

Symbiotic Habitat: Multilayer Audio-Visual Generation and Graphic Notation Based on Ecological Soundscape Analysis

Wenyue Jiang, Zimo Dong *

Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai, China

* Corresponding author: Zimo Dong

Abstract: This paper offers Multilayer Ecological Soundscape Notation (MESN) to make the acoustic base of generative ecological soundscape graphics traceable. Environmental recordings from four Chongming habitats, Song recordings from 25 bird species, and human activity noises are evaluated independently using RMS Energy, Spectral Centroid, Onset Strength, and low/mid/high frequency-band energy. After source-wise normalization and temporal alignment, MESN maps acoustic information to graphical variables and temporal properties to graphic primitives on a shared notation timeline. Case studies indicate that identical mapping principles create diverse outputs while keeping correlation to substantial acoustic changes. Applied in Symbiotic Habitat, MESN provides a traceable mechanism for ecological soundscape notation.

Keywords: Ecological Soundscape; Audio-Visual Generation; Graphic Notation; Acoustic Feature Analysis.

1. Introduction

Common ways to display the temporal, spectral and energy structure in ecoacoustic analysis are spectrograms, acoustic indices and signal processing measures. Audio-visual production is often insufficient to explain the translation of acoustic dimensions into graphics. This hampers traceability for environmental, avian and human sources in ecological soundscapes and prevents the preservation of temporal linkages among sources.

The study proposes a Multilayer Ecological Soundscape Notation (MESN) based on field recordings of farmland, reedbed, mudflat, and woodland in Chongming, 25 kinds of bird songs, and sounds of human activities. The sources are aligned at the notation stage and analysed independently. RMS Energy, Spectral Centroid, Onset Strength and frequency-band energy are transferred to graphical thickness, trajectory, diffusion and spectral texture respectively, while temporal continuity and discreteness define lines, arcs and points. MESN considers these features as descriptors of the sound structure, not as indications of ecological quality.

2. Literature Review

Pijanowski et al. defined soundscape ecology as a system at landscape size consisting of biological, geophysical and anthropogenic sound [1]. Sueur and Farina defined ecoacoustics as the ecological study and interpretation of ambient sound [2], and Gasc et al. highlighted the importance of ornithology to soundscape metrics and disturbance analysis [3]. Sethi et al. have shown that single acoustic indices and standard ML models are not reliable predictors of species richness across datasets [4]. Our study supports the approach of using acoustic features in our study as structural descriptors rather than direct biodiversity proxies.

Koch et al. developed a multidimensional birdsong analysis with an emphasis on interpretability and generalisability [5]. Shim et al. and Kim et al. exhibited sound-driven visual production [6-7], but acoustic-to-graphic

linkages are still mostly entrenched in latent representations.

Finke stressed point, line and space as structural aspects of graphic notation [8]. The Decibel ScorePlayer structured continuous and multi-part temporal forms [9] and Karchkhadze et al. transferred visual notation into a computational graphic-semantic-sound pipeline [10]. MESN closes the remaining gap by analysing recorded ecology sound by source type and translating explicit acoustic properties to a standardised notation timeline.

3. Method

MESN consists of six operations: sound-material organization, independent source analysis, feature normalisation, acoustic-graphic mapping, graphic notation language and multilayer temporal alignment (Figure 1).

3.1. Sound Materials and Data Organization

The sonic materials are composed of three types of sources: outdoor sound, birdsong and sounds of human activity. For the avian part, we selected recordings of 25 species that were classed as Song (S) to account for vocalisation category and avoid mixing Song recordings with alarm, contact, or other vocal behaviours.

Environmental noises were recorded from field recordings in four representative habitats in Chongming: farmland, reedbed, mudflat and woodland. Sounds of human activity included agricultural work, discussion, boats, intertidal foraging activity, footsteps and children's activities.

The features were retrieved separately from each source; the 3 source categories were not pre-mixed before to analysis. Here, "separate analysis" refers to independent processing of source files but with common analysis parameters and not to blind source separation. This reduces masking of dominating sources and preserves the temporal and spectral structure of each source before they are combined on a common notation timeline.

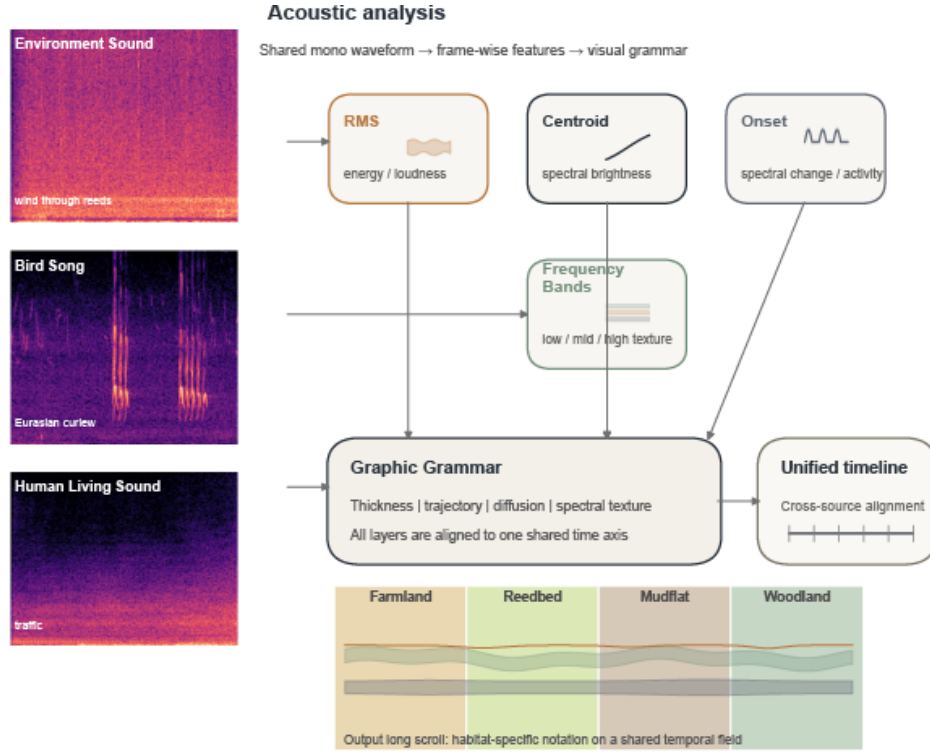


Figure 1. Overall framework of MESN acoustic analysis and graphic notation

3.2. Audio Preprocessing and Feature Extraction

All audio was read as mono and transformed into continuous float32 waveform arrays, preserving the original sample rate without resampling to a uniform sample rate. The analysis frame length was set to 4096 samples for sampling rates of 48 kHz or higher, 2048 samples for 22.05 to below 48 kHz and 1024 samples below 22.05 kHz to get a suitable balance between time and frequency resolution for recordings recorded at different rates. Librosa [11] was used to build the primary acoustic features. In the particular implementation used in this study, the temporal sampling-density parameter `line_density` was set at 1.50 which corresponding to an analysis frame rate of:

$$\text{fps} = \text{round}(32 \times 1.50) = 48 \quad (1)$$

The hop length was determined as:

$$H = \max\left(64, \left\lfloor \frac{L_{\text{sr}}}{48} \right\rfloor\right) \quad (2)$$

where `sr` is the original sample rate of each audio file. Under normal recording rates this corresponds to a temporal sampling density of ~48 frames/s; the hop length is `sr` - dependent.

All features were estimated on the basis of center alignment. The short-time Fourier transform (STFT) was performed with a `n_fft` equal to the adaptive frame length. Frequency-band energy was computed from the STFT power spectrum:

$$P(f, t) = |X(f, t)|^2 \quad (3)$$

This power spectrum was then utilised to compute low, middle and high frequency band energies. Since all features had the same hop length and center-alignment setting, their temporal indices could be directly aligned during the graphical mapping.

3.3. Multidimensional Acoustic Feature Extraction

The RMS Energy, Spectral Centroid, Onset Strength and Low/Mid/High Frequency Energy feature groups were processed independently. Separation refers to the analysis process, not statistical independence.

3.3.1. RMS Energy

RMS Energy was used to describe the overall energy of the signal over time:

$$\text{RMS}(t) = \sqrt{\frac{1}{N} \sum_{n=1}^N x_t[n]^2} \quad (4)$$

where `N` denotes the analysis frame length. The RMS sequence was first smoothed using a Savitzky-Golay filter with a window length of 9, followed by

$$R' = \log(1 + 24R) \quad (5)$$

for nonlinear compression. Then the compressed sequence was linearly transformed to [0, 1] using the 5th and 98th percentiles as lower and upper bounds respectively. In future mapping, RMS served as a relative-energy descriptor for within-recording variability rather than absolute loudness comparison across recordings.

3.3.2. Spectral Centroid

Spectral Centroid was used to characterize the center of the spectral energy distribution:

$$C(t) = \frac{\sum_k f_k |X(k, t)|}{\sum_k |X(k, t)|} \quad (6)$$

where `fk` is the frequency of the `k`th spectral bin. The raw sequence was smoothed with a Savitzky-Golay filter with window length 13, followed by:

$$C' = \log_2[\max(C, 1)] \quad (7)$$

and scaled to [0, 1] with the 4th and 96th percentiles as the lower and upper limits. To preserve both the global trend and

the local variance, the normalised sequence was further processed with two levels of Gaussian smoothing, with standard deviations determined as:

$$\sigma_{\text{macro}} = \max\left(2.2, \frac{N_f}{52}\right) \quad (8)$$

$$\sigma_{\text{macro}} = \max\left(1.0, \frac{N_f}{132}\right) \quad (9)$$

where N_f is the length of the feature sequence. The macro- and micro-smoothed sequences were combined with weights of 0.72 and 0.28, respectively, and interpolated at seven times the original sampling density using a CubicSpline with natural boundary conditions to produce a continuous spectral-centroid trajectory.

The Spectral Centroid and RMS were computed and normalised separately, so the graphical trace of the centroid was not directly driven by the RMS energy levels.

3.3.3. Onset Strength

We utilised Onset Strength to define the significance of spectral-energy shift between neighbouring time frames in characterising transitory activity and event boundaries.

Raw Onset Strength sequence was first upper-clipped at the 99.2nd percentile and then smoothed using a Savitzky-Golay filter with window length 5 followed by

$$O' = \log(1 + 80O) \quad (10)$$

for compression and scaling to $[0, 1]$ using the 5th and 98th percentiles as lower and upper bounds. To suppress the local numerical jitter while maintaining a somewhat fast transient variation, the succeeding smoothing parameter was set to:

$$\sigma_{\text{reactive}} = \max\left\{0.45, \frac{N_f}{280}\right\} \quad (11)$$

The sequence was then interpolated with four times the original sample density using a CubicSpline with natural boundaries.

In the graphical mapping used in this investigation, Onset Strength was employed to pinpoint syllable onsets, footfall, impacts, and other quick auditory changes. While RMS measures wider energy variation, Onset Strength considers changes over shorter time intervals.

The software still has an F0/Pitch analysis interface based on `librosa.pyin()`; but complex environmental recordings naturally contain a high degree of non-periodic or unstable

content, and so F0 was not employed as a core mapping feature for the final notation.

3.3.4. Frequency-Band Energy

The frequency band was separated into three analytical areas according to the power spectrum of STFT, low frequency (20–300 Hz), mid frequency (300–2000 Hz) and high frequency (>2000 Hz). Energy of frequency band b at time t was defined as:

$$E_b(t) = \sum_{f_k \in B_b} |X(k, t)|^2 \quad (12)$$

Each of the three band energy sequences were smoothed with a Savitzky-Golay filter with window length 9, followed by

$$E'_b(t) = \log(1 + E_b(t)) \quad (13)$$

for compression and independent scaling to $[0, 1]$ with the 4th and 98th percentiles as lower and upper bounds.

The three-band divide described the coarse spectral makeup inside each recording. It was not used as direct category classifier for source.

3.4. Mapping Acoustic Features to Graphical Variables

MESN explicitly maps acoustic features to traceable graphical variables (Figure 2).

Graphical thickness represents relative energy change for RMS Energy, vertical position of the trajectory for Spectral Centroid, local expansion and diffusion for Onset Strength, and continuous textures for the three frequency bands for Low/Mid/High Frequency Energy. These relations are defined by rules rather than learned from a latent-variable model.

Thus, the four groupings of features correspond to four analytical dimensions: relative energy, spectral location, transient activity and spectral composition.

In this mapping design, statistical independence of acoustic features is not assumed; instead, different graphical dimensions are prevented from being driven by the same energy parameter, so that substantial graphical changes may be tracked to the corresponding acoustic computations.

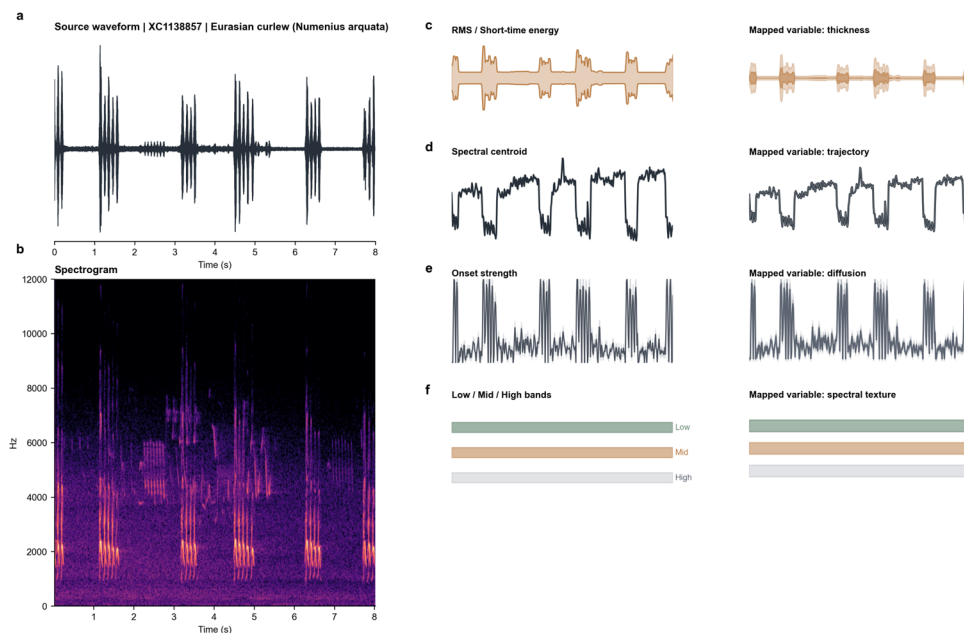


Figure 2. Interpretable mapping from acoustic features to graphical variables

3.5. Graphic Notation Grammar

Graphic primitives are assigned according to event continuity and discreteness (Figure 3).

Linear structures describe continuous sounds. Long lines or continuous bands depict persistent sound fields. Short lines and arcs represent shorter sounds with clear motion or rhythmic variation. Points represent discrete, instantaneous

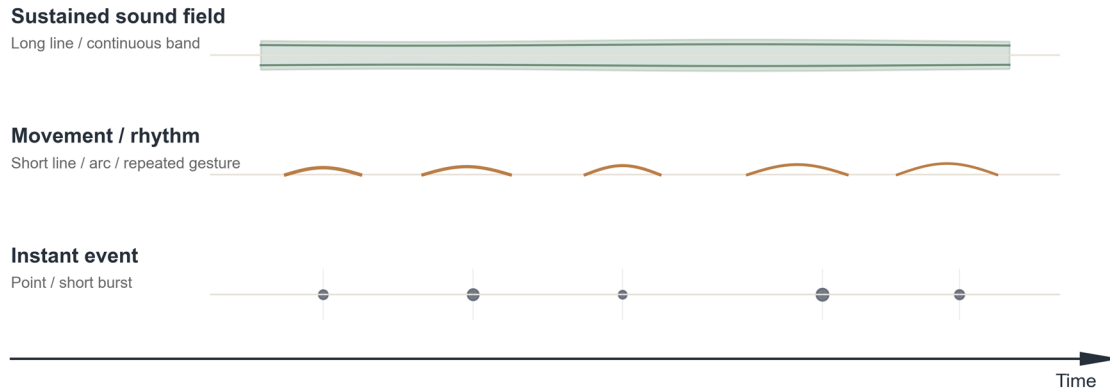


Figure 3. Point-, line-, and field-based graphic notation grammar along the time axis

3.6. Multilayer Temporal Alignment and Graphical Composition

The results of individual analyses and graphical representations of environmental sound, birdsong and human activity sounds are mapped into a single notation-time coordinate. The same horizontal coordinates mean the same point in the created notation timeline and various layers maintain continuity, gaps, overlap and interweaving.

If the source materials are independent recordings, the common timeline is a temporal alignment at the level of

sound events. The event detection is performed using a threshold of -42 dB and binary opening and closing operations for temporal consolidation. The minimum duration of the events is about 0.08 s and neighbouring events with time separation of less than about 0.12 s are combined. The images generated by this process are not direct tracings of the original waveform or spectrogram, but rather a rule-based re-encoding of temporal structure and acoustic attributes.

creative arrangement and notation, but not natural co-occurrence as documented.

4. Results and Application

4.1. Results of Acoustic Feature Mapping

In the 8 s birdsong sample shown in Figure 4, recurrent syllable clusters correspond to increasing RMS thickness and dense peaks in Onset Strength. The Spectral Centroid distinguishes vocal from silent intervals, and the three frequency bands keep a rough spectral composition.

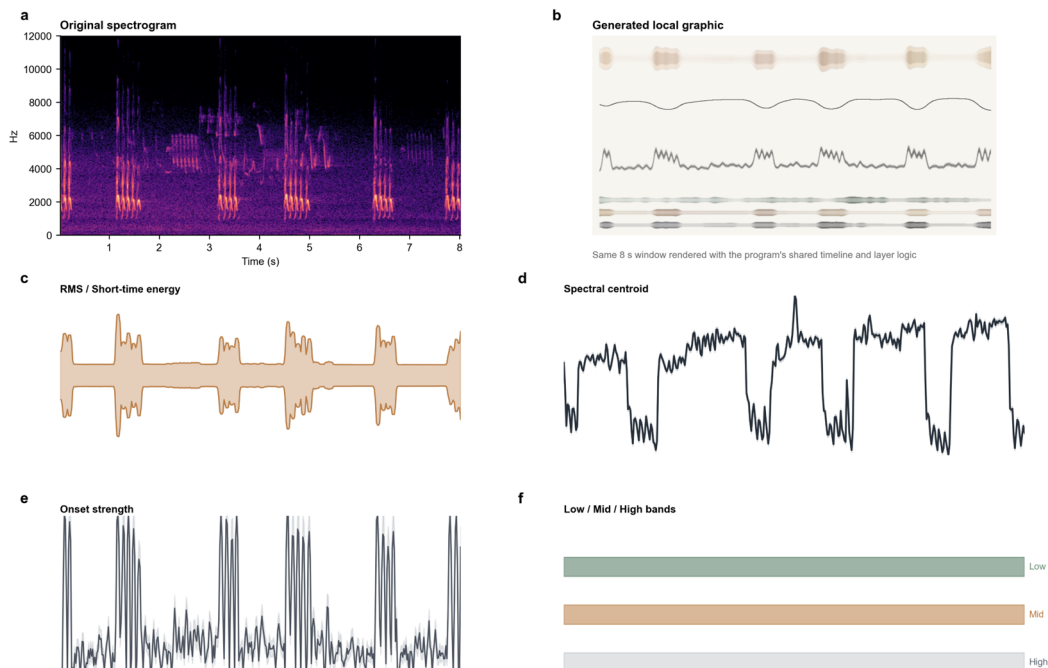


Figure 4. Acoustic analysis and local graphical generation from a single audio segment

Changes of major thickness and of direction and local expansions still correlate to the respective intervals of acoustic-feature variation. The four feature groups are computed and normalised individually, so local RMS peaks

do not directly influence centroid height, and onset peaks do not influence calculated frequency-band energy.

The research shows rule-level traceability across graphical variations, aural features and temporal positions; it does not

test species identity, ecological state or perceptual effectiveness.

4.2. Results of Multilayer Ecological Soundscape Graphic Notation

Environmental sound, birdsong and human activity were independently analysed within a 12 s notation frame (Figure 5). Representative shared locations are marked with vertical dashed lines. The ambient layer is a continuous background structure, the birdsong is an intermittent energy and fleeting

events. The human activity layer is formed by longer length segments.

The composite spectrogram at the top of Fig. 5 is solely provided as an aid to visualise temporal alignment and overlap; it does not contribute to the acoustic-feature calculation for any particular source layer. The composite result thus does not change feature values acquired from individual analysis. MESN preserves source-specific continuity, transience and spectral states while establishing cross-layer relationships using the common horizontal coordinate.

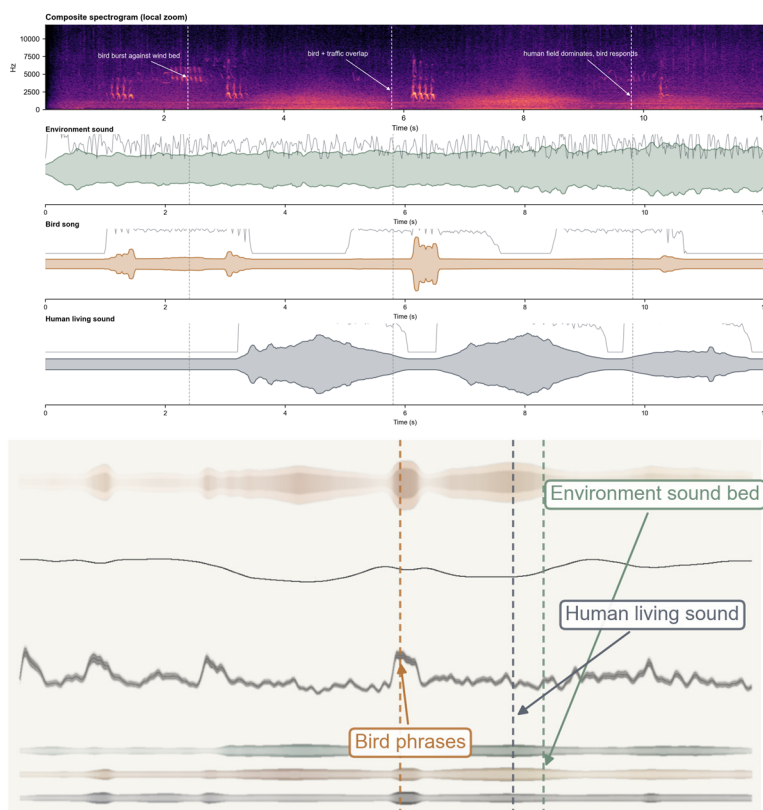


Figure 5. Multi-source temporal alignment and composite graphic notation

4.3. Matched-Scale Comparison Across Four Habitat Samples

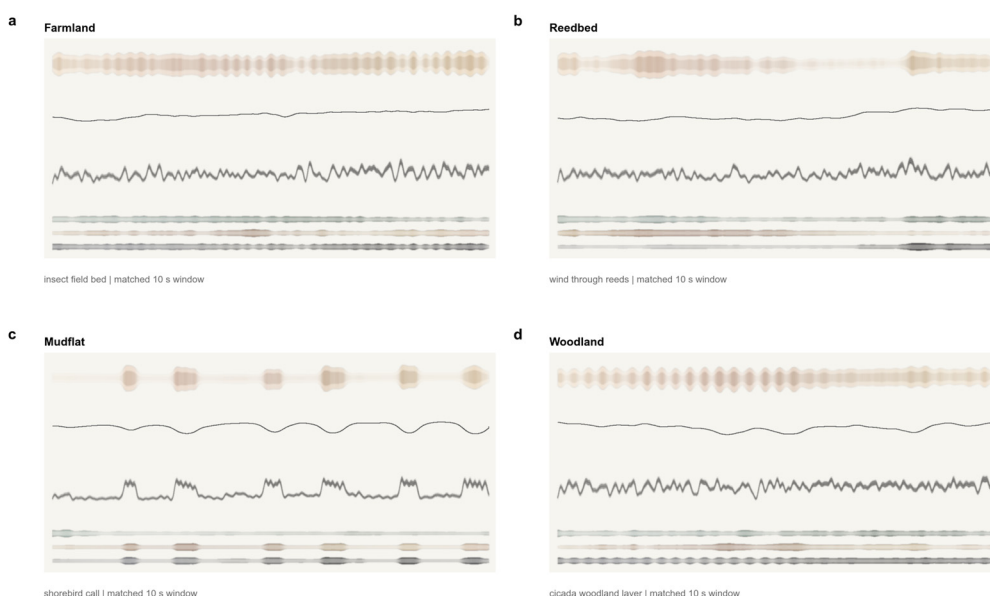


Figure 6. Matched-scale comparison of generated results across four habitat samples

Figure 6 shows the 10 s farm, reedbed, mudflat, and woodland samples with the same time window, generating parameters, and display scale. The examples illustrated range from farmland as dense small scale fluctuations over a continuous background to reedbed as longer sustained structures with localised variation, mudflat as more discrete event groups separated by clear intervals and woodland as a denser sequence of continuous transient activity. These variances result in unique combinations of RMS thickness, centroid trajectory, onset activity, and frequency-band texture. The observed graphical discrepancies are due to the acoustic temporal structures of the input samples for a certain rule set. They are not statistical habitat signatures or markers of ecological quality, but rather exemplary qualitative examples.

MESN creates explicit links between acoustic parameters and graphic variables, such as thickness, trajectory, diffusion and texture of the frequency-band. The interpretability relies on the traceability of the mapping rules and data sources. MESN is a documented depiction of the acoustic structure rather than a typical system of performance notation. This is achieved by translating the recorded sound through a sound-feature-graphic process.

Individual analysis keeps good layers of environment, birds and humans before their merger in a common chronology. Recordings are normalised independently, so graphical amplitudes are not absolute representations of changes in loudness, and aligned timelines do not always mean that events occurred in synchrony.

Limitations of the method are the recording settings, coarse segmentation into frequency bands and smoothing methods which may lose detail in the near term. The habitat types and 10 s examples are illustrative, not statistically significant. Further quantitative and perceptual confirmation is thus needed.

To summarise, MESN integrates separate source analysis, explicit feature mapping and visual notation in a common temporal framework. Applied in Symbiotic Habitat it gives a traceable and reproducible way of representing ecological soundscape.

References

- [1] Pijanowski, B. C., Villanueva-Rivera, L. J., Dumyahn, S. L., & et al. (2011). Soundscape ecology: The science of sound in the landscape. *BioScience*, 61(3), 203-216. <https://doi.org/10.1525/bio.2011.61.3.6>.
- [2] Sueur, J., & Farina, A. (2015). Ecoacoustics: The ecological investigation and interpretation of environmental sound. *Biosemiotics*, 8(3), 493-502. <https://doi.org/10.1007/s12304-015-9248-x>.
- [3] Gasc, A., Francomano, D., Dunning, J. B., & Pijanowski, B. C. (2017). Future directions for soundscape ecology: The importance of ornithological contributions. *The Auk: Ornithological Advances*, 134(1), 215-228. <https://doi.org/10.1642/AUK-16-124.1>.
- [4] Sethi, S. S., Bick, A., Ewers, R. M., & et al. (2023). Limits to the accurate and generalizable use of soundscapes to monitor biodiversity. *Nature Ecology & Evolution*, 7, 1373-1378. <https://doi.org/10.1038/s41559-023-02148-z>.
- [5] Koch, T. M. I., Marks, E. S., & Roberts, T. F. (2025). A deep learning approach for the analysis of birdsong. *eLife*, 13, RP101111. <https://doi.org/10.7554/eLife.101111.3>.
- [6] Shim, J. Y., Kim, J., & Kim, J. K. (2023). Audio-to-visual cross-modal generation of birds. *IEEE Access*, 11, 27719-27729. <https://doi.org/10.1109/ACCESS.2023.3257565>.
- [7] Kim, S. B., Senocak, A., Ha, H., & et al. (2023). Sound to visual scene generation by audio-to-visual latent alignment. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6430-6440. <https://doi.org/10.1109/CVPR52729.2023.00622>.
- [8] Finke, G. (2019). Partituren zum Lesen und Schauen: Bildlichkeit als Merkmal graphischer Notation. *Zeitschrift der Gesellschaft für Musiktheorie*, 16(1), 21-39. <https://doi.org/10.31751/1001>.
- [9] Hope, C., Vickery, L., Wyatt, A., & James, S. (2015). The Decibel ScorePlayer: A digital tool for reading graphic notation. *Proceedings of the First International Conference on Technologies for Music Notation and Representation (TENOR 2015)*, 58-69.
- [10] Karchkhadze, T., Shao, K., & Dubnov, S. (2024). Interpreting graphic notation with MusicLDM: An AI improvisation of Cornelius Cardew's Treatise. *Proceedings of the 2024 IEEE International Conference on Big Data (BigData)*, 3181-3190. <https://doi.org/10.1109/BigData62323.2024.10825824>.
- [11] McFee, B., Raffel, C., Liang, D., & et al. (2015). librosa: Audio and music signal analysis in Python. *Proceedings of the 14th Python in Science Conference*, 18-24. <https://doi.org/10.25080/Majora-7b98e3ed-003>.