ENHANCED SPECTROGRAM PROCESSING WITH TEMPORAL SEQUENCES FOR ANTARCTIC WHALE CALL DETECTION USING YOLOV11

Technical Report

Ragib Amin Nihal, Benjamin Yen, Takeshi Ashizawa, Kazuhiro Nakadai

Institute of Science Tokyo
Systems and Control Engineering
Tokyo, Japan
ragib@ra.sc.e.titech.ac.jp

ABSTRACT

We present a modified spectrogram processing approach combined with temporal sequence analysis for Antarctic whale call detection in the BioDCASE 2025 Challenge Task 2. Building on the baseline YOLO approach, We implement enhanced spectrogram preprocessing through pre-filtering, magnitude inversion, and 98th percentile normalization. We created temporal awareness by generating 3-frame RGB sequences from consecutive spectrogram frames, allowing a YOLOv11m detector to process temporal information. Class-specific confidence thresholds are applied based on validation performance analysis. On the validation set, the approach achieves 59.45% F1-score, 72.88% precision, and 51.93% recall, representing an improvement over the baseline YOLO performance of 43% F1-score.

***Index Terms*—** whale call detection, spectrogram preprocessing, temporal sequences, YOLO, passive acoustic monitoring

1. INTRODUCTION

Antarctic whale call detection presents challenges due to the low presence rate of calls ($\leq$ 10% of recording time) and variable underwater acoustic environments across different Antarctic sites and time periods. The BioDCASE 2025 Challenge Task 2 focuses on detecting seven types of Antarctic blue and fin whale calls, grouped into three categories for evaluation, as illustrated in Fig. 1.

The baseline YOLO approach provided by the challenge organizers achieves 43% F1-score on the validation set. Standard spectrogram-based approaches face difficulties with whale calls due to their low-frequency characteristics and potential masking by background noise. We addressed these issues by modifying the spectrogram preprocessing pipeline and adding temporal sequence processing to the baseline YOLO framework.

2. DATASET CHARACTERISTICS

2.1. AcousticTrends_BlueFinLibrary Dataset

The dataset consists of 1880 hours of Antarctic whale recordings from the AcousticTrends_BlueFinLibrary (ATBFL) [2], spanning 11 site-year datasets collected between 2005-2017 across Antarctica. The training set includes 6007 audio files from 8 datasets, while validation uses 587 files from 3 datasets (Casey2017, Kerguelen2014, Kerguelen2015).

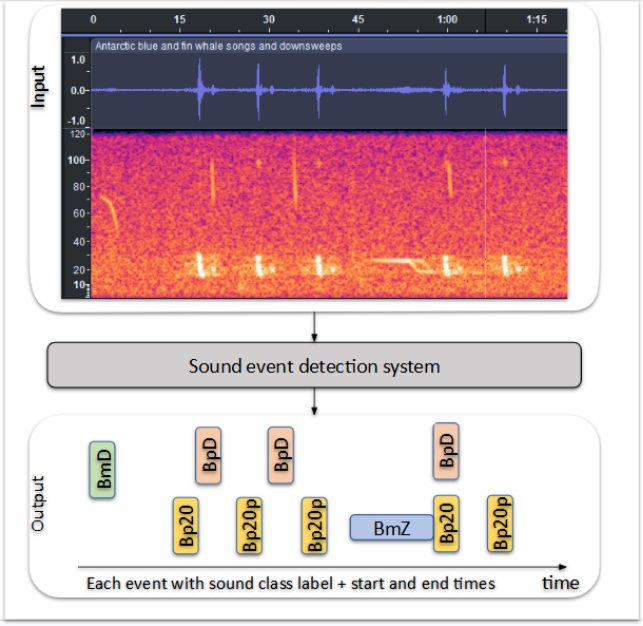


Figure 1: Overview of the sound event detection task showing different whale call types: Blue whale D and Z calls (BmD and BmZ), and Fin whale calls including 20 Hz Pulse with (Bp20p) or without (Bp20) overtone and 40Hz downsweep (BpD). (Image adapted from BioDCASE Challenge 2025[1])

2.2. Dataset Challenges

The dataset presents several characteristics that complicate detection:

Low Event Density: Whale calls occur in only 5.1% of the total recording time, creating a highly imbalanced detection problem.

Acoustic Variability: Recordings come from different hydrophone deployments, water depths, and time periods, resulting in varying background noise characteristics and propagation conditions.

Annotation Complexity: The dataset includes potential annotation inconsistencies due to subjective interpretation of call boundaries, fragmentation due to multipath propagation, and analyst variations in marking call starts/ends.

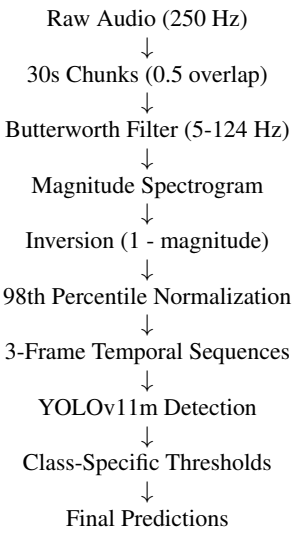


Figure 2: Processing pipeline showing the sequence of operations from raw audio to final whale call predictions.

Site-Specific Issues: Some datasets (Elephant Island, Balleny Island) have known annotation gaps, while others (Casey, Maud Rise) feature strong chorus bands that complicate single call detection [6].

2.3. Call Type Distribution

The three evaluation categories show uneven distribution: blue whale calls (bmabz) are most common, followed by 20Hz+ calls (bp), with D-calls (d) being least frequent. This imbalance motivates the use of class-specific confidence thresholds in post-processing.

3. METHOD

3.1. Processing Pipeline Overview

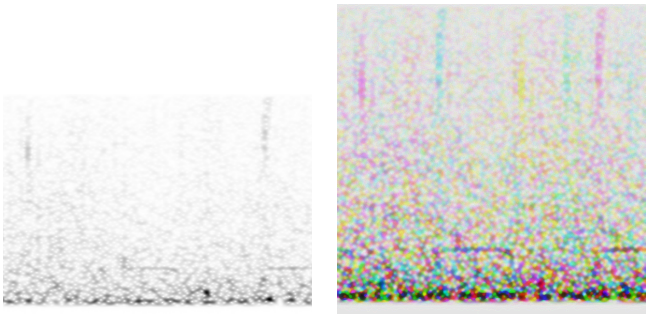
Fig. 2 illustrates the complete processing pipeline from raw audio to final predictions.

3.2. Spectrogram Preprocessing Modifications

Based on analysis of whale call characteristics in the training data, several preprocessing modifications are implemented to the standard spectrogram generation. Fig. 3 illustrates the visual impact of these modifications on whale call visibility:

Pre-filtering: Applied a 5-124 Hz Butterworth bandpass filter to remove frequencies outside the whale call range before spectrogram computation. This reduces noise and focuses processing on relevant spectral content.

Magnitude Inversion: Inverted the spectrogram magnitude using $sxx = 1 - sxx$. This modification changes the typical spectrogram appearance where whale calls appear as darker regions against lighter backgrounds.



(a) Baseline spectrogram (b) Enhanced spectrogram

Figure 3: Comparison of spectrogram preprocessing approaches showing a blue whale Z-call. (a) Baseline: Whale calls appear as faint horizontal patterns that may be difficult to distinguish from background noise. (b) Enhanced: Pre-filtered (5-124 Hz), magnitude-inverted, and 98th percentile normalized spectrogram. The whale call (visible around 20-25 Hz) appears as a prominent dark contour, making detection more reliable.

98th Percentile Normalization: Instead of standard min-max normalization, used:

$$\begin{aligned} per &= \text{percentile}(sxx.flatten(), 98) \\ sxx &= \frac{sxx - sxx.min()}{per - sxx.min()} \\ sxx[sxx > 1] &= 1 \end{aligned} \tag{1}$$

This approach reduces the impact of outlier values that may vary across different recording conditions.

3.3. Temporal Sequence Generation

To incorporate temporal information, created 3-frame sequences:

Frame Selection: For each spectrogram frame, combined it with the previous and next frames.

RGB Encoding: The three frames are arranged as RGB channels (previous=R, current=G, next=B).

Overlap Strategy: Adjacent sequences share frames, ensuring continuous temporal coverage with 50% overlap between consecutive triplets.

This allows the YOLO detector to process temporal patterns while using standard computer vision architectures designed for RGB images.

3.4. Object Detection Configuration

We use YOLOv11m [7] as the base detector with standard configuration:

Model Architecture: YOLOv11m with approximately 20M parameters
Input Format: RGB images (temporal sequences)
Classes: 3 categories (bmz, bpd, bp20plus)
Initialization: Pre-trained COCO weights

Training Parameters:

- Batch size: 16
- Training time: 8 hours on H100 GPU

Table 1: Overall Performance Comparison

Method	F1-score	Precision	Recall
Baseline YOLO	43.0%	67.0%	32.0%
This approach	59.45%	72.88%	51.93%

- Standard YOLO loss functions
- No additional data augmentation beyond temporal sequences

3.5. Post-processing

Based on validation set analysis, applied class-specific confidence thresholds:

- **bpd calls**: 0.45 threshold
- **bmz calls**: 0.30 threshold
- **bp20plus calls**: 0.30 threshold

Applied Non-Maximum Suppression with IoU threshold 0.45, followed by coordinate conversion to the required BioDCASE submission format.

4. EXPERIMENTAL SETUP

4.1. Implementation Details

Audio Processing: All audio files are processed in 30-second chunks with 0.5 overlap step (15-second spacing between chunk starts). Audio is resampled to 250 Hz to match dataset specifications.

Spectrogram Parameters:

- nfft: 512
- win_length: 256
- hop_length: 20
- Window: Hann
- Scaling: density

Temporal Processing: Each audio chunk generates multiple spectrogram frames. Valid temporal sequences require at least 3 consecutive frames.

4.2. Evaluation Protocol

We followed the official BioDCASE evaluation methodology [1]:

- Temporal IoU with 0.3 threshold for matching predictions to ground truth
- Application of joining dictionary: {bma,bmb,bmz}→bmabz, {bmd,bpd}→d, {bp20,bp20plus}→bp
- Per-class and per-dataset metrics computation

5. RESULTS

5.1. Validation Set Performance

Table 1 shows the overall performance compared to the baseline:

The approach shows improvement in both F1-score (+16.45%) and recall (+19.93%) while maintaining similar precision levels.

Table 2: Per-Dataset Results

Dataset	F1-score	Precision	Recall
Casey2017	55.73%	72.64%	52.23%
Kerguelen2014	61.78%	75.54%	43.76%
Kerguelen2015	60.67%	72.21%	60.49%

Table 3: Per-Class Results After Joining

Class	F1-score	Precision	Recall	Notes
bmabz	71.77%	73.38%	70.23%	Blue whale calls
d	48.24%	75.52%	35.43%	D-calls (challenging)
bp	58.33%	69.73%	50.13%	20Hz+ calls

5.2. Per-Dataset Performance

Performance varies across datasets, with Kerguelen2015 showing the best recall (60.49%) and Kerguelen2014 the best precision (75.54%).

5.3. Per-Class Performance

Blue whale calls (bmabz) achieve the strongest performance, while D-calls show the lowest recall, motivating the higher confidence threshold for this class.

6. DISCUSSION

6.1. Analysis of Results

The preprocessing modifications provide measurable improvements over the baseline YOLO approach. As illustrated in Fig. 3, the enhanced spectrogram processing significantly improves whale call visibility compared to standard spectrograms. The 98th percentile normalization appears to help with the varying acoustic conditions across different Antarctic sites. The temporal sequences allow the model to process call dynamics that single-frame spectrograms cannot capture.

6.2. Method Limitations

Several limitations affect the approach:

Fixed Temporal Window: The 3-frame window may not capture the full duration of longer whale calls, which can extend beyond three consecutive frames.

Class Imbalance: D-calls remain challenging to detect (35.43% recall), suggesting the preprocessing may be less effective for this call type.

Computational Overhead: The temporal sequence generation increases data volume by 3x compared to single-frame processing.

6.3. Future Improvements

Potential improvements could include:

- Adaptive temporal window sizes based on call type
- More sophisticated normalization schemes for different acoustic environments

- Ensemble methods combining different preprocessing approaches
- Extension to detect all seven individual call types rather than three grouped categories

7. CONCLUSION

Here we present a modified approach for Antarctic whale call detection that builds on the baseline YOLO framework with spectrogram preprocessing modifications and temporal sequence processing. The method achieves 59.45% F1-score on the BioDCASE challenge validation set, representing a 16.45% improvement over the baseline performance. The approach demonstrates that domain-specific preprocessing and temporal modeling can provide benefits for marine acoustic detection tasks, though limitations remain, particularly for D-call detection. The results provide a foundation for further development of automated whale call detection in passive acoustic monitoring applications.

8. REFERENCES

[1] BioDCASE Challenge 2025, "Task 2: Supervised Detection of Strongly-Labelled Whale Calls," 2025. [Online]. Available: https://biodcase.github.io/challenge2025/

[2] B. Miller et al., "An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors," *Sci. Rep.*, vol. 11, p. 806, 2021.

[3] I. Castro et al., "Beyond counting calls: estimating detection probability for Antarctic blue whales reveals biological trends in seasonal calling," *Front. Mar. Sci.*, 2024.

[4] E. Schall et al., "Deep learning in marine bioacoustics: a benchmark for baleen whale detection," *Remote Sens. Ecol. Conserv.*, vol. 10, pp. 642-654, 2024.

[5] G. Dubus et al., "Improving automatic detection with supervised contrastive learning: application with low-frequency vocalizations," *Workshop DCLDE*, 2024.

[6] E. Schall and C. Parcerisas, "A Robust Method to Automatically Detect Fin Whale Acoustic Presence in Large and Diverse Passive Acoustic Datasets," *J. Mar. Sci. Eng.*, vol. 10, no. 12, p. 1831, 2022.

[7] G. Jocher et al., "YOLOv11: An Improved Real-Time Object Detection Algorithm," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics