A THREE-CHANNEL PHYSICS-INFORMED TCN FOR CROSS-SITE ANTARCTIC BLUE AND FIN WHALE CALL DETECTION

Technical Report

George-Daniel Gherasim

ENSTA, Institut Polytechnique de Paris
CSN, Brest, France
george-daniel.gherasim@ensta.fr

ABSTRACT

We describe a system for Sound Event Detection (SED) of Antarctic blue and fin whale calls submitted to BioDCASE 2026 Task 2 [1]. The task involves seven fine-grained call types grouped into three super-classes, evaluated at temporal IoU ≥ 0.3 (macro-F1) on mono 250 Hz hydrophone recordings. The primary challenge is cross-site generalization: models are trained on eight recording sites and evaluated on four entirely unseen sites with different instruments and noise environments. Our approach uses a physics-informed three-channel frontend — each channel tuned to the time-frequency signature of one super-class — with Per-Channel Energy Normalization (PCEN) [2] for site-level gain normalization, followed by a compact dilated Temporal Convolutional Network (TCN) [3] with three multi-label sigmoid heads. We achieve a robust validation macro-F1 of 0.516 (IoU$\geq$ 0.3), and a val-tuned configuration reaches 0.542. Both outperform the published challenge baselines, also measured on the validation set (YOLOv11: 0.44, ResNet18: 0.32). The system operates at the three-superclass granularity used for scoring and emits a fixed representative fine label per group; it does not perform seven-way fine classification. The design priority throughout is robustness over held-out val performance.

Index Terms— bioacoustics, sound event detection, whale calls, temporal convolutional network, cross-site generalization

1. INTRODUCTION

BioDCASE 2026 Task 2 [1] is a Sound Event Detection challenge on underwater acoustic recordings of Antarctic blue whales (Balaenoptera musculus intermedia) and fin whales (Balaenoptera physalus). Audio is mono, 250 Hz sample rate, with recordings spanning multiple Southern Ocean sites at different seasons. Seven fine-grained call types — bma, bmb, bmz, bmd, bpd, bp20, bp20plus — are provided in annotations, but scoring groups them into three super-classes:

- bmabz: blue whale tonal calls (bma, bmb, bmz) — narrow-band, slowly frequency-modulated signals in the 16–30 Hz range with durations of tens of seconds;
- d: downsweep calls (bmd, bpd) — broadband 20–120 Hz downsweeps with finer temporal structure;
- bp: fin whale 20 Hz pulses (bp20, bp20plus) — pulse trains with a fundamental near 14–30 Hz and a harmonic overtone at 80–120 Hz.

The official evaluation metric is temporal IoU at threshold 0.3, with greedy 1-to-1 matching per class (IoU@0.3), aggregated as macro-F1 across the three super-classes.

The central difficulty of this task is domain shift across recording sites: models are trained on eight sites, validated on three sites (all different from training), and evaluated on four additional sites never seen during development. Instruments, deployment depths, ambient noise spectra, and signal-to-noise ratios all vary significantly across sites. A 2025 edition top-ranked transformer ensemble achieved val macro-F1 0.64 but dropped to 0.50 on the blind evaluation set [1], a -0.14 gap attributable to site-level overfit. This motivates our design emphasis on regularization and cross-site robustness over val-set peak performance.

2. METHOD

2.1. Physics-Informed Three-Channel Frontend

All processing operates at 250 Hz (no resampling), exploiting the full useful bandwidth (0–124 Hz, Nyquist at 125 Hz). The frontend extracts three complementary feature channels, each tuned to the time-frequency structure of one super-class, on a single shared temporal grid with hop $H = 32$ samples (0.128 s/frame, ≈ 7.8 frames/s). The choice of $H = 32$ (rather than larger hops) was the single most important design decision and is discussed in Section 3.

Channel 1 – ABZ narrowband (for bmabz): STFT with $N_{FFT} = 512$ (2.048 s window, frequency resolution 0.49 Hz), cropped to 12–35 Hz (47 bins). The long analysis window provides fine frequency resolution matched to the slow tonal sweeps of blue whale A/B/Z calls.

Channel 2 – DDswp broadband (for d): STFT with $N_{FFT} = 128$ (0.512 s window, frequency resolution 1.95 Hz), cropped to 20–124 Hz (53 bins). The short window provides the temporal resolution needed to resolve fast downsweeps spanning 20–120 Hz.

The author is a student at ENSTA, Institut Polytechnique de Paris, which co-organizes BioDCASE Task 2. The author had no role in organizing the task, no access to the evaluation-set annotations, and used only the publicly available development data; all training targets and decoding thresholds were derived solely from the development split.

Channel 3 – bp harmonic (for bp): STFT with $N_{FFT} = 256$ (1.024 s window), with explicit harmonic decomposition. The fundamental band 14–30 Hz is split into 8 sub-bands; the overtone band 80–120 Hz is split into 5 sub-bands. A coincidence feature $c = \log(1 + \bar{E}_{\text{fund}} \cdot \bar{E}_{\text{over}})$ captures harmonic co-occurrence (higher SNR for true harmonic pulses vs. mono-band noise), yielding 14 dimensions total.

Total feature dimension: $47 + 53 + 14 = 114$.

Each channel is independently normalized with Per-Channel Energy Normalization (PCEN) [2], an adaptive gain control with time constant $\tau = 0.4$ s. PCEN suppresses the slowly varying noise floor, which differs across recording sites and depths, providing level-invariant representations that improve cross-site transfer. All outputs are sanitized (NaN/inf $\to$ 0) before the model.

2.2. Dilated TCN

The model is a dilated Temporal Convolutional Network [3] operating on the sequence $(T, 114)$. Architecture: 6 residual blocks, 96 channels, kernel size 3, dilation 2^i $(i = 0, \ldots, 5)$, GroupNorm + GELU activations, non-causal (symmetric receptive field), dropout 0.3, weight decay 10^{-4}. Three sigmoid output heads (one per super-class) produce frame-level multi-label probabilities.

Training objective: binary cross-entropy with positive class weight $w^+ = 8$ (compensates class imbalance). Optional boundary heads — three additional linear layers producing onset and offset Gaussian-bump targets, in the spirit of dual-objective onset/frame transcription [4] — extend the output from $(B, T, 3)$ to $(B, T, 9)$. When boundary heads are active, the loss is:

$$\mathcal{L} = \text{BCE}(p_{\text{pres}},\, y;\, w^+) + 0.5\,\text{BCE}(p_{\text{on}},\, y_{\text{on}}) + 0.5\,\text{BCE}(p_{\text{off}},\, y_{\text{off}}), \quad (1)$$

where y_{on} and y_{off} are Gaussian bumps (standard deviation 0.19 s) centered at event start and end frames, respectively, with overlap resolved by element-wise maximum.

2.3. Decoding and Submission

Post-processing applies, per super-class: (1) probability threshold, (2) median filter (3 frames), (3) minimum duration and minimum gap merging. For the boundary-head variant, an additional gate-and-snap stage uses onset/offset peaks (detected via scipy.find_peaks) to filter false-positive segments (gate: a segment is suppressed unless an onset/offset peak exists within a class-specific window) and to refine boundary positions (snap).

Two configurations were submitted:

- Robust: presence-only decoding with lightly swept thresholds (bmabz 0.9 / d 0.6 / bp 0.8). Val macro-F1 0.5165. Designed for cross-site robustness.
- Tuned: boundary heads with gate-and-snap and duration clamps, thresholds grid-searched on val. Val macro-F1 0.5416. Carries val-overfit risk on unseen eval sites.

For the official fine-label submission format, each super-class detection is emitted with the most frequent fine label of its group: bmabz $\to$ bma, d $\to$ bmd, bp $\to$ bp20. The official scorer reaggregates fine labels back to the three super-classes.

Table 1: Validation macro-F1 (IoU@0.3) by configuration.

Configuration	Macro	bmabz	d	bp
TCN hop=64 baseline	0.502	0.592	0.352	0.562
hop=32, presence-only	0.516	0.578	0.399	0.573
hop=32 + boundary heads	0.542	0.627	0.418	0.580

3. EXPERIMENTS AND RESULTS

3.1. Setup

Data: development set from [1]. Train: 8 sites (ballenyislands2015, casey2014, elephantisland2013/2014, greenwich2015, kerguelen2005, maudrise2014, rosssea2014). Val: 3 sites (kerguelen2014, kerguelen2015, casey2017). Eval: 4 blind sites (ddu2019, ddu2021, kerguelen2019, kerguelen2020). All splits are site-disjoint by design (domain-shift evaluation). Approximately 24k/19k/15k annotated events per super-class in training.

Evaluation uses the official 1D temporal IoU scorer with threshold 0.3, greedy 1-to-1 matching, macro-F1. A note on training label noise: 155 training file basenames collide across sites (timestamp-only filenames), causing $\approx 3.6\%$ label merge noise in the training set. Validation and eval sets have zero collisions; submitted predictions are unaffected.

3.2. Main Results

Table 1 reports val macro-F1 for three configurations.

The official challenge baselines, evaluated on the same validation set, are: YOLOv11 macro-F1 0.44 (P 0.67, R 0.32) and ResNet18 macro-F1 0.32 (P 0.29, R 0.36) [1]. Our robust configuration (0.516) exceeds both at equal scope (val, IoU@0.3). The Whale-VAD system (Stellenbosch, 2025 top-6) on the same val sites achieves macro-F1 ≈ 0.37 (d-class F1 0.12 vs. our 0.40).

3.3. What Worked and What Did Not

Fine temporal resolution (hop=32, +0.015 macro). This was the most promising lever among those tried, but we are careful not to overstate it. The macro gain (0.502 $\to$ 0.516) is small, and is concentrated almost entirely in the d-class (0.352 $\to$ 0.399, +0.047), while bmabz in fact decreases slightly (0.592 $\to$ 0.578) and bp is essentially flat. The proposed mechanism is class-specific and physically plausible: short downsweep calls span only a few frames at hop=64, so finer temporal resolution aligns their predicted extents to the annotation boundaries and lifts the IoU above the 0.3 threshold. However, the headline macro delta (+0.015) is comparable to the ± 0.04 per-class single-run variance we treat as noise elsewhere in this study, and each configuration was trained once on a single 3-site val split. We therefore present hop=32 as the most promising lever, not a statistically established one; the per-class pattern is consistent with the mechanism, but multi-seed and leave-one-site-out quantification (Section 3.5) would be required to confirm it. The hop=32 model is also the only setting in which the boundary heads help; at hop=64 they were rejected.

Onset/offset boundary heads (+0.007, partly val-tuned). Adding Gaussian-bump onset/offset supervision and boundary-gated decoding yields modest gains in bmabz (gating in "either" mode reduces false positives) and d/bp. The gain is partly attributable to the large val decode grid search, so we conservatively treat 0.516 (presence-only) as the deployable number.

Augmentation (rejected). Standard spectral augmentations (time stretch, masking) over-corrupted the compact PCEN feature representations and degraded val performance.

Conformer / larger model (rejected). Replacing the TCN with a Conformer [5] and increasing capacity led to train-site overfit with no cross-site gain. This reinforces that capacity is not the bottleneck; cross-site regularization is.

GBM stacking (rejected). Gradient-boosted stacking of TCN out-of-fold probabilities with hand-crafted spectral features (spectral centroid, subband energies) degraded performance: the spectral features carry per-site noise rather than transferable signal.

Phase features and 10 Hz high-pass (neutral). Instantaneous frequency deviation (IFD) and group delay features were tested in all three phase modes; the apparent per-class deltas (±0.04) were within single-run seed variance and considered noise. A 10 Hz high-pass pre-filter similarly showed no reliable improvement.

Methodological note. Throughout the lever study, single-run per-class deltas of ±0.04 were treated as seed noise, and every candidate was compared against an identical control run. Improvements were accepted only when the macro gain was positive and the per-class pattern was mechanistically consistent. We did not, however, perform multi-seed significance testing; as noted above, the hop=32 macro gain itself sits near this single-run noise threshold (Section 3.5).

3.4. Generalization Analysis

The val→eval gap is a central concern. Evidence from the 2025 edition shows that a 3×Swin-Transformer ensemble achieved val macro-F1 0.64 but only 0.50 on the blind eval set [1], a −0.14 drop consistent with site-level overfit. Our design choices — a small (96-channel, 6-block) TCN, strong dropout (0.3) and weight decay (10^{-4}), PCEN normalization per channel, no ensembling, and conservative threshold sweeps — are intended to minimize this gap. We therefore report 0.516 as the primary deployable figure and expect the tuned 0.542 to shrink toward it on the blind eval sites. We emphasize that this is a reasoned prediction, not a measured result: we do not have access to the blind eval scores at submission time, so the size of the val→eval gap for our system remains unverified.

The remaining weakness across all configurations is bmabz precision (≈ 0.50, i.e., approximately half of bmabz detections are false positives). The clearest next lever not yet explored is training with explicit pure-negative files (sites with no whale calls), which could reduce the false-positive rate without compromising recall.

3.5. Limitations

We state the main limitations explicitly. (i) No multi-seed statistics. Each configuration was trained once; we report single-run macro-F1 without mean±standard deviation over seeds. Small deltas — including the hop=32 gain (+0.015) — are therefore not significance-tested and should be read as suggestive. Reporting mean±std over 5–10 seeds, and a leave-one-site-out estimate of inter-site variance over the eight training sites, is the most important follow-up and would be needed to upgrade "most promising lever" to "established lever". (ii) PCEN not ablated. PCEN carries our cross-site robustness argument by design, but we did not compare it against log-mel or per-channel z-score normalization, so its quantitative contribution is asserted rather than measured. (iii) Coincidence feature not ablated. The harmonic coincidence feature is motivated physically but its specific contribution to bp detection is untested. (iv) val→eval gap unverified. As noted above, the predicted shrinkage of the tuned configuration toward the robust one is a reasoned expectation, not a measured result. (v) Three-class scope. The system detects at the three-superclass granularity used for scoring and emits a fixed representative fine label per group; it does not discriminate the seven fine call types.

4. CONCLUSION

We presented a compact physics-informed system for cross-site Antarctic whale call detection. A three-channel STFT frontend — ABZ narrowband, DDswp broadband, and bp harmonic — with PCEN normalization provides physically motivated features that are robust to inter-site gain and noise variation. A dilated TCN with three multi-label sigmoid heads operates on the 114-dimensional feature sequence at a fine temporal resolution of 0.128 s/frame.

The most promising design decision was temporal resolution: reducing the hop from 64 to 32 samples produced the largest macro improvement (+0.015), concentrated in the short-downsweep d-class, although this gain sits near our single-run noise threshold and is not yet significance-tested (Section 3.5). Optional onset/offset boundary heads provide modest additional gains in val-tuned decode mode. All other capacity or complexity increases were rejected due to cross-site overfit.

The robust submitted configuration achieves val macro-F1 0.516, exceeding the published baselines (YOLOv11: 0.44, ResNet18: 0.32). The remaining principal weakness is high false-positive rate on the bmabz class (precision ≈ 0.50); addressing this via pure-negative training examples is the clearest remaining lever.

Acknowledgment

The author thanks the BioDCASE 2026 organizers for providing the dataset, annotations, and official evaluation scorer. As noted on the first page, the author is affiliated with ENSTA, a co-organizing institute of this task; to prevent any unfair advantage, the author had no involvement in the task organization and no access to the blind evaluation annotations, and developed the system exclusively from the public development data.

5. REFERENCES

[1] BioDCASE 2026 Organizing Committee, "BioDCASE 2026 challenge task 2: Antarctic blue and fin whale sound event detection," https://biodcase.github.io/, 2026, accessed: June 2026.

[2] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 5670–5674.

[3] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018. [Online]. Available: https://arxiv.org/abs/1803.01271

[4] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in Proc. International Society for Music Information Retrieval Conference (ISMIR), Paris, France, 2018, pp. 50–57.

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in Proc. Interspeech, Shanghai, China, 2020, pp. 5036–5040.