Tracking Time-Varying Syntax in Birdsong with a Sequential Autoencoder

Nadav Elami, Yarden Cohen
Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel
nadav.elami@weizmann.ac.il, Yarden.J.Cohen@weizmann.ac.il

Abstract

Songbirds are excellent models for studying sensorimotor sequence learning. Their songs are composed of vocal units, called syllables. The ordering of syllables in song is governed by syntax rules that determine syllable transition probabilities. We recently used regression analysis to show that canaries, a seasonal songbird, change transition probabilities across days and afford a new model for studying how the brain adapts syntax rules. But, regression analyses, which calculate transition probabilities in neighboring song batches, are noise-limited in small subsets of songs.

Here, to overcome this limitation and study the dynamics of syntax rules in fine temporal resolution we develop a neural filtering approach that infers time-varying transition probabilities from birdsong sequences. Inspired by deep learning methods for analyzing neural spiking data, we designed an autoencoder that treats each song as an observation from a probabilistic syntax model whose parameters change between song bouts. We carried simulated experiments, modeling both simple Markov and second-order dependencies of transitions, and demonstrate that our method accurately tracks syntax changes.

These findings underscore the potential of our approach to reveal the neural mechanisms underlying dynamic sensorimotor sequence generation.

Keywords: songbird syntax; time-varying models; Markov processes; neural filtering; autoencoder; sequence learning

Introduction

Songbirds offer valuable models for studying the neural mechanisms underlying sensorimotor behaviors. They acquire their vocal repertoire through memorizing tutor songs and iteratively refining their vocal output (Fee, 2004). The resulting song is comprised of distinct vocal syllables. In variable singers like Bengalese finches and canaries, the sequential arrangement of syllables, the syntax, exhibits probabilistic rules where syllable transitions depend on the identity and order of preceding syllables. In canaries, these syntax dependencies can bridge as many as 50 syllables and extend over several seconds of behavior (Markowitz, Ivie, Kligler, & Gardner, 2013; Cohen et al., 2020).

Birdsong syntax is often assumed to reflect a stationary Markov process that sets fixed transition probabilities between syllables (Zhang, Wittenbach, Jin, & Kozhevnikov, 2017). However, we recently showed that canaries alter their transition probabilities across days and weeks during their spring mating season (Levin & Cohen, 2024)—perhaps influenced by hormonal fluctuations, seasonal cues, or subtle learning mechanisms. These syntax rule dynamics were captured using regression-based methods that compute daily-averaged transition probabilities from batches comprising roughly 100 songs per day. Such methods face an inherent trade-off between temporal resolution and statistical power. Moreover, enforcing a single regression timescale and functional form may obscure the system's true dynamics.

To address these limitations, we developed a *neural* filtering approach. Rather than imposing strong parametric constraints on how syntax rules change over time, we leverage a Latent Factor Analysis via Dynamical Systems (LFADS) framework to learn the dynamics of transition probabilities directly from the data (Pandarinath et al., 2018). We create an autoencoder that processes observed syllable-to-syllable transition counts (bigrams), and generates latent (i.e., not directly observed but inferred from the data) transition logits that evolve flexibly over time.

We demonstrate the effectiveness of this approach using simulated data that mimics canary and Bengalese finch songs. In these simulations, transitions vary at different rates. Some follow monotonic trends, while others exhibit oscillatory patterns. We find that the model consistently outperforms traditional estimates derived from bigram counts – whether these counts are used directly (raw) or after applying a simple smoothing procedure – highlighting its ability to detect finer variations in song production in greater temporal resolution.

By revealing these changes in a data-driven manner, our framework can offer new insights into the neurobiological mechanisms underlying birdsong flexibility. Future applications of this approach may extend to large datasets of canary or Bengalese finch recordings, helping to uncover how neural circuits balance robust song production with ongoing syntax rule changes. More broadly, this method could provide insights into sequence generation in other domains, such as human speech or during motor skill acquisition.

Methods

Neural Network–Based Optimal Filtering

Our goal is to predict latent transition logits that underlie observed bigram counts. In our notation, hat-marked quantities (e.g., $\hat{\mathbf{Z}}_k$, $\hat{P}(\hat{\mathbf{Z}}_k)$) denote model predictions; those without a hat (e.g., $\mathbf{Z}_k$, $P(\mathbf{Z}_k)$) are ground truth from the simulated data.

We define our model through a recurrent mapping:

$$\hat{\mathbf{Z}}_k = f_\Sigma(\mathbf{y}_k, \hat{\mathbf{Z}}_{k-1}),$$

where f_Σ is generally a smooth function parameterized by Σ and implemented as a recurrent artificial neural network. Here, $\mathbf{y}_k$ is the observed bigram counts for batch k computed from 10–60 songs per batch.

The predicted latent logits, $\hat{\mathbf{Z}}_k$, are converted into transition probabilities using a row-wise softmax $g(\cdot)$:

$$\text{Prob}(i \to j) = \hat{P}(\hat{\mathbf{Z}}_k)_{ij} = g(\hat{\mathbf{Z}}_k)_{ij} = \frac{\exp((\hat{Z}_k)_{ij})}{\sum_{j'} \exp((\hat{Z}_k)_{ij'})}.$$

where i and j are syllable types. We learn the model by minimizing the cross-entropy loss,

$$\mathcal{L}(\Sigma) = -\sum_k \sum_{i,j} P(\mathbf{Z}_k)_{ij} \log \hat{P}(\hat{\mathbf{Z}}_k)_{ij},$$

using backpropagation through time.

Autoencoder Architecture

Our neural autoencoder (Figure 1), inspired by the LFADS framework, employs a bidirectional encoder-decoder structure. Bigram counts from song batches are first processed by a bidirectional GRU encoder network to infer initial latent states. A second recurrent controller network dynamically updates inferred latent inputs. These latent variables feed into a bidirectional GRU-based generator network, producing latent factors that map to transition logits. Row-wise softmax normalization transforms logits into probability transition matrices. Training optimizes cross-entropy reconstruction loss augmented by a Kullback–Leibler (KL) divergence penalty between the inferred latent distribution and a Gaussian prior, encouraging the latent space to remain smooth and well-regularized, thereby capturing the underlying syntax variations.

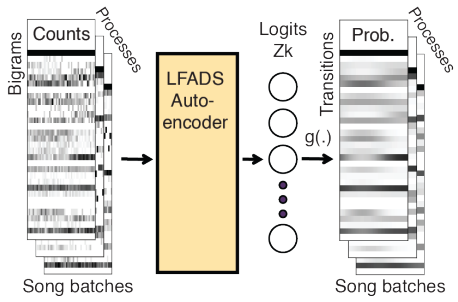

Figure 1: **Network architecture.** Autoencoder (LFADS-inspired) encoding batch bigram counts into latent transition logits, row-wise softmax normalization to transition probabilities, and reconstruction.

Simulation of Birdsong Sequences

To evaluate the model, we generated synthetic song sequences of 3–12 syllables using a six-syllable alphabet (including start and end markers). The transition probabilities were governed by either a first- or second-order Markov process. For each simulated process, the batch-wise transition matrices were used to draw a single realization of song sequences. Changes in the transition probabilities were imposed over 7–100 sequential batches, following either a monotonic trajectory or a periodic trajectory (see Figure 2 for examples).

The simulated dataset comprised between 20K and 50K independent processes (with one realization per process), which were then split into a training set (80%) and a test set (20%). We compared the model's predicted transition probabilities against two baseline methods: (1) raw bigram counts derived directly from the simulated sequences and (2) smoothed bigram counts obtained by applying a running average over a window of 5 batches to reduce noise.

Results

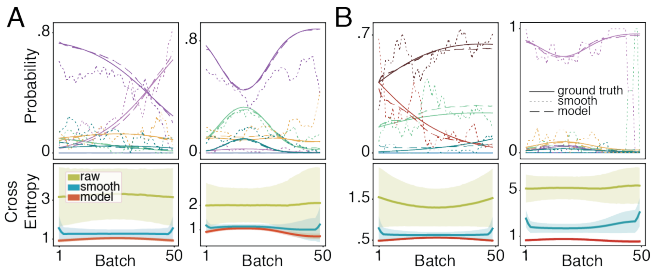

Figure 2: **Simulation results.** Transition probabilities (**top**) and cross-entropy errors (**bottom**) for first (**A**) and second-order (**B**) processes evolving monotonically (**left**) or periodically (**right**) over 50 batches of 10-60 songs each. Autoencoder estimates (dashed) closely track true transitions (solid), outperforming smoothed (dotted) and raw (not shown) baselines.

Figure 2 summarizes the core findings. While raw bigram estimates are noisy and smoothed estimates introduce bias, our autoencoder predictions more accurately track both gradual and oscillatory shifts in the ground-truth probabilities. This advantage in performance appears robust across different dependency orders (first vs. second) and across varying transition evolution patterns (monotonous vs. periodic). Quantitatively, the model maintains lower cross-entropy with the true transition probabilities and exhibits lower variability across simulated processes (shaded regions).

Discussion

Most studies of birdsong treat syntax rules as static, yet accumulating evidence suggests that the underlying transition probabilities may vary over time. By modeling each batch of songs as a time-varying probabilistic process, our autoencoder offers an approach to track changes in transition probabilities that significantly improve accuracy and temporal resolution compared to methods that assume stationarity or impose strict regression windows.

While our current work does not yet demonstrate this on large-scale real-world recordings, the method has the potential to enable such analyses and reveal how changing environments or internal states shape birdsong syntax rules. Beyond songbirds, our approach may generalize to other domains where hidden dynamics influence sequence generation—such as sensorimotor learning, or alterations in speech syntax associated with aging and disease.

Acknowledgments

This work was supported by a research grant from the Latin American Hub for New Scientists, by a personal research grant (N. 2401/22 to YC) from the Israel Science Foundation, and by an ERC grant (*NeuralSyntax*, 101170729, to YC).

References

Cohen, Y., Shen, J., Semu, D., Leman, D. P., Liberti, W. A., Perkins, L. N., ... Gardner, T. J. (2020). Hidden neural states underlie canary song syntax. *Nature*, *582*(7813), 539–544.

Fee, M. S. (2004). The role of auditory feedback in learned vocal behaviors. *Current Opinion in Neurobiology*, *14*(6), 736–741.

Levin, S., & Cohen, Y. (2024). Canary mating season songs move between order and disorder. *bioRxiv*, 2024–10.

Markowitz, J. E., Ivie, E., Kligler, L., & Gardner, T. J. (2013). Long-range order in canary song. *PLoS computational biology*, *9*(5), e1003052.

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., ... others (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, *15*(10), 805–815.

Zhang, Y. S., Wittenbach, J. D., Jin, D. Z., & Kozhevnikov, A. A. (2017). Temperature manipulation in songbird brain implicates the premotor nucleus hvc in birdsong syntax. *Journal of Neuroscience*, *37*(10), 2600–2611.