Behavioral/Cognitive

Partially Observable Markov Models Inferred Using Statistical Tests Reveal Context-Dependent Syllable Transitions in Bengalese Finch Songs

Jiali Lu,[1]* Sumithra Surendralal,[2]* Kristofer E. Bouchard,[3,4] and Dezhe Z. Jin[1]

[1]Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA, [2]Symbiosis School for Liberal Arts, Symbiosis International (Deemed University), Pune 411014, Maharashtra, India, [3]Scientific Data Division and Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, and [4]Helen Wills Neuroscience Institute & Redwood Center for Theoretical Neuroscience, University of California at Berkeley, Berkeley, California 94720

Generative models have diverse applications, including language processing and birdsong analysis. In this study, we demonstrate how a statistical test, designed to prevent overgeneralization in sequence generation, can be used to infer minimal models for the syllable sequences in Bengalese finch songs. We focus on the partially observable Markov model (POMM), which consists of states and the probabilistic transitions between them. Each state is associated with a specific syllable, with the possibility that multiple states may correspond to the same syllable. This characteristic differentiates the POMM from a standard Markov model, where each syllable is linked to a single state. The presence of multiple states for a syllable suggests that transitions between syllables are influenced by the specific contexts in which these transitions occur. We apply this method to analyze the songs of six adult male Bengalese finches, both before and after they were deafened. Our results indicate that auditory feedback plays a crucial role in shaping the context-dependent syllable transitions characteristic of Bengalese finch songs.

Significance Statement

Generative models are proficient at representing sequences where the order of elements, such as words or birdsong syllables, is context-dependent. In this study, we demonstrate that a probabilistic model, inspired by the neural encoding involved in song production in songbirds, effectively captures the context-dependent transitions of syllables in Bengalese finch songs. Our findings reveal that the absence of auditory input, as observed in deafened Bengalese finches, reduces these context dependencies, indicating that auditory feedback is essential for establishing context-dependent sequencing in their songs. This method can be applied to various behavioral sequences, offering valuable insights into the neural mechanisms underlying the statistical patterns that govern these sequences.

Introduction

Behavioral sequences, ranging from human language to birdsong, follow probabilistic rules. The success of large language models, such as GPT, demonstrates that the transition probabilities between words in a sentence are highly influenced by the preceding words (OpenAI, 2023). However, it remains unclear how such context-dependent probabilistic rules are encoded in the brain. Studies of birdsongs provide insights into this issue. Variable songs of species such as the Bengalese finch and canary exhibit context-dependent syllable transitions (Okanoya, 2004; Jin and Kozhevnikov, 2011; Jin, 2013; Markowitz et al., 2013), and the neural correlates of these dependencies have been studied using advanced imaging techniques to visualize the neural activity in the brain areas controlling the song (Cohen et al., 2020).

The probabilistic rules governing syllable transitions, or song syntax, can be effectively modeled using state transition models (Okanoya, 2004; Jin, 2013; Markowitz et al., 2013). The partially observable Markov model (POMM) (Jin and Kozhevnikov, 2011), which consists of states associated with individual syllables and probabilistic transitions between them, is closely related to the neural control of singing in the sensory-motor area HVC (used as a proper name) in songbirds (Fee et al., 2004; Jin, 2009). Experiments have demonstrated that a synaptic chain

Received March 19, 2024; revised Oct. 15, 2024; accepted Dec. 26, 2024.

Author Contributions: J.L., S.S., and D.Z.J. performed research; J.L., S.S., and D.Z.J. analyzed data; J.L., S.S., and D.Z.J. edited the paper; K.E.B. contributed unpublished reagents/analytic tools; D.Z.J. designed research; D.Z.J. wrote the first draft of the paper; D.Z.J. wrote the paper.

Research was supported by NSF award EF-1822476 (DZJ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

*These authors contributed equally to the project.

The authors declare no competing financial interests.

Correspondence should be addressed to Dezhe Z. Jin at dzj2@psu.edu.

https://doi.org/10.1523/JNEUROSCI.0522-24.2024

network, which supports the propagation of ultra-sparse burst spike sequences (Hahnloser et al., 2002; Long et al., 2010; Egger et al., 2020), drives downstream motor neurons to produce a specific syllable (Fee et al., 2004). In a POMM, each state corresponds to one such "syllable-chain". Transitions between the states generate syllable sequences. A single syllable can be associated with multiple states, allowing the Markovian dynamics of state transitions to produce non-Markovian, context-dependent syllable transitions (Jin, 2009; Jin and Kozhevnikov, 2011). Context dependency is, therefore, closely related to state multiplicity.

Auditory feedback plays an important role in shaping the syllable sequences of Bengalese finch songs (Okanoya and Yamaguchi, 1997; Woolley and Rubel, 1997, 2002; Sakata and Brainard, 2008; Wittenbach et al., 2015). Within a few days after deafening, the syllable sequences become more random (Okanoya and Yamaguchi, 1997; Woolley and Rubel, 1997). Additionally, there is a significant decrease in syllable repetitions (Wittenbach et al., 2015). When altered auditory feedback is provided to Bengalese finches during singing, particularly at the branching points of syllable transitions, it can significantly influence the probabilities of these transitions (Sakata and Brainard, 2006, 2008). But it is unclear how auditory feedback impact the context dependencies of syllable transitions in the song of the Bengalese finch.

In this paper, we develop a novel statistical method for inferring POMMs from a set of observed syllable sequences. This principled method is based on an interpretable statistical measure and is fully automated to infer the POMM with the minimal number of states while remaining compatible with the observed syllable sequences. It represents a significant advancement over the previous heuristic approach (Jin and Kozhevnikov, 2011), which required manual interventions and did not guarantee the inference of a minimal POMM. Using this new method, we construct minimal POMMs from the songs of six Bengalese finches both before and shortly after deafening. Our results show that deafening reduces state multiplicity in the POMMs, indicating that auditory feedback plays a crucial role in creating context dependencies in Bengalese finch songs.

Materials and Methods

Dataset. The data set in this work was previously used to analyze syllable repetitions in Bengalese finch songs (Wittenbach et al., 2015) (accessible for download at http://www.dezhejinlab.org/SharedData/). Specific details regarding the recording of songs, the annotation of syllables, the procedure for deafening, as well as the Ethics Statement, are available in the published paper (Wittenbach et al., 2015).

Deafening was performed by bilateral cochlear removal, with the completeness of the removal verified through visual inspection under a dissecting microscope. Songs from the deafened Bengalese finches were recorded 2 to 4 days post-deafening (Wittenbach et al., 2015). The data used in this study were collected from six male adult Bengalese finches, identified as bfa14, bfa16, bfa19, bfa7, o10bk90, and o46bk78, both before and after deafening.

Within the data set, syllables are labeled from *a* to *l* and from *x* to *z*. Ambiguous syllables are denoted by the symbols 0 and −, and these are excluded from the analysis. Typically, Bengalese finch song bouts begin with short introductory notes, labeled as *i*, *j*, and *k*. Syllable sequences are defined as segments of syllables bracketed by periods of introductory notes and the end of the recordings.

POMM. A POMM is characterized by a state vector $V = [\alpha, \omega, s_3, s_4, \ldots, s_n]$, where $s_1 = \alpha$ and $s_2 = \omega$ represent the start and end states, respectively. Here, n denotes the total number of states, and for $i = 3, \ldots, n$, s_i corresponds to the syllable symbol associated with the ith state. It is important to note that the same syllable symbol can occur multiple times within the state vector.

Transitions between these states are governed by a transition matrix T, where each element T_{ij} specifies the probability of transitioning from state i to state j. There are two constraints: no transitions lead back to the start state ($T_{i1} = 0$), and no transitions occur from the end state ($T_{2j} = 0$).

The sequence generation from a POMM begins at the start state. At each state i, the next state j is selected based on the probabilities T_{ij} from among the potential states (s_2 to s_n). Once the next state is selected, the symbol s_j is appended to the sequence. This process continues until the end state is reached, at which point the sequence generation terminates.

Markov model. A Markov model can be viewed as a special case of a POMM, where each syllable symbol appears only once in the state vector. In this case, the transition probabilities T are calculated as follows:

$$T_{ij} = \frac{N_{ij}}{N_i},$$

where N_i represents the total number of occurrences of the state (or syllable) s_i in the set of syllable sequences Y, and N_{ij} denotes the total number of times the subsequence $s_i\, s_j$ (i.e., the transition from s_i to s_j) appears in Y.

Additionally, note that:

$$N_i = \sum_{j=1}^{n} N_{ij}.$$

Therefore, to compute the transition probabilities, it is only necessary to determine the values of N_{ij}.

Baum-Welch algorithm. Due to state multiplicity, computing the transition matrix T for a POMM from the set of syllable sequences Y is more complex than for a standard Markov model, though the general approach remains similar. The process begins by assigning random state transition probabilities. Using these initial probabilities, the state transition sequences corresponding to the syllable sequences in the set Y are determined. The transition probabilities are then updated as follows:

$$T_{ij} = \frac{N_{ij}}{N_i},$$

where N_i represents the number of times state i appears in the state sequences, and N_{ij} is the number of times the subsequence of states ij appears. The procedure is repeated with updated transition probabilities T until the change in T becomes smaller than 10^{-6}. Given that the final result may depend on the initial randomization of T, the process is repeated 100 times with different random seeds, and the transition matrix T that maximizes the probability of generating Y from the POMM is selected.

This computation is efficiently implemented using the Baum-Welch algorithm (Rabiner, 1989). Consider a sequence $y_1\, y_2 \cdots y_t \cdots y_m$ in the set Y, where t is the step within the sequence, and m is the maximum length of the sequence. The algorithm is divided into three steps:

1. *Forward Probability* $\alpha_i(t)$: This is the probability of being at state i at step t, given that the preceding sequence is $y_1\, y_2 \cdots y_{t-1}$. It is computed iteratively using

$$\alpha_i(t+1) = \delta_i(y_{t+1}) \sum_{j=1}^{n} \alpha_j(t) T_{ji},$$

where the initial conditions are $\alpha_1(0) = 1$ and $\alpha_j(0) = 0$ for all $j \neq 1$. The indicator function $\delta_i(y_{t+1}) = 1$ if the symbol y_{t+1} matches the symbol s_i associated with state i, and $\delta_i(y_{t+1}) = 0$ otherwise.

2. *Backward Probability* $\beta_i(t)$: This is the probability of being at state i at step t, with the subsequent sequence being $y_{t+1}, \ldots, y_m$. It is calculated iteratively using:

$$\beta_i(t) = \delta_i(y_t) \sum_{j=1}^{n} T_{ij}\beta_j(t+1).$$

The initial conditions are $\beta_2(m+1) = 1$ and $\beta_j(m+1) = 0$ for all $j \neq 2$.

3. *Calculation of* N_i *and* N_{ij}: The forward and backward probabilities $\alpha_i(t)$ and $\beta_i(t)$ must be calculated for each sequence in Y. The number of transitions from state i to state j is given by

$$N_{ij} = \sum_{Y} \sum_{t=1}^{m} \alpha_i(t) T_{ij} \beta_j(t+1).$$

For a given sequence $y_1\, y_2 \cdots y_m$, the probability that the POMM generates the sequence is

$$P_y = \alpha_2(m+1),$$

which corresponds to the forward probability of reaching the end state at step $m+1$.

The total probability for the set Y is

$$P_Y = \prod_{y \in Y} P_y.$$

It is often more convenient to work with the log-likelihood, expressed as follows:

$$L_Y = \log P_Y = \sum_{y \in Y} \log P_y.$$

By iteratively updating the transition probabilities based on the forward and backward probabilities, the Baum-Welch algorithm adjusts T to increase the likelihood of reproducing the set of sequences Y (Rabiner, 1989).

Statistical tests. For comparing distributions of paired data, we use the Wilcoxon signed-rank one-side test, implemented in the Python module `scipy` via `scipy.stats.wilcoxon` (Virtanen et al., 2020). This test is non-parametric, hence there is no degree of freedom. Instead, the median difference is reported.

State merging. States i and j associated with the same syllable can be merged by eliminating state j. The transition probability from state i to state k ($k \neq i, j$) is recomputed as follows:

$$T_{ik} = \frac{N_{ik} + N_{jk}}{N_i + N_j}.$$

Here, N_i and N_j represent the number of times states i and j are visited, respectively, and N_{ik} and N_{jk} are the number of times the transitions from state i to state k and from state j to state k occur. These counts are initially obtained during the process of constructing a higher-order Markov model that is statistically compatible with the observed syllable sequences (see Results).

The transition probability from state k ($k \neq i, j$) to state i is updated as follows:

$$T_{ki} \rightarrow T_{ki} + T_{kj}.$$

The number of times state i is visited is updated to

$$N_i \rightarrow N_i + N_j.$$

After these updates, state j is removed from the POMM.

Results

The dataset of songs from six Bengalese finches, both before and shortly after deafening, was used in a previous study focusing on the phenomenon of syllable repetitions, particularly the influence of auditory feedback on these repetitions. Our findings suggest that long sequences of repeated syllables are best described by a non-Markovian process (Jin and Kozhevnikov, 2011). This process is characterized by a gradual decrease in the probability of a syllable repeating as the sequence progresses. We identified the underlying neural mechanism as synaptic adaptation in the auditory feedback pathway. Specifically, strong and adaptive auditory feedback is crucial for sustaining long syllable repetitions. Supporting this, we observed that removing auditory feedback through deafening led to a significant shortening of syllable repetitions (Wittenbach et al., 2015). It is important to note, however, that our analysis was specifically confined to syllable repetitions and did not extend to other aspects of song structure.

In this work, we investigate the context dependencies in the syllable sequences of six Bengalese finches and how auditory feedback might contribute to these dependencies using POMMs. Since syllable repetitions are best described by non-Markovian state transition models, we focus on the non-repetitive versions of the sequences, where only the first instance of any repeated syllable is retained. For example, if the syllable sequence is *ABBBC*, then the non-repetitive version is *ABC*. Throughout the rest of this paper, the term "syllable sequences" specifically refers to these non-repetitive versions. Each syllable sequence is typically preceded by a variable number of introductory notes, which are excluded from the analysis (Materials and Methods).

POMM for syllable sequences

A POMM for a set of syllable sequences consists of states and transitions between them. Each state is associated with a single syllable. Additionally, there are a start state and an end state. State transitions begin at the start state, with the next state being selected based on the transition probabilities from the current state. The process stops once the end state is reached. A path from the start state to the end state generates one syllable sequence.

A POMM is visualized using directed graphs (Fig. 1) generated with the software Graphviz (Ellson et al., 2001). Following the convention introduced previously (Jin and Kozhevnikov, 2011), the start state is represented as a pink node marked with the symbol α. All other states are shown as nodes labeled with their associated syllables, and syllables with multiple states are marked in red font. A state's color is cyan if it can transition to the end state and white otherwise. To reduce clutter, the end state is not displayed. When necessary, the symbol ω is used to denote the end of sequences.

State transitions are depicted with arrows color-coded according to their transition probabilities P. Strong transitions ($0.5 \leq P \leq 1$) are shown in red; medium transitions ($0.1 \leq P < 0.5$) in green; and weak transitions ($0.01 \leq P < 0.1$) in gray. To further reduce clutter, only transitions with $P \geq 0.01$ are shown.

Two types of context dependency

We construct two simple examples to demonstrate the existence of different types of context dependencies in syllable transitions (Fig. 1). These examples involve five syllables: *A*, *B*, *C*, *D*, and *E*. The transitions from *C* to either *D* or *E* depend on whether *C* is preceded by *A* or *B*. Example 1 illustrates a scenario where a syllable transition is either allowed or prohibited depending on the context. The observed sequences consist of two unique sequences: *ACD* and *BCE*, each with a probability of 0.5. The transition $C \rightarrow D$

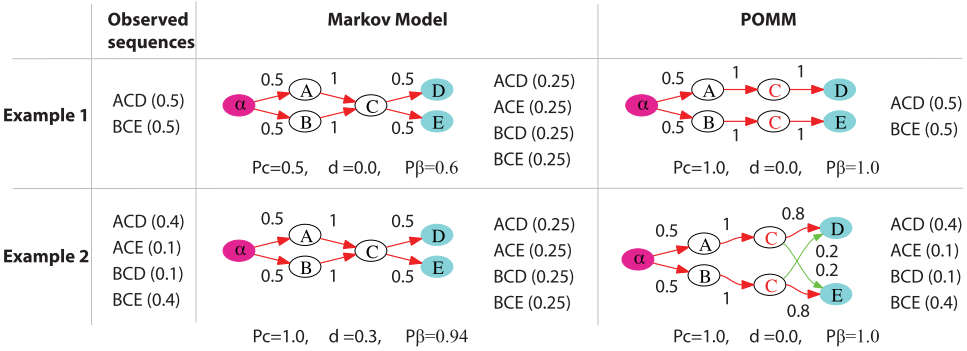

Figure 1. Two examples illustrating two types of context dependency. In Example 1, there are two unique sequences with equal probabilities (shown in parentheses) in the observed set. The Markov model overgeneralizes, resulting in two unobserved sequences and a sequence completeness of $P_c = 0.5$. The POMM with two states for syllable *C* avoids overgeneralization, achieving $P_c = 1$ and $P_\beta = 1$. In Example 2, the observed set comprises four unique sequences, with two being more frequent. The Markov model fails to capture these frequency differences (total variation distance $d = 0.3$), even though $P_c = 1$. However, the POMM with two states for syllable *C* successfully captures the probabilities of the unique sequences, indicated by $d = 0$ and $P_\beta = 1$. The color scheme for transition probabilities *P* in POMM diagrams: red arrows $P \geq 0.5$; green arrows $0.5 > P \geq 0.1$; gray arrows $0.1 > P \geq 0.01$.

occurs only when *C* is preceded by *A*, while the transition $C \rightarrow E$ occurs only when *C* is preceded by *B*. Consequently, sequences such as *ACE* and *BCD* are not observed. We refer to this form of context dependence as *type I context dependence*.

Example 2 illustrates a case where context dependence is reflected in the transition probabilities. The observed sequences consist of four unique sequences: *ACD* with a probability of 0.4, *ACE* with a probability of 0.1, *BCD* with a probability of 0.1, and *BCE* with a probability of 0.4. The transitions $C \rightarrow D$ and $C \rightarrow E$ occur regardless of the preceding syllable. However, the transition probabilities vary depending on whether *A* or *B* precedes *C*. We refer to this type of context dependence as *type II context dependence*.

A simple model that can be inferred from the observed sequences is the Markov model, where each syllable is associated with a single state. However, in both examples, the Markov model is too simplistic to capture the context dependencies present in the transitions.

The Markov model for Example 1 can be inferred by calculating transition probabilities from the observed sequences (Fig. 1). The sequences begin with either *A* or *B*, each with equal probability, so the start state transitions to the *A*-state or *B*-state with a probability of 0.5. Both of these states then transition to the *C*-state with a probability of 1. Since *C* can be followed by either *D* or *E*, the *C*-state transitions to the *D*-state or the *E*-state, each with a probability of 0.5. Finally, the *D*-state and the *E*-state transition to the end state with a probability of 1.

The Markov model fails because it overgeneralizes. Starting from the start state, there are four possible state transition paths, generating the sequences *ACD*, *ACE*, *BCD*, and *BCE*, each with a probability of 0.25. However, the sequences *ACE* and *BCD* are not observed in the data.

The issue of overgeneralization can be quantified using the concept of sequence completeness P_c, which is defined as the total probability of the model generating all unique sequences in the observed set:

$$P_c = \sum_{i=1}^{M} P_i,$$

where *M* is the number of unique sequences, and P_i is the probability of the *i*th unique sequence as computed by the model. The degree of overgeneralization is represented by $1 - P_c$, which corresponds to the total probability of the model generating sequences that are not present in the observed set. For Example 1, the unique sequences in the data are *ACD* and *BCE*. The Markov model assigns a probability of 0.25 to each of these sequences. Therefore, we calculate $P_c = 0.5$.

To avoid overgeneralization, we need to infer a more complex model, specifically a POMM where *C* is associated with two distinct states. The *A*-state and the *B*-state transition separately to these respective *C*-states (Fig. 1). This POMM generates the two sequences *ACD* and *BCE*, each with a probability of 0.5, resulting in $P_c = 1$ for the observed set. In other words, the model does not overgeneralize.

In Example 2, the Markov model fails not because of overgeneralization but due to misaligned probabilities with the data (Fig. 1). While the model generates all observed unique sequences, resulting in $P_c = 1$, it incorrectly represents the sequence probabilities. The less frequent sequences *ACE* and *BCD*, which each have a probability of 0.1 in the data, are assigned a higher probability of 0.25 by the model. Conversely, the more probable sequences *ACD* and *BCE*, each with a probability of 0.4 in the data, are underrepresented in the model with a probability of 0.25.

A simple measure of the differences in probabilities is the total variation distance (Gibbs and Su, 2002), defined as follows:

$$d = \frac{1}{2}\sum_{i=1}^{M} |P_{i,o} - P_{i,m}|.$$

Here,

$$P_{i,o} = \frac{N_i}{N}$$

represents the observed probability of the *i*th unique sequence, where N_i is the number of times the sequence appears and *N* is the total number of observed sequences.

On the other hand, $P_{i,m}$ is the normalized probability of the sequence computed with the model, defined as follows:

$$P_{i,m} = \frac{P_i}{P_c}.$$

The normalization ensures that

$$\sum_{i=1}^{M} P_{i,m} = 1,$$

which is necessary because we are comparing $P_{i,m}$ with $P_{i,o}$, where $\sum_{i=1}^{M} P_{i,o} = 1$. The value of d is 0 when the probabilities perfectly match, and 1 for a complete mismatch.

In Example 2, the Markov model has a total variation distance of $d = 0.3$. A more complex model, which includes two distinct states for C (as shown in Fig. 1), achieves $d = 0$, indicating that it effectively captures the context-dependent changes in transition probabilities.

The total variation distance may not fully reveal type I context dependence. In Example 1, the Markov model generates the two observed sequences, *ACD* and *BCE*, each with a probability of 0.25. After normalization, these probabilities become 0.5. As a result, the Markov model yields $d = 0$.

To capture both type I and type II context dependencies, we combine P_c and d into a single measure:

$$P_\beta = (1 - \beta)P_c + \beta(1 - d),$$

where β is the weight assigned to the total variation distance and is a value between 0 and 1. A perfect model would have $P_\beta = 1$. We refer to this combined measure as the *augmented sequence completeness*.

As we will show below, model selection based on P_β is not sensitive to the value of β, provided it is neither too close to 0 nor 1. A suitable choice of β should balance the variances of P_c and d to ensure that the variance of P_β is not dominated by either component. Since P_c is a sum of probabilities, it tends to be less sensitive to measurement errors, as positive and negative errors often offset one another. In contrast, d, which aggregates the absolute values of errors, is likely to be more sensitive to measurement inaccuracies. For this reason, we choose $\beta < 0.5$ to account for the difference in sensitivities. In this study, we set $\beta = 0.2$. This consideration is particularly important when the number of observed sequences is small. When a large number of sequences are available, probability measurements are more accurate, making the specific choice of β less critical.

The POMMs with two states for C in both Example 1 and Example 2 yield $P_\beta = 1$, regardless of the chosen value for β.

Inference of a minimal POMM from observed sequences

In this section, we demonstrate how a POMM can be inferred from a set of observed syllable sequences using a statistical test. The inferred POMM is considered minimal in the sense that it contains the fewest number of states among all POMMs that are statistically compatible with the observed sequences.

To illustrate this process, we first construct a "ground-truth" POMM, from which we generate a set of "observed sequences" (Fig. 2). We then apply the inference process to determine the minimal POMM. Finally, we compare the inferred POMM with the original ground-truth POMM for validation.

Statistical test of a POMM

To identify a POMM that aligns with the observed set of syllable sequences, it is necessary to develop a method for statistically evaluating the model's fit. This evaluation can be framed as a hypothesis test, where the null hypothesis posits that the observed set is generated by the POMM. The measure P_β can be used for this evaluation. Ideally, the P_β value for the observed set, as computed with the POMM, should be 1. This would indicate that the POMM accurately generates all the unique sequences present in the observed set and does not produce unobserved sequences. Additionally, it suggests that the probabilities of these unique sequences are consistent with the observations. However, in practice, due to the finite number N of observed sequences, the observed set might not include all the sequences that a bird is capable of producing. As a result, a P_c value less than 1 could be due to the limited size of N, rather than overgeneralization by the model. Moreover, discrepancies in the probabilities of the unique sequences may also arise from imprecise probability measurements when N is finite.

To account for the finite N effect, we generate random sets of N sequences from the POMM. For each generated set, we compute P_β using the POMM. The distribution of P_β for these generated sets helps assess the likelihood that the observed P_β is part of this distribution. Specifically, we calculate the probability p that the observed P_β exceeds the P_β values of the generated sets. If $p < 0.05$, we infer that the observed P_β is unlikely to have been drawn from this distribution, then leading to the rejection of the POMM as a model for the observed set. Conversely, if $p \geq 0.05$, the POMM is not statistically rejected and is therefore accepted as a model for the observed set. To construct the P_β distribution, we generate 10,000 random sets of N sequences from the POMM.

To demonstrate this process, we use the "ground truth model" shown in Figure 2*a*. This model consists of two states for syllables *A* and *C*, and one state for each of the syllables *B*, *D*, and *E*. The model can generate seven unique sequences with the following associated probabilities: *A* with a probability of 0.1, *ACD* with a probability of 0.36, *ACE* with a probability of 0.04, *BCD* with a probability of 0.05, *BCE* with a probability of 0.2, *BAE* with a probability of 0.125, and *BA* with a probability of 0.125. The sequences produced by the ground truth model exhibit both type I and type II context-dependent syllable transitions.

The ground-truth model is evidently non-Markovian. Ideally, when N is large, the Markov model inferred from the observed sequences should be rejected by the statistical test using P_β. To demonstrate this, we generate three sets of observed sequences from the ground-truth model with $N = 10$, $N = 30$, and $N = 90$. For each set, we infer the Markov models and subsequently test their fit.

Markov models are inferred by analyzing the observed sequences. We count the number of transitions between syllables, as well as transitions from the start state to the syllables and from the syllables to the end state. These counts are then converted into transition probabilities through normalization (see Materials and Methods). The Markov model inferred from the set with $N = 30$ is illustrated in Figure 2*b*. To assess the validity of a Markov model, we generate 10,000 random sets of N sequences using the model. For each set, we compute P_β, resulting in a distribution of P_β, as shown in Figure 2*c*. We then compare this distribution to the P_β of the observed set.

As N increases, the distribution of P_β shifts towards 1. This shift occurs because a larger N ensures that most of the unique sequences the model can generate are included in a generated set, leading to $P_c \to 1$ for these sets. Additionally, the probabilities of the unique sequences computed from the generated sets align more closely with those calculated from the model.

However, the P_β values for the observed sets, indicated by red lines in the figure, remain relatively unchanged as N increases. This stability arises because the observed sets are generated using the ground-truth model, which is non-Markovian. Therefore, increasing N does not improve P_c for the observed sets. As N grows, the P_β for the observed sets consistently falls below that of the generated sets, indicating that the Markov models are not statistically compatible with the observed sets.

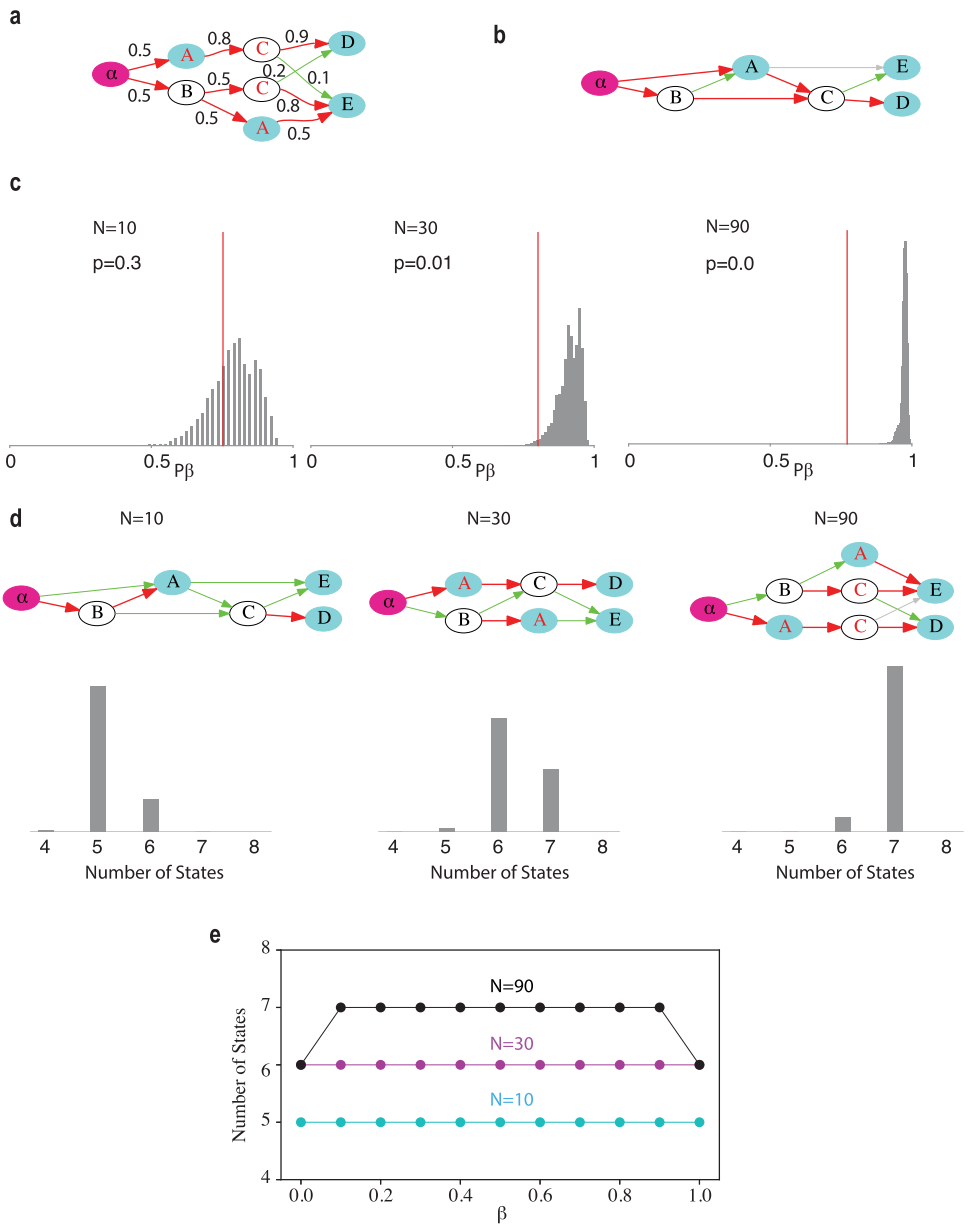

Figure 2. An example for statistically testing and inferring POMMs from a finite set of observed sequences. ***a***, The ground truth POMM from which the observed sequences are generated. The numbers near the arrows are the transition probabilities. ***b***, The Markov model for a set of $N = 30$ observed sequences. ***c***, Statistical tests of the Markov models for $N = 10$, 30, 90. The gray bars are the distributions of P_β for the generated sets. The red lines are the P_β for the observed sets. ***d***, POMMs inferred from the observed sets, and distributions of the number of states in the POMMs for 100 runs. ***e***, The impact of the choice of β. The median number of states in the inferred POMMs for various values of β.

In the examples shown in the figure, the p-values are $p = 0.3$ for $N = 10$, $p = 0.01$ for $N = 30$, and $p = 0$ for $N = 90$. The results may vary due to differences in the samples of the "observed sequences" generated from the ground-truth model. To illustrate these fluctuations, we repeated the process 100 times for each N. For $N = 10$, the results were $p = 0.3 \pm 0.3$ (mean ± standard deviation); for $N = 30$, $p = 0.01 \pm 0.03$; and for $N = 90$, $p = 0 \pm 0$.

Using the criterion of $p < 0.05$, the Markov model is rejectable for $N = 30$ and $N = 90$. However, for $N = 10$, the model cannot be rejected, despite the ground-truth model being non-Markovian. This occurs because, when N is too small, there is insufficient evidence to conclusively reject the Markov model.

If the ground-truth model is Markovian, increasing N does not lead to the rejection of the Markov model, then which is consistent with expectations.

While the Markov model is used as an example, the statistical testing process based on P_β can be applied to evaluate any POMM.

Searching the state space for a minimal POMM

Using the statistical test described above, we can identify the minimal POMM that is statistically compatible with a set of observed sequences. This is achieved by exploring the state space of possible POMMs. First, we identify a POMM that passes the statistical test. Then, we simplify the model by merging and deleting states until the POMM is reduced to its simplest form that still satisfies the statistical test requirements.

To search for a POMM compatible with the observed set, we begin by constructing higher-order Markov models. For the mth-order Markov model, we first flank each sequence in the

observed set with the start symbol α and the end symbol ω. We then collect unique subsequences of length m, as well as subsequences up to length m that start from α. For example, with $m = 2$, the sequence $\alpha A\ C\ D\omega$ yields the unique subsequences: α, αA, AC, CD, and $D\omega$.

Each unique subsequence is assigned to a state. The subsequence α is assigned to the start state, while all subsequences ending with ω are assigned to the end state. The remaining unique subsequences are assigned to distinct states, with the final syllable of each subsequence serving as the symbol for that state. This assignment transforms each observed sequence into a sequence of states, and the transition probabilities between these states are calculated by counting the number of transitions. This method produces a POMM equivalent to the mth order Markov model.

We apply the statistical test to determine if the POMM satisfies the acceptance criterion of $p \geq 0.05$. Starting from $m = 1$, which corresponds to the basic Markov model, we incrementally increase m until the POMM is accepted.

We next simplify the POMM by merging and deleting states associated with the same syllable. State transition probabilities are re-computed based on the counts of transitions between states (see Materials and Methods) (Jin and Kozhevnikov, 2011). A merge is retained if the resulting POMM passes the statistical test. After no further merges are accepted, we proceed to reduce the POMM through state deletion. If a syllable is associated with more than one state, then we reduce the number of states for that syllable by one. Transition probabilities between states are then computed by maximizing the log-likelihood that the model generates the observed sequences, using the Baum–Welch algorithm (see Materials and Methods) (Rabiner, 1989). To avoid local minima in the algorithm, we perform 100 runs with random seeds and select the run with the highest log-likelihood. If the reduced POMM is accepted, the deletion process continues. Through this state reduction procedure, we obtain a POMM with the minimal number of states that passes the statistical test based on P_β.

After state reduction, we simplify the transitions between the states in the POMM. We systematically remove each transition and recalculate the transition probabilities using the Baum–Welch algorithm. If the log-likelihood of the observed sequences after a cut remains above a predefined threshold, the cut is accepted; otherwise, the transition is retained. The threshold is set to the log-likelihood before the cuts, minus an estimate of the fluctuation in the log-likelihood due to inaccuracies in computing the transition probabilities. This estimate is the standard deviation of the log-likelihoods before the cuts, calculated from the 100 runs of the Baum-Welch algorithm with random seeds. If the POMM after the cuts does not pass the statistical test, then the cuts are reversed.

Evaluation with the ground truth model

We evaluate the above procedure using the ground truth model (Fig. 2*a*). A set of N "observed sequences" is generated from the ground truth model, from which we infer the minimal POMMs. To assess the impact of sampling, this process is repeated 100 times. The results for $N = 10$, $N = 30$, and $N = 90$ are shown in Figure 2*d*. We present typical inferred POMMs along with the distributions of the total number of states in the POMMs inferred from the 100 sets.

For $N = 10$, the total number of states is predominantly 5, and the Markov model is generally accepted. Some models have 4 states, as syllables D or E may not appear in the observed sequences due to the small sample size. When $N = 30$, the total number of states varies between 5 and 7, with a typical POMM having 6 states, as illustrated in the figure. For $N = 90$, the total number of states is primarily 7, and the inferred POMMs closely resemble the structure of the ground truth model.

These results were obtained with $\beta = 0.2$. We also tested the impact of different values of β. Across a wide range of β, the median number of states in the inferred POMMs remained largely insensitive to the choice of β (Fig. 2*e*).

This example shows that our procedure tends to infer a simpler POMM than the ground truth model when N is small. Conversely, when N is large, the procedure successfully recovers the ground truth model. Notably, the procedure does not generate models more complex than the ground truth model.

POMMs of Bengalese finch songs in the normal condition and after deafening

To investigate the effect of auditory input on the song syntax of the Bengalese finch, we analyzed the songs of six adult Bengalese finches both before and 2–4 days after deafening. This dataset has been previously utilized for analyzing syllable repeats (Wittenbach et al., 2015). In this study, we focus on the non-repetitive versions of the syllable sequences.

Test of Markov models on the Bengalese finch songs

We tested whether Markov models are statistically compatible with the observed syllable sequences using the $p \geq 0.05$ criterion. The results are shown in Figure 3. For three birds, the syllable sequences were found to be incompatible with the Markov models both before and after deafening (o10bk90, normal: $p = 0$, deafened: $p = 0$; bfa16, normal: $p = 0$, deafened: $p = 0$; o46bk78, normal: $p = 0$, deafened: $p = 0$). For the other three birds, the sequences were incompatible with the Markov models before deafening, but post-deafening, the Markov models were accepted (bfa7, normal: $p = 0$, deafened: $p = 0.4$; bfa14, normal: $p = 0$, deafened: $p = 0.5$; bfa19, normal: $p = 0.03$, deafened: $p = 0.3$). These results suggest that deafening may reduce the Bengalese finch song syntax from non-Markovian to Markovian for some birds, though this effect is not universal.

Deafening results in the creation of novel transitions between syllables, as well as new starting and ending syllables. The transition probabilities for these novel transitions tend to be low, with a median of 0.04. However, 22% of these novel transitions exceed a probability of 0.1, representing 18 out of 81 transitions. The majority of these novel transitions are observed in two birds (27 for bfa14; 21 for bfa19). Additionally, a small number (8) of transitions disappear following deafening, with a median transition probability of 0.02 in the normal condition.

As observed in the previous studies, deafening increases sequence variability (Okanoya and Yamaguchi, 1997; Woolley and Rubel, 1997). The variability of transitions from a given syllable i (or the start state) is quantified using the transition entropy, $S_i = -\sum_{j=1}^{M} p_{ij} \log_2 p_{ij}$, where M represents the number of branches of the transitions, and p_{ij} is the probability of the jth branch. If $M = 1$, then the transition is stereotypical, and S_i equals zero. For a given M, the entropy reaches its maximum when the transition probabilities for all branches are equal, and this maximum entropy increases with M. The median of transition entropies is significantly higher after deafening (0.95 ± 0.55) than before (0.35 ± 0.51; Wilcoxon signed-rank one-sided test, $p = 5.4 \times 10^{-6}$, median difference: 0.31). Similarly, the number of branches M is also significantly larger after deafening (4 ± 1.5, median $\pm$ s.d.)

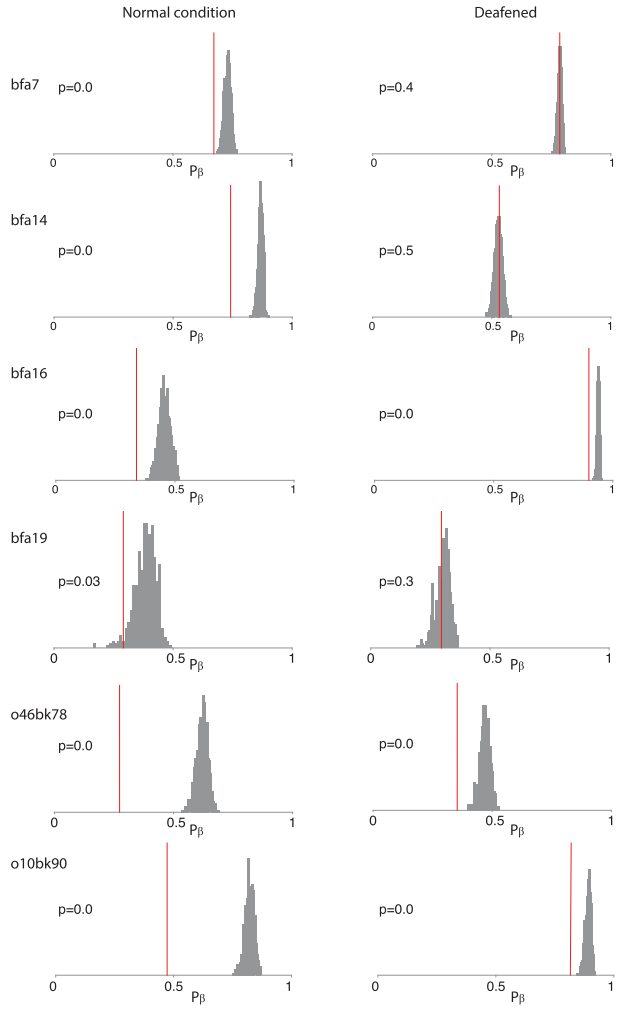

Figure 3. Statistical tests of the Markov models for all birds. The distributions of P_β for the generated sets and P_β of the observed set (red line) are shown. The *p*-values are displayed.

compared to before (2 ± 0.90; Wilcoxon signed-rank one-sided test, $p = 9.8 \times 10^{-7}$, median difference: 1.0).

POMMs of the Bengalese finch songs

We inferred minimal POMMs from the observed syllable sequences before and after deafening in six birds (Figs. 4, 5). In the normal condition, the birds' songs comprised 44 syllables, of which 26 required 1 state, 13 required 2 states, 2 required 3 states, and 3 required 4 states. Thus, the majority of syllables necessitated 1 or 2 states. The POMMs encompassed 76 states in total. When considering only transition branches with probabilities greater than 0.01, most states had up to 3 outgoing branches, with 29, 31, and 11 states having 1, 2, and 3 branches, respectively.

After deafening, there were 43 syllables (syllable *g* for bfa7 dropped out post-deafening). Most syllables (40) required only 1 state, with the remaining 3 requiring 2 states. In total, the POMMs comprised 52 states. Counting only transition branches with probabilities greater than 0.01, the states had up to 7 outgoing branches (there were 2, 19, 7, 13, 6, 3, and 2 states with 1 to 7 branches, respectively).

Deafening significantly reduces state multiplicity, as evidenced by the decrease in the number of extra states (defined as the difference between the number of states for the syllables and the number of syllables) (Fig. 6*a*; Wilcoxon signed-rank one-sided test, $p = 0.016$, median difference: −2.5). Additionally, the mean normalized transition entropy between states, which is the transition entropy divided by $\log_2 M$, is notably higher after deafening in all but one bird (Fig. 6*b*; Wilcoxon signed-rank one-sided test, $p = 0.031$, median difference: 0.21). These findings indicate that deafening reduces context dependencies in syllable sequences, as demonstrated by the diminished state multiplicity. Furthermore, the transitions between states become more random post-deafening.

In our previous work, n-gram distributions, representing the probabilities of n-length subsequences, were used for fitting POMMs (Jin and Kozhevnikov, 2011). In contrast, our current method does not rely on n-gram distributions. Nonetheless, the 3- to 7-gram distributions of the sequences generated by the POMMs for the six birds, in both normal and deafened conditions, agree with those observed in the actual syllable sequences (Figs. 7, 8). This provides further validation of the POMMs we have inferred here.

Neural mechanisms of the state multiplicity

Experiments that have recorded neural activity of the HVC neurons projecting to downstream motor areas in the zebra finch have demonstrated that each syllable is driven by sequential bursts of a specific set of projection neurons (Hahnloser et al., 2002; Long et al., 2010; Lynch et al., 2016; Picardo et al., 2016; Egger et al., 2020; Moll et al., 2023). A projection neuron bursts once during a specific syllable but remains inactive during others, suggesting that different syllables are driven by distinct sets of neurons with no overlap. These neural dynamics are effectively modeled using synaptic chain networks of projection neurons (Fig. 9) (Jin et al., 2007; Long et al., 2010; Miller and Jin, 2013; Egger et al., 2020; Tupikov and Jin, 2021). Both experimental evidence and computational models strongly support the idea that syllable-chains in the HVC underlie the production of individual syllables (Fee et al., 2004; Jin, 2009; Moll et al., 2023).

Within this framework, variable syllable transitions at branching points can be understood as the result of winner-take-all competitions mediated by inhibitory HVC interneurons between the syllable-chains for the alternative syllables (Chang and Jin, 2009; Jin, 2009; Wittenbach et al., 2015). The activation of competing syllable-chains can be driven by the activity of the preceding syllable-chain (Jin, 2009), and/or by auditory feedback from the preceding syllables (Wittenbach et al., 2015).

The intrinsic model provides a simple way to correlate a POMM with the networks of projection neurons in the HVC. In this model, each state corresponds to a syllable-chain, with syllables being activated primarily by the activities of preceding syllables. If the POMM assigns multiple states to the same syllable, then it implies that there are multiple corresponding syllable-chains driving that syllable. For instance, in Example 1 (Fig. 1), where two unique observed sequences, *ACD* and *BCE*, are present, the POMM assigns two distinct states for syllable *C*. Thus, the intrinsic model requires two distinct syllable-chains for *C*, as shown in Figure 9*a*. In this configuration, the syllable-chain for *A* (chain-A) connects to one of the chain-Cs, which then connects to chain-D. Similarly, chain-B connects to the other chain-C, which subsequently connects to chain-E. Each syllable-chain is thus responsible for driving a specific syllable, with different chains for *C* depending on the preceding syllable (either *A* or *B*).

However, since the intrinsic model does not account for the role of auditory feedback, it is unable to explain the reduction in state multiplicity observed after deafening. As a result, the

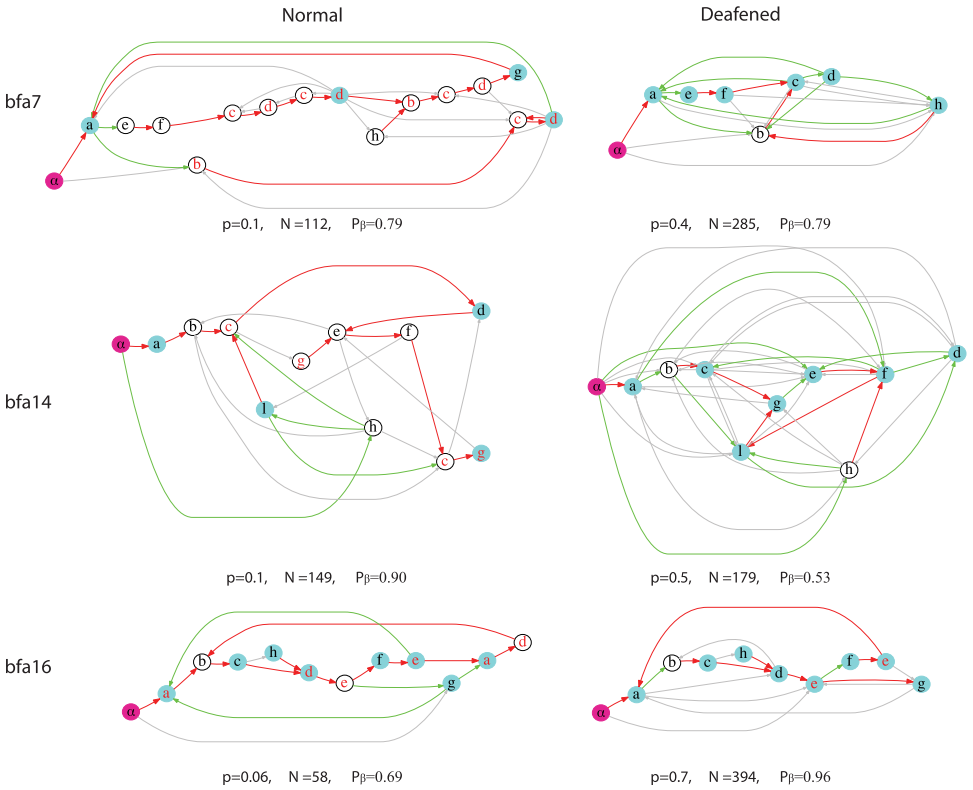

Figure 4. POMMs before and after deafening for three birds. The results for bfa7, bfa14, and bfa16 are shown. The *p*-values, the number of sequences in the observed sets, and P_β of the observed sets are shown.

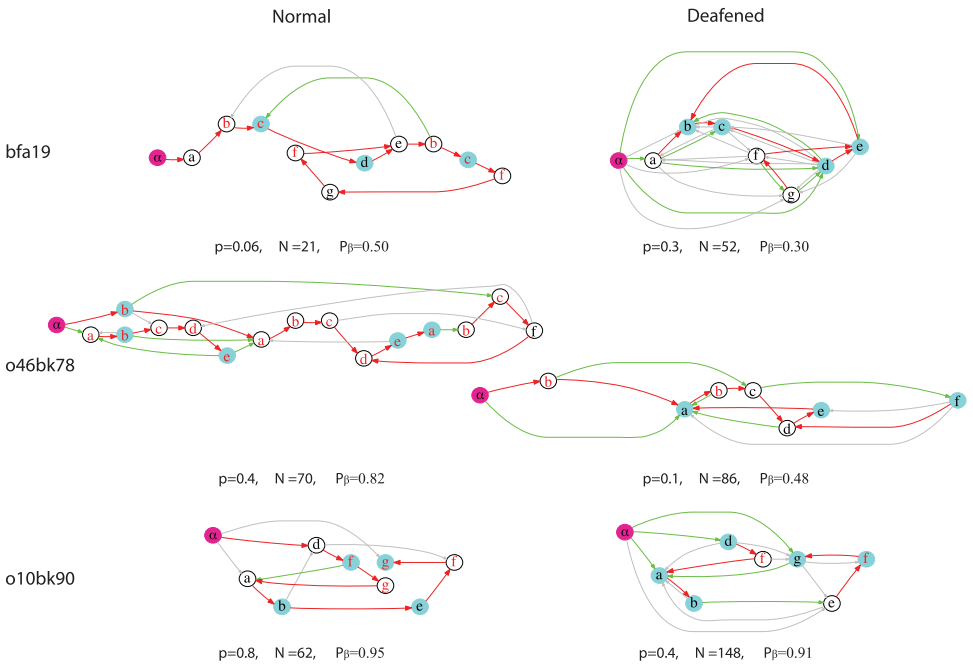

Figure 5. POMMs before and after deafening for the other three birds. The results for bfa19, ok46bk78, and o10bk90 are shown.

intrinsic model fails to capture this critical aspect of the data, and thus can be ruled out.

An alternative to the intrinsic model is the reafferent model, in which auditory feedback from preceding syllables determines the syllable transitions (Sakata and Brainard, 2006, 2008; Hanuschkin et al., 2011; Wittenbach et al., 2015). In this model, each syllable has only one syllable-chain. The presence of multiple states for the same syllable in the POMMs is attributed to the differential effects of auditory feedback from preceding syllables on the subsequent transitions.

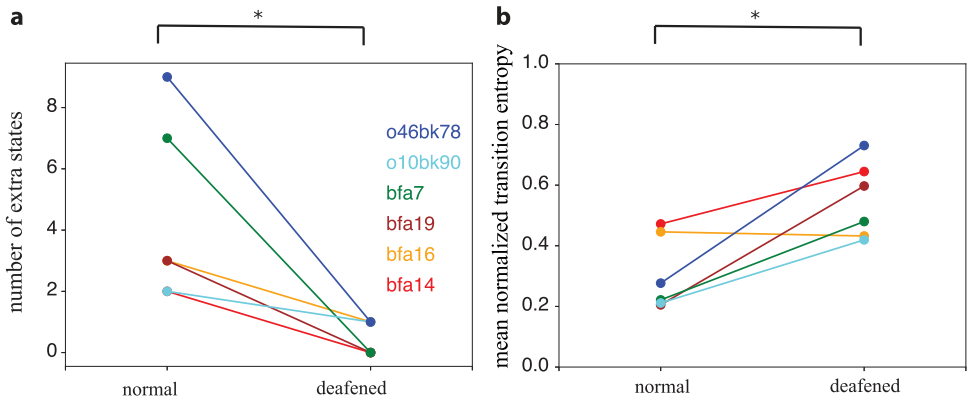

Figure 6. Effects of deafening on the POMMs. ***a***, The numbers of extra states for the syllables are significantly reduced. ***b***, The transitions from the states become more random after deafening.

This concept is illustrated with Example 1 in Figure 9*b*. In this case, there is a single syllable-chain for *C*, which connects to both chain-D and chain-E. However, the activation of either chain-D or chain-E is determined by the reafferent auditory inputs (Sakata and Brainard, 2006, 2008; Hanuschkin et al., 2011; Wittenbach et al., 2015). Specifically, the auditory feedback from syllable *A* is directed toward chain-D, while the auditory feedback from syllable *B* is directed toward chain-E (Fig. 9*b*). These auditory inputs bias the transitions from chain-C to either chain-D or chain-E (Jin, 2009; Hanuschkin et al., 2011; Wittenbach et al., 2015). With sufficiently strong auditory inputs, the probability of transitioning from *C* to *D* approaches 1 when *C* is preceded by *A*. Conversely, when *C* is preceded by *B*, the transition probability to *E* approaches 1.

Two pieces of evidence cast doubt on the reafferent model. First, while deafening significantly reduces the state multiplicity in the POMMs, it does not completely eliminate it. This suggests that multiple states cannot be entirely explained by auditory feedback alone, implying that other factors contribute to the state multiplicity. Second, the durations of syllables and the silent gaps between them can be longer than the time required for auditory feedback to reach the HVC. By the time transitions to the next syllables occur, the auditory feedback from preceding syllables may have faded away, making it unlikely that it provides sufficient context for syllable transitions. This temporal gap raises questions about the reafferent model's capacity to explain context-dependent syllable transitions.

We illustrate the second point with the example in Figure 9*b*. For the activity in chain-A to effectively bias the transition from chain-C toward chain-D, the auditory feedback signal from syllable *A* must still be present at the time of the transition from chain-C. This requires that the "round-trip" delay–from the activity in chain-A to the corresponding auditory feedback reaching the HVC–be longer than the duration of chain-C. In this case, the duration of activity propagation in chain-C should encompass both the duration of syllable *C* and the preceding silent gap. If the round-trip delay is shorter than the combined duration of syllable *C* and the silent gap, then the auditory feedback from syllable *A* will have faded before it can influence the transition from syllable *C*, thereby challenging the viability of auditory feedback to consistently govern context-dependent transitions.

In Figure 10, we plot the distributions of the gap-syllable durations for all syllables that require multiple states under the normal condition for the six birds. The median values of these durations range approximately from 80 ms to 180 ms. For syllables that are repeated, the auditory feedback must be available after the completion of the syllable repetitions. Including these repetitions, the median values of the durations extend approximately from 80 ms to 600 ms. These ranges establish a lower limit on the delays of the auditory feedback required for the reafference model to function effectively.

Experiments that perturbed auditory feedback during singing in Bengalese finches found that sound perturbations altered HVC activity approximately 44 ms after the onset of the perturbation, with no significant changes in HVC activity detectable beyond 80 ms (Sakata and Brainard, 2008). Given that the premotor delay from HVC to syllable production is around 50 ms (Schmidt, 2003), the round-trip delay of auditory feedback from a syllable is limited to approximately 130 ms. Therefore, the reafference mechanism can be ruled out for at least 28% of syllables with state multiplicity, as the median durations of gap-syllables for these syllables exceed 130 ms (Fig. 10). This fraction increases to 44% when syllable repetitions are included, further challenging the viability of the reafference model for explaining state multiplicity in these cases.

A model that combines elements of both the intrinsic and reafferent models can explain the effects of deafening on state multiplicity without requiring long delays of auditory feedback (Fig. 9*c*). In this hybrid model, similar to the intrinsic model, each state in the POMM corresponds to a synaptic-chain network. Therefore, for a syllable requiring multiple states, there are multiple corresponding syllable-chains for that syllable. However, unlike the intrinsic model, the connections between the syllable-chains are more divergent. In this model, auditory feedback from a syllable "tunes" the syllable transitions that follow it, reinforcing transitions that encode context dependence. When deafening occurs, this feedback is lost, "de-tuning" the transitions and reducing context dependence, thus explaining the observed reduction in state multiplicity following deafening.

In Figure 9*c*, we illustrate this auditory-tuning model using Example 1. Chain-A and chain-B have intrinsic connections to both chain-Cs. The auditory feedback from chain-A biases the transition toward the chain-C that leads to chain-D, while the auditory feedback from chain-B biases the transition to the chain-C that leads to chain-E. When auditory feedback is lost due to deafening, the network reverts to relying solely on the intrinsic connections. This allows transitions from both chain-A and chain-B to either of the chain-Cs, resulting in less context-specific transitions and a reduction in state multiplicity.

The auditory-tuning model suggests that the reduction of the state multiplicity in POMMs after deafening can be explained by deafening-induced changes in the transition probabilities between

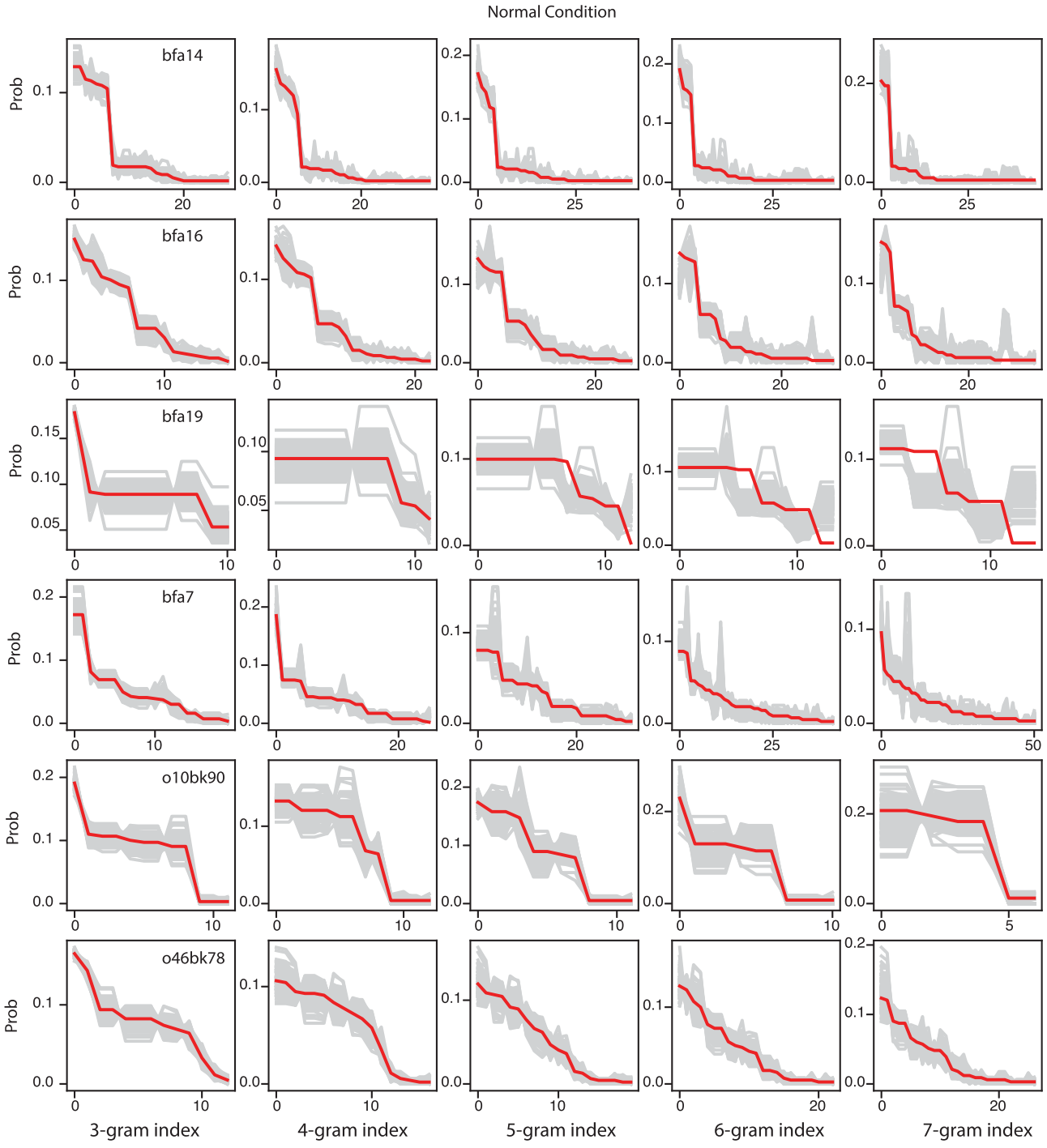

Figure 7. Comparisons of n-gram distributions in the normal condition. The redlines are the probabilities of n-grams in the observed sets. The n-grams are ordered in a descending order of probabilities in the observed sets. The gray lines are the probabilities of the ordered n-grams computed from 100 sets of sequences generated from the POMMs. The number of sequences in each generated set is matched to the number of sequences in the observed set. The red line is mostly within the range defined by the gray lines.

the states. To investigate this possibility, we consider each state in the POMMs under the normal condition as syllable-chains, then modify the transition probabilities between them. These modifications are informed by the pairwise syllable transition probabilities observed after deafening. If a novel transition from syllable x to syllable y emerges after deafening (with a probability of $\geq$0.01), then new transitions are established from all states associated with x to all states associated with y. Conversely, if the transition from x to y is eliminated after deafening, then we remove the transitions from all states associated with x to all states associated with y. We further refine the transition probabilities between the states. From a state associated with x, if transitions exist to states associated with y, then the probabilities of these transitions are scaled to ensure the total transition probability matches the transition probability from x to y after deafening. These modifications of the transition probabilities could diminish the context dependencies, potentially simplifying the POMMs to versions with reduced state multiplicity, which we refer to as reduced POMMs. Importantly, these modifications do not increase state multiplicity, as the number of syllable-chains remains unchanged.

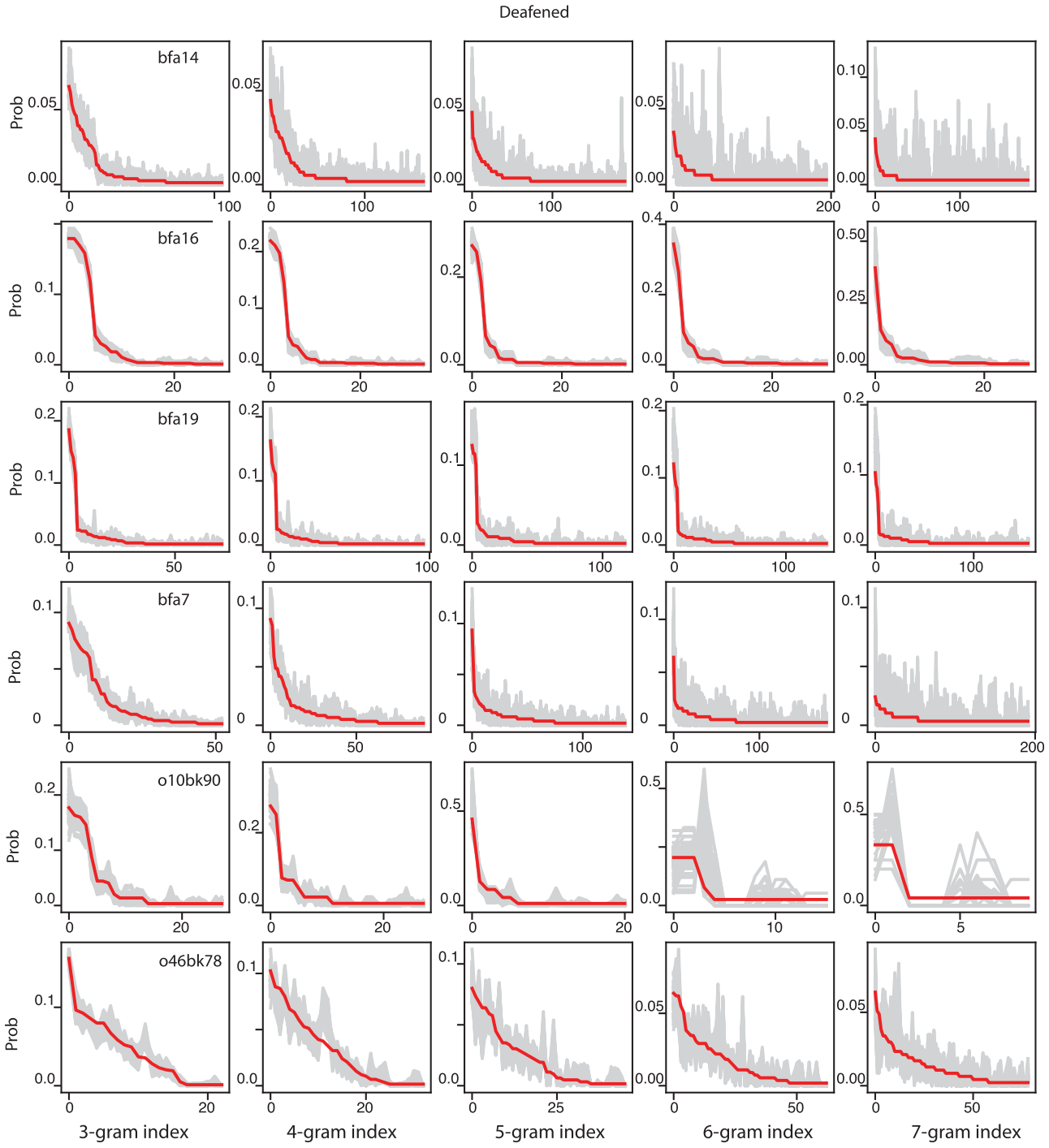

Figure 8. Comparisons of n-gram distributions after deafening. The same as in Figure 7 for the deafened case.

The reduction process unfolds as follows: With a modified POMM, we generate 100 sets of N sequences, where N represents the number of sequences observed for the deafened bird. From these sets, we select the set with the maximum probability given the modified POMM, designating it as the "observed syllable sequences." Using these "observed sequences," we infer a minimal POMM, which represents the reduced POMM after modifying the transitions. This process is illustrated in Figure 11*a–c*, using o10bk90 as an example. The reduced POMM for this bird matches the POMM observed after deafening in terms of the number of states for each syllable, although the transition probabilities differ slightly due to fluctuations introduced by the finite N. Similarly, the reduced POMMs for all other birds align with the POMMs post-deafening, with the exception of o46bk78, where syllable *c* retains an additional state, as shown in Figure 12. These results provide supporting evidence for the auditory-tuning model as the neural correlate of the POMMs.

Comparison to higher-order Markov models

As discussed above, a higher-order Markov model is equivalent to a POMM. We construct the higher-order Markov model with the minimal order that is compatible with the observed

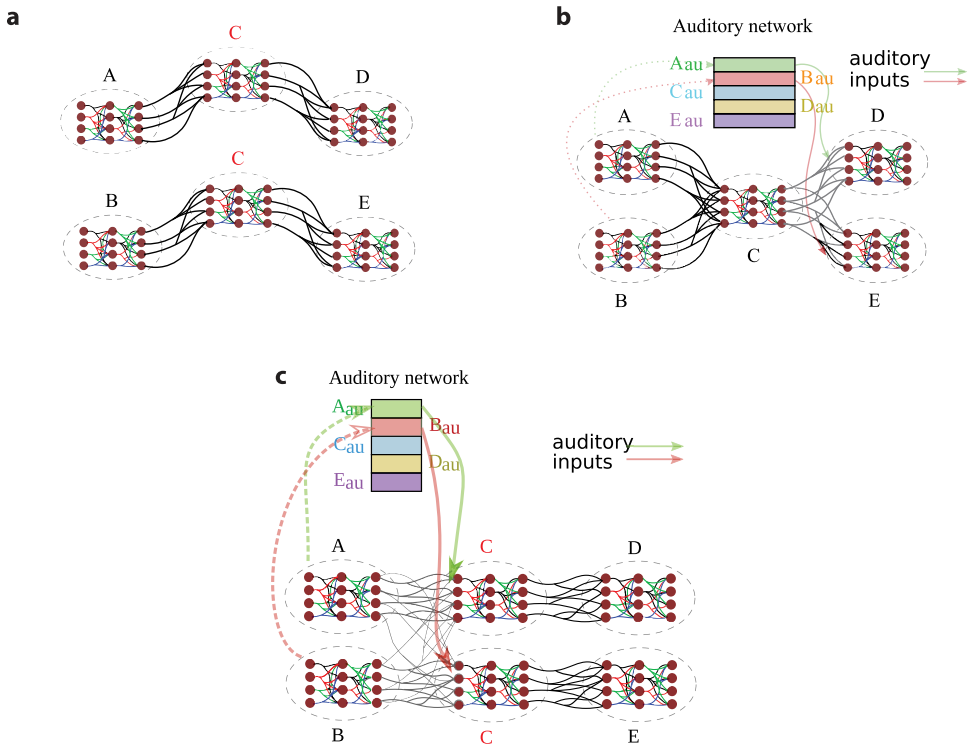

Figure 9. Neural mechanisms for the POMM in Example 1 (Fig. 1). There are two unique sequences, *ACD* and *BCE*, in the observed set. ***a***, The intrinsic mechanism for the multiple states of syllable *C*. Two syllable-chains encode the two states for *C*. ***b***, The reafference mechanism for the multiple states of syllable *C*. There is one syllable-chain for *C*, with the multiple states arising due to the auditory feedback from preceding syllables differentially influencing the syllable transitions from *C*. ***c***, The auditory-tuning model for the multiple states of syllable *C*. There are two syllable-chains for *C*. Intrinsic connections exist from chain-A and chain-B to the two chain-Cs. Auditory feedback from syllable A biases the transition from chain-A to the upper chain-C, while auditory feedback from syllable B biases the transition from chain-B to the lower chain-C.

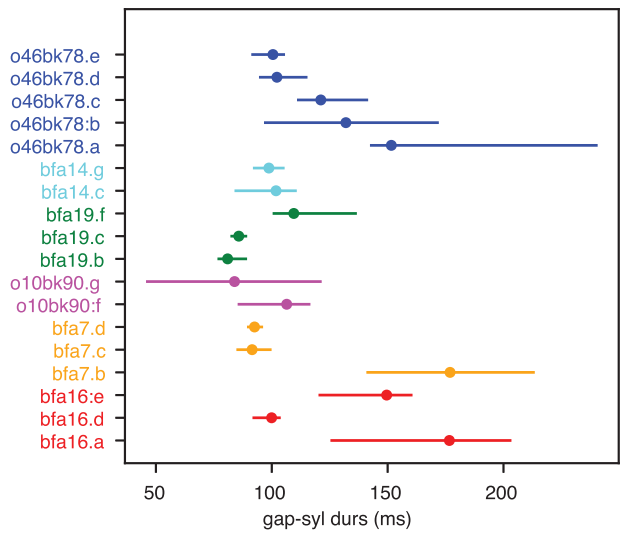

Figure 10. Distributions of gap-syllable durations. Durations of syllables plus the preceding gaps are shown for all syllables with multiple states in the POMMs for all birds. The dots indicate the median values and the bars indicate the 5%–95% ranges of the distributions.

syllable sequences, using the statistical test based on P_β. Subsequently, we reduce this initial POMM by merging and deleting states until we arrive at the minimal POMM, which contains the fewest number of states while still remaining compatible with the observed sequences. Higher-order Markov models have been widely used to model birdsong sequences (Dobson and Lemon, 1979; McLean and Roach, 2021). Therefore, it is useful to compare the initial POMMs, which are equivalent to the higher-order Markov models compatible with the observed sequences, with the minimal POMMs inferred for all birds, both before and after deafening (Figs. 4, 5).

As shown in Tables 1 and 2, the number of states in the initial POMMs is much higher than in the minimal POMMs. Furthermore, the state multiplicity in the initial POMMs does not accurately capture the context dependencies in syllable sequences. For example, consider the bird o10bk90 after deafening: the inferred POMM has an extra state for syllable *f* (Fig. 5). In contrast, the initial POMM has 5, 3, 4, and 2 extra states for syllables *a*, *e*, *g*, and *f*, respectively.

The orders of the higher-order Markov models ranged from 2 to 6 under normal conditions (Table 1), but from 1 to 2 after deafening (Table 2). This reduction in order supports our conclusion that deafening reduces context dependencies in syllable transitions. However, the higher-order Markov models are overly complex and fail to identify which syllables are crucial for generating these context dependencies.

Discussion

POMMs are generative models well-suited for capturing the structure of birdsong syllable sequences (Jin and Kozhevnikov, 2011). POMMs address context dependencies by associating multiple states with the same syllable. A state transition path, from the start state to the end state, generates a sequence of syllables. A syllable with state multiplicity can transition in various ways depending on its specific state, thereby participating in context-dependent transitions to other syllables. If a POMM lacks adequate state multiplicity, then it may overgeneralize,

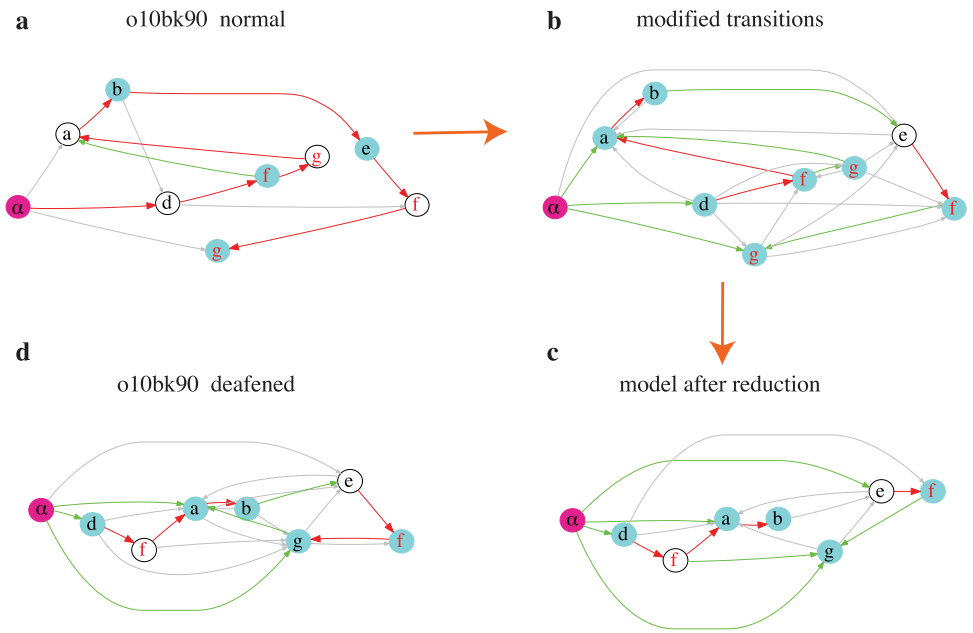

Figure 11. Reduction of a POMM after modifications of the transition probabilities. ***a***, The POMM of o10bk90 in the normal condition. ***b***, The POMM after modifications of the transitions probabilities using the syllable transition probabilities after deafening. ***c***, Reduced POMM inferred from the $N = 62$ "observed sequences" generated from modified POMM. ***d***, The POMM after deafening shown for comparison with the reduced POMM.

producing sequences that are either never observed (type I context dependence) or disproportionately increasing the likelihood of certain sequences compared to what is observed (type II context dependence). By statistically testing whether a POMM overgeneralizes relative to the observed syllable sequences, we can identify a minimal POMM—one that has the fewest states while still being compatible with the observed set. In an analysis of minimal POMMs for Bengalese finch songs, under normal conditions and shortly after deafening, we demonstrate that auditory feedback plays a crucial role in generating context-dependent syllable transitions in Bengalese finch songs.

To evaluate the fit of a POMM, we first construct the distribution of augmented sequence completeness, P_β, based on sequences sampled from the candidate POMM. This distribution is then used to calculate the p-value of the observed set's P_β, as computed by the POMM. We reject the POMM if $p < 0.05$. Lowering the p-value threshold allows for the acceptance of simpler POMMs with fewer extra states.

Our method is conservative. When the number of observed sequences N is small, the inferred POMM may underestimate the true number of states due to the limited representation of context dependencies in the observed sequences. A practical approach to assess whether N is sufficiently large is to examine if the sequence completeness P_c, as calculated using POMM, approaches 1 for the observed sequences. The difference, $1 - P_c$, can serve as a rough estimate of the total probability of missing unique sequences.

Two common methods for model selection are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Zucchini and MacDonald, 2009). For a POMM, the AIC is defined as $2k - 2L$, where k is the number of transition probabilities and L is the log-likelihood of the observed sequences given the POMM. The BIC is defined as $k \log N - 2L$, where N is the number of observed sequences. Among candidate POMMs, the model with the lowest AIC or BIC is selected to balance model fit and complexity. In tests using the simple example shown in Figure 2, we find that both AIC and BIC tend to underestimate state multiplicity compared to our method. AIC and BIC require comparison between models and often necessitate the enumeration of possible models to identify the minimal one. In contrast, our method can directly evaluate a single model, allowing the initial acceptance of a complex model, which can then be simplified by merging or deleting states.

Hidden Markov models (HMMs) are commonly used to model sequences (Rabiner, 1989), including Bengalese finch songs (Katahira et al., 2011). Like POMMs, HMMs involve probabilistic state transitions. However, unlike a POMM, a state in an HMM can emit any syllable, with the emission probabilities needing to be fitted based on the observed sequences.

While the flexibility of emission probabilities in HMMs could, in principle, reduce the number of required states for certain syllable sequences, this is not always the case. For example, in a song consisting of a single sequence, *ABC*, such as in the zebra finch song, the HMM would still require five states (the start and end states, along with three states emitting syllables *A*, *B*, and *C*, respectively) to avoid overgeneralization. Thus, the number of states in the HMM remains the same as in the POMM. However, the HMM includes more free parameters because the emission probabilities must also be fitted. As a result, for birdsong, HMMs offer no clear advantage over POMMs and may be unnecessarily more complex due to the increased number of parameters required.

Another drawback of HMMs is that their states are less interpretable in terms of neural activity in the HVC. During singing, distinct sets of neurons are responsible for producing each syllable (Fee et al., 2004), and the same set of neurons does not probabilistically drive the production of multiple syllables. This makes the emission probabilities in HMMs difficult to interpret within the context of song production.

There are many models designed to efficiently capture statistical regularities in sequences. Sparse Markov models (Jääskinen et al., 2014) and variable length Markov chains (VLMCs)

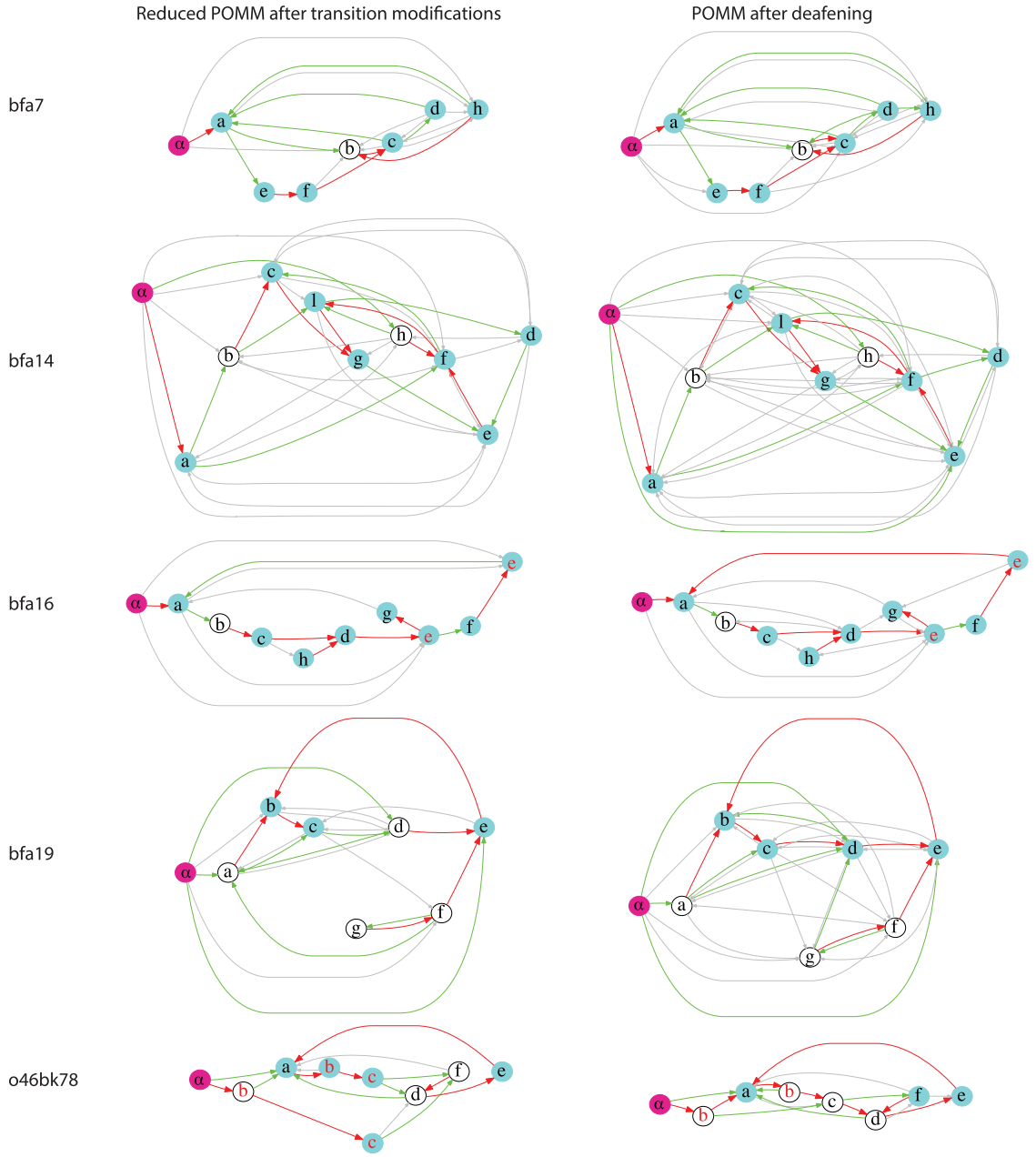

Figure 12. Comparisons of the reduced POMMs and the POMMs after deafening. The results for bfa7, bfa14, bfa16, bfa19, and o46bk78 are shown. The POMMs after deafening are the same as those presented in Figures 4 and 5, but the states have been laid out differently to facilitate easy comparison with the reduced POMMs.

Table 1. Comparison of the inferred POMMs with the higher-order Markov models, which are equivalent to the initial POMMs before state merging and deletion, for all six birds under normal conditions

Bird	Inferred POMM	Higher-order Markov model	Order
bfa7	17	42	5
bfa14	13	35	3
bfa16	13	17	2
bfa19	12	16	4
o46bk78	17	40	6
o10bk90	10	17	3

The number of states and the order of the higher-order Markov models are listed.

Table 2. Same comparison as in Table 1 but for the condition after deafening

Bird	Inferred POMM	Higher-order Markov model	Order
bfa7	9	9	1
bfa14	11	11	1
bfa16	11	21	2
bfa19	9	9	1
o46bk78	9	17	2
o10bk90	9	22	2

(Bühlmann and Wyner, 1999) focus on reducing higher-order Markov models into simpler forms that still retain the critical context dependencies within sequences. In this regard, both models share similarities with POMMs. However, unlike sparse Markov models and VLMCs, POMMs provide a biologically interpretable framework for birdsong. This interpretability is a key strength of POMMs in the study of birdsong syntax, as they allow for connections to biological structures such as

syllable-chains in the HVC. POMMs may not be well suited for other types of animal vocal sequences, such as mouse ultrasonic vocalizations (Wu et al., 2024), where such biological interpretations may not apply.

Under normal conditions, the songs of Bengalese finches in our study exhibit context-dependent syllable transitions. The POMMs for these songs show varying levels of state multiplicity, as shown in Figures 4 and 5, highlighting significant individual differences. Investigating the origins of these differences could be insightful, particularly in understanding the role of learning in shaping the formation of context dependencies in the songs.

Deafening significantly reduces state multiplicity, which can be explained by a model where the randomness of syllable transitions, driven by intrinsic connections between syllable chains, is mitigated by auditory feedback (Figs. 9, 11, 12). In this auditory-tuning model, auditory feedback modulates transitions between syllable chains, particularly those following context-providing syllables. This feedback refines transition probabilities and enhances context dependencies. This mechanism aligns with observations that disruptions to auditory feedback in Bengalese finch songs increase randomness in syllable sequences and introduce novel transitions (Sakata and Brainard, 2006). Consistent with the auditory-tuning model, adjusting the POMMs under normal conditions to reflect post-deafening syllable transition probabilities significantly simplifies the POMMs, making them closely resemble those observed after deafening (Figs. 11, 12).

The reafferent model is an alternative in which state multiplicity arises entirely from auditory feedback (Fig. 9*b*). However, based on syllable durations and the time scale required for auditory feedback signals to be detectable in the HVC (Sakata and Brainard, 2006), we have argued that the reafferent model is not feasible. Further experiments employing a variety of techniques could help clarify whether the auditory feedback time scale is indeed too short to account for the time scales needed to explain context-dependent syllable transitions in Bengalese finch songs. Additionally, neural signals other than auditory feedback may also contribute to context dependencies.

Experiments imaging calcium dynamics in the HVC of singing canaries have shown that different sets of HVC neurons can be active for the same syllable, depending on the context-dependent syllable transitions (Cohen et al., 2020). This finding supports the neural interpretation of state multiplicity in POMMs. Applying similar techniques to the Bengalese finch should allow for a quantitative test of POMMs. For an individual bird, the POMM inferred from its songs should provide a lower limit on the number of distinct sets of HVC neurons driving each syllable. These predicted numbers can then be compared to experimental observations. Such an experiment would directly test the validity of POMMs for Bengalese finch songs.

Previous studies on the effects of deafening in Bengalese finches have highlighted the rapid loss of sequence stereotypy, emphasizing the necessity of online auditory feedback for maintaining stereotyped syllable sequences (Okanoya and Yamaguchi, 1997; Woolley and Rubel, 1997). Our findings are consistent with these observations, as we also observe increased randomness in syllable sequences following deafening. On average, across the birds studied, there is a marked increase in transition entropy at syllable transition branching points after deafening (Fig. 6*b*). This increase is mainly due to the previously dominant transitions at these branching points becoming more evenly distributed, resulting in branches with more similar transition probabilities. Notably, similar outcomes have been reported in studies involving perturbations of auditory feedback (Sakata and Brainard, 2006), cooling of the HVC (Zhang et al., 2017), and enhancing inhibition within the HVC (Isola et al., 2020) in Bengalese finches. Investigating the potential for a unified neural mechanism underlying these diverse manipulations is an intriguing direction for future research.

In conclusion, we have developed a method for inferring minimal POMMs from observed sequences. When applied to the syllable sequences of Bengalese finch songs, both before and after deafening, our results indicate that the auditory system plays a crucial role in establishing context dependencies in syllable transitions. This method has broad applicability and could be useful for analyzing behavioral sequences in other animals as well.

References

Bühlmann P, Wyner AJ (1999) Variable length Markov chains. Ann Stat 27: 480–513.

Chang W, Jin DZ (2009) Spike propagation in driven chain networks with dominant global inhibition. Phys Rev E 79:051917.

Cohen Y, Shen J, Semu D, Leman DP, Liberti WA, Perkins LN, Liberti DC, Kotton DN, Gardner TJ (2020) Hidden neural states underlie canary song syntax. Nature 582:539–544.

Dobson CW, Lemon RE (1979) Markov sequences in songs of American thrushes. Behaviour 68:86–105.

Egger R, Tupikov Y, Elmaleh M, Katlowitz KA, Benezra SE, Picardo MA, Moll F, Kornfeld J, Jin DZ, Long MA (2020) Local axonal conduction shapes the spatiotemporal properties of neural sequences. Cell 183:537–548.

Ellson J, Gansner E, Koutsofios L, North SC, Woodhull G (2001) Graphviz—open source graph drawing tools. In: *Graph Drawing. GD 2001. Lecture Notes in Computer Science* (Mutzel P, Jünger M, Leipert S, eds) vol 2265. Berlin, Heidelberg: Springer.

Fee MS, Kozhevnikov AA, Hahnloser RH (2004) Neural mechanisms of vocal sequence generation in the songbird. Ann N Y Acad Sci 1016:153–170.

Gibbs AL, Su FE (2002) On choosing and bounding probability metrics. Int Stat Rev 70:419–435.

OpenAI (2023) Gpt-4 technical report.

Hahnloser RH, Kozhevnikov AA, Fee MS (2002) An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature 419:65.

Hanuschkin A, Diesmann M, Morrison A (2011) A reafferent and feed-forward model of song syntax generation in the bengalese finch. J Comput Neurosci 31:509–532.

Isola GR, Vochin A, Sakata JT (2020) Manipulations of inhibition in cortical circuitry differentially affect spectral and temporal features of bengalese finch song. J Neurophysiol 123:815–830.

Jääskinen V, Xiong J, Corander J, Koski T (2014) Sparse Markov chains for sequence data. Scand J Stat 41:639–655.

Jin DZ (2009) Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. Phys Rev E 80:051902.

Jin DZ (2013) The neural basis of birdsong syntax. In: *Progress in cognitive science: from cellular mechanisms to computational theories*, Beijing, China: Peking University Press.

Jin DZ, Kozhevnikov AA (2011) A compact statistical model of the song syntax in bengalese finch. PLoS Comput Biol 7:e1001108.

Jin DZ, Ramazanoğlu FM, Seung HS (2007) Intrinsic bursting enhances the robustness of a neural network model of sequence generation by avian brain area HVC. J Comput Neurosci 23:283–299.

Katahira K, Suzuki K, Okanoya K, Okada M (2011) Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. PLoS One 6:e24516.

Long MA, Jin DZ, Fee MS (2010) Support for a synaptic chain model of neuronal sequence generation. Nature 468:394.

Lynch GF, Okubo TS, Hanuschkin A, Hahnloser RH, Fee MS (2016) Rhythmic continuous-time coding in the songbird analog of vocal motor cortex. Neuron 90:877–892.

Markowitz JE, Ivie E, Kligler L, Gardner TJ (2013) Long-range order in canary song. PLoS Comput Biol 9:e1003052.

McLean LC, Roach SP (2021) Markov dependencies in the song syntax of Hermit Thrush (Catharus guttatus). J Ornithol 162:469–476.

Miller A, Jin DZ (2013) Potentiation decay of synapses and length distributions of synfire chains self-organized in recurrent neural networks. Phys Rev E 88:062716.

Moll FW, Kranz D, Asensio AC, Elmaleh M, Ackert-Smith LA, Long MA (2023) Thalamus drives vocal onsets in the zebra finch courtship song. Nature 616:132–136.

Okanoya K (2004) The bengalese finch: a window on the behavioral neurobiology of birdsong syntax. Ann N Y Acad Sci 1016:724–735.

Okanoya K, Yamaguchi A (1997) Adult bengalese finches (Lonchura striata var. domestica) require real-time auditory feedback to produce normal song syntax. J Neurobiol 33:343–356.

Picardo MA, Merel J, Katlowitz KA, Vallentin D, Okobi DE, Benezra SE, Clary RC, Pnevmatikakis EA, Paninski L, Long MA (2016) Population-level representation of a temporal sequence underlying song production in the zebra finch. Neuron 90:866–876.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77:257–286.

Sakata JT, Brainard MS (2006) Real-time contributions of auditory feedback to avian vocal motor control. J Neurosci 26:9619–9628.

Sakata JT, Brainard MS (2008) Online contributions of auditory feedback to neural activity in avian song control circuitry. J Neurosci 28:11378–11390.

Schmidt MF (2003) Pattern of interhemispheric synchronization in HVC during singing correlates with key transitions in the song pattern. J Neurophysiol 90:3931–3949.

Tupikov Y, Jin DZ (2021) Addition of new neurons and the emergence of a local neural circuit for precise timing. PLoS Comput Biol 17: e1008824.

Virtanen P, et al. (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. Nat Methods 17:261–272.

Wittenbach JD, Bouchard KE, Brainard MS, Jin DZ (2015) An adapting auditory-motor feedback loop can contribute to generating vocal repetition. PLoS Comput Biol 11:e1004471.

Woolley SM, Rubel EW (1997) Bengalese finches Lonchura Striata domestica depend upon auditory feedback for the maintenance of adult song. J Neurosci 17:6380–6390.

Woolley SM, Rubel EW (2002) Vocal memory and learning in adult bengalese finches with regenerated hair cells. J Neurosci 22:7774–7787.

Wu Y, Jarvis ED, Sarkar A (2024) Bayesian semiparametric Markov renewal mixed models for vocalization syntax. Biostatistics 25:648–665.

Zhang YS, Wittenbach JD, Jin DZ, Kozhevnikov AA (2017) Temperature manipulation in songbird brain implicates the premotor nucleus HVC in birdsong syntax. J Neurosci 37:2600–2611.

Zucchini W, MacDonald IL (2009) *Hidden Markov models for time series: an introduction using R*. New York: Chapman and Hall/CRC.