Shared acoustic manifolds for exploratory comparison of passerine vocalizations

Lucio Arese
Independent Researcher, Italy
contact@lucioarese.net
ORCID: 0009-0009-2019-7326

This study presents a fixed-parameter pipeline designed to support reproducible embedding of frame-level representations of multiple passerine vocalizations within shared low-dimensional spaces. Three passerine species are considered: Eurasian Wren, Tree Pipit and Common Chaffinch, with a selection of four individuals for each species group. Vocalization frames from each species group are mapped into a single three-dimensional coordinate system to allow comparison between individuals while preserving temporal continuity. The pipeline operates under a controlled protocol with an unsupervised, geometry-first exploratory approach. Two feature representations are used: MFCC (40 coefficients with delta and delta-delta) and 80-bin chroma vectors. The two feature sets provide complementary analytical lenses on the signal, ranging from spectral-envelope dynamics to relative frequency organization, without imposing discrete musical categories. The dimensionality-reduction process features a PCA-20 preconditioning step followed by a UMAP embedding, resulting in a total of six manifolds (two feature spaces × three species). The resulting embeddings are visualized as continuous trajectories in two separate layouts: a view with individual identity separated by solid coloring and another augmented view with descriptor overlays as color coding, applied post-embedding. The descriptors include spectral centroid and a chroma-derived concentration measure (Chroma Energy Concentration or CEC, introduced in this work), visualized as scalar fields on the manifold geometry. A supplementary case study demonstrates event-level backtracking from localized manifold regions to the underlying audio, enabling identification of recurring vocal events concentrated in specific embedding regions. The framework operates independently of labeling or categorization: it provides a descriptive interface intended to complement spectrogram-based analysis, supporting qualitative comparison and hypothesis generation.

1 Introduction

Animal vocalizations form a highly diverse and multi-faceted acoustic domain. In order to analyze it, the discipline of bioacoustics has developed and established practices relying on spectrogram inspection with the use of acoustic descriptors such as fundamental frequency, bandwidth and measures of spectral complexity[1]. These approaches created vast catalogs of repertoires and classifications, and they have been widely useful for purposes such as species identification and comparative behavioral studies[2, 3]. At the same time, it is common practice to summarize complex vocalizations using descriptor sets and segment-level measurements, leaving their per-frame structure only indirectly accessible and, in some cases, less systematically explored. Recent unsupervised pipelines address this by constructing frame-level representations and latent spaces for exploratory analysis[4].

From a purely acoustical standpoint, reducing a continuous, highly structured signal into an

array of discrete numbers tends to emphasize certain aspects of that signal while compressing others[5, 6]. As a result, its fine-grained spectral and temporal evolution may be difficult to inspect as a continuous object, and it could become a sensible limitation when the aim is not simply to characterize a single recording, but to compare vocalizations across individuals while preserving continuity, revealing recurrent motifs, and making differences visually assessable without forcing early interpretation.

Recent advancements in computational methods provide new opportunities to reorganize such frame-level detail into a coherent geometric space. Dimensionality-reduction techniques such as UMAP[7] (and related approaches such as t-SNE[8]) enable embedding of high-dimensional vectors into low-dimensional spaces with the aim of preserving local neighborhood relations. Applying such mappings to time-ordered sequences may let continuous trajectories emerge, with spatial proximities reflecting similarities in acoustic structure. In bioacoustics, related approaches have increasingly used latent or embedded representations for visualization, comparison, and exploration of vocal repertoires[9, 10, 4]. However, their use to compare multiple individuals within a shared geometric space remains less standardized: normalization choices, fitting strategies, and interpretive conventions may vary between implementations[4, 11].

This study develops and documents a fixed-parameter analysis pipeline for mapping frame-level acoustic features into continuous, low-dimensional trajectories, with extensive parameter description to support reproducibility. The exploratory work unfolds through a case study covering three different passerine species, each one including four different individuals. Frame-level representations of their vocalizations are embedded into a common three-dimensional space using two complementary feature sets: MFCC (timbral structure)[12] and chroma vectors (octave-wrapped log-frequency amplitude profiles, 80 bins per octave)[6, 13]. The resulting manifolds are presented as explicitly exploratory representations and the vocalizations are treated as unlabeled acoustic signals. No behavioral patterns are inferred, nor new taxonomic groupings are proposed; the focus is instead on documenting the geometric consistencies emerging between individuals, outlining a visual framework that may support future research in the investigation of vocal structure[4].

By making frame-level organization comparable across individuals and feature spaces, this framework is intended to support future work in hypothesis generation, targeted annotation, and repertoire-level investigation of vocal structure. In this sense, the present study positions an analytical acoustic framework within a bioacoustic context with a methodological and descriptive focus: its low-dimensional embeddings are proposed as an additional inspection interface alongside established spectrogram-based practices, providing an analytical pipeline that could integrate naturally into existing bioacoustic workflows.

2 Methods

The study analyzes vocalizations from three different passerine species: Tree Pipit (Anthus trivialis), Eurasian Wren (Troglodytes troglodytes), and Common Chaffinch (Fringilla coelebs). Each species group contains a selection of four different individuals. The three species present distinct song organizations, ranging from rapid broadband syllable sequences to more tonal, phrase like patterns: overall, they provide a compact test set to evaluate the pipeline across different vocal morphologies, while keeping the design simple and reproducible. Moreover, Eurasian Wren, Tree Pipit and Common Chaffinch are widely recorded in Europe and have substantial publicly available material, which facilitates the selection of multiple individuals with consistent recording metadata and sufficiently long high-quality segments.

All twelve recordings come from the Xeno-canto database[14]. The selection criteria were based on an adequate signal-to-noise ratio, consistent and comparable vocalization patterns and a recording length sufficient to edit a 60 s sequence of vocalizations for each recording. These criteria were applied independently of the geographical origin of the birds, resulting in a set of individuals distributed across different regions of Europe. All selected excerpts correspond to song-type vocalizations as labeled in the source archive. All recording metadata and sources are provided in Supplementary Tables S1–S3.

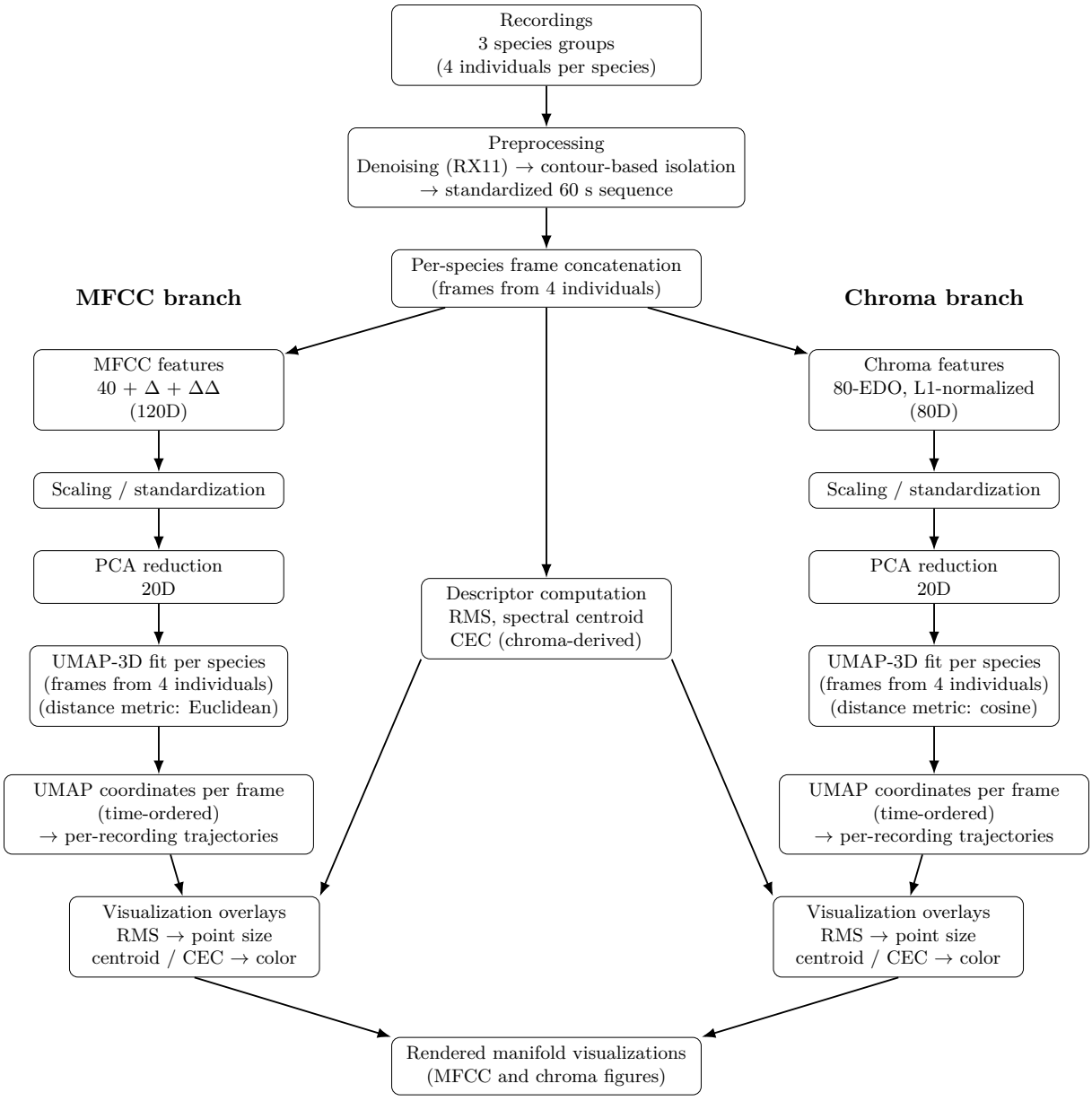

Figure 1: Overview of the analysis pipeline. For each species, frames from four individuals are concatenated to fit shared UMAP embeddings separately for MFCC and chroma features. Embedded frames form time-ordered trajectories per recording. In parallel, per-frame descriptors are computed (RMS and spectral centroid from audio frames; CEC from chroma vectors) and used only as visualization overlays (point size and color), without entering the embedding pipeline.

2.1 Audio preprocessing

Given the heterogeneous field conditions (different recordists, microphones and acoustic environments) variability in spectral coloration and signal-to-noise ratio must be taken into account, especially for timbre-based descriptors such as MFCC. To mitigate these effects, a standardized multistage denoising and signal isolation pipeline was applied to all the recordings, with the aim of reducing background noise and improving spectral clarity to facilitate comparison across recordings. Nevertheless, this process cannot fully eliminate all biases related to the recording conditions.

2.1.1 Denoising procedure

The denoising procedure was conceived as a standardized preprocessing step rather than a signal restoration process. All parameters were defined prior to analysis and applied identically to all recordings, without any file-specific adjustment. This choice was made in order to preserve comparability between recordings, rather than to maximize noise suppression for individual files.

The same preprocessing pipeline was applied to all the recordings, and it was implemented using a spectral denoising workflow on iZotope RX 11. It consisted of five different spectral denoising passes (two using adaptive noise prints and three with manual noise profiles). Denoising passes used both the “Advanced” and "Advanced+Extreme" algorithm mode with FFT window length of 50 ms and multi-resolution enabled, and the parameters were set in order to ensure a balance between noise suppression and preservation of the harmonic and noisy components of the vocalizations. Full parametrization for the denoising procedure is provided in Supplementary Table S4.

2.1.2 Sensitivity checks

In order to verify that the denoising process did not alter the spectral structure, two complementary checks were performed. First, a batch of SNR-proxy metrics was computed on the trimmed raw segments to determine the baseline signal-background separability across recordings (Supplementary Table S5). Second, denoising sensitivity was quantified by comparing MFCC and chroma features extracted from the denoised audio, against the corresponding raw segments evaluated on the same frame set, using the same frame indices defined by amplitude gating on raw frame-RMS (frames in the upper 30% and 70% of the raw frame-RMS distribution; Supplementary Tables S6–S7). The 30% gate emphasizes the highest-RMS vocalization frames, whereas the 70% gate includes a broader set of lower-RMS frames and therefore provides a more stringent test against residual background influence. Sensitivity for MFCC and chroma is reported through cosine similarity statistics (mean with 5th percentile in parentheses) along with descriptor-level shifts in spectral centroid, bandwidth and CEC (a chroma-derived descriptor to be discussed later).

The analysis was performed on two denoised variants for each recording: full5, applying all the denoising passes in order, and noP4, applying the same chain while omitting pass 4, which was the most aggressive pass in the sequence (highest reduction strength and "Advanced+Extreme" algorithm settings; Supplementary Table S4). Results of these checks support that the overall structure of both feature spaces was preserved by the denoising procedure, and that the subsequent manifold organization is not determined by preprocessing artifacts.

2.1.3 Signal isolation and audio standardization

Following the denoising process, the final step consisted of signal isolation through manual tracing of its time-frequency contours for the vocalizations on the spectrogram, replacing the non-vocal regions with digital silence to remove residual background activity. This operation was strictly limited to vocal contours masking and did not modify the waveform within the retained vocal regions, nor the internal spectral or temporal structure of the signal.

After contour-based isolation, a single continuous interval of its isolated vocalizations was used to construct a standardized 60 s sequence, which served as the input for all subsequent frame-level analyses. No discontinuous selection or montage of preferred events was performed. The time ranges used for each recording are reported in Supplementary Tables S1–S3. The end result is a uniformly preprocessed audio segment, suitable for frame-level analysis and ensuring a comparable contribution from each individual to the embedding process. Supplementary figure S7 provides an example of the described workflow, including spectrogram views of the various denoising stages and the final standardized 60 s excerpt.

2.2 Feature computation

2.2.1 MFCC

Mel-Frequency Cepstral Coefficients (MFCC) were used because of their frame-level timbral information and for the compact description of spectral envelope shape they provide, with some reduced sensitivity to overall gain compared to raw spectra[12, 15].

Every audio signal was converted into a sequence of short-time frames with a window size of 4096 samples and a hop size of 384 samples, and those same parameters were applied throughout all the recordings. For each frame, 40 coefficients were computed together with their first- and second-order temporal derivatives (delta and delta-delta), in order to provide a detailed description of the spectral envelope. The resulting 120-dimensional feature space makes it possible for the embedding algorithms to evaluate both the instantaneous spectral shape and its local temporal evolution. Full MFCC parametrization (window type, mel bands, $f_{\min}/f_{\max}$, pre-emphasis, lifter, scaling) is listed in Supplementary Table S8.

2.2.2 Chroma 80-bins

Concurrently, each frame was also represented by an 80-bin chroma vector computed over an 80 equal-division-of-the-octave (80-EDO) grid[6, 13]. Unlike the standard 12-bin chroma representation commonly used in musical analysis, this higher resolution setting was adopted to better capture the fine-grained frequency variations and micro-interval structure characteristic of animal vocalizations, without imposing assumptions of harmonic organization or discrete categories.

The octave-folding here is used as a pragmatic normalization emphasizing relative frequency organization and frequency-modulation patterns, while reducing sensitivity to absolute frequency shifts across individuals and recording conditions. It was conceived as a modeling choice to facilitate cross-individual comparison under heterogeneous recording environments. It should not be interpreted as a perceptual or production-based tuning system, but rather as a high resolution, octave-normalized sampling providing a continuous representation of octave-folded spectral amplitude distribution across frequency-class bins.

The chroma vectors were normalized by their L1 norm, so that only the relative chroma-bin amplitude distribution (and not the global amplitude) can be taken into account in the embeddings. The same window and hop sizes were applied to all recordings (respectively 4096 and 384 samples, consistent with the MFCC feature computation).

Octave folding necessarily removes absolute frequency information, implying that this representation should be used as a complementary view rather than a replacement for absolute frequency descriptors. The complete parametrization for chroma (80-EDO mapping, smoothing, normalization, tuning reference, etc.) is listed in Supplementary Table S8.

Complementary roles of feature sets. By computing both MFCC-based and chroma-based embeddings for each vocalization, it was possible to enable two complementary acoustic perspectives for each recording. MFCC features provide a coherent description of timbral structure (reflecting changes in spectral envelope, noise-related components and short-term temporal modulation), while microtonal chroma features provide an octave-normalized view capturing how amplitude is distributed across frequency bins (after octave folding) and how that distribution evolves over time.

In short, they reflect two distinct yet complementary aspects that make up the signal. Moreover, their reduced sensitivity to absolute amplitude provides a reasonable detachment from the specific recording conditions of each individual audio, mitigating recording-related bias. Additional stabilization is achieved through an intermediate computation step.

2.3 PCA 20-dimensional reduction

Between features computation and non-linear embedding an intermediate dimensionality-reduction step was taken, consisting of the reduction to a Principal Component Analysis (PCA) 20-dimensional space. This step was motivated by two main considerations.

First, as previously stated, the recordings are of heterogeneous nature, coming from different microphones and distances, with different noise floors and spectral coloration. PCA was utilized to concentrate the dominant modes of variation into orthogonal components and to reduce redundancy and small amplitude variation that can destabilize subsequent neighborhood-based embeddings; in this sense, it was intended as a preconditioning stage rather than an information-maximizing compression step. Secondly, given the sensitivity of UMAP to local noise and redundant dimensions, a 20-dimensional PCA reduction can be viewed as a form of smoothing/regularization step in feature space.

2.3.1 Choice of 20 dimensions

The choice of 20 components for the reduction in dimensionality was done as a practical compromise observed across datasets: PCA-20 preserved an average 55.83% variance retention for MFCC features and 86.19% for chroma features (see Supplementary Table S9 for variance retention values with PCA-15/20/30). The lower retention for MFCC is expected, given the higher dimensionality and correlation structure determined by the introduction of its derivative delta and delta-delta features which distributed variance across a larger number of components. In this pipeline, the retention of the dominant modes of variation and the reduction of variability at fine scale is intended to obtain more interpretable shared embeddings for the purposes of this study.

2.3.2 Uniform protocol across MFCC and chroma branches

It is also important to note that the same PCA applied to very different feature sets like MFCC and chroma was not been done with the purpose to make such features comparable to each other, but rather to make each of them internally stable for UMAP embedding under a single controlled protocol. Also, the choice of applying a 20-dimensional reduction to both feature sets relies on the principle that in a preliminary exploratory study a common compromise would be better than having two separately tuned pipelines.

All those reasons combined make a 20-dimensional PCA reduction a reasonable compromise choice in order to provide a common intermediate geometry before UMAP embedding. This procedure was uniformly applied to all recordings for the creation of such intermediate feature space before UMAP embedding, both for MFCC and chroma branches of the pipeline.

2.4 UMAP embedding

Uniform Manifold Approximation and Projection (UMAP) was selected as the non-linear embedding method for this work. Animal vocalizations are complex signals evolving continuously over time, and the embeddings produced by UMAP may exhibit continuous trajectories, since it is designed to preserve local neighborhoods and often yields extended coherent structures in low-dimensional space. It can also handle a large number of frames with practical computational cost, and allows fine-tuning and control over neighborhood scale and minimum distance constraints through its hyperparameters[7]. For these reasons, UMAP was deemed suitable for the purposes of the present study.

2.4.1 Shared manifolds construction

As noted previously, a set of four different individual recordings was processed to compute MFCC and chroma features, with a subsequent intermediate 20-dimensional PCA reduction. Subsequently, a shared embedding was constructed by concatenating all the resulting frames from the various individuals into a single feature matrix: one for MFCC, one for chroma. A UMAP model was fitted over this concatenated dataset, in order to create a three-dimensional common manifold for each species group and feature type[4]. Frames retained the recording IDs, allowing recovery of the trajectories of each individual within the shared geometric space.

UMAP was run with n_neighbors = 30, min_dist = 0.1, n_components = 3, and random_state = 42. For MFCC embeddings, the Euclidean metric was used; for chroma embeddings, the cosine metric was applied. All embeddings used a shared scaling and normalization procedure (including a global unit-cube normalization) as specified in Supplementary Table S8, which reports the full parameter set and command-line arguments.

No Procrustes alignment or post-hoc rotation was applied to the embeddings.

2.4.2 Cosine and Euclidean metrics

As a final note, a brief explanation is provided about the choice of Euclidean distance for the MFCC embeddings and cosine for the chroma ones, since it is the only intentional divergence in an otherwise parameter-shared pipeline. As MFCC vectors encode the spectral envelope shape and its temporal evolution, Euclidean metrics were used to outline the absolute differences between coefficient dimensions. In contrast, chroma vectors are L1-normalized and represent an

octave-folded amplitude distribution across frequency-class bins[6], so their relative orientation carries more importance than absolute magnitude; therefore, the cosine distance provides a proper metric to compare chroma orientations in feature space. Such choices were considered reasonable for the feature representation according to their intrinsic structure. Alternative distance metrics and other parameter configurations were explored in the present work, as the goal was to maintain methodological consistency rather than parameter optimization.

2.5 Spectral descriptors

In parallel with the dimensional reduction steps described above, two descriptors were computed and used exclusively for visualization purposes: frame-level amplitude (RMS), used to modulate point size in the manifold renderings, and spectral centroid, used as a scalar field color coding to provide an interpretable frequency-balance anchor.

In addition, a chroma-derived descriptor termed Chroma Energy Concentration (CEC) was computed from the per-frame chroma vector as a measure of concentration in the chroma-bin amplitude distribution (ratio between the maximum chroma bin and the sum of all bins); it is used only as a scalar field visualization and is discussed further in the Results section as well as in Supplementary Materials.

3 Results

As a result of the procedure described above, a total of six shared manifolds (two feature sets across three species) were computed, each containing four individual trajectories evolving within a common low-dimensional space, allowing their visual comparison within the embedded geometry. Each trajectory consists of a sequence of points ordered in time, representing the evolution of the vocalization. Proximity and commonly traversed paths can be read as shared acoustic regimes between individuals, whereas excursions or branching departures reflect shifts in the feature profile over time.

Visualization of the manifolds was obtained by constructing three-dimensional geometric representations from the embedded coordinate data. In these models, each frame is represented by a point whose size is modulated by the RMS amplitude of the signal. Amplitude scaling was applied to harmonize point-size ranges across individuals; all relative amplitude ratios within each recording have been preserved, and no amplitude information was used in the embedding process. Solid coloring differentiates the individual trajectories within each manifold.

In addition to solid-color representations, a second set of three-dimensional visualizations was constructed by overlaying two scalar descriptors on the manifolds. Spectral centroid was included as a familiar visual reference for spectral balance, color-coded on the low-dimensional geometry. Chroma energy concentration (CEC), introduced in this work, is shown as a complementary scalar field quantifying the amplitude concentration (after octave folding) in the chroma-bin amplitude distribution, computed as the ratio of the maximum chroma-bin amplitude to the sum of all bins. These scalar fields are applied after the embedding process and therefore do not influence the manifold geometry.

In the following, a systematic paired description of the MFCC-based and chroma-based embeddings of each species group is provided, to facilitate comparison between the two feature representations and to allow a consistent cross-species comparison under uniform visualization settings. The descriptions that follow are not intended to support functional or behavioral inter-

pretation,nor do they indicate any inferences at the population level, given the limited sample size; all observed patterns are reported as instances of the general behavior of the embedding framework rather than defining characteristics of the examined species.

3.1 Eurasian Wren group

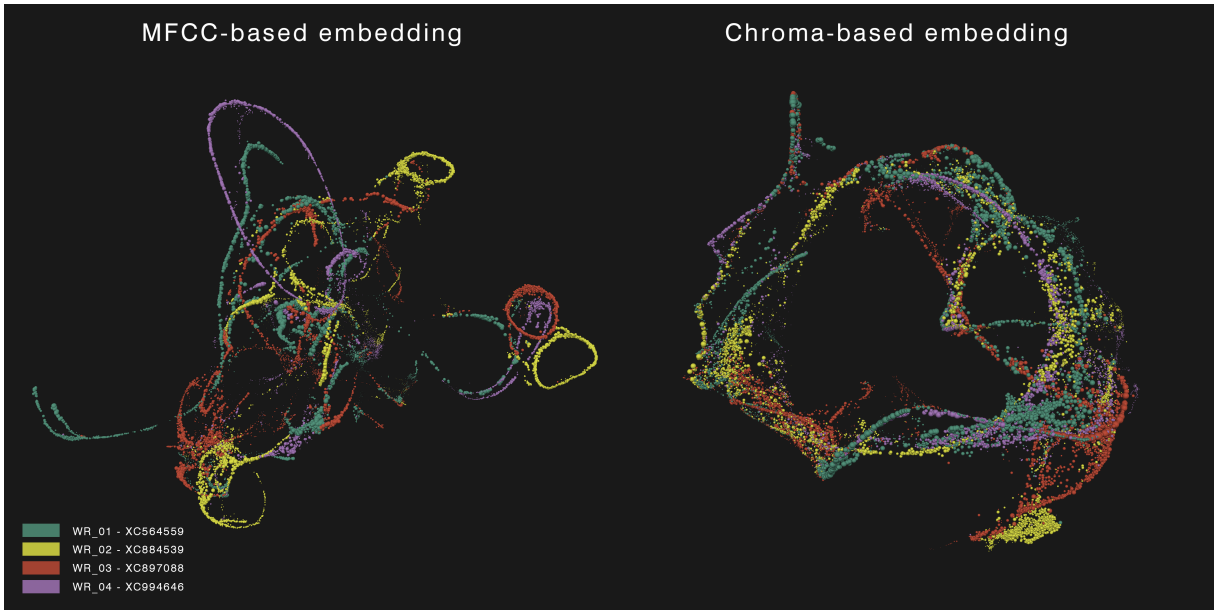

Figure 2: Eurasian Wren group. Paired low-dimensional manifolds computed from frame-level features for four individuals (solid color trajectories; legend reports recording IDs). Left: MFCC-based embedding. Right: chroma-based embedding. Each frame is rendered as a point in the embedded space and ordered in time along the trajectory; point size is modulated by the RMS amplitude of the signal for visualization (relative amplitude dynamics are preserved within each recording and harmonized across individuals).

MFCC-based embedding. The MFCC-based manifold for the Eurasian Wren group displays a structure that is highly articulated and spatially extended (Fig. 2, left), composed of multiple intertwined strands and several loop-like excursions. The internal regions of the embedding show an interconnected core traversed by the various individuals. Each trajectory develops differently elongated peripheral trajectories: they branch away from and, in several cases, rejoin the core. Many loop-like structures from different individuals occur in proximity within the shared space, and areas of partial overlap are observed on some of the strands characterizing the geometry.

Chroma-based embedding. The chroma-based embedding for the Eurasian Wren group reveals a geometry that differs substantially from its MFCC-based counterpart. It exhibits an apparently cohesive three-dimensional organization characterized by a surface-like structure with a pronounced central void, resembling a hollow shell or ring (Fig. 2, right). The four trajectories show substantial overlap across the manifold, with no single region exclusively occupied by one recording, although local density differences and peripheral protrusions are visible.

3.2 Tree Pipit group

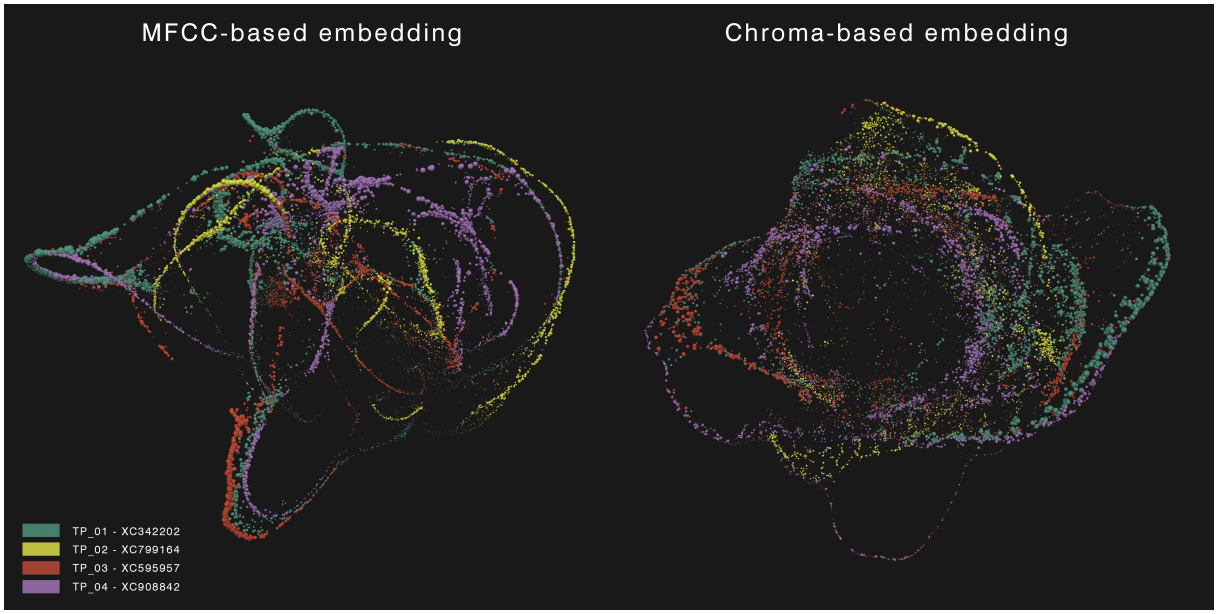

Figure 3: Paired MFCC-based (left) and chroma-based (right) manifolds for four individuals, displayed using the same conventions as Fig. 2. The chroma-based embedding is further examined in the Supplementary case study (Supplementary section S3), which demonstrates a backtracking workflow from geometry to signal, recovering recurring acoustic events from a localized region of the manifold.

MFCC-based embedding. The MFCC-based embedding for the Tree Pipit group forms a broad, interconnected structure with a dense central region and elongated exterior excursions (Fig. 3, left). The four individual trajectories overlap substantially through the core of the manifold, while several peripheral arcs are preferentially occupied by specific individuals. Overall, the geometry is continuous and strand-like, with repeated returns to common paths and several long-range deviations that expand the occupied volume.

Chroma-based embedding. The chroma-based manifold yields a more globally coherent geometry dominated by a curved ring/shell-like structure with four peripheral lobes (Fig. 3, right). Most trajectories co-occupy a common surface, and differences between individuals appear primarily as local departures rather than separate clusters.

3.3 Common Chaffinch group

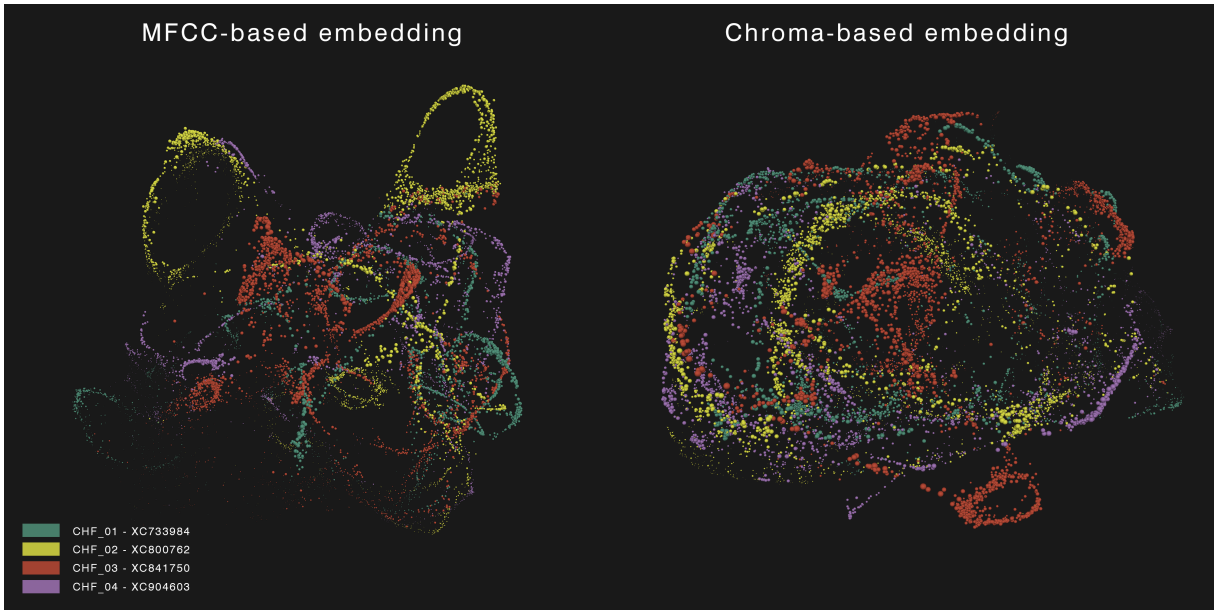

Figure 4: Common Chaffinch group. Paired MFCC-based (left) and chroma-based (right) manifolds for four individuals, displayed using the same conventions as Fig. 2.

MFCC-based embeddings. The MFCC-based embedding for Common Chaffinch forms a loop-rich structure with a densely interwoven interior, where the trajectories extensively overlap (Fig. 4, left). Multiple peripheral loop excursions extend outward and are, in places, preferentially occupied by specific trajectories, indicating localized individual bias within an otherwise continuous manifold.

Chroma-based embeddings. The chroma-based embedding (Fig. 4, right) forms a cohesive, bounded envelope. Strong intermixing is observed across the trajectories. Individual differences appear mainly as localized density concentrations and peripheral protrusions rather than separable regions. In the displayed orientation, a small detached loop-like excursion is prominently occupied by one trajectory, while the remainder of the manifold remains broadly shared.

3.4 Spectral descriptors as scalar fields

In the following visualizations, additional layers of information are applied to the manifolds by mapping descriptor values as color coding for each point, while individual identity is intentionally collapsed. Instead, a single spatial field is created that highlights the descriptor value distribution along the geometries of the embeddings.

Two descriptors were employed: spectral centroid, a widely used measure of the frequency-weighted center of mass of the spectrum, provides a familiar reference and a visual anchor for how spectral balance shifts along the trajectories; chroma energy concentration (CEC) is provided as a complementary view, quantifying how strongly chroma-bin amplitudes are concentrated within the octave-folded representation: higher values indicate a more concentrated chroma profile (greater concentration into fewer bins), while lower values reflect a more distributed profile. Both descriptors are displayed as continuous scalar fields shaped by the geometry of the embedding.

In order to maintain consistency and comparability, fixed global scales were used throughout all visualizations: spectral centroid is displayed over a 2–9 kHz range and CEC over a 0–0.6 range. These ranges were selected to encompass the full distribution of values observed across all individuals and species included in the study, and are applied uniformly in both the main figures and in the descriptor visualizations present in the additional materials.

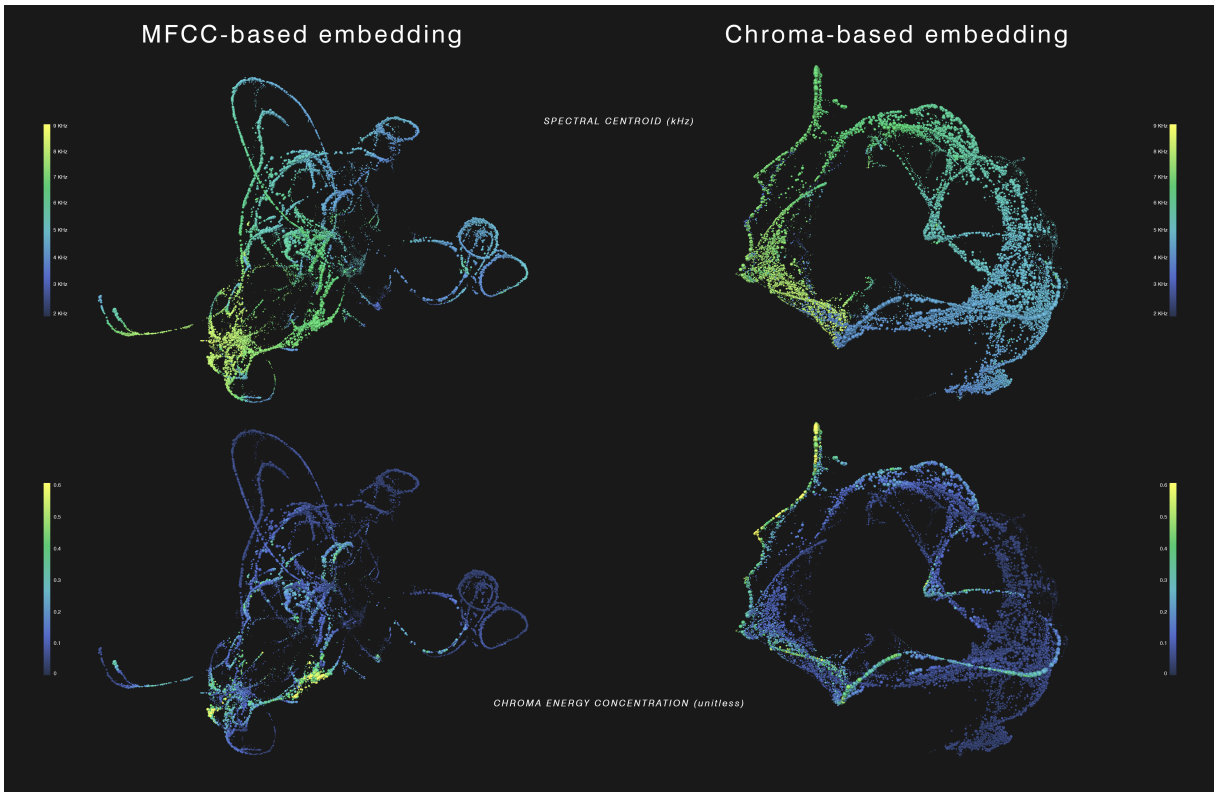

Figure 5: MFCC-based (left) and chroma-based (right, 80-EDO) manifolds colored by spectral centroid (top, 2–9 kHz) and chroma energy concentration (CEC; bottom, 0–0.6). Individual identity is collapsed to emphasize descriptor distributions as continuous scalar fields.

Spectral centroid. Applying spectral centroid as a scalar field to the MFCC-based and chroma-based manifolds of the Eurasian Wren group yields structured, spatially coherent gradients across both embedding geometries (Fig. 5, top row). In the MFCC-based manifold (Fig. 5, top left), elevated centroid values concentrate within the densest region of the embedding and extend along several major arcs departing from it, while lower values occupy broad portions of the remaining trajectories, producing smooth transitions along the manifold paths.

In the chroma-based manifold (Fig. 5, top right), centroid values exhibit a clear large-scale anisotropy around the shell-like geometry: higher values are concentrated along a subset of boundary arcs (notably on the left and upper portions in the displayed orientation), whereas lower values dominate the opposite side and the lower extension, forming gradual gradients around the manifold.

Chroma energy concentration. Chroma energy concentration (CEC) applied as a scalar field highlights a different organization (Fig. 5, bottom row). In the MFCC-based embedding (Fig. 5, bottom left), higher CEC values occur as localized patches and short segments along selected trajectories, while most extended loops and peripheral excursions remain comparatively low. In the chroma-based embedding (Fig. 5, bottom right), elevated CEC values are more

spatially constrained and tend to occur along specific boundary arcs and folded regions, with the majority of the manifold showing lower CEC values.

Overall, these scalar fields reveal complementary distributions across the same geometries: spectral centroid varies smoothly at broad spatial scales, whereas CEC appears more intermittent and localized, with both patterns shaped by the different configurations of the MFCC-based and chroma-based low-dimensional spaces.

3.5 Results extensions in the Supplementary Materials

The Supplementary Materials provide additional visualizations and worked examples for completeness and for qualitative inspection. These items document the analysis workflow and the outputs of the fixed-parameter pipeline in greater detail, without extending the main Results section.

3.5.1 Additional descriptor overlays

In addition to the descriptor-coated views of the Eurasian Wren group illustrated above, the Supplementary Materials provide additional descriptor overlays for the two remaining groups (Supplementary section S2, Fig. S1–S2). These renderings support visual examination of descriptor distributions across the shared geometries under the same fixed global scales (Fig. S3). These views are constructed post-embedding and do not influence the computed manifold geometries.

3.5.2 Supplementary case study

A supplementary case study focused on the Tree Pipit group demonstrates a backtracking workflow from geometry to signal (Supplementary section S3, Fig. S4–S6, Table S10). While the Results section focused on describing the manifold geometries, this analysis provides a concrete example of how shared embeddings can be used as a practical inspection interface. Here, a compact region of the chroma-based embedding was identified by qualitative inspection of the geometry and its descriptor overlays: backtracking from the region of interest to the audio signal recovered a set of 17 short syllable events, recurring across individuals and occurring in two consistent frequency regions.

3.5.3 Supplementary videos

Supplementary Videos V1–V3 provide multi-angle animated views to complement the static figures. The videos include both solid-color, descriptor-coated views, and a detail view of the region of interest (ROI) used in the Supplementary case study, together with the corresponding trajectory segments of the recovered acoustic events.

4 Discussion

4.1 Low-dimensional manifolds as exploratory representations

4.1.1 Fixed-parameter shared embeddings

This study presented a fixed-parameter pipeline for mapping frame-level representations of passerine vocalizations into shared low-dimensional spaces, with extensive parameter documentation to support reproducibility. Three distinct species groups were considered: for each species, frames from four individuals are projected into a single coordinate system under a fixed protocol designed to preserve temporal continuity and stable geometry between individuals. The pipeline is deliberately non-optimized and non-categorical: neither supervised labeling nor any parameter tuning specific to individual recordings or species were used in the process.

4.1.2 Relation to embedding-based bioacoustics

Related embedding-based approaches have been used to organize and visualize animal vocalizations, often through learned or spectrogram-derived representations, but conventions about the shared space and interpretive framing vary between studies [9, 4, 10]. The present framework differs in that it does not learn a representation: it embeds fixed, interpretable frame-level features under a transparent protocol with fixed parameters, prioritizing geometry-first comparison across individuals.

4.1.3 Exploratory scope

The resulting manifolds should be considered as exploratory representations. They do not constitute or define any sort of classification, taxonomy or model of vocal production; rather, they offer a coherent geometric view of how acoustic features evolve, represented as continuous trajectories evolving in low-dimensional space. In this perspective, they act as a descriptive interface to inspect the high-dimensional space of acoustic features, complementing traditional spectrogram-based analysis rather than replacing it.

4.2 Complementary roles of MFCC and chroma feature spaces

Across all species considered, MFCC and chroma-based embeddings produced qualitatively different geometries, despite being generated using the same fixed-parameter pipeline. This controlled dual-representation design provided two different views of each recording.

MFCC-based manifolds produced articulated and often filamentary geometries, emphasizing variations in spectral envelope, broadband structure and temporal modulation. Chroma-based manifolds instead exhibited more compact geometries overall, ranging from shell-like shapes to more folded organizations, highlighting octave-folded amplitude distribution across frequency-class bins.

Such intrinsic differences between these representations are the outcome of two distinct analytical lenses applied to the same signal, revealing aspects of it which can be considered as complementary rather than hierarchical.

4.3 Inter-individual consistency in the shared manifolds

The shared embeddings enabled a visual inspection of the inter-individual consistency in embedded vocalizations within a single shared space, without relying on specific similarity scoring or classification objectives. It was observed how the embeddings captured a common acoustic structure with the individual trajectories occupying overlapping regions of low-dimensional space; at the same time, the trajectories retained distinct paths for every individual. This balance between correspondence and differentiation suggests that such shared manifold constructions could be useful for exploratory comparisons of individuals, populations or recordings contexts, while remaining agnostic to functional or biological interpretations[4, 11].

4.4 Descriptor overlays as scalar fields on manifolds

4.4.1 Overlays as post-hoc data layers

Overlaying spectral descriptors as color-coded scalar fields onto the embedded manifolds illustrates how additional layers of information can be integrated into the framework, without influencing the embedding process. Across feature spaces, they were applied as a color coating over the low-dimensional geometry, highlighting patterns or specific subsets of frames without inducing clustering or reorganization of the embedding. Although the spatial distribution of descriptor values across the embeddings of the examined species groups cannot be considered an indicator of function or vocal strategy in itself, it can nevertheless be considered as a further data layer that can be inspected alongside the manifold structures.

4.4.2 Chroma Energy Concentration

Notably, chroma energy concentration (CEC) was introduced in this study, defined per-frame as the ratio between the maximum chroma-bin amplitude and the chroma-sum. It quantifies how concentrated the chroma-bin amplitude distribution is within the chroma representation: higher values indicate dominance by a narrow subset of bins, while lower values reflect a pattern more evenly spread through the chroma profile. CEC is used here as a heuristic measure of concentration rather than as an unequivocal indicator of narrowband, tonal emissions; the Supplementary case study (Supplementary section S3) provides an example of its use in complementarity with spectral centroid, where their joint inspection on the Tree Pipit chroma-based manifold geometry supported identification and comparison of concentrated events across different frequency regions. In addition, chroma features exhibited consistently high cosine similarity in the denoising sensitivity checks (Supplementary table S6–S7), complementing the MFCC representation with a feature space that is relatively insensitive to preprocessing operations.

4.4.3 Decoupling embedding geometry from descriptor visualization

The separation between manifold construction and descriptor visualization is central to the framework and is implemented by design. Descriptor overlays are computed separately and applied after the embedding process; this decoupling provides an immediate and intuitive view of how the selected values vary across the geometry. While spectral centroid and CEC were used here as practical examples, it is possible to apply virtually any scalar descriptor, manually defined annotations, or any frame-aligned sets of values derived from external analyses.

4.5 Methodological considerations and limitations

The present study is intentionally scoped as a small, exploratory case study with three species groups of four individuals each. It was designed to document the behavior of a shared-embedding pipeline rather than to support statistical generalizations, and there are limitations that must be acknowledged.

4.5.1 Fixed parameters and lack of optimization

First, the specific choice of embedding parameters, intermediate dimensionality reduction and distance metrics was intended as a common controlled compromise rather than an optimized configuration. While additional parameter exploration could have yielded alternative geometries, this study prioritized reproducibility and comparability between species.

4.5.2 Compressed audio and codec artifacts

Secondly, a number of recordings (8 out of 12, the remainder in WAV format) were available only in MP3 format. Given the passerine material analyzed here, the most salient spectral content lies in the mid-frequency range, so it is less likely for high-frequency roll-off to be a dominant factor in the resulting embeddings; however, other codec-dependent artifacts (e.g., quantization noise shaping, transient smearing) may still affect fine spectral structure[16].

4.5.3 Recording conditions and residual confounds

Although a uniform process of denoising, sensitivity checks, and contour-based isolation has been applied throughout all the audio material, variability in recording conditions remains a potential source of bias that cannot be fully eliminated. Standardized preprocessing may not fully remove condition-specific signatures (habitat background, microphone/distance effects) that can influence computational analyses; prior work on acoustic individual identification highlights the importance of explicitly testing for such confounding effects when interpreting algorithmic outcomes[17]. Related observations in large-scale bird audio systems further emphasize recording quality and domain-shift effects across recording contexts[18].

4.5.4 Scope of interpretation

Finally, the low-dimensional manifolds presented here should not be conflated with topological data analysis, nor do they encode any information about causal or generative models of vocal production.

4.6 Outlook

The results of the present study suggest that shared low-dimensional manifolds may serve as a useful exploratory tool in bioacoustics research. Potential applications may include analysis of individuals, populations and comparison of recording contexts, as well as integration with annotation, segmentation or quantitative measures on the embedded geometry.

Future work could extend this framework to larger datasets, additional species and alternative feature representations. It would also be possible to investigate how geometric properties of the manifolds could relate to independent annotation systems or controlled experimental conditions. However, these extensions fall outside the scope of the present study.

5 Conclusion

This study documented a fixed-parameter pipeline for constructing shared low-dimensional manifolds from frame-level representations of passerine vocalizations, enabling direct visual comparison of multiple individuals within a common coordinate system while preserving temporal continuity. By applying the same protocol across three species groups and two complementary feature spaces (MFCC and chroma), the work establishes a reproducible baseline for manifold-based, geometry-first inspection of vocal structure.

The framework is intended as a descriptive interface that complements established spectrogram- and feature-based practices: it supports exploratory inspection, targeted backtracking from geometry to signal, and hypothesis generation without imposing categorical labels or biological interpretation. It is offered as an exploratory lens and a methodological contribution: a flexible workflow for examining vocalizations that leaves biological interpretation and analytical extensions explicitly open.

6 Acknowledgements

This research was conducted independently by the author Lucio Arese without any external funding. The author expresses gratitude to the Xeno-canto community and to all the individual recordists credited in the Supplementary Materials for making their audio resources publicly available, enabling the analyses performed in the paper. Visual models were developed in TouchDesigner, with preprocessing and analysis performed with custom Python scripts. Analyses were implemented in Python using NumPy, SciPy, pandas, scikit-learn, and UMAP (umap-learn).[19, 20, 21, 22, 7] Audio denoising procedures were executed in iZotope RX 11. ChatGPT was used for programming assistance and editorial suggestions to improve the clarity and readability of the manuscript. The author retains responsibility for all analyses, interpretations and written content of this paper.

6.1 Contributors

Conceptualization, methodology, software, data curation, formal analysis, visualization, manuscript writing, review and editing: Lucio Arese.

6.2 Data availability

Processed analysis outputs underlying the figures (shared-embedding coordinates and associated frame-aligned descriptors) are available as CSV files via Zenodo (DOI: 10.5281/zenodo.18332166). The deposit includes two CSV files per species group and feature space (MFCC and chroma), plus the case study events table.

6.3 Code availability

The paper documents a fixed-parameter pipeline designed for reproducibility. Custom Python code used for preprocessing, feature extraction, and embedding is available from the author upon reasonable request, and can be provided to editors and reviewers during peer review.

6.4 Competing interests

The author declares no competing interests.

Supplementary Materials

Supporting information and additional sections

Table S0. Overview of supplementary materials

Section	Description
S1. Supplementary Tables	Recording metadata, denoising parameters, feature computation, and PCA statistics (Tables S1–S9).
S2. Additional descriptor overlays	Additional manifold visualizations with spectral centroid and chroma energy concentration (CEC) overlays; descriptor comparison plots (Figures S1–S3).
S3. Tree Pipit case study	Observation-driven exploration of a localized manifold region, hypothesis formulation and verification through frame-level audio inspection; includes events data table and figures (Table S10; Figures S4–S6).
S4. Denoising and standardization	Example of the five-pass denoising workflow, spectral preservation, and standardized output (Figure S7).
S5. Supplementary Videos	Rotating 3D manifold animations showing low-dimensional geometry of embedded trajectories (Videos V1–V3).

S1. Supplementary Tables

Supplementary Table S1 — Common Chaffinch group recording metadata

Table S1: Common Chaffinch (*Fringilla coelebs*) recordings. All recordings were converted to mono and resampled to 48 kHz prior to analysis (see Methods).

Individual ID	XC ID	Recordist	Date	Time	Country	Location	Latitude	Longitude	Elev. (m)	Voc. type	Dur. (s)	Orig. SR (Hz)	Ch.	Type	Analyzed seg. (s)
CHF_01	XC733984	Hannu Varkki	2022-05-24	06:10	Finland	Säräisniemi, Vaala, Kainuu	64.4424	26.7783	130	song	144.9	48000	mono	mp3	0–70
CHF_02	XC800762	Daniele Baroni	2023-05-06	20:15	Finland	Savojärvi, Turku, Varsinais-Suomi	60.7435	22.3894	80	song	172.9	48000	stereo	wav	0–107
CHF_03	XC841750	Cedric Mroczko	2023-05-16	19:10	Ukraine	Svalovychi, Lyubeshivs'kyi district, Volyn Oblast	51.8731	25.6489	140	song	282.0	44100	stereo	mp3	0–86
CHF_04	XC904603	Martin Billard	2024-05-12	09:12	France	Phare de la pointe d'Agon (near Agon-Coutainville), Manche, Normandy	49.0030	-1.5770	10	song	108.0	44100	stereo	wav	0–85

Supplementary Table S2 — Tree Pipit group recording metadata

Table S2: Tree Pipit (*Anthus trivialis*) recordings. All recordings were converted to mono and resampled to 48 kHz prior to analysis (see Methods).

Individual ID	XC ID	Recordist	Date	Time	Country	Location	Latitude	Longitude	Elev. (m)	Voc. type	Dur. (s)	Orig. SR (Hz)	Ch.	Type	Analyzed seg. (s)
TP_01	XC342202	Patrik Åberg	2013-04-28	05:29	Sweden	Almömosse, Hjo Västra Götalands län	58.2662	14.1886	220	song	177.2	44100	stereo	mp3	56–138
TP_02	XC799164	Thomas Bergman	2023-05-06	06:39	Sweden	Kålsö, Mörkö, Södermanland	58.9462	17.6537	10	song	139.7	48000	stereo	mp3	0–77
TP_03	XC595957	Simon Elliott	2010-05-05	12:00	United Kingdom	Harwood Forest, Northumber-land, England	55.2311	-2.0089	280	song	128.6	48000	stereo	mp3	34–116
TP_04	XC908842	Olivier Swift	2024-05-16	09:29	France	Arrondissement de Bernay (near Glos-sur-Risle), Eure, Normandie	49.2652	0.6970	110	song	177.5	48000	stereo	mp3	0–95

Supplementary Table S3 — Eurasian Wren group recording metadata

Table S3: Eurasian Wren (*Troglodytes troglodytes*) recordings. All recordings were converted to mono and resampled to 48 kHz prior to analysis (see Methods).

Individual ID	XC ID	Recordist	Date	Time	Country	Location	Latitude	Longitude	Elev. (m)	Voc. type	Dur. (s)	Orig. SR (Hz)	Ch.	Type	Analyzed seg. (s)
WR_01	XC564559	Peter Boesman	2020-05-31	05:50	Belgium	Mechels Broek, Mechelen, Antwerpen	51.0177	4.5160	0	song	99.4	44100	stereo	mp3	0–99
WR_02	XC884539	Beatrix Saadi-Varchmin	2024-02-25	08:02	Germany	beaver marsh behind Klingelbächel, near Thaining, Landsberg am Lech, Oberbayern, Bayern	47.9699	10.9755	650	song	100.2	44100	stereo	wav	0–74
WR_03	XC897088	Lennart Jeppsson	2024-04-17	10:00	Sweden	Östratorp, Baskemölla, Simrishamn Municipality, Skåne län	55.5864	14.2804	70	song	128.3	48000	stereo	mp3	0–99
WR_04	XC994646	João Tomás	2025-04-23	10:02	Spain	Sayago (near Peñausende), Zamora, Castile and León	41.3052	-5.8789	800	song	140.0	44100	stereo	wav	0–87

Supplementary Table S4 — Denoising parameters

Table S4: Spectral denoising passes (iZotope RX, Spectral De-noise). Parameters used for the five-pass denoising chain. Each pass was rendered sequentially on the same trimmed segment. Two denoised variants were exported for sensitivity analysis: *full5* = passes 1–5 applied in order; *noP4* = passes 1–3 and 5 applied in order (pass 4 omitted).

Pass	Preset name	Adaptive	Thr	Red	Smooth	Artifact	Algorithm	FFT (ms)	MR	BehavSm.	Synth	Enh	White	Mask	Rel (ms)
1	`Birdsong_AdaptiveCleaning`	Y	3.8	8.7	1.4	3.8	Advanced	50	Y	1.5	8.3	6.1	3.1	1.0	80
2	`Birdsong_ManualPrint`	N	3.4	8.7	1.4	3.8	Advanced	50	Y	1.5	8.3	5.3	2.0	1.0	80
3	`Birdsong_Smoothing`	Y	3.8	5.5	6.0	3.8	Advanced	50	Y	1.5	8.3	6.1	3.1	1.0	80
4	`Birdsong_Extreme`	N	3.4	12.0	4.6	3.8	Adv.+Extr.	50	Y	1.5	7.0	5.3	2.0	1.0	80
5	`Birdsong_Final`	N	3.4	7.2	4.6	3.8	Adv.+Extr.	50	Y	1.5	7.0	5.3	2.0	1.0	80

Notes. Column abbreviations correspond to iZotope RX *Spectral De-noise* UI controls: Adaptive = Adaptive mode (Y/N); Thr = Threshold; Red = Reduction; Smooth = Smoothing; Artifact = Artifact control; Algorithm = Algorithm selection; FFT = FFT size (ms); MR = Multi-resolution (Y/N); BehavSm. = Algorithm Behavior: Smoothing; Synth = Noise Floor: Synthesis; Enh = Noise Floor: Enhancement; White = Noise Floor: Whitening; Mask = Masking; Rel (ms) = Release time (ms). For adaptive-mode passes, Learn time = 2.5 s. Quality = Best (D*), Reduction curve enabled, and Dynamics Knee = 1.5 were held constant across passes.

Supplementary Table S5 — SNR proxy screening for all recordings (raw audio)

Table S5: SNR proxy batch metrics computed on trimmed raw recordings. SNR proxy is estimated as $20 \log_{10}(\text{median RMS}_{\text{loud}}/\text{median RMS}_{\text{quiet}})$ using frame-wise RMS, with quiet and loud defined as the lowest and highest 20% of frames, respectively. Spectral flatness is reported on quiet frames as a proxy for background noise character.

Species	ID	File	SR (Hz)	Frames	SNR proxy (dB)	Flatness (quiet)
Chaffinch	CHF_01	XC733984	48000	8740	35.903	0.0284
Chaffinch	CHF_02	XC800762	48000	13365	36.242	0.0597
Chaffinch	CHF_03	XC841750	48000	10740	35.728	0.0375
Chaffinch	CHF_04	XC904603	48000	10615	38.668	0.0163
Tree Pipit	TP_01	XC342202	48000	10240	44.440	0.0425
Tree Pipit	TP_02	XC799164	48000	9615	48.026	0.0710
Tree Pipit	TP_03	XC595957	48000	10240	45.140	0.0629
Tree Pipit	TP_04	XC908842	48000	11865	37.933	0.0491
Wren	WR_01	XC564559	48000	12411	46.790	0.1647
Wren	WR_02	XC884539	48000	9240	41.393	0.1014
Wren	WR_03	XC897088	48000	12365	41.378	0.0249
Wren	WR_04	XC994646	48000	10865	37.592	0.0032

Supplementary Table S6 — Denoising sensitivity checks - Amplitude gate 30%

Table S6: Denoising sensitivity (amp gate 30%). For each recording, features extracted from the denoised audio are compared to the corresponding raw audio on the same gated frame set (frames in the upper 30% of the raw frame-RMS distribution; RMS percentile threshold). Results are reported for two denoising variants: **full5** (passes 1–5) and **noP4** (identical chain excluding pass 4). MFCC and chroma cosine similarities (per-frame) are reported as mean with the 5th percentile in parentheses. Δ values report the change in the mean of spectral centroid (Hz), spectral bandwidth (Hz), and CEC (unitless) relative to raw.

Species	ID	XC ID	Dur. (s)	Frames kept	$\text{Centroid}_{\text{raw}}$ (Hz)	BW_{raw} (Hz)	CEC_{raw}	**full5**					**noP4**				
								MFCC cos_{120}	Chroma cos	ΔCentroid (Hz)	ΔBW (Hz)	ΔCEC	MFCC cos_{120}	Chroma cos	ΔCentroid (Hz)	ΔBW (Hz)	ΔCEC
Common Chaffinch	CHF_01	XC733984	70	2622	4053	584	0.050	0.783 (0.720)	0.999 (0.993)	19	-42	0.002	0.788 (0.721)	0.999 (0.993)	19	-44	0.002
Common Chaffinch	CHF_02	XC800762	107	4010	3758	889	0.045	0.826 (0.725)	0.986 (0.926)	182	-68	0.004	0.830 (0.731)	0.986 (0.926)	177	-79	0.004
Common Chaffinch	CHF_03	XC841750	86	3222	4196	677	0.056	0.879 (0.812)	0.997 (0.989)	23	-33	0.004	0.881 (0.812)	0.997 (0.989)	23	-35	0.004
Common Chaffinch	CHF_04	XC904603	85	3185	4155	558	0.058	0.802 (0.706)	0.993 (0.951)	69	-101	0.005	0.803 (0.709)	0.992 (0.941)	69	-108	0.006
Tree Pipit	TP_01	XC342202	82	3072	5020	337	0.098	0.872 (0.813)	0.999 (0.999)	3	-2	0.002	0.869 (0.809)	0.999 (0.999)	3	-2	0.002
Tree Pipit	TP_02	XC799164	77	2885	4507	406	0.073	0.953 (0.926)	0.999 (0.998)	-2	-6	0.001	0.955 (0.928)	0.999 (0.998)	-2	-6	0.001
Tree Pipit	TP_03	XC595957	82	3072	4705	392	0.089	0.888 (0.846)	1.000 (0.999)	2	-12	0.003	0.886 (0.841)	1.000 (0.999)	2	-13	0.003
Tree Pipit	TP_04	XC908842	95	3560	5126	453	0.079	0.876 (0.780)	0.997 (0.981)	30	-17	0.003	0.878 (0.783)	0.997 (0.981)	30	-17	0.004
Eurasian Wren	WR_01	XC564559	99	3710	5847	312	0.162	0.910 (0.868)	1.000 (0.999)	0	-9	0.004	0.909 (0.864)	1.000 (0.999)	0	-10	0.005
Eurasian Wren	WR_02	XC884539	74	2772	5477	505	0.087	0.902 (0.826)	1.000 (0.998)	8	-33	0.003	0.901 (0.827)	1.000 (0.998)	8	-33	0.004
Eurasian Wren	WR_03	XC897088	99	3710	5945	447	0.138	0.833 (0.785)	0.999 (0.997)	12	-13	0.006	0.834 (0.781)	0.999 (0.997)	12	-14	0.006
Eurasian Wren	WR_04	XC994646	87	3260	5534	424	0.114	0.937 (0.894)	0.999 (0.994)	18	-74	0.004	0.937 (0.895)	0.999 (0.994)	18	-74	0.004

Denoising sensitivity checks - Amplitude gate 70%

Table S7: Denoising sensitivity (amp gate 30%). For each recording, features extracted from the denoised audio are compared to the corresponding raw audio on the same gated frame set (frames in the upper 70% of the raw frame-RMS distribution; RMS percentile threshold). Results are reported for two denoising variants: **full5** (passes 1–5) and **noP4** (identical chain excluding pass 4). MFCC and chroma cosine similarities (per-frame) are reported as mean with the 5th percentile in parentheses. Δ values report the change in the mean of spectral centroid (Hz), spectral bandwidth (Hz), and CEC (unitless) relative to raw.

Species	ID	XC ID	Dur. (s)	Frames kept	$\text{Centroid}_{\text{raw}}$ (Hz)	BW_{raw} (Hz)	CEC_{raw}	**full5**					**noP4**				
								MFCC cos_{120}	Chroma cos	ΔCentroid (Hz)	ΔBW (Hz)	ΔCEC	MFCC cos_{120}	Chroma cos	ΔCentroid (Hz)	ΔBW (Hz)	ΔCEC
Common Chaffinch	CHF_01	XC733984	70	6118	3541	1223	0.045	0.706 (0.574)	0.955 (0.827)	567	-580	0.014	0.714 (0.592)	0.952 (0.815)	557	-595	0.015
Common Chaffinch	CHF_02	XC800762	107	9355	2902	1636	0.039	0.779 (0.685)	0.934 (0.802)	835	-499	0.017	0.788 (0.697)	0.933 (0.804)	797	-533	0.017
Common Chaffinch	CHF_03	XC841750	86	7518	3758	987	0.051	0.822 (0.740)	0.971 (0.898)	146	-278	0.015	0.824 (0.744)	0.968 (0.886)	135	-302	0.017
Common Chaffinch	CHF_04	XC904603	85	7430	3922	1141	0.050	0.741 (0.654)	0.935 (0.812)	695	-403	0.023	0.752 (0.671)	0.934 (0.807)	714	-252	0.025
Tree Pipit	TP_01	XC342202	82	7168	4433	829	0.088	0.815 (0.697)	0.980 (0.908)	170	-290	0.014	0.816 (0.703)	0.978 (0.901)	159	-305	0.015
Tree Pipit	TP_02	XC799164	77	6730	4272	646	0.063	0.933 (0.886)	0.989 (0.950)	11	-91	0.006	0.935 (0.890)	0.988 (0.949)	7	-96	0.006
Tree Pipit	TP_03	XC595957	82	7168	4387	1788	0.070	0.843 (0.704)	0.958 (0.802)	366	-877	0.009	0.842 (0.713)	0.957 (0.796)	387	-867	0.010
Tree Pipit	TP_04	XC908842	95	8305	3640	1054	0.065	0.842 (0.740)	0.927 (0.692)	730	-329	0.012	0.846 (0.753)	0.926 (0.690)	714	-331	0.013
Eurasian Wren	WR_01	XC564559	99	8655	5975	1945	0.098	0.854 (0.736)	0.964 (0.850)	-527	-1174	0.014	0.857 (0.756)	0.959 (0.828)	-547	-1141	0.016
Eurasian Wren	WR_02	XC884539	74	6468	4733	1020	0.073	0.844 (0.704)	0.983 (0.912)	136	-338	0.009	0.843 (0.702)	0.982 (0.911)	134	-341	0.009
Eurasian Wren	WR_03	XC897088	99	8655	5057	1065	0.090	0.775 (0.652)	0.979 (0.905)	385	-243	0.011	0.774 (0.652)	0.976 (0.894)	345	-270	0.012
Eurasian Wren	WR_04	XC994646	87	7605	4071	948	0.109	0.900 (0.814)	0.912 (0.522)	840	-390	-0.018	0.902 (0.820)	0.912 (0.519)	834	-391	-0.017

Supplementary Table S8 — Technical parameters

Table S8: Technical parameters for MFCC and chroma feature computation and embedding pipeline. The same settings were applied uniformly to all recordings.

Stage	MFCC branch	Chroma branch
Audio standardization	Resampling to 48,000 Hz. Stereo recordings converted to mono by channel averaging. Working format standardized prior to feature extraction (see Methods).	
Frame / STFT settings	Window: Hann, 4096 samples (85.33 ms). Hop: 384 samples (8.00 ms; 125 frames/s). FFT size: 4096.	
Core representation	MFCC coefficients: 40. Mel bands: 128. Frequency range: 20 Hz to Nyquist.	Microtonal chroma: 80 bins on an 80 equal-division-of-the-octave (80-EDO) grid. Reference tuning: A4 = 440 Hz.
Pre-processing	Pre-emphasis filter: $\alpha = 0.97$.	Prewhitening: $\alpha = 0.0$ (disabled).
Coefficient computation	DCT: type-II (orthonormal). Lifter: $L = 22$.	Chroma bins accumulated from STFT magnitude $\|X(f)\|$ (linear amplitude; no power weighting). Per-frame L1 normalization enabled.
Temporal derivatives	Delta + delta-delta enabled, yielding 120-dimensional vectors (40×3).	Not applied.
Delta computation method	First and second temporal derivatives computed per coefficient using `numpy.gradient` over frames (central differences; forward/backward at boundaries).	Not applied.
Smoothing / stabilization	Not applied beyond PCA preconditioning.	Circular Gaussian smoothing applied to chroma bins ($\sigma = 0.3$ bins). EMA smoothing: 1 (disabled).
Intermediate reduction	PCA reduction to 20 dimensions prior to UMAP embedding (PCA-20).	
UMAP (shared-manifold) settings	Metric: Euclidean. `n_neighbors` = 30. `min_dist` = 0.1. Seed = 42. Output dimension: 3D.	Metric: Cosine. `n_neighbors` = 30. `min_dist` = 0.1. Seed = 42. Output dimension: 3D.
Coordinate normalization	Reducer script normalization: `-norm cube` with `-global_cube` enabled (shared unit-cube normalization across embeddings; overrides per-embedding scaling).	
Reducer script arguments	`-scale -pca_dim_for_umap 20 -umap_neighbors 30 -umap_min_dist 0.1 -umap_metric euclidean -seed 42 -norm cube -global_cube`	`-scale -pca_dim_for_umap 20 -umap_neighbors 30 -umap_min_dist 0.1 -umap_metric cosine -seed 42 -norm cube -global_cube`
Chroma-derived descriptor	Chroma Energy Concentration (CEC) computed per frame as $\mathrm{CEC} = \frac{\max(\mathbf{c})}{\sum \mathbf{c}}$, where $\mathbf{c}$ is the 80-bin chroma vector. Used only as a scalar-field visualization (coating), not as an input to PCA/UMAP.	

Supplementary Table S9 — PCA variance retention

Table S9: PCA variance retention. Cumulative explained variance retained (%) by PCA dimensionality reduction (PCA-15/20/30) for MFCC and chroma feature sets, computed on concatenated frame-level datasets per species group.

Species group	Feature set	PCA-15 (%)	PCA-20 (%)	PCA-30 (%)
Eurasian Wren	MFCC	50.22	58.66	72.35
Eurasian Wren	Chroma	78.42	84.49	91.18
Tree Pipit	MFCC	49.39	58.57	72.20
Tree Pipit	Chroma	78.98	84.87	91.34
Common Chaffinch	MFCC	42.17	50.26	63.17
Common Chaffinch	Chroma	86.17	89.21	93.03
Mean (3 species groups)	MFCC	47.26	55.83	69.24
Mean (3 species groups)	Chroma	81.19	86.19	91.85

S2. Additional descriptor overlays

The main Results section presented spectral-descriptors (centroid and Chroma Energy Concentration, CEC) scalar-field visualizations overlaid as color coding on the Eurasian Wren group embeddings. The following supplementary figures apply the same visualization procedure to the remaining Tree Pipit and Common Chaffinch groups.

In each case, centroid and CEC are meant to provide interpretable anchors of spectral balance and octave-folded concentration of chroma-bin amplitudes, while leaving the UMAP embedded geometry unchanged. These figures are provided for completeness, to support qualitative inspection across the three species groups under a common protocol.

S2.1 Tree Pipit group

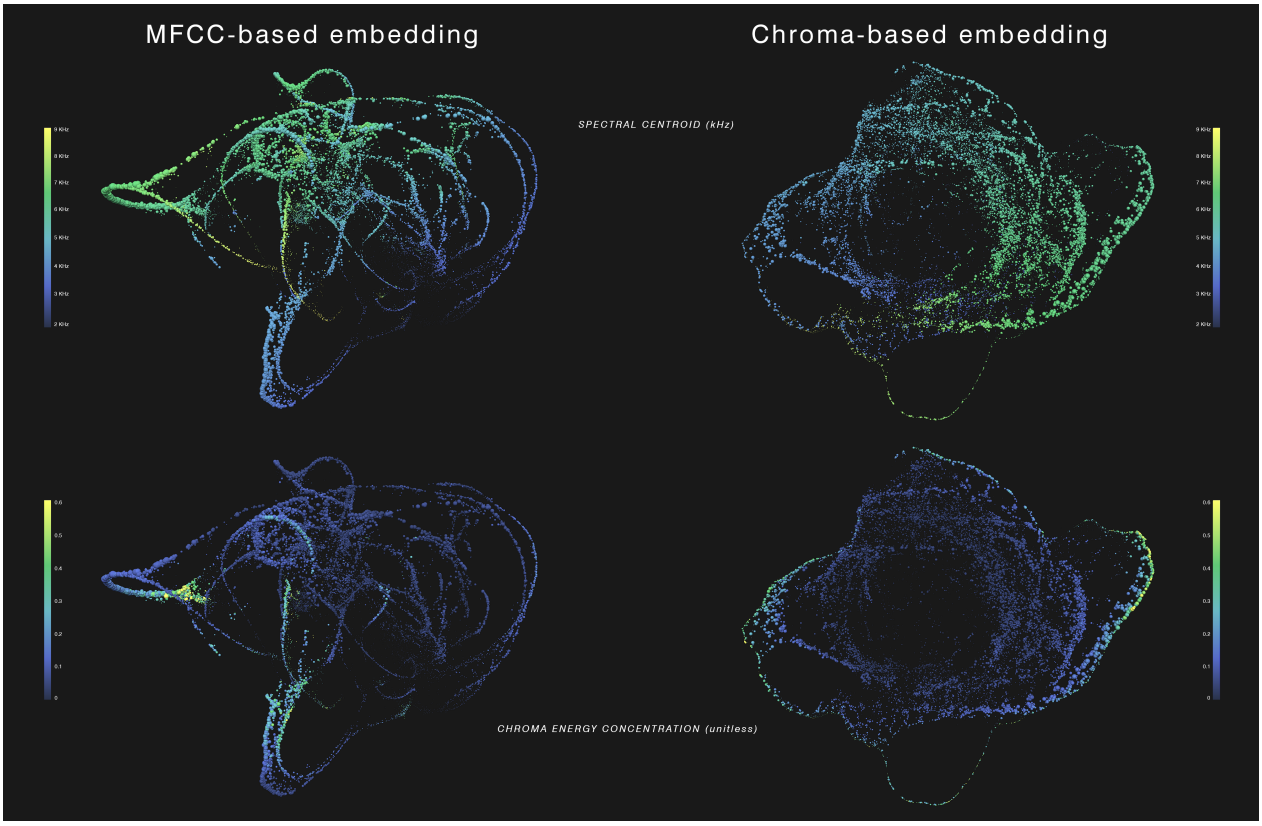

Figure S1: Tree Pipit group. Scalar-field visualizations on the Tree Pipit manifolds: MFCC-based embedding (left) and 80-EDO chroma embedding (right) colored by spectral centroid (top, 2–9 kHz) and CEC (bottom, 0–0.6). Individual labeling is intentionally collapsed so descriptor patterns appear as continuous distributions in the shared space.

Spectral centroid Applying spectral centroid as a scalar field to the MFCC-based and chroma-based manifolds of the Tree Pipit group reveals smooth, structured gradients across both embedding geometries (Fig. S1, top row). In the MFCC-based manifold (Fig. S1, top left), higher centroid values concentrate along the leftward extension and portions of the densest central region, while lower values dominate broad sections of the right-hand arcs, producing gradual transitions along the main trajectories. In the chroma-based manifold (Fig. S1, top right), centroid values show a clear large-scale organization across the envelope-like structure, with elevated values concentrated on the right-hand bulge and portions of the upper boundary in the displayed orientation, and lower values occupying extended regions on the opposite side.

Chroma energy concentration Chroma energy concentration (CEC) as a scalar field highlights a more intermittent and localized pattern (Fig. S1, bottom row). Overall, CEC remains predominantly low across both manifolds, with higher values emerging only in restricted regions of the embedding. In the MFCC-based manifold (Fig. S1, bottom left), elevated CEC values appear as localized patches and short streaks, most prominently at the far-left tip and along selected segments of the lower extension, while most of the manifold remains comparatively low. In the chroma-based manifold (Fig. S1, bottom right), higher CEC values are strongly concentrated along portions of the outer boundary (especially around the right-side lobe) whereas the majority of the embedded surface shows low CEC values. Under the current uniform comparative scaling, these distributions indicate that higher chroma-bin concentration occurs only within limited regions of the shared low-dimensional geometry.

S2.2 Common Chaffinch group

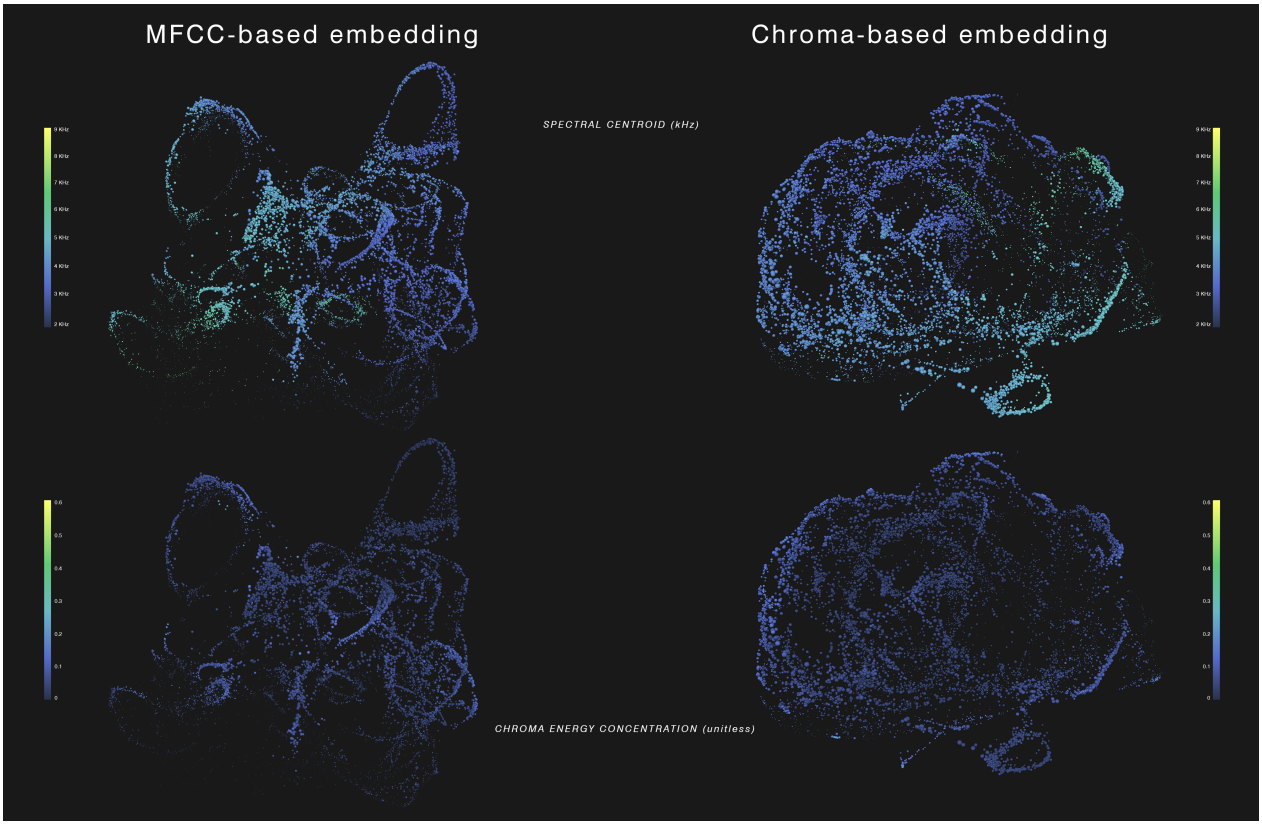

Figure S2: Common Chaffinch group. Scalar-field visualizations on the Common Chaffinch manifolds, displayed using the same conventions as Fig. S1

Spectral centroid. Applying spectral centroid as a scalar field to the MFCC-based and chroma-based manifolds of the Common Chaffinch group yields a smooth but comparatively low-contrast distribution under the current fixed global scaling (Fig. S2, top row). In the MFCC-based manifold (Fig. S2, top left), centroid values remain largely in the lower portion of the displayed range, with moderate increases localized to a few central trajectory bundles and selected arc segments, while broad peripheral loops and extended excursions remain dominated by lower values. In the chroma-based manifold (Fig. S2, top right), centroid values likewise show limited dynamic range, with slightly elevated values concentrated along a restricted set of boundary and protruding regions (notably toward the right side in the displayed orientation), whereas most of the envelope is occupied by lower centroid values.

Chroma energy concentration. Chroma energy concentration (CEC) as a scalar field reinforces the same low-contrast behavior (Fig. S2, bottom row). Across both embeddings, CEC remains predominantly low and visually flattened under the chosen comparative scale, with only small localized increases along a few MFCC trajectory segments (Fig. S2, bottom left). In the chroma-based embedding (Fig. S2, bottom right), CEC shows minimal differentiation across the manifold, with only faint boundary-localized increases and the majority of the structure remaining at low concentration values. Overall, within the current uniform scaling, both centroid and CEC provide continuous, interpretable scalar fields without sharp discontinuities, but CEC exhibits a particularly restricted effective dynamic range for this species.

S2.3 Pairwise relationships between spectral descriptors

Figure S3: Pairwise relationships between visualization descriptors across species groups. For each species group (rows: Eurasian Wren, Tree Pipit, Common Chaffinch), frame-level values are shown for *spectral centroid vs. RMS* (left), *CEC vs. RMS* (middle), and *CEC vs. spectral centroid* (right). Points are colored by individual (legend). Dotted guides mark the fixed visualization ranges used for the manifold overlays (centroid and CEC color scales), enabling direct verification that the chosen common scaling captures the predominant value ranges across groups while preserving visibility of outliers (CEC clipping: Eurasian Wren 0.27%, Tree Pipit 0.15%). The right-column panels summarize how octave-folded chroma concentration (CEC) relates to spectral balance (centroid), contextualizing localized high-CEC regions discussed in the descriptor-overlay figures.

S3. Supplementary case study - Tree Pipit group

S3.1 Overview

This supplementary case study illustrates a hypothesis generated from examination of the embedding geometry and its descriptor distributions, which was subsequently verified through frame-level analysis of raw and preprocessed audio data. The initial observation, hypothesis formulation, and verification were carried out without prior knowledge of the specific time intervals involved; the relevant segments were identified only after selecting a localized region in the embedding space and mapping its frames back to audio. This emphasizes the exploratory nature of the finding, which emerged directly from observation of the manifold structure.

S3.2 Manifold-based observation and hypothesis formulation

During qualitative inspection of the Tree Pipit chroma-based manifold, a localized region apparently traversed by multiple trajectories was observed, exhibiting highly divergent spectral centroid values and elevated chroma energy concentration (CEC) values. Under the fixed comparative scaling, centroid values in this region formed two clusters visually aligned near ∼4 kHz and ∼8 kHz, while relatively high CEC values (∼0.4–0.6) suggested concentrated chroma profiles. The co-occurrence of high CEC and two-clusters centroid pattern within a compact multi-trajectory region enabled formulation of an observation-driven hypothesis: the presence of recurring narrowband acoustic events in two separate frequency regions consistent with an approximate factor of two separation.

This reasoning assumes that, under high-CEC conditions, spectral centroid is more likely to track the dominant frequency than broadband coloration or sparse high-frequency components. This hypothesis motivated a backtracking step from geometry to signal, to test whether the localized region corresponded to recurring acoustic events.

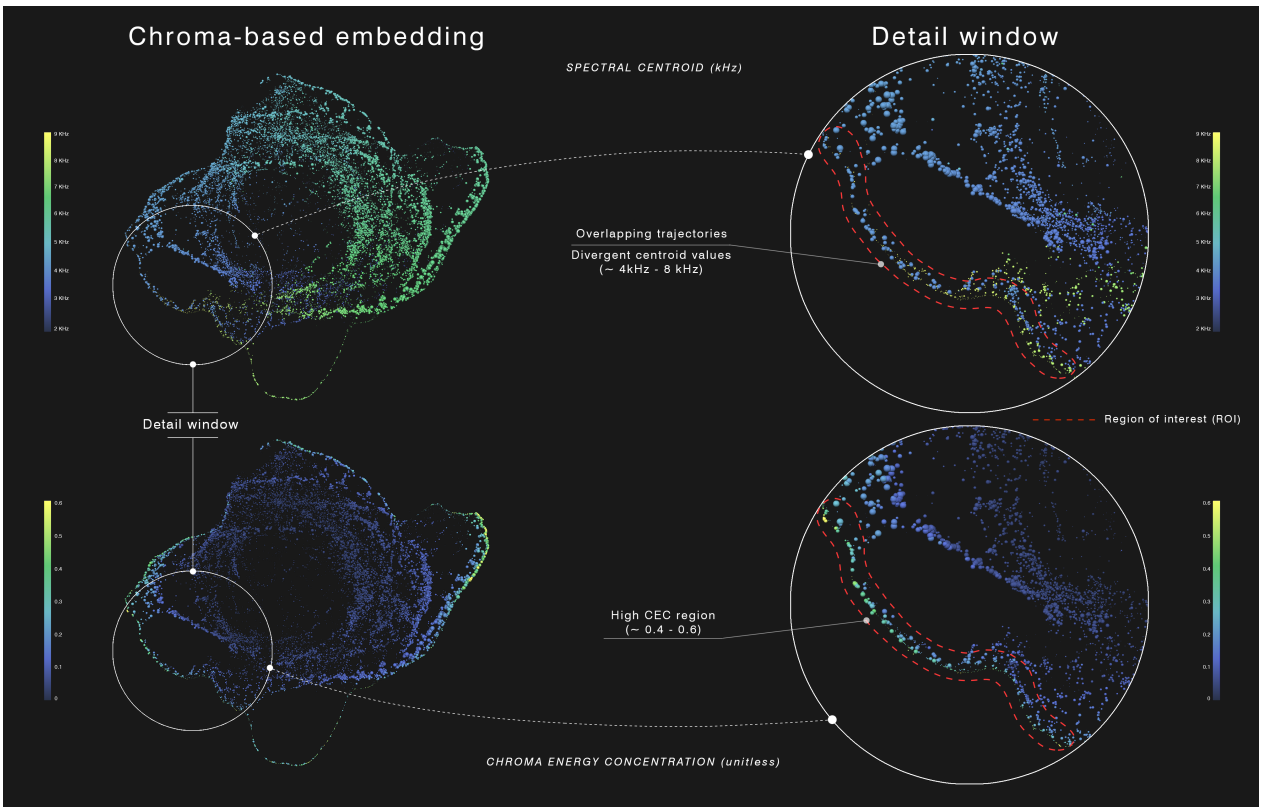

Figure S4: Localized region of the Tree Pipit chroma-based embedding used for hypothesis generation. Left: full chroma-based manifold with the region of interest indicated. Right: magnified view of the detail window, showing overlapping trajectories and two centroid regions near ∼4 kHz and ∼8 kHz, under elevated CEC values (∼0.4–0.6). The descriptor overlays are used here as visual cues to guide targeted backtracking from embedding space to audio.

S3.3 Hypothesis verification

The region of interest (ROI) was manually selected by visual inspection of the chroma-based embedding and its descriptor coatings, focusing on the compact multi-trajectory area shown in Figure S4, and then mapped back to audio. The backtracking of ROI frames to audio confirmed the presence of sequences of short, narrowband, frequency-modulated syllables occupying consistent frequency regions. They were produced by all four individuals at different points in their vocalizations. The syllables shared qualitatively similar spectral-envelope profiles and occupied two frequency regions near ∼4 kHz and ∼8 kHz, consistent with the bifurcated centroid pattern that motivated the hypothesis.

S3.4 Results

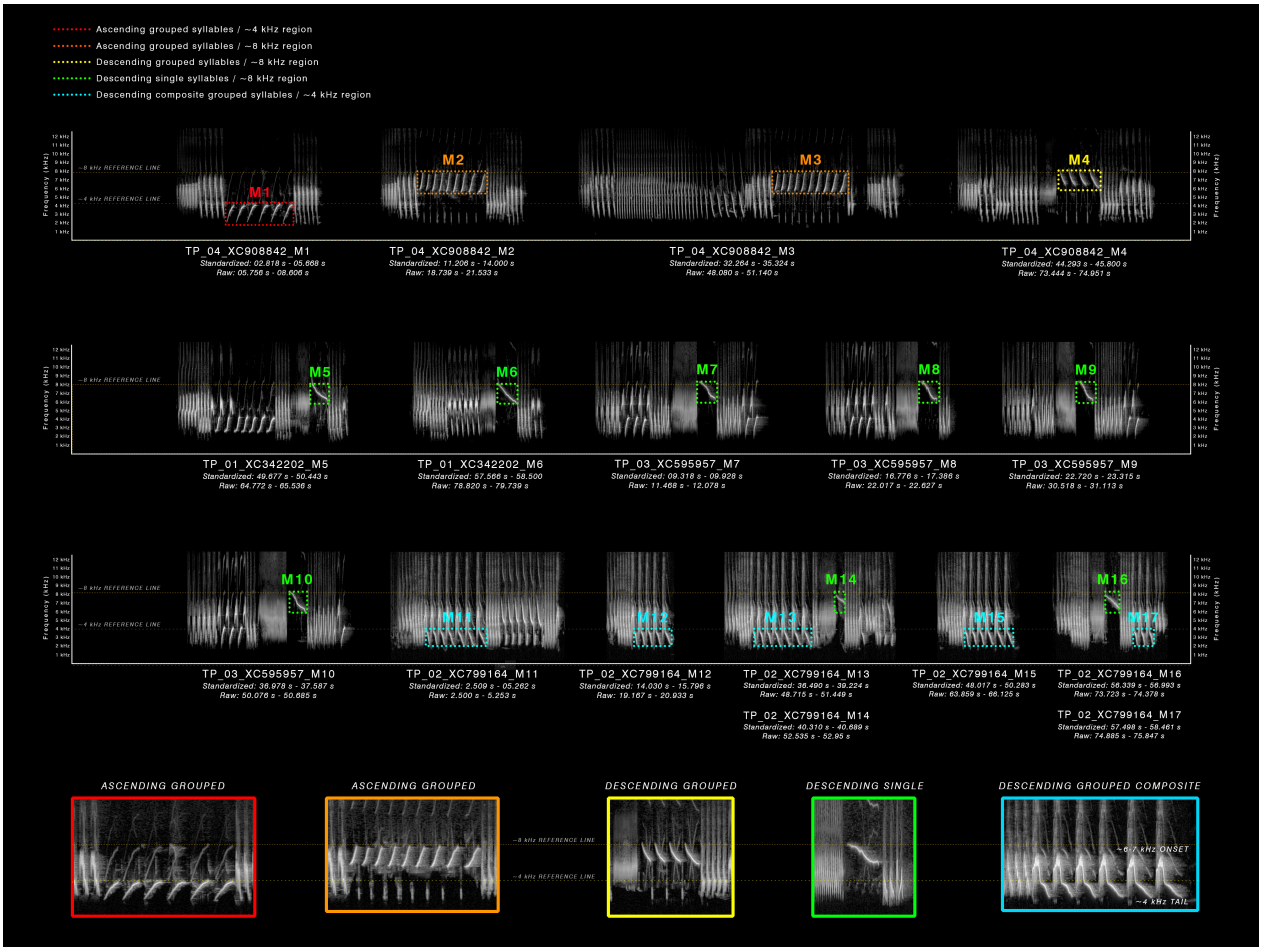

Figure S5: Representative event set recovered from the localized embedding region. Ascending (M1–M3) and descending events (M4–M17). Each panel is labeled by individual and Xeno-canto recording reference, with time intervals reported for both the standardized audio segment used for embedding and the original raw recording. The bottom row provides enlarged exemplars for each event category to highlight typical morphology. In particular, *descending grouped composite* events exhibit a higher frequency onset (∼6–7 kHz) followed by a lower-frequency tail near ∼4 kHz; the highlighted ROI-linked portion corresponds to this lower tail.

Seventeen acoustic events corresponding to the localized manifold region were identified by backtracking from ROI to audio (Fig. S5). The events occur predominantly in the ∼4 kHz and ∼8 kHz frequency regions. Here, the ∼4 kHz and ∼8 kHz labels are used as shorthand for the approximate upper extent of the dominant sweep (approximately spanning ∼2–4 kHz and ∼6–8 kHz, respectively). They are not to be interpreted as strict frequency bounds.

Across the recovered set, the events are grouped in two contour families defined by the direction of their frequency modulation: ascending (M1–M3) and descending (M4–M17). Within the descending

set, three descriptive sub-families were defined: a grouped unit in the upper region (M4); single-syllable upper-frequency hooks (M5–M10, M14, M16), and groups of composite syllables (M11, M12, M13, M15, M17), characterized by a higher-frequency onset around ∼6–7 kHz followed by a lower-frequency tail near ∼4 kHz. In these composite events, their tail corresponds to the ROI-traversing portion of the manifold anchored to the ∼4 kHz region, while centroid coatings also indicate an upper-frequency regime.

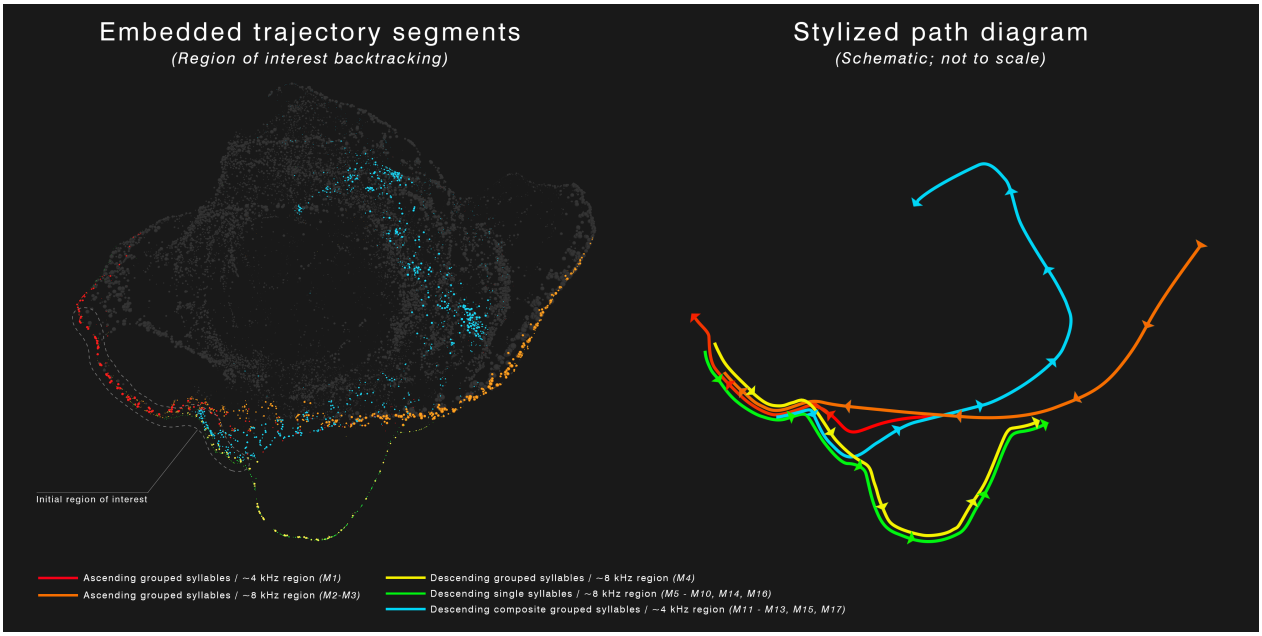

Figure S6: Event-derived trajectory segments and schematic path diagram. Trajectory segments corresponding to the events in Fig. S5 projected back into the full chroma-based manifold. Colored points indicate event-associated frames grouped by contour family and frequency region; the remaining manifold is shown in low opacity for context. The grey dashed outline indicates the initial ROI used for backtracking. Right: stylized path diagram summarizing the dominant shared ridge and branch directions (schematic; not to scale).

When projected back into the embedding space, the event-associated trajectory segments from all categories converge onto the same compact ridge-like region used for hypothesis formulation, and then depart along a small number of recurrent branches (Fig. S6, left). The stylized path diagram illustrates these dominant directions in a schematic manner (Fig. S6, right).

Under the fixed viewpoint used for visualization, ascending-contour segments (M1–M3, red and orange arrows) traverse the ridge predominantly in one projected direction, whereas descending-contour segments (M4–M17, yellow, green and cyan arrows) traverse it in the opposite projected direction, consistent with the reversal in sweep polarity, while retaining forward time progression.

Notably, upper-frequency single-syllable events from three individuals (TP_01, TP_02, TP_03) align along a narrow filament-like subpath (M5–M10, M14, M16; green), while the upper-frequency grouped syllable from TP_04 follows along the same route (M4, yellow), making this subpath common to all four individuals. In contrast, grouped composite syllables associated to TP_02 (M11–M13, M15, M17, cyan) populate a distinct excursion direction linked to the onset-plus-tail morphology. Together, these observations show that multiple spectro-temporal event types, spanning different frequency regimes and contour polarities, can map onto shared and repeatable subpaths of the chroma-based manifold under the fixed pipeline.

S3.5 Conclusion

The aim of this exploratory case study is to illustrate how low-dimensional acoustic embeddings can assist inspection of fine-grained structural patterns within multi-individual vocal datasets, without inferring cognitive or behavioral mechanisms. It provides an example of how manifold-based visual exploration can guide targeted backtracking from geometry to waveform and spectrogram, allowing

the recovery of consistent, frequency-specific events.

The Tree Pipit chroma-based embedding highlighted a compact corridor traversed by multiple individuals, associated with a bifurcated centroid pattern and elevated CEC values. Backtracking confirmed the recurrence of narrowband events occurring in two frequency regions with an approximate factor of two separation and exhibiting consistent contour families (including both ascending and descending modulation patterns). The full set of identified events (timing, duration, class, and frequency region) is summarized in Supplementary Table S10.

Furthermore, the case in question highlights the potential of chroma-based representations to identify subtle organizational features in vocalization structures, which may serve as a bridge between raw acoustic data and more structured patterns. Such visualizations may offer a complementary means of hypothesis generation regarding signal organization, independent of manual segmentation or categorical labeling.

Table S10: Details of Tree Pipit (*Anthus trivialis*) events corresponding to the region highlighted in Figure S4. Standardized times refer to the preprocessed one-minute excerpts used for embedding; raw times correspond to the original Xeno-Canto recordings. Frequency regions are approximate and refer to the dominant frequency of the highlighted event; for composite events, the highlighted portion corresponds to the lower-frequency tail near ∼4 kHz, while the onset occupies a higher frequency near ∼6–7 kHz.

Event ID	Individual (XC ID)	Std. (s)	Raw (s)	Dur. (s)	Event type	Region (kHz)	Notes
M1	TP_04 – XC908842	02.818–05.668	05.756–08.606	2.85	Ascending grouped	4	6 repeated syllables
M2	TP_04 – XC908842	11.206–14.000	18.739–21.533	2.79	Ascending grouped	8	9 repeated syllables
M3	TP_04 – XC908842	32.264–35.324	48.080–51.140	3.06	Ascending grouped	8	10 repeated syllables
M4	TP_04 – XC908842	44.293–45.800	73.444–74.951	1.51	Descending grouped	8	4 repeated syllables
M5	TP_01 – XC342202	49.677–50.443	64.772–65.536	0.77	Descending single	8	Single syllable
M6	TP_01 – XC342202	57.566–58.500	78.820–79.739	0.93	Descending single	8	Single syllable
M7	TP_03 – XC595597	09.318–09.928	11.468–12.078	0.61	Descending single	8	Single syllable
M8	TP_03 – XC595597	16.776–17.386	22.017–22.627	0.61	Descending single	8	Single syllable
M9	TP_03 – XC595597	22.720–23.315	30.518–31.113	0.60	Descending single	8	Single syllable
M10	TP_03 – XC595597	36.978–37.587	50.076–50.685	0.61	Descending single	8	Single syllable
M11	TP_02 – XC799164	02.509–05.262	02.500–05.253	2.753	Descending composite grouped	4	7 repeated syllables
M12	TP_02 – XC799164	14.030–15.796	19.167–20.933	1.766	Descending composite grouped	4	5 repeated syllables
M13	TP_02 – XC799164	36.490–39.224	48.715–51.449	2.734	Descending composite grouped	4	7 repeated syllables
M14	TP_02 – XC799164	40.310–40.689	52.535–52.915	0.38	Descending single	8	Single syllable
M15	TP_02 – XC799164	48.017–50.283	63.859–66.125	2.266	Descending composite grouped	4	6 repeated syllables
M16	TP_02 – XC799164	56.339–56.993	73.723–74.378	0.655	Descending single	8	Single syllable
M17	TP_02 – XC799164	57.498–58.461	74.885–75.847	0.963	Descending composite grouped	4	3 repeated syllables

S4. Denoising and standardization workflow example

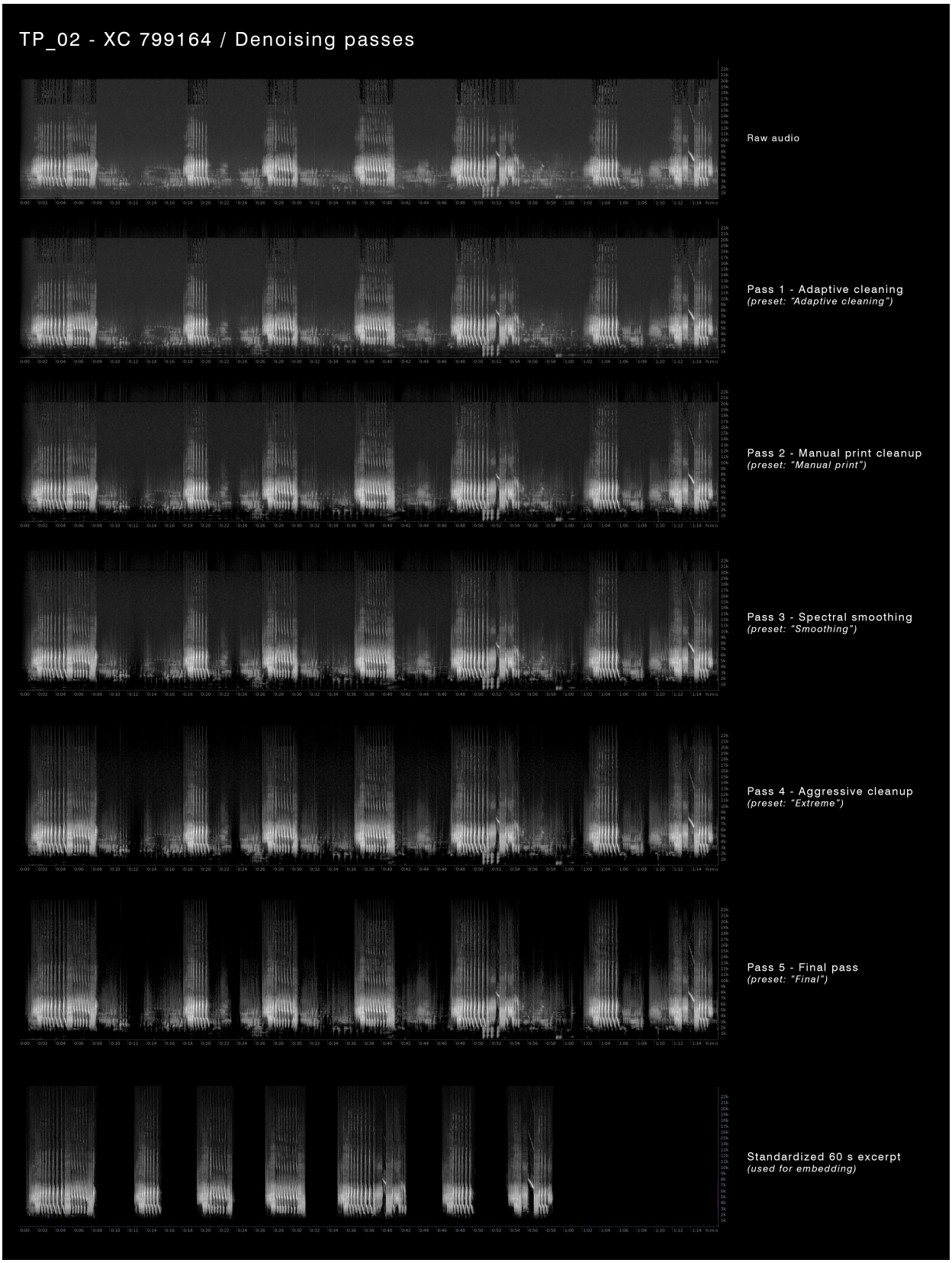

Figure S7: Example of the denoising and standardization workflow (Tree Pipit; TP_02 — XC799164). From top to bottom: spectrograms of the original recording (*raw audio*), five sequential iZotope RX11 denoising stages (*Pass 1–Pass 5*; user-defined preset labels shown for traceability), and the *standardized 60 s excerpt* used for feature extraction and embedding. Standardization is performed by manually selecting vocalization events from the denoised recording and assembling them into a fixed-duration sequence in which all non-vocal intervals are replaced by digital silence. This yields a uniform 60 s input across recordings, ensuring comparable contribution from each individual to the shared manifolds while preserving the temporal ordering and internal structure of the retained vocal segments.

S5. Supplementary videos

Three supplementary videos provide animated views of the manifolds, to facilitate qualitative inspection beyond fixed points of view.

- **Supplementary Video V1 (37 s).** *Paired shared manifolds for all species groups (solid-color individuals).* Montage of the six shared embeddings (three species groups × two feature spaces: MFCC and chroma). Individuals are shown with solid colors to visualize overlap, inter-individual consistency, and trajectory organization within each shared space.
- **Supplementary Video V2 (28 s).** *Descriptor-coated manifolds.* Selection of shared embeddings visualized with individual identity collapsed, with spectral centroid and Chroma Energy Concentration (CEC) shown as scalar-field color coatings.
- **Supplementary Video V3 (33 s).** *Supplementary case study workflow (from observation of region of interest to recovered events).* Video companion to the Tree Pipit supplementary case study. First, a close view of the region of interest is shown in the chroma embedding with centroid and CEC coatings, showing where the observation originated (Fig. S4). The video then visualizes the resulting trajectory segments corresponding to the recovered events (Fig. S6). The full manifold is shown in transparency for context.

All videos are 1920x1080 resolution, 30 fps, mp4 h264 format. Links are available via Zenodo (DOI: 10.5281/zenodo.18332166).

References

[1] Arik Kershenbaum, Daniel T. Blumstein, Marie A. Roch, Çağlar Akçay, Gregory Backus, Mark A. Bee, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 91(1):13–52, 2016. Published online 2014-11-26.

[2] Christopher K. Catchpole and Peter J. B. Slater. *Bird Song: Biological Themes and Variations*. Cambridge University Press, 2008.

[3] Peter Marler. Bird calls: their potential for behavioral neurobiology. *Annals of the New York Academy of Sciences*, 1016(1):31–44, 2004.

[4] Mara Thomas, Frants H. Jensen, Baptiste Averly, Vlad Demartsev, Marta B. Manser, Tim Sainburg, Marie A. Roch, and Ariana Strandburg-Peshkin. A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *Journal of Animal Ecology*, 91(8):1567–1581, 2022.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[6] Meinard Müller. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer Cham, 2015.

[7] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[8] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[9] Tim Sainburg, Marvin Thielk, and Timothy Q. Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology*, 16(10):e1008228, 2020.

[10] Jack Goffinet, Samuel Brudner, Richard Mooney, and John Pearson. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife*, 10:e67855, 2021.

[11] Francisco J. Bravo Sanchez, Nathan B. English, Md Rahat Hossain, and Steven T. Moore. Improved analysis of deep bioacoustic embeddings through dimensionality reduction and interactive visualisation. *Ecological Informatics*, 81:102593, 2024.

[12] Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[13] Meinard Müller and Sebastian Ewert. Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 215–220, 2011.

[14] Willem-Pier Vellinga. Xeno-canto - bird sounds from around the world. Occurrence dataset, 2024. Accessed via GBIF.org on 2026-01-10.

[15] Lawrence R. Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[16] Karlheinz Brandenburg, Christof Faller, Jürgen Herre, James D. Johnston, and W. Bastiaan Kleijn. Perceptual coding of high-quality digital audio. *Proceedings of the IEEE*, 101(9):1905–1919, September 2013.

[17] Dan Stowell, Tereza Petrusková, Martin Šálek, and Pavel Linhart. Automatic acoustic identification of individual animals: Improving generalisation across species and recording conditions. *Journal of the Royal Society Interface*, 16(153):20180940, 2019.

[18] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

[19] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, et al. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

[21] Wes McKinney. Data structures for statistical computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 56–61, 2010.

[22] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.