WHALE CALL DETECTION WITH A FREQUENCY-DYNAMIC CRNN ENSEMBLE

Technical Report

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

Institute of Science Tokyo
Department of Systems and Control Engineering
Tokyo, Japan

1. ABSTRACT

We address BioDCASE 2026 Task 2, the detection of Antarctic blue whale (bma/bmb/bmz/bmd) and fin whale (bp20/bp20plus/bpd) calls in long-duration, low-frequency passive-acoustic recordings. Audio is processed at 250 Hz in 32-second chunks. Each chunk is transformed to a log-magnitude STFT (`n_fft=256`, `hop=32`). A per-file noise-floor bias subtraction corrects the large ($\sim$47 dB) cross-site PSD differences that cause dev-to-test collapse. PCEN normalization is then applied. The detector is a Frequency-Dynamic CRNN (FDY-CRNN) with a seven-channel convolutional stack, frequency-dynamic basis kernels, and a bidirectional GRU. It is trained as a seven-class multi-label problem using masked focal binary cross-entropy. Predictions are multi-label (no argmax) and allow polyphony of up to four simultaneous calls. The submission ensembles five members: four site-disjoint cross-validation folds (two with mean-teacher test-time adaptation) and one all-sites model. We average logits across members, then apply per-class threshold-and-duration event decoding and IoU-0.3 non-maximum suppression to comply with the strict 2026 duplicate-handling rule. On four-fold site-disjoint cross-validation the system reaches a mean macro F_1 of 0.538. We submit four threshold variants that span precision-leaning and recall-leaning operating points.

2. SYSTEM OVERVIEW

2.1. Task and data.

The development set holds 6,591 WAVs across 11 site-years, about 1,880 h in total, with file lengths from 5 to 60 min. The evaluation set covers 4 unseen site-years (`ddu2019`, `ddu2021`, `kerguelen2019`, `kerguelen2020`), 795 WAVs in all. Annotators marked 7 raw call types with start and end times. The official evaluator collapses these to 3 scored categories. Up to 4 calls can overlap in time. We therefore treat the classes as multi-label sigmoid outputs throughout training, and apply the 7→3 collapse only at evaluation.

Eval category	Raw call types
`bmabz`	`bma`, `bmb`, `bmz` (Antarctic blue A/B/Z tonals)
`d`	`bmd`, `bpd` (blue D-call + fin 40 Hz downsweep)
`bp`	`bp20`, `bp20plus` (fin 20 Hz pulse ± overtone)

2.2. Three properties of the data shape the design.

(P1) Cross-site PSD bias of about 47 dB. Different hydrophone gains shift the long-term spectra by up to two orders of magnitude. We attribute this to the bias, since raw spectrograms do not transfer across sites.
(P2) Calls are sparse, present in under 10% of the audio. Uniform sampling then gives batches that are over 90% negative and starve the rare `d` class, so batches need to be anchored on positives.
(P3) Per-class difficulty spans about 0.15 F1 (`bmabz` ≈0.61, `d` ≈0.47). One global threshold cannot serve all three classes, so we decode each class separately.

2.3. Bias-aware spectrogram with PCEN

The STFT uses n_{fft}=256 with hop 32. The window is 1.02 s, close to the 1 s minimum call duration. This gives a 7.8125 Hz frame rate and 250 frames per 32 s chunk. The cross-site normalisation step (P1) is **per-file noise-floor bias subtraction**. We estimate a 30 s noise floor for each file. In training this comes from an annotation-derived estimate; at evaluation it comes from the first 30 s. We resample the floor to 129 STFT bins and subtract it from each log-magnitude spectrogram. We then add back a fixed global reference of $\rho^\star$=39.509 dB, calibrated from 200 random chunks, so the network always sees the same zero-point. If a file's in-band noise differs from the site median by more than 3·MAD, we use the site-median bias instead, which guards against outliers. A Phase-1 ablation showed that **PCEN** (with fixed librosa parameters) already removes most of the cross-site shift on its own. The extra effect of the bias term is small, near 0 macro F1. We keep it as a cheap normaliser rather than a main driver. The 5–124 Hz band is applied as a frequency mask inside the model, not as a CPU filter.

2.4. Model: FDY-CRNN

Frequency-Dynamic convolution keeps K learnable basis kernels and combines them with attention that depends on frequency position. The layer can apply different effective kernels to different bands. This suits the task. The three categories sit in characteristic, partly overlapping bands: `bmabz` long tonals peak near 26 Hz, `d` short downsweeps cover the mid-band, and `bp` 20 Hz pulses concentrate around 17–26 Hz. Band-conditioned kernels can specialise for these regions, while a plain CNN has to learn to ignore the wrong bands at every position. The model has 4.77 M parameters. It is a 7-block convolutional stack `[16,32,64,128,128,128,256]` with K=4 FDY basis maps at `pool_depths = [1,3,5]`, time pooling, a bidirectional GRU

(256 units, 2 layers), and a 7-output sigmoid head. All batch-norm layers use `track_running_stats=False`. A Phase-1 ablation explains why. The class-anchored sampler creates a train/eval distribution shift that corrupts the running statistics. With running stats on, eval-mode `bma` F1 fell to 0.00 while train-mode F1 was 0.68 on the same input. Using batch statistics in both modes removes the problem.

2.5. Training

Class-anchored sampler (P2). Each batch of 16 has 7 anchored slots, one guaranteed positive per class, and 9 random slots. The random slots are weighted by hour of day within a site and by file duration. This way the rare `d` class, which covers well under 1% of the audio at most sites, appears in every batch instead of being starved by uniform sampling.

Pre-merge. For the four short-pulse classes (`bmd,bp20,bp20plus,bpd`) we merge adjacent annotations that fall within 1.5 s before chunking. This adds about 0.02 F1.

Loss. We use per-class focal BCE (γ=2, α=1), multi-label with no argmax. A per-(site,class) mask removes the gradient for classes a site never labelled.

Splits. We apply SpecAugment to the PCEN spectrogram. The 4 folds each hold out 2 site-years and train on 9, with rosssea2014, casey2014, and greenwich2015 always in training. Fold 2 (casey2017 + kerguelen2005) is the hardest split. It runs first as a gate, requiring site-disjoint macro F1 ≥ 0.30.

TTA. We add a mean-teacher consistency loss only for folds that pass a proxy gate, defined as a gain above 0.005 macro F1 when we adapt on a holdout and evaluate on that holdout. Folds 2 and 4 pass ($+0.015$, $+0.009$). Folds 1 and 3 over-adapt, so they keep their baseline checkpoints.

2.6. Decoding

An event is a run of consecutive frames above threshold that lasts longer than a class minimum duration. Both values are per-class (P3). We pick them with a leave-one-fold-out (LOO) sweep that maximises mean held-out F1. For deployment we use the per-class value that maximises F1 averaged over all 4 folds:

class	τ	$d^{\min}$
`bmabz`	0.47	2.0 s
`d`	0.47	1.0 s
`bp`	0.38	1.0 s

The thresholds are stable across folds, and τ_{bp}=0.38 is selected by every fold. The single deployment set is therefore close to loss-free against the LOO mean. The patched 2026 evaluator, re-vendored after the organiser fix on 2026-06-02, scores N predictions on one ground-truth event as 1 TP and $(N-1)$ FP. This is stricter than 2025. We enable a flag-gated time-IoU NMS (IoU ≥ 0.3, grouped by `(dataset,file,annotation)`) as a safety net. In practice the per-class run decoding produces non-overlapping events.

2.7. Ensemble and four submissions

The ensemble averages the $7 \times T$ logits of 5 members before the sigmoid and threshold step. The total is 23.83 M parameters. The members are the fold-1 baseline, fold-2 TTA, fold-3 baseline, fold-4 TTA, and an all-sites model. The all-sites model has no site-disjoint validation, so we use it only for selection. BioDCASE allows four submissions. We use them to vary the ensemble composition and the threshold strictness:

#	ensemble	per-class thresholds	NMS	rationale
1	5-model	deployment (0.47/0.47/0.38)	on	primary submission
2	5-model	loose (-0.05 each)	on	recall hedge
3	4-fold only	deployment	on	drops all-sites
4	5-model	tight ($+0.05$ each)	on	precision hedge

3. RESULTS AND ABLATIONS

3.1. 4-fold leave-one-out cross validation

The held-out fold F1 (with per-class optimal thresholds) is:

fold	held-out sites	per-class optimal F1
1	elephantisland2014, maudrise2014	0.588
2	casey2017, kerguelen2005	0.504
3	elephantisland2013, ballenyislands2015	0.538
4	kerguelen2014, kerguelen2015	0.521
	4-fold mean macro F1	**0.538**

Across the 4 folds, per-class F1 is about 0.61 for `bmabz`, 0.47 for `d`, and 0.55 for `bp`. The `d` class is the main remaining source of error. The top 2025 leaderboard score under the lenient evaluator was 0.50 (Marolt, multi-resolution Swin + LEAF). Our 0.538 uses the same lenient evaluator. It is a slight upper bound on what the stricter 2026 evaluator will give for the same predictions. The gap is under 0.005 in our tests, because our decoder produces no within-class duplicates above IoU 0.3.

3.2. Per-site test prediction counts

On the 795 evaluation WAVs, submission 1 produces 15,064 events across all 4 sites. 621 of 795 files have at least one event. The class share is 42.5% `bmabz`, 37.1% `d`, and 20.4% `bp`.

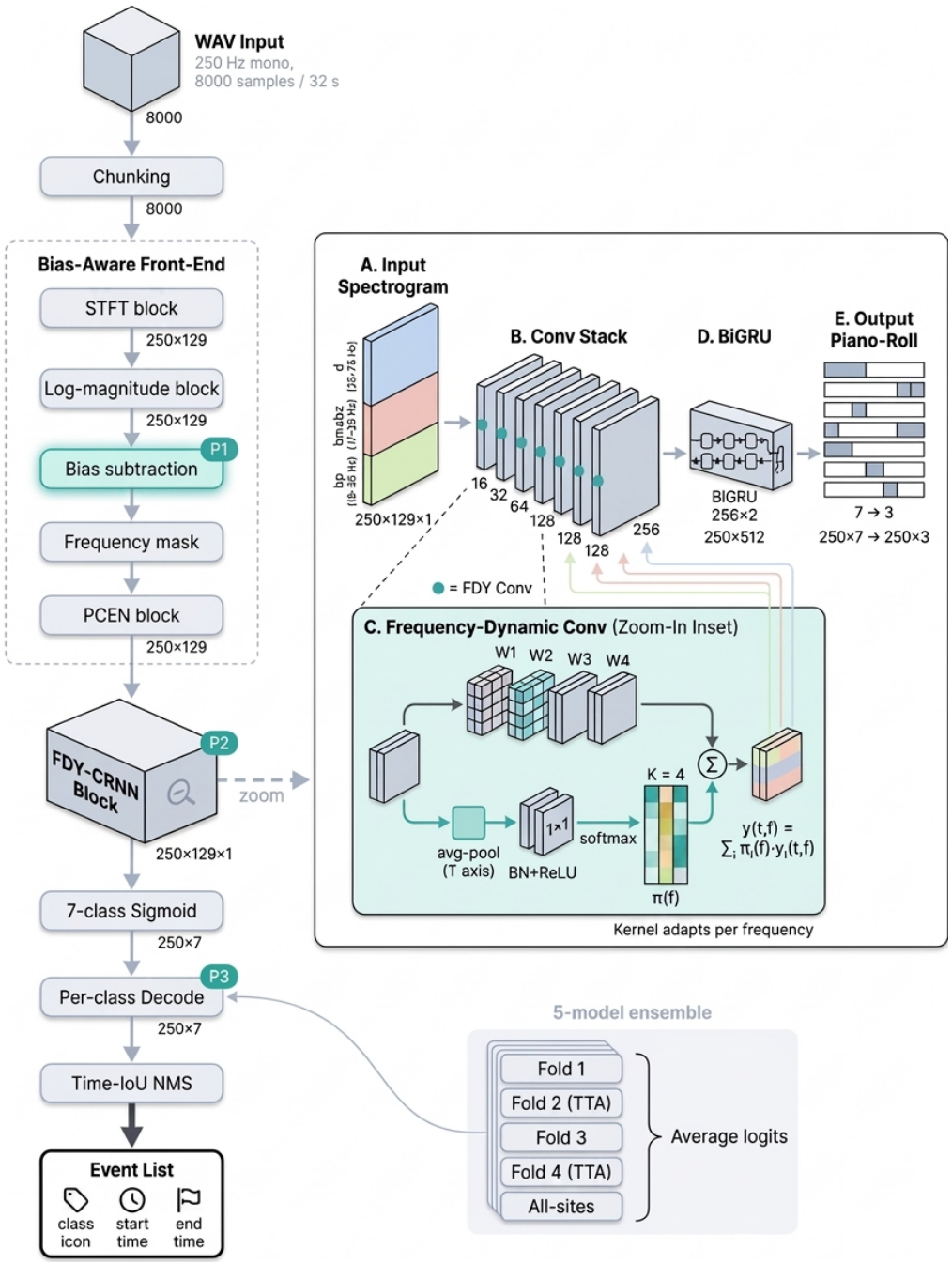

Figure 1: End-to-end architecture of the Whale FDY–CRNN system. **Left spine (data flow, top to bottom):** 32 s of 250 Hz mono audio (8000 samples) is chunked, then passed through the bias-aware front-end—STFT (n_{fft}=256, hop 32, yielding 250 × 129), log-magnitude, per-file noise-floor bias subtraction (P1, teal), the 5–124 Hz frequency mask, and PCEN. The resulting $250 \times 129 \times 1$ representation enters the FDY-CRNN (P2), whose 7-class sigmoid head emits frame-level posteriors (250×7). Per-class threshold and minimum-duration decoding (P3) followed by time-IoU NMS produces the final event list (class, start, end). **Right panel (FDY-CRNN detail, A–E):** (A) the band-coded input spectrogram, with the three scored categories (`d`, `bmabz`, `bp`) occupying characteristic, partly overlapping frequency regions; (B) the seven-block convolutional stack (channels $16 \to 32 \to 64 \to 128 \to 128 \to 128 \to 256$), with FDY blocks marked by teal dots; (C) the frequency-dynamic convolution operator, in which K=4 learnable basis kernels ($W_1 - W_4$) are combined by a time-pooled, frequency-dependent attention $\pi(f)$ to form an effective kernel that adapts per frequency band, $y(t,f) = \sum_i \pi_i(f)\, y_i(t,f)$; (D) a bidirectional GRU (256 units, 2 layers; 250×512); and (E) the 7-track output piano-roll, with the $7 \to 3$ category collapse applied only at evaluation. **Lower right:** the five-member ensemble (four site-disjoint folds, two with mean-teacher TTA, plus an all-sites model) is combined by averaging logits before the decode step. Teal marks the two learned normalisation/attention pathways (bias subtraction and frequency attention); the three band hues link spectrogram bands to effective kernels to output tracks.