# Joint time–frequency scattering-enhanced representation for bird vocalization classification

**Yimeng Min**
Department of Computer Science
Cornell University
Ithaca, New York, USA
min@cs.cornell.edu

**Eliot T. Miller**
Lab of Ornithology
Cornell University
Ithaca, New York, USA
etm45@cornell.edu

**Daniel Fink**
Lab of Ornithology
Cornell University
Ithaca, New York, USA
daniel.fink@cornell.edu

**Carla P. Gomes**
Department of Computer Science
Cornell University
Ithaca, New York, USA
gomes@cs.cornell.edu

## Abstract

Neural Networks (NNs) have been widely used in passive acoustic monitoring. Typically, audio is converted into a Mel Spectrogram as a preprocessing step before being fed into NNs. In this study, we investigate the Joint Time-Frequency Scattering transform as an alternative preprocessing technique for analyzing bird vocalizations. We highlight its superiority over the Mel Spectrogram because it captures intricate time-frequency patterns and emphasizes rapid signal transitions. While the Mel Spectrogram often gives similar importance to all sounds, the scattering transform differentiates between rapid and slow variations better. We use a Convolution Neural Network architecture and an attention-based transformer. Our results demonstrate that both the NN architectures can benefit from this enhanced preprocessing, where scattering transform can provide a more discriminative representation of bird vocalizations than the traditional Mel Spectrogram.

## 1 Introduction

Biodiversity is progressively endangered by anthropogenic change, making it essential to monitor animal populations to build sustainable plans. Utilizing passive acoustic monitoring (PAM) has become a novel method for observing biodiversity. Benefiting from recent advancements in automatic recording devices and related technologies Wood et al. [2019], Ruff et al. [2020]. This approach has been successfully applied in studies of detection of elephant calls, insects, primates, and density estimates of birds Ganchev et al. [2007], Sanders and Mennill [2014], Zwart et al. [2014], Heinicke et al. [2015], Wrege et al. [2017], Bjorck et al. [2019].

A primary challenge in PAM is identifying vocalizations of specific species within vast amounts of audio data. Researchers have developed pattern recognition techniques to find specific predefined sounds from recordings. However, unlike human speech recognition, which often follows the structured rules of a particular language with its vocabulary and grammar, animal sounds don't always adhere to such predictable patterns. Furthermore, Xie et al. [2022] suggested that although numerous studies have shown promising results for identifying individual species in short audio clips, the automated detection of various species in extended, noisy field recordings still presents challenges.

Computational Sustainability Workshop at NeurIPS 2023.

**Our Contributions** In this article, we delve into the challenge of identifying bird vocalizations using artificial neural networks. Here, calls from various bird species might intersect and overlap. To enhance our approach, we leverage the advancements in joint time-frequency scattering (JTFS) transform Mallat [2012], Bruna and Mallat [2013]. Different from other data-driven methods that utilize the Mel Spectrogram as input, **(1)** we propose to use **scattering transform representation** for bird vocalizations and **(2)** our results show how it significantly **augments the performance of both convolutional neural networks and transformer-based neural networks in the identification of bird vocalizations.** This underscores the potential of the scattering transformation in constructing a robust and optimized representation for bird vocalization, offering a promising avenue for future research and applications in PAM.

## 2 Building Representations for Bird Vocalization

**Background** Previous research has shown that models based on time-varying sinusoids are effective for classifying and recognizing standard bird sounds Harma and Somervuo [2004]. Yet, a significant category of bird sounds exists that don't precisely fit the mould of pure sinusoids. Instead, these sounds exhibit a distinct harmonic spectral composition. To address this challenge, researchers have aimed to create representations for various bird species, and recent studies have leveraged data-driven approaches in the domain of biodiversity monitoring Tolkova et al. [2021], Cohen et al. [2022], Xie et al. [2022].

**Neural Network-based methods** Neural Network-based (NN-based) approaches have revolutionized audio recognition Hinton et al. [2012]. Unlike traditional signal processing techniques that often depend on manually crafted parameters, NN-based methods take spectrograms as input and can recognize bird species in an end-to-end fashion. McIlraith and Card [1997] studied the recognition of songs of six species common to Manitoba, Canada, using backpropagation learning in two-layer perceptrons. Selouani et al. [2005] improved the NN approach by adding a feedback loop to the multilayer perceptron (MLP) network, and they tested the classification of sixteen Canadian bird species. Ruff et al. [2020] demonstrated the use of a deep convolutional neural network (CNN) for automating the detection of owl vocalizations in spectrograms generated from field recordings from Artuso et al. [2014].

In NN-based approaches, preprocessing plays an important role. Preprocessing generates representations that serve as the inputs for the NNs. We then train the NNs to extract features from these representations for classification tasks. Consequently, efficient representations can significantly enhance the performance of the NNs.

Different representations are used in NN-based method for bird vocalization. In McIlraith and Card [1997], the author built the representation using linear predictive coding and windowed Fourier transforms. In Ruff et al. [2020], they generated multiple spectrograms with different parameters as the input of their CNN. Among various representations, the Mel Spectrogram is the most commonly used one. It visually depicts how a signal's frequency spectrum changes over time Gong et al. [2021], Xie et al. [2019]. Recently, Wang et al. [2022] demonstrated that the joint time–frequency scattering (JTFS) transform outperforms the performance of a Mel-frequency representation in chick call recognition. This finding inspires the exploration of scattering transforms for species classification tasks that display similar spectrotemporal patterns.

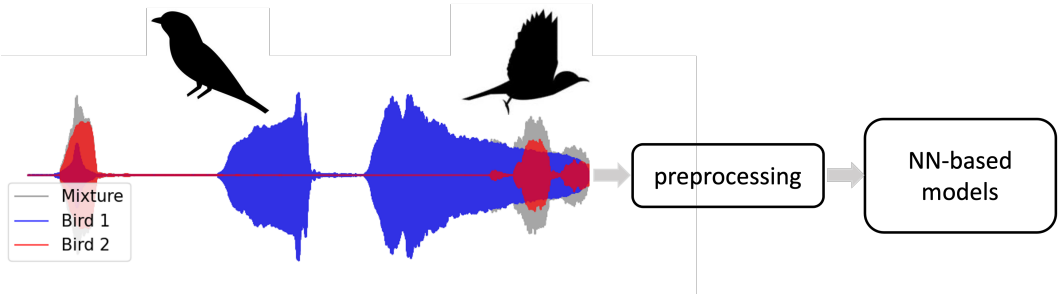


Figure 1: Illustation of the input data and our method.

**Dataset** In this paper, we focus on developing NN-based techniques for multi-label classification of bird vocalizations. We initially build a dataset using various recordings of individual bird vocalizations. In our preparation process, we choose bird calls from two different species and position them randomly within a 1-second time frame. These bird calls might start at different times within this span. We then merge them to form the new clip. In the dataset, every clip is a combination of two birds chosen from a set of 15 species: olsfly, grepew, wewpew, eawpew, yebfly, acafly, aldfly, wilfly, leafly, hamfly, gryfly, dusfly, pasfly, corfly, and bubfly. An illustration of our dataset is shown in Figure 1. Given a mixture of two birds, we aim to refine the preprocessing phase to construct a more effective representation suitable for NN-based models. This helps us identify which bird species are vocalizing within that clip.

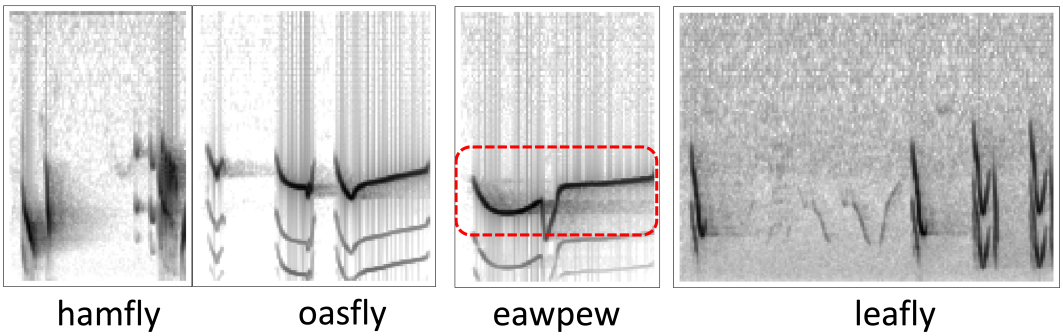


Figure 2: Mel Spectrogram $\mathbf{M}$ of different kinds of birds. The red box indicates a transition from a quickly varying pitch to a slower one. The Mel Spectrogram assigns similar amplitudes to the entire sound. Hamfly: Hammond's Flycatcher; oasfly: Olive-sided Flycatcher; eawpew: Eastern Wood-Pewee; leafly: Least Flycatcher.

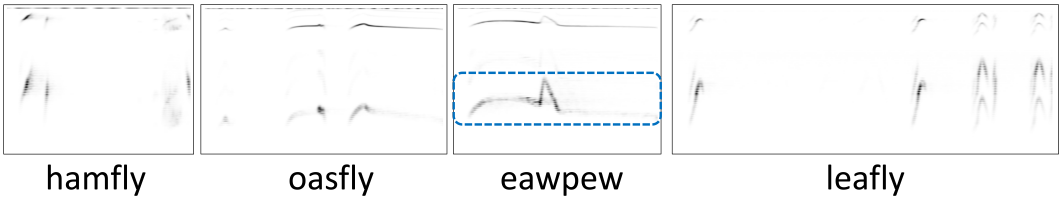


Figure 3: Scattering representation $\mathbf{S}$ of different kinds of birds. The blue box indicates a transition from a quickly varying pitch to a slower one. In the scattering transform, faster varying components are emphasized with greater amplitudes, while the slower varying parts are given diminished amplitudes. We also observed similar patterns in oasfly vocalization.

**Method** The Joint time–frequency scattering transform decomposes an audio waveform by a wavelet filterbank. The main idea behind scattering transforms is to capture hierarchical structures within the signal. When we apply them jointly in time and frequency, they can tease out intricate patterns that might be missed by other types of transformations. Given a audio clip $x(t)$, we consider 3 orders of scattering transform, order 0, order 1 and order 2, denoted as $S_0(J, Q, t)$, $S_1(J, Q, t)$ and $S_2(J, Q, t)$, where $J$ is the averaging scale and $Q$ is the number of wavelets per octave. Here $S_0$ corresponds to the original signal's average, $S_1$ is obtained by applying a wavelet transform followed by a modulus operation to the original signal, which provides time and frequency information separately. $S_2$ is calculated by further processing the $S_1$ coefficients, which captures interactions between time and frequency localized structures, revealing joint time-frequency patterns that the first order might miss, we refer more details of scattering transform to Mallat [2012]. For our dataset, we specify $J = 6$ and $Q = 16$. Considering a 1-second audio clip sampled at 44,100 Hz, the resulting dimensions are as follows: $S_0$ has a shape of (1, 345), $S_1$ is (63, 345), and $S_2$ is (158, 345). We subsequently concatenate $S_0$, $S_1$, and $S_2$ to construct the comprehensive scattering representation, $\mathbf{S} \in \mathbb{R}^{222\times345}$.

Here, we compare the scattering representation $\mathbf{S}$ and the commonly used Mel Spectrogram $\mathbf{M}$ of different birds. As illustrated in Figure 2 and Figure 3, the Mel Spectrogram $\mathbf{M}$ contains a significant amount of environmental noise. In contrast, within the scattering transformation, such noise, which remains relatively consistent over time, is primarily encapsulated by $S_0$. Both $S_1$ and $S_2$ capture the patterns without losing on temporal resolution. Moreover, the scattering transform excels in discerning swift or ephemeral sound structures. As illustrated in the case of the eawpew bird vocalization: the call typically starts with a rapidly fluctuating pitch, which then transitions to a slower variation. In the Mel Spectrogram $\mathbf{M}$, both the fast and slow varying sound structures are represented with similar amplitudes. However, when utilizing the scattering representation $\mathbf{S}$, the fast varying sound is manifested with a more pronounced signal (a greater amplitude) while the slow varying one appears more subdued. This distinction arises primarily due to the second-order scattering, $S_2$. It captures rapid temporal fluctuations in a signal, such as the transition from a fast varying to a slow varying sound, which is a signature characteristic observed in certain bird vocalizations.

## 3 Results

To evaluate the enhancement in performance offered by the scattering-based approach across various neural network models, we test it on two architectures: a convolutional neural network (CNN) and an attention-based method (transformer). The CNN structure we use has two 3x3 convolutional layers (with 4 and 8 channels respectively), followed by ReLU activations and 2x2 max-pooling. The output is flattened and passed through two linear layers for classification. The transformer is adapted from Gong et al. [2021], employing a purely attention-based approach based on the Vision Transformer Dosovitskiy et al. [2020].

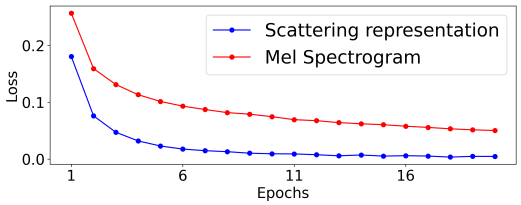


Figure 4: Training loss using CNN.

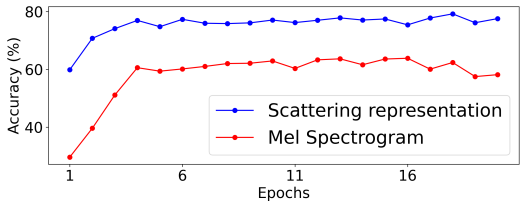


Figure 5: Validation accuracy using CNN.

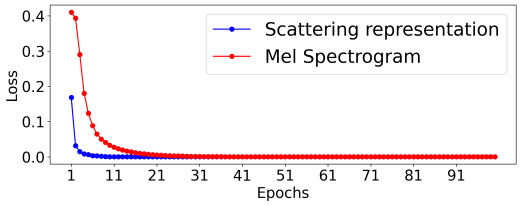


Figure 6: Training loss using transformer.

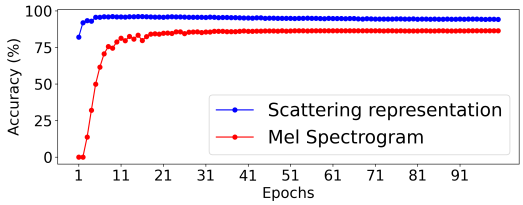


Figure 7: Validation accuracy using transformer.

We compare the CNN's performance and transformer's performance using scattering representation $\mathbf{S}$ and the commonly used Mel Spectrogram $\mathbf{M}$. The results are shown in Figure 4, 5, 6 and 7. We observe that using scattering representation accelerates convergence for both CNN and transformer during training. During validation, the scattering representation outperforms the Mel Spectrogram. Specifically, using CNN with the scattering representation, the model achieves a validation accuracy of 59.88% in the first epoch and reaches 77.58% after 20 epochs. In contrast, using the Mel Spectrogram, the final accuracy is only 58.17%. When using transformer, the model achieves a validation accuracy of 81.97% in the first epoch and 94.12% after 100 epochs, while Mel Spectrogram achieves 86.33% after 100 epochs. Our findings indicate that instead of using complicated NN-based methods like transformers, applying signal processing techniques (scattering) can also significantly boost performance in bird vocalization tasks.

## 4 Conclusion

In this paper, we discussed the Joint Time-Frequency Scattering (JTFS) transform and how it differs from the Mel Spectrogram in the context of bird vocalization analysis. We use the JTFS transform to preprocess bird vocalizations and improve the performance of NN-based models in bird vocalization recognition tasks. The JTFS's ability to represent complex time-frequency interactions makes it more suitable for the task of bird vocalization analysis. We hope that this work will inspire researchers to develop advanced signal processing techniques, enhancing input representations for NN-based methods in biodiversity monitoring.

## 5 Discussion

**Mel Spectrogram and JTFS** Mel Spectrogram focuses on representing the spectral information of a signal in a way that's more aligned with human hearing. The strength of JTFS lies in providing a highly detailed analysis, capturing both time and frequency characteristics while preserving the signal's hierarchical structure. This makes JTFS particularly useful in tasks requiring a deep understanding of signal properties. Taking the eawpew as an example, the Mel Spectrogram effectively captures how the frequency varies over time, providing a clear spectral representation. On the other hand, the JTFS offers a more detailed analysis, such as the rate at which the frequency changes, delving deeper into the dynamics of the sound's characteristics. The detailed and extensive analysis provided by JTFS makes it particularly suitable for various animal vocalization tasks, especially in scenarios where the frequency changes rapidly. For tasks like human speech, where frequency changes are less rapid, the Mel Spectrogram should suffice as it efficiently captures the necessary spectral information.

## 6 Future Work

This study focuses on a two-class classification problem using JTFS as preprocessing. We plan to explore multi-label classification problems, integrating JTFS with various methods, ranging from classical approaches such as Independent Component Analysis (ICA) to more neural network-based structures.

**Acknowledgement** We appreciate the constructive suggestions provided by the reviewers. This project is partially supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program; the National Science Foundation (NSF) and the National Institute of Food and Agriculture (NIFA); the Air Force Office of Scientific Research (AFOSR); the Department of Energy; and the Toyota Research Institute (TRI).

## References

Connor M Wood, Viorel D Popescu, Holger Klinck, John J Keane, RJ Gutiérrez, Sarah C Sawyer, and M Zachariah Peery. Detecting small changes in populations at landscape scales: a bioacoustic site-occupancy framework. *Ecological Indicators*, 98:492–507, 2019.

Zachary J Ruff, Damon B Lesmeister, Leila S Duchac, Bharath K Padmaraju, and Christopher M Sullivan. Automated identification of avian vocalizations with deep convolutional neural networks. *Remote Sensing in Ecology and Conservation*, 6(1):79–92, 2020.

Todor Ganchev, Ilyas Potamitis, and Nikos Fakotakis. Acoustic monitoring of singing insects. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–721. IEEE, 2007.

Claire E Sanders and Daniel J Mennill. Acoustic monitoring of nocturnally migrating birds accurately assesses the timing and magnitude of migration through the great lakes. *The Condor: Ornithological Applications*, 116(3):371–383, 2014.

Mieke C Zwart, Andrew Baker, Philip JK McGowan, and Mark J Whittingham. The use of automated bioacoustic recorders to replace human wildlife surveys: an example using nightjars. *PloS one*, 9 (7):e102770, 2014.

Stefanie Heinicke, Ammie K Kalan, Oliver JJ Wagner, Roger Mundry, Hanna Lukashevich, and Hjalmar S Kühl. Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods in Ecology and Evolution*, 6(7):753–763, 2015.

Peter H Wrege, Elizabeth D Rowland, Sara Keen, and Yu Shiu. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods in Ecology and Evolution*, 8(10): 1292–1301, 2017.

Johan Bjorck, Brendan H Rappazzo, Di Chen, Richard Bernstein, Peter H Wrege, and Carla P Gomes. Automatic detection and compression for passive acoustic monitoring of the african forest elephant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 476–484, 2019.

Jiangjian Xie, Yujie Zhong, Junguo Zhang, Shuo Liu, Changqing Ding, and Andreas Triantafyllopoulos. A review of automatic recognition technology for bird vocalizations in the deep learning era. *Ecological Informatics*, page 101927, 2022.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65 (10):1331–1398, 2012.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

A Harma and Panu Somervuo. Classification of the harmonic structure in bird vocalization. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–701. IEEE, 2004.

Irina Tolkova, Brian Chu, Marcel Hedman, Stefan Kahl, and Holger Klinck. Parsing birdsong with deep audio embeddings. *arXiv preprint arXiv:2108.09203*, 2021.

Yarden Cohen, David Aaron Nicholson, Alexa Sanchioni, Emily K Mallaber, Viktoriya Skidanova, and Timothy J Gardner. Automated annotation of birdsong with a neural network that segments spectrograms. *Elife*, 11:e63853, 2022.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Alex L McIlraith and Howard C Card. Birdsong recognition using backpropagation and multivariate statistics. *IEEE Transactions on Signal Processing*, 45(11):2740–2748, 1997.

S-A Selouani, M Kardouchi, E Hervet, and D Roy. Automatic birdsong recognition based on autoregressive time-delay neural networks. In *2005 ICSC Congress on Computational Intelligence Methods and Applications*, pages 6–pp. IEEE, 2005.

CC Artuso, CS Houston, DG Smith, C Rohner, and A Poole. The birds of north america online. 2014.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Jie Xie, Kai Hu, Mingying Zhu, Jinghu Yu, and Qibing Zhu. Investigation of different cnn-based models for improved bird sound classification. *IEEE Access*, 7:175353–175361, 2019.

Changhong Wang, Emmanouil Benetos, Shuge Wang, and Elisabetta Versace. Joint scattering for automatic chick call recognition. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 195–199. IEEE, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.