SongMAE: Fine-Grained Syllable Discovery in Birdsong Using Asymmetric Patches

George Vengrovski[1,2]
Timothy J. Gardner[1,2*]
[1]Institute of Neuroscience and Department of Biology
[2]Phil and Penny Knight Campus for Accelerating Scientific Impact
University of Oregon, Eugene, OR, USA
{georgev,timg}@uoregon.edu

Abstract

Self-supervised bioacoustic encoders have been used for species classification but so far have not addressed syllable-level structure in birdsong. We introduce SongMAE, a compact MAE-ViT (Masked Autoencoder Vision Transformer) that operates on mel spectrograms with 2 ms temporal resolution. SongMAE is pretrained with masked spectrogram reconstruction on diverse bioacoustic recordings. Despite a 14M-parameter footprint and 2 s context, its embeddings cluster some canary, Zebra Finch, and Bengalese Finch syllable types, yielding syllable-separable latent spaces and indicating the possibility of zero-shot, syllable-level analysis suitable for on-device ecological monitoring. We discuss limitations (lack of quantitative benchmarks and comparisons) and outline directions—patch aspect-ratio ablations, larger pretraining, and multi-resolution training—to enable unsupervised analysis of song components with a general pretrained model.

1 Introduction and Motivation

In the field of bioacoustics, there has been a proliferation of pretrained self-supervised encoding models. Motivated by classification and detection tasks in data-sparse contexts, models such as BirdMAE [1], AVES [2] and BEAT [3] have been pretrained on large corpora of bioacoustic data. The primary objective of these models has been to create embeddings of ecological recordings that are useful for the detection and classification of species. These approaches are state-of-the-art for detection and classification tasks; however, they were not designed with the intention of creating embeddings for the individual components of song, such as notes, syllables, or calls. The ability to create useful embeddings for these song elements remains an open problem that requires models specifically optimized for capturing fine temporal structure rather than global acoustic patterns.

All current bioacoustic embedding models segment audio, either via 1D convolution of raw samples (Whisper, HuBERT; used in WhisperSeg, AVES) [2], [4], [5] into tokens, or in the case of ViT-like models, spectrograms into patch embeddings [6]. A fundamental trade-off in temporal resolution exists: classification tasks benefit from coarse resolution, which reduces the memory footprint given transformers' quadratic attention scaling. However, segments then represent temporal spans that exceed birdsong's shortest inter-syllable gaps. The models are then inadequate for song structure analysis which requires fine-grained resolution to capture the rapid acoustic transitions defining individual vocalizations. No general bioacoustic models optimize for such resolution. TweetyBERT [7] demonstrates the value of high temporal resolution for syllable discovery but remains canary-specific. Furthermore, the TweetyBERT approach applies masks spanning all frequencies in specific

[*] Corresponding author: timg@uoregon.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Non-Human Animal Communication.

timeblocks, which removes the flexibility of experimenting with masking only portions of frequencies – which we believe is important for pretraining on species with non stereotyped song. Non-deep ML approaches for granular birdsong analysis exist, but they suffer from a low representational capacity [8].

To address this gap, we introduce SongMAE, a Masked Autoencoder (MAE) vision transformer (ViT) pretrained on a portion of the BirdSet dataset. We stress that this is preliminary work, as our scope is to test whether a general syllable-level model can learn representations that generalize to species and recording environments outside its training set, using canaries, Zebra Finch, and Bengalese Finch (for which we have high-quality labels) as our test case. Our key technical innovation lies in spectrogram generation and patch parameters: we compute spectrograms with 2ms time bin resolution and employ asymmetric patches of 32 mels × 1 time bin that preserve fine temporal granularity. SongMAE generates meaningful syllable-level representations for canary, Zebra Finch, and Bengalese Finch vocalizations, producing UMAP embeddings that spatially segregate human-annotated syllables in embedding space. Notably, the model was never trained on canaries and never encountered the laboratory recording conditions used for evaluation of any of these species.

2 Related Work

With respect to masked autoencoders, [9] demonstrated that asymmetric patches favoring temporal resolution (16 mels × 8 timebins, 16 mels × 4 timebins) outperformed square (16 mels ×16 timebins) patches across most audio tasks, with particularly strong gains in speech and music recognition. It has also been observed that for speech tasks, as opposed to audio event tasks, full frame level patches (spanning all mel bins) proved to be superior [10]. Research on wav2vec2 has shown that reducing temporal resolution while maintaining model size degrades performance, though this can be mitigated by increasing model capacity [11].

3 Methods

3.1 Architectural Overview

SongMAE follows the canonical Masked Autoencoder (MAE) architecture, consisting of an asymmetric encoder-decoder design, and it operates on spectrograms [12], [13]. The representation is learned through a self-supervised reconstruction task: spectrograms are divided into patches through a convolutional operation, positionally encoded, and then randomly masked. The encoder processes only the non-masked patches and passes these encoded representations, along with learned mask placeholder patches, to the decoder. The decoder then reconstructs the original masked patches using mean squared error loss.

3.2 Patch Parameters and Masking

Birdsong shares speech's characteristic of discrete, rapidly transitioning acoustic units. However, unlike human speech, birdsong operates on timescales approximately 10 times faster. Most self-supervised audio models employed in bioacoustics are directly derived from those used in human speech analysis and thus may be insufficient for analyzing fast-changing bird syllables. The key architectural innovation in SongMAE is the use of asymmetric patches optimized for fine temporal resolution. Unlike square patches (16×16) used in standard audio MAEs, which span 160 ms, we employ rectangular patches of 32 mel bins × 1 timebin. Combined with our 2 ms spectrogram timebins (versus BirdMAE's 10 ms bins), SongMAE achieves 80× finer temporal resolution than existing bioacoustic MAE models [14]. This comes at the cost of reduced frequency resolution (32 mel bins per patch versus 16), a tradeoff necessary to maintain manageable sequence lengths given GPU memory constraints, and something we aim to experiment with in further studies.

Our masking strategy employs contiguous masking to address the challenge of 2ms-width asymmetric patches, where adjacent unmasked patches could trivially reconstruct missing content through local correlations. We mask blocks of 32 consecutive patches along the temporal dimension, forcing the model to learn longer-range dependencies rather than rely on local interpolation. Then, we add additional single patch masks to obtain exactly 75% masking.

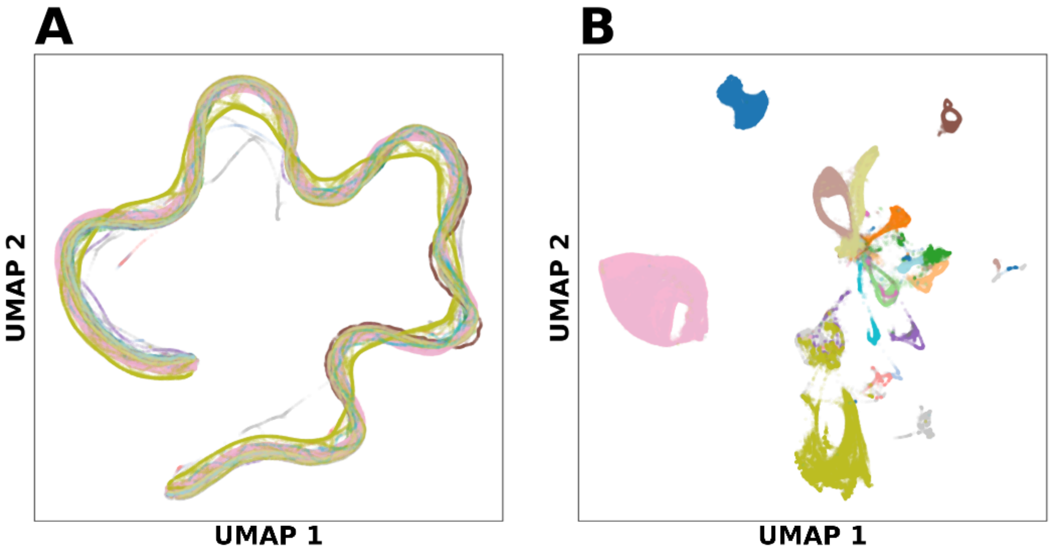

Figure 1: SongMAE latent embeddings of canary song. (A): UMAP applied directly to encoder embeddings. (B): Embedding after de-positioning. Colors represent canary syllable types.

3.3 Pretraining Dataset

The pretraining dataset consists of a subset of the XCL portion of the BirdSet dataset [15], derived from Xeno-Canto recordings [16]. Notably, we exclude all atlantic canary recordings from the training set to enable zero-shot evaluation on canary vocalizations; these recordings are moved into the validation set. After filtering for detected bird vocalizations, the final pretraining dataset contained 512,620 recording snippets, and the validation set contained 100,000 recordings.

3.4 Spectrogram Generation

Audio files are resampled to 32 kHz and converted to 128-dimensional log-mel spectrograms. We use an FFT size of 1024 with a hop length of 64 samples, producing time bins of 2ms. Spectrograms are normalized using dataset-wide z-scoring statistics (mean: -50.14, std: 20.66). During pretraining, spectrograms are segmented into fixed contexts of 1024 time bins (2.048 seconds), with shorter segments zero-padded as needed.

3.5 Training Details

We adopt a 75% masking ratio and train for 150,000 batches using the AdamW optimizer with a learning rate of 2e-4 and a cosine annealing schedule [17], [18]. The reconstruction loss follows the standard MAE approach: normalized patch-wise MSE computed only on masked patches. We use a batch size of 256 and enable automatic mixed precision training for training speed and efficient memory usage [19].

3.6 Latent Space Generation

We obtained ground-truth syllable annotations from laboratory-recorded datasets for canary [20], Bengalese Finch [21], and Zebra Finch [22]. For latent space generation, we employed an automated song detection system to identify and exclude spectrogram segments lacking songs, or extended periods of silence, ensuring our embeddings represent only meaningful song content [7].

To visualize the learned representations, we processed canary spectrograms through the SongMAE encoder and extracted patch embeddings for analysis. For each spectrogram chunk of 1024 timebins (2.048 seconds), the encoder produces a grid of patches with dimensions H×W×D, where H represents frequency patches (4 for 128 mel bins with 32-mel patches), W represents temporal patches (1024 for 1-timebin patches), and D is the embedding dimension (384).We reshape these patches to preserve

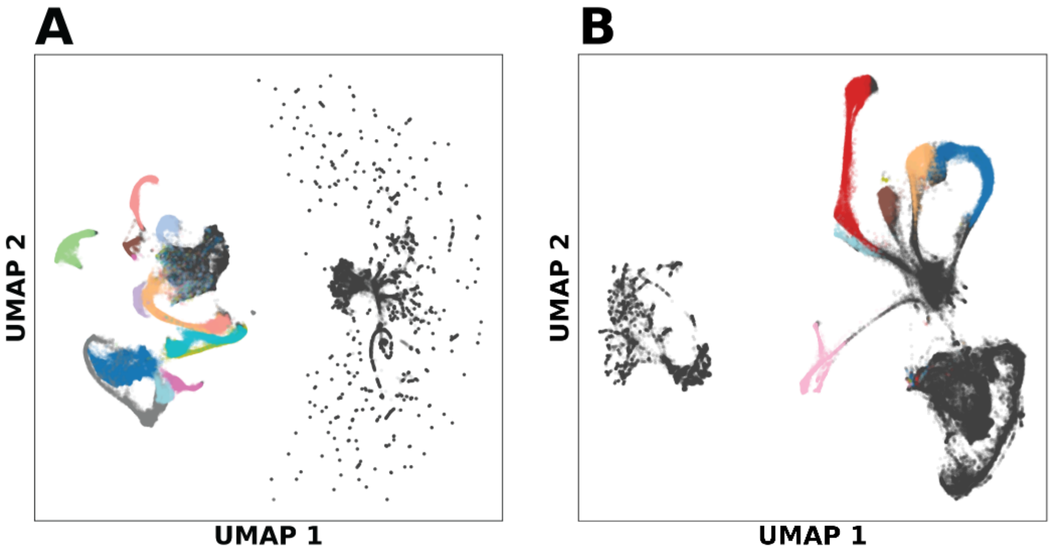

Figure 2: (A): Bengalese Finch embedding after de-positioning. (B): Zebra Finch embedding after de-positioning. Colors represent ground truth syllable types, black represents intersyllabic silence.

temporal relationships while stacking frequency information, resulting in embeddings of shape W×(H·D). This frequency-stacking approach maintains the temporal sequence while combining multi-frequency information at each time point into a single high-dimensional vector.

The model's learned positional encodings introduce a strong positional bias that initially dominates the latent structure. To reveal syllable-level patterns, we apply a "de-positioning" procedure: for each of the 1024 possible positions in the context window, we compute the mean embedding across all occurrences of that position in the dataset, then subtract this position-specific mean from each embedding. This effectively removes the positional signal while preserving acoustic content.

We perform UMAP dimensionality reduction with the following parameters: 2 components, 200 neighbors, cosine distance metric, and minimum distance of 0.1.

4 Results

4.1 Latent Space Generation

Despite never encountering canary vocalizations during training, SongMAE's learned representations successfully capture canary syllable structure in an unsupervised manner. We collected 256,000 timebins (8.5 mins) of canary recordings, removed non-song elements, encoded it, and used UMAP to visualize the latent encoding. Figure 1 shows the resulting two-dimensional projection of SongMAE's latent space. The left panel displays the raw UMAP projection, which exhibits organization but appears dominated by a "rope-like" structure caused by the learned positional encoding. After applying our de-positioning procedure, distinct syllable clusters emerge , with each cluster corresponding primarily to a single syllable type as colored by ground-truth labels. We observe similar syllable-level clustering for Bengalese and Zebra Finch (Figure 2).

5 Discussion

Our results, although preliminary, suggest that SongMAE can learn generalizable representations of the individual components of song such as notes and syllables. Reconstructions from species entirely absent from the pretraining set demonstrate that SongMAE can infer large missing fragments of song structure from small contextual clues, something that can only be done if a strong and general representation of birdsong is learned. Visualizations and clustering of the latent space of the model demonstrate that it can be used to extract syllable-level representations of some canary vocalizations despite never being exposed to them during the pretraining process.

These findings suggest practical applications for bioacoustic monitoring. With further refinement of the model, through training optimizations and a larger dataset, SongMAE could be used for the discovery of song units in species without human annotations—or at least to provide an initial set of soft labels for scientists that can accelerate analysis and labeling efforts by researchers. Another potential application is dialect mapping and species-abundance quantification. Detection of shifts and variations in specific song units or calls could be used to differentiate populations or individuals in ecological monitoring settings [23]. Furthermore, the small size of the model (approximately 14 million total parameters) makes it particularly suitable for embedded devices used in remote ecological monitoring. In such settings, computational resources are limited and internet connectivity is often weak. Storing months of raw audio recordings is impractical, making the model's compressed embedding outputs advantageous. SongMAE's compact size makes it well-suited for on-device deployment, enabling the potential for real-time analysis in remote areas with low throughput network connectivity [24].

However, because this investigation is preliminary and largely exploratory, there are several limitations that must be addressed before practical use. The primary limitation is that we have only evaluated the vocalizations of three bird species, all recorded in lab settings, and we are not confident in the model's ability to discover units of song in other species, particularly those in noisy settings or with highly variable songs. Even though the reconstruction quality for validation species demonstrates that the model has learned meaningful acoustic representations, it is possible - even likely - that more sophisticated analysis methods beyond dimensionality reduction and clustering would be required to extract syllable-level structure from the model's latent space for species with more complex vocalizations.

Our immediate next steps consist of exploring patch size and aspect ratios on latent embeddings and linear probes, to better understand the optimal trade-off between frequency and temporal resolution. We also want to apply this to different species to assess whether completely unsupervised syllable discovery is achievable, with the primary difficulty being collecting individual unit labels. SongMAE is currently undertrained, and augmentations, data curation, and increasing the size of spectrograms in the pretrain dataset will likely provide significant improvements in the model.

Acknowledgments and Disclosure of Funding

This work was funded by NIH R01NS118424.

References

1. Rauch L, Moummad I, Heinrich R, Joly A, Sick B, Scholz C. Can masked Autoencoders also listen to birds? arXiv [cs.LG]. 2025. Available: `http://arxiv.org/abs/2504.12880`
2. Hagiwara M. AVES: Animal Vocalization Encoder based on Self-Supervision. arXiv. 2022 [cited 28 May 2024]. Available: `https://github.com/earthspecies/aves`
3. Miron M, Alizadeh M, Gilsenan-McMahon E, Narula G, Chemla E, Robinson D, et al. What matters for bioacoustic encoding. arXiv [cs.SD]. 2025. Available: `http://arxiv.org/abs/2508.11845`
4. Hsu W-N, Bolte B, Tsai Y-HH, Lakhotia K, Salakhutdinov R, Mohamed A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv [cs.CL]. 2021. Available: `http://arxiv.org/abs/2106.07447`
5. Gu N, Lee K, Basha M, Ram SK, You G, Hahnloser RHR. Positive Transfer of the Whisper Speech Transformer to Human and Animal Voice Activity Detection. bioRxiv. 2023. p. 2023.09.30.560270. doi:10.1101/2023.09.30.560270
6. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv [cs.CV]. 2020. Available: `http://arxiv.org/abs/2010.11929`
7. Vengrovski G, Hulsey-Vincent MR, Bemrose MA, Gardner TJ. TweetyBERT: Automated parsing of birdsong through self-supervised machine learning. *Animal Behavior and Cognition*. bioRxiv. 2025. Available: `https://www.biorxiv.org/content/10.1101/2025.04.09.648029v1.full.pdf`
8. Goffinet J, Brudner S, Mooney R, Pearson J. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *eLife*. 2021;10. doi:10.7554/eLife.67855
9. Niizumi D, Takeuchi D, Ohishi Y, Harada N, Kashino K. Masked Spectrogram Modeling using Masked Autoencoders for learning general-purpose audio representation. arXiv [eess.AS]. 2022. Available: `http://arxiv.org/abs/2204.12260`
10. Baade A, Peng P, Harwath D. MAE-AST: Masked Autoencoding Audio Spectrogram Transformer. arXiv [eess.AS]. 2022. Available: `http://arxiv.org/abs/2203.16691`
11. Wu F, Kim K, Pan J, Han K, Weinberger KQ, Artzi Y. Performance-efficiency trade-offs in unsupervised pre-training for speech recognition. arXiv [cs.CL]. 2021. Available: `http://arxiv.org/abs/2109.06870`
12. Huang P-Y, Xu H, Li J, Baevski A, Auli M, Galuba W, et al. Masked Autoencoders that Listen. arXiv [cs.SD]. 2022. Available: `http://arxiv.org/abs/2207.06405`
13. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. arXiv [cs.CV]. 2021. Available: `http://arxiv.org/abs/2111.06377`
14. Schwinger R, Rauch L, Huseljic D, Wirth M, Herde M, Heinrich R, et al. Foundation models for bioacoustics – a comparative review. arXiv [cs.SD]. 2025. Available: `http://arxiv.org/abs/2508.01277`
15. Rauch L, Schwinger R, Wirth M, Heinrich R, Huseljic D, Herde M, et al. BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics. arXiv [cs.SD]. 2024. Available: `http://arxiv.org/abs/2403.10380`
16. Vellinga W, Planqué R. The Xeno-canto collection and its relation to sound recognition and classification. *CLEF*. 2015. Available: `https://www.academia.edu/download/71184984/The_Xeno-canto_collection_and_its_relati20211003-3195-131pdy3.pdf`
17. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv [cs.LG]. 2014. Available: `http://arxiv.org/abs/1412.6980`
18. Loshchilov I, Hutter F. SGDR: Stochastic gradient descent with warm restarts. arXiv [cs.LG]. 2016. Available: `http://arxiv.org/abs/1608.03983`
19. Micikevicius P, Narang S, Alben J, Diamos G, Elsen E, Garcia D, et al. Mixed Precision Training. International Conference on Learning Representations. 2018. Available: `https://openreview.net/pdf?id=r1gs9JgRZ`
20. Cohen Y, Nicholson DA, Sanchioni A, Mallaber EK, Skidanova V, Gardner TJ. Automated annotation of birdsong with a neural network that segments spectrograms. *eLife*. 2022;11. doi:10.7554/eLife.63853
21. Nicholson D, Queen JE, Sober SJ. Bengalese Finch song repository. figshare. 2023. doi:10.6084/m9.figshare.4805749.v9
22. Koch TMI, Marks ES, Roberts TF. AVN: A deep learning approach for the analysis of birdsong. *eLife*. 2024. doi:10.7554/eLife.101111

23. Marler P, Tamura M. Song ‘dialects’ in three populations of white-crowned sparrows. *Condor*. 1962;64(5):368–377

24. Höchst J, Brensing C, Richly S, Kuchenbecker M, Penkert M, Beyer N, et al. Bird@edge: Bird species recognition at the edge. In: Lecture Notes in Computer Science. Cham: Springer International Publishing; 2022. pp. 69–86

Appendix

Hyperparameter	SongMAE Enc.	SongMAE Dec.	BirdMAE ViT-B Enc.	BirdMAE ViT-L Enc.	BirdMAE ViT-H Enc.
Hidden Dim	384	192	768	1024	1280
# Heads	6	6	12	16	16
# Layers	6	3	12	24	32
FFN Dim	1536	768	3072	4096	5120
Pos Enc Type	Learned	Learned	Fixed sinusoidal	Fixed sinusoidal	Fixed sinusoidal
Total Params	~12.5M	~1.5M	86M	307M	632M
Context Length	~2 s	~2 s	5 s	5 s	5 s
Effective Patch	32 mels × 2 ms		16 mels × 160 ms		

Appendix Table 1: Architectural specifications of SongMAE versus BirdMAE models. SongMAE's fine temporal resolution (2 ms vs 160 ms patches) necessitates compact architectures to manage memory constraints from longer sequence lengths. BirdMAE decoder size is not shown as it is not described in the paper.