

Relational acoustic geometry of over 20,000 common nightingale strophes: a framework for continuous vocal analysis

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

Abstract

Large collections of complex animal vocalizations are often organized through predefined song categories, which provide powerful descriptive units but do not directly characterize continuous and representation-dependent relationships within the acoustic material. This study presents a field-grounded framework for examining such relationships without assigning vocalizations to categories in advance. During a 68-day field campaign in the Cuneo Fluvial Park, Italy, 227 georeferenced recordings comprising approximately 42.5 h of audio were collected across ten operational observation areas. Of these, 161 recordings were retained, yielding 20,075 indexed nightingale strophes. Each strophe was represented in parallel as high-dimensional MFCC-40 and microtonal Chroma-80 time-feature matrices, referred to here as acoustic terrains, and projected into separate globally fitted PCA spaces. Three-dimensional UMAP embeddings supported interactive navigation and acoustic inspection, while quantitative analyses were conducted in the retained PCA coordinates using observation-area centroids, within-area dispersion, and exact nearest-neighbor relationships under Euclidean and cosine metrics.

Strophes from the ten observation areas occupied broadly shared but non-uniform acoustic domains. Their organization was expressed through differences in centroid position and dispersion and through directional enrichment within local neighborhoods, rather than through isolated source-specific clusters. MFCC generally retained stronger same-area concentration than chroma, while Euclidean and cosine distances produced substantially different neighborhood compositions. Relational organization was therefore conditional both on the acoustic properties emphasized by the representation and on the geometrical criterion used to define proximity. Observation-area-associated structure remained measurable after candidates from the focal recording were excluded, and the broader ordering of directional enrichment relationships was substantially preserved across neighborhood sizes and under equal-contribution subsampling, although their magnitude remained sensitive to corpus composition and analytical scale. Denoising sensitivity analyses additionally showed high directional agreement between feature profiles before and after the final denoising pass.

Because observation areas represent operational field units rather than confirmed individual identities, these relationships should not be interpreted as definitive individual repertoires, conventional strophe types, or functional vocal classes. The study instead establishes a documented corpus and a reproducible framework in which relationships among complete vocal events can be examined before categorical assignment, providing a basis for subsequent categorical, sequential, temporal, and longitudinal investigation of nightingale song and for empirical evaluation in other complex animal vocal systems.

1 Introduction

Common nightingales (*Luscinia megarhynchos*) possess large and internally differentiated song repertoires whose organization cannot be described by repertoire size alone [1, 2]. Repertoire

size varies with age and can change between breeding seasons [3, 4]; it has also been related to body measures, arrival date, and paternal effort [5, 6]. Acoustic differences among song categories may additionally contribute to their communicative functions [2]. The repertoire is sequentially organized: songs are not produced as independent or random draws, but recur according to structured transition patterns and higher-order retrieval rules [7, 8]. Nightingale song therefore poses a problem not only of inventory—determining which vocal forms occur—but also of organization: how recurrent forms vary, relate to one another, and are distributed through time.

A substantial part of nightingale repertoire research begins by assigning vocalizations to discrete song or strophe types [1, 3, 9, 2]. Its primary question can be formulated as:

What kind of vocalization is this?

This classificatory approach identifies recurrent vocal forms and defines discrete units that can subsequently be related to biological, behavioral, and sequential properties. Song-type classifications have enabled estimates of repertoire size and composition, comparisons of repertoire sharing among individuals, and investigations of the acoustic and communicative properties of particular song categories [1, 9, 2]. They also provide the units required for analyses of repertoire sequencing. Network-based analysis, for example, has represented song types as nodes and successive occurrences as transitions, revealing structural properties that are not accessible from repertoire inventories alone [8].

Classification is therefore more than a descriptive exercise: it provides a powerful abstraction through which vocal variation can be related to source attribution, behavioral context, repertoire composition, and sequential organization. Constructing such categories nevertheless requires decisions about which acoustic differences define separate types, which variation remains within a type, and how intermediate or ambiguous vocalizations are assigned. Categorical organization consequently introduces an interpretive boundary between within-type variation and between-type difference.

The approach developed in this study adopts a different analytical starting point. Its guiding question can be formulated as:

How is this vocalization acoustically related to others?

The organizing criterion shifts from categorical identity to relational acoustic structure. Each complete strophe is represented as a temporally structured acoustic object and placed in relation to other strophes according to the similarities and differences expressed within a specified acoustic representation and distance geometry. Categories, strophe-type interpretations, source associations, and contextual meanings can subsequently be examined in relation to this organization, but they are not required in advance for the analytical spaces to be constructed.

This distinction does not imply that categorical classification and relational acoustic structure are opposing or mutually exclusive descriptions. They operate at different analytical levels, and their correspondence may vary. Renditions assigned to the same strophe type may occupy a compact acoustic region when they share a stable internal morphology, but may be more widely distributed when substantial within-type variation is present. Conversely, different strophe types may appear acoustically close under a particular representation, while continuous relationships may cross boundaries treated as discrete by a classificatory system. Classification and relational acoustic structure are therefore complementary, but not necessarily equivalent, descriptions whose correspondence can itself be examined empirically.

Related work in computational bioacoustics has also used continuous or low-dimensional acoustic

representations to organize animal vocalizations and examine graded acoustic structure alongside or beyond predefined categorical descriptions [10, 11, 12]. Such approaches provide a broader methodological context for examining relationships among vocal events in acoustic space, while leaving open the question of how that organization corresponds to conventional categorical labels.

The framework was developed and evaluated using an extensive field corpus of common nightingale song collected in the Cuneo Fluvial Park, northwestern Italy, during the spring 2026 breeding season. Across a 68-day campaign from 18 April to 24 June, 227 georeferenced recordings comprising approximately 42.5 h of raw audio were obtained from ten operational observation areas along the Gesso River corridor. Of these recordings, 161 were retained for analysis, yielding 20,075 indexed strophes. The observation areas were established through repeated field localization and are treated as operational source units rather than as confirmed individual identities. Each indexed strophe remained linked to its source recording, sequential position, recording date and time, spatial metadata, and associated recording-level annotations. The corpus therefore combines large-scale acoustic material with preserved temporal, spatial, and observational provenance.

In this study, the relational approach required complete strophes to be compared without reducing their internal development to a small set of global summary measurements. Nightingale strophes are internally organized into temporally ordered elements, syllables, phrases, and sections, and may include narrow-band whistles, broadband trills, buzz-like sounds, repeated elements, and other modulated components [13, 6]; their acoustic morphology therefore depends not only on which features are present, but also on how those features develop through time. Each strophe was consequently encoded as a fixed-length, temporally normalized terrain of frame-level acoustic features. MFCC [14] and a high-resolution microtonal chroma representation [15, 16] provided two distinct acoustic descriptions from which separate corpus-wide PCA spaces were constructed. Retained PCA coordinates formed the basis for quantitative comparison, while low-dimensional UMAP embeddings [17] supported navigation, inspection, and acoustic retrieval and were not treated as intrinsic maps of repertoire structure.

The conceptual architecture preceded the field campaign, while its empirical realization and final analytical form developed through the campaign itself. Before fieldwork began, the central proposition was that complete vocal events could be represented as temporally structured acoustic objects, compared within corpus-wide relational spaces, and kept explicitly connected to their source material. The number and spatial distribution of active singers, their recurrent singing locations, their persistence across the season, and the operational observation areas that could be supported through repeated localization were not known in advance. No particular separation, overlap, neighborhood structure, or global geometry was therefore specified as an expected outcome.

Accordingly, this study asks whether a large field corpus of nightingale strophes can be organized and quantitatively examined as a system of continuous acoustic relationships without requiring prior assignment to conventional strophe types. It investigates how corpus organization appears across MFCC and microtonal-chroma representations; how centroid position, within-area dispersion, and local-neighborhood composition describe complementary aspects of that organization; and how the resulting relationships depend on Euclidean or cosine distance, neighborhood scale, recording-level recurrence, and unequal source contribution. Throughout these analyses, every represented strophe retains a direct connection to its source recording, contextual metadata, sequential position, and original acoustic material, allowing relationships identified at corpus scale to be returned to individual vocal events for inspection and interpretation.

2 Methods

2.1 Study area and field corpus

Fieldwork was conducted during spring 2026 within the Cuneo Fluvial Park in northwestern Italy, along a section of the Gesso River corridor covering an approximately 6.5-ha observation space. The study area includes riverine woodland, vegetated margins and more open transitional zones with recurrent nightingale singing locations at relatively short distances from one another (Fig. 1).

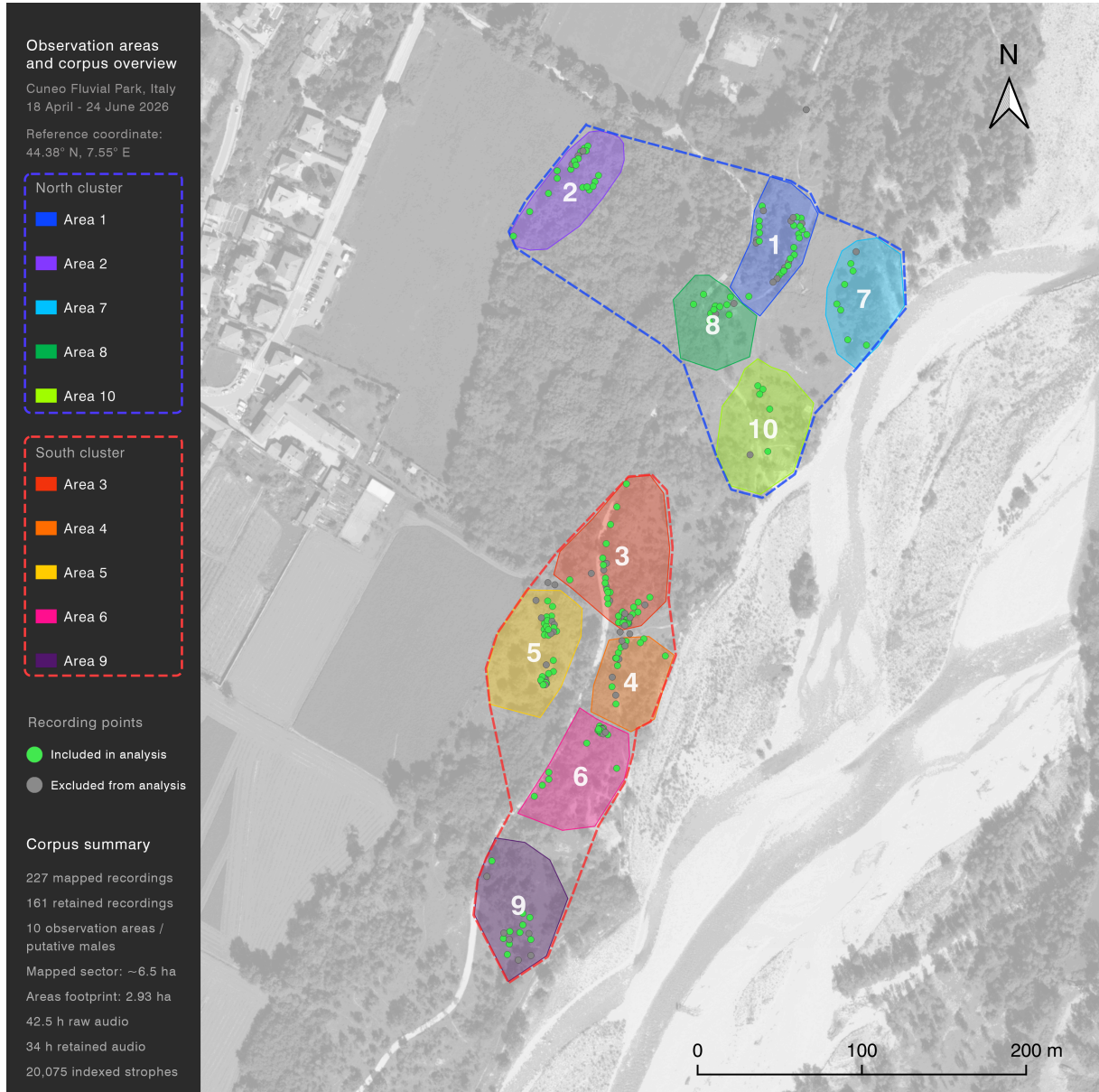


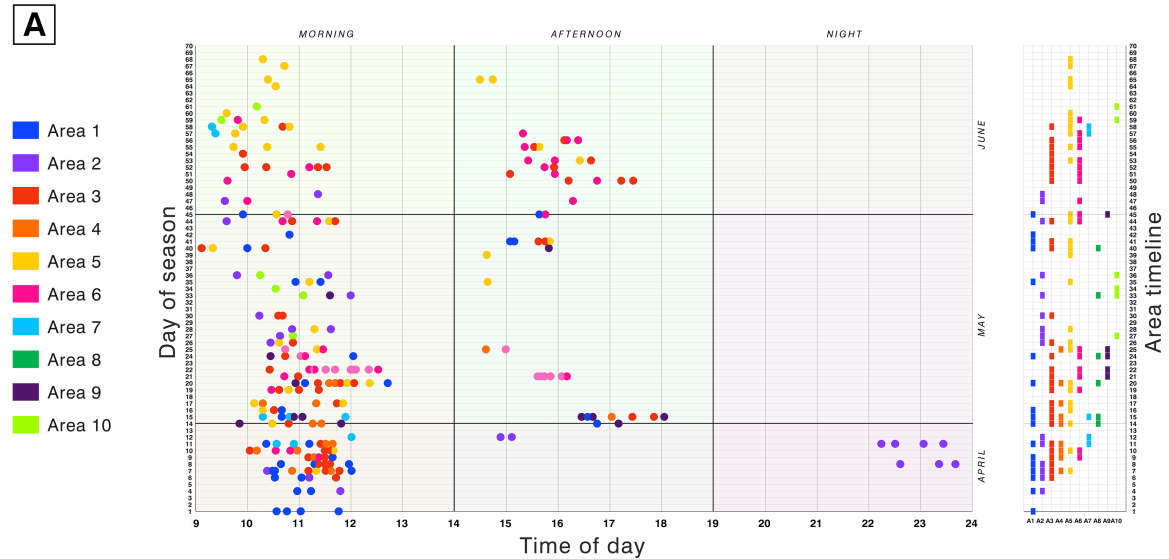
Figure 1: Observation areas and corpus overview in Cuneo Fluvial Park, Italy, from 18 April to 24 June 2026. Colored polygons indicate the ten operational observation areas associated with putative focal males, while dashed outlines indicate two broader northern and southern spatial groupings used for operational organization of the study area. Green and grey points identify recordings retained in and excluded from analysis, respectively. Observation areas represent recurrent focal-source activity units inferred from repeated spatially coherent field observations. They are not interpreted as formally established territories or confirmed individual identities. The mapped study extent covers approximately 6.5 ha, with a combined observation-area footprint of 2.93 ha. The complete corpus comprises 227 mapped recordings, of which 161 were retained for analysis, yielding 20,075 indexed strophes. Basemap: AGEA 2024 orthophoto, Regione Piemonte / CSI Piemonte WMS service regp-agea.2024; © AGEA 2024. Study overlays and map composition by the author; prepared in QGIS 4.0.1.

Recordings took place over a 68-day observation period going from 18 April to 24 June; they

covered the local seasonal arc from the early establishment of singing males up until the late phase of spring breeding activity (Fig. 2). Recording locations and observation areas were mapped in QGIS and enriched with extensive annotations, creating a geographical framework containing organized information about recording locations, observation areas and acoustic metadata.

Temporal distribution of the recording corpus

Full recording corpus



Retained recordings

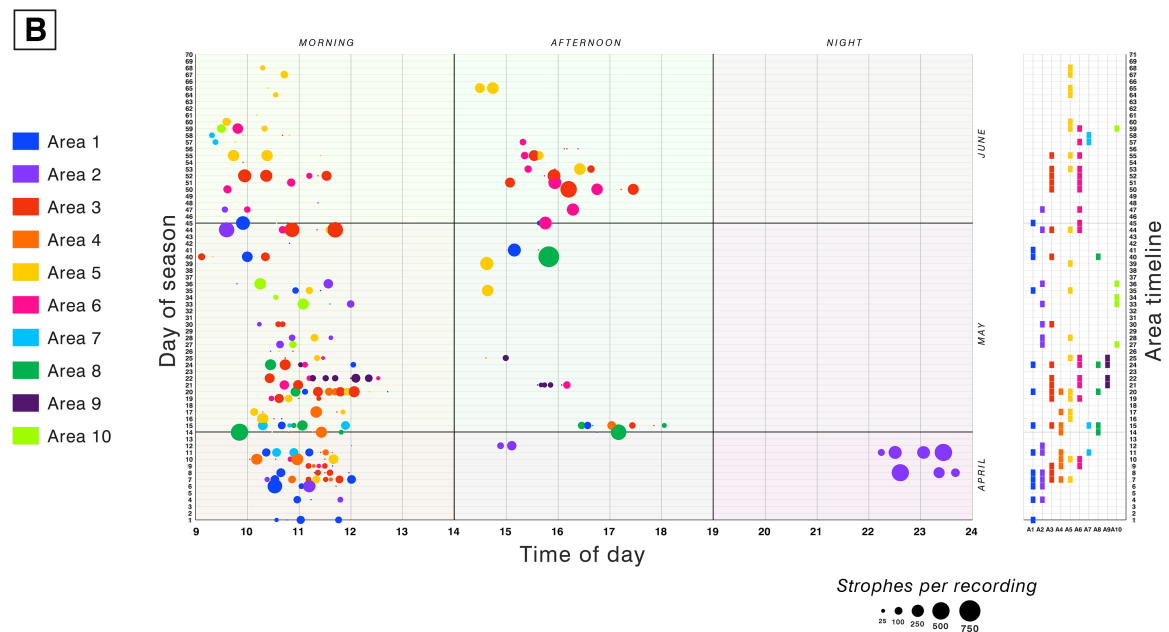


Figure 2: Temporal distribution of the recording corpus and retained analytical material. (A) Full recording corpus, with recordings positioned by time of day and day of season and colored according to observation area. (B) Recordings retained for analysis after corpus curation; point size indicates the number of indexed strophes contributed by each recording. Day of season is measured from the beginning of field observations on 18 April 2026. Shaded regions indicate morning, afternoon and night recording windows and the progression through April, May and June. The narrow panels at right summarize the seasonal distribution of recordings within each observation area.

The use of observation areas as operational source units was informed by the breeding ecology of

the common nightingale. Territorial males usually concentrate singing activity within spatially restricted singing areas and repeatedly use localized song posts, while neighbouring singing territories show little spatial overlap [18]. Longer-term studies further indicate that males may occupy the same territories across successive breeding seasons, although territory switching also occurs [19]. Repeated vocal activity within a restricted patch across multiple observation dates therefore provided a biological basis for hypothesizing that an area-specific recording series was frequently associated with a recurrent focal male.

Because the birds were unmarked, however, continuity of individual identity could not be formally demonstrated, and occasional intrusion, turnover, or replacement cannot be excluded. All analyses were consequently conducted and reported using observation-area attribution rather than assumed individual identity. Attribution confidence was assessed qualitatively from the cumulative field evidence supporting continuity of the putative focal source within each observation area. *High* confidence indicated repeated spatially coherent detections across multiple visits, recurrent use of localized singing posts or restricted activity areas, and clear separation from neighbouring sources. *Moderate* confidence indicated broadly coherent repeated localization but weaker continuity, fewer observations, or greater spatial ambiguity. *Low* confidence indicated more limited or less continuous evidence, such that source continuity across recordings remained comparatively uncertain. These levels were used only to describe the strength of source-attribution evidence and did not enter the acoustic analyses. Detailed field-construction information and photographic documentation of the observation areas are provided in Supplementary Section S1.

Observation area	Retained recordings	Indexed strophes	Corpus contribution (%)	Attribution confidence
Area 1	21	2,232	11.1	High
Area 2	20	3,218	16.0	High
Area 3	31	4,439	22.1	High
Area 4	11	1,238	6.2	High
Area 5	24	2,806	14.0	High
Area 6	22	2,124	10.6	High
Area 7	7	566	2.8	Moderate
Area 8	10	2,339	11.7	Moderate
Area 9	10	504	2.5	Low
Area 10	5	609	3.0	Moderate
Total	161	20,075	100.0	–

Table 1: Contribution of the ten observation areas to the retained analytical corpus. Retained recordings are those included after corpus curation, and corpus contribution expresses the percentage of the complete set of 20,075 indexed strophes contributed by each observation area. Attribution confidence provides a qualitative assessment of the continuity of the putative focal source associated with each observation area, based on cumulative spatial and observational evidence obtained during fieldwork.

2.1.1 Recording protocol

Recordings were performed using a RØDE Videomic Pro shotgun condenser microphone, coupled with a Panasonic HC-X1000 video camera as recorder, at 48 kHz and 32-bit floating-point depth. Before feature computation, the processed audio was exported as 48-kHz, 16-bit PCM WAV files, retaining the original 32-bit recordings as source masters. The recording procedures were completely non-invasive: individuals were not marked and no playback or other acoustic stimulation was used. Source attribution relied entirely on repeated spatial and behavioral observation. Acquisition was performed at close range whenever field conditions permitted, with a mean distance of 14.53 m from the singing source, IQR 11.0–17.6 m and range 4.9–31.2 m (see Fig. S2 for the estimated positions of singing spots across all observation areas). Recordings at longer distances were generally avoided because preservation of fine spectral structure was

considered essential for subsequent acoustic analysis. Recording gain and acquisition settings were standardized after the initial field sessions and subsequently maintained uniform throughout the campaign. The sampling extended across multiple dates and daily windows, with most recordings obtained during morning and afternoon activity, with a smaller number during targeted nocturnal sessions. The main objective of the recording campaign was to obtain repeated high-quality samples of active sources rather than a balanced coverage throughout the day, and in some observation areas sampling was constrained by accessibility, environmental contamination, source movement, and safe night-time access.

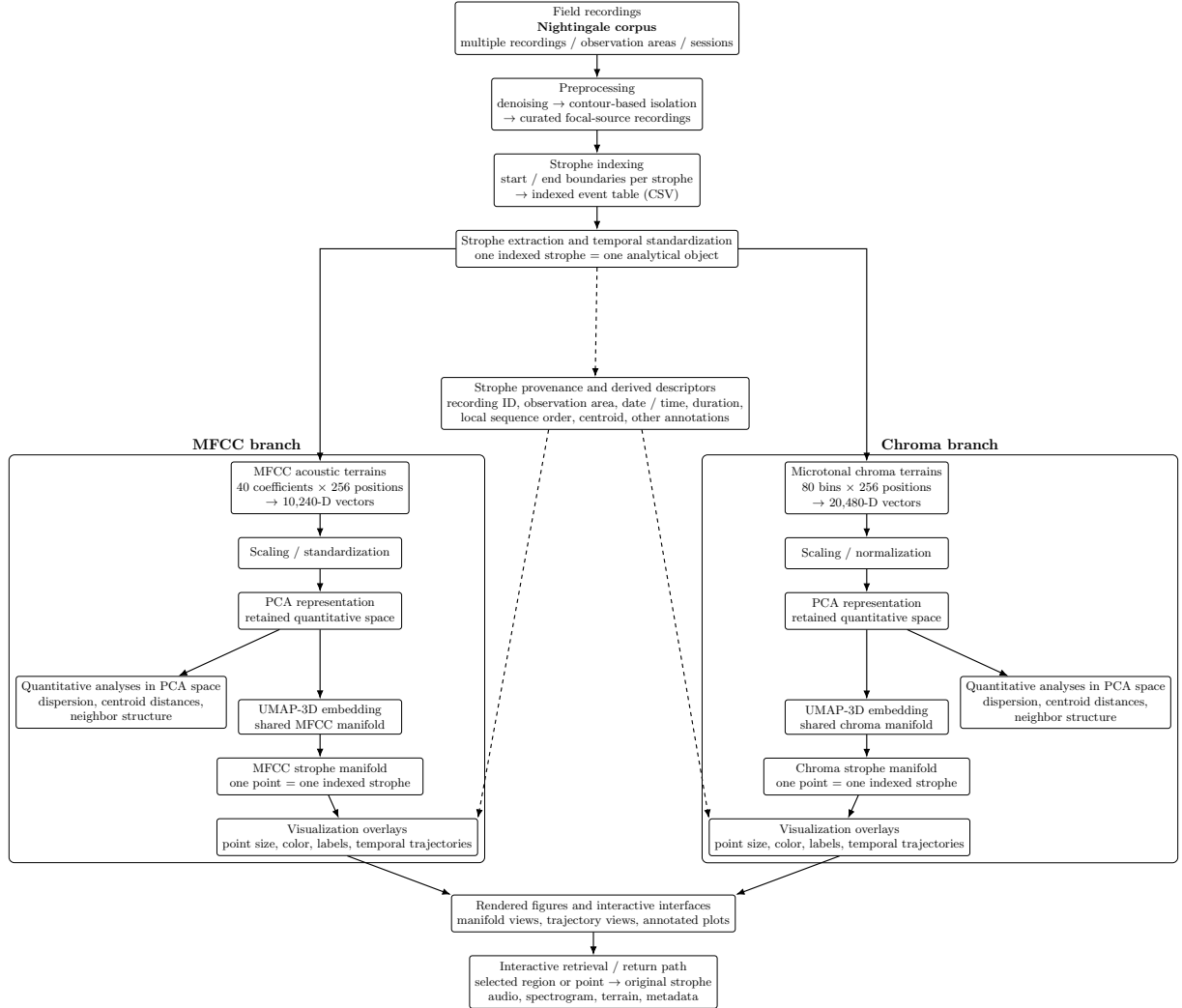


Figure 3: Overview of the strophe-level analysis pipeline. Field recordings are denoised and manually contoured to isolate focal-source vocal material, then indexed into individual strophes. Each indexed strophe is extracted and temporally standardized, yielding one analytical object per vocal event. Two parallel whole-strophe representations are then constructed: MFCC acoustic terrains and high-resolution microtonal chroma terrains. After vectorization, each representation is reduced into a retained PCA space, which serves as the principal quantitative representation for measurements such as within-area dispersion, centroid relationships, and local neighborhood organization. In parallel, the retained PCA coordinates are embedded in UMAP to generate navigable low-dimensional manifolds for MFCC and chroma. Strophe provenance and derived descriptors (e.g. recording identity, observation area, time, duration, and other annotations) are maintained separately and used only as visualization overlays or retrieval keys; they do not enter PCA or UMAP construction. Manifold regions or individual points can then be traced back to the original indexed strophes for audio, spectrogram, terrain, and metadata inspection.

2.2 Audio preprocessing and strophe indexing

2.2.1 Terminology and analytical units

Throughout this study, *strophe* denotes a complete nightingale vocal unit and constitutes the primary unit of analysis. Adult nightingale singing is characteristically organized as a succession of discrete songs, or strophes, normally alternating with silent inter-strophe intervals [20, 13]. This unit is commonly referred to as a *song* in the nightingale literature; accordingly, the term *strophe type* is used here for categorical groupings corresponding to the conventionally termed *song types*. The term *song* is reserved in the present study for nightingale singing as a broader behavioral and repertoire phenomenon, while *strophe sequence* denotes an ordered succession of strophes. *Vocalization* is used only as a broader descriptive term when the analytical unit is not being specified.

2.2.2 Preprocessing sensitivity assessment

All retained recordings were processed through a uniform two-stage denoising procedure, applied independently of observation area, recording provenance, perceived quality, or subsequent analytical outcome. Both stages used manually selected noise prints. The first pass targeted the low-frequency acquisition floor between 0 and 1 kHz, dominated by camera-preamplifier noise, while leaving the frequency range containing the nightingale signal of interest (approximately 1–15 kHz) unaffected by the denoising operation. A second, deliberately mild denoising pass was then applied across the full spectrum to reduce residual broadband environmental and recording noise. No file-specific optimization was performed to improve the appearance or separation of subsequent acoustic representations. The floor-denoised recordings were used as the baseline for sensitivity analyses, allowing the effect of the second broadband denoising pass to be assessed independently of the initial removal of low-frequency acquisition noise. Full denoising configurations, sensitivity-analysis procedures, recording-level results, and corpus-wide summaries are reported in Supplementary Section S2.

2.2.3 Strophe indexing

Manual contouring. Strophe indexing followed a hybrid half-manual and half-automatic procedure designed to combine acoustic inspection with standardized event detection. Following denoising, each retained recording was manually contoured to isolate the vocal activity attributed to the focal source; regions outside the areas of interest were replaced by digital silence, including environmental noise, interfering conspecific activity and heterospecific vocalizations. The resulting files contained retained vocal material from the focal source embedded within otherwise silent audio. All contouring profiles were preserved in the corresponding iZotope files for future audit, inspection and revision.

Particular treatment was required for adjacent strophes produced with strongly reduced or acoustically indistinct intervening pauses. Adult nightingale singing is normally organized as an alternation of strophes and silent inter-strophe intervals [20], while increases in song rate are achieved primarily through shortening these pauses [2]. Experimentally induced high-intensity territorial responses have likewise been associated with reduced pause duration and increased song rate [21]. In the present corpus, such temporal compression occasionally resulted in successive acoustically distinct strophes being delivered with very short or no clearly discernible intervening pause.

Where two such strophes could nevertheless be distinguished on the basis of their internal acoustic structure, the boundary was identified manually and a marker was inserted in the iZotope project file. Each marker was assigned a qualitative confidence level (low, medium, or good), reflecting the certainty of the segmentation decision. A fixed 1-s interval of digital silence was then inserted at the identified boundary in the indexing copy. This interval served exclusively as a technical delimiter for automated detection.

Strophes separated in this manner were subsequently treated identically to all other indexed strophes: their origin in a temporally compressed delivery was not encoded as an acoustic feature or used during construction of the relational spaces, although it remained recoverable from the original audio, editing projects, and segmentation annotations.

Automated strophe indexing. Following denoising and manual contouring, the curated recordings were processed using a standardized temporal indexing procedure. Since the contoured recordings already had non-target acoustic material removed, this stage was focused on partitioning the retained focal signal into discrete temporal event indices. Detection was performed within the 1–15 kHz range using framewise root-mean-square (RMS) amplitude, calculated with a 20-ms window and an 8-ms hop. For each recording independently, an activity floor was estimated from the 20th percentile of framewise RMS levels, and frames exceeding this floor by 10 dB were classified as active.

Successive active frames were grouped into contiguous regions without merging across intervening inactive frames. Regions shorter than 700 ms were discarded, and no additional temporal padding was applied to the resulting boundaries. Each retained event received a stable identifier and sequential position together with its temporal coordinates and associated provenance. Complete parameters for the indexing script are reported in Supplementary Table S13.

Recordings belonging to the same analytical block were indexed within a single batch procedure. The resulting event indexes were then used directly to assemble stitched analytical blocks without a second detection pass, preserving the original ordering and source identity of the indexed events. Successive events were separated by 200 ms of digital silence introduced solely as a technical separator. Explicit mappings between block-level events and their source recordings, identifiers and original temporal positions were retained throughout subsequent feature extraction and analysis. Across the complete analytical corpus, the procedure yielded 20,075 indexed strophes.

2.2.4 Strophe provenance

Each indexed strophe retained an explicit link to its source recording and field context throughout the analytical pipeline. Provenance information included the original recording identifier, onset and offset coordinates, duration, and local sequential position within the source recording, together with the corresponding recording date and time, observation area, and geospatial metadata associated with the recording. All observational annotations at recording-level were preserved and could therefore be associated with every derived strophe. These included information concerning recording conditions, vocal activity, conspecific presence and other field observations relevant to subsequent inspection. Supplementary Tables S11A, S11B, and S11C provide the complete metadata scheme.

This provenance structure was maintained through subsequent acoustic feature extraction, terrain construction, dimensionality reduction and embedding. Consequently, any strophe represented in the analytical spaces could be traced back to its original audio, position within the local singing sequence, recording session and spatial source. The relational representations were

therefore constructed from individually indexed acoustic objects without detaching them from their temporal, spatial and observational context.

2.3 Acoustic terrain representations

2.3.1 Whole-strophe representation

The unit of analysis in this study is the complete nightingale strophe, operationally defined as the continuous vocal segment bounded by adjacent inter-strophe pauses, consistent with the characteristic alternation of strophes and silent intervals in adult nightingale song [20]. Instead of reducing each strophe to a set of global summary descriptors, the framework preserves its internal temporal development by retaining frame-level acoustic feature profiles in their original sequence across the complete vocal unit.

Each indexed strophe is initially represented as a two-dimensional feature matrix, with one axis corresponding to successive temporal frames and the other to the dimensions of the chosen acoustic representation. Each matrix cell therefore contains the value of one acoustic-feature dimension at a particular point in time. For visualization, the same matrix can be rendered as a three-dimensional surface in which time and feature dimension define the horizontal plane and feature magnitude determines height. Because these surfaces resemble topographic reliefs, they are referred to here as *acoustic terrains* (Fig. 4).

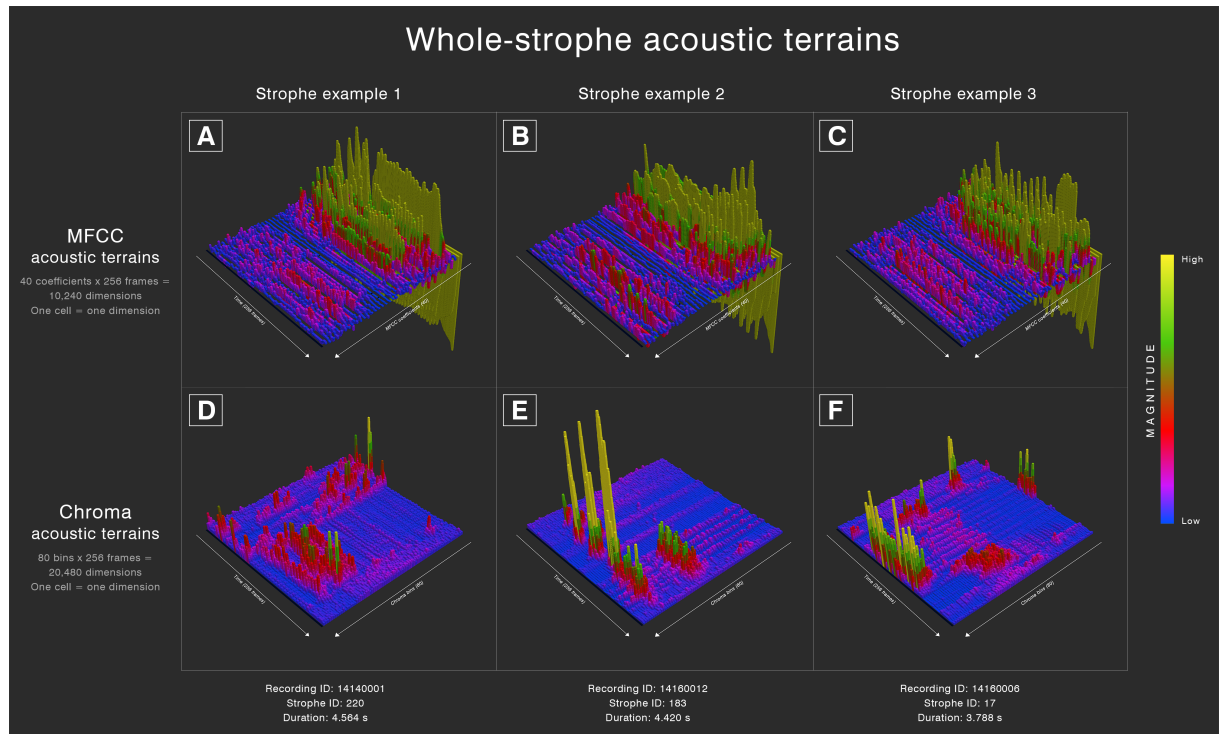


Figure 4: Representative whole-strophe acoustic terrain representations. Three indexed strophes are shown using the two distinct acoustic representations employed in the analysis. **(A–C)** MFCC terrains, comprising 40 coefficients temporally standardized to 256 positions and yielding 10,240-dimensional flattened representations. **(D–F)** High-resolution microtonal chroma terrains, comprising 80 chroma bins standardized to the same 256 temporal positions and yielding 20,480-dimensional flattened representations. Each column shows the same indexed strophe across the two acoustic representations, with source recording, strophe identifier, and original duration reported below. Surface height represents feature value across normalized strophe time and feature dimension. Colour provides an indicative representation of relative feature value. Vertical terrain scaling is strongly exaggerated for visual legibility.

Two distinct representations were constructed for every indexed strophe: one based on mel-

frequency cepstral coefficients (MFCCs) [14], describing the evolving spectral-envelope structure, and one based on high-resolution microtonal chroma [16], describing the evolving distribution of octave-folded relative-frequency spectral magnitude. The two representations were maintained as separate analytical branches and subsequently subjected to the same temporal standardization procedure before dimensionality reduction. Complete feature-extraction parameters are reported in Supplementary Table S14.

2.3.2 MFCC terrains

MFCC terrains were constructed from 40 base cepstral coefficients computed on successive frames of each strophe. Delta and delta-delta coefficients were not used in the present strophe-level analysis. A strophe containing T source frames therefore formed an initial $T \times 40$ matrix.

Each matrix was subsequently resampled along its temporal axis to a fixed length of $N = 256$ positions. The resulting 256×40 terrain was then linearized into a 10,240-dimensional vector for subsequent PCA analysis (Fig. 4A–C).

2.3.3 High-resolution microtonal chroma terrains

The second representation used an 80-bin equal-division-of-the-octave microtonal chroma description. In this study, chroma is not used primarily as a conventional musical pitch-class descriptor. Rather, octave folding provides a fine relative log-frequency representation in which spectral magnitude is resolved at substantially finer intervals than standard 12-bin chroma. Absolute octave information is intentionally discarded, while relative spectral position within the octave and its frame-to-frame evolution are retained.

The representation therefore provides an acoustic lens different from MFCCs: MFCC terrains emphasize the temporal evolution of spectral-envelope structure, while microtonal chroma terrains emphasize the temporal evolution of fine relative-frequency organization. A strophe containing T source frames initially formed a $T \times 80$ matrix. Following temporal standardization to $N = 256$ positions, each strophe was represented by a 256×80 chroma terrain and subsequently linearized into a 20,480-dimensional vector (Fig. 4D–F).

Dynamic inspection of successive MFCC and microtonal-chroma terrains during synchronized source-recording playback is provided in Supplementary Videos V1 and V2 (Table S10).

2.3.4 Temporal standardization and vectorization

Because indexed strophes varied in duration, their frame-level matrices contained different numbers of temporal observations and could not be compared directly as fixed-dimensional objects. Each acoustic-feature dimension was therefore linearly interpolated from its original T frames onto a common normalized temporal axis containing $N = 256$ positions. Temporal order and relative within-strophe development were thereby retained while absolute duration was removed from the terrain geometry.

A target length of 256 positions was selected as a practical fixed-resolution representation during framework development, following power-of-two dimensionalities used in preliminary testing. At the feature-extraction rate used here, a 384-sample hop at 48 kHz produces 125 frames per second, so 256 positions also provides a temporal sampling density close to that of a two-second

vocal segment (250 frames). The value of 256 is not intended to represent a biologically defined strophe duration: shorter and longer strophes were both interpolated across the same normalized temporal axis.

After temporal standardization, the matrices were linearized deterministically while preserving their ordered feature–time structure, producing 10,240-dimensional MFCC vectors and 20,480-dimensional chroma vectors for subsequent dimensionality reduction. Original strophe duration remained available separately through the provenance metadata and was not encoded by the length of the standardized terrain.

2.4 Principal component representation

2.4.1 PCA construction

Principal component analysis (PCA) was applied independently to the MFCC and high-resolution chroma representations, using the complete set of retained strophes to construct one shared linear coordinate system for each acoustic lens. Following temporal standardization and vectorization, each strophe was represented by a 10,240-dimensional MFCC vector and a 20,480-dimensional chroma vector. PCA was fitted globally within each representation, so that all 20,075 strophes were expressed within the same feature-specific coordinate system.

For the MFCC representation, the flattened terrain coordinates were standardized globally before PCA using feature-wise z-scores computed across the complete strophe corpus. This step compensated for differences in scale among MFCC coefficients and temporal positions. The chroma terrains, already normalized during frame-level feature computation, were passed to PCA without an equivalent additional feature-wise standardization. No observation-area, individual, temporal, spatial, or other contextual information entered either PCA construction.

A 1000-component PCA representation was retained for both acoustic branches. At this dimensionality, 92.54% of the total variance was retained for MFCC and 81.06% for high-resolution chroma across the complete 20,075-strophe corpus. These values refer specifically to the globally fitted corpus-wide PCA models and therefore describe the covariance structure of the full analytical dataset.

Because the amount of variance captured by a given number of principal components depends on the covariance structure and heterogeneity of the data to which PCA is fitted, component counts and cumulative explained-variance values obtained from smaller subsets are not directly comparable with those of the corpus-wide models. A separate reconstruction analysis was therefore used only to illustrate qualitatively how increasing PCA dimensionality progressively recovers temporal-acoustic terrain morphology under truncation (Fig. 5); the retained-variance values shown in that analysis refer to an independently fitted PCA constructed from a single 330-strophe recording and not to the complete corpus.

PCA was treated as an intermediate quantitative representation and as a preprocessing step for nonlinear embedding. Because it provides an explicit linear transformation of the original terrain vectors, PCA scores retain a direct and reproducible relationship with the acoustic feature space and can be transformed back into approximate terrain representations. Quantitative comparisons of strophe relationships were therefore performed in the retained PCA spaces, while the subsequent UMAP representation was used primarily for low-dimensional navigation, visualization, and hypothesis generation.

2.4.2 Terrain reconstruction across PCA dimensionalities

Inverse terrain reconstruction was used to examine qualitatively how progressive PCA truncation affects the temporal-acoustic morphology retained in the original feature representation. Because reconstruction and visualization of multiple high-dimensional PCA states for the complete 20,075-strophe corpus would produce very large intermediate datasets, the reconstruction series shown in Fig. 5 was generated from a representative recording containing 330 indexed strophes. A separate PCA model was fitted to this subset; its component counts and cumulative explained-variance values therefore characterize only that recording-specific covariance structure and were not used as direct numerical equivalents of the corpus-wide PCA dimensionality.

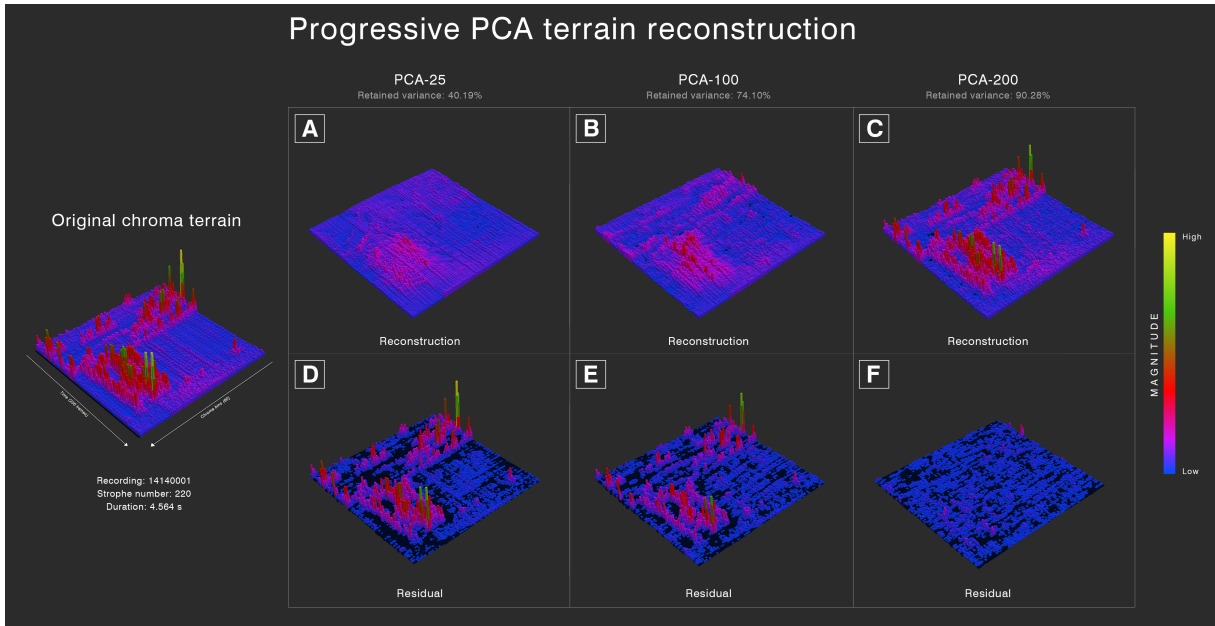


Figure 5: Progressive PCA reconstruction of a representative high-resolution chroma terrain within a 330-strophe recording subset. A separate PCA model was fitted to the chroma terrains belonging to this recording for the purpose of illustrating the effect of progressive dimensionality reduction on terrain morphology; this PCA model is independent of the corpus-wide PCA used for subsequent analysis. The original 80-bin $\times$ 256-position chroma terrain is shown at left. **(A–C)** Reconstructions obtained using the first 25, 100, and 200 principal components of the subset PCA, retaining 40.19%, 74.10%, and 90.28% of the variance within this 330-strophe subset, respectively. **(D–F)** Corresponding residual terrains, calculated as the difference between the original representation and each PCA reconstruction. Increasing PCA dimensionality progressively restores finer structure in the original terrain while reducing the remaining residual structure. These explained-variance values describe only the illustrated recording-specific chroma PCA and are not directly comparable with the variance retained by the corpus-wide PCA models fitted to all 20,075 strophes.

For a selected component count K , PCA scores beyond the first K components were omitted and the remaining representation was transformed back into the original flattened feature space using the inverse PCA transformation. Where feature-wise standardization preceded PCA, reconstructed coordinates were subsequently returned to the original flattened feature scale using the corresponding inverse standardization transform. Reconstructed vectors were then reshaped to their original $N \times 40$ MFCC or $N \times 80$ chroma terrain configuration.

Reconstructions at progressively increasing dimensionalities were compared with the corresponding original representations to visualize how larger linear subspaces recover the temporal-acoustic morphology of individual strophes (Fig. 5A–C). The comparison was used to inspect the preservation of broad terrain organization, localized structures, temporal transitions, and fine-scale feature variation across different levels of compression. Residual terrains, obtained by subtracting each reconstruction from the corresponding original terrain, provided a complementary view of the structure not recovered at each dimensionality (Fig. 5D–F).

The reconstruction analysis therefore complements cumulative explained variance without equating the dimensional requirements of the illustrated subset with those of the complete corpus. Explained variance quantifies the proportion of variation retained within the specific dataset on which a PCA model is fitted, whereas inverse reconstruction makes the morphological consequences of truncation directly inspectable in the acoustic representation from which that PCA space is derived. The subset reconstruction was thus used as an illustrative and interpretive assessment of information loss under PCA compression, while all subsequent corpus-wide quantitative analyses were performed using the independently fitted 1000-component PCA representations of the complete 20,075-strophe dataset.

Supplementary Video V3 provides a synchronized dynamic demonstration of the same reconstruction procedure, allowing progressive recovery of terrain morphology to be inspected during playback and strophe switching (Table S10).

2.5 UMAP embedding and relational spaces

2.5.1 Embedding construction

Uniform Manifold Approximation and Projection (UMAP) [17] was used to construct low-dimensional navigational representations of the two corpus-wide PCA spaces for the complete 20,075-strophe corpus. Separate embeddings were fitted for the MFCC and high-resolution chroma branches, using the corresponding PCA score matrices as input. In each case, all retained strophes were embedded jointly within a single corpus-wide space in which each point in the geometry represents a single strophe. The resulting three-dimensional coordinates therefore describe relational organization among the complete set of strophes under the two distinct acoustic representations.

UMAP was configured with $n_{\text{neighbors}} = 15$, $\text{min_dist} = 0.02$, cosine metric, and a fixed random seed of 42. The same UMAP dimensionality and hyperparameter configuration were used for both the MFCC and chroma representations. This common parameter policy was adopted to employ a uniform parametrization across the embeddings produced by the two acoustic lenses without introducing representation-specific tuning at the embedding stage. Differences between the resulting manifolds therefore are more attributable to their respective acoustic feature representations and preceding transformations than to deliberately different UMAP configurations.

Observation-area identity, recording provenance, spatial metadata and other field annotations were not supplied to UMAP and were associated with the resulting coordinates only after embedding for visualization and subsequent analysis. Post hoc overlays of recording-level and QGIS-derived field metadata, and of derived acoustic descriptors, on the completed embedding coordinates are demonstrated in Supplementary Videos V5 and V6 (Table S10).

2.5.2 Parameter policy

Parameter selection followed a restrained corpus-wide policy. Once the strophe-level representations were defined, the same analytical configuration was applied to the complete corpus. Parameters were not adjusted for individual recordings, observation areas, subsets of strophes, or particular regions of an embedding, and the same UMAP settings were used for the MFCC and chroma representations. No iterative tuning was performed to increase visual separation, produce more compact neighborhoods, or improve agreement with recording provenance or pre-existing

strophe classifications. The resulting UMAP embeddings are therefore treated as reproducible, parameter-dependent views of corpus organization and not regarded as uniquely correct maps of the acoustic material.

2.5.3 Interpretation of analytical spaces

PCA and UMAP were assigned distinct analytical roles throughout the study. The retained PCA representation was treated as the principal quantitative space for evaluating relationships among strophes. Distances, dispersion, centroid relationships, and other numerical comparisons were therefore calculated from PCA coordinates rather than from the low-dimensional UMAP embedding. PCA remains a representation-dependent space—its geometry is determined by the selected acoustic features, temporal normalization, scaling and retained component dimensionality—but its linear construction provides an explicit and reproducible relationship with the standardized terrain vectors from which it is derived.

UMAP was instead used as a low-dimensional navigational and exploratory representation of these higher-dimensional relationships. Its three-dimensional embeddings were used to visualize corpus-scale organization, inspect the distribution of recordings and strophe sequences, identify candidate neighborhoods and transitional regions, and retrieve acoustically related strophes for subsequent listening and inspection. Patterns observed in UMAP space were therefore treated as sources of hypotheses and targets for acoustic interrogation and not as direct quantitative measurements of repertoire structure.

Accordingly, distances, shapes, apparent separations, or other geometrical features of the UMAP embeddings were not treated as metrically faithful properties of repertoire organization. UMAP functions here primarily as a low-dimensional interface to the corpus, preserving a navigable connection between the represented points and their source audio, acoustic terrains, spectrograms, and associated metadata, while quantitative relational comparisons are performed in the retained PCA spaces.

Dynamic inspection of the complete three-dimensional MFCC and microtonal-chroma embeddings from multiple viewpoints is provided in Supplementary Video V4 (Table S10).

2.5.4 Sequential trajectories through acoustic space

Although strophes entered PCA and UMAP as individual acoustic objects, their original sequential position within each source recording was retained through the provenance structure described above. Temporal order could therefore be reintroduced after construction of the relational spaces by linking successive strophes according to their indexed sequence. The resulting ordered series are referred to here as *acoustic-space trajectories*.

Trajectories were visualized within an interactive TouchDesigner interface using the three-dimensional UMAP embeddings as an exploratory representation of how successive strophes from a recording or selected vocal episode were distributed through the shared acoustic space. Connecting segments indicate temporal succession between indexed strophes rather than metrically interpreted paths. Accordingly, apparent trajectory length, curvature, or spatial extent in the UMAP embedding was not treated as a quantitative property of the vocal sequence.

In the present study, these trajectories were used for visualization and interactive inspection, and not as metrics for quantitative sequence analysis. Because the indexed ordering of every strophe is preserved, corresponding ordered coordinates remain available in the retained PCA spaces

for future analyses of displacement, recurrence, revisitation, or other sequence-level properties. An interactive example of recording-level trajectory inspection is provided in Supplementary Video V7 (Table S10).

2.6 Retrieval and acoustic inspection

Within the same TouchDesigner interface, the low-dimensional embeddings were coupled to direct access to the underlying acoustic material. Individual embedded strophes and user-defined local selections could be traced to their indexed source events and retrieved for inspection. Retrieval provided access to the corresponding audio, source recording identifier, sequential position, temporal coordinates, recording metadata, MFCC and chroma terrains, and spectrographic representation.

This return path allowed the low-dimensional representation to function as a navigational interface. Embedded points could be selected visually and immediately returned to concrete vocal events for listening and comparison, while provenance information retained their temporal and field context. The retrieval interface was used in the present study primarily for exploratory inspection and methodological demonstration; quantitative relational analyses were performed independently in the retained PCA spaces.

Pointwise strophe inspection and free-form selection retrieval of local acoustic neighborhoods are demonstrated in Supplementary Videos V8 and V9 (Table S10).

2.7 Quantitative analyses

Quantitative analyses were performed directly in the retained PCA representations, with MFCC and chroma analyzed independently. PCA scores were not whitened before distance calculation, so that the relative scaling of the retained principal components remained that of the corresponding globally fitted PCA representation.

For the complete corpus, the centroid of each observation area was calculated as the arithmetic mean of all corresponding strophe coordinates. Euclidean distance from each strophe to its own-area centroid was used to characterize within-area dispersion through the median radial distance, root-mean-square (RMS) radius, interquartile range, and upper radial quantiles. Pairwise Euclidean distances between observation-area centroids were calculated separately, thereby distinguishing within-area dispersion from between-area centroid separation. Because MFCC and chroma define different PCA coordinate systems and undergo different upstream scaling procedures, absolute distance magnitudes were interpreted only within the corresponding representation and were not compared numerically across MFCC and chroma spaces.

Local neighborhood structure was evaluated by exact k -nearest-neighbor search in the retained PCA coordinates. For each focal strophe, all eligible corpus strophes were ranked according to the specified distance metric, and the first k observations were retained. Neighborhood analyses were calculated for $k = 10, 25, 50$, and 100 , with $k = 25$ used for the principal corpus-wide comparisons reported in the main text. In all reported neighborhood analyses, strophes originating from the focal strophe's own raw recording were excluded from the candidate pool before neighbor selection, while strophes from other recordings belonging to the same observation area remained eligible. This procedure limited the immediate contribution of within-recording recurrence without removing same-area relationships across separate recordings.

Neighborhood organization was evaluated independently under Euclidean and cosine distance.

Euclidean distance measures absolute separation within the retained PCA coordinates, while cosine distance measures angular similarity between PCA score vectors and reduces sensitivity to differences in overall vector magnitude. Neither metric was treated as a uniquely privileged criterion of acoustic similarity. Instead, the two metrics were used as distinct definitions of relational proximity within the same globally fitted PCA spaces, allowing both shared and metric-dependent aspects of local organization to be examined.

For every focal strophe, same-area fraction was calculated as the proportion of its k neighbors originating from the same observation area. Neighborhood diversity was quantified using the Shannon entropy of observation-area composition, normalized by $\log(10)$ [22], with larger values indicating more even representation of the ten observation areas and smaller values indicating greater concentration. Corpus-wide summaries were calculated across focal strophes, so that each strophe contributed equally. Area-level neighborhood mixing matrices were obtained by pooling neighbor assignments across focal strophes belonging to each observation area.

Because observation areas contributed unequal numbers of strophes to the empirical corpus, observed neighborhood composition was additionally evaluated relative to candidate availability. For each focal strophe, the expected proportion of candidates from each observation area was calculated from the corresponding eligible candidate pool after same-recording exclusion. These expected proportions were then aggregated by focal observation area. Enrichment was defined as the ratio between observed and expected neighborhood proportions, with values greater than 1 indicating overrepresentation and values below 1 indicating underrepresentation relative to candidate availability. Mixing and enrichment matrices were directional: rows represented the focal observation area and columns the observation area of neighboring strophes.

Sensitivity analyses examined the effects of neighborhood scale and unequal observation-area contribution. Persistence across neighborhood scales was evaluated by comparing the directional off-diagonal enrichment values obtained at $k = 10, 50$, and 100 with those from the principal $k = 25$ analysis. Comparisons used Spearman rank correlations across the 90 directional off-diagonal observation-area pairs; diagonal self-enrichment values were excluded from these calculations.

To assess the effect of unequal source contribution, 500 strophes were sampled randomly without replacement from each of the ten observation areas, using a fixed random seed of 42 and yielding an equal-contribution subset of 5,000 strophes. A sample size of 500 strophes per area was selected as a near-maximal common sample, constrained by the smallest observation-area corpus (Area 9, $n = 504$). The same sampled strophes were used in all four representation–metric combinations. They retained their coordinates in the original globally fitted PCA-1000 spaces; PCA was not refitted to the balanced subset. Same-recording exclusion was retained, and the neighborhood summaries and enrichment matrices were recalculated on the balanced candidate pool. Correspondence between full-corpus and balanced enrichment structures was evaluated using Spearman rank correlations across the 90 directional off-diagonal area pairs. Complete neighborhood-scale and equal-contribution results are reported in the Supplementary Material.

3 Results

3.1 Analytical corpus and corpus-wide acoustic organization

Following corpus curation, 161 of the 227 mapped recordings were retained for analysis, yielding a total of 20,075 indexed strophes distributed across the ten observation areas. Contributions were

unequal among observation areas and recording periods, reflecting the field-oriented sampling design described previously (Figs. 1–2).

The corpus-wide PCA representations retained 92.54% of total variance for the MFCC terrains and 81.06% for the high-resolution chroma terrains at 1,000 principal components. These PCA representations formed the higher-dimensional coordinate spaces used for subsequent quantitative analysis. Separate three-dimensional UMAP embeddings were generated to provide a navigational view of the corresponding corpus-wide relational organization. The corresponding global embedding geometries can be inspected dynamically in Supplementary Video V4 (Table S10).

In both embeddings, strophes from all ten observation areas extensively occupied shared regions of the respective acoustic spaces. Neither representation resolved into ten discrete observation-area-specific groups. Instead, observation-area identities were distributed throughout the global geometries, while locally uneven concentrations and partially shared regions were visible within both representations.

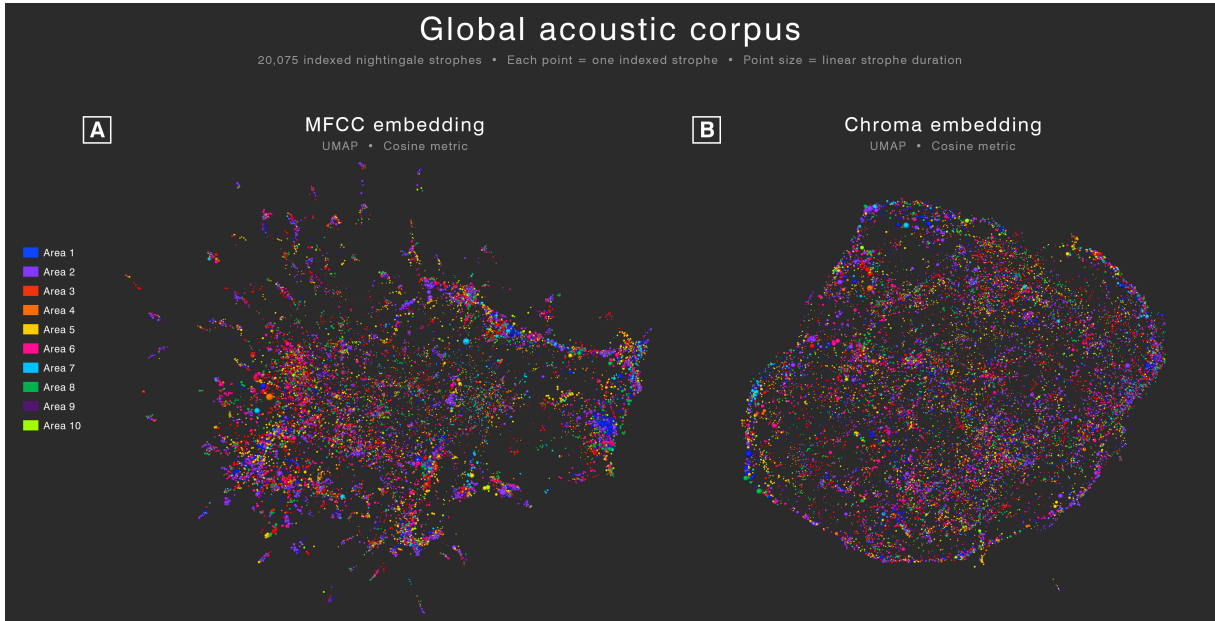


Figure 6: Global low-dimensional embeddings of the indexed nightingale corpus. (A) UMAP embedding derived from MFCC-based whole-strophe representations. (B) UMAP embedding derived from high-resolution microtonal chroma whole-strophe representations. The complete analytical corpus comprises 20,075 indexed strophes, with each point representing one indexed strophe. Point colour indicates observation area of provenance, and point size is scaled linearly with strophe duration.

The two acoustic representations nevertheless produced markedly different low-dimensional organizations. The MFCC embedding exhibited a highly extended structure containing multiple branches (Fig. 6A), peripheral extensions and locally dense regions; conversely, the chroma embedding formed a more continuous and spatially consolidated configuration with extensive interspersions of observation-area identities (Fig. 6B). These differences were treated descriptively at the UMAP level and were not interpreted as metric differences in acoustic distance; quantitative comparisons were therefore performed directly in the corresponding PCA-1000 spaces.

3.2 Within-area dispersion and centroid organization in PCA space

Observation areas differed in both their internal dispersion and their relative positions within the two PCA representations (Fig. 7). These properties described distinct aspects of corpus

organization: comparatively compact within-area distributions could nevertheless have centroids substantially displaced from other areas, while greater within-area dispersion did not necessarily imply greater between-area separation.

Within the MFCC PCA-1000 representation, RMS radius ranged from 91.7 to 99.9 PCA units (Fig. 7A). Area 7 showed the lowest RMS dispersion (91.7), whereas Area 2 showed the highest (99.9), indicating comparatively limited variation in overall RMS dispersion despite more pronounced differences in the interquartile and upper-tail extents of the area-specific radial distributions.

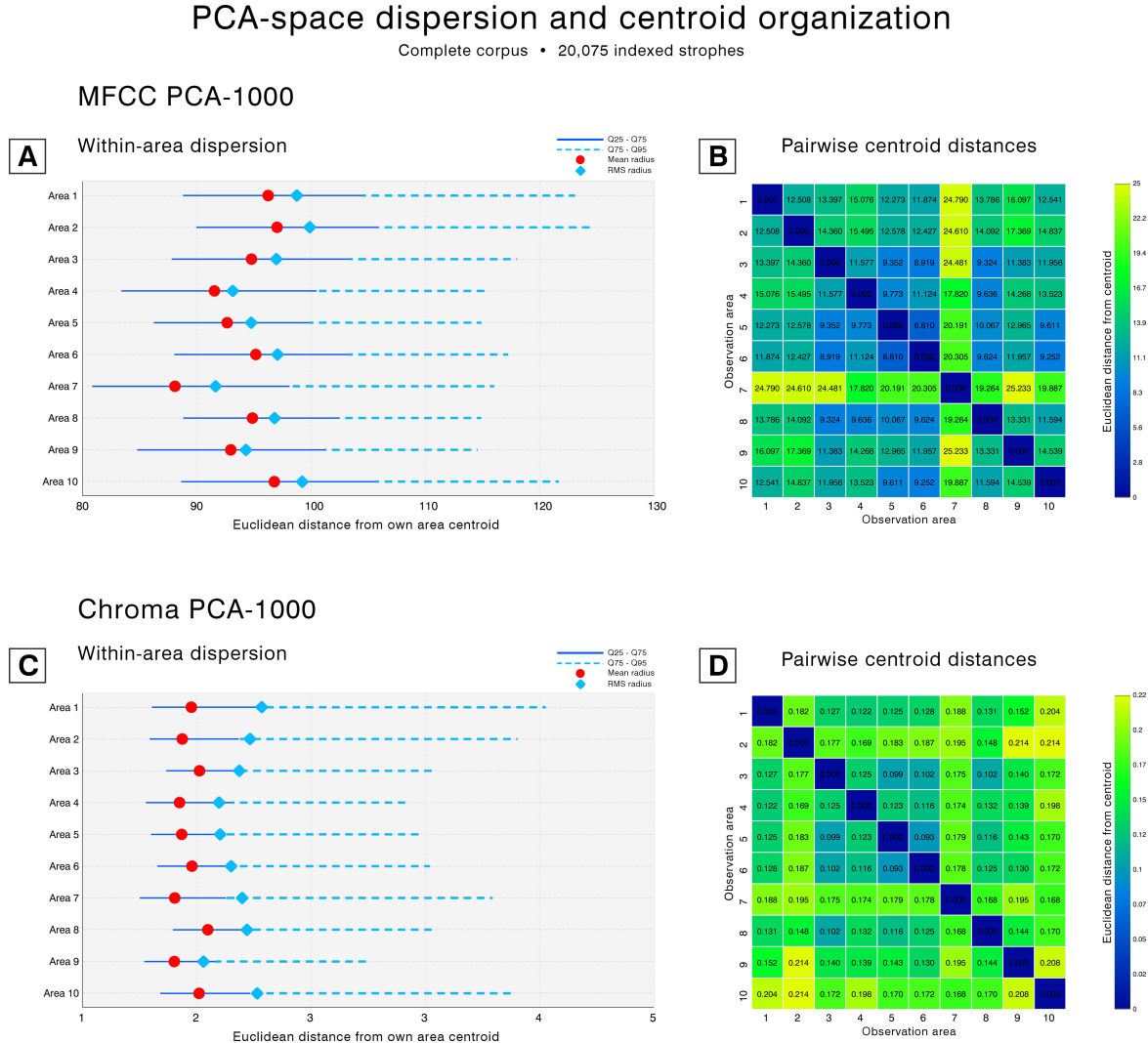


Figure 7: PCA-space dispersion and centroid organization across the complete 20,075-strophe corpus. (A) MFCC PCA-1000 within-area dispersion, expressed as Euclidean distance of individual strophes from the centroid of their corresponding observation area. Solid intervals indicate the 25th–75th percentiles, dashed intervals extend from the 75th to the 95th percentile, red circles indicate mean radial distance, and blue diamonds indicate RMS radius. **(B)** Pairwise Euclidean distances among the ten observation-area centroids in MFCC PCA-1000 space. **(C)** Chroma PCA-1000 within-area dispersion, summarized using the same statistics. **(D)** Pairwise Euclidean distances among observation-area centroids in chroma PCA-1000 space. Distance magnitudes are specific to each acoustic representation and are therefore interpreted within the corresponding MFCC or chroma PCA space. They are not compared directly across representations.

Pairwise centroid relationships were substantially more heterogeneous. Area 7 participated in several of the largest MFCC centroid separations, including Area 7–Area 9 (25.23), Area 1–Area 7 (24.79), Area 2–Area 7 (24.61), and Area 3–Area 7 (24.48). In contrast, Areas 5 and 6 formed

the closest centroid pair (6.81) (Fig. 7B). Area 7 therefore provides a clear example of the distinction between within-area dispersion and between-area position: its strophes showed the lowest overall MFCC RMS dispersion while its centroid remained strongly separated from several other observation areas.

The chroma PCA representation produced a different ordering of within-area dispersion (Fig. 7C) and centroid relationships (Fig. 7D), showing that these geometric properties depended on the acoustic representation. Absolute distance values between the MFCC and chroma spaces were not compared because their coordinate scales differ. Within the chroma representation itself, Areas 5 and 6 again formed the closest centroid pair (0.093), while the largest separations involved Area 2 with Areas 9 and 10 (approximately 0.214). Thus, the closest centroid relationship recurred across representations, while the broader ordering of centroid separations changed substantially with the acoustic representation used to describe the corpus.

3.3 Local neighborhood organization in PCA space

Local organization was examined using exact k -nearest-neighbor searches in the retained PCA spaces. The principal analyses used $k = 25$ and excluded all strophes originating from the focal strophe’s own raw recording before neighbor selection, while retaining eligible material from other recordings belonging to the same observation area. Neighborhood organization was evaluated independently under Euclidean and cosine distance, providing two explicit definitions of proximity within the same retained PCA coordinates.

The two metrics were treated as alternative relational geometries instead of a primary analysis and a secondary robustness test. Euclidean distance characterizes absolute separation within the retained PCA coordinates; cosine distance characterizes similarity in PCA-vector direction while reducing sensitivity to vector magnitude. The resulting neighborhood structures were therefore compared both within and across the MFCC and chroma acoustic representations.

3.3.1 Euclidean neighborhood organization

Under Euclidean distance, neighborhoods were extensively mixed across observation areas in both acoustic representations (Fig. 8). In the MFCC PCA space, the corpus-wide mean same-area fraction was 18.6% across the 25 nearest eligible neighbors (Fig. 8A). The corresponding mean normalized Shannon entropy was 0.782 (Fig. 8B), indicating that local neighborhoods generally contained material from several observation areas rather than being dominated by a single source unit.

Same-area occupancy was strongly heterogeneous among observation areas. Mean MFCC same-area fractions ranged from approximately 2.1% for Area 10 and 6.5% for Area 9 to approximately 23–24% for Areas 2 and 3 and 22.2% for Area 7 (Fig. 8A). The observation-area structure of the MFCC space therefore did not consist of uniformly compact or uniformly mixed local groups.

Local neighborhoods were more extensively mixed in the chroma PCA representation. The corpus-wide mean same-area fraction decreased to approximately 14.4%, while normalized neighborhood entropy increased to approximately 0.822 (Fig. 8D–E). Chroma therefore showed weaker aggregate observation-area locality than MFCC under Euclidean distance, although substantial between-area heterogeneity remained.

3.3.2 Euclidean enrichment relative to candidate availability

Because observation areas contributed unequal numbers of strophes to the corpus, raw neighborhood composition alone could not distinguish preferential local association from differences in candidate availability. Observed neighbor frequencies were therefore compared with the corresponding observation-area proportions in the eligible candidate pool. The resulting enrichment matrices revealed structured departures from abundance-based expectation in both acoustic representations (Fig. 8C and F).

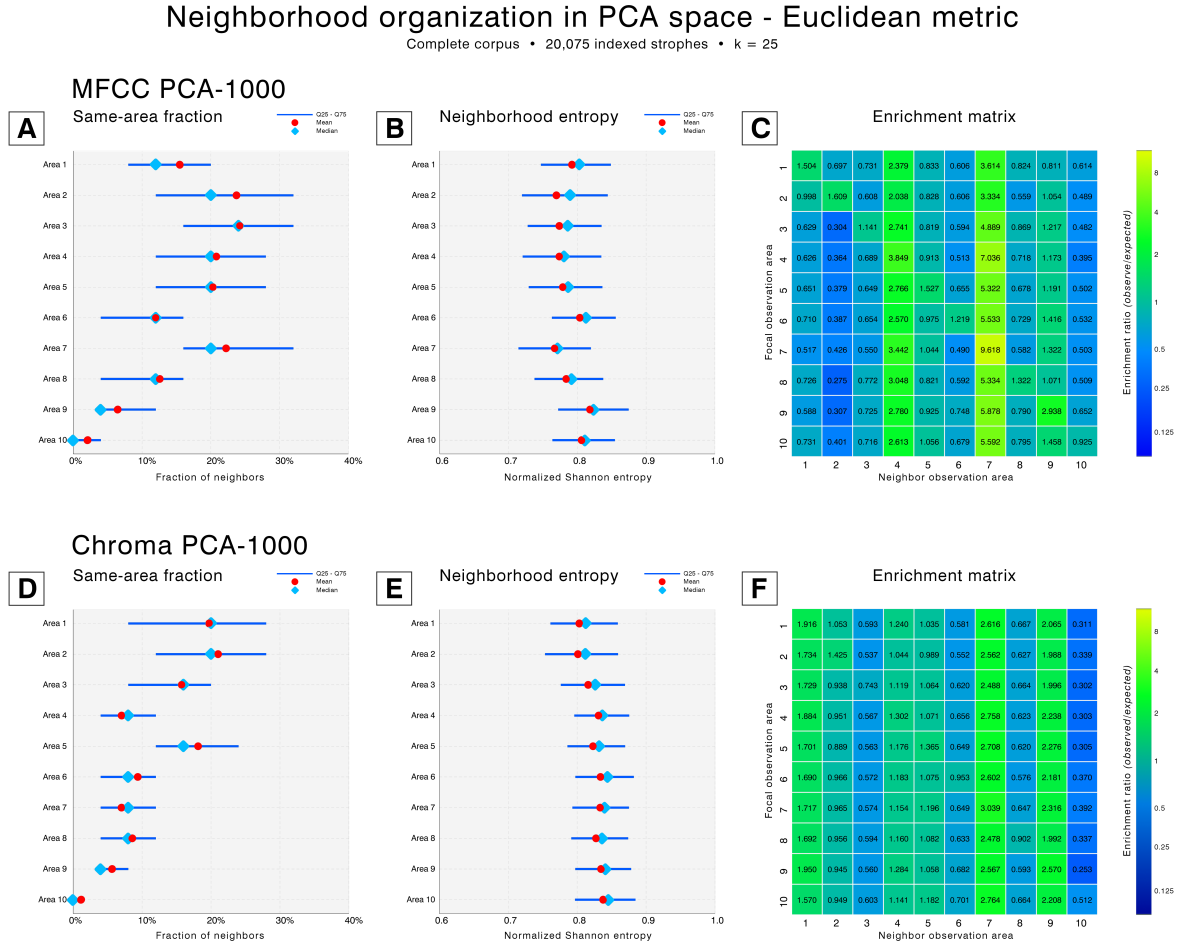


Figure 8: Euclidean neighborhood organization in the retained PCA spaces for the complete 20,075-strophe corpus ($k = 25$; same-recording candidates excluded). (A) MFCC PCA-1000 same-area fraction, summarizing for each observation area the fraction of neighboring strophes belonging to the focal area. (B) MFCC PCA-1000 neighborhood entropy, expressed as normalized Shannon entropy of observation-area composition. (C) MFCC PCA-1000 enrichment matrix, showing the observed-to-expected ratio of neighboring strophes from each observation area relative to their availability in the eligible candidate pool. (D) Chroma PCA-1000 same-area fraction. (E) Chroma PCA-1000 neighborhood entropy. (F) Chroma PCA-1000 enrichment matrix. For the same-area fraction and neighborhood entropy summaries, horizontal intervals indicate the 25th–75th percentiles, red circles indicate means, and blue diamonds indicate medians. Enrichment values greater than 1 indicate overrepresentation relative to candidate availability; values below 1 indicate underrepresentation.

The largest and most heterogeneous Euclidean enrichment ratios occurred in the MFCC representation. Area 7 was substantially overrepresented in the neighborhoods of every focal observation area. For example, Area 7 accounted for approximately 20% of the selected neighbors of Area 4 strophes despite constituting only approximately 2.8% of the corresponding eligible candidate pool, producing an enrichment of approximately $7.0\times$. Area 7 also showed strong self-enrichment ($9.62\times$) after strophes from the same raw recording had been removed from consideration (Fig. 8C).

Area 4 was likewise recurrently overrepresented within MFCC neighborhoods, although less strongly than Area 7. Conversely, several area combinations occurred substantially less frequently than expected from candidate availability. The MFCC Euclidean neighborhood geometry therefore contained directed local association patterns that were not apparent from corpus abundance alone.

The chroma enrichment ratios occupied a narrower range. Departures from candidate availability remained present, but the strongest values were generally lower than in the MFCC representation. Area 7 remained recurrently overrepresented and showed a self-enrichment of approximately $3.04\times$, while Area 9 was also consistently overrepresented in the neighborhoods of multiple focal areas (Fig. 8F). The detailed Euclidean relational organization was therefore representation dependent: MFCC produced stronger and more heterogeneous departures from candidate availability, whereas chroma produced a more broadly mixed local geometry.

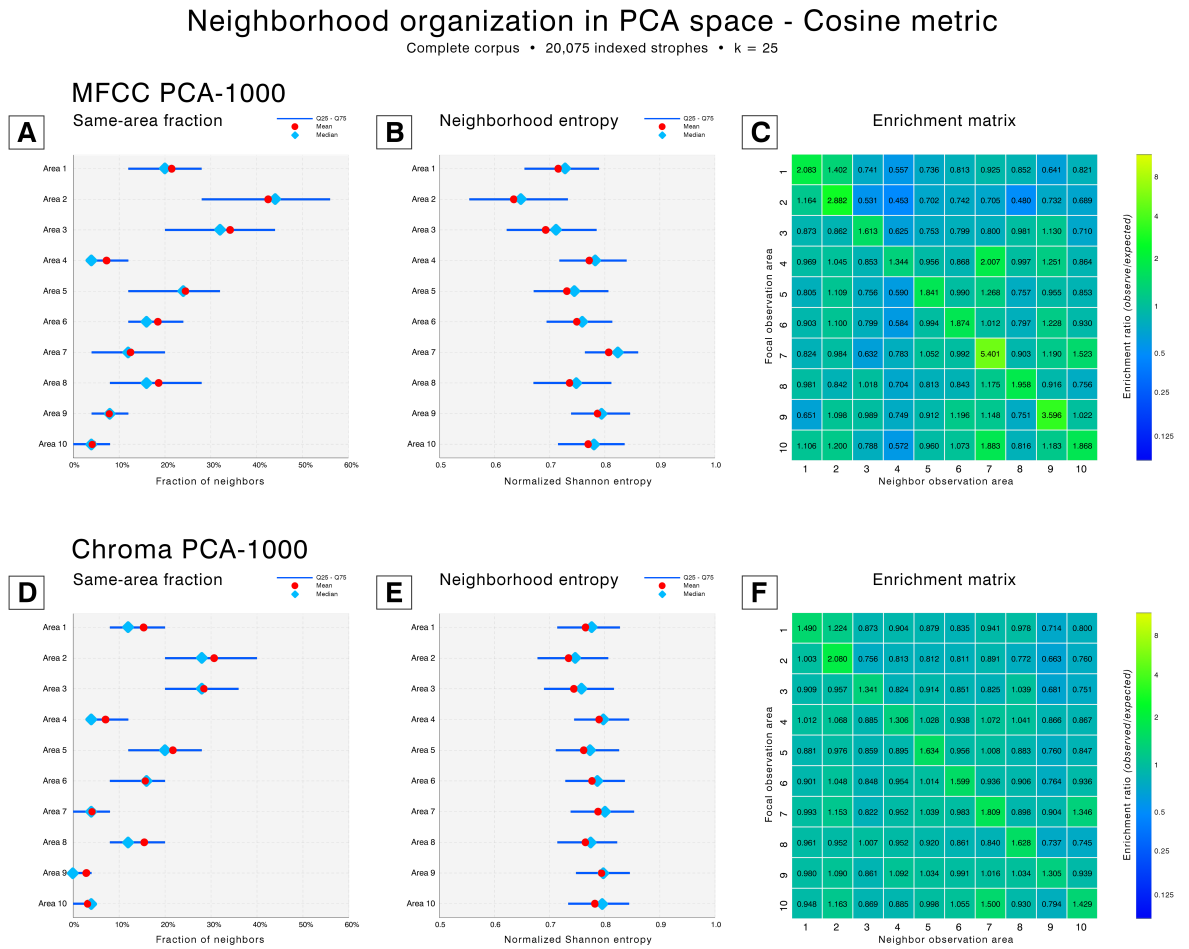


Figure 9: Cosine neighborhood organization in the retained PCA spaces for the complete 20,075-strophe corpus ($k = 25$; same-recording candidates excluded). (A) MFCC PCA-1000 same-area fraction. (B) MFCC PCA-1000 normalized neighborhood entropy. (C) MFCC PCA-1000 enrichment matrix, expressed as the observed-to-expected ratio relative to eligible candidate availability. (D) Chroma PCA-1000 same-area fraction. (E) Chroma PCA-1000 normalized neighborhood entropy. (F) Chroma PCA-1000 enrichment matrix. Graphical summaries and enrichment scaling follow the conventions used in Fig. 8.

3.3.3 Cosine neighborhood organization

Cosine distance produced a distinct local organization in both PCA representations (Fig. 9). Aggregate observation-area locality increased relative to the Euclidean analysis: the corpus-wide

mean same-area fraction reached approximately 25.4% for MFCC (Fig. 9A) and 20.1% for chroma (Fig. 9D), while mean normalized neighborhood entropy decreased to approximately 0.715 (Fig. 9B) and 0.760 (Fig. 9E), respectively.

The increase in MFCC locality was strongly heterogeneous among observation areas. Mean same-area fractions reached approximately 42.5% for Area 2 and 34.2% for Area 3, while Area 5 also showed substantial locality (approximately 24.4%). By contrast, Areas 10, 4, and 9 remained much less locally concentrated, with mean same-area fractions of approximately 4.2%, 7.3%, and 7.9%, respectively (Fig. 9A). Cosine MFCC neighborhoods therefore did not simply produce a uniform increase in same-area association, but redistributed locality unevenly among observation areas.

The MFCC cosine enrichment matrix likewise showed structured but substantially different departures from candidate availability. All observation areas showed self-enrichment above unity, with particularly strong diagonal enrichment for Area 7 (5.40 $\times$), Area 9 (3.60 $\times$), and Area 2 (2.88 $\times$) (Fig. 9C). Off-diagonal relationships were generally more moderate than the strongest Euclidean associations, and the prominent Euclidean attraction toward Area 7 was substantially reduced under cosine distance.

Cosine neighborhoods in chroma space remained more extensively mixed than their MFCC counterparts. Same-area fractions were highest for Area 2 (approximately 30.7%) and Area 3 (28.4%), while Areas 9, 10, and 7 showed particularly low values of approximately 2.9%, 3.2%, and 4.2%, respectively (Fig. 9D). Chroma self-enrichment nevertheless remained above unity for all observation areas, while departures from expectation in the overall enrichment matrix remained comparatively modest (Fig. 9F). As under Euclidean distance, MFCC therefore retained stronger aggregate observation-area locality than chroma, but the detailed distribution of that locality depended strongly on the metric.

3.3.4 Comparison between Euclidean and cosine neighborhood geometries

The two metrics produced similar broad trends but markedly different local relational topologies. Under both Euclidean and cosine distance, MFCC showed a greater mean same-area fraction and lower neighborhood entropy than chroma. Changing from Euclidean to cosine distance increased mean same-area fractions from 18.6% to 25.4% for MFCC and from 14.4% to 20.1% for chroma, while normalized neighborhood entropy decreased from 0.782 to 0.715 and from 0.822 to 0.760, respectively.

At the level of individual cross-area relationships, however, correspondence between the two metrics was weak. Spearman correlation between off-diagonal enrichment values was approximately $\rho = 0.04$ for MFCC and $\rho = 0.14$ for chroma. Consequently, the increase in aggregate observation-area locality under cosine distance did not imply preservation of the specific cross-area enrichment structure identified under Euclidean distance.

The same retained PCA coordinates therefore supported substantially different local organizations depending on how proximity was defined. The distinction between Euclidean and cosine neighborhoods is thus not limited to changes in the overall strength of source association, but extends to the identity and relative prominence of the relationships composing the local neighborhood structure.

3.3.5 Neighborhood sensitivity

The principal neighborhood results were evaluated for sensitivity to unequal observation-area contribution and to the choice of neighborhood size. Equal-contribution subsampling altered absolute same-area fractions and neighborhood entropy, but directional off-diagonal enrichment structures remained substantially concordant with those of the complete corpus. Enrichment structure was even more stable across neighborhood sizes ($k = 10, 25, 50$, and 100). Full sensitivity results are reported in Supplementary Section S3 and Supplementary Tables S4– S7.

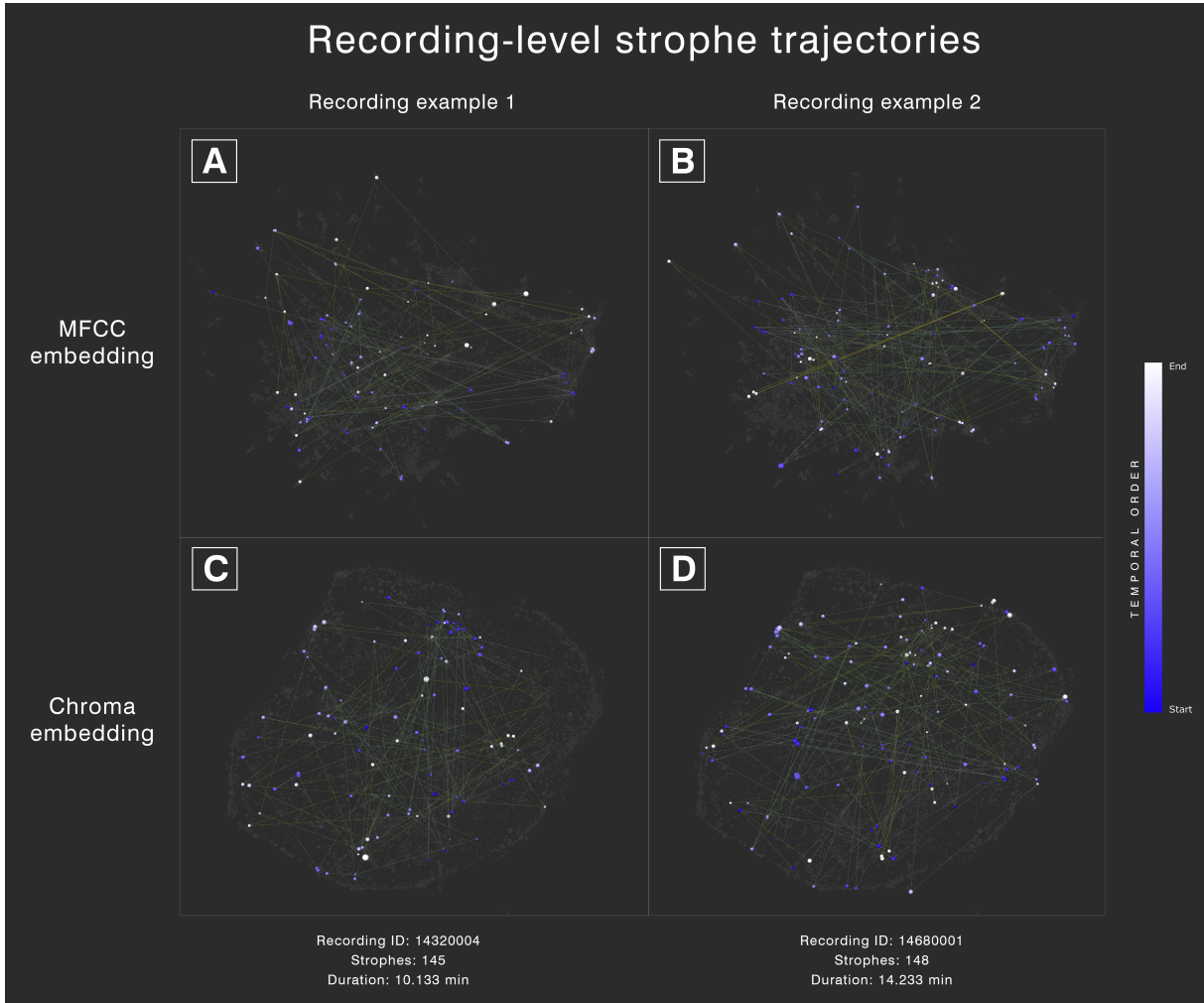


Figure 10: Recording-level trajectories through the low-dimensional acoustic spaces. Two representative recordings are shown across the MFCC and high-resolution chroma UMAP embeddings. **(A–B)** MFCC trajectories for recording examples 1 and 2. **(C–D)** Corresponding chroma trajectories for the same recordings. Indexed strophes belonging to each recording are highlighted against the global embedding and connected according to their original sequential order within the recording. Point colour indicates temporal progression from the beginning to the end of each recording. Connecting lines indicate sequence order rather than metric paths within the embedding. Recording identifiers, number of indexed strophes, and recording durations are reported below each column. The complete corpus-wide embeddings are shown in transparency for spatial context.

3.4 Recording-level retrieval and sequential trajectories

Because strophe provenance was preserved throughout the analytical pipeline, individual recordings could be retrieved within the corpus-wide embeddings without reconstructing or refitting the underlying spaces. Figure 10 illustrates two recordings containing 145 and 148 indexed strophes, respectively, displayed within both the shared MFCC and chroma embeddings (Fig. 10A–D).

In both acoustic representations, strophes from the selected recordings were distributed across multiple portions of the corpus-wide embeddings instead of remaining confined to a single local region. Reintroducing their original indexed order allowed successive strophes to be followed from the beginning to the end of each recording, providing a recording-level view of sequential progression through the low-dimensional representation. The same source sequence could therefore be inspected under two distinct acoustic representations while retaining common recording provenance and temporal order.

Connecting segments represent temporal succession between consecutive indexed strophes. The trajectories were consequently used as visual and navigational representations and were not interpreted quantitatively in terms of path length, curvature, or displacement. Each displayed strophe remained linked to its source audio, position within the recording, acoustic terrains, spectrogram, and associated field metadata, allowing individual points and sequence positions to be returned directly to the original acoustic material for listening and inspection. An interactive example of this recording-level trajectory navigation is provided in Supplementary Video V7 (Table S10).

The trajectory representation therefore demonstrates how corpus-wide relational spaces can retain access to the sequential structure of their source recordings: individual strophes can be viewed within the global representation while remaining recoverable as ordered acoustic events.

4 Discussion

4.1 A shared but non-uniform acoustic domain

The analyses of the complete corpus describe an organization that is neither subdivided into observation-area-specific domains nor reducible to homogeneous mixing among the sampled sources. In both acoustic representations, strophes from all ten observation areas occupied broadly shared portions of the corpus-wide spaces, while the quantitative PCA analyses revealed systematic differences in centroid position, dispersion, and neighborhood composition. The principal organization emerging from the corpus is therefore one of extensive shared acoustic occupancy combined with persistent local differentiation.

That is an important distinction, because broad occupation of common acoustic space does not imply equivalence of the sampled repertoires. The present analysis does not estimate repertoire saturation for individual sources, and it does not establish that acoustically neighboring strophes would necessarily be categorized under the same strophe type. Instead, it shows that a substantial portion of the sampled acoustic variation is organized within broadly shared relational structures. Within those structures, different observation areas contribute at varying frequencies and local concentrations. The appropriate interpretation is therefore shared organization of the sampled acoustic material rather than complete repertoire sharing among putative individuals.

This interpretation is reinforced by the measurements of within-area dispersion and between-area centroid separation. In MFCC PCA space, RMS within-area radii ranged from 91.7 to 99.9 PCA units, whereas the largest pairwise centroid separations were approximately 25 units. Chroma showed the same broad relationship at its own representation-specific scale, with within-area radial dispersion large relative to centroid displacement. These relative magnitudes alone cannot demonstrate overlap between broad and potentially multimodal distributions; however, the independent neighborhood analyses directly showed extensive cross-area proximity among individual strophes. Taken together, these global and local measurements support an organization

in which observation-area distributions occupy substantially shared acoustic domains rather than isolated acoustic compartments.

Shared occupancy did not, however, imply uniform occupancy. Area 7 provides an especially informative example under the MFCC representation. It showed the lowest RMS within-area dispersion while simultaneously participating in several of the largest centroid separations, including those with Areas 9, 1, 2, and 3. Areas 5 and 6, by contrast, formed the closest centroid pair in both MFCC and chroma PCA representations. These relationships illustrate why within-area dispersion, centroid position, and between-area separation should be treated as distinct geometric properties rather than different expressions of a single quantity. An observation-area corpus may therefore be comparatively compact while occupying a relatively displaced position, or comparatively broad without having an unusually separated centroid.

The local-neighborhood results add another level to this organization. Under Euclidean distance and after excluding candidates originating from the focal recording, 18.6% of MFCC neighbors and 14.4% of chroma neighbors, on average, originated from the focal observation area. Mean normalized Shannon entropy was also high in both representations, at 0.782 and 0.822, respectively. At the same time, same-area occupancy varied substantially among observation areas, and the enrichment analyses revealed recurrent departures from what would be expected from candidate availability alone.

The resulting organization can therefore be understood as differential occupancy within a broadly shared acoustic domain. Several sources may occupy common portions of acoustic space while contributing to those portions at markedly different relative frequencies. Under this interpretation, a region need not be exclusive to one observation area in order to contribute to source-associated differentiation. Conversely, the occurrence of acoustically similar strophes across several observation areas does not require that those strophes occur with comparable prevalence, sequential context, or behavioral significance. Source-related organization may consequently be expressed through differential occupancy of a shared acoustic domain rather than through the formation of sharply separated source-specific spaces.

4.2 Acoustic representation and metric as analytical lenses

A second major result is that relational organization was conditional both on the acoustic description of the strophe and on the geometrical criterion used to define proximity. MFCC and high-resolution microtonal chroma produced partially corresponding but clearly non-identical organizations. Changing the distance metric within either fixed PCA representation produced an additional reorganization, particularly pronounced in the local-neighborhood analyses. These dependencies should not be treated merely as technical complications surrounding an otherwise unique relational geometry. Rather, they show that different acoustic representations and proximity criteria emphasize different aspects of corpus organization.

4.2.1 Agreement and divergence between MFCC and microtonal chroma

The two acoustic representations emphasize substantially different properties of the original signal. MFCC terrains describe the temporal development of spectral-envelope structure; microtonal chroma terrains emphasize fine octave-folded relative-frequency organization. The resulting PCA spaces therefore do not organize all strophes identically. What is informative is the combination of agreement and divergence observed between them.

Some broad relationships persisted across the two representations. Most notably, Areas 5 and 6 formed the closest centroid pair in both PCA spaces despite their different coordinate scales and feature constructions. Both representations also produced extensive cross-area neighborhood mixing and non-uniform source-associated enrichment. Such correspondences indicate that some aspects of corpus organization recur across substantially different descriptions of the acoustic signal.

The differences were nevertheless pronounced. Under Euclidean distance, MFCC neighborhoods retained a larger same-area component than chroma neighborhoods (18.6% versus 14.4%) and showed lower normalized entropy (0.782 versus 0.822). MFCC enrichment was also more heterogeneous. Area 7, for example, reached approximately $7.0\times$ enrichment around Area 4 and $9.62\times$ self-enrichment after same-recording exclusion, whereas its chroma self-enrichment was approximately $3.04\times$. Chroma therefore produced a more extensively mixed local organization under the same Euclidean neighborhood definition.

These differences should not be interpreted as competing estimates of the same acoustic quantity. Two strophes may share pronounced spectral-envelope morphology while differing in relative-frequency organization, or exhibit the converse relationship. MFCC and microtonal chroma consequently function as complementary acoustic lenses rather than alternative attempts to reconstruct a single feature-independent notion of similarity.

Relationships recurring across both representations have a different interpretive status from relationships that depend strongly on one acoustic description. Conversely, representation-specific relationships can themselves be informative by showing that a given similarity depends on the acoustic properties emphasized by that representation. Agreement and divergence between the two spaces are therefore both analytical outcomes rather than one being necessarily interpreted as confirmation and the other as failure.

4.2.2 Euclidean and cosine distance reveal different local organizations

Metric dependence was particularly pronounced in local neighborhood organization. Replacing Euclidean with cosine distance increased mean same-area neighborhood occupancy from 18.6% to 25.4% in MFCC and from 14.4% to 20.1% in chroma. Neighborhood entropy decreased correspondingly, from 0.782 to 0.715 for MFCC and from 0.822 to 0.760 for chroma. Cosine distance therefore produced stronger aggregate observation-area locality in both acoustic representations.

This increase did not correspond to preservation of the detailed Euclidean organization. Off-diagonal enrichment patterns showed very weak correspondence between metrics, with Spearman correlations of approximately $\rho = 0.04$ for MFCC and $\rho = 0.14$ for chroma. The identity and relative strength of cross-area associations were consequently highly conditional on the metric used to define local proximity.

This divergence has a direct geometrical interpretation. Euclidean distance measures absolute displacement between strophes in retained PCA space and, because the PCA coordinates were not whitened before distance calculation, retains the variance-dependent scaling of the principal-component axes. Cosine distance instead measures the angular relationship between PCA score vectors, disregarding differences in their overall magnitude. Because the PCA scores are centered on the corpus mean, cosine distance compares the directions in which strophes deviate from that mean. Euclidean distance, by contrast, is translation-invariant and measures absolute pairwise separation while retaining the variance-dependent scaling of the unwhitened principal-component axes.

The weak correspondence between their neighborhood organizations should therefore not be interpreted simply as a lack of robustness. Euclidean and cosine distance pose different relational questions to the same acoustic representation. Two strophes that are close under cosine but not under Euclidean distance may exhibit similar directions of acoustic deviation from the corpus mean despite differing substantially in magnitude. Conversely, proximity under Euclidean distance may depend more strongly on their absolute positions within the retained PCA geometry.

Metric choice consequently forms part of the scientific definition of the relationship being measured rather than constituting a purely technical parameter applied after the acoustic representation has already determined the relevant notion of similarity.

4.2.3 Beyond a single “true manifold”

Taken together, the representation- and metric-dependent results argue against treating relational analysis and dimensionality reduction as procedures for recovering a single latent geometry that can be regarded as the uniquely correct organization of the vocal corpus. In the present study, the same indexed strophes produced partially corresponding but non-identical relational structures under MFCC and microtonal chroma representations, while changing the distance metric within a fixed PCA space produced a further and, in some cases, pronounced reorganization of local neighborhoods. The observed geometry is therefore conditional on both the acoustic information encoded by the representation and the criterion through which similarity is defined. This interpretation is consistent with broader work on latent-space representations of animal vocalizations, where the organization obtained is understood to depend on the acoustic representation, similarity criterion, and dimensionality-reduction procedure used to construct the space [12, 10].

This conditionality does not imply that the resulting geometries are arbitrary or equally informative for every question. Each representation emphasizes specified properties of the acoustic signal, and each metric implements a defined geometrical relationship among the resulting observations. Their consequences can therefore be examined explicitly rather than remaining hidden within the analytical pipeline. Relationships that recur across representations, metrics, subsampling conditions, or neighborhood scales acquire a different evidential status from relationships that depend strongly on a particular analytical construction. Conversely, such dependence can itself be informative when it can be related to the acoustic or geometrical property emphasized by that construction.

The analytical objective is consequently not to identify one manifold that supersedes all alternative descriptions, but to determine which aspects of corpus organization remain stable across analytical views and which emerge only under particular acoustic or relational lenses. This distinction is especially important when low-dimensional embeddings produce visually compelling structures. Apparent clusters, branches, separations, or transitional regions may be stable within a particular representation and analytical construction without thereby constituting intrinsic properties of the vocal material independently of that construction.

For this reason, the framework treats quantitative geometry, low-dimensional visualization, and acoustic verification as related but distinct analytical levels. PCA provides retained high-dimensional coordinate systems in which explicit relational measurements can be performed, while UMAP provides a navigational projection whose visible geometry is interpreted qualitatively and conditionally. Alternative acoustic representations, distance metrics, subsampling conditions, and neighborhood scales are therefore treated as different analytical views rather than solely as competing attempts to recover a single hidden organization of the corpus.

4.3 Relational before categorical

The present framework establishes a specific analytical order: acoustic relationships among complete vocal events are constructed before categorical membership is specified. The present results show why this ordering matters. Without prior assignment of strophes to conventional types, the corpus exhibited measurable source-associated structure in centroid position, dispersion, neighborhood composition, and directional enrichment. At the same time, observation areas did not form isolated acoustic domains, and local neighborhoods remained extensively mixed across sources. Relational organization was therefore detectable without requiring categorical boundaries to be specified in advance.

This does not establish how the resulting geometry corresponds to conventional nightingale strophe types. That correspondence remains an independent empirical question. A direct extension of the present framework would be to assign strophe types independently and project those labels onto the established relational spaces. Compact type-associated distributions would indicate categories that are also acoustically coherent under the selected representation, whereas distributed or overlapping distributions could indicate substantial within-type heterogeneity, continuity between assigned classes, or acoustic properties emphasized by the representation that are not captured by the categorical criterion itself.

The present study therefore provides the relational side of that comparison. Its contribution is not to replace categorical description, but to show that structured acoustic relationships can be constructed and measured before categorical assignment. Classification can then be introduced as an additional analytical layer whose correspondence with the relational organization is tested rather than built into the geometry from the outset.

4.4 Quantitative geometry, navigation, and return to the acoustic source

A central property of the framework is that corpus-scale acoustic geometry remains connected to the individual observations from which it is constructed. This connection is particularly important in exploratory high-dimensional analysis, where mathematical or visual salience cannot by itself establish biological salience. A relationship may be well defined geometrically while its acoustic basis, categorical interpretation, or biological relevance remains uncertain. Interactive embedding approaches in bioacoustics have similarly emphasized the value of coupling low-dimensional acoustic organization with direct inspection of the underlying vocal material [23].

The recording-level trajectories illustrate how this connection can be preserved across analytical scales. Individual recordings occupied multiple regions of both MFCC and chroma UMAP embeddings, and reintroduction of indexed temporal order allowed successive strophes to be followed through the corpus-wide representations. These trajectories are not interpreted as metrically faithful paths: their apparent length, curvature, and extent remain properties of the nonlinear projection. Their value instead lies in combining global acoustic organization with recoverable recording-level sequence context.

More generally, every represented strophe retains a direct path back to its source recording, original temporal position, acoustic terrain, spectrographic representation, and associated field metadata. Consequently, a conspicuous UMAP region, an unusual position in PCA space, or a locally enriched neighborhood can be traced back to the underlying signal for listening and inspection rather than interpreted exclusively through its geometric position. This return path provides an important interpretive safeguard by allowing apparent structure to be checked against the original acoustic material. This is especially important when a pattern depends strongly on

a particular representation or metric, or when visual prominence in UMAP is not matched by strong separation in the retained PCA space.

The analytical process is therefore bidirectional. Acoustic observations are transformed into indexed representations and relational geometries, while relationships identified within those geometries remain auditable against the original sound and its recording context. Quantitative PCA geometry, low-dimensional UMAP navigation, and source-linked retrieval consequently serve different but complementary roles: measurement, exploration, and acoustic verification. Their separation allows a corpus containing more than 20,000 indexed strophes to remain computationally tractable and relationally measurable without becoming detached from the empirical material through which those relationships must ultimately be interpreted.

4.5 Sampling structure, source attribution, and inferential limits

The relational patterns described in this study emerge from an observational field corpus rather than from a balanced experimental design. Observation areas contributed unequal numbers of recordings and indexed strophes, and their temporal and diel coverage also differed. These asymmetries reflect the progressive discovery and repeated detection of singing sources, accessibility, recording opportunity, acoustic contamination, and changes in detected vocal activity across the season.

The neighborhood analyses addressed unequal contribution in two complementary ways. Enrichment evaluated observed local composition relative to the material available within each eligible candidate pool, while the equal-contribution analysis repeated the measurements using a fixed balanced subset. Balancing altered the absolute same-area fractions and neighborhood entropies, confirming that these aggregate quantities were sensitive to corpus composition. Nevertheless, directional enrichment patterns remained positively concordant between the complete and balanced corpora and were broadly stable across the tested neighborhood sizes (Supplementary Tables S4 and S7). The broader ordering of cross-area relationships was therefore substantially preserved under both equal-contribution subsampling and changes in neighborhood scale, although their magnitude and detailed expression remained sensitive to corpus composition. These sensitivity analyses apply specifically to neighborhood composition; centroid and within-area dispersion measurements remain descriptions of the empirical distributions in the complete corpus rather than estimates from a balanced experimental design.

Candidates originating from the focal strophe’s raw recording were excluded from all reported neighborhood analyses. Same-area association and enrichment therefore reflected relationships across separate recordings rather than repeated strophes within the focal raw recording. This exclusion does not, however, establish statistical independence among the remaining observations. Recordings assigned to the same observation area may still share season, daypart, behavioral context, spatial location, acquisition conditions, and other sources of dependence.

The relevant unit of inference is therefore the observation-area corpus rather than the biological individual. As established in the Methods, observation areas were species-informed operational source hypotheses supported by repeated field evidence, but they did not constitute confirmed individual identities. The persistence of area-associated organization across separate recordings is compatible with recurrent activity of focal males, but does not establish individual-specific acoustic organization. Because attribution confidence varied among observation areas, the strength of any source-level interpretation is likewise graded. Stronger inference at the individual level would require independent identity evidence and/or longitudinal designs explicitly addressing continuity and turnover.

Temporal sampling imposes an additional limitation. The corpus provides broad seasonal coverage but was collected predominantly during morning and afternoon activity; systematic dawn sampling was not undertaken, and most nocturnal material originated from Area 2. Differences among observation areas therefore cannot be separated completely from possible effects of season, diel context, behavioral state, or recording conditions. The preservation of recording-level provenance makes these influences accessible to subsequent analysis, but resolving them will require designs that model the associated dependencies explicitly.

The resulting geometries should consequently be interpreted as measurable organization within the sampled acoustic material, not as definitive individual signatures, conventional strophe types, or functional vocal classes. Extensive cross-area occupancy does not establish complete repertoire sharing, and neighborhood enrichment does not by itself demonstrate individual preference, territorial function, or communicative meaning. The present results instead identify observation-area-associated structure whose expression remains conditional on acoustic representation, distance metric, sampling composition, and analytical scale.

These constraints define the next analytical stage rather than diminishing the present findings. The preserved provenance and relational spaces can support subsequent comparisons with conventional strophe classifications, analyses of vocal sequences and recording-level recurrence, explicit temporal and contextual models, and longitudinal comparisons across future breeding seasons. The present study establishes the empirical and analytical baseline from which those more narrowly defined biological questions can be tested.

4.6 Transferability to other animal vocal systems

Although developed here through a field corpus of common nightingale song, the analytical framework is not intrinsically tied to nightingale strophes or to conventional song-type organization. Its more general principle is the representation of each segmented vocal event as a temporally structured acoustic object that can be transformed into a fixed-dimensional whole-event representation and positioned within a corpus-wide relational space. In the present implementation, internal acoustic morphology is represented through MFCC and high-resolution microtonal chroma terrains. The same analytical architecture could in principle be applied to other vocal systems in which variation of interest is expressed through the internal temporal and acoustic organization of discrete vocal events. More broadly, learned and low-dimensional acoustic representations have been applied across multiple animal vocal systems as tools for describing repertoire structure, similarity, and vocal variation [24, 25].

Transferability does not imply that the feature representations or parameter choices used here should remain unchanged. Different animal vocal systems may require different definitions of the analytical event and may express relevant variation through spectral envelope, absolute or relative frequency structure, modulation, temporal organization, harmonicity, or other signal properties. MFCCs, microtonal chroma, and other frame-level acoustic representations can therefore be regarded as distinct analytical layers within a broader relational framework rather than as universal representations. The appropriate representation, temporal standardization, and distance geometry should be selected and evaluated according to the structure of the vocal material and the biological question being addressed.

The associated provenance and retrieval architecture is similarly transferable: each represented event can remain linked to its source audio, sequential position, and contextual metadata, allowing relational patterns to be inspected against the underlying signal. Nightingale song consequently provides the initial field-grounded application of the present framework rather than

defining its intended limit. Applications across other taxa and vocal systems could determine which organizational properties are specific to particular vocal repertoires and which emerge more generally when complete vocal events are compared as structured acoustic objects. Such transferability remains an empirical question and should be demonstrated through application rather than assumed from the present results alone.

5 Conclusion

This study establishes a field-grounded corpus and an analytical framework for examining large collections of nightingale vocalizations without requiring song categories to be defined in advance. Across 161 retained recordings and 20,075 indexed strophes from ten observation areas, complete vocal events were represented through MFCC and high-resolution microtonal-chroma terrains and placed within corresponding corpus-wide PCA spaces. Centroid position, within-area dispersion, and local-neighborhood composition provided distinct views of corpus organization, while preserved provenance maintained a direct connection between relational geometry and the underlying acoustic material.

The resulting organization was extensively shared but non-uniform. Strophes associated with the ten observation areas did not form isolated acoustic compartments, yet their distributions were not independent of observation-area attribution. Source-associated structure appeared through differences in centroid position and dispersion and through non-uniform composition and enrichment within local acoustic neighborhoods. MFCC and microtonal chroma produced partially corresponding but non-identical organizations, while Euclidean and cosine distance substantially reorganized local relationships within the same retained PCA spaces. Acoustic organization was therefore conditional both on the properties encoded by the representation and on the geometrical criterion through which proximity was defined.

Sensitivity analyses further showed that the broader ordering of directional enrichment relationships remained substantially preserved under equal-contribution subsampling and across neighborhood sizes, although the absolute composition of local neighborhoods remained sensitive to corpus composition and analytical scale. These checks do not transform the observational corpus into a balanced experimental design, nor do the observation-area labels establish confirmed individual identity. The present results should therefore be interpreted as measurable organization within the sampled acoustic material rather than as definitive individual repertoires, strophe types, or functional vocal classes.

The preserved coordinate spaces, sequential provenance, and source-linked retrieval architecture provide a basis for more targeted analyses of this organization. Independently assigned strophe classifications can be compared with the relational spaces; temporal and sequential structure can be examined explicitly; recording-level and hierarchical analyses can address contextual dependence; and additional field seasons can test the persistence of source-associated relationships through time. The same analytical architecture can also be evaluated in other animal vocal systems, with event definitions, acoustic representations, and distance geometries adapted to the structure of each vocal domain. Such transferability remains an empirical question rather than an assumption of the present study.

The principal contribution of this work is therefore not a finished taxonomy of nightingale song or a nightingale-specific classification pipeline, but a reproducible and adaptable relational foundation from which such questions can be formulated and tested. Relationships are examined before categories are assigned, and the analytical space is constructed before its resulting organization is known. Within those spaces, nightingale strophes form overlapping and locally differentiated

relational structures: sufficiently coherent to support quantitative comparison, yet sufficiently dependent on acoustic representation, distance geometry, sampling structure, and scale that no single analytical view can be assumed to constitute a uniquely correct organization of the corpus.

6 Statements and declarations

Data and code availability

The data and audiovisual materials supporting this study are publicly available in a versioned Figshare repository (<https://doi.org/10.6084/m9.figshare.33497311>). The deposit contains the recording-level master table; a documented, recording-stratified sample of 500 indexed strophes, comprising 50 strophes from each observation area, with corresponding audio files and selection manifests; strophe-level MFCC and chroma PCA-1000 coordinates, UMAP coordinates, computed descriptors, and associated metadata for all 20,075 retained strophes; the aggregated outputs underlying the reported centroid, dispersion, neighborhood, equal-contribution, and denoising-sensitivity analyses; and the nine supplementary videos. Detailed supplementary tables are included within the preprint.

The complete raw and processed acoustic corpus is not included in this initial public deposit. Primary follow-up analyses based on the 2026 field corpus remain in active development, and requests for access to additional non-public acoustic material may therefore be considered on a case-by-case basis for defined research purposes. Requests substantially overlapping with primary analyses already in progress may be deferred until those analyses have been completed. The custom Python code used to generate the reported representations and analytical outputs is likewise not included in this initial release and may be provided separately for replication or validation purposes upon reasonable request.

Ethics and permissions

This study was based exclusively on non-invasive field observation and passive acoustic recording of free-living common nightingales (*Luscinia megarhynchos*). No animals were captured, handled, restrained, marked, biologically sampled, or subjected to experimental manipulation. No playback, baiting, or other induced stimulation was used, and no nests or breeding sites were physically interfered with. The study therefore involved no procedures constituting animal experimentation, and formal animal ethics approval was not required. Field recordings were conducted in accordance with the access conditions specific to the study area.

Author contributions

Lucio Arese (author): Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization, Writing—original draft, Writing—review and editing, and Project administration. The author conducted all fieldwork, including source localization, georeferencing, field observation, and audio acquisition; constructed and curated the recording corpus and associated metadata; performed audio preprocessing, manual contouring, strophe indexing, quality control, and preparation of the deposited data package; developed and implemented the analytical and visualization framework; designed and produced all figures, maps,

analytical visualizations, and supplementary audiovisual materials; and developed the custom TouchDesigner interfaces used for interactive inspection, navigation, retrieval, and visualization of the acoustic data.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of generative AI assistance

During the preparation of this work, the author used ChatGPT (OpenAI; GPT-5.5, GPT-5.6, and GPT-5.6 Sol) for programming assistance, organization of analytical outputs and manuscript structure, and editorial revision of draft manuscript prose. All field data collection, methodological decisions, analytical choices, factual verification, interpretation of results, and final crafting, editing, and revision of the manuscript text were performed by the author. All AI-assisted code and textual outputs were reviewed, tested where applicable, corrected as necessary, and incorporated only at the author's discretion. The final selection, formulation, editing, and revision of the manuscript content were carried out by the author. The author retains full responsibility for the content of the publication.

Competing interests

The author declares no competing interests.

Supplementary materials

Table S0: Overview of supplementary materials

Section	Contents
S1. Field corpus, sampling structure and dataset construction	Construction of the field corpus across the 68-day campaign; emergence and definition of the operational observation areas; recurrent singing locations and activity areas; temporal and acquisition constraints; corpus curation, sampling asymmetries, field documentation, and recording-level provenance.
S2. Preprocessing and denoising sensitivity	Documentation of the two denoising configurations and their frequency-dependent reduction curves; comparison of MFCC and chroma representations before and after the second denoising pass; summary and recording-level sensitivity results; and a representative spectrographic view of the preprocessing and indexing sequence.
S3. PCA neighborhood sensitivity analyses	Evaluation of neighborhood organization across $k = 10, 25, 50$, and 100 , and under equal observation-area contribution. The balanced analysis uses a fixed subset of 500 strophes per area ($n = 5,000$) within the original globally fitted PCA spaces, with comparisons across MFCC and chroma representations and Euclidean and cosine distance.
S4. Field-metadata overlays on the MFCC manifold	Descriptive overlays of selected recording-level and spatial metadata on a representative MFCC embedding. The section connects recording events represented as mapped field locations with the corresponding groups of strophes distributed as point clouds within the acoustic geometry.
S5. Supplementary data package and audiovisual materials	Description of the deposited data package, including the recording master table, representative 500-strophe audio subset, strophe-level PCA and UMAP coordinate tables, quantitative PCA-analysis outputs, and denoising-sensitivity results. Supplementary Videos V1–V9 document acoustic terrains, manifold navigation, metadata inspection, strophe retrieval, and return to the source audio.
S6. Supplementary tables	Complete recording inventory and annotation definitions; recording-level denoising-sensitivity results; strophe-indexing parameters; technical settings for MFCC and chroma computation, terrain construction, PCA, and UMAP; and software and library versions.

S1 Field corpus, sampling structure and dataset construction

S1.1 Corpus overview and development of sampling structure

The main Methods describe the field corpus, observation-area definitions and recording procedures used in the analysis. This section provides additional documentation of how the corpus developed during fieldwork and of the spatial, temporal and acquisition asymmetries retained in the final dataset.

The 2026 field campaign covered 68 days from mid-April to late June and produced 227 geo-referenced recordings across ten observation areas. Of these, 161 recordings were retained for analysis, yielding 20,075 indexed strophes from approximately 42.5 h of raw field audio. Recording metadata, inclusion status and annotation variables are reported in Supplementary Tables S11A– S11C, and the complete inventory is additionally supplied as a machine-readable dataset.

Sampling structure emerged progressively rather than from predefined recording stations. At the beginning of the campaign, the local distribution of singing nightingales was largely unknown. Accessible portions of the riverine corridor were therefore surveyed, active singers were located, and candidate source areas were reassessed across successive visits. Recording positions were comparatively mobile during this exploratory phase, and the observation-area boundaries retained in the final corpus had not yet been established.

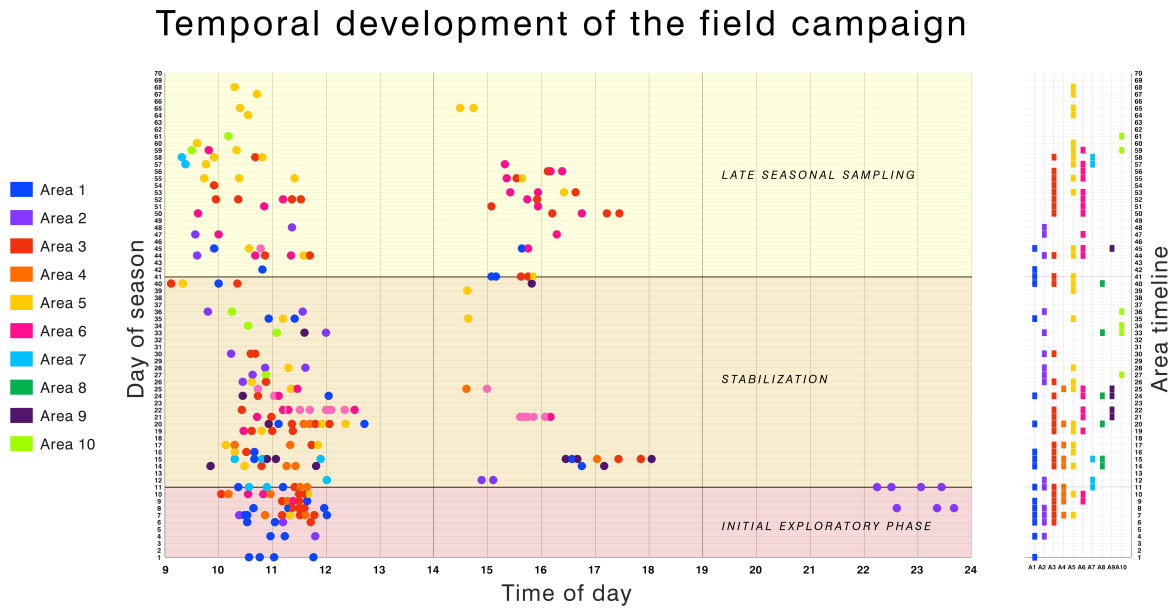


Figure S1: Temporal development of the 2026 field campaign across season day, time of day and observation area. Points represent individual recordings and are coloured by observation area; the panel at right summarizes their seasonal distribution within each area. Shading identifies the exploratory, stabilization and late-season phases of fieldwork. These phases describe the empirical progression of sampling and were not used as formal sampling strata.

Repeated localization gradually produced a more stable field structure. Recurrent singing positions, spatially restricted movements and simultaneous activity from neighboring singers allowed operational source areas to be distinguished with increasing confidence. Recording effort consequently became organized around these recurrent locations and the associated areas of movement.

The distribution of effort changed again as detected vocal activity declined unevenly late in

the season. Singing was no longer detected during later observation visits to several previously sampled areas, whereas it continued to be detected in a smaller set of locations, particularly within the southern portion of the study system. Late-season recording therefore concentrated on sources that remained detectable during the available observation windows. These patterns describe the temporal distribution of field detections and should not be interpreted as continuous measures of singing activity outside the periods in which each area was observed. The exploratory, stabilization and late-season phases shown in Fig. S1 describe the empirical development of the campaign and were not used as formal sampling strata.

S1.2 Spatial source documentation

The biological and operational basis of observation-area attribution is defined in the main Methods. The supplementary material provided here documents the spatial observations underlying those operational source hypotheses and preserves distinctions among different levels of localization evidence.

Mapped recurrent singing posts represent the most spatially precise source anchors available from field observation. Where singing activity could not be associated consistently with a single post, restricted or broader activity areas represent spaces within which a singer was observed or acoustically tracked moving among positions. Observation-area polygons provide a wider operational envelope for recurrent activity and are therefore not interchangeable with these more precise source anchors (Fig. S2).

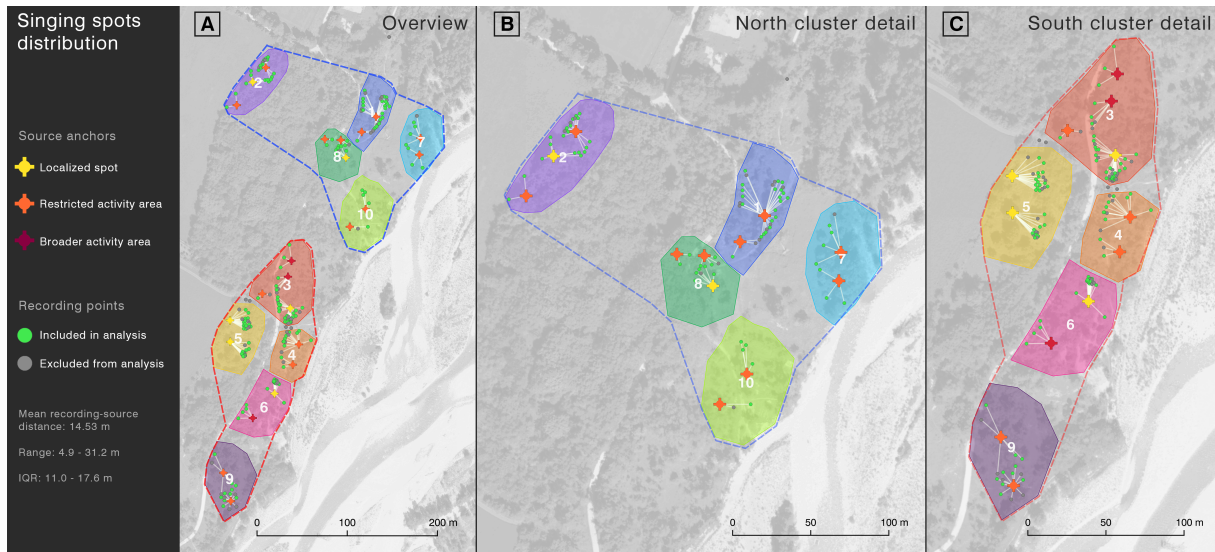


Figure S2: Spatial distribution of recurrent singing locations, activity areas and recording points across the ten observation areas. (A) complete study system, (B) northern grouping and (C) southern grouping. Source anchors distinguish localized singing posts from restricted and broader activity areas. Green and grey points indicate recordings retained in and excluded from the analytical corpus. QGIS-derived distances from recording positions to the corresponding mapped focal singing locations averaged 14.53 m (IQR: 11.0–17.6 m; range: 4.9–31.2 m). Observation-area polygons provide operational envelopes for recurrent activity rather than formal territorial boundaries, and the northern and southern groupings are shown for spatial organization of the mapped study area rather than as biological units. Basemap: AGEA 2024 orthophoto, Regione Piemonte / CSI Piemonte WMS service regp_agea_2024; © AGEA 2024. Study overlays and map composition by the author; prepared in QGIS 4.0.1.

Representative views of the ten observation areas are shown in Fig. S3. These photographs provide visual documentation of the vegetation structure, openness and access conditions encountered during fieldwork. They were not collected through a standardized habitat-sampling design and were not used as environmental variables in the analyses.

Field notes additionally preserve episodes of simultaneous singing, close-range interaction and observed confrontation among neighboring birds. These observations were not collected under



Figure S3: Representative field views of the ten operational observation areas. The photographs provide qualitative visual context for the vegetation structure, openness and access conditions encountered during fieldwork. They were not acquired through a standardized habitat-sampling protocol, were not used as environmental covariates and should not be interpreted as measurements of habitat composition or formal territorial boundaries.

a dedicated behavioral protocol and are retained only as contextual evidence associated with source localization.

S1.3 Acquisition context, visual documentation and corpus provenance

Temporal and diel sampling were uneven. Recording effort was concentrated in the morning, with a substantial additional contribution from afternoon observations, while systematic dawn sampling was not undertaken. Night recording was constrained by safe field access; Area 2, adjacent to accessible open terrain lit by the nearby city lights, consequently contributed most of the nocturnal material. The corpus therefore provides broad seasonal coverage but not a balanced sample across dayparts.

All field material was acquired as synchronized audiovisual recordings, with the video camera serving as the audio recorder for the external condenser shotgun microphone. Consequently, every audio recording retained in the field corpus has a corresponding video record. The video material provides recording-level visual provenance even when the singing bird is not visible, documenting the recording position and, to varying degrees, surrounding vegetation, environmental conditions and other contextual features present during acquisition. Audio tracks were extracted for acoustic processing, while the original video files were retained as part of the source record. In a subset of recordings, the singing bird was visible within the video frame, providing direct visual documentation associated with the corresponding acoustic event (Fig. S4). Such observations confirm the presence of a singing nightingale during the illustrated recording but do not establish persistent individual identity across recordings or dates.

Field acquisition generally prioritized close-range recording, typically within approximately 10–20 m of the mapped focal singing location. QGIS-derived distances had a mean of 14.53 m, an interquartile range of 11.0–17.6 m, and an overall range of 4.9–31.2 m (Fig. S2). This produced strong focal signals while placing the observer within the immediate environment of the singer. Possible observer influence was therefore annotated during corpus review. These observer-effect annotations are retained as descriptive contextual variables, and they are not formal behavioral measurements (Supplementary Tables S11B and S11C).

Recordings were retained for analysis when the focal vocal material could be isolated reliably for strophe indexing and acoustic representation. Most exclusions resulted from overlapping biological

sound, particularly blackbird and blackcap vocalizations, or other interference preventing reliable separation of the nightingale signal. Low focal-signal level was not the dominant cause of exclusion.



Figure S4: Representative video-documented visual contacts with singing nightingales. Frames extracted from synchronized audiovisual field recordings show singing birds in Area 1 (recording ID 14160004), Area 4 (14240010), Area 6 (14530005) and Area 8 (14250025). Additional direct visual contacts occurred in Areas 3 and 5 but did not yield video frames showing the bird. Each illustrated frame documents the presence of a singing nightingale during the corresponding recording event; it does not establish that recordings obtained from the same observation area on other dates originated from the same individual. Recording identifiers refer to the complete field-corpus inventory; analytical inclusion status is reported in Supplementary Table S11A.

Recording effort and indexed-strophe counts consequently remained unequal among observation areas. These differences reflect source discovery, accessibility, variation in detected vocal output, persistence of detections across the season, recording opportunity and corpus-curation criteria. The imbalance was retained as a property of the empirical field corpus rather than removed during corpus construction; its consequences for neighborhood measurements are evaluated separately through the equal-contribution sensitivity analysis in Supplementary Section S3.

Supplementary Table S11A provides the complete recording-level inventory, including spatial, temporal, duration and analytical-inclusion metadata. Tables S11B and S11C document the associated annotation variables and their definitions. Together with the machine-readable master dataset, these records preserve the provenance required to filter the field corpus and reconnect subsequent analytical results to the recordings from which they were derived.

S2 Preprocessing and denoising sensitivity

S2.1 Denoising sensitivity checks

Audio preprocessing comprised two successive applications of iZotope RX 11 Spectral De-noise. For each recording, a noise profile was learned manually from a selected noise-only interval. Adaptive processing was disabled, while all remaining controls—including the custom frequency-dependent reduction curves—were held constant across the corpus. The first pass was designed to remove the dominant low-frequency acquisition-noise floor, introduced primarily by the camera preamplifier, while limiting processing within the principal song-bearing range of approximately 1–15 kHz. Its output constituted the floor-stripped baseline. The second pass applied a deliberately mild, broader denoising treatment and produced the final-denoised recordings used for feature computation. The two configurations are shown in Fig. S5, with complete parameter values reported in Supplementary Tables S2 and S3.

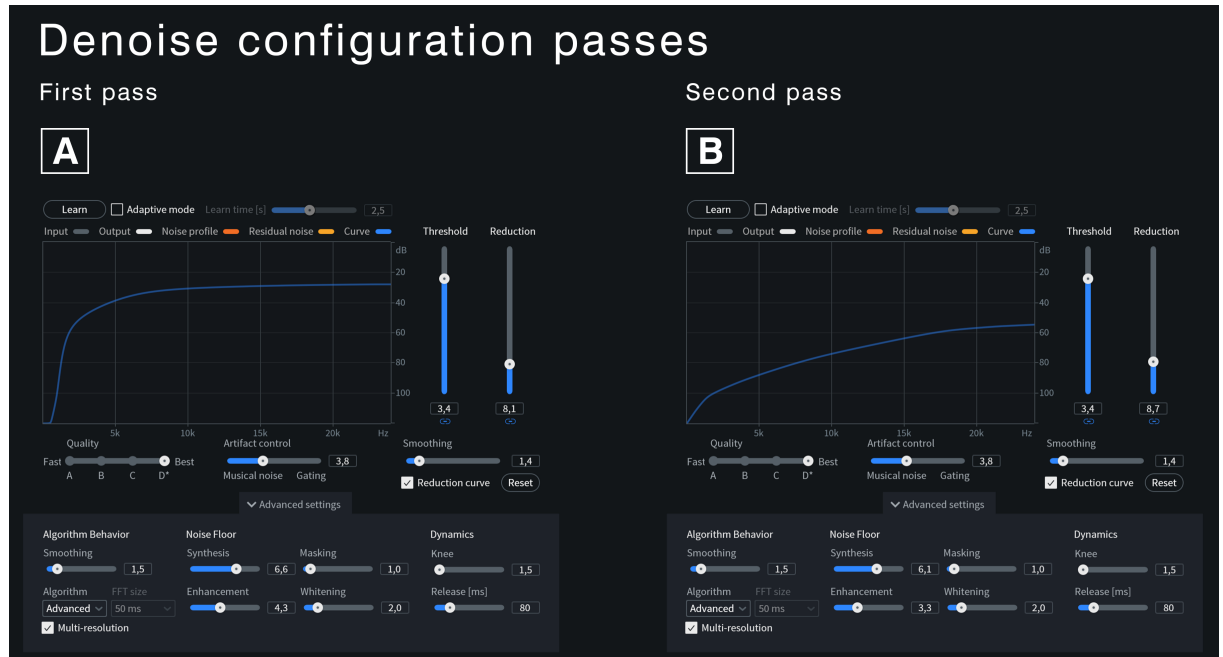


Figure S5: Fixed iZotope RX 11 Spectral De-noise configurations used during the two-pass preprocessing procedure. (A) First pass, configured to concentrate noise reduction toward the low-frequency acquisition-noise floor while applying progressively less reduction through the principal song-bearing range. (B) Second pass, applying a broader frequency-dependent reduction profile. Blue lines show the custom reduction curves held constant across recordings; lower curve positions correspond to greater noise reduction and higher positions to less reduction in the native software display. Recording-specific noise-profile and signal traces have been hidden. Complete numerical and categorical settings are reported in Supplementary Tables S2 and S3.

To isolate the incremental effect of the second pass, the floor-stripped and final-denoised versions of all 161 retained recordings were compared directly. Using the raw recordings as the reference would instead have caused the measurements to be dominated by removal of the low-frequency technical-noise floor during the first pass. MFCC-40 and L1-normalized Chroma-80 representations were calculated for corresponding frames in each paired recording. Comparisons were performed using the highest-RMS 35% and 75% of frames, with the gates defined from the floor-stripped baseline.

The two gates provided a more restrictive signal-focused condition and a broader condition containing a greater contribution from lower-amplitude vocal and environmental material.

Three complementary measurements were calculated at each gate: recording-profile cosine similarity, the median cosine similarity between corresponding frames, and relative Frobenius change [26]. The cosine measurements evaluated directional agreement at recording and frame levels, whereas relative Frobenius change retained sensitivity to numerical differences in feature magnitude. Complete recording-level results are reported in Supplementary Table S12.

Across both gates, MFCC and chroma retained consistently high profile and framewise cosine agreement (Table S1). The broader 75% gate produced a modest reduction in corresponding-frame agreement, particularly for chroma, while relative Frobenius change showed measurable numerical differences, especially for MFCC. The results therefore support substantial preservation of feature-profile organization after the second pass, rather than an absence of numerical change.

Table S1: Corpus-wide denoising-sensitivity results across the 161 retained recordings.

Representation Gate Metric			Mean $\pm$ SD	Median [IQR]	P05–P95
MFCC-40	35%	Profile cosine	0.98749 $\pm$ 0.00777	0.98925 [0.98540–0.99176]	0.97483–0.99443
MFCC-40	35%	Framewise cosine median	0.97773 $\pm$ 0.00550	0.97897 [0.97577–0.98135]	0.96701–0.98441
MFCC-40	35%	Relative Frobenius change	0.397 $\pm$ 0.038	0.393 [0.374–0.422]	0.336–0.456
MFCC-40	75%	Profile cosine	0.98602 $\pm$ 0.00869	0.98792 [0.98429–0.99100]	0.97121–0.99396
MFCC-40	75%	Framewise cosine median	0.97531 $\pm$ 0.00624	0.97665 [0.97347–0.97968]	0.96282–0.98186
MFCC-40	75%	Relative Frobenius change	0.411 $\pm$ 0.035	0.407 [0.391–0.426]	0.361–0.473
Chroma-80	35%	Profile cosine	0.99745 $\pm$ 0.00506	0.99916 [0.99753–0.99970]	0.98901–0.99992
Chroma-80	35%	Framewise cosine median	0.99560 $\pm$ 0.00486	0.99719 [0.99418–0.99902]	0.98519–0.99969
Chroma-80	35%	Relative Frobenius change	0.220 $\pm$ 0.077	0.212 [0.160–0.277]	0.106–0.355
Chroma-80	75%	Profile cosine	0.99664 $\pm$ 0.00457	0.99842 [0.99560–0.99929]	0.98818–0.99974
Chroma-80	75%	Framewise cosine median	0.98725 $\pm$ 0.00637	0.98730 [0.98328–0.99171]	0.97633–0.99759
Chroma-80	75%	Relative Frobenius change	0.280 $\pm$ 0.051	0.276 [0.248–0.304]	0.198–0.375

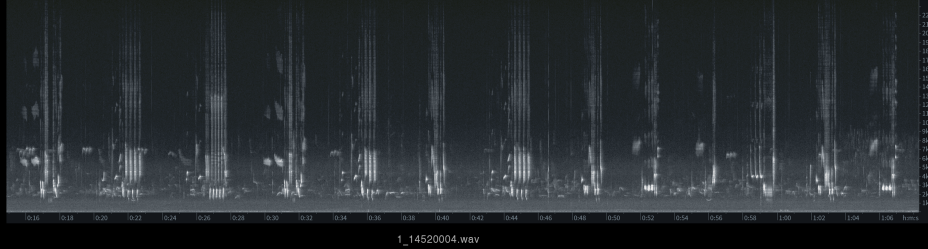
Notes. Statistics treat recordings as equal-contribution units ($n = 161$). SD is the population standard deviation of the 161 recording-level values. *Framewise cosine median* denotes the median corresponding-frame cosine calculated within each recording before corpus-level summarization. P05–P95 gives the 5th–95th percentile range.

Figure S6 provides a complementary qualitative view of the complete sequence from raw recording to indexed output. The numerical sensitivity comparison corresponds specifically to the floor-stripped baseline and final-denoised stages shown in panels B and C; the subsequent contouring and indexing stages were not included in that comparison.

Representative spectrographic view of the preprocessing and indexing pipeline

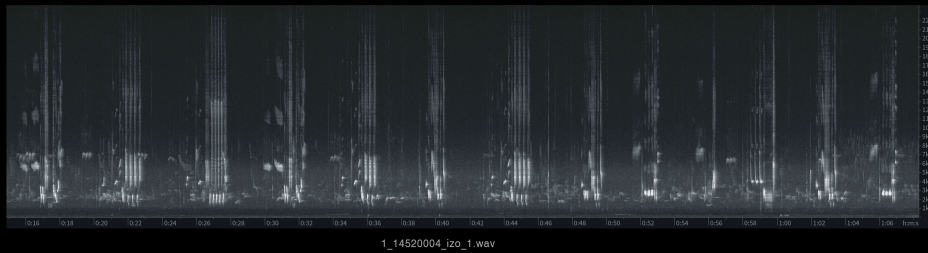
Recording ID: 14520004 (excerpt)

A



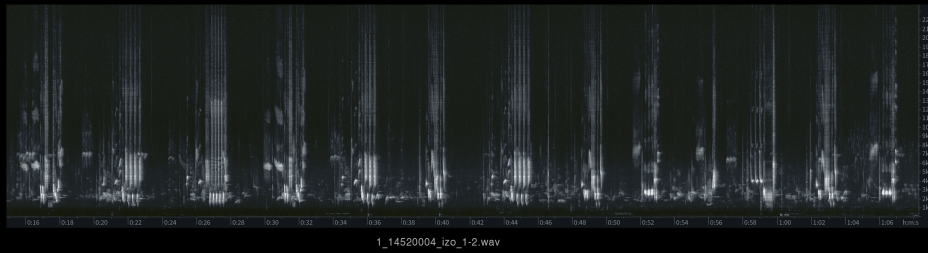
Raw audio

B



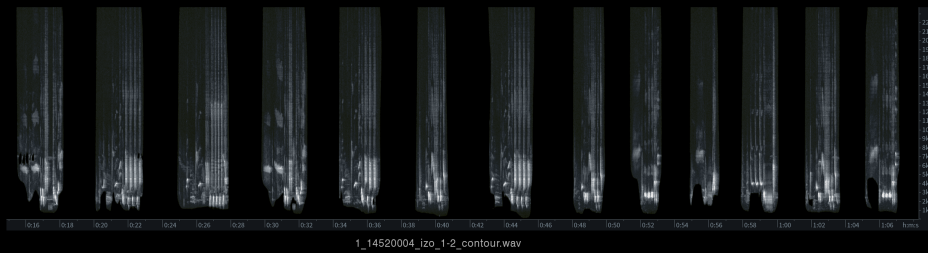
First pass:
floor-denoised baseline

C



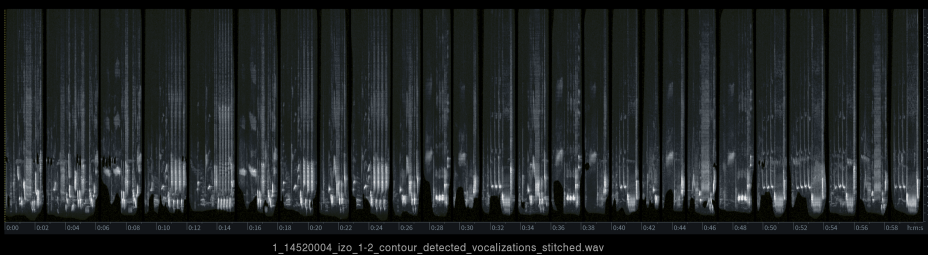
Second pass:
final denoised audio

D



Contoured audio

E



Indexed strophes:
stitched audio output

Figure S6: Representative spectrographic view of the preprocessing and indexing pipeline. The sequence shows (A) raw audio, (B) the floor-stripped baseline following the first denoising pass, (C) the final audio following the second denoising pass, (D) manually contoured audio, and (E) indexed strophes concatenated into the stitched output. Panels A–D show the same interval on the original recording timeline, whereas panel E follows the reconstructed timeline of the indexed output. The quantitative sensitivity analysis compares stages B and C; the spectrograms provide a qualitative illustration of the broader preprocessing sequence.

Table S2: Fixed iZotope RX 11 Spectral De-noise settings for the first preprocessing pass, which produced the floor-denoised baseline.

Parameter	Corpus setting
Module	iZotope RX 11 Spectral De-noise
Preset identifier	<code>Birdsong_lowfloor_strip</code>
Noise-profile acquisition	Manual Learn from a selected noise-only interval
Adaptive mode	Disabled
Adaptive learn time	Inactive; 2.5 s displayed while adaptive mode was disabled
Threshold	3.4 dB
Reduction	8.1 dB
Frequency-dependent reduction curve	Enabled; custom first-pass curve (Fig. S5A)
Quality	D* (Best)
Artifact control	3.8
Main smoothing	1.4
Algorithm	Advanced
Algorithm-behavior smoothing	1.5
Multi-resolution processing	Enabled
FFT size	50 ms displayed; inactive in multi-resolution mode
Noise-floor synthesis	6.6
Noise-floor enhancement	4.3
Noise-floor masking	1.0
Noise-floor whitening	2.0
Dynamics knee	1.5
Dynamics release	80 ms
Output noise only	Disabled

Table S3: Fixed iZotope RX 11 Spectral De-noise settings for the second preprocessing pass, which produced the final-denoised recordings used for feature computation.

Parameter	Corpus setting
Module	iZotope RX 11 Spectral De-noise
Preset identifier	<code>Birdsong_lowfloor_2ndpass_lighter</code>
Noise-profile acquisition	Manual Learn from a selected noise-only interval
Adaptive mode	Disabled
Adaptive learn time	Inactive; 2.5 s displayed while adaptive mode was disabled
Threshold	3.4 dB
Reduction	8.7 dB
Frequency-dependent reduction curve	Enabled; custom second-pass curve (Fig. S5B)
Quality	D* (Best)
Artifact control	3.8
Main smoothing	1.4
Algorithm	Advanced
Algorithm-behavior smoothing	1.5
Multi-resolution processing	Enabled
FFT size	50 ms displayed; inactive in multi-resolution mode
Noise-floor synthesis	6.1
Noise-floor enhancement	3.3
Noise-floor masking	1.0
Noise-floor whitening	2.0
Dynamics knee	1.5
Dynamics release	80 ms
Output noise only	Disabled

Notes. Parameters without explicit physical units are reported on the native unitless scale displayed by iZotope RX. A noise profile was learned separately for each recording and was therefore not a fixed corpus parameter; all controls and frequency-dependent reduction curves listed above were held constant. The curves are documented graphically in Fig. S5; their control-point coordinates are not reported numerically by the interface.

S3 PCA neighborhood sensitivity analyses

The principal neighborhood analyses used $k = 25$ and preserved the empirical composition of the field corpus, in which observation areas contributed unequal numbers of indexed strophes. Two sensitivity analyses were therefore performed to evaluate the dependence of neighborhood measurements on (i) neighborhood size and (ii) unequal observation-area contribution. All analyses used the retained globally fitted PCA-1000 coordinate spaces, and candidates originating from the focal strophe’s own raw recording were excluded throughout.

Neighborhood measurements were repeated at $k = 10, 25, 50$, and 100 for both MFCC and

chroma representations under Euclidean and cosine distance. Increasing k generally broadened neighborhood composition, with mean same-area fractions tending to decrease and normalized neighborhood entropy to increase (Supplementary Tables S5 and S6). Despite these changes in absolute neighborhood composition, the directional off-diagonal enrichment structure remained highly stable relative to the principal $k = 25$ analysis. Spearman correlations across the 90 directional cross-area enrichment relationships ranged from $\rho = 0.901$ to 0.994 across the alternative neighborhood sizes and four representation–metric combinations (Supplementary Table S7). Stability was highest for Euclidean chroma ($\rho = 0.988$ – 0.994), while the lowest correspondence was observed for cosine chroma at $k = 100$ ($\rho = 0.901$). The principal ordering of cross-area enrichment relationships was therefore only weakly dependent on the choice of $k = 25$ within the range examined.

Sensitivity to unequal observation-area contribution was evaluated using a balanced subset obtained by sampling 500 strophes without replacement from each of the ten observation areas ($n = 5,000$; seed 42). The same sampled strophes were used in all four representation–metric combinations. Their coordinates were retained from the globally fitted PCA-1000 spaces and the PCA models were not refitted, so that the comparison evaluated neighborhood organization under equal-contribution subsampling within the established acoustic coordinate systems. Same-recording exclusion was retained.

At $k = 25$, equal-contribution subsampling reduced the corpus-wide mean same-area fraction in all four analytical spaces. MFCC same-area fraction decreased from 18.58% to 12.92% under Euclidean distance and from 25.42% to 15.25% under cosine distance. The corresponding chroma values decreased from 14.38% to 9.82% and from 20.13% to 10.94%, respectively. Normalized neighborhood entropy responded differently. It decreased from 0.782 to 0.747 for Euclidean MFCC, but increased from 0.715 to 0.844 for cosine MFCC, from 0.822 to 0.849 for Euclidean chroma, and from 0.760 to 0.888 for cosine chroma (Supplementary Table S4). Same-area fraction and entropy therefore responded differently to corpus equalization, consistent with their characterization of distinct aspects of local neighborhood composition.

Directional enrichment structure nevertheless remained substantially concordant between the complete and balanced corpora. Spearman correlations across the 90 off-diagonal directional area pairs were $\rho = 0.877$ for Euclidean MFCC, 0.773 for cosine MFCC, 0.969 for Euclidean chroma, and 0.724 for cosine chroma (Supplementary Table S4). Thus, equal-contribution subsampling altered absolute neighborhood concentration and evenness more strongly than it altered the rank ordering of cross-area enrichment relationships, although the degree of correspondence varied among acoustic representations and distance metrics.

Taken together, these analyses indicate that the detailed numerical composition of local neighborhoods was sensitive to both corpus composition and neighborhood scale, whereas the broader directional enrichment structure was comparatively stable. The balanced subset is therefore used here as a sensitivity analysis of the empirical corpus rather than as a replacement analytical dataset.

Table S4: Effect of equal area contribution on $k = 25$ PCA neighborhood measurements.

Representation	Metric	Same-area fraction (%)		Normalized entropy		Enrichment concordance
		Full	Balanced	Full	Balanced	Spearman ρ
MFCC-40	Euclidean	18.58	12.92	0.782	0.747	0.877
MFCC-40	Cosine	25.42	15.25	0.715	0.844	0.773
Chroma-80	Euclidean	14.38	9.82	0.822	0.849	0.969
Chroma-80	Cosine	20.13	10.94	0.760	0.888	0.724

Notes. Full-corpus values are weighted across all 20,075 focal strophes. Balanced values use the same reproducible fixed-seed random sample of 500 strophes from each of the ten observation areas ($n = 5,000$; seed 42) in all four representation-metric combinations. Sampled strophes retain their coordinates from the globally fitted PCA-1000 spaces; the PCA models were not refitted. Same-area fraction is the mean percentage of the 25 eligible neighbors originating from the focal strophe’s observation area. Normalized entropy is the Shannon entropy of neighborhood area composition divided by $\log(10)$. Enrichment-structure ρ is the Spearman correlation between the full and balanced analyses across the 90 off-diagonal directional area-pair enrichment values. Candidates originating from the focal strophe’s raw recording were excluded throughout.

Table S5: Mean same-area neighborhood fraction across neighborhood sizes in the retained PCA spaces.

Area	Euclidean				Cosine			
	$k = 10$	$k = 25$	$k = 50$	$k = 100$	$k = 10$	$k = 25$	$k = 50$	$k = 100$
<i>A. MFCC-40</i>								
1	19.5	15.5	13.7	12.3	27.5	21.4	18.7	16.8
2	34.9	23.7	18.3	14.9	54.5	42.5	35.5	30.7
3	29.3	24.2	22.3	21.4	43.4	34.2	29.7	26.8
4	22.0	20.8	19.9	18.7	8.9	7.3	6.5	6.2
5	24.8	20.3	18.5	17.6	31.8	24.4	20.5	18.2
6	16.0	12.0	10.5	9.8	25.3	18.4	15.3	13.4
7	26.2	22.2	19.2	16.2	15.0	12.5	11.1	10.0
8	18.1	12.6	9.9	8.2	26.5	18.6	14.9	12.7
9	7.2	6.5	6.1	5.5	8.8	7.9	7.3	6.4
10	2.8	2.1	1.9	1.9	5.8	4.2	3.5	3.0
<i>Corpus</i>	23.9	18.6	16.2	14.8	33.0	25.4	21.6	19.0
<i>B. Chroma-80</i>								
1	21.4	19.7	19.1	18.7	19.3	15.3	13.5	12.3
2	27.0	21.0	19.0	18.1	40.1	30.7	25.8	22.5
3	20.1	15.7	14.0	12.8	35.4	28.4	25.3	23.2
4	7.1	7.0	7.3	7.4	8.3	7.1	6.4	6.1
5	21.3	18.1	16.7	15.9	27.2	21.7	18.9	17.0
6	11.7	9.4	8.7	8.6	21.0	15.7	13.5	12.1
7	8.4	7.0	6.0	5.4	5.6	4.2	3.7	3.3
8	10.8	8.6	7.6	6.7	20.2	15.5	13.1	11.6
9	6.1	5.6	5.5	4.9	3.5	2.9	2.5	2.4
10	1.5	1.1	1.0	1.0	3.8	3.2	2.8	2.6
<i>Corpus</i>	17.5	14.4	13.2	12.5	25.6	20.1	17.5	15.7

Notes. Values are percentages and represent the mean, across focal strophes from each observation area, of the k nearest eligible neighbors originating from the same area. The corpus row is weighted across all 20,075 focal strophes. All analyses used the retained PCA-1000 coordinates; candidates originating from the focal strophe's raw recording were excluded before neighbor selection.

Table S6: Mean normalized neighborhood entropy across neighborhood sizes in the retained PCA spaces.

Area	Euclidean				Cosine			
	$k = 10$	$k = 25$	$k = 50$	$k = 100$	$k = 10$	$k = 25$	$k = 50$	$k = 100$
<i>A. MFCC-40</i>								
1	0.641	0.792	0.852	0.888	0.559	0.715	0.793	0.839
2	0.581	0.770	0.848	0.888	0.438	0.635	0.741	0.805
3	0.609	0.774	0.844	0.882	0.508	0.692	0.782	0.834
4	0.636	0.774	0.834	0.874	0.627	0.772	0.834	0.868
5	0.627	0.779	0.846	0.884	0.567	0.731	0.809	0.854
6	0.660	0.804	0.862	0.894	0.594	0.749	0.817	0.857
7	0.636	0.767	0.828	0.865	0.663	0.807	0.865	0.897
8	0.631	0.784	0.846	0.880	0.565	0.736	0.813	0.855
9	0.678	0.818	0.874	0.902	0.659	0.787	0.841	0.870
10	0.678	0.806	0.862	0.893	0.632	0.770	0.828	0.863
<i>Corpus</i>	0.625	0.782	0.848	0.885	0.546	0.715	0.796	0.843
<i>B. Chroma-80</i>								
1	0.675	0.816	0.870	0.896	0.617	0.765	0.830	0.868
2	0.647	0.808	0.864	0.892	0.551	0.735	0.816	0.861
3	0.670	0.819	0.874	0.898	0.575	0.744	0.818	0.861
4	0.696	0.826	0.875	0.897	0.656	0.790	0.844	0.874
5	0.692	0.829	0.879	0.901	0.609	0.762	0.830	0.868
6	0.696	0.833	0.880	0.900	0.628	0.777	0.840	0.874
7	0.702	0.830	0.875	0.898	0.658	0.788	0.846	0.875
8	0.692	0.825	0.876	0.899	0.620	0.765	0.830	0.867
9	0.702	0.832	0.877	0.898	0.672	0.794	0.851	0.881
10	0.703	0.834	0.879	0.900	0.650	0.782	0.840	0.872
<i>Corpus</i>	0.680	0.822	0.874	0.898	0.604	0.760	0.828	0.867

Notes. Values are means across focal strophes from each observation area. Shannon entropy of neighborhood area composition was normalized by $\log(10)$, producing values from 0 (one area only) to 1 (maximally even representation of all ten areas). The corpus row is weighted across all 20,075 focal strophes. Candidates originating from the focal strophe's raw recording were excluded.

Table S7: Stability of off-diagonal neighborhood-enrichment structure across neighborhood sizes.

Representation	Metric	$k = 10$ vs. 25	$k = 50$ vs. 25	$k = 100$ vs. 25
MFCC-40	Euclidean	0.983	0.989	0.956
MFCC-40	Cosine	0.968	0.977	0.942
Chroma-80	Euclidean	0.988	0.994	0.991
Chroma-80	Cosine	0.948	0.968	0.901

Notes. Values are Spearman rank correlations calculated across the 90 off-diagonal directional area-pair enrichment values, using the principal $k = 25$ enrichment matrix as the reference. All analyses used the complete 20,075-strophe corpus in the retained PCA-1000 coordinates and excluded candidates originating from the focal strophe’s raw recording. The diagonal self-enrichment values were omitted so that the coefficients describe the persistence of cross-area relational structure.

S4 Field-metadata overlays on the MFCC manifold

S4.1 Linking field metadata and acoustic geometry

The field and acoustic components of the corpus are connected through the recording event. A recording event is defined here as one retained field recording, identified by its unique raw recording identifier and associated in the QGIS dataset with a geographic position, date, time, observation-area attribution and contextual scores. On the field map, each recording event is represented by a single point. Following preprocessing and indexing, however, the same recording becomes a group containing a variable number of individual strophes. The relationship between the two representations is therefore one-to-many: one geographically located recording point corresponds to a group of strophes distributed through the acoustic embedding.

Each point in the embedding represents one indexed strophe rather than one complete recording. Every strophe retains the identifier of its source recording and inherits the recording-level metadata associated with that event. When a metadata layer is applied, all strophes originating from the same recording receive the same colour, although their positions in the embedding are determined independently by their acoustic representations. A single point on the map consequently expands into a potentially dispersed point cloud in acoustic space. This distribution provides a qualitative view of the internal acoustic diversity contained within the recording and of its overlap with strophes originating from other recording events.

Figure S7 illustrates this correspondence using the cosine-metric MFCC UMAP embedding as a representative acoustic space. The coordinates of all 20,075 strophes remain identical across the four panels; only the metadata used for colouring changes. Panel A distinguishes recordings belonging to the northern and southern observation-area clusters. Panel B applies the season day associated with each recording event. Panels C and D apply the vocal-intensity and interaction-presence scores recorded in the field dataset. In each case, the accompanying map shows where the recording-level attribute originated, while the embedding shows how the corresponding groups of strophes are distributed across the shared acoustic geometry.

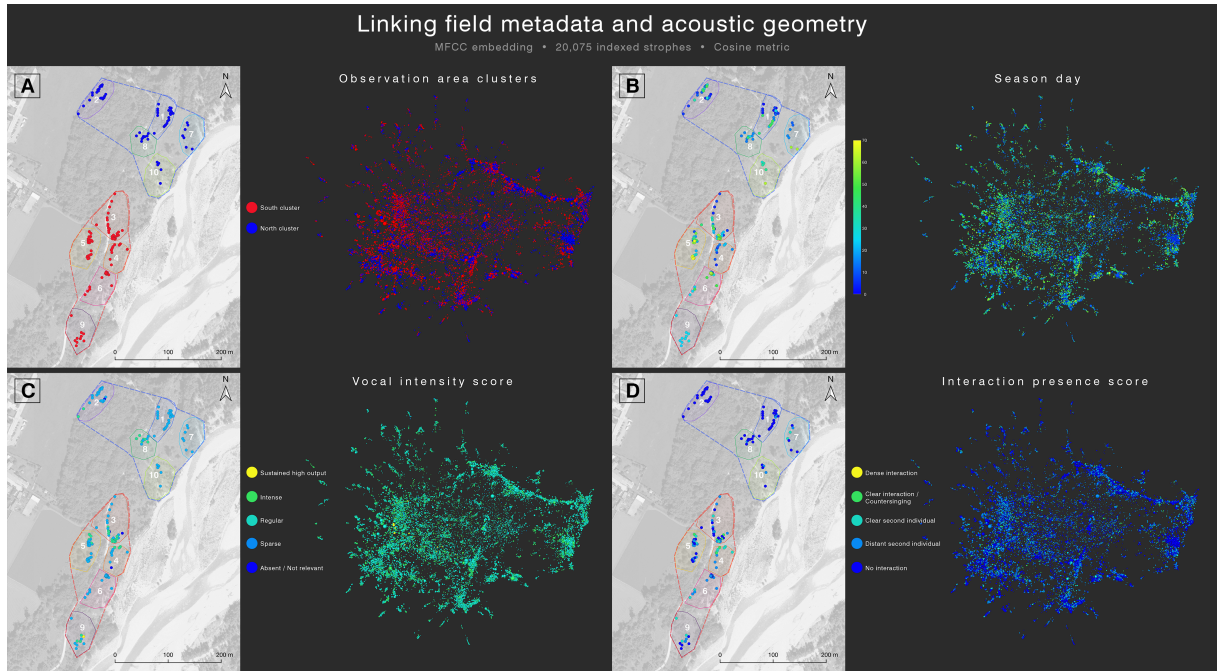


Figure S7: Linking field metadata and acoustic geometry through a representative MFCC embedding. Each panel juxtaposes the spatial distribution of recordings in the QGIS field dataset with the corresponding metadata colouring of the 20,075 indexed strophes in the cosine-metric MFCC UMAP embedding: **(A)** northern and southern observation-area clusters, **(B)** season day, **(C)** vocal-intensity score, and **(D)** interaction-presence score. All embedding panels show the same strophes at identical coordinates; only the metadata colouring changes. Season day, vocal intensity and interaction presence are recording-level attributes inherited by each strophe from its source recording rather than measurements made independently at strophe level. The overlays provide a descriptive link between field context and acoustic organization and identify metadata dimensions available for future targeted analyses; they are not presented here as inferential tests of their effects on acoustic geometry.

These metadata were not used to construct the PCA or UMAP spaces and therefore exerted no influence on the positions of the strophes. They were joined to the completed embedding afterwards through the source-recording identifier. Vocal intensity and interaction presence should likewise be understood as contextual annotations of the recording event, not as measurements independently assigned to every constituent strophe. The colour-coded embeddings are consequently descriptive rather than inferential. Their visual composition is also affected by differences in the number of retained recordings, the number of strophes contributed by each recording and the frequency of the metadata categories represented in the corpus.

The integration nevertheless establishes a direct and auditable link between field observation and acoustic organization. It allows any region of the embedding to be traced back from individual strophes to their source recordings, geographic locations and contextual metadata. In the present study, the overlays provide an informative overview of the field variables already connected to the acoustic corpus. In subsequent work, the same structure can support targeted analyses of temporal change, spatial relationships, vocal intensity, interaction context and other recording-level attributes, without requiring the acoustic space or the underlying corpus to be reconstructed.

S5 Supplementary data package and audiovisual materials

The supplementary package deposited with this study provides the machine-readable materials underlying corpus documentation, acoustic-space construction, quantitative PCA analyses, denoising sensitivity checks, and representative acoustic inspection. It also contains nine videos documenting dynamic and interactive components of the framework that cannot be represented

fully through static figures. The complete package is archived at <https://doi.org/10.6084/m9.figshare.33497311>.

The deposited release corresponds to the analytical corpus reported in the manuscript: 20,075 indexed strophes from 161 retained recordings assigned to ten operational observation areas. Stable recording and strophe identifiers preserve links among the recording-level master table, the representative audio subset, the embedding tables, the quantitative outputs, and the source information displayed in the supplementary videos. Each principal data directory contains a plain-text README describing its files and relevant field definitions.

S5.1 Deposited data package

The repository contents are summarized in Table S8. The package exposes the complete machine-readable geometry and quantitative outputs used in the study, together with a representative 500-strophe audio subset. The complete source-recording and indexed-audio corpus is not included in this release; its access conditions are stated in the Data and code availability section of the manuscript.

Table S8: Contents of the deposited supplementary data and audiovisual package. Detailed schemas, file relationships, and analytical conditions are documented in the README supplied with each principal directory.

Component	Deposited contents	Scope and function
Recording-level master table	One CSV containing all 227 field recordings	Spatial and temporal metadata, corpus-retention status, raw and indexed durations, annotation scores, strophe and technical-delimiter counts, and mapped singing locations
Representative audio subset	Ten area-level WAV files, corresponding reconstructed CSV indices, a combined 500-strophe index, and selection manifests	Fifty strophes from each observation area, selected with approximately balanced representation of the 161 retained recordings. This subset supports direct inspection of the acoustic material but is not a substitute for the complete corpus.
PCA and UMAP embedding data	Twenty-four strophe-level CSV files: 12 operational blocks for MFCC and 12 for microtonal chroma	Source and indexing provenance, acoustic diagnostic descriptors, recording metadata, three-dimensional UMAP coordinates, and the 1,000 retained PCA coordinates for every indexed strophe.
PCA-derived quantitative analyses	Dispersion and centroid outputs, pairwise centroid distances, neighborhood summaries and matrices, and balanced-sampling sensitivity outputs	Machine-readable results for the MFCC and chroma PCA-1000 spaces, including Euclidean and cosine neighborhood analyses at $k = 10, 25, 50$, and 100 and the fixed equal-contribution subset of 5,000 strophes.
Denosing sensitivity outputs	Per-recording results, quality-control files, and four-row summaries for each of the ten observation areas, together with the corpus-level recap	Comparison of first-pass floor-stripped and final-denoised recordings for MFCC-40 and L1-normalized Chroma-80 at the 35% and 75% baseline-RMS gates. The corpus recap was calculated from the pooled recording-level values rather than by averaging area summaries.
Supplementary videos	Nine audiovisual demonstrations, V1–V9	Dynamic terrain inspection, PCA reconstruction, embedding navigation, metadata and descriptor overlays, sequential trajectories, pointwise retrieval, and lasso-based neighborhood retrieval.

The recording-level master table supplies the common reference for field metadata through the

original recording identifier. Strophe-level identifiers and source-recording fields provide the corresponding links among the audio sample, embedding tables, and quantitative outputs. The PCA and UMAP files are distributed across operational blocks for storage and interaction, but the coordinates belong to two separately fitted corpus-wide spaces. Quantitative PCA outputs should therefore be interpreted according to their stated representation, metric, sampling condition, and neighborhood size rather than as independent block-level embeddings.

S5.2 Operational corpus blocks in the deposited embeddings and interfaces

For processing and interactive handling, the retained corpus was organized into 12 operational blocks corresponding to the ten observation areas. Eight observation areas were represented by a single block, whereas Areas 2 and 3 were each subdivided into two blocks. Area 2 was separated into diel and nocturnal recordings, while Area 3 was divided into earlier and later recording subsets.

These subdivisions were used for data organization, filtering, trajectory inspection, and interactive retrieval only. They did not define additional observation areas or alter the contribution of individual strophes to the shared PCA and UMAP representations. All 20,075 retained strophes entered the corpus-wide transformations jointly, with each strophe represented once. The 12-block organization should therefore be understood as an operational partition of the same ten-area corpus rather than as a distinct biological or inferential grouping.

Table S9: Operational corpus blocks used for processing, data deposition, and interactive inspection. Areas 2 and 3 were subdivided into two blocks, while all other observation areas were represented by a single block. These subdivisions did not constitute additional observation areas or modify the contribution of individual strophes to the corpus-wide PCA or UMAP representations.

Operational block	Recordings	Indexed strophes	Composition
Area 1	21	2,232	All retained recordings
Area 2 diel	13	1,313	Diel recordings
Area 2 night	7	1,905	Nocturnal recordings
Area 3 early	20	1,701	Earlier recording subset
Area 3 late	11	2,738	Later recording subset
Area 4	11	1,238	All retained recordings
Area 5	24	2,806	All retained recordings
Area 6	22	2,124	All retained recordings
Area 7	7	566	All retained recordings
Area 8	10	2,339	All retained recordings
Area 9	10	504	All retained recordings
Area 10	5	609	All retained recordings
Total	161	20,075	

The interactive manifold interfaces display these blocks as independently selectable corpus components. Selection or exclusion of a block changes only the material displayed or retrieved during inspection and does not reconstruct the underlying shared acoustic space.

S5.3 Supplementary audiovisual materials

The supplementary videos document visual and interactive components of the analytical framework performed with custom-built Touchdesigner interfaces. They provide dynamic inspection of the whole-strophe acoustic terrains, their reconstruction across PCA dimensionalities, the

corpus-wide low-dimensional embeddings, post hoc visualization layers, recording-level sequential trajectories, and the retrieval of individual strophes or local acoustic neighborhoods.

The videos are methodological and inspection materials rather than additional quantitative analyses. Three-dimensional UMAP is used as a navigational and retrieval interface, whereas the quantitative measurements reported in the study are calculated in the retained PCA representations. Unless otherwise specified, filtering, colouring, trajectory display, and interactive selection operate on already constructed coordinates and do not refit or modify the PCA or UMAP spaces.

S5.3.1 Acoustic-terrain inspection and PCA reconstruction

Videos V1–V3 document the high-dimensional whole-strophe representations from which the relational spaces were constructed. Video V1 shows successive MFCC terrains during synchronized playback of an example source recording, allowing the evolving spectral-envelope organization of complete vocal events to be inspected. Video V2 provides the corresponding demonstration for the 80-bin microtonal-chroma terrains and their evolving relative-frequency organization.

Video V3 demonstrates inverse reconstruction of a microtonal-chroma terrain across progressively increasing PCA dimensionalities. During synchronized playback and strophe switching, the progressive recovery of broad morphology, localized structure, and finer acoustic detail can be inspected dynamically. This recording-specific reconstruction corresponds to the illustrative PCA analysis described in the main text and is not a separate corpus-wide quantitative analysis.

S5.3.2 Embedding geometry, overlays, trajectories, and source retrieval

Videos V4–V9 progress from global views of the MFCC and microtonal-chroma embeddings to contextual overlays, restoration of recording-level sequence order, and direct retrieval of indexed acoustic material.

Video V4 presents the complete three-dimensional embedding geometries from multiple orientations. All 20,075 strophes are retained, with each point representing one strophe and point size scaled to original strophe duration.

Video V5 applies recording-level and QGIS-derived metadata as post hoc colour layers on the fixed coordinates.

Video V6 applies derived acoustic descriptors in the same manner. Neither class of overlay enters the PCA or UMAP transformations.

Video V7 restores recording-level strophe order by connecting successive events within selected source recordings. The displayed sequence window can be adjusted interactively to inspect how successive vocal events occupy the embedding. Connecting segments indicate indexed temporal succession and are not interpreted as metric paths through UMAP space.

Video V8 demonstrates pointwise inspection through the probe-click interface. Movement across a saved geometry view highlights corresponding strophes, while clicking a point retrieves the associated vocal event for playback and spectrographic inspection together with its source metadata.

Video V9 demonstrates free-form lasso selection of a local region, export of the selected strophe set, and generation of a stitched audio sequence with an associated source-information CSV.

These operations provide a direct return path from corpus-scale geometry to individual sounds and acoustically localized sets without requiring prior categorical assignment.

Table S10: Overview of the nine supplementary videos deposited with the data package.

All Supplementary Videos V1–V9 are deposited in the accompanying Figshare repository (DOI: 10.6084/m9.figshare.33497311). The identifiers V1–V9 correspond directly to the deposited video files.

ID	Title	Principal function
V1	Real-time MFCC terrain inspection during recording playback	Displays successive MFCC acoustic terrains in synchrony with the indexed source recording.
V2	Real-time microtonal-chroma terrain inspection during recording playback	Displays successive high-resolution microtonal-chroma terrains during synchronized recording playback.
V3	Real-time PCA reconstruction of microtonal-chroma terrains	Demonstrates progressive inverse reconstruction of terrain morphology across tested PCA dimensionalities.
V4	Global views of the MFCC and microtonal-chroma embeddings	Provides dynamic inspection of the complete corpus-wide geometries and their principal low-dimensional visual differences.
V5	Field-metadata overlays on the acoustic embeddings	Links QGIS-derived and recording-level metadata to fixed embedding coordinates for post hoc visual inspection.
V6	Acoustic-descriptor overlays on the acoustic embeddings	Shows the distribution of selected derived acoustic descriptors across the completed geometries.
V7	Recording-wise strophe trajectories and sequential inspection	Restores recording-level strophe order and demonstrates interactive adjustment and progressive inspection of sequence windows.
V8	Pointwise strophe inspection using the probe-click interface	Demonstrates real-time highlighting, pointwise retrieval, spectrographic inspection, source metadata, and audio playback.
V9	Lasso selection and retrieval of local acoustic neighborhoods	Selects groups of strophes, exports the selected set, and generates stitched audio with corresponding source-information outputs.

The identifiers below correspond to the video filenames in the repository. The data tables and videos together provide complementary routes through the same corpus: the machine-readable files expose the deposited coordinates, metadata, and quantitative outputs, while the audiovisual materials demonstrate how those structures can be inspected and returned to their source-linked acoustic events.

S6 Supplementary tables

Supplementary Table S11A. Recording inventory and analytical contribution

Recording ID	Area	Date	Season day	Time	Duration	Included strophes	Mergers	Spot
14080004	1	18/04	1	10:34	02:57	Y	17	0
14080006	1	18/04	1	10:46	08:31	N	—	1
14080008	1	18/04	1	11:02	25:42	Y	92	0
14080010	1	18/04	1	11:46	05:18	Y	61	0
14130001	1	21/04	4	10:58	15:47	Y	75	0
14130002	1	21/04	4	11:14	14:43	N	—	—
14130004	2	21/04	4	11:48	08:54	Y	40	0
14140001	1	23/04	6	10:32	31:35	Y	342	58
14140002	1	23/04	6	11:03	04:53	Y	40	0
14140003	2	23/04	6	11:12	22:38	Y	240	2
14140005	3	23/04	6	11:43	13:12	N	—	13
14150001	2	24/04	7	10:23	04:48	Y	29	0
14150002	1	24/04	7	10:29	01:26	N	—	1
14150003	1	24/04	7	10:32	11:12	Y	105	2
14150005	4	24/04	7	10:52	11:23	Y	81	2
14150008	3	24/04	7	11:11	04:26	Y	36	0
14150010	5	24/04	7	11:20	09:23	Y	83	1
14150011	3	24/04	7	11:31	05:11	Y	30	0
14150012	4	24/04	7	11:37	03:16	Y	15	0
14150013	3	24/04	7	11:47	08:56	Y	84	3
14150015	1	24/04	7	12:01	14:23	Y	109	1
14160001	1	25/04	8	10:39	12:07	Y	113	1
14160004	1	25/04	8	11:18	01:06	N	—	1
14160006	3	25/04	8	11:22	06:53	Y	42	2
14160007	3	25/04	8	11:30	03:47	N	—	13
14160008	3	25/04	8	11:36	17:12	Y	52	0
14160010	1	25/04	8	11:58	02:05	N	—	1
14160011	2	25/04	8	22:38	44:47	Y	501	0
14160012	2	25/04	8	23:23	19:24	Y	189	0
14160013	2	25/04	8	23:42	27:04	Y	106	0
14170005	3	26/04	9	11:11	05:00	Y	38	3
14170006	4	26/04	9	11:17	02:26	Y	12	0
14170008	6	26/04	9	11:23	04:32	Y	19	1
14170009	3	26/04	9	11:30	06:36	Y	35	0
14170010	1	26/04	9	11:39	01:57	N	—	1
14190001	3	27/04	10	10:03	07:03	N	—	14
14190002	4	27/04	10	10:11	18:05	Y	187	4
14190004	6	27/04	10	10:33	16:06	N	—	21
14190006	6	27/04	10	10:50	07:19	Y	33	0
14190007	4	27/04	10	10:58	25:14	Y	224	9
14190008	3	27/04	10	11:30	06:22	N	—	—
14190009	3	27/04	10	11:37	03:06	N	—	—
14190010	5	27/04	10	11:40	08:53	Y	141	1
14200003	1	28/04	11	10:22	11:44	Y	96	0
14200004	7	28/04	11	10:34	18:08	Y	111	3
14200005	7	28/04	11	10:54	17:55	Y	111	1
14200006	1	28/04	11	11:12	09:22	Y	89	0
14200009	3	28/04	11	11:25	05:47	N	—	13
14200010	4	28/04	11	11:31	07:48	Y	47	1
14200011	4	28/04	11	11:39	05:02	N	—	17
14210001	2	28/04	11	22:16	15:38	Y	64	0
14210002	2	28/04	11	22:32	32:26	Y	275	0
14210003	2	28/04	11	23:05	23:29	Y	260	0
14210004	2	28/04	11	23:28	47:45	Y	510	0
14230003	7	29/04	12	12:01	14:51	N	—	9
14230004	2	29/04	12	14:54	13:08	Y	60	0
14230005	2	29/04	12	15:07	18:05	Y	133	2

Recording ID	Area	Date	Season day	Time	Duration	Included strophes	Mergers	Spot
14240002	8	01/05	14	09:51	33:05	Y	502	123
14240005	5	01/05	14	10:29	01:12	N	—	19
14240007	3	01/05	14	10:48	17:59	N	—	13
14240009	4	01/05	14	11:16	06:19	N	—	17
14240010	4	01/05	14	11:26	12:47	Y	194	10
14240012	8	01/05	14	11:49	03:10	Y	26	3
14240017	1	01/05	14	16:46	02:17	N	—	1
14240022	8	01/05	14	17:11	29:08	Y	416	87
14250004	7	02/05	15	10:18	15:50	Y	138	3
14250006	1	02/05	15	10:40	07:36	Y	82	2
14250007	7	02/05	15	10:49	03:06	Y	18	1
14250009	8	02/05	15	10:54	02:57	Y	30	5
14250010	8	02/05	15	11:04	12:15	Y	162	28
14250011	7	02/05	15	11:54	15:22	Y	111	0
14250013	8	02/05	15	16:28	06:42	Y	75	13
14250014	1	02/05	15	16:35	05:51	Y	73	2
14250016	8	02/05	15	16:41	01:54	N	—	8
14250018	4	02/05	15	17:03	23:15	N	—	17
14250021	4	02/05	15	17:03	07:01	Y	110	8
14250019	3	02/05	15	17:27	04:10	Y	54	10
14250024	3	02/05	15	17:52	09:47	N	—	13
14250025	8	02/05	15	18:04	03:03	Y	27	2
14260005	5	03/05	16	10:18	12:18	Y	197	4
14260006	3	03/05	16	10:31	04:15	N	—	13
14260007	1	03/05	16	10:40	03:25	N	—	1
14280003	5	04/05	17	10:08	08:15	Y	83	0
14280004	4	04/05	17	10:18	07:38	N	—	17
14280007	4	04/05	17	11:20	16:35	Y	209	8
14280009	3	04/05	17	11:44	05:18	N	—	14
14280011	5	04/05	17	11:51	01:51	Y	27	0
14300001	6	06/05	19	10:28	04:52	Y	35	2
14300002	3	06/05	19	10:37	09:09	Y	122	9
14300003	5	06/05	19	10:48	07:38	Y	78	10
14300004	3	06/05	19	11:00	05:45	N	—	—
14300006	3	06/05	19	11:23	04:18	Y	24	0
14310001	8	07/05	20	10:56	10:27	Y	145	28
14310002	1	07/05	20	11:07	05:11	Y	44	2
14310004	3	07/05	20	11:22	12:19	Y	142	7
14310005	4	07/05	20	11:35	06:36	Y	92	3
14310006	4	07/05	20	11:42	04:30	Y	67	0
14310007	3	07/05	20	11:48	07:49	Y	130	20
14310008	5	07/05	20	11:56	05:49	Y	82	1
14310009	3	07/05	20	12:04	11:56	Y	192	21
14310011	5	07/05	20	12:22	10:55	N	—	19
14310013	1	07/05	20	12:43	13:41	N	—	1
14320003	6	08/05	21	10:43	11:58	Y	130	2
14320004	3	08/05	21	10:59	10:08	Y	145	34
14320008	9	08/05	21	15:37	03:00	N	—	23
14320009	9	08/05	21	15:41	02:45	Y	22	1
14320010	9	08/05	21	15:45	03:06	Y	30	0
14320011	9	08/05	21	15:52	04:41	Y	34	0
14320012	9	08/05	21	16:05	01:20	N	—	23
14320013	6	08/05	21	16:11	06:57	Y	76	6
14330001	3	09/05	22	10:26	09:08	Y	143	14
14330005	6	09/05	22	11:12	05:53	Y	35	1
14330011	9	09/05	22	11:16	04:00	Y	52	7
14330006	6	09/05	22	11:18	02:09	N	—	21

Supplementary Table S11A (continued)

Recording ID	Area	Date	Season day	Time	Duration	Included strophes	Mergers	Spot
14330010	9	09/05	22	11:31	04:41	Y	48	11 24
14330012	9	09/05	22	11:42	04:34	Y	50	5 24
14330013	9	09/05	22	12:00	02:15	N	—	— 24
14330014	9	09/05	22	12:02	01:17	N	—	— 23
14330015	9	09/05	22	12:06	08:49	Y	115	12 23
14330018	9	09/05	22	12:21	09:00	Y	80	4 24
14330019	6	09/05	22	12:32	04:00	Y	16	0 22
14340001	8	11/05	24	10:27	15:31	Y	188	36 6
14340002	3	11/05	24	10:44	12:48	Y	182	21 13
14340003	9	11/05	24	11:02	04:24	Y	27	0 23
14340004	6	11/05	24	11:07	05:03	Y	40	0 22
14340011	1	11/05	24	12:03	04:14	Y	37	6 1
14350001	9	12/05	25	10:44	10:17	N	—	— 23
14350005	5	12/05	25	11:21	05:00	Y	46	1 20
14350006	6	12/05	25	11:28	02:24	Y	12	0 21
14350007	4	12/05	25	14:37	02:57	N	—	— 18
14350009	9	12/05	25	15:00	06:31	Y	46	1 23
14360001	2	13/05	26	10:27	04:01	N	—	— 3
14360002	5	13/05	26	10:37	07:09	N	—	— 20
14360004	3	13/05	26	10:53	11:48	N	—	— —
14370001	2	14/05	27	10:38	09:16	Y	80	0 3
14370002	10	14/05	27	10:53	07:15	Y	71	8 11
14380001	2	15/05	28	10:52	08:04	Y	53	0 3
14380003	5	15/05	28	11:18	08:05	Y	84	3 20
14380005	2	15/05	28	11:37	03:01	Y	21	0 4
14390001	2	17/05	30	10:14	03:32	Y	22	0 3
14390002	3	17/05	30	10:36	05:15	Y	40	2 14
14390003	3	17/05	30	10:41	06:09	Y	38	3 14
14410002	10	20/05	33	11:05	30:04	Y	197	3 11
14410003	8	20/05	33	11:36	05:52	N	—	— 6
14410005	2	20/05	33	12:00	08:48	Y	76	0 4
14420001	10	21/05	34	10:33	13:45	Y	28	0 11
14430001	1	22/05	35	10:56	10:25	Y	53	0 1
14430002	5	22/05	35	11:12	08:42	Y	69	0 19
14430003	1	22/05	35	11:25	07:22	N	—	— 2
14430004	5	22/05	35	14:39	18:30	Y	212	1 19
14440001	2	23/05	36	09:48	09:39	N	—	— 3
14440002	10	23/05	36	10:15	29:29	Y	207	5 11
14440004	2	23/05	36	11:34	14:24	Y	132	0 3
14450010	5	26/05	39	14:38	27:48	Y	286	0 19
14460001	3	27/05	40	09:07	11:37	Y	64	0 13
14460002	5	27/05	40	09:20	12:37	N	—	— 19
14460004	1	27/05	40	10:00	16:24	Y	171	4 2
14460005	3	27/05	40	10:21	09:29	Y	108	1 16
14460006	8	27/05	40	15:50	53:06	Y	768	89 6
14470001	1	28/05	41	15:05	02:44	Y	23	1 1
14470002	1	28/05	41	15:10	23:25	Y	275	4 1
14470003	3	28/05	41	15:38	03:20	N	—	— 16
14470004	3	28/05	41	15:46	03:16	N	—	— 13
14470005	5	28/05	41	15:51	08:26	N	—	— 19
14490003	1	29/05	42	10:49	08:37	N	—	— 2
14510001	2	31/05	44	09:36	33:15	Y	381	3 3
14510003	6	31/05	44	10:41	09:16	Y	76	2 21
14510004	3	31/05	44	10:52	27:21	Y	357	23 13
14510005	6	31/05	44	11:21	10:51	N	—	— 21
14510006	5	31/05	44	11:35	05:51	Y	70	0 19
14510007	3	31/05	44	11:42	24:32	Y	404	76 13

Recording ID	Area	Date	Season day	Time	Duration	Included strophes	Mergers	Spot
14520001	1	01/06	45	9:55	26:20	Y	309	8 1
14520002	5	01/06	45	10:34	08:40	N	—	— 19
14520003	9	01/06	45	10:47	02:00	N	—	— 24
14520004	1	01/06	45	15:39	02:48	Y	26	0 1
14520005	6	01/06	45	15:46	23:23	Y	268	25 22
14530001	2	03/06	47	09:34	04:37	Y	46	0 3
14530004	6	03/06	47	10:00	05:42	Y	52	4 21
14530005	6	03/06	47	16:18	58:23	Y	241	14 21
14540004	2	04/06	48	11:22	15:21	N	—	— 3
14550001	6	06/06	50	09:37	11:58	Y	93	0 21
14550003	3	06/06	50	16:13	32:26	Y	467	72 13
14550004	6	06/06	50	16:46	20:52	Y	210	12 21
14550006	3	06/06	50	17:14	13:44	N	—	— 13
14550007	3	06/06	50	17:28	12:02	Y	191	19 13
14560002	6	07/06	51	10:51	11:30	Y	95	0 21
14560005	3	07/06	51	15:05	22:01	Y	147	27 13
14560006	6	07/06	51	15:57	28:48	Y	275	12 21
14570001	3	08/06	52	09:57	17:39	Y	273	15 13
14570002	3	08/06	52	10:22	16:09	Y	228	25 13
14570003	6	08/06	52	11:12	08:47	Y	45	0 21
14570004	3	08/06	52	11:22	04:19	N	—	— 13
14570008	3	08/06	52	11:32	10:46	Y	149	11 13
14570009	6	08/06	52	15:45	09:12	N	—	— 21
14570010	3	08/06	52	15:56	16:01	Y	268	36 13
14580002	6	09/06	53	15:26	29:51	Y	73	5 21
14580003	6	09/06	53	15:57	23:31	N	—	— 21
14580005	5	09/06	53	16:26	12:31	Y	215	7 19
14580006	3	09/06	53	16:39	05:03	Y	74	5 13
14590001	3	10/06	54	09:55	06:42	N	—	— 13
14600002	5	11/06	55	09:44	16:26	Y	199	0 19
14600003	5	11/06	55	10:23	18:30	Y	193	0 19
14600005	5	11/06	55	11:25	07:09	N	—	— 20
14600008	6	11/06	55	15:22	09:31	Y	77	4 21
14600009	3	11/06	55	15:33	12:56	Y	180	23 13
14600011	5	11/06	55	15:39	08:01	Y	106	6 20
14610005	3	12/06	56	16:08	01:34	N	—	— 13
14610006	6	12/06	56	16:11	06:19	N	—	— 21
14610007	6	12/06	56	16:24	42:19	N	—	— 21
14620005	7	13/06	57	09:23	06:09	Y	39	1 10
14620006	5	13/06	57	09:46	05:19	N	—	— 20
14620008	6	13/06	57	15:20	07:28	Y	56	3 21
14630004	7	14/06	58	09:19	06:59	Y	38	0 10
14630007	5	14/06	58	09:55	05:30	N	—	— 20
14630008	3	14/06	58	10:41	04:40	N	—	— 13
14630009	5	14/06	58	10:49	10:15	N	—	— 19
14640002	10	15/06	59	09:30	15:06	Y	106	2 12
14640003	6	15/06	59	09:49	25:57	Y	167	1 21
14640005	5	15/06	59	10:20	04:50	Y	44	0 20
14650001	5	16/06	60	09:36	07:12	Y	93	0 20
14660001	10	17/06	61	10:11	04:39	N	—	— 12
14660002	5	17/06	61	10:24	22:08	N	—	— 20
14670001	5	20/06	64	10:33	04:01	Y	30	0 20
14680001	5	21/06	65	14:30	14:14	Y	148	0 20
14680002	5	21/06	65	14:45	17:44	Y	215	0 20
14690001	5	23/06	67	10:43	08:51	Y	76	0 20
14700003	5	24/06	68	10:18	04:39	Y	29	0 19

Notes. Strophe/merger counts apply only to retained recordings; durations include 1-s technical delimiters. Used segments and full metadata are provided in the master CSV.

Supplementary Table S11B. Recording-level annotation scores

Recording ID	Audio quality	Target clarity	Other birds	Human noise	Signal overlap	Vocal intensity	Vocal continuity	Interaction presence	Observer effect
14080004	3	3	1	0	0	2	2	0	0
14080006	3	3	1	0	0	4	4	0	0
14080008	3	3	3	0	0	2	3	0	0
14080010	3	4	2	0	0	2	3	0	0
14130001	3	4	1	0	1	2	2	0	0
14130002	3	3	4	0	4	2	3	2	0
14130004	3	3	1	1	0	2	2	0	0
14140001	3	3	1	0	1	2	3	0	0
14140002	3	4	0	1	0	2	3	0	0
14140003	3	3	2	0	0	2	3	0	0
14140005	0	4	3	1	1	3	4	2	0
14150001	3	3	1	1	1	1	2	0	1
14150002	3	3	0	2	0	2	3	0	0
14150003	3	4	1	0	0	2	3	0	0
14150005	3	3	0	0	0	2	2	2	1
14150008	3	4	1	0	1	2	2	2	1
14150010	3	2	1	0	1	2	2	2	0
14150011	3	3	2	0	2	1	2	2	1
14150012	3	3	1	1	1	1	2	1	1
14150013	3	3	1	0	1	2	2	0	0
14150015	3	3	1	0	1	2	2	0	0
14160001	3	3	1	1	1	2	3	0	0
14160004	3	4	1	0	1	2	3	0	0
14160006	3	3	0	0	0	2	2	0	0
14160007	3	2	0	1	0	1	1	1	1
14160008	3	3	1	0	0	1	2	1	1
14160010	3	2	2	0	2	1	1	0	2
14160011	4	4	0	0	0	3	3	0	0
14160012	4	4	0	0	0	3	2	0	3
14160013	4	4	0	0	0	2	2	0	4
14170005	3	2	1	0	1	2	3	1	0
14170006	3	2	0	0	0	2	2	1	1
14170008	3	3	0	0	0	2	2	1	1
14170009	3	2	1	1	1	2	2	1	0
14170010	3	3	1	0	1	2	2	0	1
14190001	2	2	2	0	2	2	2	2	0
14190002	3	3	0	0	0	2	3	1	0
14190004	3	3	3	0	2	2	2	1	0
14190006	3	2	2	0	2	2	2	2	0
14190007	3	3	1	0	0	2	3	1	0
14190008	3	2	1	1	2	2	2	2	0
14190009	2	2	2	0	2	2	2	0	1
14190010	3	3	2	0	2	3	3	3	0
14200003	3	3	1	0	1	2	3	1	0
14200004	2	2	1	0	1	2	2	0	1
14200005	2	2	0	0	0	2	3	1	0
14200006	2	3	1	0	1	2	3	1	0
14200009	3	2	3	0	4	3	2	4	0
14200010	3	3	1	0	1	2	2	2	0
14200011	2	2	3	0	3	3	2	3	0
14210001	4	4	0	0	0	2	2	0	3
14210002	4	4	0	0	0	2	3	0	2
14210003	4	4	0	0	0	2	3	0	0
14210004	4	4	0	0	0	2	3	0	0
14230003	2	2	2	0	3	1	2	1	0
14230004	2	3	3	0	3	2	3	0	0
14230005	3	3	1	0	2	2	2	0	0

Recording ID	Audio quality	Target clarity	Other birds	Human noise	Signal overlap	Vocal intensity	Vocal continuity	Interaction presence	Observer effect
14240002	3	3	1	0	2	3	3	1	0
14240005	3	2	1	0	2	2	2	1	0
14240007	2	3	3	0	3	3	3	2	0
14240009	2	3	3	0	3	3	2	2	0
14240010	3	3	2	0	3	3	3	3	0
14240012	3	3	0	1	0	3	3	0	0
14240017	3	3	0	0	2	2	2	0	0
14240022	3	3	1	1	1	2	3	0	0
14250004	3	3	2	0	2	2	2	2	0
14250006	2	3	2	0	2	2	3	0	0
14250007	2	3	1	0	1	2	2	2	0
14250009	3	4	1	0	0	3	3	0	0
14250010	3	3	2	0	2	3	3	0	0
14250011	2	2	1	0	0	2	2	0	1
14250013	3	3	2	0	2	2	2	0	0
14250014	3	3	1	1	1	2	3	0	0
14250016	3	2	1	0	1	3	2	0	0
14250018	0	0	3	0	4	3	2	3	0
14250021	2	3	2	0	2	2	3	2	0
14250019	2	3	1	1	2	3	3	1	0
14250024	2	3	2	0	3	3	3	3	0
14250025	3	3	1	0	2	2	3	0	0
14260005	2	3	1	0	1	2	3	1	0
14260006	3	3	1	0	1	3	3	0	0
14260007	2	2	1	0	1	2	3	0	0
14280003	3	3	1	0	1	2	3	0	0
14280004	3	3	2	0	1	2	3	0	0
14280007	3	3	3	2	3	2	3	3	0
14280009	3	4	1	0	1	2	3	0	0
14280011	3	3	1	2	0	2	3	1	0
14300001	2	2	2	0	2	2	3	0	0
14300002	2	3	2	0	2	3	3	3	0
14300003	2	3	2	0	2	2	3	0	0
14300004	2	3	4	0	4	2	3	4	0
14300006	3	3	0	0	0	2	2	0	0
14310001	2	3	0	0	0	3	3	0	0
14310002	2	3	0	0	0	2	3	0	0
14310004	3	4	3	0	3	2	3	3	0
14310005	3	4	2	0	2	3	3	2	0
14310006	3	4	1	0	2	3	3	2	0
14310007	2	3	2	0	1	3	3	0	0
14310008	2	3	1	0	0	3	3	0	0
14310009	4	4	2	0	2	3	3	0	0
14310011	3	3	1	0	3	2	3	2	0
14310013	4	4	0	1	1	2	3	0	0
14320003	3	3	1	0	1	2	3	1	0
14320004	4	4	2	0	3	3	3	3	0
14320008	3	3	0	0	0	3	2	0	1
14320009	3	3	1	0	0	2	3	0	1
14320010	2	2	0	0	0	2	3	0	0
14320011	2	2	0	0	0	2	2	0	0
14320012	2	2	0	0	0	2	2	0	0
14320013	2	2	1	0	2	2	3	2	0
14330001	3	3	1	0	0	3	3	1	0
14330005	3	3	2	0	2	2	2	0	0
14330011	2	3	2	1	2	3	2	2	0
14330006	3	2	1	0	1	2	2	1	0

Supplementary Table S11B (continued)

Recording ID	Audio quality	Target clarity	Other birds	Human noise	Signal overlap	Vocal intensity	Vocal continuity	Interaction presence	Observer effect
14330010	3	3	2	0	2	4	2	0	0
14330012	3	3	2	0	2	3	3	2	0
14330013	2	2	2	0	2	2	2	0	0
14330014	3	3	2	0	2	3	3	0	0
14330015	2	2	1	0	2	3	3	1	0
14330018	2	2	1	0	2	2	3	2	0
14330019	2	3	1	0	1	2	2	1	0
14340001	3	3	0	0	0	3	3	0	0
14340002	3	4	1	0	1	3	3	0	0
14340003	3	4	1	0	1	2	2	1	0
14340004	3	3	1	0	1	2	3	2	0
14340011	3	4	2	0	2	2	3	0	0
14350001	1	2	1	0	1	2	3	1	0
14350005	2	3	2	0	2	2	3	1	0
14350006	2	2	1	0	0	2	2	1	0
14350007	3	3	2	0	2	3	3	3	0
14350009	2	3	1	0	0	2	2	0	1
14360001	2	2	3	0	3	2	2	0	0
14360002	2	3	2	0	2	2	3	1	0
14360004	2	2	4	0	4	3	3	4	0
14370001	3	3	1	0	1	2	3	0	0
14370002	2	3	2	0	2	2	3	0	0
14380001	2	3	1	0	0	2	3	0	0
14380003	3	3	1	0	2	2	3	0	0
14380005	2	3	2	0	2	2	3	0	0
14390001	2	2	1	0	1	2	3	0	0
14390002	2	2	1	2	1	2	3	1	0
14390003	2	3	1	0	1	2	2	0	0
14410002	3	3	1	0	1	2	2	1	0
14410003	3	3	2	0	2	4	2	0	0
14410005	2	3	2	0	2	2	3	0	0
14420001	3	3	1	0	1	2	3	0	0
14430001	2	2	1	1	2	2	3	0	0
14430002	3	3	0	0	1	2	3	0	0
14430003	2	2	1	0	2	2	3	0	0
14430004	2	3	1	0	1	2	3	0	0
14440001	2	2	2	0	2	2	3	0	0
14440002	3	3	1	0	1	2	2	1	0
14440004	2	2	1	2	2	2	2	0	0
14450010	3	2	1	0	1	2	2	1	0
14460001	2	2	2	1	2	2	3	0	0
14460002	2	3	3	1	3	2	3	0	0
14460004	2	3	2	0	2	2	3	0	0
14460005	2	3	1	0	2	2	3	0	0
14460006	3	3	1	1	1	2	3	0	0
14470001	2	3	2	0	2	2	3	0	0
14470002	3	3	1	0	2	2	3	0	0
14470003	3	2	1	0	1	2	2	0	0
14470004	3	3	2	1	2	2	3	0	0
14470005	2	2	2	0	1	2	3	1	0
14490003	3	3	3	0	3	2	3	0	0
14510001	3	3	1	0	1	2	3	0	0
14510003	3	2	1	0	1	2	3	0	0
14510004	3	3	1	0	1	3	3	0	0
14510005	3	2	1	0	1	2	2	1	0
14510006	2	2	1	0	2	3	3	2	0
14510007	3	3	1	0	1	3	3	0	0

Recording ID	Audio quality	Target clarity	Other birds	Human noise	Signal overlap	Vocal intensity	Vocal continuity	Interaction presence	Observer effect
14520001	3	4	2	0	2	2	3	0	0
14520002	3	3	1	0	1	2	3	1	0
14520003	2	2	0	0	0	1	2	0	0
14520004	4	4	2	0	2	2	3	0	0
14520005	3	2	2	0	2	2	3	1	0
14530001	3	3	1	0	2	2	3	0	0
14530004	3	2	2	0	2	2	3	1	0
14530005	2	2	2	2	2	2	2	1	0
14540004	2	2	2	0	2	2	3	0	0
14550001	3	3	2	0	2	2	3	0	0
14550003	3	4	1	0	1	3	3	1	0
14550004	3	4	1	0	1	2	3	1	0
14550006	3	3	2	0	2	3	3	1	0
14550007	3	4	2	0	2	3	3	1	0
14560002	3	3	1	0	1	2	3	1	0
14560005	3	3	1	0	2	3	3	1	0
14560006	3	3	2	0	2	2	3	1	0
14570001	2	3	1	0	1	3	3	1	0
14570002	3	4	1	0	1	3	3	1	0
14570003	2	2	1	0	0	2	3	1	0
14570004	3	3	0	0	0	3	3	0	0
14570008	3	4	1	0	1	3	3	1	0
14570009	2	2	2	1	2	1	1	0	0
14570010	3	4	2	0	2	2	3	1	0
14580002	3	3	1	0	2	2	3	1	0
14580003	2	3	1	0	2	2	3	1	0
14580005	3	3	2	0	2	3	3	2	0
14580006	3	3	1	0	2	3	3	2	0
14590001	4	4	1	0	4	3	3	4	0
14600002	2	2	1	0	2	2	3	1	0
14600003	2	2	1	1	2	2	3	1	0
14600005	2	2	1	0	3	2	3	1	0
14600008	2	2	2	1	3	2	3	0	0
14600009	3	3	1	1	1	2	3	0	0
14600011	2	2	1	2	1	2	3	1	0
14610005	2	2	1	1	0	2	2	0	0
14610006	2	2	1	0	1	2	3	0	0
14610007	2	3	2	1	3	3	3	1	0
14620005	3	3	1	0	2	2	3	0	0
14620006	3	3	3	0	3	2	3	0	0
14620008	3	3	1	0	1	2	3	1	0
14630004	2	2	0	0	0	2	2	0	0
14630007	2	3	2	0	2	3	3	0	0
14630008	3	3	1	1	2	3	2	1	0
14630009	2	2	1	0	1	2	3	0	0
14640002	2	3	1	0	1	2	2	0	0
14640003	2	3	1	0	1	2	2	1	0
14640005	2	3	1	0	2	2	3	0	0
14650001	2	2	1	0	2	2	3	1	0
14660001	3	3	1	1	2	2	3	0	0
14660002	1	2	2	3	3	2	3	0	0
14670001	2	2	1	0	1	2	2	0	0
14680001	2	3	1	1	1	2	3	1	0
14680002	2	3	1	1	1	2	3	1	0
14690001	2	2	1	1	1	2	3	0	0
14700003	2	3	1	1	0	2	2	0	0

Notes. All annotation variables are ordinal scores from 0 to 4. Score definitions are provided in Supplementary Table S11C.

Supplementary Table S11C. Annotation score definitions

Annotation	Score	Definition
Audio quality	0	unusable
	1	poor
	2	usable but limited
	3	good
	4	excellent / high-SNR
Target clarity	0	absent / unusable
	1	very weak or highly ambiguous
	2	present and usable with caution
	3	clear
	4	very clear / dominant target
Other birds	0	absent or negligible
	1	light
	2	moderate
	3	strong
	4	dominant / blocking
Human noise	0	absent or negligible
	1	light
	2	moderate
	3	strong
	4	dominant / blocking
Signal overlap	0	no overlap
	1	minor / occasional
	2	moderate
	3	strong
	4	dominant / analysis-blocking
Vocal intensity	0	absent / not relevant
	1	sparse
	2	regular
	3	intense
	4	very intense / sustained high-output
Vocal continuity	0	absent / not relevant
	1	isolated strophes
	2	intermittent
	3	mostly continuous
	4	near-continuous / long sustained sequence
Interaction presence	0	no evident interaction
	1	possible distant/weak second individual
	2	clear second individual, limited interaction
	3	clear interaction / alternation / counter-singing
	4	dense interaction / intertwined multi-singer episode
Observer effect	0	no evident effect
	1	possible but uncertain
	2	mild likely effect
	3	clear effect
	4	strong immediate effect / movement or interruption

Notes. Definitions apply to the ordinal annotation scores reported in Supplementary Table S11B.

Supplementary Table S12. Recording-level denoising-sensitivity results

Recording ID	Area	MFCC-40						Chroma-80					
		35%			75%			35%			75%		
		P	F	R	P	F	R	P	F	R	P	F	R
14080004	1	0.9959	0.9785	0.399	0.9948	0.9799	0.396	0.9886	0.9857	0.357	0.9956	0.9860	0.322
14080008	1	0.9875	0.9803	0.427	0.9844	0.9776	0.456	0.9994	0.9992	0.168	0.9997	0.9921	0.237
14080010	1	0.9938	0.9792	0.425	0.9916	0.9762	0.435	0.9996	0.9967	0.237	0.9979	0.9846	0.256
14130001	1	0.9828	0.9720	0.451	0.9802	0.9668	0.478	0.9944	0.9949	0.295	0.9964	0.9810	0.332
14130004	2	0.9939	0.9807	0.434	0.9939	0.9819	0.428	0.9990	0.9808	0.301	0.9967	0.9840	0.269
14140001	1	0.9838	0.9738	0.468	0.9812	0.9694	0.488	0.9922	0.9953	0.301	0.9943	0.9855	0.334
14140002	1	0.9809	0.9624	0.465	0.9817	0.9606	0.473	0.9941	0.9948	0.294	0.9946	0.9867	0.302
14140003	2	0.9874	0.9798	0.438	0.9876	0.9785	0.445	0.9996	0.9962	0.240	0.9991	0.9880	0.249
14150001	2	0.9851	0.9768	0.434	0.9854	0.9763	0.433	0.9986	0.9898	0.241	0.9951	0.9810	0.240
14150003	1	0.9813	0.9779	0.406	0.9771	0.9735	0.439	0.9987	0.9988	0.188	0.9967	0.9873	0.249
14150005	4	0.9870	0.9767	0.442	0.9859	0.9777	0.424	0.9992	0.9840	0.297	0.9962	0.9821	0.269
14150008	3	0.9886	0.9790	0.403	0.9907	0.9796	0.420	0.9972	0.9990	0.208	0.9982	0.9885	0.249
14150010	5	0.9207	0.9457	0.557	0.9211	0.9398	0.604	0.9671	0.9917	0.403	0.9834	0.9802	0.379
14150011	3	0.9857	0.9756	0.441	0.9876	0.9751	0.453	0.9995	0.9966	0.210	0.9993	0.9850	0.263
14150012	4	0.9877	0.9777	0.443	0.9866	0.9783	0.424	0.9984	0.9826	0.316	0.9948	0.9784	0.289
14150013	3	0.9910	0.9792	0.433	0.9902	0.9784	0.441	0.9969	0.9958	0.270	0.9990	0.9849	0.276
14150015	1	0.9847	0.9725	0.470	0.9863	0.9704	0.473	0.9992	0.9879	0.305	0.9984	0.9850	0.282
14160001	1	0.9904	0.9792	0.378	0.9888	0.9739	0.405	0.9974	0.9986	0.207	0.9950	0.9863	0.264
14160006	3	0.9901	0.9810	0.411	0.9913	0.9830	0.401	0.9996	0.9903	0.220	0.9981	0.9895	0.198
14160008	3	0.9837	0.9670	0.434	0.9780	0.9618	0.455	0.9936	0.9826	0.331	0.9973	0.9872	0.284
14160011	2	0.9886	0.9795	0.398	0.9901	0.9769	0.423	0.9993	0.9992	0.172	0.9979	0.9939	0.251
14160012	2	0.9826	0.9781	0.424	0.9831	0.9782	0.432	0.9984	0.9986	0.211	0.9990	0.9886	0.244
14160013	2	0.9900	0.9725	0.413	0.9921	0.9757	0.427	0.9949	0.9941	0.244	0.9975	0.9892	0.222
14170005	3	0.9845	0.9752	0.455	0.9857	0.9756	0.458	0.9994	0.9914	0.263	0.9994	0.9855	0.266
14170006	4	0.9838	0.9681	0.408	0.9765	0.9732	0.406	0.9954	0.9899	0.269	0.9971	0.9869	0.250
14170008	6	0.9927	0.9757	0.390	0.9928	0.9778	0.393	0.9913	0.9882	0.313	0.9967	0.9872	0.280
14170009	3	0.9822	0.9750	0.475	0.9853	0.9779	0.462	0.9989	0.9814	0.299	0.9974	0.9856	0.255
14190002	4	0.9898	0.9772	0.408	0.9892	0.9768	0.399	0.9979	0.9929	0.343	0.9982	0.9799	0.375
14190006	6	0.9916	0.9822	0.385	0.9911	0.9803	0.398	0.9997	0.9975	0.194	0.9992	0.9900	0.259
14190007	4	0.9831	0.9737	0.422	0.9788	0.9699	0.418	0.9995	0.9922	0.308	0.9988	0.9833	0.308
14190010	5	0.9855	0.9795	0.363	0.9843	0.9763	0.391	0.9997	0.9995	0.105	0.9980	0.9958	0.195
14200003	1	0.9895	0.9813	0.336	0.9874	0.9776	0.361	0.9980	0.9990	0.150	0.9921	0.9885	0.243
14200004	7	0.9847	0.9733	0.406	0.9786	0.9692	0.426	0.9953	0.9925	0.324	0.9980	0.9763	0.347
14200005	7	0.9842	0.9686	0.390	0.9843	0.9652	0.387	0.9879	0.9898	0.327	0.9851	0.9736	0.349
14200006	1	0.9915	0.9810	0.379	0.9907	0.9797	0.384	0.9955	0.9985	0.253	0.9932	0.9859	0.360
14200010	4	0.9935	0.9798	0.380	0.9924	0.9809	0.371	0.9976	0.9955	0.293	0.9986	0.9837	0.324
14210001	2	0.9759	0.9631	0.406	0.9721	0.9584	0.432	0.9926	0.9907	0.214	0.9866	0.9783	0.272
14210002	2	0.9854	0.9772	0.370	0.9797	0.9687	0.393	0.9957	0.9970	0.224	0.9891	0.9793	0.269
14210003	2	0.9844	0.9786	0.356	0.9793	0.9730	0.372	0.9994	0.9995	0.144	0.9945	0.9937	0.254
14210004	2	0.9757	0.9711	0.408	0.9680	0.9623	0.442	0.9991	0.9994	0.142	0.9921	0.9940	0.258
14230004	2	0.9833	0.9790	0.408	0.9819	0.9764	0.424	0.9980	0.9976	0.208	0.9948	0.9830	0.266

Recording ID	Area	MFCC-40						Chroma-80					
		35%			75%			35%			75%		
		P	F	R	P	F	R	P	F	R	P	F	R
14230005	2	0.9905	0.9787	0.425	0.9893	0.9775	0.423	0.9991	0.9957	0.292	0.9992	0.9818	0.343
14240002	8	0.9913	0.9824	0.346	0.9913	0.9802	0.369	1.0000	0.9998	0.108	0.9997	0.9985	0.221
14240010	4	0.9854	0.9795	0.380	0.9841	0.9759	0.396	0.9998	0.9986	0.156	0.9995	0.9915	0.283
14240012	8	0.9848	0.9779	0.369	0.9845	0.9760	0.391	0.9998	0.9985	0.173	0.9969	0.9885	0.248
14240022	8	0.9888	0.9801	0.370	0.9873	0.9763	0.397	0.9999	0.9994	0.133	0.9993	0.9947	0.250
14250004	7	0.9832	0.9768	0.400	0.9825	0.9756	0.398	0.9976	0.9905	0.277	0.9959	0.9725	0.325
14250006	1	0.9935	0.9782	0.387	0.9929	0.9788	0.389	0.9945	0.9971	0.246	0.9908	0.9840	0.361
14250007	7	0.9838	0.9759	0.394	0.9833	0.9740	0.401	0.9975	0.9966	0.205	0.9931	0.9751	0.292
14250009	8	0.9945	0.9850	0.326	0.9940	0.9818	0.350	0.9998	0.9996	0.105	0.9998	0.9961	0.218
14250010	8	0.9856	0.9714	0.431	0.9871	0.9720	0.437	0.9999	0.9978	0.160	0.9996	0.9910	0.270
14250011	7	0.9861	0.9768	0.405	0.9858	0.9775	0.398	0.9985	0.9874	0.347	0.9986	0.9776	0.356
14250013	8	0.9918	0.9826	0.353	0.9913	0.9790	0.377	0.9982	0.9990	0.180	0.9954	0.9884	0.287
14250014	1	0.9917	0.9816	0.366	0.9914	0.9804	0.366	0.9989	0.9985	0.166	0.9964	0.9914	0.275
14250019	3	0.9796	0.9720	0.388	0.9760	0.9657	0.399	0.9912	0.9910	0.243	0.9804	0.9693	0.324
14250021	4	0.9868	0.9792	0.344	0.9838	0.9749	0.358	0.9969	0.9971	0.162	0.9883	0.9840	0.250
14250025	8	0.9846	0.9733	0.353	0.9817	0.9671	0.375	0.9936	0.9961	0.229	0.9873	0.9851	0.293
14260005	5	0.9922	0.9831	0.340	0.9912	0.9793	0.361	0.9999	0.9991	0.133	0.9994	0.9930	0.219
14280003	5	0.9873	0.9785	0.359	0.9841	0.9740	0.386	0.9995	0.9991	0.157	0.9986	0.9870	0.234
14280007	4	0.9899	0.9821	0.371	0.9902	0.9800	0.388	0.9996	0.9984	0.147	0.9990	0.9932	0.231
14280011	5	0.9873	0.9806	0.331	0.9854	0.9757	0.364	0.9999	0.9996	0.091	0.9977	0.9943	0.211
14300001	6	0.9913	0.9801	0.395	0.9914	0.9763	0.405	0.9985	0.9945	0.279	0.9990	0.9823	0.338
14300002	3	0.9895	0.9824	0.351	0.9891	0.9798	0.368	0.9998	0.9988	0.155	0.9989	0.9910	0.267
14300003	5	0.9895	0.9813	0.378	0.9900	0.9797	0.382	0.9996	0.9959	0.235	0.9991	0.9867	0.269
14300006	3	0.9748	0.9663	0.442	0.9738	0.9635	0.442	0.9974	0.9897	0.349	0.9977	0.9762	0.348
14310001	8	0.9928	0.9816	0.383	0.9902	0.9768	0.392	0.9997	0.9966	0.223	0.9993	0.9897	0.281
14310002	1	0.9754	0.9754	0.386	0.9644	0.9667	0.431	0.9560	0.9879	0.410	0.9648	0.9838	0.415
14310004	3	0.9869	0.9793	0.384	0.9857	0.9767	0.407	0.9992	0.9990	0.181	0.9987	0.9935	0.284
14310005	4	0.9882	0.9813	0.382	0.9873	0.9784	0.397	0.9988	0.9984	0.206	0.9988	0.9884	0.302
14310006	4	0.9852	0.9783	0.371	0.9844	0.9733	0.399	0.9995	0.9984	0.176	0.9992	0.9885	0.273
14310007	3	0.9845	0.9777	0.413	0.9844	0.9749	0.420	0.9989	0.9960	0.251	0.9989	0.9841	0.337
14310008	5	0.9920	0.9821	0.359	0.9900	0.9767	0.380	0.9994	0.9990	0.175	0.9978	0.9893	0.247
14310009	3	0.9887	0.9811	0.384	0.9887	0.9781	0.411	0.9997	0.9993	0.141	0.9978	0.9954	0.283
14320003	6	0.9944	0.9844	0.370	0.9952	0.9824	0.379	0.9998	0.9982	0.177	0.9996	0.9891	0.272
14320004	3	0.9933	0.9853	0.331	0.9931	0.9828	0.356	1.0000	0.9997	0.105	0.9994	0.9972	0.242
14320009	9	0.9930	0.9821	0.386	0.9921	0.9783	0.393	0.9985	0.9961	0.241	0.9991	0.9875	0.286
14320010	9	0.9901	0.9775	0.419	0.9858	0.9747	0.426	0.9983	0.9914	0.324	0.9989	0.9875	0.319
14320011	9	0.9920	0.9759	0.423	0.9874	0.9766	0.399	0.9988	0.9728	0.362	0.9991	0.9859	0.303
14320013	6	0.9876	0.9773	0.427	0.9862	0.9725	0.449	0.9994	0.9945	0.227	0.9993	0.9828	0.289
14330001	3	0.9926	0.9845	0.371	0.9922	0.9828	0.370	0.9997	0.9976	0.205	0.9996	0.9883	0.279
14330005	6	0.9870	0.9806	0.413	0.9854	0.9775	0.422	0.9996	0.9974	0.217	0.9992	0.9874	0.290

Supplementary Table S12 (continued)

Recording ID	Area	MFCC-40						Chroma-80					
		35%			75%			35%			75%		
		P	F	R	P	F	R	P	F	R	P	F	R
14330010	9	0.9915	0.9774	0.392	0.9878	0.9705	0.422	0.9998	0.9965	0.220	0.9991	0.9846	0.319
14330011	9	0.9887	0.9752	0.442	0.9863	0.9741	0.440	0.9938	0.9943	0.297	0.9956	0.9834	0.364
14330012	9	0.9896	0.9736	0.436	0.9858	0.9716	0.440	0.9977	0.9901	0.341	0.9977	0.9751	0.396
14330015	9	0.9905	0.9797	0.400	0.9879	0.9753	0.421	0.9993	0.9960	0.248	0.9993	0.9855	0.304
14330018	9	0.9879	0.9667	0.430	0.9827	0.9663	0.421	0.9890	0.9852	0.403	0.9902	0.9821	0.429
14330019	6	0.9872	0.9647	0.482	0.9837	0.9684	0.456	0.9960	0.9781	0.380	0.9976	0.9799	0.366
14340001	8	0.9907	0.9793	0.404	0.9883	0.9760	0.418	0.9988	0.9970	0.227	0.9987	0.9914	0.296
14340002	3	0.9905	0.9752	0.437	0.9893	0.9772	0.422	0.9974	0.9935	0.295	0.9969	0.9868	0.351
14340003	9	0.9894	0.9725	0.375	0.9897	0.9725	0.399	0.9976	0.9993	0.202	0.9902	0.9922	0.353
14340004	6	0.9634	0.9703	0.452	0.9577	0.9635	0.493	0.9907	0.9978	0.269	0.9902	0.9901	0.284
14340011	1	0.9911	0.9832	0.362	0.9904	0.9818	0.386	0.9992	0.9994	0.182	0.9991	0.9935	0.249
14350005	5	0.9854	0.9770	0.415	0.9843	0.9752	0.421	0.9997	0.9956	0.219	0.9993	0.9807	0.288
14350006	6	0.9928	0.9776	0.415	0.9908	0.9775	0.410	0.9986	0.9920	0.292	0.9995	0.9830	0.302
14350009	9	0.9921	0.9747	0.389	0.9897	0.9722	0.403	0.9872	0.9926	0.355	0.9897	0.9872	0.376
14370001	2	0.9890	0.9823	0.377	0.9884	0.9803	0.377	0.9983	0.9983	0.213	0.9948	0.9830	0.271
14370002	10	0.9633	0.9688	0.456	0.9562	0.9628	0.500	0.9997	0.9991	0.136	0.9987	0.9930	0.221
14380001	2	0.9873	0.9783	0.414	0.9873	0.9767	0.434	0.9992	0.9984	0.190	0.9963	0.9830	0.240
14380003	5	0.9892	0.9820	0.396	0.9867	0.9785	0.429	0.9999	0.9996	0.131	0.9995	0.9913	0.242
14380005	2	0.9894	0.9814	0.423	0.9903	0.9812	0.418	0.9990	0.9970	0.248	0.9984	0.9850	0.312
14390001	2	0.9870	0.9779	0.412	0.9854	0.9749	0.410	0.9943	0.9946	0.245	0.9917	0.9809	0.281
14390002	3	0.9738	0.9700	0.448	0.9707	0.9659	0.470	0.9971	0.9957	0.231	0.9951	0.9833	0.270
14390003	3	0.9748	0.9632	0.392	0.9721	0.9566	0.407	0.9878	0.9842	0.259	0.9793	0.9658	0.288
14410002	10	0.9930	0.9758	0.374	0.9915	0.9717	0.393	0.9937	0.9972	0.266	0.9927	0.9807	0.315
14410005	2	0.9910	0.9809	0.372	0.9876	0.9784	0.415	0.9998	0.9993	0.130	0.9995	0.9955	0.203
14420001	10	0.9886	0.9809	0.304	0.9840	0.9736	0.334	0.9984	0.9981	0.162	0.9963	0.9910	0.191
14430001	1	0.9902	0.9772	0.347	0.9916	0.9766	0.387	0.9988	0.9996	0.137	0.9958	0.9966	0.239
14430002	5	0.9854	0.9798	0.373	0.9879	0.9801	0.370	0.9969	0.9955	0.248	0.9923	0.9772	0.286
14430004	5	0.9870	0.9795	0.379	0.9857	0.9756	0.402	0.9986	0.9993	0.196	0.9989	0.9917	0.282
14440002	10	0.9949	0.9836	0.333	0.9952	0.9816	0.356	0.9958	0.9989	0.222	0.9962	0.9883	0.285
14440004	2	0.9939	0.9856	0.383	0.9935	0.9843	0.400	0.9998	0.9985	0.179	0.9992	0.9892	0.255
14450010	5	0.9923	0.9796	0.430	0.9904	0.9785	0.416	0.9995	0.9882	0.325	0.9990	0.9855	0.304
14460001	3	0.9902	0.9794	0.407	0.9901	0.9777	0.402	0.9993	0.9942	0.289	0.9979	0.9760	0.353
14460004	1	0.9928	0.9846	0.356	0.9917	0.9832	0.385	0.9998	0.9998	0.106	0.9993	0.9971	0.184
14460005	3	0.9747	0.9692	0.472	0.9648	0.9592	0.506	0.9982	0.9928	0.283	0.9974	0.9777	0.301
14460006	8	0.9905	0.9826	0.414	0.9899	0.9811	0.418	0.9998	0.9972	0.203	0.9995	0.9895	0.297
14470001	1	0.9868	0.9811	0.382	0.9849	0.9782	0.405	0.9995	0.9989	0.166	0.9989	0.9913	0.281
14470002	1	0.9894	0.9820	0.378	0.9883	0.9791	0.408	0.9999	0.9996	0.140	0.9997	0.9942	0.253
14510001	2	0.9858	0.9806	0.390	0.9842	0.9780	0.420	0.9999	0.9996	0.118	0.9993	0.9947	0.222
14510003	6	0.9920	0.9766	0.409	0.9915	0.9742	0.416	0.9961	0.9953	0.310	0.9948	0.9834	0.379
14510004	3	0.9900	0.9828	0.388	0.9889	0.9805	0.405	0.9999	0.9987	0.154	0.9998	0.9929	0.241

Recording ID	Area	MFCC-40						Chroma-80					
		35%			75%			35%			75%		
		P	F	R	P	F	R	P	F	R	P	F	R
14510006	5	0.9893	0.9823	0.389	0.9879	0.9798	0.415	0.9997	0.9992	0.155	0.9995	0.9918	0.229
14510007	3	0.9887	0.9809	0.388	0.9885	0.9788	0.410	0.9998	0.9989	0.149	0.9997	0.9942	0.243
14520001	1	0.9916	0.9821	0.374	0.9903	0.9804	0.407	0.9993	0.9998	0.112	0.9978	0.9977	0.219
14520004	1	0.9866	0.9802	0.401	0.9837	0.9771	0.415	0.9992	0.9981	0.207	0.9985	0.9832	0.308
14520005	6	0.9898	0.9789	0.422	0.9894	0.9751	0.435	0.9995	0.9962	0.212	0.9997	0.9893	0.295
14530001	2	0.9893	0.9822	0.380	0.9896	0.9802	0.397	0.9993	0.9990	0.187	0.9989	0.9883	0.258
14530004	6	0.9865	0.9747	0.413	0.9862	0.9708	0.434	0.9983	0.9977	0.210	0.9981	0.9871	0.292
14530005	6	0.9961	0.9812	0.301	0.9938	0.9762	0.335	0.9997	0.9996	0.086	0.9984	0.9981	0.175
14550001	6	0.9950	0.9851	0.366	0.9955	0.9833	0.376	0.9998	0.9987	0.171	0.9997	0.9913	0.264
14550003	3	0.9904	0.9828	0.351	0.9888	0.9809	0.379	0.9999	0.9998	0.081	0.9998	0.9987	0.171
14550004	6	0.9835	0.9787	0.384	0.9790	0.9739	0.416	0.9999	0.9995	0.132	0.9995	0.9938	0.222
14550007	3	0.9696	0.9690	0.429	0.9634	0.9629	0.471	0.9997	0.9993	0.131	0.9987	0.9955	0.226
14560002	6	0.9931	0.9786	0.382	0.9919	0.9725	0.406	0.9990	0.9988	0.188	0.9919	0.9879	0.324
14560005	3	0.9918	0.9834	0.349	0.9903	0.9815	0.377	0.9999	0.9997	0.093	0.9996	0.9982	0.180
14560006	6	0.9894	0.9822	0.365	0.9883	0.9784	0.390	0.9999	0.9993	0.139	0.9997	0.9911	0.247
14570001	3	0.9970	0.9845	0.309	0.9964	0.9818	0.343	0.9995	0.9998	0.113	0.9986	0.9985	0.196
14570002	3	0.9907	0.9785	0.343	0.9910	0.9756	0.374	0.9988	0.9997	0.127	0.9949	0.9978	0.245
14570003	6	0.9843	0.9737	0.433	0.9846	0.9737	0.426	0.9966	0.9913	0.328	0.9981	0.9786	0.338
14570008	3	0.9935	0.9846	0.375	0.9937	0.9832	0.387	0.9999	0.9992	0.147	0.9997	0.9919	0.251
14570010	3	0.9947	0.9818	0.354	0.9948	0.9798	0.387	0.9999	0.9997	0.111	0.9998	0.9976	0.202
14580002	6	0.9937	0.9775	0.391	0.9904	0.9713	0.415	0.9998	0.9970	0.208	0.9988	0.9875	0.294
14580005	5	0.9940	0.9826	0.368	0.9939	0.9805	0.399	0.9999	0.9995	0.107	0.9999	0.9966	0.190
14580006	3	0.9915	0.9819	0.363	0.9912	0.9799	0.394	0.9998	0.9995	0.129	0.9996	0.9960	0.206
14600002	5	0.9928	0.9820	0.385	0.9936	0.9799	0.401	0.9998	0.9991	0.162	0.9997	0.9926	0.258
14600003	5	0.9960	0.9812	0.392	0.9958	0.9804	0.389	0.9997	0.9965	0.236	0.9997	0.9836	0.294
14600008	6	0.9892	0.9776	0.385	0.9900	0.9763	0.400	0.9960	0.9980	0.208	0.9941	0.9913	0.305
14600009	3	0.9933	0.9813	0.325	0.9932	0.9781	0.356	0.9994	0.9998	0.089	0.9963	0.9984	0.209
14600011	5	0.9820	0.9640	0.429	0.9827	0.9621	0.449	0.9993	0.9991	0.112	0.9956	0.9953	0.262
14620005	7	0.9863	0.9792	0.398	0.9866	0.9786	0.399	0.9993	0.9965	0.234	0.9986	0.9787	0.285
14620008	6	0.9899	0.9813	0.394	0.9882	0.9797	0.408	0.9995	0.9971	0.191	0.9989	0.9908	0.227
14630004	7	0.9767	0.9690	0.440	0.9712	0.9677	0.436	0.9994	0.9881	0.269	0.9951	0.9780	0.293
14640002	10	0.9933	0.9763	0.421	0.9914	0.9759	0.424	0.9845	0.9926	0.393	0.9875	0.9787	0.437
14640003	6	0.9920	0.9822	0.399	0.9923	0.9809	0.409	0.9996	0.9975	0.220	0.9996	0.9869	0.274
14640005	5	0.9859	0.9770	0.421	0.9832	0.9746	0.437	0.9991	0.9972	0.242	0.9994	0.9852	0.297
14650001	5	0.9936	0.9827	0.381	0.9943	0.9817	0.391	0.9998	0.9983	0.206	0.9998	0.9892	0.258
14670001	5	0.9943	0.9819	0.393	0.9923	0.9813	0.408	0.9979	0.9965	0.296	0.9991	0.9797	0.329
14680001	5	0.9879	0.9788	0.414	0.9869	0.9772	0.418	0.9991	0.9971	0.255	0.9994	0.9864	0.280
14680002	5	0.9926	0.9767	0.412	0.9890	0.9764	0.406	0.9861	0.9926	0.358	0.9882	0.9855	0.411
14690001	5	0.9746	0.9718	0.470	0.9732	0.9719	0.477	0.9991	0.9948	0.205	0.9975	0.9824	0.259
14700003	5	0.9891	0.9759	0.428	0.9882	0.9756	0.420	0.9992	0.9907	0.301	0.9995	0.9833	0.295

Table S13: Technical configuration of automatic strophe detection, indexing and stitched-block assembly.

Stage	Parameter	Corpus setting and implementation	Source
Input	Analytical input	Final-denoised, manually contoured WAV recordings; non-target intervals were represented by digital silence.	Procedure
	Sampling	Native input sampling rate retained; all corpus inputs were 48,000 Hz. No resampling was performed by the indexer.	Fixed
	Detection channels	Multichannel input was averaged arithmetically to a mono <code>float32</code> analysis copy. Exported audio was sliced from the original input array.	Fixed
Detection front end	Frequency band	1,000–15,000 Hz (<code>--fmin 1000 --fmax 15000</code>).	Explicit
	Band-pass filter	Butterworth band-pass (<code>scipy.signal.butter</code> , $N = 4$, second-order-section form), applied forward and backward with <code>sosfiltfilt</code> .	Fixed
	RMS window	20 ms, corresponding to 960 samples at 48 kHz.	Default
	Frame hop	384 samples, corresponding to 8 ms and 125 RMS frames s^{-1} at 48 kHz.	Default
	RMS scale	$r_i = \sqrt{\text{mean}(x_i^2) + 10^{-12}}$; $d_i = 20 \log_{10}(r_i + 10^{-12})$.	Fixed
Adaptive threshold	Noise-floor estimate	20th percentile of the band-limited framewise RMS-dB distribution, calculated independently for each recording (<code>--floor-percentile 20</code>).	Explicit
	Activity threshold	A frame was active when its RMS level was strictly greater than the recording-specific floor plus 10 dB (<code>--threshold-db 10</code>).	Explicit
Event definition	Region construction	Consecutive active frames formed one half-open detection interval. Inactive gaps were not merged.	Fixed
	Minimum duration	700 ms (<code>--min-duration-ms 700</code>); shorter regions were discarded.	Explicit
	Boundary padding	0 ms before and after each detected interval (<code>--pad-ms 0</code>).	Explicit
	Sample boundaries	For active frames $[f_0, f_1)$, start sample = $f_0 H$ and end sample = $(f_1 - 1)H + W$, clipped to the available audio, where H is the hop and W the RMS window.	Fixed
Stitching	Per-recording audition gap	150 ms of digital silence between detected slices.	Default
	Final-block assembly	Enabled (<code>--finalize-block</code>). The final block was assembled from the per-recording index CSVs without re-running detection.	Explicit + fixed
	Final-block gap	200 ms of digital silence between successive slices (<code>--block-gap-ms 200</code>); no gap followed the final event.	Explicit
	Ordering and identifiers	Natural filename order across recordings and original event order within recordings; source recording and vocalization identifiers were preserved.	Default + fixed
	Audio preservation	Unfiltered source slices were written using the input sample rate, channel count and WAV subtype; finalization required these properties to match across the block.	Fixed
Outputs	Index and provenance	Per-recording event CSV, detection preview, stitched audition WAV and JSON summary; final-block WAV, compatible index CSV, source-to-block mapping CSV and JSON summary.	Fixed
	Numerical indexing	Times exported to six decimal places together with exact source and block sample indices and 384-sample-hop frame indices.	Fixed

Notes. “Explicit” denotes an option supplied on the command line for every corpus batch; “Default” denotes a parser default that remained active because it was not overridden; and “Fixed” denotes behavior encoded directly in the script. Manual acoustic contouring preceded detection. Where adjacent strophes required separation, the manually inserted one-second digital-silence delimiters described in the Methods were already present in the indexing input. The per-recording CSVs were the frozen source of truth for final-block construction, so block finalization preserved event count and order rather than detecting the stitched audio again.

Table S14: Technical parameters for MFCC-40 and Chroma-80 feature computation, terrain construction, PCA and UMAP embedding. The same branch-specific settings were applied uniformly across all retained recordings and indexed strophes.

Stage	MFCC branch	Chroma branch
Audio input and decoding	48-kHz, 16-bit PCM WAV analysis files derived from the 32-bit floating-point source recordings. Audio was read with <code>scipy.io.wavfile</code> , converted to <code>float32</code> , and reduced from stereo to mono by sample-wise channel averaging. No sample-rate conversion was performed by the feature script.	48-kHz, 16-bit PCM WAV analysis files derived from the 32-bit floating-point source recordings. Audio was read with the Python <code>wave</code> module, converted to <code>float32</code> , and reduced from stereo to mono by sample-wise channel averaging; SoundFile provided fallback support for other WAV subtypes. No sample-rate conversion was performed by the feature script.
Frame and FFT settings	Periodic Hann window, 4096 samples (85.33 ms); hop 384 samples (8.00 ms; 125 frames s^{-1}); FFT size 4096. Frames began at sample zero; no centring, padding or incomplete terminal frames.	Symmetric Hann window, 4096 samples (85.33 ms); hop 384 samples (8.00 ms; 125 frames s^{-1}); FFT size 4096. Frames began at sample zero; no centring, padding or incomplete terminal frames.
Core representation	40 MFCC coefficients from 128 triangular mel filters spanning 20 Hz–24 kHz.	80-bin microtonal chroma on an 80 equal-division-of-the-octave grid. Positive FFT frequencies through 24 kHz were octave-folded; A4 = 440 Hz and detuning offset = 0 cents.
Spectral and coefficient processing	Pre-emphasis $\alpha = 0.97$; power spectrum $ X(f) ^2/4096$; natural-log mel energies with floor 10^{-12} ; orthonormal type-II DCT; lifter $L = 22$.	Linear-magnitude weighting $ X(f) $; nearest-bin 80-EDO assignment; power weighting disabled; prewhitening $\alpha = 0$; per-frame L1 normalization enabled; circular smoothing $\sigma = 0$; EMA coefficient 1.
Derivatives and export	Delta and delta-delta coefficients disabled. Each frame contained 40 values. Times and coefficients exported to CSV with six decimal places.	No temporal derivatives. Each frame contained 80 values. Times and chroma bins exported to CSV with six decimal places.
Strophe extraction	Start times mapped using floor and end times using ceiling. Frame ranges were uncentred, clipped to the feature sequence and required at least four source frames.	Start times mapped using floor and end times using ceiling. Frame ranges were uncentred, clipped to the feature sequence and required at least four source frames.
Temporal normalization and terrain	Linear interpolation to 256 temporal positions, producing 256×40 terrains and 10,240 flattened values per strophe.	Linear interpolation to 256 temporal positions, producing 256×80 terrains and 20,480 flattened values per strophe.
Pre-PCA scaling	Every flattened terrain dimension was globally standardized across the 20,075 strophes using <code>StandardScaler</code> .	No additional scaling after per-frame L1 normalization.
PCA	One global PCA fitted across all MFCC terrains; 1000 components; <code>whiten=False</code> , <code>svd_solver=auto</code> , <code>random_state=42</code> .	One global PCA fitted across all chroma terrains; 1000 components; <code>whiten=False</code> , <code>svd_solver=auto</code> , <code>random_state=42</code> .
UMAP input	Unwhitened PCA-1000 scores without further scaling; duration weight = 0 and spectral-centroid weight = 0.	Unwhitened PCA-1000 scores without further scaling; duration weight = 0 and spectral-centroid weight = 0.
UMAP	Cosine metric; <code>n_neighbors=15</code> ; <code>min_dist=0.02</code> ; <code>n_components=3</code> ; <code>random_state=42</code> . Other parameters retained at package defaults.	Cosine metric; <code>n_neighbors=15</code> ; <code>min_dist=0.02</code> ; <code>n_components=3</code> ; <code>random_state=42</code> . Other parameters retained at package defaults.
Coordinate and metadata handling	Raw UMAP coordinates retained without post-embedding normalization. QGIS metadata and diagnostic fields were added only after embedding.	Raw UMAP coordinates retained without post-embedding normalization. QGIS metadata and diagnostic fields were added only after embedding.

Notes. MFCC and chroma were fitted in independent global PCA and UMAP coordinate systems. Post-embedding metadata and diagnostic variables did not contribute to either geometry. Software versions are reported in Supplementary Table S15.

Table S15: Software and primary computational libraries used for corpus construction, acoustic analysis, visualization and interactive inspection.

Component	Software/library	Version	Role
Runtime	Python	3.10.18	Script execution environment
Numerical computation	NumPy	1.26.4	Array operations, FFT calculations, interpolation and vectorized numerical processing
Scientific computation	SciPy	1.15.2	WAV input, signal filtering, windowing, discrete cosine transforms and statistical functions
Data handling	pandas	2.3.3	CSV processing, metadata alignment and result aggregation
Scaling, PCA and neighborhoods	scikit-learn	1.7.2	Feature standardization, principal-component analysis and exact nearest-neighbor search
Manifold embedding	umap-learn	0.5.9.post2	Construction of the three-dimensional UMAP embeddings
Audio input and output	SoundFile / libsndfile	0.13.1 / 1.2.2	Audio decoding, segment extraction, stitched-output construction and WAV writing
Diagnostic visualization	Matplotlib	3.10.8	Generation of strophe-detection preview plots
Spatial data	QGIS	4.0.1	Recording-metadata management, georeferencing and cartographic visualization
Audio preprocessing	iZotope RX	11	Two-pass denoising and manual audio preparation
Interactive inspection	TouchDesigner	2025.30060	Acoustic-terrain and manifold navigation, audio retrieval and supplementary-video production

Notes. Only directly used software and primary computational libraries are listed. Python standard-library modules and transitive dependencies are omitted. SoundFile used libsndfile as its audio input/output backend.

References

- [1] Silke Kipper, Roger Mundry, Henrike Hultsch, and Dietmar Todt. Long-term persistence of song performance rules in nightingales (*Luscinia megarhynchos*): A longitudinal field study on repertoire size and composition. *Behaviour*, 141(3):371–390, 2004.
- [2] Hansjoerg P. Kunc, Valentin Amrhein, and Marc Naguib. Acoustic features of song categories and their possible implications for communication in the common nightingale (*Luscinia megarhynchos*). *Behaviour*, 142(8):1083–1097, 2005.
- [3] Sarah Kiefer, Anne Spiess, Silke Kipper, Roger Mundry, Christina Sommer, Henrike Hultsch, and Dietmar Todt. First-year common nightingales (*Luscinia megarhynchos*) have smaller song-type repertoire sizes than older males. *Ethology*, 112(12):1217–1224, 2006.
- [4] Sarah Kiefer, Christina Sommer, Constance Scharff, Silke Kipper, and Roger Mundry. Tuning towards tomorrow? common nightingales *Luscinia megarhynchos* change and increase their song repertoires from the first to the second breeding season. *Journal of Avian Biology*, 40(2):231–236, 2009.
- [5] Silke Kipper, Roger Mundry, Christina Sommer, Henrike Hultsch, and Dietmar Todt. Song repertoire size is correlated with body measures and arrival date in common nightingales (*Luscinia megarhynchos*). *Animal Behaviour*, 71(1):211–217, 2006.

- [6] Conny Bartsch, Michael Weiss, and Silke Kipper. Multiple song features are related to paternal effort in common nightingales. *BMC Evolutionary Biology*, 15:115, 2015.
- [7] Dietmar Todt and Henrike Hultsch. How songbirds deal with large amounts of serial information: retrieval rules suggest a hierarchical song memory. *Biological Cybernetics*, 79(6):487–500, 1998.
- [8] Michael Weiss, Henrike Hultsch, Iris Adam, Constance Scharff, and Silke Kipper. The use of network analysis to study complex animal communication systems: a study on nightingale song. *Proceedings of the Royal Society B: Biological Sciences*, 281(1785):20140460, 2014.
- [9] Philipp Sprau and Roger Mundry. Song type sharing in common nightingales, *Luscinia megarhynchos*, and its implications for cultural evolution. *Animal Behaviour*, 80(3):427–434, 2010.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] Denise Jäckel, Kim G. Mortega, Ulrich Brockmeyer, Gerlind U. C. Lehmann, and Silke L. Voigt-Heucke. Unravelling the stability of nightingale song over time and space using open, citizen science and shared data. *Frontiers in Ecology and Evolution*, 10:778610, 2022.
- [14] 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.
- [15] 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.
- [16] Lucio Arese. Shared acoustic manifolds for exploratory comparison of passerine vocalizations. EcoEvoRxiv, January 2026. Preprint. doi: 10.32942/X2W65N.
- [17] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.
- [18] Marc Naguib, Rainer Altenkamp, and Benjamin Griessmann. Nightingales in space: Song and extra-territorial forays of radio tagged song birds. *Journal für Ornithologie*, 142(3):306–312, 2001.
- [19] Tobias Roth and Valentin Amrhein. Estimating individual survival using territory occupancy data on unmarked animals. *Journal of Applied Ecology*, 47(2):386–392, 2010.
- [20] S. Wendlandt and Dietmar Todt. Ontogeny of time structure in nightingale song. *Bioacoustics*, 6(4):322–323, 1996. Abstract.

- [21] Philipp Sprau, Tobias Roth, Rouven Schmidt, Valentin Amrhein, and Marc Naguib. Communication across territory boundaries: distance-dependent responses in nightingales. *Behavioral Ecology*, 21(5):1011–1017, 2010.
- [22] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.
- [23] 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.
- [24] Tim Sainburg and Timothy Q. Gentner. Toward a computational neuroethology of vocal communication: From bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers in Behavioral Neuroscience*, 15:811737, 2021.
- [25] Veronica Morfi, Robert F. Lachlan, and Dan Stowell. Deep perceptual embeddings for unlabelled animal sound events. *The Journal of the Acoustical Society of America*, 150(1):2–11, 2021.
- [26] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, 4 edition, 2013.