

Final Project Proposal: Few-Shot Detection of Bioacoustic Events

Vicente José Escarigo Miranda
UAM, Universidad Autonoma de Madrid
Lisbon, Portugal

Abstract—Few-shot learning is a promising paradigm for bioacoustic sound event detection, where only a small number of labeled examples are available for a target species or call type. This report proposes a modular and lightweight baseline for DCASE 2023 Task 5 (Few-shot Bioacoustic Event Detection). The method follows a general pipeline that converts raw audio into time–frequency representations, extracts frame-level embeddings with a shared encoder, performs support-guided scoring of query audio, applies temporal post-processing to form event predictions, and evaluates performance using the official event-based F-measure based on IoU and bipartite matching. The goal of this project is to understand the task constraints in depth and deliver a reproducible baseline system that can be extended with alternative modeling choices.

Index Terms—few-shot learning, bioacoustics, sound event detection, embeddings, prototypical methods, event-based evaluation

I. INTRODUCTION

Passive acoustic monitoring is increasingly used to study biodiversity and animal behavior, but labeling large volumes of field recordings remains expensive and often infeasible. Few-shot learning addresses this limitation by enabling detection and classification of a target sound category given only a few labeled exemplars. DCASE 2023 Task 5 focuses on few-shot bioacoustic event detection in realistic environments, requiring systems to detect sparse and short vocalisations in long recordings under strict data-use constraints.

This project aims to (i) study the task specification and evaluation protocol in depth, (ii) implement a clean end-to-end baseline system aligned with the official rules, and (iii) establish a reproducible framework for future experiments on representation learning, few-shot classification, and post-processing strategies.

II. TASK DESCRIPTION AND DATA

A. Task setup

The task is defined as a 5-shot detection problem. For each recording in the validation/evaluation sets, only the first five positive (POS) events of a single class of interest are provided as exemplars (support). Systems must then detect additional occurrences of the class in the remainder of the recording (query region). Each file must be treated independently, and the official evaluation ignores all time before the end of the fifth POS event.

B. Development set

The development data include a multi-source training set with multi-class annotations (POS/NEG/UNK) and a validation set with single-class annotations (POS/UNK) for a hidden class of interest per file. The recordings vary in sampling rate, acoustic conditions, and event sparsity, which makes robust detection challenging.

C. Annotation format and UNK labels

Training annotations provide event intervals with POS/NEG/UNK labels across classes, while validation/evaluation annotations provide POS/UNK for the single target class. UNK denotes ambiguous regions and is treated separately in evaluation to avoid penalizing predictions in uncertain segments.

III. EVALUATION METRIC

The task uses an event-based, macro-averaged F-measure. Predicted events are matched to ground-truth events using an intersection-over-union (IoU) criterion, followed by bipartite matching to enforce one-to-one pairing between predictions and references. Precision and recall are computed from matched events, and the final score is the macro-averaged F-measure over the evaluation set. The metric ignores the support portion of each file (up to the end of the fifth POS event) and handles UNK regions separately.

IV. PROPOSED BASELINE SYSTEM

A. Overview

The planned system follows a general strategy that does not commit to a specific architecture: (1) audio preprocessing, (2) time–frequency feature extraction, (3) embedding extraction with a shared encoder, (4) support-guided scoring of query frames, (5) temporal post-processing into event intervals, and (6) official evaluation.

B. Preprocessing and features

Audio is resampled if required, amplitude-normalized, and converted to log-mel spectrogram features. Feature extraction settings (window length, hop size, and number of mel bins) will be selected to preserve short event structure while keeping computational cost low.

C. *Embeddings and support-guided scoring*

A shared encoder maps spectrogram frames (or short context windows) into an embedding space. The support exemplars are used to build class representations (e.g., prototype means or simple templates). Query frames are scored by similarity to the support-derived representation, producing frame-level confidence values for the class of interest.

D. *Post-processing*

Frame-level scores are converted into event predictions using thresholding and temporal smoothing (e.g., median filtering), followed by merging consecutive detections and enforcing minimum event duration constraints. Post-processing is expected to strongly affect the precision–recall trade-off and event boundary quality.

V. EXPERIMENTAL PLAN

Experiments will be conducted on the official development set. Validation files will be processed under the same 5-shot constraint as the evaluation set. The initial goal is to reproduce the official baseline behavior and then run controlled ablations on (i) feature extraction choices, (ii) encoder capacity, (iii) similarity metrics, and (iv) post-processing parameters. Performance will be reported using the official event-based F-measure, alongside precision and recall for diagnostic analysis.

VI. CONCLUSION

This project proposes a reproducible baseline framework for DCASE 2023 Task 5 that emphasizes correctness, modularity, and computational efficiency. By implementing the full pipeline from audio to event predictions and official evaluation, the project will establish a strong foundation for exploring alternative encoders and few-shot scoring strategies while remaining aligned with the task constraints.

ACKNOWLEDGMENT

REFERENCES

- [1] DCASE 2023 Challenge, Task 5: Few-shot Bioacoustic Event Detection.
- [2] DCASE Task 5 evaluation metric repository: event-based F-measure with IoU and bipartite matching.