Generative 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 and Center for Neural Engineering, The Pennsylvania State University, University Park, PA, USA
2 Symbiosis School for Liberal Arts, Symbiosis International (Deemed University), Pune, Maharashtra, India
3 Scientific Data Division and Biological Systems & Engineering Division, Lawerence Berkeley National Laboratory
4 Helen Wills Neuroscience Institute & Redwood Center for Theoretical Neuroscience, UC Berkeley

† These authors contributed equally to the project
∗ Corresponding author: Dezhe Z. Jin, dzj2@psu.edu

Abbreviated title: **Generative models for Bengalese finch songs**

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements: 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

Abstract

Generative models are widely used in modeling sequences from language to birdsong. Here we show that a statistical test designed to guard against overgeneralization of a model in generating sequences can be used to infer minimal models for the variable syllable sequences of Bengalese finch songs. Specifically, the generative model we consider is the partially observable Markov model (POMM). A POMM consists of states and probabilistic transitions between them. Each state is associated with a syllable, and one syllable can be associated with multiple states. This multiplicity of association from syllable to states distinguishes a POMM from a simple Markov model, in which one syllable is associated with one state. The multiplicity indicates that syllable transitions are context-dependent. The statistical test is used to infer a POMM with minimal number of states from a finite number of observed sequences. We apply the method to infer POMMs for songs of six adult male Bengalese finches before and shortly after deafening, and show that auditory feedback plays an important role in creating context-dependent syllable transitions in Bengalese finch songs.

Author Summary

Large language models, such as the one underlying ChatGPT, are probabilistic generative models that can capture context-dependent sequencing of words. We show that a probabilistic model inspired by how the brain encodes song production in songbirds can accurately capture context-dependent syllable transitions in the songs of Bengalese finches. The models reveal that deafening reduces the context dependencies, suggesting that auditory feedback plays a significant role in creating such dependencies in Bengalese finch songs. Our approach can be applied to other behavioral sequences and help to understand the neural mechanisms underlying statistical regularities in the sequences.

Introduction

Large language models, such as ChatGPT, are probabilistic models that can capture the regularities in human language (OpenAI, 2023). The neural mechanisms that underlie these regularities are not well understood. Research on birdsong can shed light on this issue since birdsong is similar to human speech in that it consists of sequences of stereotypical syllables and is a learned behavior (Doupe and Kuhl, 1999). The syllable sequences of many songbird species are variable, and follow probabilistic rules (or syntax) that can be described with state transition models (Okanoya, 2004; Jin and Kozhevnikov, 2011; Jin, 2013; Markowitz et al., 2013). For Bengalese finch songs, it has been shown that the syllable sequences are well described by partially observable Markov models (POMMs) (Jin and Kozhevnikov, 2011). A POMM consists of states, the transition probabilities between the states, and the syllables associated with the states. Each state is associated with one syllable. The state transitions are Markovian: the transition probabilities between the states are fixed. Thus the paths of state transitions leading to a state do not affect the subsequent transitions from the state. A path of state transitions generates a syllable sequence by keeping track of associated syllables. Although a state is associated with one syllable, the converse is not necessarily true. In a POMM, one syllable can be associated with multiple states. This multiplicity of syllable to state association enables a POMM to describe context dependencies in syllable transitions: the transition probabilities between syllables depend on the preceding syllable sequences (Jin and Kozhevnikov, 2011). The Markov model is a special case of POMM, in which there is one-to-one correspondence between the states and the syllables. Markov models are not capable of describing context dependencies in syllable transitions.

POMM is motivated by the idea that birdsong is driven by synaptic chains in the premotor nucleus HVC (proper name) of the song system (Hahnloser et al., 2002; Fee et al., 2004; Jun and Jin, 2007; Jin et al., 2007; Jin, 2009; Long et al., 2010; Wittenbach et al., 2015; Lynch et al., 2016; Picardo et al., 2016; Jin, 2013; Zhang et al., 2017; Egger et al., 2020; Tupikov and Jin, 2021). Specifically, the HVC neurons that project to the downstream motor areas form

feedforward synaptic chain networks within HVC. Bursts of spikes propagate along a chain, with each projection neuron bursting once during the propagation, driving the production of one syllable through the projections to the downstream motor areas (Fee et al., 2004; Jin, 2009). The activation of one such "syllable-chain" can be identified as the neural correlate of one state in a POMM (Jin, 2009; Jin and Kozhevnikov, 2011; Wittenbach et al., 2015). Within this paradigm, inferring POMMs from observed syllable sequences can shed light on the neural dynamics in HVC that underlie the production of variable syllable sequences.

Auditory feedback has been shown to affect Bengalese finch song syntax (Okanoya and Yamaguchi, 1997; Woolley and Rubel, 1997; Woolley and Rubel, 2002; Sakata and Brainard, 2008; Wittenbach et al., 2015). A few days after deafening, the syllable sequences become more random (Okanoya and Yamaguchi, 1997; Woolley and Rubel, 1997), and the number of repetitions of long repeating syllables becomes smaller (Wittenbach et al., 2015). Altered auditory feedback to intact singing birds delivered at branching points of syllable transitions can change the transition probabilities (Sakata and Brainard, 2006; Sakata and Brainard, 2008). These observations demonstrate that auditory feedback could play an important role in creating context dependencies in syllable transitions in Bengalese finch song.

In this paper, we analyze context dependencies in the songs of six Bengalese finches before and shortly after deafening. We first devise a novel method for inferring a POMM from a set of observed syllable sequences. The method uses the concept of *sequence completeness*, which is the sum of the probabilities that the POMM generates each of the unique sequences in the observed set. Sequence completeness is further augmented with the differences of the probabilities of the unique sequences computed with the observed set and computed with the model, leading to the augmented sequence completeness, P_β. The method is designed to find the smallest number of states for each syllable that makes the POMM statistically compatible with the observed sequences. This is done by testing the hypothesis that the observed sequences are generated by the POMM, using the distribution of P_β for sets of POMM-generated sequences. Compared to the previous heuristic method (Jin and Kozhevnikov, 2011), our new method is much simpler

and more principled.

Using this method, we show that auditory feedback contributes significantly to the context dependencies in Bengalese finch songs. Specifically, we infer minimal POMMs for syllable sequences of 6 Bengalese finches in the normal condition and shortly after deafening. In the normal condition, the POMMs of all birds require multiple states for some syllables. Shortly after deafening, the POMMs either degenerate completely to Markov models (3 out of 6 birds) or have the state multiplicity drastically reduced (3 out of 6 birds). Hence the loss of hearing weakens context dependencies in syllable transitions.

Results

In this paper, we analyze the dataset collected for a previous study (Wittenbach et al., 2015), which showed that syllable repeats in Bengalese finch songs, especially for those syllable types with a variable number of repeats, are best modeled as the re-activation of syllable-chains with auditory feedback, with the feedback strength reduced after each repetition (Wittenbach et al., 2015). In this work we focus on the non-repeat versions of the sequences, in which only the first syllable of any repetition is retained. For example, if the syllable sequence is $ABBBC$, the non-repeat version is ABC. In the rest of the paper, syllable sequences refer to the non-repeat versions.

Each syllable sequence is typically led by a variable number of introductory notes. These introductory notes are excluded in the analysis. All sequences have definite starts and ends. Thus the POMMs have two special states. One is the Start-state, from which all state transitions begin, and the other is the End-state, at which all state transitions terminate. The POMMs are visualized with directed graphs (Fig. 1a). Following the convention introduced previously (Jin and Kozhevnikov, 2011), we denote the Start-state as a pink node marked with the symbol α. All other states are represented as nodes marked with associated syllables. Syllables with multiple states are marked with red fonts. The color of a state is cyan if it can transition to the end state, and is white otherwise. To reduce clutter in the graph, the End-state is not shown (we

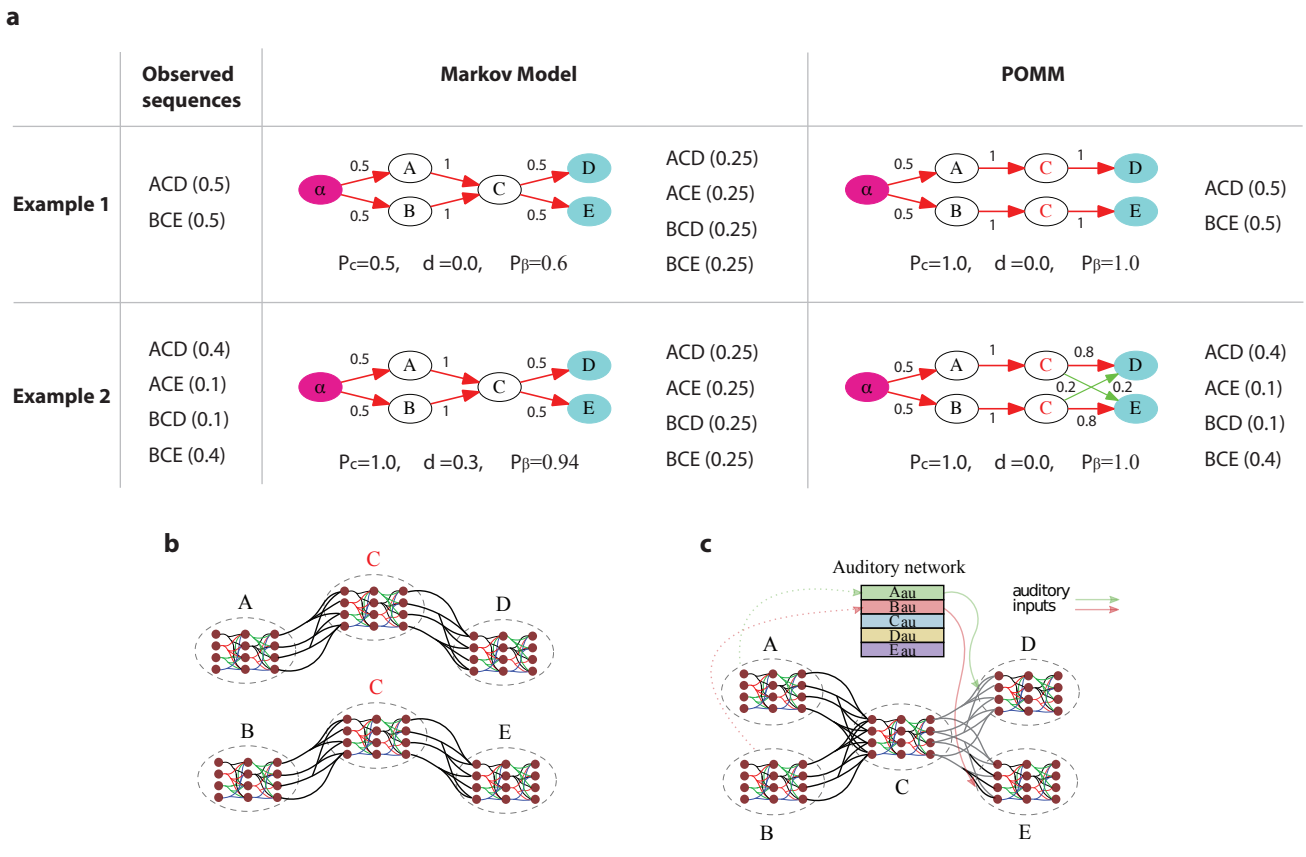


Figure 1: **Concepts illustrated with examples**. **a**. In Example 1, there are two unique sequences with equal probabilities (shown in parentheses) in the observed set. The Markov model overgeneralizes, creating two unobserved sequences (the sequence completeness is $P_c = 0.5$). The POMM with two states for syllable C does not overgeneralize ($P_c = 1$). In Example 2, there are four unique sequences in the observed set. Two of them are more frequent than the others. The Markov model fails to capture the differences in the probabilities of the unique sequences (the total variation distance is $d = 0.3$), although $P_c = 1$. The POMM with the two states for syllable C fully captures the probabilities of the unique sequences ($d = 0$ and the augmented sequence completeness is $P_\beta = 1$). **b**. The intrinsic mechanism for the multiple states for syllable C. Separate syllable-chains encode the multiple states. **c**. The reafference mechanism for the multiple states for syllable C. There is one syllable-chain for C. The multiple states are due to the auditory feedback from preceding syllables differentially influencing syllable transitions from C.

use symbol ω to mark the end of sequences when necessary). State transitions are shown with arrows color coded for the 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

Context dependencies in syllable transitions can take two forms. In one form, a syllable transition can be allowed or prohibited depending on the context. A simple example is that the observed set contains two unique sequences: ACD and BCE, each with probability 0.5 (Fig. 1a, Example 1). The transition $C \rightarrow D$ only occurs if C is preceded by A; and the transition $C \rightarrow E$ only occurs if C is preceded by B. Consequently, sequences ACE and BCD are unobserved. We call this form the *type I context dependence*.

In the other form, context dependence manifests in the probabilities. A simple example modified from Example 1 is that the observed set contains sequences ACD, with probability 0.4; ACE, with probability 0.1; BCD, with probability 0.1; and BCE, with probability 0.4 (Fig. 1a, Example 2). The transitions $C \rightarrow D$ and $C \rightarrow E$ are observed regardless of the preceding syllable; however, the transition probabilities are different when A precedes C than when B precedes C. We call this form the *type II context dependence*.

With the two examples we show that sufficient state multiplicity is required for capturing context dependencies in syllable transitions. For Example 1, consider constructing the Markov model for the set of observed sequences, which only requires calculating the transition probabilities between the syllables. The graph of the Markov model is shown in Fig. 1a. The sequences can start with either A or B with equal probability, hence the Start-state transitions to the states associated with A or B (A-state or B-state) with probability 0.5. These two states transition to the C-state with probability 1. Since C can be followed by either D or E, the C-state transitions to either the D-state or the E-state with probability 0.5. Finally, the D-state and the E-state transition to the End-state with probability 1. From the Start-state to the End-state, there are

four possible state transition paths, generating four sequences ACD, ACE, BCD, and BCE, each with probability 0.25. However, ACE and BCD are not in the observed set. Thus the Markov model overgeneralizes.

To characterize the overgeneralization of a POMM, we introduce the concept of *sequence completeness* P_c, which is defined as the total probability of the POMM 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 computed with the model. For Example 1, we have $P_c = 0.5$. The amount of overgeneralization is $1 - P_c$, which is the total probability of all unique sequences that the model can generate but are not in the observed set. The Markov model clearly does not capture the type I context dependence in this example. A more complex model has two states for syllable C, and the A-state and the B-state transition separately to each of these states (Fig. 1a). This POMM generates two sequences ACD and BCE with probabilities 0.5 each, and $P_c = 1$ for the observed set. In other words, the model does not overgeneralize.

Because P_c is the total probability of all unique sequences in the set, it is insensitive to the probabilities of individual unique sequences. Consider the Markov model for Example 2, which is the same as for Example 1 (Fig. 1a). The Markov model generates all observed unique sequences, hence $P_c = 1$. Although the model does not generate unobserved sequences, it misses the type II context dependence in Example 2. For this we need to compare probabilities of the unique sequences between the model and the observation.

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

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

Here

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

is the observed probability of the i-th unique sequence, and is the ratio of the copy number N_i of this sequence in the observed set of N sequences; and $P_{i,m}$ is the normalized probability of the sequence computed with the model

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

The normalization is to ensure that

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

which is required since we are comparing $P_{i,m}$ to $P_{i,o}$, and $\sum_{i=1}^{M} P_{i,o} = 1$. For Example 2, the Markov model has $d = 0.3$. The model with two states for C, as shown in Fig. 1a, can perfectly capture this type II context dependence with $d = 0$.

The total variation distance may not reveal type I context dependence. For Example 1, the Markov model generates the two observed sequences ACD and BCE with probabilities 0.25. However, after normalization the probabilities are 0.5. Hence we have $d = 0$ for the Markov model.

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 given to the total variation distance, and is a number between 0 and 1. We call this combined measure the *augmented sequence completeness*. In this paper we set $\beta = 0.2$. We find that this choice gives a good balance in discovering both types of context dependencies in syllable transitions. A perfect model would have $P_\beta = 1$.

Neural correlates of state multiplicity

Within the framework of syllable-chains in HVC, it is natural to assume that the multiple states associated with one syllable correspond to multiple syllable-chains in HVC that drive the

production of the same syllable (Jin, 2009; Cohen et al., 2020). In Example 1 discussed above, the POMM that fits the observed sequences has two states for syllable C. With two syllable-chains for C (referred to as chain-C), the sequences ACD and BCE can be wired into two separate chains, as shown in Fig. 1b. This is the *intrinsic mechanism* for the state multiplicity in POMMs. This mechanism can account for the type II context dependence in Example 2 by introducing weaker connections from the end of the chain-C in ACD to the start of the chain-E, and from the end of the chain-C in BCE to the start of chain-D, since the transition probabilities depend on the connection strength (Jin, 2009).

An alternative mechanism uses auditory feedback. In this case there is one syllable-chain for C, which connects to chain-D and chain-E. However, the activations of chain-D and chain-E are determined by the reafferent auditory inputs (Sakata and Brainard, 2006; Sakata and Brainard, 2008; Hanuschkin et al., 2011; Wittenbach et al., 2015). The auditory feedback from syllable A is sent to chain-D; while the auditory feedback from syllable B is sent to chain-E (Fig. 1d). The auditory inputs can bias the transitions to chain-D or to chain-E from chain-C (Jin, 2009; Hanuschkin et al., 2011; Wittenbach et al., 2015). With strong enough auditory inputs, the probability of transition from C to D should approach 1 when C is preceded by A. When C is preceded by B, the transition probability to D should approach 1. This is the *reafference mechanism* for the state multiplicity.

These two mechanisms have different predictions for the effects of deafening. The intrinsic mechanism predicts that the state multiplicity remains unchanged after deafening. The reafference mechanism predicts that all state multiplicity disappears after deafening, and the song syntax will become Markovian. These predictions can be tested by inferring POMMs for the observed syllable sequences before and shortly after deafening.

Statistical test of POMM

To find the POMM that is compatible with the observed set of syllable sequences, we need to devise a way of statistically evaluating the fit of the POMM. This problem can be cast as a

hypothesis test, in which the null hypothesis is that the observed set is generated by the POMM. We can use P_β for this purpose. Ideally, P_β of the observed set computed with the POMM should be 1, which indicates that the POMM generates all of the unique sequences in the observed set, and importantly, does not generate unobserved sequences; moreover, the probabilities of the unique sequences agree with the observations. In practice, due to the finite number N of sequences observed, it is possible that the observed set does not contain all possible sequences that the bird is capable of producing. Therefore, $P_c < 1$ could be due to the smallness of N, and not because the model overgeneralizes. Additionally, the mismatches in the probabilities of the unique sequences could be due to the inaccurate measurements of the probabilities when N is finite.

To take into account the finite N effect, we generate random sets of N sequences from the POMM. For each generated set, we compute P_β with the POMM. The P_β distribution of the generated sets can be used to gauge the likelihood that the P_β of the observed set is drawn from the distribution. Specifically, we compute the probability p that the observed P_β is greater than the P_β of the generated sets. If $p < 0.05$, we conclude that the observed P_β is not likely drawn from the distribution, and the POMM is rejected as the model for the observed set. If $p \geq 0.05$, the POMM is not statistically rejected, and it is accepted as the model for the observed set. In this work, we build the P_β distribution by generating 10000 random sets of N sequences from the POMM.

We illustrate this process with an example. In Fig. 2a, we show the "ground truth model". It has 2 states for syllables A and C, and one state for each of syllables B, D, E. The model generates 7 unique sequences: A, probability 0.1; ACD, probability 0.36; ACE, probability 0.04; BCD, probability 0.05; BCE, probability 0.2; BAE, probability 0.125; and BA, probability 0.125. From the model, we generate three sets of "observed sequences" with $N = 10$, $N = 30$ and $N = 90$. Sequences generated from the ground truth model contain both type I and type II context-dependent syllable transitions.

We construct Markov models from the observed sets by computing the probabilities of start-

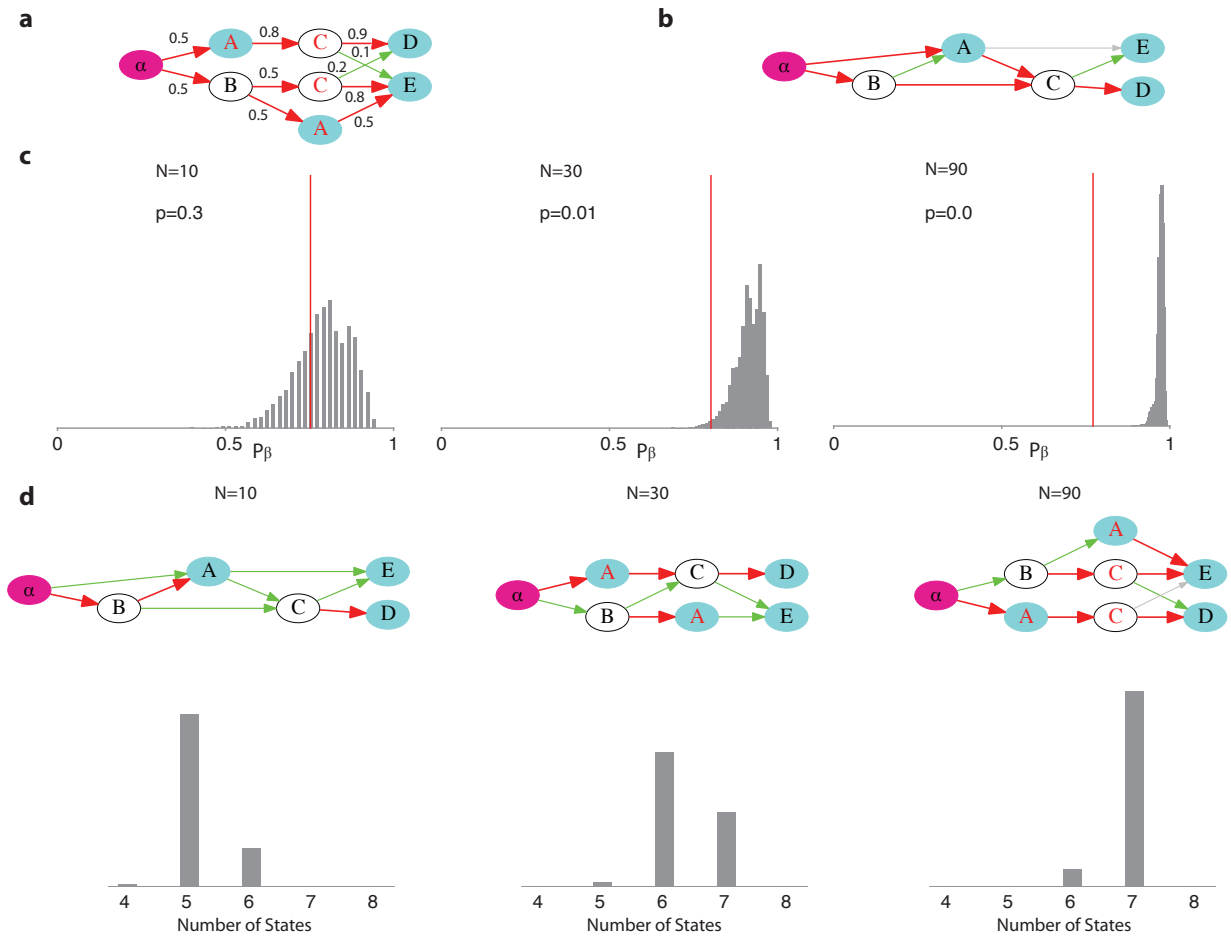


Figure 2: **An example for statistically testing and inferring POMMs from finite observed sequences**. **a**. The ground truth POMM from which the observed sequences are generated. The numbers near the arrows are the transition probabilities. **b**. The Markov model for a set of $N = 30$ observed sequences. **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.

ing or ending at each syllable, and the probabilities of transitioning from one syllable to another. The Markov model constructed from the $N = 30$ set is shown in Fig. 2b. We generate 10000 random sets of N sequences from the Markov models, and compute P_β of these generated sets. The distributions of P_β are shown in Fig. 2c. The distributions shift towards 1 as N increases.

This is because a large N ensures that a generated set contains most of the unique sequences that the model can generate, hence $P_c \to 1$ for the generated sets; moreover, the probabilities of the unique sequences computed with the generated sets agree better with those computed with the model. In contrast, the P_β of the observed sets computed with the Markov models, indicated with red lines in the figure, do not change much as N increases. This is because the observed sets are generated with the ground truth model, which is not a Markov model. Increasing N does not improve P_c for the observed sets. As N increases, the P_β for the observed sets is clearly below those of the generated sets, indicating that the Markov models are not statistically compatible with the observed sets.

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 $N = 90$. We run this process for 100 observed sets generated from the ground truth model for each N, and compute the p-value distributions. For $N = 10$, we find that $p = 0.3 \pm 0.3$ (mean $\pm$ standard deviation); for $N = 30$, $p = 0.01 \pm 0.03$; and for $N = 90$, $p = 0 \pm 0$. Therefore, for $N = 30$ and $N = 90$, the Markov model can be rejected based on the $p < 0.05$ criterion. For $N = 10$, however, the Markov model cannot be rejected, even though the ground truth model is non-Markovian.

If the ground truth model is Markovian, increasing N does not lead to rejection of the Markov model, as expected (supplementary Fig. S1).

Although we used the Markov model as an example, this process of statistical testing based on P_β can be used to test any POMM.

Inferring POMM from observed sequences

Given a set of observed syllable sequences, we infer a POMM that is statistically compatible with the set. We also require that the POMM is a minimal model, such that the number of states for each syllable is as small as possible, and the transitions between the states are sparse. This is achieved through a procedure that consists of initialization with higher-order Markov models, state reduction through state merging and deletion, and pruning of transitions between

the states. We illustrate this procedure through the example shown in Fig. 2a.

We first construct initial POMMs from higher order Markov models (Supplementary Fig. S2). To derive the m-th order Markov model, we flank each sequence in the observed set with the start symbol α and the end symbol ω, and collect unique subsequences of length m. We additionally collect unique subsequences of length up to m from the start symbol α. For example, for $m = 2$, the observed sequence $\alpha ACD\omega$ gives unique subsequences α, αA, AC, CD, and $D\omega$. The subsequence α is assigned to the Start-state. All unique subsequences ending with ω are assigned to the End-state. For the rest, each unique subsequence is assigned to a unique state, and the state is associated with the last syllable of the subsequence. With this assignment, each observed sequence gives a sequence of states. By counting the number of transitions, we compute the transition probabilities between the states. This procedure results in a POMM that is equivalent to the m-th order Markov model. We statistically test whether the POMM is accepted with the criterion $p \geq 0.05$. Starting from $m = 1$, which is the Markov model, we increase m until the corresponding POMM is accepted.

We next reduce the POMM by merging and deleting states associated with the same syllable. The state transition probabilities are re-computed using the counts of transitions between the states (Jin and Kozhevnikov, 2011). A merge is kept if the resulting POMM is accepted. After the merging stops, we further reduce the POMM with state deletion. If a syllable is associated with more than one state, we reduce the state number for the syllable by 1. We compute the transition probabilities between the states by maximizing the log-likelihood that the model generates the observed sequences using the Baum-Welch algorithm (Rabiner, 1989). To avoid local minima that the algorithm may encounter, we consider 100 runs of the algorithm with random seeds, and select the run with the largest likelihood. If the reduced POMM is accepted, the deletion continues. With these state reduction procedures, we arrive at a POMM with 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 cut every transition and recalculate the transition probabilities using the Baum-

Welch algorithm. If the log-likelihood of the observed sequences after the cut is larger than a 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 of the log-likelihood due to inaccuracies in computing the transition probabilities. The estimate is set to be the standard deviation of the log-likelihoods before the cuts in the 100 runs of the Baum-Welch algorithm with random seeds. If the POMM after the cuts does not pass the statistical test, the cuts are reversed.

We show the accuracy of the above procedure by inferring POMMs from 100 sets of N observed sequences generated from the ground truth model (Fig. 2a). The results for $N = 10, 30, 90$ are shown in Fig. 2d. We display typical POMMs inferred, and 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 mostly 5, and the Markov model is accepted. This is because for most sets of $N = 10$, the Markov model passes the statistical test. Some models have 4 states because syllables D or E may not appear in the observed sequences due to the small N. For $N = 30$, the total number of states ranges from 5 to 7. A typical POMM with 6 states is shown in the figure. For $N = 90$, the total number of states is mostly 7, and the inferred POMMs have the same structure as the ground truth model.

This example shows that our procedure tends to infer a simpler POMM than the ground truth model when N is small. When N is large, the procedure uncovers the ground truth model. Importantly, the procedure does not create more complex models than the ground truth model.

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

To see how auditory input affects the Bengalese finch song syntax, we analyze songs of six adult Bengalese finches before and two days after deafening. The dataset was used previously for analyzing syllable repeats (Wittenbach et al., 2015). Here we focus on the non-repeat versions of the syllable sequences.

We first test if Markov models are statistically compatible with the observed syllable se-

quences using the $p \geq 0.05$ criterion. The results are shown in Supplementary Figs. S3-S4. For three birds the syllable sequences are not compatible with the Markov models before and after deafening (o10bk90, normal $p = 0$, deafened $p = 0$; bfa16, normal $p = 0$, deafened $p = 0$; o46bk78, normal $p = 0$, deafened $p = 0$). Those of the other three birds are not compatible with the Markov models before deafening, but after deafening the Markov models are 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 reduces Bengalese finch song syntax from non-Markovian to Markovian for some birds but not for all.

Deafening also creates novel transitions between syllables, as well as novel starting and ending syllables. The transition probabilities of these novel transitions tend to be small (median, 0.04), but 22% have probabilities larger than 0.1 (18 transitions out of 81). The majority of these novel transitions appear in two birds (27 for bfa14; 21 for bfa19). A small number (8) of transitions also disappear after deafening (median transition probability in the normal condition, 0.02).

As observed in previous studies (Woolley and Rubel, 1997; Okanoya and Yamaguchi, 1997), deafening increases sequence variability. The variability of transitions from a given syllable i (or the start state) is quantified with the transition entropy as $S_i = -\sum_{j=1}^{M} p_{ij} \log_2 p_{ij}$, where M is the number of branches of the transitions, and p_{ij} is the probability of the j-th branch. If $M = 1$, the transition is stereotypical, and we have $S_i = 0$. For a given M, the entropy is maximum if the transition probabilities for all branches are equal. This maximum entropy increases with M. The median of transition entropies is significantly larger after deafening (0.95 ± 0.55) than before (0.35 ± 0.51; $p = 5 \times 10^{-6}$, Wilcoxon signed-rank one-sided test). The number of branches M is also significantly larger afer deafening (4 ± 1.5, median $\pm$ s.d.) than before (2 ± 0.90; $p = 9.8 \times 10^{-7}$, Wilcoxon signed-rank one-sided test).

We next construct POMMs from the observed syllable sequences before and after deafening. The inferred POMMs are shown in Figs. 3-4. In the normal condition, there are 44 syllables in the songs of the birds; among them, 26 require 1 state, 13 require 2 states, 2 require 3 states, and 3 require 4 states. So most syllables require 1 or 2 states. There are 76 states in the

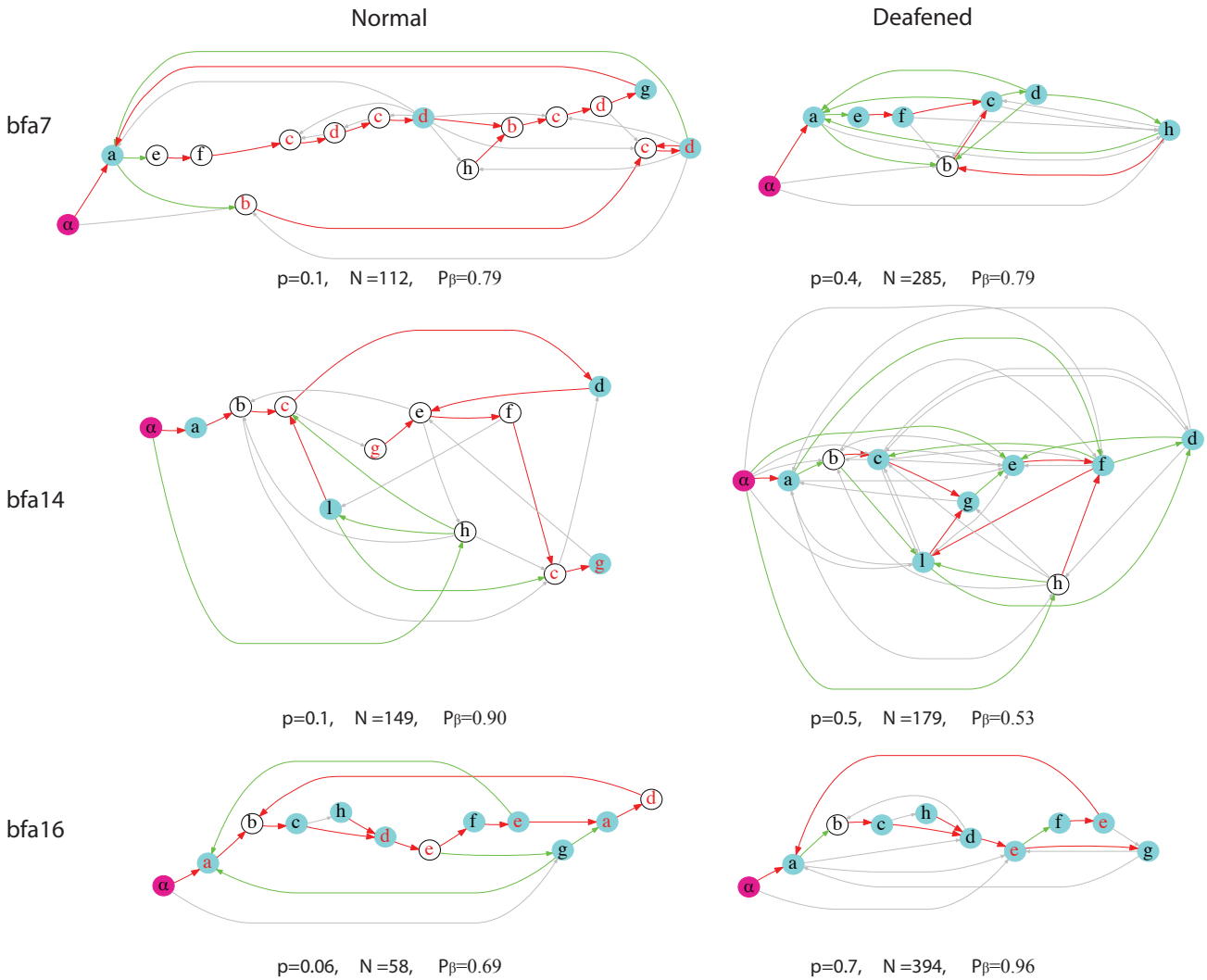


Figure 3: **POMMs before and after deafening for three birds**. The results for bfa7, bfa14, and bfa16 are shown. The p-values, the number of sequences in the observed sets, and P_β of the observed sets are shown.

POMMs. Counting only transition branches with probabilities greater than 0.01, the majority of states have up to 3 outgoing branches (there are 29, 31, 11 states with branch numbers 1, 2, 3, respectively).

After deafening, there are 43 syllables (syllable g for bfa7 drops out after deafening). Most

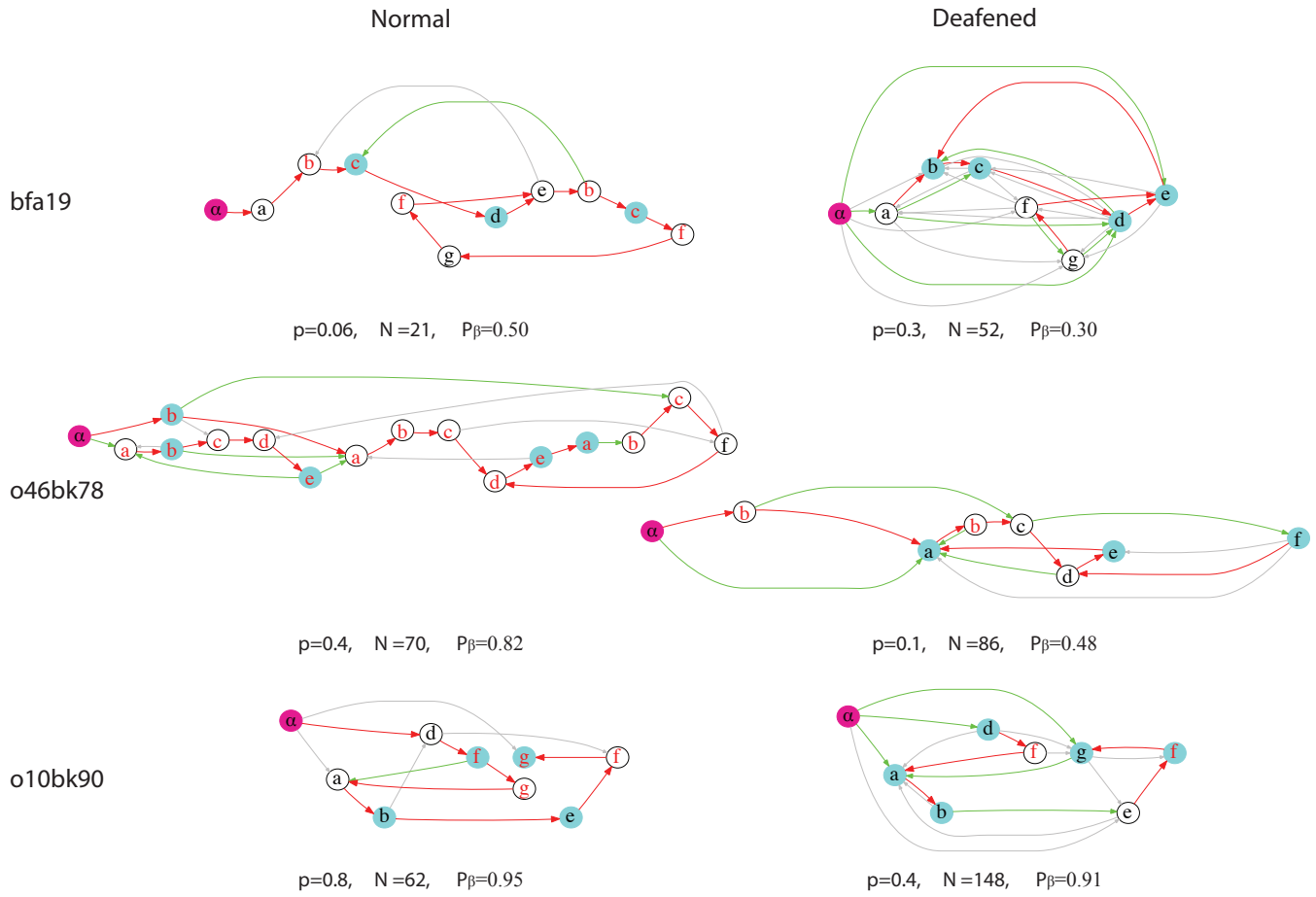


Figure 4: **POMMs before and after deafening for the other three birds**. The results for bfa19, ok46bk78, and o10bk90 are shown.

syllables (40) require only 1 state, and the remaining 3 require 2 states. There are 52 states in the POMMs. Counting only the transition branches with probabilities greater than 0.01, the states have up to 7 outgoing branches (there are 2, 19, 7, 13, 6, 3, 2 states with branch numbers 1 to 7, respectively).

Deafening significantly reduces the state multiplicity, as measured by the number of extra states (defined as the number of states for the syllables minus the number of the syllables) (Fig. 5a; the Wilcoxon signed-rank one-sided test, $p = 0.02$). The mean normalized transition

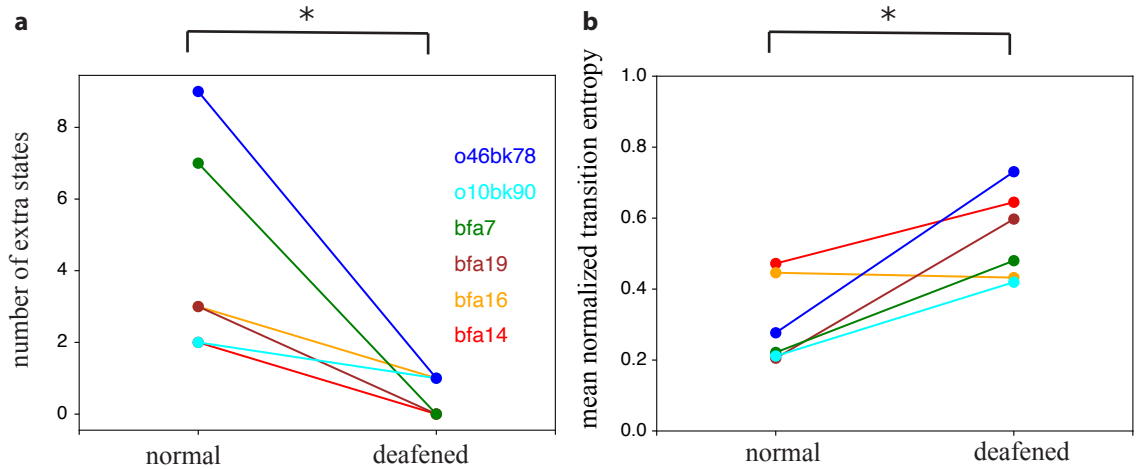


Figure 5: **Effects of deafening on the POMMs**. **a**. The numbers of extra states for the syllables are significantly reduced. **b**. The transitions from the states become more random after deafening.

entropy between the states (transition entropy divided by $\log_2 M$, where M is the number of transitions from the state) is significantly larger after deafening for all but one bird (Fig. 5b; Wilcoxon signed-rank one-sided test, $p = 0.03$, tested with all birds). Thus, deafening reduces the context dependencies in syllable sequences, as indicated by the reduction of the state multiplicity. Additionally, transitions between the states become more random.

The POMMs encode context dependencies in the syllable transitions. In the following, we show such dependencies for each bird in the normal condition and after deafening. We first show the major syllable transitions in the observed sets. We then point out how reducing the state multiplicity by merging states associated with the same syllable makes the POMM overgeneralize or significantly enhance the probabilities of some subsequences compared to the observed. This merging technique is inspired by the examples shown in Fig. 1. The state-merged POMM retains all state transition branches of the original POMM, but the transition probabilities are re-calculated with the Baum-Welch algorithm using the sequences in the observed sets. Since our goal is to show why multiple states are needed in the original POMMs, we select a few subsequences for evaluation and do not discuss all possible context dependencies in the syllable sequences.

The evaluation is done with the measure P_s, which is the fraction of sequences in a set that contain a selected subsequence. We first compute P_s for the set of N observed sequences. We then compute P_s for 10000 sets of N sequences generated from a state-merged POMM. This creates a distribution of P_s from the generated sets. We report the median P_s of this distribution to show how much the probability is enhanced compared to the P_s for the observed set. The significance of the enhancement is shown with the p-value, which is the probability p that P_s in the distribution is smaller than the P_s for the observed set. The process is analogous to the test of POMMs shown in Fig. 2.

bfa7

In the normal condition, syllable b has 2 states and syllables c and d have 4 states each (Fig. 3). The two states for b encode the following context dependence:

$$d \longrightarrow b_1 \longrightarrow c \longrightarrow d \longrightarrow g \ ,$$

$$a \longrightarrow b_2 \longrightarrow c \longrightarrow d \longrightarrow \omega \ .$$

The state multiplicity for c and d reflects the following context dependencies:

$$f \longrightarrow c_1 \longrightarrow d_1 \longrightarrow c_4 \longrightarrow d_4 \begin{array}{l} \nearrow b \\ \searrow \omega \end{array} \ ,$$

$$d \longrightarrow b \longrightarrow c_2 \longrightarrow d_2 \longrightarrow g \ ,$$

$$a \longrightarrow b \longrightarrow c_3 \longrightarrow d_3 \longrightarrow \omega \ .$$

Here ω denotes the end of the sequence, and the subscripts indicate different states for the same syllable. The consequences of merging states are summarized in Table 1. The mergers make the POMM overgeneralize, generating subsequences that are unobserved. These are examples of type I context dependencies.

Deafening leads to the appearance of $c \to a$ and $h \to a$ transitions, strengthening of $d \to a$ transition, and disappearance of $d \to c$ transition. Except for b, the sequence can now stop at all syllables. Interestingly, the $d \to g$ transition is lost and syllable g does not appear after

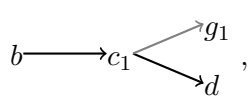
states merged	subsequence	P_s observed	median P_s after merging	p
b_1, b_2	$abcdg$	0	0.07	0.004
c_1, c_2	$fcdg$	0	0.1	0
c_1, c_3	$bcdcdb$	0	0.04	0.01
c_1, c_4	$fcdb$	0	0.08	0
c_2, c_3	$abcdg$	0	0.04	0.007
c_2, c_4	$bcdb$	0	0.04	0.003
c_3, c_4	$bcdb$	0	0.04	0.009
d_1, d_2	$fcdg$	0	0.1	0.0001
d_1, d_3	$bcdcdb$	0	0.04	0.01
d_1, d_4	$fcdb$	0	0.08	0.0001
d_2, d_3	$abcdg$	0	0.04	0.006
d_2, d_4	$bcdb$	0	0.04	0.0066
d_3, d_4	$bcdb$	0	0.04	0.009

Table 1: Consequences of pairwise merging of states with the same syllables in the POMM for bfa7 in the normal condition. Listed are the pair of states merged, subsequences examined, P_s of the subsequences in the observed set, median of the P_s distribution generated from the state-merged POMMs, and the p-value.

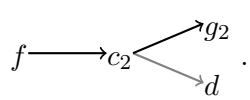
deafening. The syntax is Markovian, suggesting that there is no context dependence.

bfa14

In the normal condition, the POMM has two states for c and g (Fig. 3). The two state for c reflect the following type II context-dependent syllable transitions:



,



.

Transition from c_1 to d is strong (transition probability $P = 0.9$), but to g is weak ($P = 0.08$). On the contrary, transition from c_2 to d is weak ($P = 0.05$), but to g is strong ($P = 0.9$). Merging c_1 and c_2 significantly enhances the probability of the subsequence $b \to c \to g$ from $P_s = 0.07$ (observed) to $P_s = 0.2$ ($p = 0$); and $f \to c \to d$ from $P_s = 0.01$ (observed) to $P_s = 0.1$ ($p = 0$).

Transition from g_1 to e is stereotypical ($P = 1.0$). In contrast, transition from g_2 to e is weak ($P = 0.03$); instead, sequence is most likely stop at g_2 ($P = 0.97$ for $g_2 \to \omega$). Merging g_1 and g_2 enhances the probability of the subsequence $f \to c \to g \to e$ from $P_s = 0.01$ (observed) to $P_s = 0.05$, which is significant with $p = 0.03$.

Deafening creates numerous novel transitions with small probabilities (< 0.1) (Fig. 3). Novel transitions with large probability (≥ 0.1) also occur, which include transitions $a \to h$, $b \to l$, $h \to f$, and $l \to g$, as well as from α to syllables b, c, e, f, l. Some transitions are weakened, which include transitions $l \to c$ and $f \to c$. All context dependencies are lost, and the model becomes Markovian.

bfa16

In the normal condition, there are two states for syllables a, d, and e (Fig. 3). The two states for a encode the following context dependence:

$$\alpha \longrightarrow a_1 \longrightarrow b \ ,$$

$$e \longrightarrow a_2 \longrightarrow d \ .$$

Here α denotes the start of the sequence. Merging a_1 and a_2 creates an unobserved subsequence $\alpha \to a \to d$ with median $P_s = 0.2$ and $p = 0$.

The two states for d encodes the following context dependence:

$$c \longrightarrow d_1 \longrightarrow e \ ,$$

$$a \longrightarrow d_2 \longrightarrow b \ .$$

Merging d_1 and d_2 creates an unobserved subsequence $c \to d \to b$ with median $P_s = 0.2$ and $p = 0$.

The two states for e encodes the context dependence

$$d \longrightarrow e_1 \longrightarrow f \ ,$$

$$f \longrightarrow e_2 \longrightarrow a \ .$$

Merging e_1 and e_2 creates unobserved subsequence $d \to e \to a$ with median $P_s = 0.4$ and $p = 0$.

The major effects of deafening are the loss of the transition from e_2 to a_2; the strengthening

of the transition $e_2 \to a_1$; and the enhancement of stopping after g. The only state multiplicity left is for syllable e, which encodes the same context dependency as in the normal condition. Merging the two states for e again creates unobserved subsequence $d \to e \to a$ with median $P_s = 0.04$ and $p = 0$.

bfa19

In the normal condition, there are two states for syllables b, c, and f (Fig. 4). The state multiplicity for b and c encodes the context dependence

$$a \longrightarrow b_1 \longrightarrow c_1 \longrightarrow d \ ,$$

$$e \longrightarrow b_2 \longrightarrow c_2 \longrightarrow f \ .$$

Merging either b_1 and b_2 or c_1 and c_2 creates unobserved subsequence $a \to b \to c \to f$ with median $P_s = 0.4$ and $p = 0.0001$.

The state multiplicity for f encodes the context dependence

$$g \longrightarrow f_1 \longrightarrow e \ ,$$

$$c \longrightarrow f_2 \longrightarrow g \ .$$

Merging f_1 and f_2 creates unobserved subsequence $g \to f \to g$ with median $P_s = 0.3$ and $p = 0.0009$.

After deafening, many novel transitions appear, most notably $a \to d$. Sequences can also start with d or e with probabilities greater than 0.1. The model becomes Markovian, and all context dependencies disappear.

o46bk78

In the normal condition, the POMM has multiplicity for multiple syllables (Fig. 4). There are 4 states for b, 3 states for a and c, and 2 states for d and e, respectively. The structure of the POMM can be summarized as two motifs connected with a bridge. Motif 1 is the following cycle:

$$\begin{tikzcd} a_1 \arrow[r] & b_1 \arrow[r] & c_1 \arrow[r] & d_1 \arrow[r] & e_1 \arrow[llll, bend right] \end{tikzcd} \ .$$

states merged	subsequence	P_s observed	median P_s after merging	p
a_1, a_2	$\alpha abcdea\omega$	0	0.04	0.03
a_1, a_3	$\alpha a\omega$	0	0.06	0.01
a_2, a_3	$ba\omega$	0	0.3	0
b_1, b_2	$\alpha abcdea\omega$	0	0.04	0.03
b_1, b_3	$\alpha abcf$	0	0.07	0.008
b_1, b_4	αbcd	0.01	0.09	0.009
b_2, b_3	$babcf$	0.03	0.2	0
b_2, b_4	$babab$	0	0.1	0.0001
b_3, b_4	$eaba$	0.03	0.2	0.0004
c_1, c_2	$\alpha abcdea\omega$ or $\alpha ababcdeabcd$	0.01	0.09	0.008
c_1, c_3	$\alpha abcf$	0	0.06	0.02
c_2, c_3	$babcf$	0.03	0.3	0
d_1, d_2	$\alpha abcdea\omega$ or $\alpha ababcdeabcd$	0.01	0.09	0.02
e_1, e_2	$\alpha abcdea\omega$ or $\alpha ababcdeabcd$	0.01	0.09	0.02

Table 2: Consequences of pairwise merging of states with the same syllables for o46bk78 in the normal condition.

The start state transitions to a_1. The bridge is the subsequence

$$a_2 \longrightarrow b_2 \longrightarrow c_2 \ .$$

Motif 1 transitions to the bridge via the transitions $b_1 \to a_2$ and $e_1 \to a_2$. The bridge leads to motif 2, which is the following cycle:

$$\begin{tikzcd} d_2 \arrow[r] & e_2 \arrow[r] & a_3 \arrow[r] & b_3 \arrow[r] & c_3 \arrow[r] & f \arrow[lllll, bend right] \end{tikzcd} \ .$$

The bridge transitions to motif 2 via $c_2 \to d_2$. There is an additional state for b, which follows from the Start-state; this state transitions to the bridge via $b_4 \to a_2$, and to motif 2 via $b_2 \to c_3$. The consequences of pairwise merging of states are shown in Table 2.

After deafening, a novel transition $d \to a$ appears. Moreover, the probability of stopping after syllable a is strongly enhanced. State multiplicity disappears except for syllable b, which is still associated with two states, reflecting the context-dependent transitions

$$\begin{tikzcd} a \arrow[r] & b_1 \arrow[l, bend right] \arrow[r] & c \end{tikzcd} \ ,$$

$$\begin{tikzcd} \alpha \arrow[r] & b_2 \arrow[ur] \arrow[dr] & \\ & & c \\ & & a \end{tikzcd} .$$

This is a type II context dependence. In both cases, syllable b is followed by syllable a or c. However, from b_1 the transition to c is favored, with probability 0.9; in contrast, from b_2 the transition to a is favored with probability 0.8. Merging b_1 and b_2 enhances the subsequence $\alpha \to b \to c$ (observed $P_s = 0.2$) to median $P_s = 0.5$, which is significant ($p = 0$).

o10bk90

In the normal condition, syllables f and g are represented by two states each (Fig. 4), reflecting the following context dependencies in syllable transitions:

$$\begin{tikzcd} e \arrow[r] & f_1 \arrow[r] & g_1 \arrow[r] & \omega \end{tikzcd} ,$$

$$\begin{tikzcd} d \arrow[r] & f_2 \arrow[r] \arrow[rr, bend left] & g_2 \arrow[r] & a \end{tikzcd} .$$

Merging f_1 and f_2 creates an unobserved subsequence $e \to f \to a$ with median $P_s = 0.1$ and $p = 0.0002$. Merging g_1 and g_2 enhances the probability of the subsequence $d \to f \to g \to \omega$ from $P_s = 0.02$ to median $P_s = 0.3$, which is significant ($p = 0$).

After deafening, transition $\alpha \to d$ is weakened, and transitions $\alpha \to a$ and $\alpha \to g$ become stronger (Fig. 4). The state multiplicity for f persists, reflecting the context-dependent transitions

$$\begin{tikzcd} e \arrow[r] & f_1 \arrow[r] & g \end{tikzcd} ,$$

$$\begin{tikzcd} d \arrow[r] & f_2 \arrow[r] & a \end{tikzcd} ,$$

which is the same as in the normal condition. As in the normal condition, merging f_1 and f_2 creates the unobserved subsequence $e \to f \to a$ with median $P_s = 0.1$ and $p = 0$.

The subsequence $d \to f \to g$ becomes rare after deafening ($P_s = 0.5$, before deafening; $P_s = 0.007$, deafened), indicating that deafening makes the transition $f_2 \to g_2$ rare. Syllable g is now represented with one state only, because this does not make the subsequence $d \to f \to g \to \omega$ more frequent than observed, unlike in the normal condition.

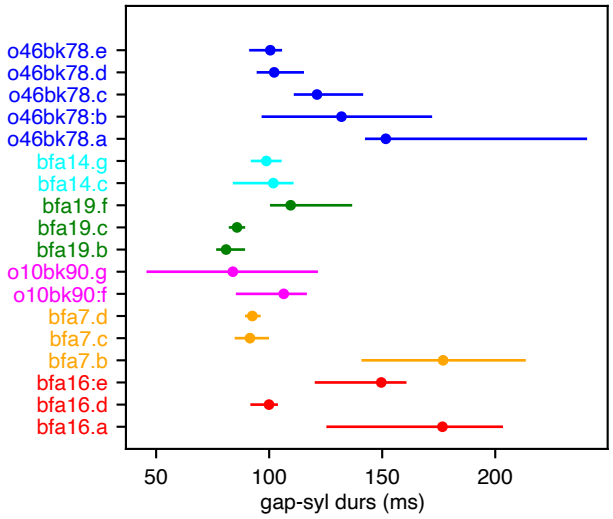


Figure 6: **Distributions of gap-syllable durations**. Durations of syllables plus the preceding gaps are shown for all syllables with multiple states in the POMMs for all birds. The dots indicate the median values and the bars indicate the 5% - 95% ranges of the distributions.

Neural mechanisms

Deafening greatly reduces the state multiplicity in the POMMs but does not completely abolish it. This suggests that the intrinsic mechanism (Fig. 1b) plays a role in creating context-dependent syllable transitions, but a major contribution comes from auditory feedback. To understand the neural mechanism of this contribution, we first examine the reafference mechanism shown in Fig. 1c, and then explore an alternative explanation.

The reafference mechanism requires that the delays of auditory feedback to HVC are long enough so that the auditory signals from the syllables that provide context are available to influence the syllable transitions after the syllable that requires multiple states. We illustrate this point with the example in Fig. 1c. For the activity in chain-A to bias the transition from chain-C towards chain-D, the feedback signal from syllable A must be available at the time of the transition. Thus the delay of the "round-trip" from the chain-A activity to the auditory feedback in HVC from syllable A must be longer than the duration of chain-C. The duration

of the activity propagation in chain-C should correspond to the duration of syllable C and the preceding silent gap. In Fig. 6, we plot the distributions of the gap-syllable durations for all syllables that require multiple states in the normal condition for the birds. The median values of the durations approximately range from 80 ms to 180 ms. Some of the syllables are repeated. In these cases the auditory feedback must be available after the syllable repetitions are finished. The distributions of the durations of gap-syllables including repetitions are shown in Fig. S5. The median values of the durations approximately range from 80 ms to 600 ms. These ranges provide a lower limit on delays of the auditory feedback for the reafference mechanism to work.

Experiments that perturbed auditory feedback during singing in Bengalese finches found that the perturbation altered HVC activity at around 44 ms after the onset of the perturbation, but no significant changes in HVC activity could be detected after 80 ms (Sakata and Brainard, 2008). Since the premotor delay from HVC to syllable production is around 50 ms (Schmidt, 2003), the round-trip delay of auditory feedback from a syllable should be limited to 130 ms. The reafference mechanism can thus be ruled out for at least 28% of syllables with state multiplicity (the median durations of gap-syllables are larger than 130 ms for these syllables (Fig. 6)). The fraction increases to 44% if syllable repetitions are included (Fig. S5).

An alternative mechanism does not require long delays of auditory feedback. Here the multiple states are encoded by separate chains as in the intrinsic mechanism (Fig. 1b), but the transitions between the chains are influenced by auditory feedback (Sakata and Brainard, 2006; Jin, 2009; Hanuschkin et al., 2011; Wittenbach et al., 2015). If the auditory feedback "tunes" the transition probabilities between the chains to create context dependencies in syllable transitions, deafening will "de-tune" them and reduce the context dependencies. In Fig. 7, we illustrate this auditory-tuning mechanism with the the simple example shown in Fig. 1. The auditory feedback from chain-A biases the transition to the chain-C that transitions to chain-D, while the auditory feedback from chain-B biases the transition to the chain-C that transitions to chain-E. Loss of auditory feedback reverts the network to rely on the intrinsic connections only, which allows transitions from chain-A and chain-B to both chain-Cs.

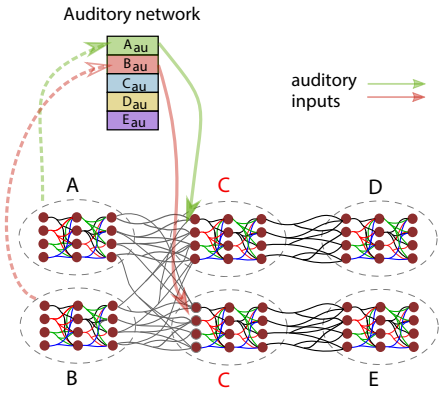


Figure 7: **The auditory-tuning mechanism**. As in Example 1 shown in Fig. 1, there are two unique sequences, ACD and BCE in the observed set. Two syllable-chains encode C. There are intrinsic connections from chain-A and chain-B to the two chain-Cs. Auditory feedback from chain-A biases the transition from chain-A to the upper chain-C, while auditory feedback from chain-B biases the transition from chain-B to the lower chain-C.

To investigate this possibility, we consider each state in the POMMs in the normal condition as syllable-chains, then modify the transition probabilities between them. The modifications are guided by the pairwise syllable transition probabilities after deafening. If a novel transition from syllable x to syllable y appears after deafening (probability must be ≥ 0.01), new transitions are created from all states associated with x to all states associated with y. If the transition from x to y is deleted after deafening, we delete the transitions from all states with x to all states with y. We further modify the transition probabilities between the states. From a state associated with x, if there are transitions to states associated with y, the probabilities of these transitions are scaled such that the total transition probability equals the transition probability from x to y after deafening. The probability of stopping at the state is adjusted so that the total probability of transitions from the state is normalized to 1. The modifications of the transition probabilities could weaken the context dependencies. This makes the POMMs reducible to simpler POMMs with less state multiplicity (referred to as reduced POMMs). Note that the modifications do not increase the state multiplicity because the number of syllable-chains are fixed and no new

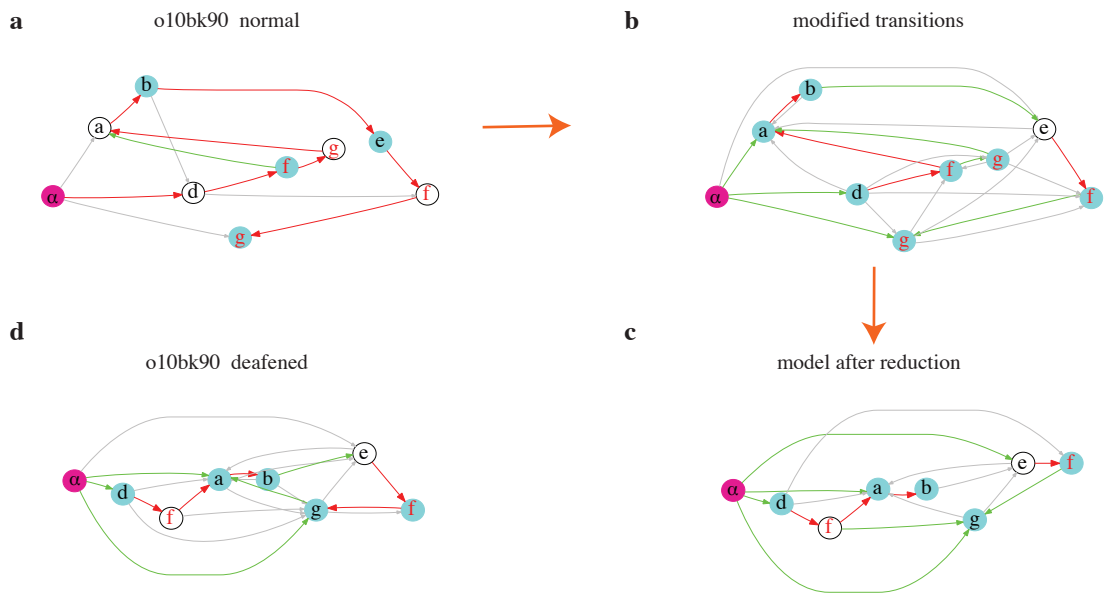


Figure 8: **Reduction of a POMM after modifications of the transition probabilities**. **a**. The POMM of o10bk90 in the normal condition. **b**. The POMM after modifications of the transitions probabilities using the syllable transition probabilities after deafening. **c**. Reduced POMM inferred from the $N = 62$ "observed sequences" generated from modified POMM. **d**. The POMM after deafening shown for comparison with the reduced POMM.

states are introduced.

The reduction process is as follows. With the modified POMMs, we generate 100 sets of N sequences, where N is the number of sequences observed for the deafened bird. Among them, we select the set with maximum probability as the "observed syllable sequences". With the generated "observed" sequences, we infer minimal POMMs, which are the reduced POMMs after the modifications of the transitions. This process is illustrated in Fig. 8a-c with o10bk90 as an example. The reduced POMM for this bird agrees with the POMM after deafening in terms of the number of states for each syllable. The transition probabilities do differ in details because the finite N introduces fluctuations. Similarly, the reduced POMMs for all other birds agree with the POMMs after deafening, except that for o46bk78 syllable c has an extra state

(Figs. S6-S7).

For the three birds whose POMMs become Markovian after deafening (bfa7, bfa19, bfa14), adding the novel transitions and deleting the dropped transitions alone, without modifying the transition probabilities between the states other than the necessary normalizations so that the total transition probability from a state is 1, already makes the reduced POMMs Markovian (Fig. S8). For o46bk78 the extra states are also greatly reduced. For the two other birds this does not change the number of extra states compared to the POMMs in the normal condition.

Discussion

Like human language, birdsong is characterized by context-dependent transitions in syllable sequences. POMMs are generative models suitable for birdsong syllable sequences (Jin and Kozhevnikov, 2011). POMMs capture context dependencies by associating multiple states to the same syllables. From the Start-state to the End-state, a state transition path generates a syllable sequence. A syllable with state multiplicity can transition differently from different states and thus be involved in context-dependent transitions to other syllables. POMMs with not enough state multiplicity may overgeneralize, creating sequences that are never observed (type I context dependency) or increase probabilities of certain sequences more than observed (type II context dependency). By statistically testing whether a POMM overgeneralizes compared to the observed set of syllable sequences, we can infer a minimal POMM that has the smallest number of states among POMMs that are compatible with the observed set. Using the minimal POMMs for Bengalese finch songs in the normal condition and shortly after deafening, we show that auditory feedback plays an important role in creating context-dependent syllable transitions in Bengalese finch songs.

To evaluate the fit of a POMM, we construct the distribution of the augmented sequence completeness P_β for the sequences sampled from the candidate POMM. This distribution is used to calculate the p-value of the P_β of the observed set computed with the POMM. We used the criterion $p < 0.05$ for rejecting the POMM. Lowering this cut-off value so that rejection is more

stringent should enable acceptance of POMMs with fewer number of extra states.

Our method is computationally intensive. The major cost comes from searching for the appropriate number of states for the syllables. It would be interesting to investigate better methods of estimating the state multiplicity. One possibility is to measure the predictive information in the syllable sequences and infer the number of parameters needed for encoding the sequence complexity (Bialek et al., 2001).

Our method is conservative. When the number N of observed sequences is small, the method tends to underestimate the true number of multiple states. This is because not all context dependencies are sufficiently represented in the observed sequences when N is small. One way to gauge whether N is large enough is to check if the sequence completeness P_c computed with the POMM is close to 1 for the observed sequences. The quantity $1 - P_c$ can be used as a rough estimate of the total probability of the missing unique sequences.

POMMs were inferred in a previous study by fitting some commonly used measures of sequences such as n-gram distributions, which are the probabilities of subsequences of length n (Jin and Kozhevnikov, 2011). This previous method involved multiple heuristic steps, and was not easy to implement. Additionally, the method required a large number of observed sequences because it relied on accurate measurements of the probabilities. In contrast, our method is principled, and can work with smaller number of observed sequences. This is because sequence completeness is a sum of probabilities and is thus not sensitive to the accuracy of probability measurements. Even though our method does not directly fit n-gram distributions, the statistics of 2- to 7-grams of the POMM generated sequences agree with those of the observed sequences for all birds (Fig. S9 and Fig. S10).

In the normal condition, the songs of Bengalese finches in our study all have context-dependent syllable transitions. The POMMs of the songs have varying degrees of multiplicity (Figs. 3-4), showing a large individual differences. It would be interesting to investigate the sources of such differences, especially whether learning plays a role in forming context dependencies in songs.

The state multiplicity is much reduced by deafening, but it persists for some syllables in some birds. Hence context-dependent syllable transitions can be produced without auditory inputs, suggesting that the intrinsic mechanism (Fig. 1b) plays at least a partial role in creating the context dependencies.

At first glance, the reduction of state multiplicity after deafening supports the reafference mechanism (Fig. 1c). However, this mechanism is limited by how long the auditory feedback signals from the context-providing syllables persist in HVC. Experiments that perturbed auditory feedback in singing Bengalese finches showed that the perturbations affected HVC activity after some delays but the effects were limited up to 80 ms after the onset of the perturbations (Sakata and Brainard, 2008). Allowing for a premotor delay from HVC activity to syllable production to be around 50 ms (Schmidt, 2003), the feedback signal from a context-providing syllable should be available up to 130 ms but not beyond. This strongly limits the "distance" of the context dependence, i.e. the number of syllable that can be in between the context-providing syllables and the context-dependent transitions. Analysis of the gap-syllable durations suggests that the distance can be at most 1, and for many cases the syllables at which the context-dependent transitions occur cannot fit into the delay limit because of the large gap-syllable durations (Figs. 6 and S5). On the other hand the distances of context dependencies can be quite long for some birds (for example o46bk78, Fig. 4 and Table 2). The upper limits of the delay of auditory feedback thus severely restrict the utility of the reafference mechanism.

The auditory-tuning mechanism combines elements of the intrinsic and the reafference mechanisms (Fig. 7). As in the intrinsic mechanism, a syllable with m states in a POMM is produced by at least m syllable-chains in HVC. This is the many-to-one mapping from syllable-chains to syllables (Jin, 2009; Jin and Kozhevnikov, 2011). As in the reafference mechanism, auditory feedback influences the transitions between the syllable-chains, but affects only the transitions immediately following the context-providing syllables. There are intrinsic connections between the syllable-chains, but they do not produce many context dependencies in the syllable sequences. The auditory feedback tunes the transition probabilities to create more context dependencies,

most likely by reducing the randomness of the transitions by the intrinsic connections. Indeed, perturbations of auditory feedback in singing Bengalese finches tend to make syllable sequences more random and even create novel transitions (Sakata and Brainard, 2006). In agreement with this mechanism, modifying the POMMs in the normal conditions with the syllable transition probabilities observed after deafening reduces the POMMs to be close to the POMMs after deafening (Figs. 8 and S6-S7).

Calcium imaging in HVC of singing birds can detect many-to-one mapping from syllable-chains to the syllables (Cohen et al., 2020). Applied to the Bengalese finch, it is possible to test the POMMs quantitatively given the auditory-tuning mechanism. For individual Bengalese finches, POMMs can be inferred from their songs. The number of states for each syllable in the POMMs should be the lower limit for the number of syllable-chains that can be found in HVC using calcium imaging for the individual birds.

Previous deafening studies in the Bengalese finch emphasized the loss of sequence stereotypy shortly after deafening and suggested that online auditory feedback is required for producing stereotyped syllable sequences (Woolley and Rubel, 1997; Okanoya and Yamaguchi, 1997). We confirm that deafening makes syllable sequences more random. On average across the birds, the transition entropy at the branching points of syllable transitions tends to increase after deafening (Fig. 5b). This increase is mostly due to the branching points that have dominant transitions becoming more "equalized", such that the branches have comparable transition probabilities. Similar effects were seen in perturbations of auditory feedback (Sakata and Brainard, 2006), cooling HVC (Zhang et al., 2017), and enhancing inhibition (Isola et al., 2020) in HVC of the Bengalese finch. It would be interesting to investigate whether there is a common neural mechanism across these manipulations.

In conclusion, we devised a method of inferring minimal POMMs from observed sequences. Application of the method to the syllable sequences of Bengalese finch songs before and after deafening suggests that the auditory system helps to create context dependencies in syllable transitions. Our method should be broadly applicable to other animal behavioral sequences.

Materials and Methods

Data set

The data set in this work was previously used for analyzing syllable repeats (Wittenbach et al., 2015) (available for download from http://personal.psu.edu/dzj2/SharedData/KrisBouchard/). Details of recording songs, annotating syllables, and deafening through bilateral cochlear removal, as well as the Ethics Statement can be found in the published paper (Wittenbach et al., 2015). We specifically used the data collected from six male adult Bengalese finches before and after deafening (labeled bfa14, bfa16, bfa19,bfa7, o10bk90, and o46bk78).

In the data set, syllables are labelled a through l, and x through z. Some ambiguous syllables are noted with symbols 0 and $-$, and they are skipped. Bengalese finch song bouts typically begin with short introductory notes. They are labeled as i, j and k. We define song sequences as segments of syllables that are bracketed by periods of introductory notes and the end of the recordings.

POMM

A POMM is specified by a state vector $V = [\alpha, \omega, s_3, s_4, \cdots, s_n]$, where $s_1 = \alpha$ and $s_2 = \omega$ are the start and the end states, n is the total number of states, and s_i for $i = 3, \cdots, n$ is the syllable symbol associated with the ith state. The same syllable symbol can appear multiple times in the state vector. Transitions between the states are described by a transition matrix T, whose element T_{ij} gives the probability of transition from state i to state j. There are no transitions to the start state, i.e. $T_{i1} = 0$; and there are no transitions from the end state, i.e. $T_{2j} = 0$.

Sequence generation from a POMM starts with the start state. At state i, the next state j is chosen with the probabilities T_{ij} among possible choices of state 2 to state n. Once chosen, the symbol s_j is added to the sequence. This process repeats until the end state is reached, at which point the sequence generation is complete.

A POMM is visualized with the software Graphviz (Ellson et al., 2001). To reduce clutter,

only transitions with probabilities larger than 0.01 are shown. Additionally, the End-state is not shown. Instead, the states that can transition to the end state are shown in cyan. The transition probability from one state to the end state is 1 minus the sum of the transition probabilities to other states. If a state does not transition to the end state with a probability larger than 0.01, the state is shown as white. The start state is shown in pink.

Markov model

A Markov model is a special case of POMM for which each syllable symbol appears only once in the state vector. The transition probabilities T can be computed as

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

where N_i is the total number of times s_i appears in the set Y of sequences, and N_{ij} is the total number of the times that the two-symbol subsequence $s_i s_j$ appears in Y. Note that

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

so we only need to compute N_{ij}.

Baum-Welch algorithm

Computing T for POMM with state multiplicity is more complicated than that for the Markov model, but the approach is similar. Starting from a set of random transition probabilities, the state transition sequences that correspond to the syllable sequences in Y are worked out. The transition probabilities are then updated according to

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

where N_i is number of times the state i appears in the state sequences, and N_{ij} is the number of times the subsequence of states ij appears. With the updated T, the process is repeated. The process stops when the changes in T is smaller than 10^{-6}. Because the result might be dependent on the initialization of T, the process is run for 100 times with different seeds for random number generator. The T that maximizes the probabilities of generating Y from the POMM is selected.

The computation is efficiently implemented with the Baum-Welch algorithm (Rabiner, 1989). Consider a sequence $y_1 y_2 \cdots y_t \cdots y_m$ in the set Y. Here t is the step in the sequence and m is the maximum length of the sequence. The algorithm consists of three parts. First, calculate the forward probability $\alpha_i(t)$, which is the probability of being at state i at step t given the proceeding sequence is $y_1 y_2 \cdots y_{t-1}$. This is computed iteratively with

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

Since all sequences start from the Start-state, the initial condition is $\alpha_1(0) = 1$ and $\alpha_j(0) = 0$ for all $j \neq 1$. Here $\delta_i(y_{t+1}) = 1$ if the symbol y_{t+1} at step $t+1$ is the same as the symbol s_i associated with state i; otherwise, $\delta_i(y_{t+1}) = 0$. Second, calculate the backward probability $\beta_i(t)$, which is the probability being at state i at step t and the follow-up sequence is $y_{t+1}, \cdots, y_m$. This is calculated iteratively with

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

Since all sequences end at the end state, the initial condition is $\beta_2(m+1) = 1$ and $\beta_j(m+1) = 0$ for all $j \neq 2$. Third, calculate N_i and N_{ij}. The forward and backward probabilities $\alpha_i(t)$ and $\beta_i(t)$ should be computed for each sequence in Y. The number of transition 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 it is given by

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

which is the forward probability of ending at the end state at step $m+1$.

The total probability of the set Y is given by

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

It is most convenient to use the log-likelihood, which is

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

Sequence completeness, total variation distance and augmented sequence completeness

For a set of sequences Y, the sequence completeness on a POMM is computed as

$$P_c = \sum_{y \in Y} P_y,$$

where y is a unique sequence in Y. The sum is over all the unique sequences in the set.

For a set of observed sequences Y_o, the total variation distance is defined as

$$d = \frac{1}{2} \sum_{y \in Y_o} |P_y - P_{y,m}|.$$

Here $P_{y,m}$ is the probability of the unique sequence y computed on the POMM and then *normalized* among the unique sequences such that

$$\sum_{y \in Y_o} P_{y,m} = 1.$$

This normalization is necessary because $P_{y,m}$ is compared to P_y, which is normalized:

$$\sum_{y \in Y_o} P_y = 1.$$

The total variation distance ranges from 0 to 1.

The augmented sequence completeness is defined as

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

Here β is a parameter that can be chosen in the range $(0, 1)$. The value of P_β ranges from 0 to 1. A perfect POMM for the observed set should yield $P_\beta = 1$ because $P_c = 1$ and $d = 0$. When N is small, the measurements of P_y are not accurate. For this case, the contribution from d should be reduced by taking a small value for β. In our work, we chose $\beta = 0.2$.

Statistical test

To test whether an observed set Y_o with N sequences could be generated from a POMM, we generate $M = 10000$ sets of N sequences, and compute the P_β of the generated sets, which gives a distribution of P_β. We also compute the augmented sequence completeness $P_{\beta,o}$ of the observed set. In the distribution, we count the number K of P_β that are smaller or equal to $P_{c,o}$. To avoid small fluctuations in P_β making K artificially small, we added 10^{-10} to $P_{\beta,o}$. The p-value is

$$p = \frac{K}{M}.$$

The POMM is rejected if $p < 0.05$, and accepted otherwise.

Inferring minimal POMM

For a given set Y of N syllable sequences, the minimal POMM is inferred through three steps: setting up the initial POMM through higher order Markov models; simplifications with state

merging followed by state deletion; and removal of transitions. Because of the inherent randomness in the inference process, the minimal POMMs from different runs may differ. We run the process 20 times and select the most common model as the final minimal POMM.

The initial POMM is constructed to be equivalent to a higher order Markov model for Y. We flank each sequence with the start symbol α and the end symbol ω, and find all subsequences of length m. Here m is the order of the Markov model. We also include all subsequences starting with the symbol α and the lengths from 1 to $m-1$. We identify all unique subsequences. Each unique subsequence is assigned a state, and the last symbol of the subsequence is associated with the state. We count the number of times N_i that the ith subsequence appears, as well as the number of times N_{ij} that the ith subsequence is followed by the jth subsequence. The transition probability from state i to state j is set to

$$T_{ij} = N_{ij}/N_i.$$

This completes the construction of the POMM that is equivalent to the mth order Markov model. Starting from $m = 1$, we increase m until the POMM passes the statistical test. This completes the initialization of the POMM.

To simplify the POMM, we first test merging states and then test deleting states. Merging states i and j with the same associated syllable eliminates state j. The transition probability from state i to state k ($k \neq i, j$) is recomputed as

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

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

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

The number of times state i is visited is updated as

$$N_i \to N_i + N_j.$$

If the POMM after the merger passes the statistical test, the merger is accepted. Otherwise the merger is reversed. To speed up this process, we first test merging all states for each syllable. If the merger is rejected, we test pairwise merging of the states for the syllable. After no further state mergers are possible, we test deleting states. For each syllable, we reduce the number of states by 1. We then compute the transition probability matrix T using the Baum-Welch algorithm to maximize the log-likelihood of Y. If the POMM passes the statistical test, the reduction continues. Otherwise the reduction is reversed. The process continues until no further state deletions are accepted.

The final step is minimization of the number of transitions in the POMM. We first remove all transitions with probability smaller than 0.001. We then remove the remaining transitions one by one, and re-compute the transition matrix T after each removal. To remove a transition from state i to state j, we set $T_{ij} = 0$ in the initial transition matrix for the Baum-Welch algorithm. The algorithm ensures that this transition element remains 0. If the log-likelihood remains within the threshold, the removal is accepted; otherwise the removal is reversed. The threshold is

$$L_\theta = L_{max} - \sigma_L,$$

where L_{max} is the log-likelihood of the original POMM before any removals, and σ_L is the standard deviation of the log-likelihood of the 100 runs of Baum-Welch algorithms with different random seeds. If after all of the accepted removals the POMM does not pass the statistical test, we revert to the POMM before any removals.

Probability of finding a subsequence

The probability P_s of finding a subsequence in a set Y is defined as

$$P_s = \frac{K}{N},$$

where N is the number of sequences in the set, and K is number of sequences that contains the subsequence.

State merging tests

To evaluate the context-dependent syllable transitions encoded by state multiplicity in a POMM, we perform pairwise state merging tests. The merged state retains all transitions to and from the two states. The transition probabilities of the state-merged POMM are recomputed using the Baum-Welch algorithm and the observed set Y_o. By examining the states transitioning into the two states, and the states that follow the two states, we find possible subsequences that can show overgeneralization after the state merger. We find a subsequence that either is unseen in the observed set ($P_{s,o} = 0$) or has small probability $P_{s,o}$. To see whether the subsequence is significantly more probable in the sequences generated from the state-merged POMM, we generate 10000 sets of N sequences from the POMM. Here N is the number of sequences in Y_o. For each generated set, we compute P_s. This creates a distribution. We count the number of P_s that is smaller than or equal to $P_{s,o} + 10^{-10}$. The p-value is the ratio of this number and 10000. We add a small number 10^{-10} to $P_{s,o}$. This is for avoiding artificially lowering p-value due to those P_s that are equal to $P_{s,o}$. For example, if the subsequence is unobserved ($P_{s,o} = 0$) and the state-merged POMM does not generate it either, we would have a situation that $P_s = 0$ for all of the sampled set. By adding the small number to $P_{s,o}$, we ensure that $p = 1$, as it should be. If $p < 0.05$, we conclude that the enhancement of P_s after state merger is significant.

Wilcoxon signed-rank test

For comparing distributions of paired data in Fig. 5, we use Wilcoxon signed-rank test using scipy.stats.wilcoxon, which is in the Python module scipy (Virtanen et al., 2020).

References

Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. *Neural computation* 13:2409–2463.

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.

Doupe AJ, Kuhl PK (1999) Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience* 22:567–631.

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 *International Symposium on Graph Drawing*, pp. 483–484. Springer.

Fee MS, Kozhevnikov AA, Hahnloser RH (2004) Neural mechanisms of vocal sequence generation in the songbird. *Annals of the New York Academy of Sciences* 1016:153–170.

Gibbs AL, Su FE (2002) On choosing and bounding probability metrics. *International statistical review* 70:419–435.

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. *Journal of computational neuroscience* 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. *Journal of Neurophysiology* 123:815–830.

Jin DZ (2009) Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus hvc. *Physical Review E* 80:051902.

Jin DZ (2013) The neural basis of birdsong syntax. *Progress in cognitive science: From cellular mechanisms to computational theories* .

Jin DZ, Kozhevnikov AA (2011) A compact statistical model of the song syntax in bengalese finch. *PLoS computational biology* 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. *Journal of computational neuroscience* 23:283–299.

Jun JK, Jin DZ (2007) Development of neural circuitry for precise temporal sequences through spontaneous activity, axon remodeling, and synaptic plasticity. *PLoS One* 2:e723.

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 computational biology* 9:e1003052.

Okanoya K (2004) The bengalese finch: a window on the behavioral neurobiology of birdsong syntax. *Annals of the New York Academy of Sciences* 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. *Journal of neurobiology* 33:343–356.

OpenAI (2023) Gpt-4 technical report.

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. *Proceedings of the IEEE* 77:257–286.

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

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

Schmidt MF (2003) Pattern of interhemispheric synchronization in hvc during singing correlates with key transitions in the song pattern. *Journal of neurophysiology* 90:3931–3949.

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

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature 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 computational biology* 11:e1004471.

Woolley SM, Rubel EW (1997) Bengalese finches lonchura striata domestica depend upon auditory feedback for the maintenance of adult song. *Journal of Neuroscience* 17:6380–6390.

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

Zhang YS, Wittenbach JD, Jin DZ, Kozhevnikov AA (2017) Temperature manipulation in songbird brain implicates the premotor nucleus hvc in birdsong syntax. *Journal of Neuroscience* 37:2600–2611.

Supplementary Material

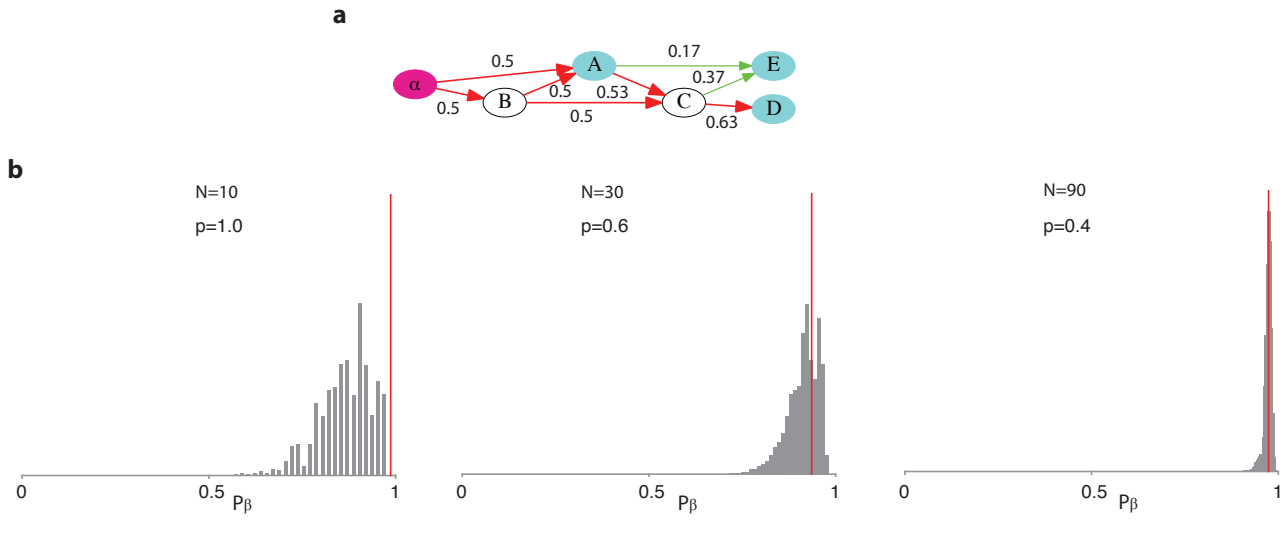


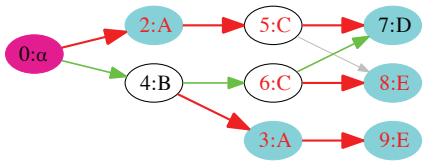
Fig. S1. **Statistical tests of the Markov models**. **a**. The ground-truth model is a Markov model. The transition probabilities are marked near the arrows. **b**. Statistical tests of the Markov models for $N = 10, 30, 90$. The gray bars are the distributions of P_β for the generated sets. The red lines are the P_β for the observed sets. The Markov model is accepted in all cases.

a

αBCEω, αBAω, αBAω, αACDω, αAω, αACDω, αBCEω, αAω, αBCEω, αACDω, ...

Sub Seq	State ID	Sym
α	0	α
Dω	1	ω
Eω	1	ω
Aω	1	ω
αA	2	A
BA	3	A
αB	4	B
AC	5	C
BC	6	C
CD	7	D
CE	8	E
AE	9	E

b



c

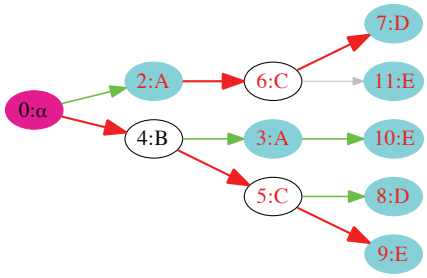


Fig. S2. **Constructing POMMs that are equivalent to higher order Markov models**. An example is shown for constructing POMMs from the second and third order Markov models. The observed sequences ($N = 30$) are generated from the ground truth POMM in Fig. 2a. **a**. The process of assigning states for the second order Markov model. Examples of sequences are shown. Each sequence is flanked by the start symbol α and the end symbol ω. In the table, the first column shows the unique subsequences. The second column shows the state IDs assigned to the subsequences. The third column shows the syllables (or symbols) associated with the states. **b**. The POMM corresponding to the second order Markov model. Both the state IDs and the associated syllables (or symbols) are displayed. **c**. The POMM corresponding to the third order Markov model. Both models have $p = 1$, hence they are compatible with the observed sequences. The POMMs have more extra states than the ground truth model.

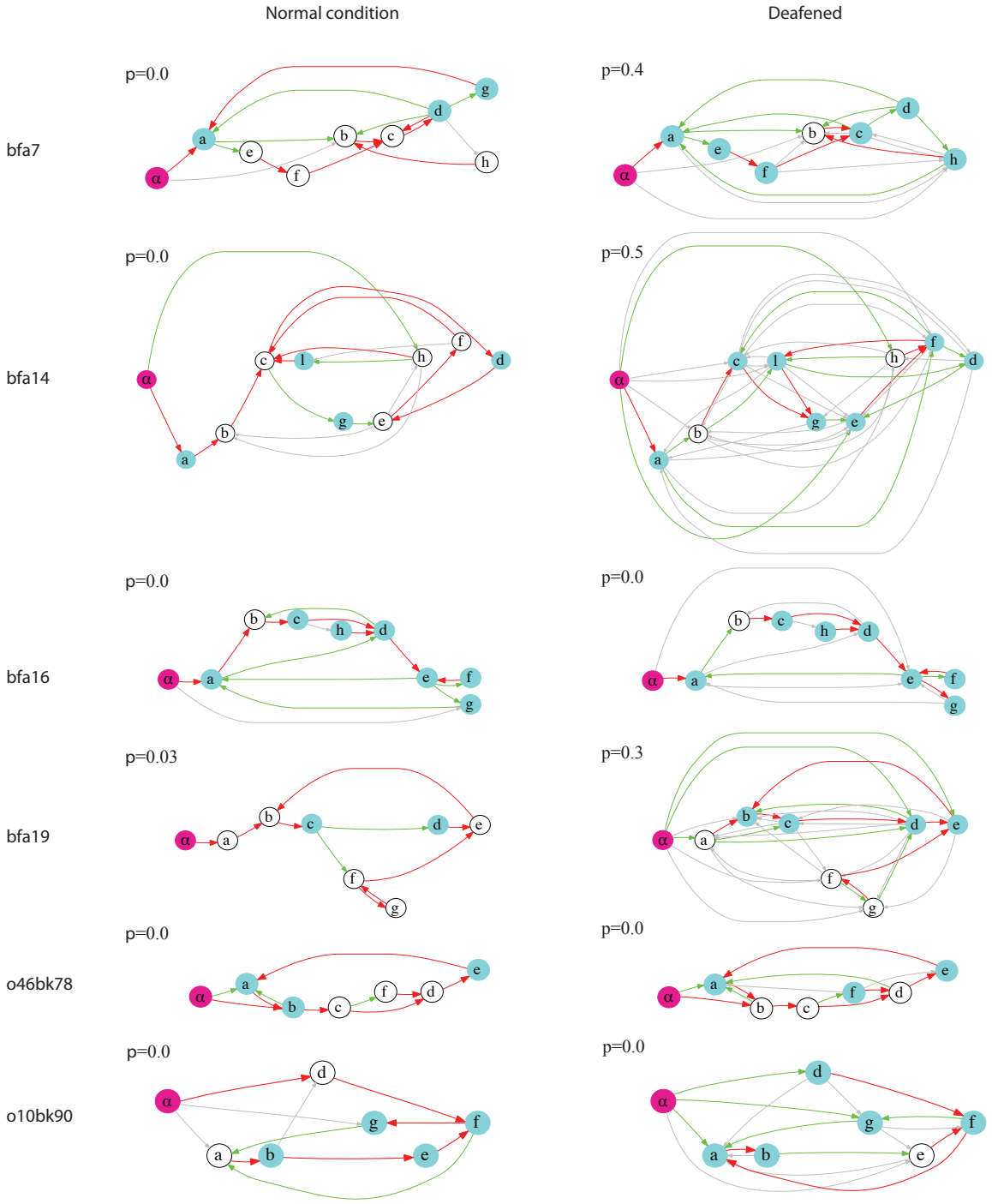


Fig. S3. **Markov models for all birds**. Bird names and the p-values are displayed.

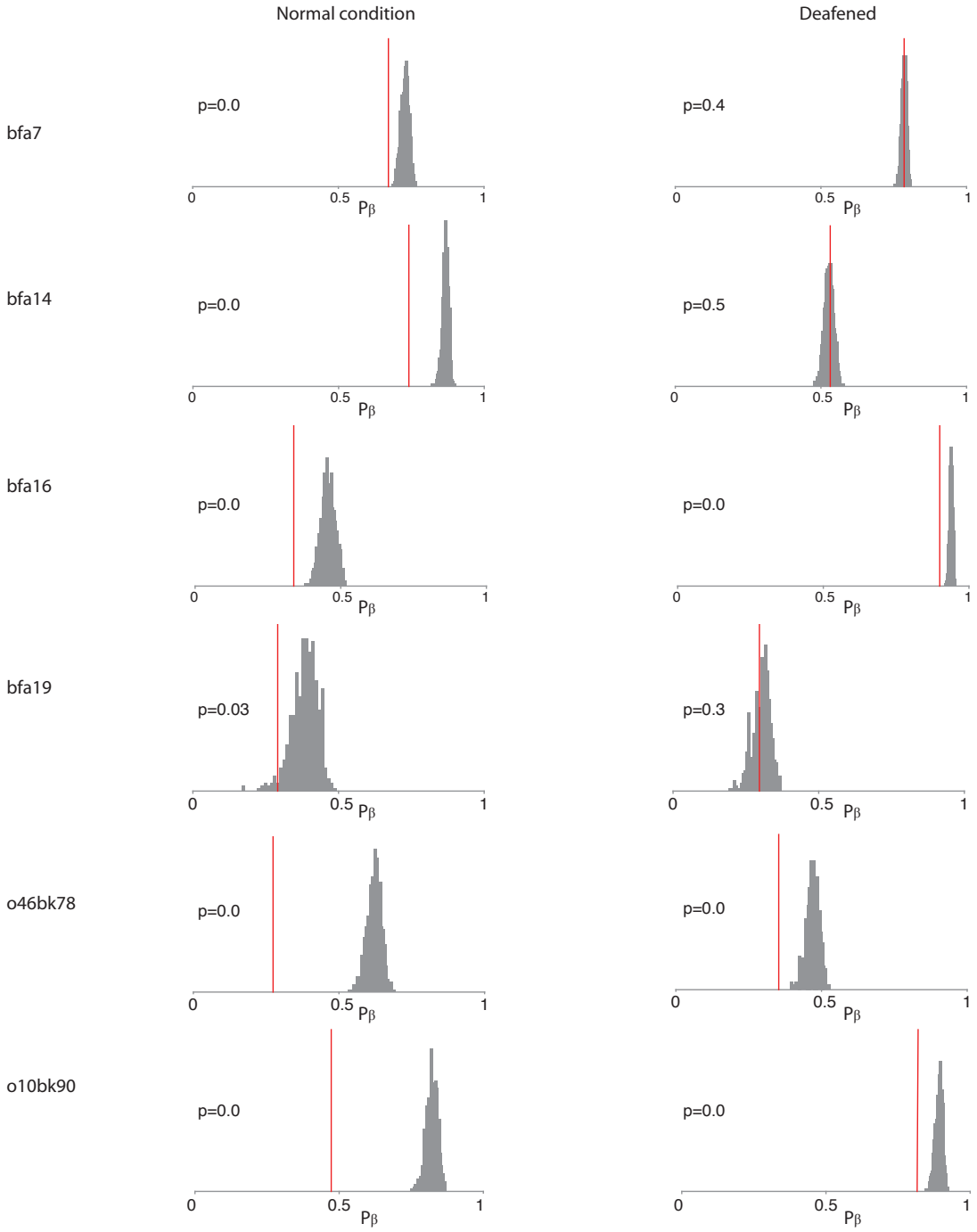


Fig. S4. **Statistical tests of the Markov models for all birds**. The distributions of P_β for the generated sets and P_β of the observed set (red line) are shown. The p-values are displayed.

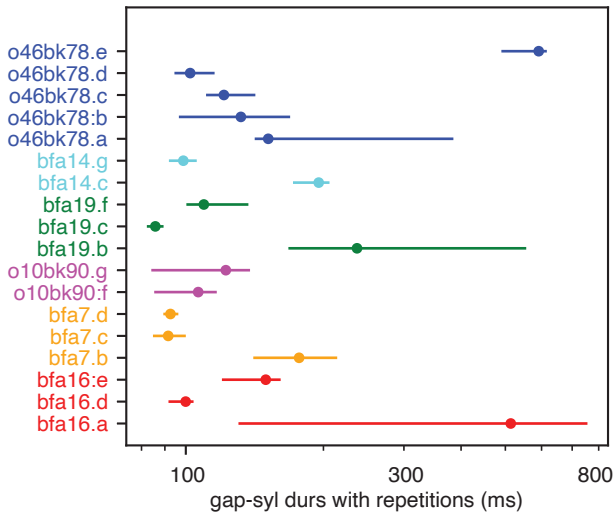


Fig. S5. **Durations of gap-syllables including repetitions**. If a syllable repeats, the duration is defined as the period from the onset of the repetition to the offset of the repetition plus the gap preceding the onset.

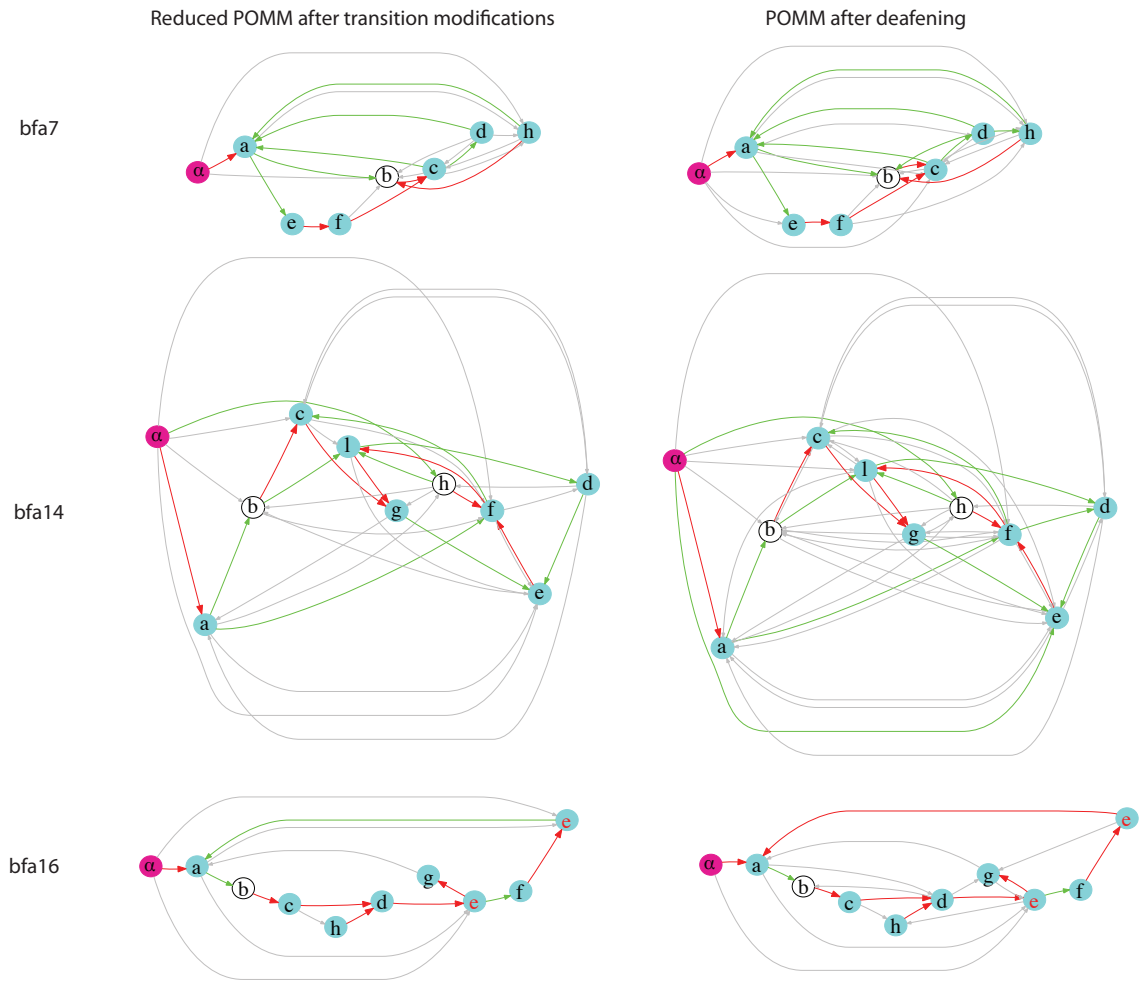


Fig. S6. **Comparisons of the reduced POMMs and the POMMs after deafening for three birds**. The results for bfa7, bfa14, and bfa16 are shown.

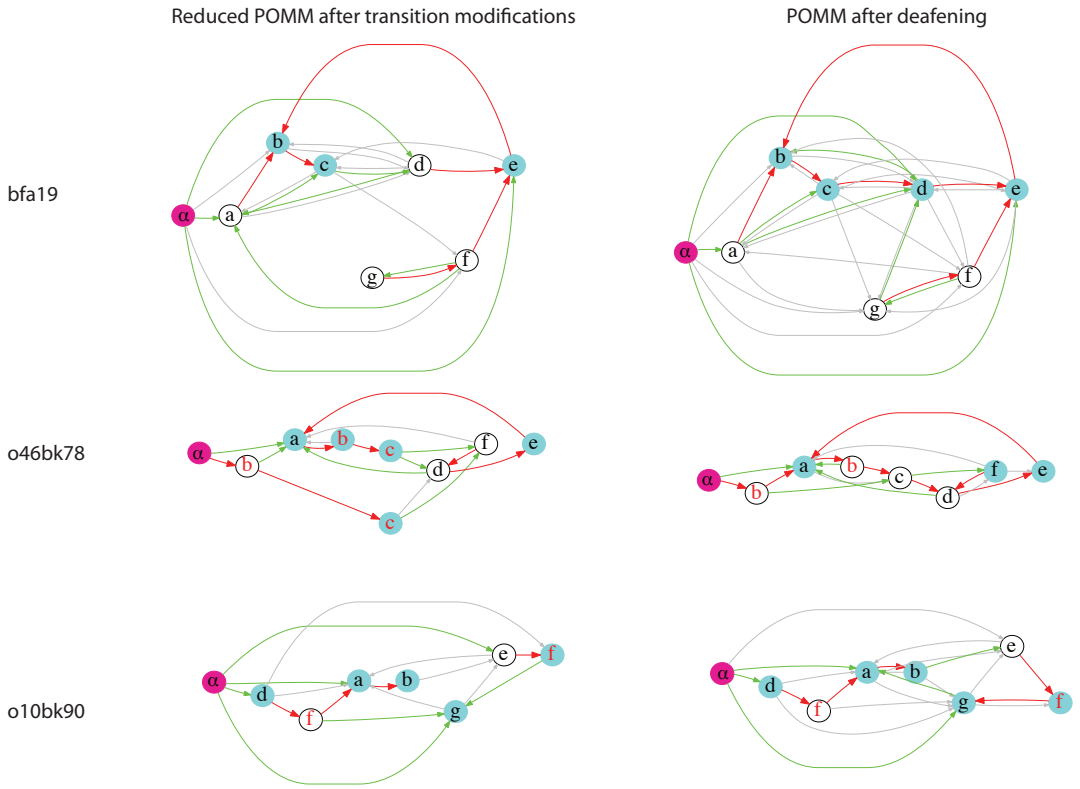


Fig. S7. **Comparisons of the reduced POMMs and the POMMs after deafening for three birds**. The results for bfa19, o46bk78, and o10bk90 are shown.

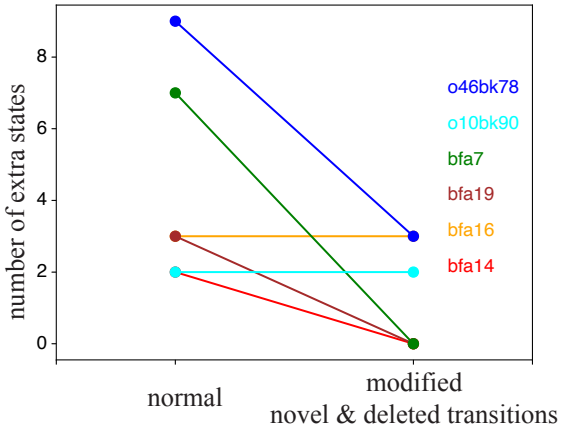


Fig. S8. **Changes of number of extra states in reduced POMMs with only novel and dropped transitions**. For bfa7, bfa19, and bfa14 the reduced POMMs become Markovian. For o46bk78 the state multiplicity is reduced compared to the POMM in the normal condition. No change in state multiplicity is seen for bfa16 and o10bk90.

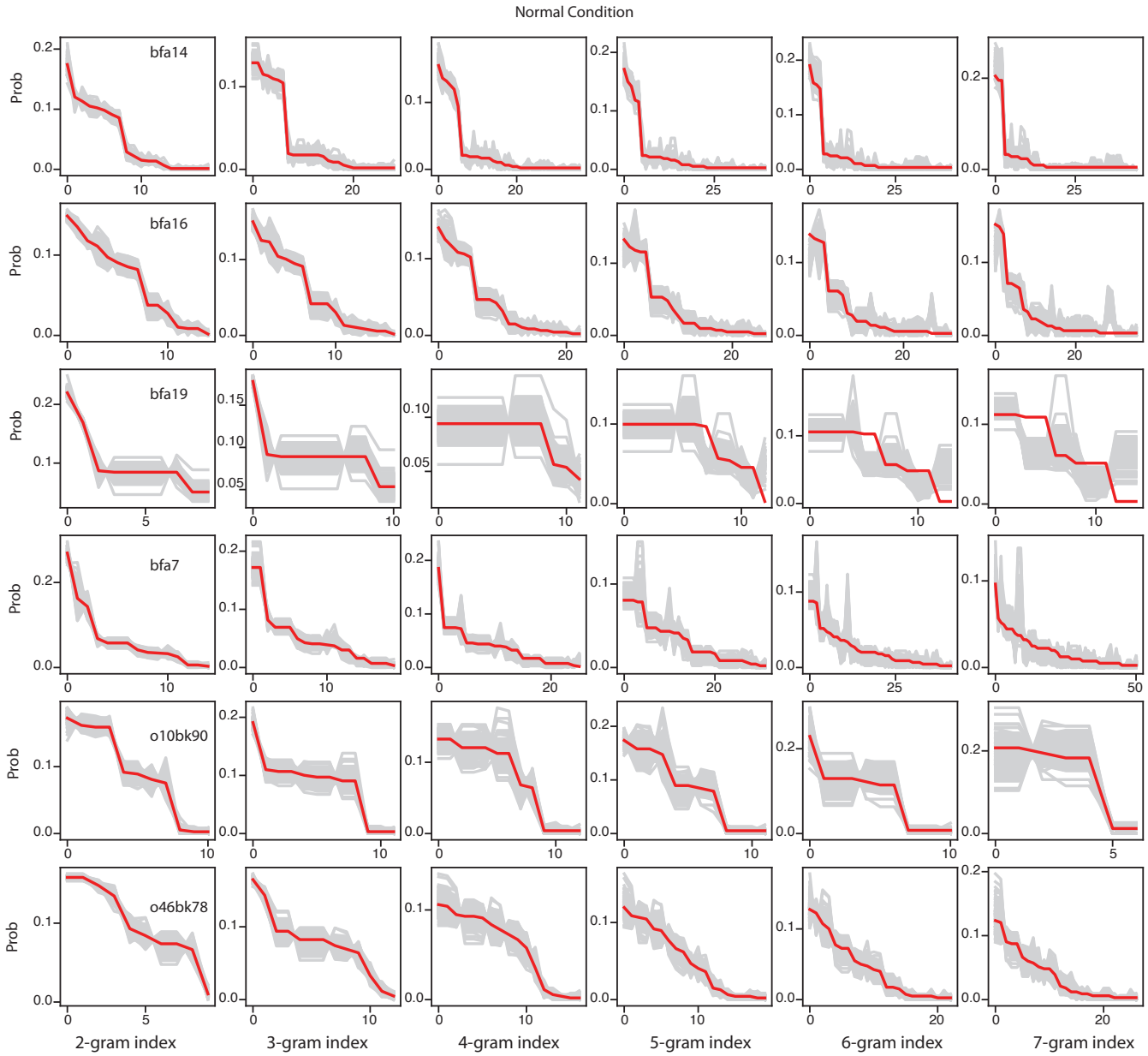


Fig. S9. **Comparisons of n-gram distributions in the normal condition**. The redlines are the probabilities of n-grams in the observed sets. The n-grams are ordered in a descending order of probabilities in the observed sets. The gray lines are the probabilities of the ordered n-grams computed from 100 sets of sequences generated from the POMMs. The red line is mostly within the range defined by the gray lines.

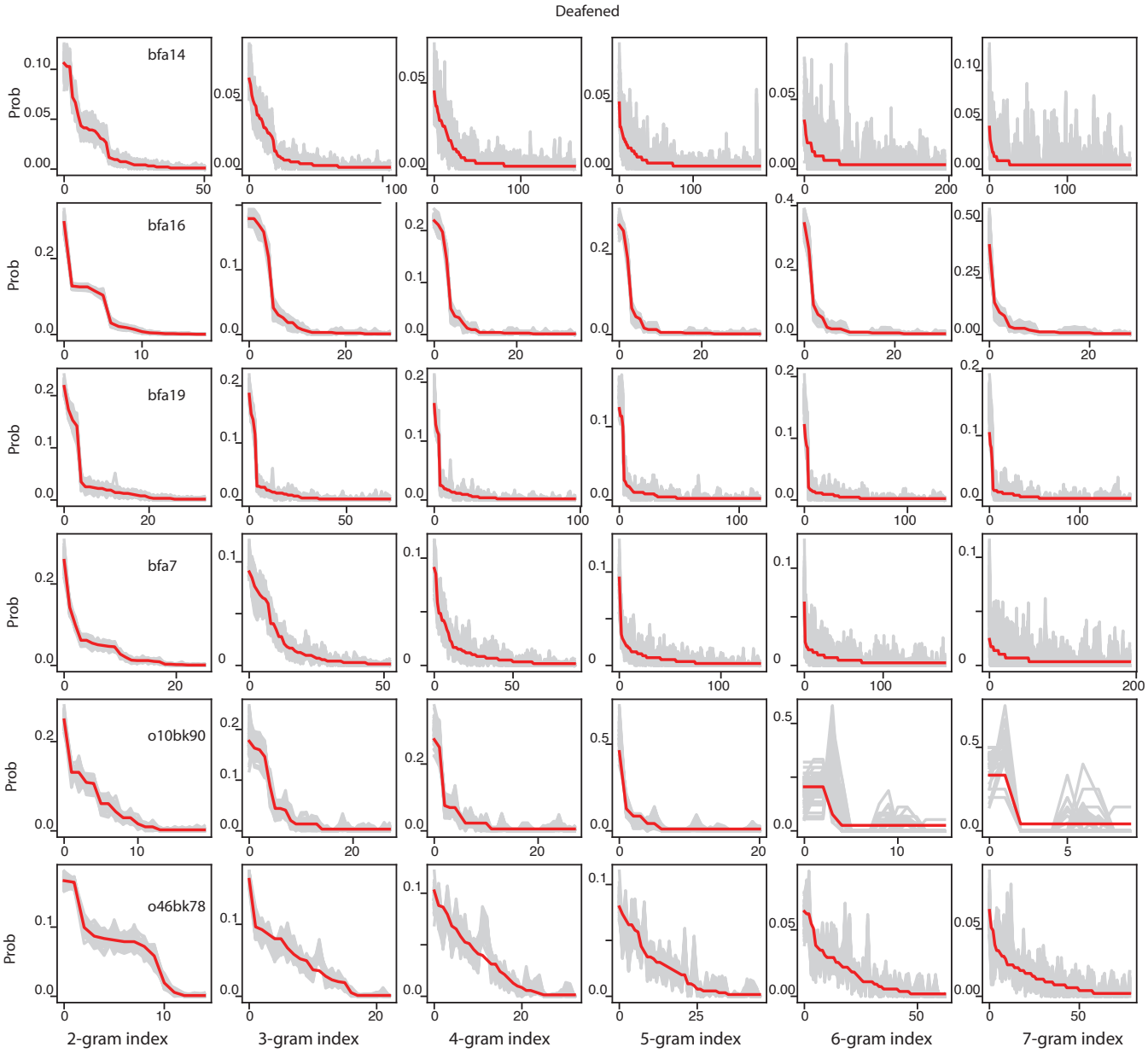


Fig. S10. **Comparisons of n-gram distributions in after deafening**. The same as in Fig. S9 for the deafened case.