Foundation-Model Comparison Report

Audio encoder bake-off for animal-vocalization clustering

Grizzly Systems — GrizCam Desktop · Generated 2026-05-01 11:55

Executive Summary

This report documents a quantitative comparison of audio **foundation-model encoders** evaluated for clustering field-recorded animal vocalizations. The aim is to discover intra-species acoustic structure — individual signature, pack identity, dialect, behavioral context — without explicit labels for those axes.

Each candidate encoder produces fixed-length per-window feature vectors, which we pool to one vector per annotation, project to 3D via UMAP (McInnes et al. 2018), and cluster with HDBSCAN (Campello et al. 2013; McInnes et al. 2017). We score each configuration on supervised accuracy (linear-probe on *sound_type* labels, 5-fold CV) and unsupervised cluster quality (V-measure, silhouette, intra-cluster cosine), plus held-out negative controls (AMI vs *aru_id*, hour, month).

Dataset

285 annotations across **10** autonomous recording units (ARUs), spanning **2024-12-12** to **2025-01-01**. Mean call duration **5.31 s** (p10–p90: 3.79–6.91 s, max 9.48 s).

Sound type	n	Mean (s)	Median (s)	p10–p90 (s)	Max (s)
howl	238	5.43	5.43	3.79–7.02	9.48
bark howl	23	4.54	4.58	3.40–5.37	6.01
growl howl	13	4.82	4.89	3.87–5.69	6.01
wavy howl	11	4.83	4.90	3.87–5.14	5.45

Per-sound-type duration percentiles. "p10–p90" = 10th to 90th percentile range, capturing the typical duration spread without single-clip outliers.

Headline findings

• Highest **linear-probe accuracy**: **Perch · band-pool · DBS** at **0.95**. The probe is a 5-fold cross-validated logistic regression on sound_type labels — it measures whether the embedding space is linearly separable by call type.

• Highest **V-measure (cluster–label agreement)**: **BEATs · band-pool** at **0.39**. V-measure scores how well the unsupervised HDBSCAN cluster assignments align with sound_type labels; a different winner from the probe means the encoder produces structure that's labelable but not naturally clustered along that axis.

• **BEATs · mean+std** showed the strongest **site-signature signal** (AMI vs aru_id = **0.61**). High AMI(aru_id) means the encoder picks up microphone / site acoustics in addition to vocalization content. This is useful for individual / pack fingerprinting (different sites = different individuals) but a confound for pure call-type clustering across sites.

• Per-class F1 winners (linear probe, by sound_type):
  • *bark howl* (n=23): best F1 = **0.91** via **Perch · band-pool**
  • *growl howl* (n=13): best F1 = **0.69** via **AVES · mean+std**
  • *howl* (n=238): best F1 = **0.99** via **Perch · band-pool · DBS**

- *wavy howl* (n=11): best F1 = **0.70** via **BEATs · band-pool**

• Fastest configuration on this dataset: **Perch · band-pool · DBS** at **0 s**. Many cells benefit from cached per-clip embeddings; per-encoder steady-state speed is reported in the Method section.

What this means in practice. Different encoders compress different aspects of the input — acoustic-event structure, taxonomic identity, fundamental-frequency contour, harmonic stack — into the embedding. The encoder choice is a *scientific decision*, not just an engineering one. An encoder that maximizes supervised *sound_type* accuracy may not be the right encoder for unsupervised individual-fingerprinting work, and vice versa. The Method section below details how each encoder was tuned; the Results section presents the head-to-head numbers.

Method

Pipeline overview

Each annotation passes through six stages: (1) audio load + sample-rate match, (2) LUFS or raw-audio normalization, (3) optional Dynamic Background Subtraction (DBS), (4) encoder forward pass, (5) per-clip pool, (6) UMAP projection + HDBSCAN clustering + linear-probe evaluation. We hold every parameter fixed across cells except the three the matrix sweeps: **encoder**, **pool variant**, and the **DBS flag**.

Fixed UMAP settings: **n_components=3**, **n_neighbors=15**, **min_dist=0.10**, **metric=cosine**. Fixed HDBSCAN: **min_cluster_size=5**. Random seed **42** throughout (UMAP, HDBSCAN, LogisticRegression). Per-ARU centering is OFF for this sweep — enabling it would conflate site-effect removal with the encoder comparison axis.

Audio preprocessing

Sample-rate matching. Audio loaded at native rate from the source ARU file (16 kHz mono in this dataset). When the selected encoder requires a different rate (Perch wants 32 kHz), we polyphase-resample via *scipy.signal.resample_poly* at the source-file level and cache the resampled array (LRU, size 4) so consecutive annotations from the same source file don't re-resample the multi-minute WAV.

LUFS loudness normalization (ITU-R BS.1770-4). The annotation slice plus a 2 s pre-context + 1 s post-context window is gain-corrected to –**23 LUFS** integrated loudness in a single pass before slicing the annotation back out, so the whole window shares one gain factor and the relative amplitude between annotation and surrounding context is preserved (critical for DBS). Annotations whose LUFS gain hits the safety bounds (clipped, too quiet, silent) are dropped from the run with a logged reason.

Per-encoder LUFS opt-out. Perch was trained on raw *librosa.load* defaults (no normalization), so forcing LUFS on its inputs shifts them out of the training distribution. The encoder registry carries an *expects_raw_audio* flag; clustering branches on it to skip LUFS for Perch and apply it to AVES, BEATs, AVES2.

Dynamic Background Subtraction (DBS, optional)

DBS is a per-clip denoising step. For each annotation we extract a **1.5 s pre-context window** from immediately before the annotation start, embed it through the same encoder, and **subtract that pre-context vector from every per-window annotation vector** before pooling. The intent: remove site / microphone / ambient signal that the encoder would otherwise carry through into the call's embedding, leaving a residual that should be more call-specific.

DBS is quality-gated. The pre-context window is rejected (and DBS skipped, falling back to plain pooling) if any of these hold: **(a)** insufficient context — the annotation starts before 1.5 s into the source file; **(b)** overlap — another annotation on the same source file intersects the pre-context window; **(c)** clipped — pre-context peak ≥ 0.989 (would inject digital-clipping artifact into the subtraction); **(d)** silent — pre-context dBFS < –75 dB (no ambient signal worth subtracting; the subtraction would be near-zero anyway).

Cells with DBS-on use a **separate cache namespace** from DBS-off so toggling the flag never returns stale embeddings. Both flags appear in the matrix so a reader can see whether DBS helps or hurts a given encoder — the answer is encoder-dependent, see the Results table and Executive Summary.

Pool variants

Encoders emit **(n_windows, embed_dim)** per annotation. Two pool strategies appear in the sweep:

mean + std — concatenates the per-window mean and per-window standard-deviation vectors, doubling the embedding dimensionality (e.g. 768 → 1536 for AVES / BEATs / AVES2; 1536 → 3072 for Perch). The mean captures the average content; the std captures temporal variability across windows (modulation, harmonic structure stability). Concat preserves both, which a downstream linear probe can read.

band-energy-weighted (band_energy) — weights each window's vector by the per-window energy inside the annotator-drawn frequency box, then mean-pools the weighted vectors. The frequency box (low_freq_hz, high_freq_hz) drives a 4th-order Butterworth band-pass filter applied per window; energy = RMS of the band-pass output. Effect: windows where the call's frequency band is acoustically dominant contribute more to the pooled vector than windows of mostly-silence or competing-band energy. Falls back to mean+std when the freq box is missing or zero-width.

Per-encoder window strategy

AVES (HuBERT-style raw-audio): 1 s windows with 0.5 s hop, ingests raw float32 mono. **BEATs** (patch-ViT, AudioSet-pretrained): same 1 s / 0.5 s hop but the model has a ~10 s receptive field with attention masking that handles short clips. **AVES2 / EAT**: fixed 1024-frame mel-patch grid (~10 s). Clips shorter than 10 s zero-pad up; clips longer truncate. **Perch v2**: fixed 5 s window (160 000 samples at 32 kHz), 2.5 s hop. We slide each encoder's window across the annotation slice, batch all windows of one clip into a single forward call (AVES / BEATs), and additionally batch up to 32 clips per *model.embed* call for Perch.

Linear probe (supervised evaluation)

scikit-learn LogisticRegression (*solver=lbfgs, max_iter=2000, C=1.0, class_weight=balanced*) trained on the per-clip pooled embedding to predict *sound_type*. Evaluation: **5-fold StratifiedKFold cross-validation** with random seed 42; class minimum-membership constraint of **5 clips per class** (classes below the threshold are dropped from the probe with a logged reason). Reported *accuracy* is the cross-validated out-of-fold accuracy; per-class *precision*, *recall*, and *F1* are the cross-validated means. The probe is a measure of **linear separability** in the embedding space — high probe accuracy means a simple linear classifier can separate the classes; it does NOT mean the unsupervised clusters align with those labels (see V-measure).

Cluster quality metrics

V-measure — harmonic mean of homogeneity (each HDBSCAN cluster contains only one sound_type) and completeness (each sound_type lives in a single cluster). 0 = no relation; 1 = perfect agreement. Differs from probe accuracy: V-measure scores the *unsupervised* partition; probe accuracy scores *linear separability under supervision*. An encoder can have high probe accuracy and low V-measure when its embedding space is labelable but its natural cluster structure is shaped by something other than the label.

Silhouette score on the embedding space (scikit-learn cosine metric). Higher = better-separated clusters. Pathological signal: a silhouette near 1.00 combined with intra-cluster cosine near 1.00 indicates embeddings have collapsed to a near-point; clusters are trivially "separated" because everything is in one tight ball. We report intra-cluster cosine alongside silhouette to surface this failure mode.

AMI vs aru_id (held-out negative control). Adjusted Mutual Information between HDBSCAN cluster assignments and ARU site identifier. AMI is corrected for chance (unlike raw mutual information). Use: a high AMI(aru_id) value indicates the encoder is picking up site / microphone signature in addition to vocalization content. For pure call-type clustering this is a confound; for individual / pack fingerprinting it is a feature. AMI(month), AMI(hour), AMI(year) are similar negative controls computed from the source-file timestamp.

Encoders evaluated

AVES (aves-base-bio)

Hagiwara, Earth Species Project, ICASSP 2023

HuBERT-style self-supervised encoder trained on the FSD50K-BIO + AudioSet-BIO subset (~360 h, multi-taxa: mammals, birds, anurans, insects). 768-d output, 16 kHz mono, 1 s native window. No species labels in pretraining, so transfer is structure-driven.

Measured per-clip wall-clock (warm, cache-miss): **35 ms/clip**; one-time model-load cost: **5 s**.

ESP-AVES2-EAT-BIO

Earth Species Project; Chen et al. (EAT), EAT: ICASSP 2024 (arxiv:2401.03497)

Efficient Audio Transformer backbone (Chen et al. 2024) with ESP's bioacoustic fine-tune. Fixed 1024-frame (~10 s) mel patch grid; shorter clips zero-pad. 768-d output, 16 kHz.

Measured per-clip wall-clock (warm, cache-miss): **110 ms/clip**; one-time model-load cost: **8 s**.

BEATs iter3

Chen, Wang, Wu, Ge, Wei, et al.; Microsoft Research Asia, NeurIPS 2022

Patch-ViT trained on AudioSet (~5.8 k h, multi-domain) with masked acoustic modeling. 768-d output, 16 kHz, ~10 s receptive field. Strong on within-clip acoustic-event discrimination.

Measured per-clip wall-clock (warm, cache-miss): **142 ms/clip**; one-time model-load cost: **3 s**.

Perch 2.0

van Merrienboer et al.; Google DeepMind, arxiv:2508.04665 (August 2025)

EfficientNet-B3 trunk + prototype-classifier head trained on ~1.5 M recordings spanning ~14,795 species (birds + mammals + amphibians + insects + environmental). 1536-d output, 32 kHz, 5 s native window. SOTA on BEANS bioacoustic benchmark.

Measured per-clip wall-clock (warm, cache-miss): **318 ms/clip**; one-time model-load cost: **5 s**.

Tuning notes

Window length matters per encoder. AVES (raw audio, no fixed input length) and BEATs (attention-masked patch-ViT) tolerate variable durations. AVES2 and Perch operate on a fixed mel-patch grid; clips shorter than the grid (10 s for AVES2, 5 s for Perch) get zero-padded, which can dilute the embedding when most of the input is silence. With ~5 s mean wolf-call duration, AVES2's 10 s grid receives ~50 % zero-pad per call; Perch's 5 s grid fits nearly tight.

Multi-threading is conditional on batch size. Pre-Phase-2 measurements showed multi-thread + batch=1 *anti*-optimizing TF-SavedModel encoders by up to 3× (per-call dispatch / oneDNN-thread-coordination overhead dominates at batch=1, with no parallel work to offset). Threading is only beneficial when paired with batched inference, which is now the default for all encoders. Threading is capped at **75 % of logical CPU cores** per desktop-app policy.

Cache-key autobinding. The per-clip embedding cache key includes encoder ID, model version, pool mode, frequency-band coordinates, source-file size, AND registry-derived window/hop/sample-rate. Changing any of those in the registry automatically invalidates stale embeddings without manual version bumps — means a preprocessing tweak can't silently reuse old embeddings.

Reproducibility. All runs use random seed 42 for UMAP, HDBSCAN, and LogisticRegression. Cells with cached embeddings are bit-identical to fresh runs (modulo ~1e-7 oneDNN reduction-order noise on encoders with multi-threaded forward). Embedding cache lives at

<input_folder>/cache/clustering/embeddings/<encoder>_<pool>[_dbs]/<hash>.npy; projection cache (UMAP+HDBSCAN result) lives at *<input_folder>/cache/clustering/projections/<hash>.pkl*. Both are content-addressed and survive across sessions.

Results

16 of **16** sweep cells completed successfully. Cells are ranked by linear-probe accuracy on sound_type labels, with V-measure as a tiebreaker.

Rank	Configuration	Probe	V-meas	Sil	AMI(aru)	Wall (s)
★ #1	Perch · band-pool · DBS	0.95	0.04	0.45	0.08	0
#2	Perch · mean+std · DBS	0.95	0.04	0.45	0.08	159
#3	Perch · band-pool	0.95	0.23	0.35	0.59	0
#4	Perch · mean+std	0.95	0.23	0.35	0.59	82
#5	BEATs · band-pool	0.93	0.39	0.50	0.61	0
#6	BEATs · band-pool · DBS	0.93	0.04	0.92	0.06	0
#7	BEATs · mean+std · DBS	0.93	0.04	0.92	0.06	37
#8	BEATs · mean+std	0.92	0.38	0.48	0.61	36
#9	AVES · mean+std	0.92	0.10	0.55	0.47	44
#10	AVES · mean+std · DBS	0.92	0.04	0.83	0.06	45
#11	AVES · band-pool · DBS	0.91	0.04	0.81	0.06	45
#12	AVES · band-pool	0.91	0.38	0.63	0.55	44
#13	AVES2 · band-pool	0.61	0.10	0.55	0.28	0
#14	AVES2 · mean+std	0.61	0.10	0.55	0.28	50
#15	AVES2 · band-pool · DBS	0.60	0.04	0.91	0.06	1
#16	AVES2 · mean+std · DBS	0.60	0.04	0.91	0.06	38

Probe = 5-fold CV linear-probe accuracy on sound_type. V-meas = V-measure (HDBSCAN cluster vs label agreement). Sil = silhouette score (higher ≈ better-separated clusters; 1.00 indicates near-collapsed embeddings). AMI(aru) = adjusted mutual information vs aru_id (held-out negative control; high values flag site/microphone signature pickup). Wall = embed-loop wall-clock.

Per-class probe metrics — Perch · band-pool · DBS (winner)

Sound type	n	Precision	Recall	F1
bark howl	23	0.91	0.91	0.91
growl howl	13	0.69	0.69	0.69
howl	238	0.99	0.99	0.99
wavy howl	11	0.58	0.64	0.61

*Precision / recall / F1 of the linear probe on each sound_type label, computed as the mean across the 5-fold cross-validation. **n** = total annotations of that type in the dataset.*

Per-cell 2D UMAP scatter

Visual sense of how each encoder × pool × DBS combination organizes the same annotation set in 2D UMAP space. Color encodes HDBSCAN cluster ID; grey points are noise (annotations that didn't fall into any dense region). Cells with similar scatter topologies are producing similar embedding-space structure.

References & Further Reading

Encoder homepages and primary papers for further research. Each link below opens externally.

AVES (aves-base-bio)

Animal Vocalization Encoder via Self-Supervision

Hagiwara, Earth Species Project — ICASSP 2023

Paper: https://arxiv.org/abs/2210.14493

Code: https://github.com/earthspecies/aves

Model card: https://github.com/earthspecies/aves#pretrained-models

ESP-AVES2-EAT-BIO

Earth Species Project AVES2 with EAT backbone, bioacoustic fine-tune

Earth Species Project; Chen et al. (EAT) — EAT: ICASSP 2024 (arxiv:2401.03497)

Paper: https://arxiv.org/abs/2401.03497

Code: https://github.com/earthspecies/avex

Model card: https://huggingface.co/EarthSpeciesProject/esp-aves2-eat-bio

BEATs iter3

Audio Pre-Training with Acoustic Tokenizers

Chen, Wang, Wu, Ge, Wei, et al.; Microsoft Research Asia — NeurIPS 2022

Paper: https://arxiv.org/abs/2212.09058

Code: https://github.com/microsoft/unilm/tree/master/beats

Model card: https://github.com/microsoft/unilm/tree/master/beats#models

Perch 2.0

Perch 2.0 - The Bittern Lesson for Bioacoustics

van Merrienboer et al.; Google DeepMind — arxiv:2508.04665 (August 2025)

Paper: https://arxiv.org/abs/2508.04665

Code: https://github.com/google-research/perch-hoplite

Model card: https://www.kaggle.com/models/google/bird-vocalization-classifier

Algorithms

McInnes, L., Healy, J., Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection. arxiv:1802.03426

Campello, R., Moulavi, D., Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates (HDBSCAN). doi:10.1007/978-3-642-37456-2_14

Project documentation

Grizzly Systems Desktop — the application that produced this report. Project source repository: github.com/Grizzly-Systems-dev/grizcam_desktop