COMP 1018P Audio EDA Report

Anthony Miyaguchi

October 28, 2025

Overview

In this report, we explore the use of domain-specific birdcall classifiers trained on the BirdSet dataset to soundscape recordings from the BirdCLEF+ 2025 competition and the MBARI Pacific Sounds dataset. We hypothesize that we can extract unsupervised signals from soundscape embeddings that can be used to identify interesting events (animal calls) in the recordings. We also explore whether we can zero-shot transfer the birdcall classifier to underwater soundscapes to identify interesting events. We find that our unsupervised methods extract unmeaningful noise from the embeddings. Future work should include parametric methods for audio representation, where features are more interpretable and controllable.

Datasets

Links

- BirdCLEF+ 2025 Competition Dataset
- Pacific Sound - AWS Open Data Registry
- MBARI Soundscape Listening Room

The BirdCLEF+ 2025 dataset contains nearly 10,000 minutes of unlabeled soundscape recordings from El Silencio Natural Reserve in Colombia. Each soundscape is one minute long and sampled at 44.1 kHz. These recordings are captured with autonomous recording units (ARUs) deployed in the field, and capture birds, insects, amphibians, and mammals in their natural habitats. This data is helpful for conservation efforts by helping effectively build an acoustic census of the area.

The MBARI Pacific Sounds dataset contains almost 10 years of underwater soundscape recordings from Monterey Bay, California. The recordings are captured with hydrophones deployed on the seafloor sampled at 256kHz in 10-minute segments. The listening room has the audio processed to listen at a 20 minute delay. Animals like whales, dolphins, sea lions, and boats can be heard in the recordings. This data is helpful for monitoring the Monterey Bay ecosystem, and tracking both animal populations and human activity in the area.

Experiments and Exploration

Links

- OpenSoundscape Documentation
- Bioacoustics Model Zoo
- Distilling Spectrograms into Tokens: Fast and Lightweight Bioacoustic Classification for BirdCLEF+ 2025
- BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics
- STUMPY Documentation
- Matrix Profile VI: Meaningful Multidimensional Motif Discovery

We build out a pipeline to extract audio embeddings from the soundscape recordings and then apply motif discovery over a low-dimensional projection of the embeddings. The embeddings are treated as a multivariate time series where the assumption is that a sliding window over audio embeddings will capture meaningful, repeating patterns in the audio. We use the STUMPY library to compute the matrix profile over the embedding time series, and extract motifs and discords that we evaluate qualitatively by listening to the audio segments. We project the high-dimensional embeddings from 1280 dimensions down to 16 dimensions to make the matrix profile computation more efficient, while preserving the structure of the data.

We use the `BirdSetEfficientNetB1` model from Bioacoustics Model Zoo trained on the BirdSet dataset. This model uses a CNN backbone (EfficientNetB1) to classify 9,736 bird species from spectrogram inputs using a 1280-dimensional embedding layer over a 5-second audio window. This model achieves comparable performance to Perch and BirdNET, while only relying on PyTorch (and thus being much easier to install and use). The api is fairly straightfoward, resulting in the following schemas after processing to convert embeddings into dense vectors after sliding half-second window inference:

```
embed
 |-- file: string (nullable = true)
 |-- start_time: double (nullable = true)
 |-- end_time: double (nullable = true)
 |-- embedding: array (nullable = true)
 |    |-- element: double (containsNull = true)

predict
 |-- file: string (nullable = true)
 |-- start_time: double (nullable = true)
 |-- end_time: double (nullable = true)
 |-- predictions: array (nullable = false)
 |    |-- element: float (containsNull = true)
```

The rows look like this:

file	start_time	end_time	embedding
H04_20230501_134500	0.0	5.0	[-0.1142078265547...
H04_20230501_134500	0.5	5.5	[-0.1127882376313...
H04_20230501_134500	1.0	6.0	[-0.0942024067044...
H04_20230501_134500	1.5	6.5	[-0.0891242176294...

It's worth noting that while converting the arrow/parquet representations of the dense embeddings is worthwhile due to the downstream linear algebra operations, the sparsity patterns of the predictions makes a dense representation memory inefficient and hard to work with for large datasets. In this work, we only work with embeddings, but the predictions themselves are *very* relevant for the downstream task of identifying bird calls in soundscape recordings.

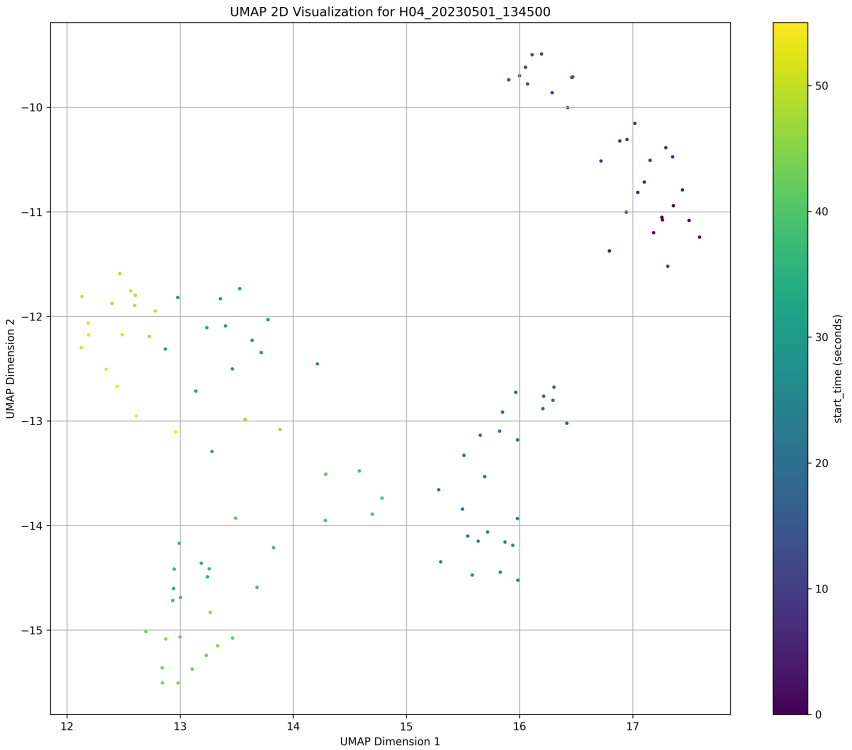

Figure 1: Scatterplot the the embedding vectors over time via UMAP.

Our analysis is easier when we have fewer dimensions to work with. One of the first things that we do is apply UMAP to visualize the embedding vectors over time in two dimensions. We color each point by the time index of the embedding

in the soundscape recording, and see the points cluster in different regions of the embedding space over time. After visualizing the data, we apply Principal Component Analysis (PCA) to reduce the dimensionality of the embedding vectors. We create a scree plot to identify the number of dimensions it takes to explain 80% of the variance in the data. We find that 16 dimensions is sufficient to explain 91.4% of the variance.

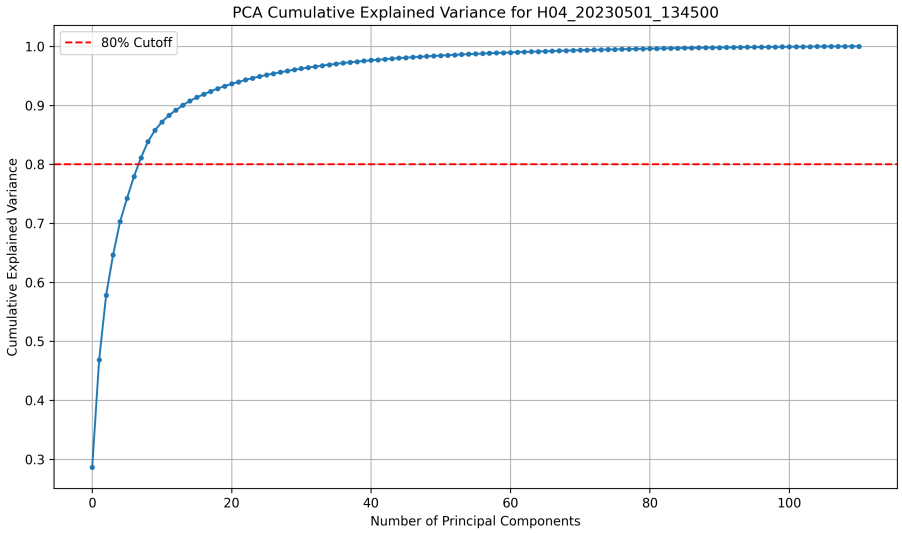

Figure 2: A scree plot demonstrating the explained cumulative variance via PCA.

We can treat this data as a histogram, where every bin in a step corresponds to a dimension in the rotated embedding space. We run the multi-dimensional matrix profile algorithm to try to identify motifs in the data. The matrix profile works by computing a distance profile between all subsequences of a given length in a time series in an efficient way. The resulting matrix profile gives you the nearest neighbor for every subsequence in the time series, and finding the minimum distance in this matrix profile allows you to identify motifs (repeating patterns) in the data. The multidimensional variant (mstump) extends this to multiple dimensions by computing the distance profiles in each dimension and aggregating them by taking advantage of the non-negativity of the z-normalized Euclidean distance. We use a motif length of 20 time steps (10 seconds) for our experiments.

We can visualize the location of the motif individually in each PCA dimension. The reason behind the location of the motif is unclear by quick visual inspection. Note that the PCA (and SVD underlying PCA) produces a basis that captures frequency-like components, with lower dimensions capturing low frequency trends and higher dimensions capturing high frequency trends.

The final thing to do is to listen to the audio segments corresponding to the

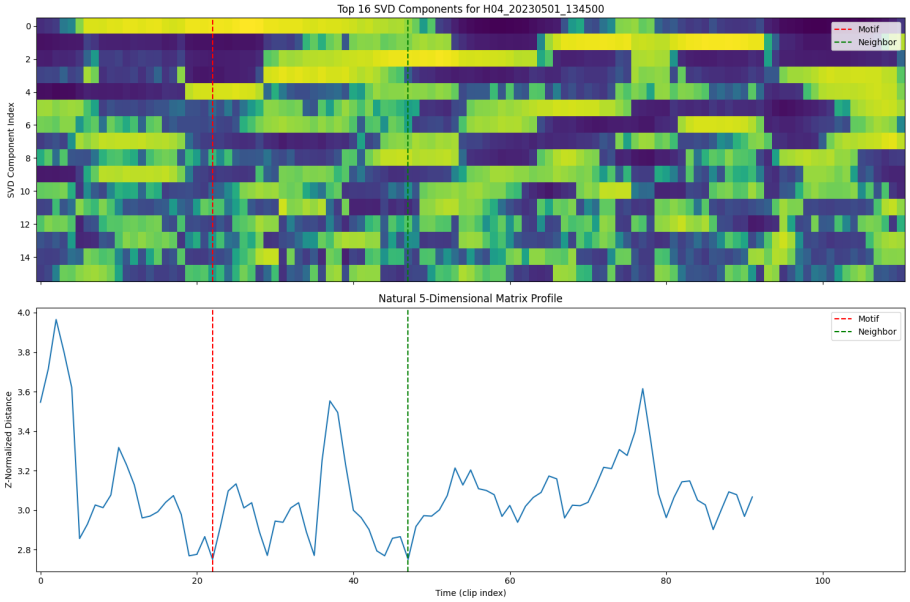

Figure 3: The extracted motif on the PCA-truncated embedding time series. The top plot shows the heatmap of the series, while the bottom shows the matrix profile and the motifs.

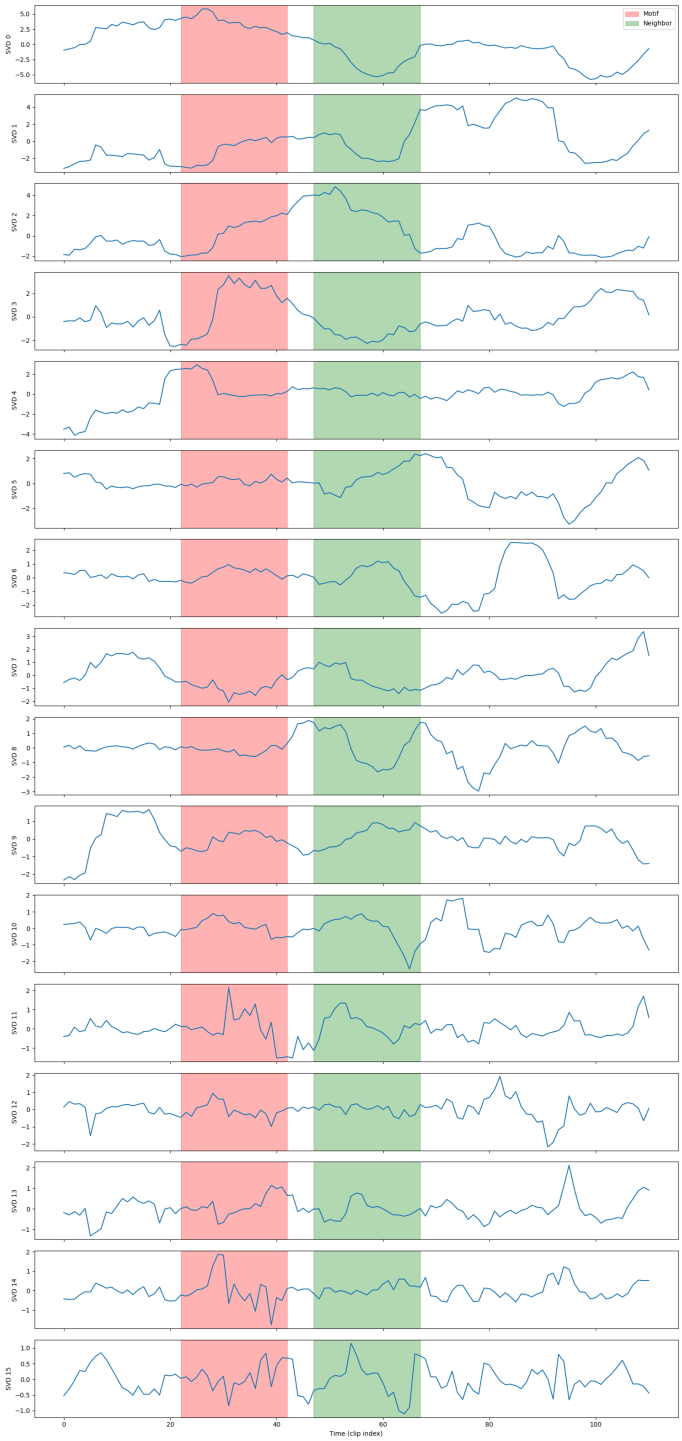

Figure 4: Motif location in each PCA dimension.

motifs. We found that all of the discovered motifs corresponded to background noise or silence in the recordings. We also listened to discords (anomalies) discovered by the matrix profile, and found that these also did not correspond to any meaningful bird calls or animal sounds. We vary the motif length to see if this affects the results, and we find a similar outcome when we use a window of 5 steps (2.5 seconds).

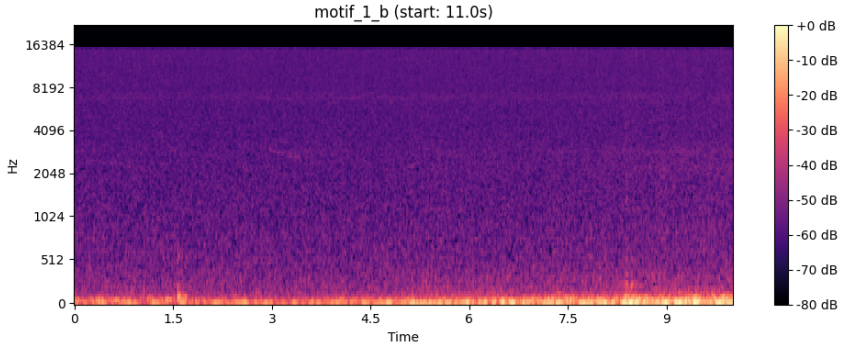

Figure 5: A motif spectrogram that sounds like background noise.

We apply the same procedure to the MBARI Pacific Sounds dataset, and find similar results where the discovered motifs correspond to background noise or silence in the recordings. Unfortunately, our unsupervised approach did not yield meaningful patterns corresponding to calls or events of interest in the soundscape recordings.

Discussion and Future Work

One of the reasons why this approach may not have worked is because the birdcall classifier is effectively smoothing the signal over a 5-second window, where-as the events we care about are likely occuring over shorter time-scales within that window. If the classifier worked on half-second windows, we might have been able to capture more meaningful patterns from the data by effectively treating the embeddings as the primary representation of the audio at the potential cost of classification performance on the time-scales needed for birdcall identification. Additionally, our selection of a 10-second motif length may have biased the results toward longer-term stationary patterns in the data. The better approach here would be to use a parametric representation i.e. the short-time Fourier transform (STFT) or Mel-frequency cepstral coefficients (MFCCs) as the primary representation of the audio, and then apply unsupervised methods on that representation directly. We have more control over these windows, and we can interpret the features more easily. This is also the representation that most of the deep-learning bioacoustic models are trained on.

Another reason why this approach may not have worked is that the represen-

tation space might be good for classifying bird calls, but not necessarily for keeping consistent representations of sounds close together over time. The ideal representation space for this would have a smooth trajectory over time for similar sounds, and a mapping of the high dimensional time series to a consistent low-dimensional manifold. There's no strong reason to believe that SVD on the embedding space would yield a good representation for motif discovery, other than the empirical effectiveness of transfer learning for classification tasks. In fact, by retaining the axes of maximum variance, we might be losing locally relevant features that are not globally significant. Low-frequency components that don't contribute much to infrequent bird call events might be overemphasized in the PCA representation.

Also, instead of using matrix profiles for motif discover, we could have just used K-NN search using the probability that the embedding is a bird call to fetch similar sounding segments of interest. This is often used in "active learning" approaches, where a human annotator is effectively in a semi-supervised loop with the model in order to build classifiers on new data.

Other things that could have been tried here involves analyzing the prediction vectors instead of the embeddings. There vectors are very sparse, but they are *very* likely to include useful locations of bird calls because this is what they were trained to predict. We could simply sum over all predictions to get a probability of bird presence over time, and extract these regions for further analysis. With fragments of bird calls, you can apply clustering or similarity search to identify unique calls or the frequency of these calls over time. There is a call-density estimation method used with the Perch model that attempts to probablistically model and characterize soundscapes by the density of bird calls over time, and this might be an interesting method to explore with the prediction data itself.

Finally, although we applied our pipeline directly to underwater soundscapes, its unclear if there is any effective transfer learning in these contexts. Prior art says that bird call classifiers are effective, but this requires a supervised finetuning step and some careful analysis on the soundscape data to implement in any reasonable way. This would be far outside the scope of a simple EDA in a two-week period without clear direction and data.

Conclusion

In this report, we looked at a complex pipeline to extract unsupervised signals from soundscapes using birdcall classifier embeddings and motif discovery via matrix profiles. We found that this approach was not effective, likely due to relevant signals being oversmoothed by the classifer, and the embedding space not being ideal for motif discovery. If we wanted to work further on matrix profiles, we would be better off using MFCCs or some other spectrogram-based representation that we have more control over. An interesting direction to take this for further work would be to apply some kind of active-learning approach by looking at the prediction vectors. Additionally, further work is needed to see

how birdcall classifiers could be used in a marine audio context.