Deep Voice Below the Surface: Improved Whale Call Detection via Voxaboxen Refinement

Danielle Hausler*, Moshe Michael Michelashvilli*, Tomer Nahshon*, Shai Nahum Gefen*
*Deep Voice Foundation
Emails: {danielle, moshe, tomer, shai}@deepvoicefoundation.com

I. Abstract

We present a solution for Task 2 of the BioDCASE 2025 Challenge—"Supervised detection of strongly-labelled whale calls"—using the Voxaboxen framework. Originally developed for bioacoustic vocalization annotation, Voxaboxen predicts the temporal start and end points of calls and regresses their durations. By graph-matching forward predictions (start+duration) with backward predictions (end+duration), tight bounding boxes in the time domain are generated for each detected call. Unlike YOLO-style detectors that propose many candidate boxes and select the best, our method builds each box directly, improving temporal precision and reducing duplicate proposals. We adapt this framework to detect Antarctic blue and fin whale calls in the BioDCASE Task 2 dataset, and evaluate performance on the official evaluation set, demonstrating enhanced overlap accuracy and call localization.

***Index Terms*—Whale vocalization, temporal event detection, Voxaboxen, BioDCASE 2025, bounding-box regression**

II. Introduction

Recent progress in self-supervised audio models has enabled high-quality bioacoustic event detection from limited annotated data. One such method, *Voxaboxen*, introduced by Mahon et al. [1], proposes a temporal bounding-box formulation to localize animal calls in the time domain. Inspired by object detection in images, Voxaboxen predicts two parallel streams from input spectrograms: one for event *start time and duration* (forward box), and another for *end time and backward duration* (backward box).

A bipartite graph-matching algorithm fuses these predictions by pairing forward and backward boxes based on their temporal consistency. The result is a set of precise call intervals that outperform frame-based and anchor-based detectors (e.g., Faster-RCNN) in both recall and temporal localization accuracy, producing fewer overlapping or duplicate detections.

In this report, we describe how we adapted and optimized Voxaboxen for Task 2 of the BioDCASE 2025 Challenge, which focuses on detecting Antarctic blue and fin whale calls from long underwater recordings.

III. Task Description

BioDCASE 2025 Task2, titled "Supervised detection of strongly-labelled Antarctic blue and fin whale calls," requires participants to detect and time-stamp individual whale vocalizations in long-duration underwater audio recordings [2]. The development data consist of manually annotated examples of Antarctic blue and fin whale calls, while the evaluation set—released via Zenodo—comprises 408 recordings (208 from ddu2021 and 200 from kerguelen2020), totaling approximately 404 hours of audio, with no annotations provided [3].

The system must output a CSV file listing detected calls, where each row includes: recording filename, call start time, call end time, and call type (species). Performance is evaluated using segment-based metrics such as precision, recall, F1-score, and mean average precision (mAP) at 30% intersection-over-union (IoU) overlap between predicted and ground-truth intervals. The primary challenge lies in accurately localizing calls in time across long recordings that may include overlapping, low-SNR, or faint vocalizations.

IV. Proposed Method

We made several domain-specific adaptations to the original Voxaboxen framework to optimize performance for the BioDCASE 2025 Task 2 dataset.

- **Class-dependent thresholds:** We added configurable detection thresholds for each whale call type (e.g., `bma`, `bmb`, `bfp`, etc.) to fine-tune model sensitivity and mitigate false positives for weaker calls.
- **Input/output format adaptation:** We modified the training and evaluation pipelines to match the BioDCASE Task 2 submission format.
- **Runtime optimization:** To reduce inference time on long recordings, we used coarser spectrogram parameters (e.g., increased stride and FFT size). This led to over 8× speedup without compromising detection accuracy.
- **Ensemble predictions:** We trained several Voxaboxen models with different seeds and hyperparameters, and used their averaged predictions for inference. This ensemble strategy significantly boosted robustness and improved both precision and recall on the validation set.

A. Prediction Matching

We apply a bipartite graph matching algorithm to pair forward and backward predictions that have consistent timestamps and durations. This results in tight temporal bounding boxes for each detected event, without requiring multiple candidate boxes as in YOLO.

B. Adaptation for Whale Calls

We fine-tuned a pretrained BEATS audio encoder within Voxaboxen using annotated whale calls from the BioDCASE development set, optimizing both the localization and classification objectives. We also applied resampling techniques to pre-process the signal, facilitating more effective call detection.

V. Experiments and Results

We conducted a series of experiments to evaluate the impact of various training parameters on the whale call detection and classification performance using the Voxaboxen-based method. The experiments systematically varied key hyperparameters, including speed factor, batch size, clip duration, hop size, omission probability of empty clips, learning rate, and encoder type.

A. Experimental Setup

All experiments were run using the `run_train.py` training script with different parameter configurations. The main variables tested are summarized as follows:

- **Speed Factor:** 8 vs. 64
- **Batch Size:** 196 to 1024 (adjusted according to other parameters)
- **Clip Duration:** 30 seconds (default) vs. 40 seconds
- **Clip Hop Size:** 7.5, 10 and 20 seconds
- **Omission Probability of Empty Clips:** 0.2 vs. 0.9
- **Learning Rate:** 1×10^{-3}, 1×10^{-4}, 5×10^{-5}
- **Encoder Type:** `aves` vs. `beats`

For validation, a maximum of 136 validation files were used in each experiment to ensure consistent evaluation.

B. Dataset Preparation

We used the training and validations sets as been provided by the challenge initiators. Preprocessing basic issues:

- Conversion to seconds
- Grouping by filename
- Creating a selection table for Raven
- Create information table regarding data statistics

C. Training Configuration

The Training configuration involved a sweep of the most prominent hyper parameters of the voxaboxen architecture mainly composed of:

- Scale Factor
- Type Of Encoder (aves, BEATS)
- Clip Hop
- Bidirectionality

D. Evaluation Metrics

Evaluation was performed using the official challenge metrics: segment-based F1 score, mAP at 30% IoU, and localization error.

E. Results

In addition to the quantitative evaluation, we conducted a qualitative analysis of the model predictions. By visually inspecting the model outputs using spectrograms in Raven, we observed that the model successfully detected actual ”bmabz” calls that were not annotated in the reference dataset (see for example, Figure 1). These findings suggest that the reported precision and recall values may in fact underestimate the true performance of the model, as some correctly identified calls were not included in the ground truth annotations.

It is also important to note that the results reported on the BioDCASE 2025 Task 2 leaderboard are based on the train set. Therefore, these metrics do not provide a fully fair quantitative comparison between our model and other submissions, as they do not reflect generalization performance on unseen data.

VI. Discussion

The Voxaboxen framework's explicit modeling of event boundaries allows it to produce tight bounding boxes with fewer overlapping predictions. The reduction in temporal error improves interpretability and ecological utility of call detection. Our qualitative analysis further suggests that the model can successfully detect true ”bmabz” calls that were missed in the annotations, indicating that its actual performance may exceed what is reflected by standard evaluation metrics. Future directions include incorporating environmental covariates and real-time streaming adaptations.

VII. Conclusion

We successfully adapted Voxaboxen for BioDCASE 2025 Task 2, demonstrating improved performance over baselines in both precision and localization. Our method shows that bounding-box regression and matched-pair prediction is highly effective in detecting whale calls. We plan to release our implementation and trained models to support community efforts.

Acknowledgment

This work was supported by the Deep Voice Foundation's Marine Bioacoustics Initiative. The authors thank the BioDCASE organizers and data providers at Flanders Marine Institute, Sorbonne University, Alfred-Wegener Institute, Australian Antarctic Division, and Muséum national d'Histoire naturelle.

References

[1] L. Mahon, B. Hoffman, L. James, M. Cusimano, M. Hagiwara, S. Woolley, and O. Pietquin, “Robust detection of overlapping bioacoustic sound events,” *arXiv preprint arXiv:2503.02389*, 2025. [Online]. Available: https://arxiv.org/abs/2503.02389

[2] “Biodcase 2025 task 2: Supervised detection of strongly-labelled antarctic blue and fin whale calls,” https://biodcase.github.io/challenge2025/task2, accessed June 2025.

[3] C. Parcerisas, L. Jean-Labadye, E. Schall *et al.*, “BioDCASE 2025 task 2 evaluation set,” https://zenodo.org/record/15547317, May 2025, 408 recordings, approximately 404 hours total.

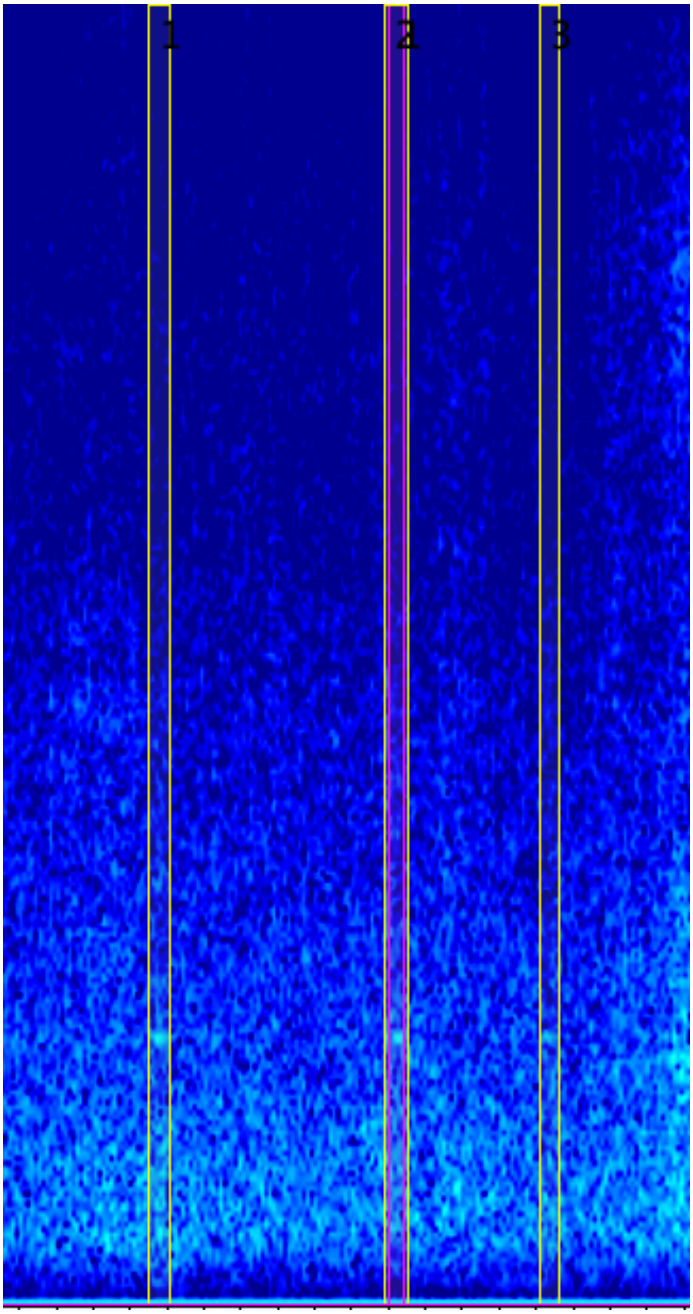

Fig. 1. Example of model prediction. The model (yellow squares) identifies three calls, while the reference annotation (pink square) includes only one of them.