# BIODCASE TASK 3: EFFICIENT BIRD SOUND RECOGNITION WITH SIGNAL-PROCESSING PRIORS AND KNOWLEDGE DISTILLATION

Technical Report

*Heitor R. Guimarães, Tiago H. Falk*

INRS-EMT, Université du Québec, Montréal, Canada

## ABSTRACT

While recent audio-based deep learning methods, such as AVES and Perch, have shown strong performance in bioacoustic tasks, their computational cost hinders deployment on resource-constrained platforms. Herein, we describe our submission to BioDCASE-Tiny 2026 Task 3, an 11-class bird sound recognition task targeting the resource-constrained ESP32-S3 microcontroller. We first evaluate Perch 2.0, a large bioacoustic model designed for bird sound classification, on the proposed task, and then distill it into a compact CNN deployable on the target device. Rather than relying on capacity, our student encodes inductive biases well-suited to bird vocalizations (e.g., spectro-temporal modulation, rhythm, and dilated temporal context) directly in its architecture: it takes a PCEN-mel front-end as input and processes frequency and temporal characteristics in separate modules. On the validation set, our model achieves $0.765$ Top-1 accuracy and $0.954$ macro ROC-AUC with $81,451$ parameters, compared to $0.563$ / $0.893$ for the provided baseline; a static int8 export retains $0.740$ / $0.952$ at roughly 153 KiB.

***Index Terms*—** Knowledge Distillation, Signal Processing, Bioacoustics, Inductive Bias

## 1. INTRODUCTION

Passive acoustic monitoring has become a key strategy in biodiversity assessment, conservation, and behavioral ecology. As Internet of Things (IoT) devices enable continuous in situ audio collection at scale, automated machine-learning-based sound recognition tools are necessary. Large bioacoustic models, such as Perch and AVES[1, 2], now achieve strong performance across a wide range of tasks, learning meaningful representations that enable downstream classification. Their computational cost, however, places them far out of reach of the low-power, memory-constrained devices that are most attractive for in-situ deployment: running inference at the edge avoids transmitting raw audio, reduces energy use, and enables real-time monitoring in environments with limited connectivity.

The BioDCASE-Tiny 2026 Task 3 challenge targets exactly this regime, posing an 11-class bird sound recognition problem (ten species plus an urban-noise background class) that must run on the ESP32-S3-Korvo-2, a microcontroller dev board with no floating-point accelerator and a few hundred kilobytes of usable RAM. Submissions are scored not only on Top-1 accuracy and macro ROC-AUC but also on resource efficiency: model size, multiply-accumulate operations, and on-device latency and peak memory. This rules out deploying a foundation model directly and instead favors compact networks, yet a small CNN trained from scratch on the limited development set leaves substantial accuracy on the table, as the provided baseline ($0.56$ accuracy) illustrates.

We bridge this gap with knowledge distillation: a frozen Perch 2.0 teacher supervises a compact student that is small enough to deploy yet inherits much of the teacher's discriminative ability. Rather than spending the student's limited capacity learning generic structure, we build in signal-processing priors that characterize bird vocalizations. Our student model, herein called **Lite-Perch**, receives as input a per-channel energy-normalized mel spectrogram as its front-end. The architecture consists of a Gabor spectro-temporal filterbank and separate frequency- and time-aggregation modules with dilated temporal convolutions. Every operator remains compliant with tflite, so the network deploys without falling back to unsupported kernels. On the validation set, this system achieves strong results, improving the baseline Top-1 accuracy and macro ROC-AUC by $20.2$ and $6.1$ absolute percentage points, respectively, while reducing the parameter count by $\approx 16\%$.

## 2. METHOD

Lite-Perch maps a 3-second, 24 kHz audio clip to one of 11 classes. The waveform is converted to a time–frequency representation by the front-end (Section 2.1), passed through a Gabor-initialized convolution (Section 2.2) and a frequency-aware module (Section 2.3), summarized over time by a dilated temporal module (Section 2.4), and classified by a small head. Throughout training, the network is supervised by a frozen Perch v2 teacher (Section 2.5); at inference, the teacher and all distillation-only components are removed, resulting in a lightweight and deployable network.

### 2.1. PCEN Front-End

From the 3-second waveform, we employ an STFT (24 kHz, window 4096, hop 512) and map it through a 40-band mel filterbank spanning 50–12000 Hz, yielding a mel-energy map $E(t, f) \in \mathbb{R}^{40\times 133}$. Instead of a logarithm compression, as in the original baseline, we apply the Per-Channel Energy Normalization (PCEN) [3, 4] dynamic compression, as defined as

$$\mathrm{PCEN}(t, f) = \left(\frac{E(t, f)}{(\varepsilon + M(t, f))^{\alpha}} + \delta\right)^{r} - \delta^{r}, \qquad (1)$$

where $M(t, f) = (1 - s)\, M(t - 1, f) + s\, E(t, f)$ is a first-order infinite impulse response (IIR) filter, with parameters $s = 0.145$, $\alpha = 0.8$, $\delta = 10$, $r = 0.25$. The output is min–max normalized per clip to $[0, 1]$. Furthermore, PCEN suppresses stationary background and equalizes per-channel loudness, improving generalization and robustness against detrimental factors.

### 2.2. Gabor-Initialized 2-D Convolution

The network's first layer is a $9 \times 9$ two-dimensional convolution (32 channels, no bias) initialized as a bank of Gabor spectro-temporal receptive fields (STRF) rather than randomly. Each kernel is a Gaussian envelope modulated by an oriented sinusoid, with the bank spanning a grid of spectral scales and temporal rates so that, at initialization, the filters respond to localized spectro-temporal modulation patterns [5, 6, 7] (i.e., joint frequency–time structure such as harmonic spacing, frequency sweeps, onset/offset rhythm) that characterize bird vocalizations. Moreover, the kernels are mean-removed and $\ell_2$-normalized but remain fully trainable, so the layer serves as a strong prior that the model can refine on task data.

### 2.3. Frequency-Aware Module

The STRF response is processed by a frequency module that contracts the spectral axis while preserving the full temporal axis. It applies ReLU, a frequency-only $\mathrm{MaxPool}(2, 1)$, then two $3 \times 3$ convolutions ($32 \to 64 \to 64$) dilated along frequency only (dilation $(2, 1)$, padding $(2, 1)$), each followed by ReLU, with a second $\mathrm{MaxPool}(2, 1)$ in between. Frequency dilation cheaply widens the spectral receptive field, while each layer retains all 133 frames. A final $\mathrm{AdaptiveAvgPool2d}(1, \mathrm{None})$ collapses the residual frequency dimension, producing a $(B, 64, 133)$ feature map, where $B$ is the batch size. Unlike a joint frequency–time global pooling, the temporal structure is preserved for the next stage.

### 2.4. Temporal Module

The 64-channel sequence is modeled by four residual blocks of depthwise-separable dilated 1-D convolutions. Each block applies the following operations:

$$x \leftarrow x + \mathrm{ReLU}\big(\mathrm{GN}(\mathrm{PW}(\mathrm{DW}_d(x)))\big), \tag{2}$$

where $\mathrm{DW}_d$ is a depthwise `Conv1d` (kernel 7, groups 64, dilation $d$), PW a pointwise $1 \times 1$ convolution, and GN GroupNorm (8 groups). Dilations $d \in \{1, 2, 4, 8\}$ give a receptive field of $\approx 1.9$ seconds, covering call-level rhythm.

The sequence is reduced by temporal statistics pooling to a 128-d vector, classified by $\mathrm{Dropout}(0.05) \to \mathrm{Linear}(128, 32) \to \mathrm{ReLU} \to \mathrm{Linear}(32, 11)$. The network has 81.5k parameters.

### 2.5. Knowledge Distillation

Herein, we use the Perch 2.0 model as our Teacher model, which consists of an EfficientNet-B3 backbone for bird sound classifier. Clips are resampled from 24 to 32 kHz; the teacher's 1536-d embedding and the logits of an 11-class linear head trained on our labels are computed offline and cached. The student is trained with three objectives:

1. Logits distillation: $\lambda T^2 \,\mathrm{KL}\big(\sigma(z_t/T) \,\|\, \sigma(z_s/T)\big)$
2. Cross-entropy (Label Smoothing): $(1 - \lambda)\,\mathrm{CE}_\epsilon(z_s, y)$
3. Embedding hint: $\beta\big[\|g(h) - \tilde{e}\|_2^2 + (1 - \cos(g(h), \tilde{e}))\big]$

Here, we use $\lambda = 0.5$, $T = 3$, $\epsilon = 0.1$, $\beta = 1$. On our notation, $z_s$, $z_t$ are the student and teacher logits, respectively, $h$ the student's 128-d pooled features before the classification head, $g$ a training-stage-only $\mathrm{Linear}(128, 1536)$ projection, and $\tilde{e}$ the teacher embedding. Note that $g(\cdot)$ is discarded at inference.

Table 1: Validation results (549 clips). Each row adds to the row above. The teacher (Perch 2.0 with an 11-class linear head, 32 kHz, frozen foundation model) is a non-deployable upper-bound.

| # | System | Acc. | AUC | Params |
|---|---|---|---|---|
| — | Teacher (Perch 2.0) | 0.8925 | 0.9925 | — |
| 0 | Baseline CNN (no KD) | 0.5628 | 0.8931 | 97k |
| 1 | + Logit distillation | 0.5974 | 0.9099 | 97k |
| 2 | + PCEN front-end | 0.6497 | 0.9296 | 97k |
| 3 | + Embedding hint | 0.6952 | 0.9402 | 97k |
| 4 | + New architecture | 0.7274 | 0.9459 | **81k** |
| 5 | + All data (XC + TAU) | **0.7650** | **0.9539** | **81k** |

### 2.6. External Data

Lastly, given the small size of the original development set, we augment the distillation training set with Perch-labeled external audio, each clip carrying its own cached teacher targets. We collect focal Xeno-canto (XC) recordings of the ten target species, sliced to 3 s at 24 kHz, and retain only clips on which the teacher agrees with the source: Perch's argmax must match the source species with confidence $\geq 0.8$. This agreement filter yields 86.3k bird clips. We further add 10k urban-noise clips for the background class from TAU Urban Acoustic Scenes 2022 [8]. To bridge the domain shift between XC and the original data, we rely on teacher labels rather than the original labels from the XC website. The submitted final model is trained on all available external data.

## 3. EXPERIMENTS AND RESULTS

### 3.1. Setup

We follow the evaluation protocol of the challenge, and the results here reported are on the defined held-out development set (549 clips). We report Top-1 accuracy and macro ROC-AUC. Across the different configurations, we try to keep the training parameter fixed as follows: Adam optimizer, batch size of 16 samples, 120 epochs, cosine schedule with 5-epoch linear warmup, peak learning rate of $10^{-3}$, no augmentation, and best accuracy checkpoint.

### 3.2. Ablation

Table 1 traces the system from the provided baseline to the submitted model, adding one component at a time. Each row is cumulative.

Three patterns stand out. First, the largest gains come from the input representation and the teacher rather than from added capacity: logit distillation, the PCEN front-end, and the embedding hint together raise accuracy from 0.563 to 0.695 at a fixed $\sim$97k parameters. Second, the architectural changes improve accuracy *while reducing* the parameter count from 97k to 81k, indicating that the gain stems from the frequency/time inductive bias rather than from model size. Third, Perch-labeled external data yields the final submitted model, which reaches $0.765$ accuracy and $0.954$ macro ROC-AUC, a 20.2-point accuracy improvement over the baseline with 16% fewer parameters, recovering roughly 70% of the teacher's accuracy advantage at a tiny fraction of its cost.

## 4. CONCLUSION

We presented Lite-Perch, a compact bird sound classifier for the BioDCASE-Tiny 2026 Task 3 challenge. The system pairs knowledge distillation from a frozen Perch 2.0 teacher with signal-processing priors built directly into the network: a PCEN front-end, a Gabor-initialized spectro-temporal stem, and separate frequency- and time-aggregation modules. On the validation set it reaches 0.765 Top-1 accuracy and 0.954 macro ROC-AUC at 81.5k parameters. Our ablations indicate that the gains come primarily from the input representation and the teacher rather than from capacity, and that the frequency/time separation improves accuracy while shrinking the model.

Although not fully explored in the report, heavy augmentation, recurrent heads, and deeper temporal modules did not lead to final performance improvements. A natural next step is to close the remaining embedded-consistency gap by replacing the firmware's integer log-mel front-end with the fixed-point PCEN implementation and measuring performance on the target hardware.

## 5. REFERENCES

[1] B. van Merriënboer, V. Dumoulin, J. Hamer, L. Harrell, A. Burns, and T. Denton, "Perch 2.0: The bittern lesson for bioacoustics," *arXiv preprint arXiv:2508.04665*, 2025.

[2] M. Hagiwara, "Aves: Animal vocalization encoder based on self-supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

[4] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2018.

[5] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[6] M. R. Schädler, B. T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134–4151, 2012.

[7] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "{LEAF}: A learnable frontend for audio classification," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=jM76BCb6F9m

[8] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Scenes and Events 2018 Workshop (DCASE2018)*, p. 9.