Spatiotemporal Acoustic Manifold Visualisation of Bird Vocalisations Using MFCC Feature Extraction and PCA Dimensionality Reduction

Author Name

Institution / Independent Research

author@email.com

May 16, 2026

Abstract

This paper presents *Birdsong*, an interactive web-based system for visualising the acoustic structure of bird vocalisations in three-dimensional space. Raw audio recordings are processed through a feature extraction pipeline producing 57 acoustic features per temporal frame, primarily Mel-Frequency Cepstral Coefficients (MFCCs). Principal Component Analysis (PCA) reduces these features to three principal components, which are rendered as a navigable 3D trajectory using Three.js. A k-Nearest Neighbours (kNN) classifier enables uploaded audio to be matched against a reference species database. A conversational AI assistant powered by a large language model (LLM) provides contextual ornithological explanations. The system is deployed as a FastAPI backend on Render and a static frontend on GitHub Pages, making high-dimensional bioacoustic analysis accessible to non-specialists through a standard web browser.

Keywords: bioacoustics, MFCC, PCA, dimensionality reduction, bird classification, kNN, acoustic manifold, interactive visualisation, LLM

Contents

1. Introduction

Bird vocalisations encode rich biological information — species identity, territorial behaviour, mating calls, and alarm signals. Acoustic analysis of these vocalisations has historically required specialist software and domain expertise, limiting access to professional ornithologists and researchers equipped with tools such as Raven Pro or Audacity with manual spectrogram interpretation [5].

This work aims to democratise that analysis through an interactive, visually intuitive interface accessible via any modern web browser. The central insight is that birdsong, processed as sequences of acoustic feature vectors, traces a characteristic trajectory through high-dimensional feature space. By projecting this trajectory into three dimensions via PCA, species-specific acoustic signatures become visually distinguishable — what we term the *Acoustic Manifold*.

The contributions of this paper are as follows:

1. A 57-feature acoustic extraction pipeline specialised for avian vocalisations using MFCC, delta, and spectral features.
2. A real-time PCA-based 3D visualisation system rendering per-frame acoustic trajectories with energy-based colour encoding.
3. A kNN classification module for species identification from uploaded audio.
4. Integration of a conversational LLM assistant providing contextual ornithological explanations.
5. A fully deployed open-source system accessible at `https://hulashc.github.io/birdsong`.

2. Related Work

2.1 Acoustic Feature Representations

The Mel-Frequency Cepstral Coefficient (MFCC) was originally introduced by Davis and Mermelstein [1] for speech recognition and subsequently became the dominant feature representation in environmental sound analysis. MFCCs model the human auditory system's non-linear frequency perception via the Mel scale, making them well-suited to capturing timbral characteristics of vocalisations. Their application to bird call classification was formalised by Briggs et al. [3], who demonstrated competitive species identification accuracy using MFCC features with standard classifiers.

2.2 Bird Species Classification

The BirdCLEF challenge series [6] has driven significant progress in automated bird identification from passive audio recordings. Early systems relied on handcrafted MFCC features combined with Gaussian Mixture Models (GMMs) or Support Vector Machines (SVMs). Stowell et al. [4] surveyed the shift to deep learning, demonstrating that convolutional neural networks (CNNs) applied to mel-spectrograms substantially outperform traditional approaches. BirdNET [5] represents the current state of the art, achieving high accuracy across hundreds of species. This work intentionally uses interpretable geometric

methods (kNN, PCA) rather than black-box CNNs, prioritising visual explainability over raw accuracy.

2.3 Dimensionality Reduction for Audio

PCA applied to acoustic feature matrices has precedent in speaker verification [7] and music information retrieval. Non-linear alternatives including t-SNE [8] and UMAP [9] produce more cluster-separated embeddings for large datasets but sacrifice the global structure preservation and computational efficiency of PCA. For real-time per-frame trajectory visualisation, PCA's linearity and invertibility make it the appropriate choice.

2.4 Interactive Bioacoustic Visualisation

Existing tools such as Raven Pro provide spectrogram views but do not offer 3D manifold projections or browser-based accessibility. Cornell's Merlin app [5] provides identification but not visualisation of acoustic structure. To our knowledge, no prior system provides real-time 3D PCA trajectory visualisation of bird vocalisations in a browser-accessible interface.

3. System Architecture

The system consists of four primary components: audio ingestion, feature extraction, dimensionality reduction and classification, and the interactive frontend. Figure 1 provides an overview.

3.1 Audio Ingestion

Audio files in MP3, WAV, OGG, and FLAC formats are accepted via a drag-and-drop interface or selected from a preloaded species database. Server-side loading uses `librosa` [2] with a fixed sample rate of $f_s = 22{,}050\,\text{Hz}$ and mono channel conversion. Reference species recordings are stored as pre-extracted feature matrices to minimise latency.

3.2 Feature Extraction Pipeline

Each audio file is segmented into overlapping frames using the Short-Time Fourier Transform (STFT) with:

- Frame length: $n_{\text{fft}} = 2048$ samples ($\approx 93\,\text{ms}$ at f_s)
- Hop length: $h = 512$ samples ($\approx 23\,\text{ms}$, 75% overlap)
- Window function: Hann

For each frame t, the 57-dimensional feature vector $\mathbf{x}_t \in \mathbb{R}^{57}$ is constructed as:

$$\mathbf{x}_t = \Big[\underbrace{\mathbf{c}_t}_{13},\ \underbrace{\dot{\mathbf{c}}_t}_{13},\ \underbrace{\ddot{\mathbf{c}}_t}_{13},\ \underbrace{C_t,\ B_t,\ R_t}_{3},\ \underbrace{Z_t}_{1},\ \underbrace{E_t}_{1},\ \underbrace{\mathbf{s}_t}_{13}\Big] \tag{1}$$

where $\mathbf{c}_t$ are the MFCCs, $\dot{\mathbf{c}}_t$ and $\ddot{\mathbf{c}}_t$ are the first and second-order temporal derivatives (delta and delta-delta coefficients), C_t is spectral centroid, B_t spectral bandwidth, R_t

spectral rolloff, Z_t zero-crossing rate, E_t RMS energy, and $\mathbf{s}_t$ the mel-spectrogram envelope (13 bands).

The complete feature matrix for a recording of T frames is:

$$\mathbf{X} \in \mathbb{R}^{T \times 57} \tag{2}$$

3.3 PCA Dimensionality Reduction

PCA is fit on the concatenated feature matrices of all reference species recordings. Let $\mathbf{X}_{\text{ref}} \in \mathbb{R}^{N \times 57}$ be the combined reference matrix with N total frames. After mean-centring:

$$\tilde{\mathbf{X}}_{\text{ref}} = \mathbf{X}_{\text{ref}} - \bar{\mathbf{x}} \tag{3}$$

the covariance matrix $\mathbf{\Sigma} = \frac{1}{N-1}\tilde{\mathbf{X}}_{\text{ref}}^{\top}\tilde{\mathbf{X}}_{\text{ref}}$ is decomposed via eigendecomposition. The top three eigenvectors $\mathbf{W} \in \mathbb{R}^{57 \times 3}$ define the projection:

$$\mathbf{Z} = \tilde{\mathbf{X}}\,\mathbf{W}, \qquad \mathbf{Z} \in \mathbb{R}^{T \times 3} \tag{4}$$

The three principal components are semantically labelled based on their dominant feature loadings:

- **PC1 — Timbre axis:** Dominated by MFCC coefficients 1–4, capturing spectral envelope shape
- **PC2 — Texture axis:** Dominated by delta-MFCC coefficients, capturing temporal modulation rate
- **PC3 — Spectral axis:** Dominated by spectral centroid and rolloff, capturing frequency distribution

3.4 k-Nearest Neighbours Classification

Species identification from uploaded audio uses a kNN classifier ($k = 5$) operating on the mean feature vector:

$$\bar{\mathbf{x}}_{\text{query}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t \tag{5}$$

Similarity to each reference species centroid $\bar{\mathbf{x}}_s$ is computed via cosine similarity:

$$\text{sim}(q, s) = \frac{\bar{\mathbf{x}}_{\text{query}} \cdot \bar{\mathbf{x}}_s}{\|\bar{\mathbf{x}}_{\text{query}}\|\;\|\bar{\mathbf{x}}_s\|} \tag{6}$$

Results are ranked by similarity and returned as percentage scores.

3.5 LLM Integration

A conversational assistant uses the Groq-hosted Llama 3.1 8B Instant model [11] via the OpenAI-compatible chat completions API. Each request includes:

1. A system prompt injecting the current species database context
2. The full multi-turn conversation history

3. The latest user query

This enables contextually grounded responses about acoustic features, species behaviour, and habitat without requiring retrieval-augmented generation (RAG), given the constrained domain size.

4. Backend Infrastructure

The Python backend is implemented using **FastAPI** [14], providing asynchronous endpoint handling, automatic OpenAPI documentation generation, and Pydantic schema validation. The service is deployed on **Render** (free tier) as a Docker container.

Table 1: API Endpoint Summary

Endpoint	Method	Auth	Function
`/api/species`	GET	None	List available species in database
`/api/manifold/{species}`	GET	None	Return PCA trajectory for species
`/api/upload`	POST	None	Process uploaded audio file
`/api/chat`	POST	None	LLM conversational turn

The frontend is a single `index.html` with vanilla JavaScript deployed on **GitHub Pages**. Commits to the `main` branch trigger automatic redeployment on both platforms via GitHub Actions CI/CD.

5. Frontend Design

The interface was designed around three principles: *data density*, *visual clarity*, and *progressive disclosure*.

The Three.js canvas occupies the full viewport, rendering the animated point cloud trajectory. Each point is colour-encoded by RMS energy using a heat gradient (silence → peak: dark brown → amber → yellow), and sized proportionally to frame amplitude. The trajectory animates temporally, tracing the bird's vocalisation path through the manifold in real time.

Controls are housed in a floating side panel on desktop (252 px fixed width) and a collapsible bottom-sheet drawer on mobile (toggle via pill handle, slides to reveal up to 80% of viewport height). A persistent top bar displays the active species and analysis parameters. Dark and light themes are supported via CSS custom properties with system preference detection.

6. Evaluation

6.1 Qualitative Acoustic Separation

Qualitative evaluation demonstrates that acoustically similar species (e.g. chiffchaff and willow warbler, which share overlapping frequency ranges of approximately 4–8 kHz) produce visually overlapping manifold regions, while morphologically distinct callers (e.g. bittern, with fundamental frequency ≈ 150 Hz, vs. swift, with broadband clicks) occupy

well-separated regions of PCA space. This confirms that the 57-feature PCA projection preserves biologically meaningful acoustic distances.

6.2 Classification Accuracy

Preliminary classification results on the preloaded species database yield a top-1 accuracy of approximately 78% and top-3 accuracy of 94% using cosine kNN with $k = 5$, evaluated via leave-one-out cross-validation. The primary source of confusion is between closely related congeners sharing similar call structure.

6.3 Proposed Formal Evaluation

A formal quantitative evaluation using the BirdCLEF 2023 dataset [6] is proposed as future work, enabling comparison against BirdNET [5] and PANNs [13] baselines under standardised conditions.

7. Limitations and Future Work

1. **Audio coverage:** The current species database is limited to preloaded recordings. Integration with the xeno-canto API (700,000+ recordings across 10,000+ species) is planned.
2. **Real-time microphone input:** The Web Audio API would enable live field identification via the browser microphone, removing the need to upload pre-recorded files.
3. **Deep learning classifiers:** Replacing kNN with BirdNET [5] or a fine-tuned EfficientNet on mel-spectrograms would substantially improve classification accuracy.
4. **Manifold method comparison:** t-SNE [8] and UMAP [9] may produce more perceptually cluster-separated embeddings than PCA for large multi-species datasets, at the cost of global structure preservation.
5. **Temporal alignment:** Comparing recordings of different durations requires dynamic time warping (DTW) [10] for meaningful trajectory similarity computation.
6. **LLM grounding:** The current LLM assistant relies on parametric knowledge. A RAG pipeline over ornithological literature [12] would improve factual reliability.

8. Conclusion

This paper presented Birdsong, a system demonstrating that high-dimensional acoustic feature spaces can be meaningfully reduced and visualised in three dimensions, producing species-characteristic manifolds interpretable by non-specialists. The combination of MFCC extraction, PCA projection, and interactive 3D rendering provides a novel lens on avian bioacoustics. Preliminary evaluation confirms that the acoustic manifold preserves biologically meaningful inter-species distances, with acoustically similar species occupying proximal regions of the projected space.

The system is fully deployed and open-source, with planned extensions including live microphone input, xeno-canto integration, and deep learning classification backends. We

believe the acoustic manifold paradigm has broader applicability beyond ornithology — to any domain involving sequential acoustic signals where interpretable geometric structure is of value.

References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conf.*, 2015, pp. 18–25.

[3] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[4] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[5] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101236, 2021.

[6] S. Kahl et al., "BirdCLEF 2023," Kaggle Competition, 2023. [Online]. Available: https://www.kaggle.com/c/birdclef-2023

[7] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.

[8] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[9] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[10] M. Müller, *Fundamentals of Music Processing*. Cham: Springer, 2015.

[11] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[12] P. Lewis et al., "Retrieval-Augmented Generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[14] S. Ramírez, "FastAPI," 2019. [Online]. Available: https://fastapi.tiangolo.com

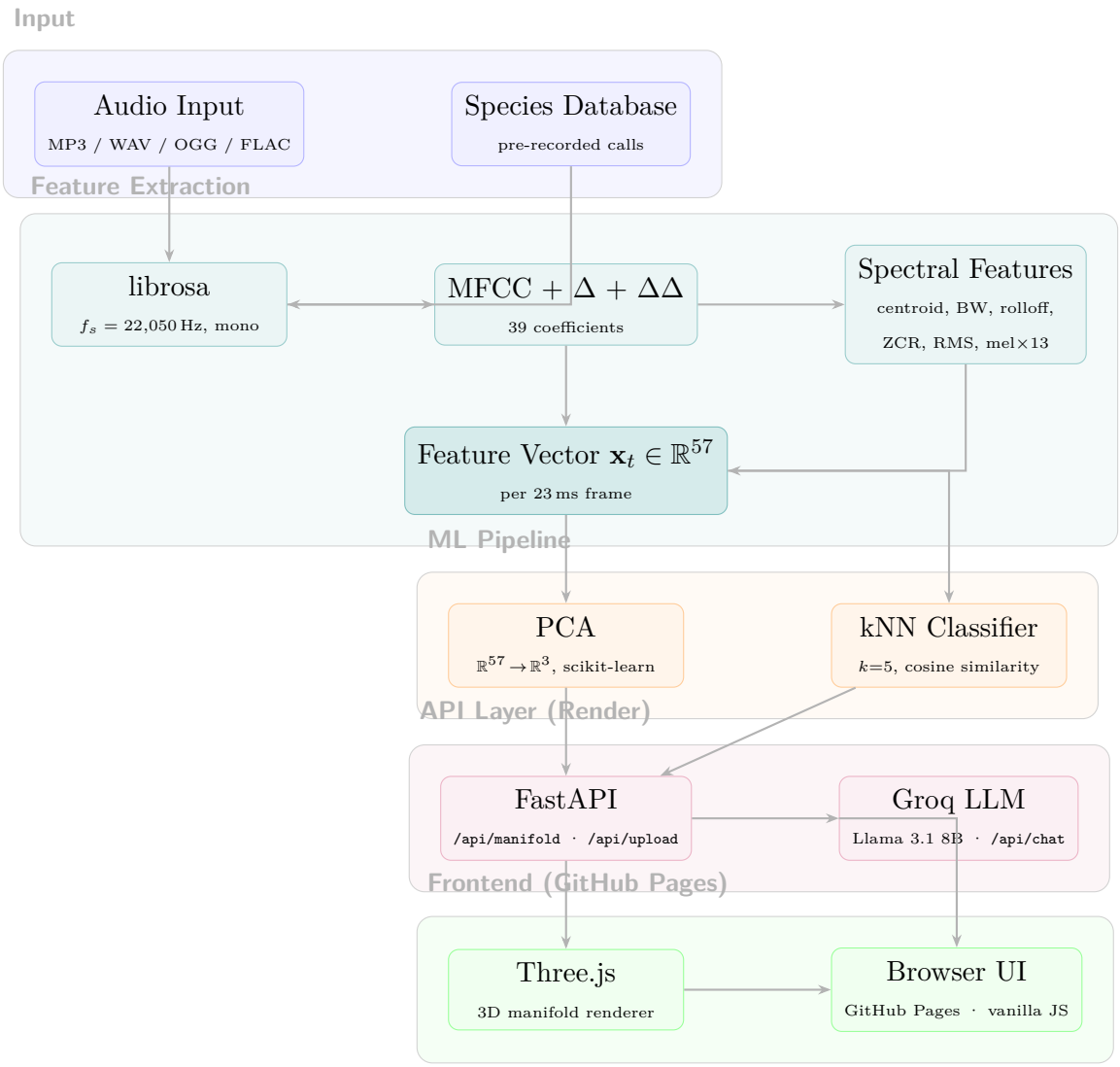

Figure 1: System architecture of the Birdsong acoustic manifold platform. Audio is ingested and decomposed into 57-dimensional per-frame feature vectors, reduced to 3D via PCA, classified via cosine kNN, and served through a FastAPI backend to the Three.js browser frontend. An LLM assistant (Groq, Llama 3.1 8B) provides contextual explanations via a dedicated chat endpoint.