects-weak-supervision-demo)

1 Introduction

Passive Acoustic Monitoring (PAM) has emerged as a significant methodology for wildlife monitoring, enabling continuous and uninterrupted observation of ecosystems [Sugai *et al.*, 2019]. Advances in sensor technology have greatly facilitated acoustic monitoring. However, analysing the large amounts of data obtained from PAM in the field has become a problem for researchers and biodiversity specialists [Gouvêa *et al.*, 2023]. Creating robust labels to identify sound events using supervised machine learning methods that require instance-level descriptions (start and end times) is labour-intensive and requires field expertise.

Museum and community-run sound collections, such as FNJV[1], the Macaulay Library[2], and Xeno-Canto[3], often contain weak descriptions indicating the species present in the recording at the file level, without temporal localisation or strong labels. These recordings, which contain such weak descriptions and are a valuable source of information, are recorded for archival purposes rather than for machine learning workflows. Therefore, target signals are often sparse, may co-occur with other signals (e.g., other species, environmental noise, or the naturalist recording metadata such as species name and a description of the recording context by speaking into the microphone). To use these collections to train acoustic detectors, descriptive vocalizations must first be localized in each recording. Previous work has addressed this challenge using heuristic signal-strength estimators to extract high-energy segments from weakly labelled focal recordings [Sprengel *et al.*, 2016; Kahl *et al.*, 2021]. Such approaches rely primarily on signal salience and assume that target vocalizations exhibit higher signal-to-noise ratios than the background. As a result, they do not explicitly model the informational content of the signal and may struggle when vocalizations are weak, overlapping, or embedded in complex soundscapes. In contrast, our workflow formulates localisation as a learning problem under weak supervision and integrates it into an interactive refinement loop, enabling adaptation beyond purely signal-based segmentation.

In our previous work, we demonstrated that weakly supervised learning methods using Multiple Instance Learning (MIL) can derive temporally localized predictions using weak labels and improve species-level detection performance in PAM datasets [Troshani *et al.*, 2024a]. We further investigated the impact of model architectures and selected pooling functions by developing an interactive interface prototype [Troshani *et al.*, 2024b]. However, these studies used weak labels derived from strongly labelled PAM data for both training and evaluation. Furthermore, the interactive portion of the interface did not involve iterative improvement of the learned instance-level models. In this demo, we extend it by presenting a scalable, interactive workflow that bridges the gap between real-world sound collections and PAM. Specifically, in this demo, we do the following:

[1]https://www2.ib.unicamp.br/fnjv/
[2]https://www.macaulaylibrary.org/
[3]https://xeno-canto.org/

- apply MIL-based localisation to weakly labelled audio collections,
- transfer the resulting instance-level models to a PAM dataset,
- present an active learning-based optimisation phase that iteratively improves localisation by using strong labels generated through user interaction, and
- provide a preliminary evaluation of transfer performance and expose refinement mechanisms for interactive investigation.

2 System Overview

Our workflow combines weakly supervised localisation using Multiple Instance Learning (MIL) with interactive refinement through active learning. The goal is to bootstrap instance-level acoustic detectors from weakly annotated sound collections and transfer them to target PAM datasets.

2.1 Weak Labels and Binary Multiple Instance Learning

Sound collection recordings are annotated at the file level with class-conditional presence statements of the form "Species c is present in this recording." These annotations do not provide onset and offset times and do not guarantee absence of other species. For a selected target species, we therefore adopt a binary formulation.

Following standard weakly supervised sound event detection (WSSED) approaches [Wang *et al.*, 2019], each recording is treated as a bag of temporal instances (frames or short windows). Frame-level predictions are produced by an instance-level model and aggregated into recording-level predictions via an average pooling function. If at least one instance contains the target vocalization, the bag is labelled positive. In prior work, we systematically compared pooling strategies and model architectures in this setting. [Troshani *et al.*, 2024a; Troshani *et al.*, 2024b]. In the present system, these components remain configurable but are not the primary focus.

2.2 Transfer via Pretrained Acoustic Embeddings

To enable efficient retraining during interaction, we leverage pretrained acoustic representations. In addition to end-to-end models such as TALNet [Wang *et al.*, 2019] and CDur [Dinkel *et al.*, 2021], we support frozen embeddings extracted from BirdNET [Kahl *et al.*, 2021] followed by a lightweight linear classifier. Prior results indicate that BirdNET embeddings transfer well to PAM datasets, including AnuraSet [Kath *et al.*, 2024; Troshani *et al.*, 2024b].

Using frozen embeddings provides two advantages in the interactive setting: (i) it reduces computational cost during iterative updates, and (ii) it stabilizes training under small amounts of newly labelled data [Yosinski *et al.*, 2014]. For these reasons, we implement interactive refinement with frozen embeddings. This design choice is critical for enabling real-time refinement within the interface.

2.3 Active Learning-Based Refinement

While MIL enables automatic localisation from weak labels, it does not achieve perfect temporal precision (table **??**). We therefore introduce an interactive refinement stage. After training an initial instance-level model using weak labels only, the system ranks candidate segments for a selected species according to an uncertainty-based query strategy. The user labels these segments in a binary presence–absence manner.

Because embeddings are frozen, only the linear classifier is updated during refinement. This allows rapid iterative retraining and immediate feedback. The refinement stage should serve two purposes: (i) improving localisation quality within the sound collection itself, and (ii) improving downstream recall when the detector is applied to an independent PAM dataset.

In contrast to traditional deep active learning pipelines, which often require repeated retraining of large models, our approach restricts adaptation to a lightweight classifier operating on fixed representations. This design makes interactive refinement computationally feasible while retaining the benefits of transfer learning.

3 Workflow

The presented system operationalizes a workflow for bootstrapping instance-level acoustic detectors from weakly annotated sound collections and transferring them to target PAM datasets. It supports both exploratory interaction and automated localisation, culminating in an iterative refinement stage.

3.1 Manual Region-of-Interest Search

As a diagnostic and exploratory entry point, the system allows users to upload a single recording and manually mark a region of interest (ROI). Acoustic embeddings are extracted for the selected segment and used to retrieve acoustically similar segments from a target PAM dataset via vector similarity search.

This mode serves two purposes: (i) it provides a rapid sanity check of the embedding space and transfer behaviour, and (ii) it offers a low-effort baseline for identifying candidate events. However, manual ROI search does not scale to large sound collections and therefore motivates the automated workflow described below.

3.2 Collection-Level Localisation via MIL

For scalable processing, users upload sound collections organized as one folder per target species. Each recording carries a weak, file-level label indicating the presence of the species.

The system trains an instance-level model under a binary MIL formulation. Users may configure:

- the instance-level model (e.g., TALNet or frozen BirdNet embeddings with a linear classifier),
- the pooling strategy used for aggregation,
- selected training hyperparameters.

Species	**TALNet** [Wang *et al.*, 2019]			**CDur** [Dinkel *et al.*, 2021]			**BirdNET** [Kahl *et al.*, 2021]		
	F1	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**
BOARAN	0.1207	0.1364	0.1082	0.0303	0.0156	**0.5411**	**0.1783**	**0.9180**	0.0988
DENMIN	0.4657	0.4581	0.4737	0.5782	0.4571	**0.7865**	**0.8027**	**0.9639**	0.6877
DENNAN	0.0193	0.7200	0.0098	0.0000	0.0000	0.0000	**0.1614**	**0.9943**	**0.0878**
LEPFUS	0.2125	0.1354	**0.4931**	**0.3782**	0.7255	0.2558	0.2604	**1.0000**	0.1497
SCIFUS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 1: Instance-level performance (F1, Precision, Recall) for five target species across three models trained on a sound collection (FNJV) and evaluated on a PAM dataset (AnuraSet).

After training, the model produces instance-level probabilities corresponding to candidate temporal segments within each recording. These localized predictions convert weak file-level labels into approximate instance-level annotations.

3.3 Active Learning-Based Refinement

Since MIL-based localisation is imperfect, the system supports an interactive refinement stage. For a selected species, candidate segments are ranked according to a query strategy (e.g., prediction uncertainty). The user labels segments in a binary presence–absence manner.

During refinement, only the instance-level model (e.g., lightweight classifier operating on frozen embeddings) is updated, enabling rapid retraining and immediate feedback within the interface. This iterative process allows users to inspect instance-level predictions interactively, supporting controlled experimentation with refinement strategies. This refinement stage is expected to improve temporal localisation within the sound collection and yield an updated detector that can be applied to independent PAM recordings.

3.4 Transfer to Target PAM Dataset

The refined detector (instance-level model) is evaluated on a target PAM dataset to assess downstream performance. By allowing a comparison between models trained with weak labels only to those refined with interactive strong labels, the system enables studying how annotation effort invested at the collection level translates into improved recall and instance-level detection performance in real-world monitoring scenarios.

4 Preliminary Evaluation

The goal of this evaluation is not to establish state-of-the-art performance, but to assess transfer behaviour when training instance-level detectors on weakly annotated sound collections and applying them to a target PAM dataset.

We selected five anuran species for which (i) weakly annotated recordings from the FNJV sound collection were available and (ii) strong annotations exist in a subset of the AnuraSet PAM dataset [Cañas *et al.*, 2023].

For each species, we trained three instance-level detectors: TALNet [Wang *et al.*, 2019], CDur [Dinkel *et al.*, 2021], and a linear classifier applied to BirdNET embeddings [Kahl *et al.*, 2021]. All models were trained using weak file-level labels from the sound collection and evaluated on AnuraSet.

We report instance-level F1 scores computed at 3-second resolution, together with precision and recall to characterise localisation behaviour under transfer.

Across the five species, models trained on weakly annotated sound collections achieved non-trivial localisation performance on AnuraSet, confirming that archival collections can serve as a viable training signal for downstream PAM detection. Performance varied substantially across model architectures and species. In particular, embedding-based models using BirdNET representations yielded consistently high precision, while recall remained sensitive to species characteristics and domain shift. For a species with rare vocalisations (*Scinax fuscomarginatus*, SCIFUS), all models struggled to detect events reliably, highlighting the challenges of transferring detectors across recording conditions.

These results illustrate that both representation choice and MIL instantiation significantly influence downstream transfer performance. The presented system enables such comparisons within an interactive workflow, supporting systematic investigation of weak supervision and model transfer under realistic data constraints.

5 Conclusion

This demo presents an integrated workflow for training instance-level acoustic detectors from weakly annotated sound collections and transferring them to PAM datasets. Building on prior work on weakly supervised localisation, the system operationalizes configurable MIL training and interactive inspection within a unified environment.

Preliminary results confirm that archival sound collections can provide a meaningful training signal for downstream detection, while revealing substantial variability across model architectures and species under real transfer conditions. The platform further enables interactive active learning-based refinement of localisation within sound collection files, allowing users to iteratively inspect and adjust instance-level predictions. By exposing model configuration, transfer behaviour, and refinement mechanisms within an interactive interface, the system provides a practical basis for systematic investigation of weak supervision in acoustic monitoring.

Acknowledgements

This research is part of the Computational Sustainability & Technology project area[4], and has been supported by the

[4]https://cst.dfki.de/

Lower Saxony Ministry of Science and Culture (MWK) in zukunft.niedersachsen program, the Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16IW26002 (AMENABLE), and the Endowed Chair of AAI at the Carl von Ossietzky University of Oldenburg.

Focal recordings kindly provided by Simone Dena (FNJV).

References

[Cañas *et al.*, 2023] Juan Sebastián Cañas, María Paula Toro-Gómez, Larissa Sayuri Moreira Sugai, Hernán Darío Benítez Restrepo, Jorge Rudas, Breyner Posso Bautista, Luís Felipe Toledo, Simone Dena, Adão Henrique Rosa Domingos, Franco Leandro de Souza, Selvino Neckel-Oliveira, Anderson da Rosa, Vítor Carvalho-Rocha, José Vinícius Bernardy, José Luiz Massao Moreira Sugai, Carolina Emília dos Santos, Rogério Pereira Bastos, Diego Llusia, and Juan Sebastián Ulloa. A dataset for benchmarking Neotropical anuran calls identification in passive acoustic monitoring. *Scientific Data*, 10(1):771, November 2023. 9 citations (Semantic Scholar/DOI) [2024-12-02].

[Dinkel *et al.*, 2021] Heinrich Dinkel, Mengyue Wu, and Kai Yu. Towards duration robust weakly supervised sound event detection. *CoRR*, abs/2101.07687, 2021.

[Gouvêa *et al.*, 2023] Thiago S. Gouvêa, Hannes Kath, Ilira Troshani, Bengt Lüers, Patricia P. Serafini, Ivan B. Campos, André S. Afonso, Sergio M. F. M. Leandro, Lourens Swanepoel, Nicholas Theron, Anthony M. Swemmer, and Daniel Sonntag. Interactive Machine Learning Solutions for Acoustic Monitoring of Animal Wildlife in Biosphere Reserves. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6405–6413, Macau, SAR China, August 2023. International Joint Conferences on Artificial Intelligence Organization.

[Kahl *et al.*, 2021] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, March 2021. 320 citations (Semantic Scholar/DOI) [2024-10-06].

[Kath *et al.*, 2024] Hannes Kath, Patricia P. Serafini, Ivan B. Campos, Thiago S. Gouvêa, and Daniel Sonntag. Leveraging transfer learning and active learning for data annotation in passive acoustic monitoring of wildlife. *Ecological Informatics*, 82:102710, September 2024.

[Sprengel *et al.*, 2016] Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. Audio Based Bird Species Identification using Deep Learning Techniques. *LifeCLEF 2016*, 2016. Meeting Name: Conference and Labs of the Evaluation Forum (CLEF) 2016.

[Sugai *et al.*, 2019] Larissa Sayuri Moreira Sugai, Thiago Sanna Freire Silva, José Wagner Ribeiro, and Diego Llusia. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, 69(1):15–25, January 2019. 346 citations (Semantic Scholar/DOI) [2024-09-15].

[Troshani *et al.*, 2024a] Ilira Troshani, Thiago S Gouvêa, and Daniel Sonntag. Leveraging Sound Collections for Animal Species Classification with Weakly Supervised Learning. In *3rd Annual AAAI Workshop on AI to Accelerate Science and Engineering (AI2ASE)*, Vancouver, Canada, 2024.

[Troshani *et al.*, 2024b] Ilira Troshani, Thiago S. Gouvêa, and Daniel Sonntag. Leveraging Weakly Supervised and Multiple Instance Learning for Multi-label Classification of Passive Acoustic Monitoring Data. In Andreas Hotho and Sebastian Rudolph, editors, *KI 2024: Advances in Artificial Intelligence*, pages 260–272, Cham, 2024. Springer Nature Switzerland.

[Wang *et al.*, 2019] Yun Wang, Juncheng Li, and Florian Metze. A Comparison of Five Multiple Instance Learning Pooling Functions for Sound Event Detection with Weak Labeling. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35, May 2019. 168 citations (Semantic Scholar/DOI) [2024-05-22] Conference Name: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ISBN: 9781479981311 Place: Brighton, United Kingdom.

[Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.