HAC-Net: Learning Natural Units from Acoustic Change

Quentin Bacquelé[1,2*], **Jean-Yves Barnagaud**[2], **Frédéric Theunissen**[3], **Nicolas Mathevon**[1,4,5]

[1]ENES Bioacoustics Research Lab, CRNL, CNRS, Inserm, University of Saint-Etienne
[2]CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD
[3]Department of Neuroscience, University of California, Berkeley
[4]École Pratique des Hautes Études - PSL, CHArt Lab, University Paris-Sciences-Lettres
[5]Institut universitaire de France

*qbacquele@gmail.com

Abstract

Animals structure their vocalizations around acoustic change points: boundaries where one element transitions to another or context shifts. These transitions reflect underlying production mechanisms and guide receiver perception. Yet most bioacoustics analyses still rely on predefined categories, energy-based rules, or generic audio codecs that ignore these natural boundaries. We propose HAC-Net, a method that discovers acoustic units by learning where patterns change in continuous recordings. The model reconstructs audio from boundaries it identifies, forcing it to place cuts at genuine transitions. It works hierarchically, finding both fine elements and larger structures. We expect this method to yield biologically grounded segmentation that supports discovery of meaningful variation and provides units suitable for sequence modeling. The resulting units will enable large-scale comparative studies across species without expert annotations, providing a consistent foundation for analyzing compositional structure, temporal organization, and downstream ecological applications.

1 Motivation

Animals produce vocal sequences that contain short stretches of relative stability in their spectrotemporal statistics, separated by abrupt changes when a new element begins or when context shifts [8]. Receivers must detect these change points to parse meaning from continuous signals [6, 12], tracking when one syllable ends and another begins, or when a territorial call shifts to courtship display [11]. Yet current analytical pipelines may not align with these units defined by shared encoder–decoder constraints [2, 15].

Therefore, the primary methodological obstacle is this unit-of-analysis dependency [5, 7]. If the goal is discovery rather than confirmation, the units themselves must be inferred from structure present in the sound. We adopt a change-based definition aligned with this principle: a unit is the longest span over which the acoustic pattern remains consistent, and a boundary occurs where that pattern shifts clearly enough to mark the onset of a new gesture or context.

This definition emerges from both production and perception constraints [15]. During vocalization, animals switch between distinct motor control patterns by engaging different syringeal or laryngeal muscles [9, 14], modulating airflow [9, 16], or repositioning articulators [10]. These production-side transitions create acoustic discontinuities. Simultaneously, receivers track local statistical

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

regularities to anticipate upcoming sounds; when these regularities break, attention shifts to process new information [13]. The boundaries where production mechanisms change thus coincide with perceptual reset points.

Recent advances in sequence modeling with learnable segmentation allow us to operationalize this framework directly on continuous audio [1, 3]. Instead of fixing units in advance, we detect content-driven changes as the signal unfolds while controlling the average number of cuts, so that comparisons across species, sites, and sensors are made at matched granularity. This moves beyond static tokenization toward dynamic segmentation that adapts to local signal statistics.

We aim to (i) discover acoustic units directly from continuous audio by learning boundaries where the acoustic pattern changes; (ii) test whether the resulting units support biologically relevant distinctions across species; and (iii) provide scalable, repertoire-free indicators suitable for comparative studies.

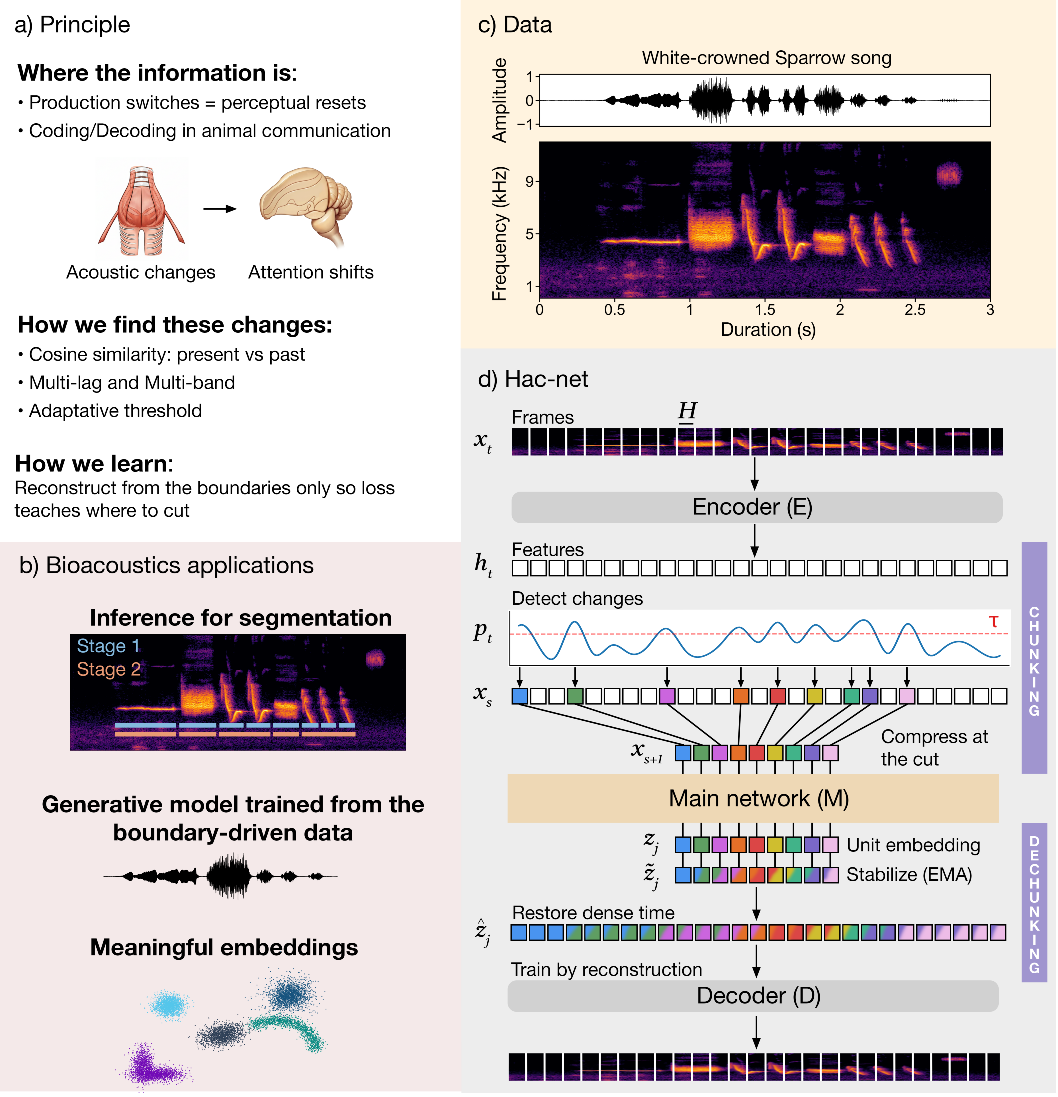

Figure 1: **HAC-Net: boundary-driven learning for bioacoustics. (a)** Principle: HAC-Net targets acoustic changes and learns cut placement by reconstructing from boundary-driven codes. **(b)** Uses: segmentation (units, motifs), optional synthesis from unit streams, and embeddings for clustering/probes. **(c)** Example White-crowned Sparrow input (waveform and spectrogram). **(d)** Detailed HAC-Net pipeline.

2 Methods

2.1 Proposed Framework: HAC-Net (Hierarchical Acoustic Chunking Network)

2.1.1 Design Objective and Scope

Goal. Building on the core H-Net architecture [1], we learn to reconstruct waveform or spectrogram from a boundary-driven, compressed representation such that minimizing reconstruction/spectral losses pressures the model to place boundaries at genuine acoustic change points. The model provides (i) **boundary times and durations** and (ii) **unit-level embeddings** suitable for discovery, clustering, and downstream bioacoustic analyses.

Architecture. HAC-Net (Hierarchical Acoustic Chunking Network) is a causal Encoder → Chunking(Router/Downsample) → Main → Dechunking(EMA/ Upsample) → Decoder. Stage 2 applies the same mechanism over the unit sequence to form motifs.

2.1.2 Notation (frames vs. units)

Audio is framed with hop H seconds. Let $x_{1:T}$ be frame inputs and h_t encoder features. The router emits change scores $p_t \in [0,1]$ (pre-gate; see S4). An adaptive threshold τ_t with a causal minimum duration of $D_{\min}$ frames yields boundary indicators $b_t \in \{0,1\}$ at times $t_{(j)}$ ($b_{t_{(j)}}$=1). Define a non-learned sentinel start at $t_{(0)}$:=1 and compute rates excluding the sentinel:

$$u_j \;=\; x_{\,t_{(j)}:\,t_{(j+1)}}, \qquad F^{(1)} = \frac{1}{T-1}\sum_{t=2}^{T} b_t, \qquad F^{\star(1)} = \frac{1}{N^{(1)}}, \qquad r^{(1)} = \frac{F^{(1)}}{H}.$$

Here $N^{(1)}$ is the target average interval length in frames (expected frames per unit). Controller statistics use pre-gate scores p_t; boundary decisions use post-gate scores $\tilde{p}_t$ (S4).

2.1.3 Core Architecture (left-to-right)

(1) Frame features (Encoder). A causal encoder E maps frames to features $h_t = E(x_t)$ that summarize local spectral shape and short-term modulation. Either a raw-waveform front-end or causal mel patches can be used (Supplement S3).

(2) Chunking: detect change and compress. *Change detection.* The router assesses whether the current acoustic pattern departs from the immediate past by comparing the present frame's features to short-lag references and aggregating evidence across several small time lags and frequency bands. Using cosine dissimilarity makes the detector largely insensitive to gain changes and background variation, while multi-lag cues capture rapid onsets and multi-band cues emphasize band-local events (e.g., a new harmonic, a frequency jump). A slowly adapting threshold controls the local cut rate, and a causal minimum-duration rule prevents micro-cuts. The result is a sequence of hard boundaries at times $t_{(j)}$ (with b_t=1 at those frames).

Compression. At each boundary, we retain a single boundary-anchored vector (e.g., a small local average around the cut) and discard interior frames, yielding a boundary-indexed sequence $\{r_j\}$ (one vector per unit), typically with far fewer elements than T. This step both decides *where to cut* and performs the actual downsampling; the resulting compact stream is the input to the main network and defines the units used in subsequent analysis.

(3) Main on the compressed stream. The main network M maps each boundary-indexed vector to a unit embedding

$$z_j \;=\; M(r_j)$$

These embeddings summarize unit-level content (e.g., band-specific energy distribution, frequency trajectory context, modulation cues). The choice of M is modular (small SSM/CNN stack, conformer, transformer).

(4) Dechunking: smooth units and restore dense time. *Stabilization.* Unit embeddings are smoothed by a causal EMA weighted by boundary confidence P_j (a downsampled function of router scores):

$$\hat{z}_j \;=\; P_j z_j \;+\; (1-P_j)\,\hat{z}_{j-1}.$$

High-confidence cuts rely more on z_j; lower-confidence cuts inherit more from $\hat{z}_{j-1}$, reducing boundary jitter.

Restoring the frame rate. We then upsample by holding the smoothed unit vector constant until the next boundary,

$$\tilde{z}_t \;=\; \hat{z}_j \qquad \text{for } t \in [t_{(j)},\, t_{(j+1)}),$$

yielding a piecewise-constant dense-time stream aligned exactly to unit edges. During training a confidence-weighted straight-through estimator ensures frame-level losses propagate through the upsampling to the unit decisions (Supplement S4). Operationally: copy the current unit code to all frames until a new unit starts.

(5) Decoder and training losses. A causal decoder D consumes a dechunked stream together with a small feature-domain residual to preserve fine detail:

$$y_{\text{pred}}(t) \;=\; D(U(\tilde{z}_t) + R\, h_t)\,,$$

where U maps dechunked codes to the decoder input domain (waveform or mel) and R is a 1×1 linear map initialized near zero. The objective combines content fidelity with global granularity and spacing control,

$$\mathcal{L} \;=\; \underbrace{\mathcal{L}_{\text{recon}} + \alpha\,\mathcal{L}_{\text{spec}}}_{\text{content fidelity}} \;+\; \underbrace{\mathcal{L}_{\text{ratio}}}_{\text{granularity control}} \;+\; \underbrace{\mathcal{L}_{\text{refr}}}_{\text{minimum-duration prior}}\;.$$

Loss components are defined in S6. Because D sees only the boundary-driven stream, improving reconstruction necessarily rewards boundaries placed at genuine acoustic transitions.

2.1.4 Hierarchy and Rate Control

Animal vocalizations are hierarchically organized (e.g., phrases comprising multiple notes). We implement hierarchy by *instantiating HAC-Net recursively at the main-network level*: the boundary-indexed stream produced by Stage 1 becomes the input to a second HAC-Net operating over unit index rather than frame index.

- **Stage 1** identifies fine-scale acoustic regimes (e.g., individual notes or syllables);
- **Stage 2** groups these elements into higher-level motifs (e.g., phrases or call types).

Both stages operate causally and share the same boundary detection mechanism, but at different temporal scales determined by their respective compression ratios. This hierarchy emerges purely from the acoustic structure, without predefined templates, allowing discovery of species-specific organizational principles.

2.1.5 Outputs and Bioacoustic Use

HAC-Net model (embeddings for analysis). We train a causal model to reconstruct vocalizations from a *boundary-driven compressed representation*, so that improving reconstruction demands placing cuts at genuine acoustic changes. The trained model exposes continuous embeddings at the unit (Stage 1) and, if enabled, motif (Stage 2) levels ($\{\hat{z}_j\}$, $\{\hat{z}^{(2)}_k\}$). These embeddings can be used directly for downstream analyses such as clustering into putative note/syllable or phrase types, linear probes for species/site/individual information, and sequence modeling over continuous vectors.

Acoustic units (boundaries and durations). After training, inference runs the encoder and router to produce boundary times $\{t_{(j)}\}$ and unit durations. These boundaries are emitted by the router's decision rule (change score, adaptive threshold, and minimum duration). They can be exported as is for rate statistics, bout/phrase structure, and alignment with behavioral annotations.

2.1.6 Contribution

We retain H-Net's core components (cosine routing, causal smoothing, confidence-weighted straight-through upsampling, ratio loss) and adapt them to continuous bioacoustic audio with: (i) multi-lag, multi-band routing, (ii) adaptive thresholding, (iii) a causal minimum-duration rule, and (iv) band-aware spectral objectives.

2.2 Evaluating Acoustic Units in Bioacoustic Contexts

2.2.1 Comparison with Traditional Segmentation Methods

Fixed-frame segmentation divides audio into uniform windows (25, 50, or 100 ms).

Energy-adaptive segmentation (deterministic) using multi-scale features (flatness, bandwidth, roll-off, entropy, energy, ZCR), silhouette-based window selection, merging short silences, low-SNR removal, and energy-envelope refinement with LOF filtering;

Neural-codec tokenizer (SoundStream) [4] with induced segments from code frames;

Random boundaries provide a statistical baseline, showing the information available from arbitrary segmentation at the target rate.

2.2.2 Unit-Level Evaluation

Reconstruction accuracy. We measure held-out reconstruction error (MSE) at matched boundary rates. Methods that place boundaries at natural transition points should achieve lower error, yielding favorable rate–distortion curves (compute and latency matched).

Biological information content. We test whether discovered units capture species-specific information through linear probing: after segmenting audio with each method, we train linear classifiers to predict species identity from unit embeddings or traditional acoustic features (frequency contours, spectral shape and modulation patterns). We report accuracy and macro-F1 at matched boundary rates.

2.2.3 Repertoire-Level Evaluation

After unit validation, we analyze repertoires that emerge when units are clustered into types (learned embeddings and acoustic features). The number of clusters is selected via information criteria.

Statistical coherence. We verify that vocal types form genuine, stable clusters by running clustering 100 times with 80% bootstrap samples and measuring the Adjusted Rand Index between solutions.

Biological relevance. For each recording, we compute a repertoire profile: the frequency distribution of vocal types used. We test whether these profiles can discriminate between species using a linear classifier with cross-validation.

3 Expected Impact and Significance

By extracting functional acoustic units from content-driven change rather than from fixed protocol, we expect to recover organization in animal communication that is less sensitive to segmentation heuristics and more closely tied to perceptual and behavioral regularities. This improved grounding should enhance the performance of sequence models that parse fine-scale vocal structure, including language model approaches.

The approach is designed to be general and scalable across taxa and recording conditions. For broader applications, it could provide standardized, repertoire-free indicators for acoustic biogeography, conservation, and rare-species monitoring. Automating boundary discovery also reduces manual annotation effort and yields reusable unit inventories that experts can build upon.

This work is a tool for discovery at the level of units, repertoires, and macro-patterns, not a replacement for hypothesis-driven bioacoustics. Behavioral experiments, expert-curated repertoires, and neurophysiological evidence remain necessary to establish ultimate meaning and function. Under these constraints, the expected impact is a more objective and functionally grounded basis for measuring and modeling the organization of animal vocal communication.

Acknowledgments

Funding. This research has been supported by the University of Saint-Etienne, the École Pratique des Hautes Études, the University Paris-Sciences-Lettres (PhD stipend to QB), the Labex CeLyA, the CNRS, the Inserm, the Institut universitaire de France (NM), and the ACOUCENE group (CESAB, French Foundation for Research on Biodiversity, French Ministry of Ecological Transition, French Office for Biodiversity).

Competing interests. The authors declare no competing interests.

References

[1] S. Hwang, B. Wang, and A. Gu. Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv preprint arXiv:2507.07955*, 2025.

[2] L. Zandberg, V. Morfi, J. M. George, D. F. Clayton, D. Stowell, and R. F. Lachlan. Bird song comparison using deep learning trained from avian perceptual judgments. *PLoS Computational Biology*, 20:e1012329, 2024.

[3] P. C. Bermant, L. Brickson, and A. J. Titus. Bioacoustic event detection with self-supervised contrastive learning. *bioRxiv*, 2022.

[4] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, M. Tagliasacchi. SoundStream: An End-to-End Neural Audio Codec. *arXiv*, 2021.

[5] K. J. Odom, M. Araya-Salas, J. L. Morano, et al. Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews*, 96:1135–1159, 2021.

[6] P. Yin, D. L. Strait, S. Radtke-Schuller, J. B. Fritz, and S. A. Shamma. Dynamics and hierarchical encoding of non-compact acoustic categories in auditory and frontal cortex. *Current Biology*, 30:1649–1663.e5, 2020.

[7] R. R. K. V. S. N., J. Montgomery, S. Garg, and M. Charleston. Bioacoustics data analysis – a taxonomy, survey and open challenges. *IEEE Access*, 8:57684–57708, 2020.

[8] A. Kershenbaum, D. T. Blumstein, M. A. Roch, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 91:13–52, 2016.

[9] C. P. H. Elemans, J. H. Rasmussen, C. T. Herbst, et al. Universal mechanisms of sound production and control in birds and mammals. *Nature Communications*, 6:8978, 2015.

[10] A. M. Taylor and D. Reby. The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, 280:221–236, 2010.

[11] K. Arnold and K. Zuberbühler. Semantic combinations in primate calls. *Nature*, 441:303, 2006.

[12] S. H. Hulse. Auditory scene analysis in animal communication. In *Advances in the Study of Behavior*, volume 31, pages 163–200. Elsevier, 2002.

[13] M. D. Hauser, E. L. Newport, and R. N. Aslin. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, 78:B53–B64, 2001.

[14] F. Goller and R. A. Suthers. Role of syringeal muscles in controlling the phonology of bird song. *Journal of Neurophysiology*, 76:287–300, 1996.

[15] T. Guilford and M. S. Dawkins. Receiver psychology and the evolution of animal signals. *Animal Behaviour*, 42:1–14, 1991.

[16] R. S. Hartley and R. A. Suthers. Airflow and pressure during canary song: direct evidence for mini-breaths. *Journal of Comparative Physiology A*, 165:15–26, 1989.

Supplementary: HAC-Net for Continuous Audio

S1. Notation and indices (frames vs. samples)

Sampling rate f_s; hop H; window $W \geq H$. Frame index $t = 1{:}T$:

$$t \;=\; 1 + \left\lfloor \tfrac{n-1}{H\,f_s} \right\rfloor.$$

Per frame t: input frame x_t, encoder feature $h_t \in \mathbb{R}^d$, router score $p_t \in [0,1]$. Lag set $\mathcal{L} \subset \mathbb{N}_+$ and causal subset $\mathcal{L}^+(t) = \{\ell \in \mathcal{L} : \; t - \ell \geq 1\}$. Initialize a *sentinel start* $t_{(0)} := 1$ (not a learned boundary). Define event indicators $b_t \in \{0,1\}$ for $t \geq 2$.

Boundaries (hard events) occur at $t_{(j)}$ where $b_{t_{(j)}} = 1$. Units are half-open sets

$$u_j \;=\; \{\, t\!: \; t_{(j)} \leq t < t_{(j+1)} \,\}, \qquad j = 1{:}J.$$

Rates and targets. Exclude the sentinel when reporting rates:

$$F^{(1)} \;=\; \frac{1}{T-1}\sum_{t=2}^{T} b_t, \quad r^{(1)} \;=\; \frac{F^{(1)}}{H}, \quad F^{\star(1)} \;=\; \frac{1}{N^{(1)}},$$

where $N^{(1)}$ is the *target average frames per unit* (interval length).

S2. Architecture and residual path

Causal stack:

$$\underbrace{\textbf{Encoder } E}_{x_t \mapsto h_t} \;\to\; \underbrace{\textbf{Chunking}}_{\text{router + downsample}} \;\to\; \underbrace{\textbf{Main } M}_{r_j \mapsto z_j} \;\to\; \underbrace{\textbf{Dechunking}}_{\text{EMA + upsample}} \;\to\; \underbrace{\textbf{Decoder } D}.$$

Residual (feature-domain). Following H-Net [1], inject a small linear projection of encoder features at the *dechunk* output:

$$\tilde{y}_t \;=\; U(\tilde{z}_t) \;+\; R\,h_t,$$

where U maps dechunked codes to the decoder input domain (waveform or mel), and R is a learnable 1×1 linear map (zero-initialized). The decoder outputs $y_{\text{pred}}(t) = D(\tilde{y}_t)$. This keeps domains aligned and mirrors the H-Net residual placement.

S3. Audio input and strict causality

Frames x_t use either: (i) left-aligned STFT (window W, hop H), mel projection, log compression; or (ii) raw waveform windows of length $W\,f_s$. All components are strictly causal, all lags satisfy $\ell \geq 1$.

S4. Dynamic chunking on audio

Router (multi-lag, multi-band). For bands $b = 1{:}B$ and causal lags $\ell \in \mathcal{L}^+(t) = \{\ell \in \mathbb{N}_+ : t - \ell \geq 1\}$,

$$q_{t,\ell}^{(b)} = W_q^{(b)} h_t, \quad k_{t-\ell}^{(b)} = W_k^{(b)} h_{t-\ell}, \quad p_t^{(b,\ell)} = \tfrac{1}{2}\left(1 - \frac{\langle q_{t,\ell}^{(b)}, k_{t-\ell}^{(b)}\rangle}{\max(\|q_{t,\ell}^{(b)}\|, \delta)\;\max(\|k_{t-\ell}^{(b)}\|, \delta)}\right) \in [0,1].$$

Context weights $w_{t,b} = \mathrm{softmax}_b(W_w h_t)$ and $a_{t,\ell} = \mathrm{softmax}_\ell(W_a h_t)$ give the pre-gate score

$$p_t = \begin{cases} \displaystyle\sum_{\ell \in \mathcal{L}^+(t)} \tilde{a}_{t,\ell} \sum_{b=1}^{B} w_{t,b}\, p_t^{(b,\ell)}, & \mathcal{L}^+(t) \neq \varnothing, \\ 1, & \mathcal{L}^+(t) = \varnothing, \end{cases} \qquad p_1 := 1.$$

Energy gate (decision-only) and clamping. Let $e_t = \log(\epsilon + \|x_t\|_2^2)$ and $g_t = \sigma(\eta(e_t - \theta_E))$ with a floor $g_{\min} \in [0.05, 0.2]$:

$$\tilde{p}_t = \mathrm{clip}(\max(g_t, g_{\min})\, p_t, \; \varepsilon, \; 1 - \varepsilon), \quad \varepsilon = 10^{-4}.$$

Controller statistics (below) use pre-gate p_t; boundary decisions use post-gate $\tilde{p}_t$.

Adaptive threshold (local rate control). With running mean on pre-gate scores

$$\bar{p}_t = (1-\rho)\bar{p}_{t-1} + \rho\, p_t,$$

set

$$\tau_t = \mathrm{clip}\Big(\sigma(\theta_\tau) + \mathrm{sg}\,[\kappa(\bar{p}_{t-1} - F^{\star(1)})],\ \tau_{\min},\ \tau_{\max}\Big),$$

targeting $G^{(1)}_{\text{pre}} \approx F^{\star(1)}$. Recommended ranges: $\rho \in [10^{-3}, 10^{-2}]$, $\kappa \in \{0.5, 1, 2\}$, $(\tau_{\min}, \tau_{\max}) \in \{(0.2, 0.8), (0.3, 0.7)\}$.

Minimum duration (refractory) and events. Let $t_{(j)}$ be the last boundary time and $r_t = \mathbf{1}\{t - t_{(j)} \ge D_{\min}\}$. Boundaries:

$$b_t = \mathbf{1}\{\tilde{p}_t \ge \tau_t\} \cdot r_t, \qquad t \ge 2.$$

Straight-through events and cumulative indices. Define the STE event surrogate

$$\tilde{b}_t = \mathbf{1}\{\tilde{p}_t \ge \tau_t\} + \mathrm{sg}\big(\sigma(\beta(\tilde{p}_t - \tau_t)) - \mathbf{1}\{\tilde{p}_t \ge \tau_t\}\big), \quad \beta \in [5, 20],$$

and cumulative unit index $u_t = \sum_{\tau=2}^{t} \tilde{b}_\tau$ (STE in backprop).

Refractory penalty (on events; optional).

$$\mathcal{L}_{\text{refr}} = \gamma \sum_{t=2}^{T} \sum_{\Delta=1}^{D_{\min}-1} \tilde{b}_t\, \tilde{b}_{t-\Delta}, \quad \gamma \in \{0, 10^{-4}, 10^{-3}\}.$$

Boundary readout, smoothing, upsampling, STE aggregation. A causal micro-pool of width $w \in \{3, 5\}$ ending at $t_{(j)}$ yields r_j, then $z_j = M(r_j)$ and

$$\hat{z}_j = P_j z_j + (1-P_j)\hat{z}_{j-1}, \qquad P_j := \tilde{p}_{\,t_{(j)}},\ \hat{z}_0 := 0.$$

Hold constant within each unit:

$$\tilde{z}_t = \hat{z}_{u_t} \quad \text{for } t \in [t_{(j)}, t_{(j+1)}).$$

Frame-level gradients are aggregated over $\{t : u_t = j\}$ and scaled by P_j to reduce jitter near uncertain cuts.

S5. Hierarchy (Stage 2)

Stage-2 operates on the *unit index*. Inputs are projected smoothed unit embeddings:

$$v_j \;=\; W^{(2)} \hat{z}_j, \quad j = 1{:}J.$$

Routing, events, EMA, and dechunking mirror Stage-1 under $t \mapsto j$, producing coarse boundaries $j_{(k)}$ and motif embeddings $z_k^{(2)} \to \hat{z}_k^{(2)}$. Targets use $F^{\star(2)} = 1/N^{(2)}$ where $N^{(2)}$ is *target units per motif*. Stage-2 statistics:

$$F^{(2)} = \frac{1}{J-1} \sum_{j=2}^{J} b_j^{(2)}, \quad G^{(2)} = \frac{1}{J} \sum_{j=1}^{J} p_j^{(2)}.$$

Stage-2 supervision. Add a light auxiliary prediction of per-motif *band-energy summaries*. For each motif k, predict $\bar{s}_k^{(2)} = A(\hat{z}_k^{(2)})$ and minimize an L1 loss to the empirical average mel magnitude pooled over $j \in [j_{(k)}, j_{(k+1)})$. Weight $\alpha^{(2)} \in [0.05, 0.2]$. This supplies training signal without duplicating the full decoder.

S6. Objectives and global rate control

Reconstruction and spectral fidelity.

$$\mathcal{L}_{\text{recon}} = \|y_{\text{pred}} - y_{\text{true}}\|_2^2, \qquad \mathcal{L}_{\text{spec}} = \sum_r \sum_{b=1}^{B} \omega_b \,\Big\|\, |S_r^{(b)}(y_{\text{pred}})| - |S_r^{(b)}(y_{\text{true}})| \,\Big\|_1,$$

with ω_b normalized (softmax).

Global granularity (ratio loss; H-Net form). For stage $s \in \{1, 2\}$, let $F^{(s)}$ be the empirical boundary fraction (excluding the sentinel) and let $G^{(s)} := \frac{1}{L_s} \sum p^{(s)}$ be the *pre-gate* mean score. With $F^{\star(s)} = 1/N^{(s)}$,

$$\mathcal{L}_{\text{ratio}}^{(s)} = \frac{N^{(s)}}{N^{(s)} - 1} \Big[(N^{(s)} - 1)\, F^{(s)} G^{(s)} + (1 - F^{(s)})(1 - G^{(s)}) \Big], \qquad \mathcal{L}_{\text{ratio}} = \sum_{s=1}^{S} \lambda_s\, \mathcal{L}_{\text{ratio}}^{(s)}.$$

Use $\lambda_s \in [0.02, 0.2]$ with the adaptive threshold of S4. Report stability diagnostics (sliding-window variance of $F^{(s)}$ and $|F^{(s)} - G^{(s)}|$) in robustness checks.

Total objective.

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \alpha\, \mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{ratio}} + \alpha^{(2)} \mathcal{L}_{\text{aux}}^{(2)} + \mathcal{L}_{\text{refr}}.$$

S7. Predictive efficiency and rate–distortion

Let $\bar{F} = F^{(1)}$, $J_{\bar{F}} = \text{round}((T - 1)\bar{F})$. Baselines *enforce* $D_{\min}$ and *exclude* the sentinel:

- **Contiguous:** place $J_{\bar{F}}$ cuts as evenly as possible on $\{2, \ldots, T\}$ with gaps $\geq D_{\min}$ (distribute remainder from the start).
- **Random:** sample $J_{\bar{F}}$ unique indices from $\{2, \ldots, T\}$, then prune to satisfy $D_{\min}$ by greedy repair.

Define

$$\text{PE}_{\text{contig}}(\bar{F}) = \frac{\mathcal{J}_{\text{contig}(\bar{F})} - \mathcal{J}_{\text{seg}}}{(T - 1)\, \bar{F}}, \quad \mathcal{J} \in \{\mathcal{L}_{\text{recon}},\ \mathcal{L}_{\text{recon}} + \alpha\, \mathcal{L}_{\text{spec}}\}.$$

Report $(\bar{F},$ held-out objective) curves with 95% bootstrap CIs over recordings.

S8. Unit embeddings and biological probes

Stage-1 probe embedding is $e_j = \hat{z}_j$. For species/site probes, train linear classifiers on $\{e_j\}$ with recording-level splits. Ablations: boundary readout r_j; mean-pooled frames over u_j.

S9. Repertoire discovery and temporal organization

Cluster $\{e_j\}$ with k-means; select k by BIC or silhouette. Stability: 100 bootstrap runs (80% subsamples), report Adjusted Rand Index. For sequence structure, fit bigram models over cluster labels; compare held-out perplexity to within-recording shuffles preserving unigram frequencies.

S10. Boundary-rate calibration for bioacoustics

Estimate $F^{\star(1)}$ on a small annotated set; pick $F^{\star(1)}$ that jointly maximizes predictive efficiency and minimizes boundary-count error. For cross-species studies, sweep $F^{\star(1)}$ on a validation mix and keep one global value for matched-rate comparisons. Keep hop H fixed across species or also report mean physical unit durations. Set $D_{\min}$ from annotations or literature and tune $(\rho, \kappa, \tau_{\min}, \tau_{\max})$ to minimize count error while keeping $|F^{(1)} - G^{(1)}|$ small (all on $\tilde{p}_t$).

S11. Dataset pipeline

Exclude clips outside $[0.25, 1.0]$ s. Cap per species to balance classes. Normalize amplitudes with a fixed global scale. Batching pads to a fixed frame budget and carries original lengths and metadata. Splits are by recording ID.

S12. Training and inference

AdamW (lr 2×10^{-4}), cosine decay, batch size 32, gradient clip 1.0. Streaming inference maintains encoder state, the last keys $\{k_{t-\ell}\}$ for $\ell \in \mathcal{L}$, and the current $\hat{z}_j$. End-to-end latency equals the encoder receptive field plus W plus one frame.

S13. Robustness checks

Noise stress: add pink noise at SNR $\{5, 10, 20\}$ dB and measure $\Delta F^{(1)}$ and $\Delta \mathrm{PE}$. Gain invariance: apply random gains in ± 6 dB. Band-limit: low-pass and high-pass to verify boundary stability under channel variation.

S14. Failure modes and diagnostics

Over-segmentation: $F^{(1)} > F^{\star(1)}$ with low confidence; mitigation: increase α (spectral loss) or lower encoder bandwidth. Under-segmentation: long units with reconstruction spikes; mitigation: raise $F^{\star(1)}$ or increase encoder capacity. Boundary drift near sweeps: strengthen EMA (reduce P_j) or use windowed readout for r_j.

S15. Complexity and resources

Router time $O(T\, d\, |\mathcal{L}|\, B)$; main/decoder $O(Td)$. Average memory $O(TF^{(1)} + d\, |\mathcal{L}|)$ per stage from boundary-indexed streams and lag states (worst-case $O(T + d\, |\mathcal{L}|)$ at high boundary rates). Report wall-clock throughput and parameter count. Baselines are compute/latency matched in rate–distortion plots.

S16. Differences from generic hierarchical models

Continuous frame-level projection replaces token embeddings; segmentation is learned from acoustic change; objectives prioritize temporal and spectral fidelity; boundary-rate calibration is tied to bioacoustic unit scales. Stage 2 operates on $v_j = W^{(2)} \hat{z}_j$ for stability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This is a proposal, claims are limited to a tokenizer-free, causal boundary-learning architecture (HAC-Net), its training objective, and an evaluation plan—without performance claims. See § Motivation and § Methods: Proposed Framework.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We state assumptions and failure modes (Supplement S10, S13, S14) and note that training/evaluation are pending, so generalization and quantitative comparisons are not yet established.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The submission contains no theorems or formal proofs, it presents an algorithmic framework and objectives with implementation details.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: No experiments have been run yet. We provide a reproducibility plan and intended settings (dataset pipeline S11, training S12, calibration S10). Full scripts and exact settings will accompany results in a later version.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This proposal reports no experimental results. Data will come from Xeno-Canto (public, contributor-specific CC licenses). Code exists and will be released with instructions after anonymization when results are added.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Experiments are not yet executed. Planned settings are outlined (S11–S13), but final splits/hyperparameters and selection procedures will be reported with results.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No empirical results are reported in this proposal. Significance reporting will accompany future experiments.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Experiments are pending; compute details (hardware, wall-clock, total FLOPs) will be documented when results are included.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public non-human animal audio: no human subjects. We will respect per-recording licenses and attribution for Xeno-Canto and discuss responsible use (e.g., avoiding harmful playback) in Broader Impacts/ethics text.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: § Expected Impact and Significance outlines positive uses (comparative bioacoustics, monitoring) and notes that behavioral meaning requires external validation. We will add risks such as misuse of synthetic playback and mitigation in the final version.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No high-risk assets are released in this proposal. If a generative model is released later, we will include watermarking, usage guidelines, and gated access as appropriate.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We will use Xeno-Canto recordings with contributor-specific Creative Commons licenses and will credit recordists and specify license terms in the supplemental material and dataset section. Preprocessing will preserve attribution and licensing metadata.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets (datasets or released model checkpoints) are introduced in this proposal. Code release is planned with results.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The work does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human-subjects research is conducted.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [NA]

    Justification: LLMs are not part of the core methodology.