OPEN ACCESS Freely available online

PLoS one

Complex Sequencing Rules of Birdsong Can be Explained by Simple Hidden Markov Processes

Kentaro Katahira[1,2,3], Kenta Suzuki[1,3,4], Kazuo Okanoya[1,3,5], Masato Okada[1,2,3]*

1 ERATO, Okanoya Emotional Information Project, Japan Science Technology Agency, Wako, Saitama, Japan, **2** Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan, **3** RIKEN Brain Science Institute, Wako, Saitama, Japan, **4** Graduate School of Science and Engineering, Saitama University, Saitama, Japan, **5** Graduate School of Arts and Sciences, The University of Tokyo, Meguro, Tokyo, Japan

Abstract

Complex sequencing rules observed in birdsongs provide an opportunity to investigate the neural mechanism for generating complex sequential behaviors. To relate the findings from studying birdsongs to other sequential behaviors such as human speech and musical performance, it is crucial to characterize the statistical properties of the sequencing rules in birdsongs. However, the properties of the sequencing rules in birdsongs have not yet been fully addressed. In this study, we investigate the statistical properties of the complex birdsong of the Bengalese finch (*Lonchura striata var. domestica*). Based on manual-annotated syllable labeles, we first show that there are significant higher-order context dependencies in Bengalese finch songs, that is, which syllable appears next depends on more than one previous syllable. We then analyze acoustic features of the song and show that higher-order context dependencies can be explained using first-order hidden state transition dynamics with redundant hidden states. This model corresponds to hidden Markov models (HMMs), well known statistical models with a large range of application for time series modeling. The song annotation with these models with first-order hidden state dynamics agreed well with manual annotation, the score was comparable to that of a second-order HMM, and surpassed the zeroth-order model (the Gaussian mixture model; GMM), which does not use context information. Our results imply that the hierarchical representation with hidden state dynamics may underlie the neural implementation for generating complex behavioral sequences with higher-order dependencies.

Citation: 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(9): e24516. doi:10.1371/journal.pone.0024516

Editor: Gonzalo G. de Polavieja, Cajal Institute, Consejo Superior de Investigaciones Científicas, Spain

Received March 23, 2011; **Accepted** August 12, 2011; **Published** September 7, 2011

Funding: This work was partially supported by Grants-in-Aid for Scientific Research (18079003, 20240020, 20650019) from the Ministry of Education, Culture, Sports, Science and Technology, Japan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: okada@k.u-tokyo.ac.jp

Introduction

Humans can generate complex sequential behaviors such as speech and musical performance. These sequences are typically composed of sequences of actions with complex sequencing rules. How our brain generates such complex sequences is difficult to understand in a straightforward manner since underlying neural circuits are complex and it is difficult to precisely explore neural circuits in human or primate brains. A solution to this issue may be given by studying songbirds [1]. In particular, Bengalese finches sing with apparently more complex sequencing rules with branching points [2,3,4], than does the zebra finch, whose songs are composed of a stereotyped syllable sequence and extensively used for birdsong studies. Bengalese finch songs have been receiving attention as a model of variable sequential behavior, from neurophysiological [5,6,7,8] and theoretical [9,10,11] view points.

To understand the mechanism for the variable sequences of the Bengalese finch song, it is important to characterize the song sequences from a statistical view point. However, the statistical properties of the Bengalese finch song have not been extensively studied. We first demonstrate that the Bengalese finch song has higher-order context dependency: the emission probability for each syllable depends on one or more than one recent syllable. For example in Figure 1A, the emission probability of syllable "c" and "d" depends not only on the adjacent syllable "b" but also on the preceding syllables "a" and "c". This property has been mentioned in previous studies [3]. However, we demonstrated its statistical significance for the first time.

We then investigated the statistical mechanism for explaining higher-order dependencies observed in Bengalese finch songs. To do this, we used the Bayesian inference method and a model selection technique. We applied hidden Markov models (HMMs) with various context dependencies to the acoustic features of a Bengalese finch song and selected a suitable model based on the Bayesian model comparison, its predictive performance, and the degree of agreement with manual annotation. As a result, we found that the first-order HMM, in which the current state appears depending only on the last state, is sufficient and suitable for describing the Bengalese finch song. Perhaps this is a counterintuitive result since the song sequences have higher-order dependency as we mentioned. This is due to a many-to-one state mapping to syllables by which the first-order HMM can generate apparently complex sequences, which we describe in this paper. These results imply that the songbird brain has parsimonious neural representation for generating apparently complex sequences. Also, these results support the branching-chain mechanism, which has been proposed in theoretical studies [9,11], for generating Bengalese finch song sequences.

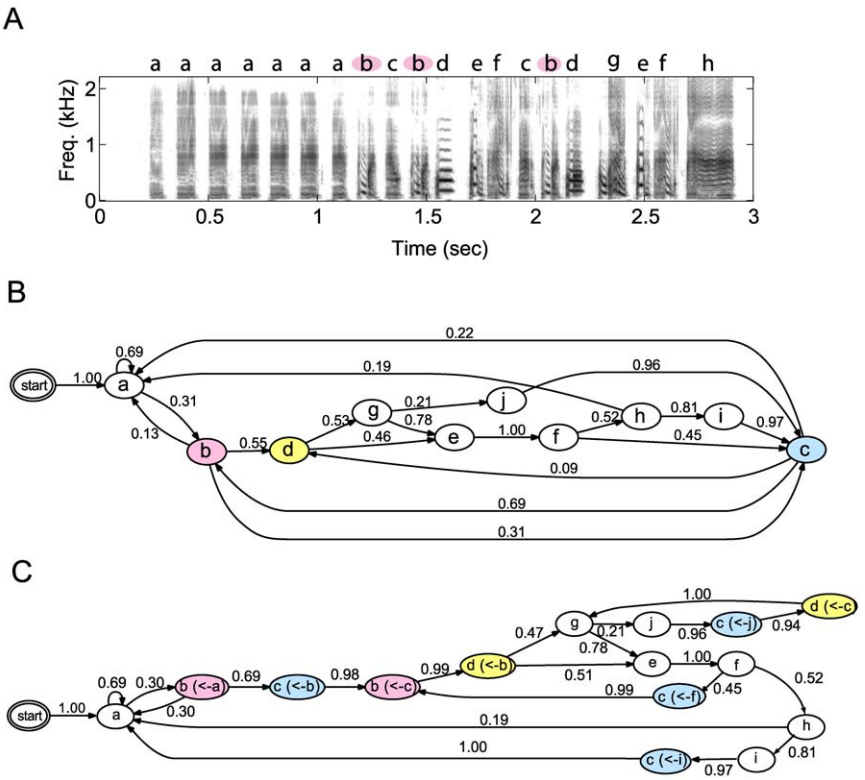

Figure 1. Example of sonogram of Bengalese finch song and its syllable label sequence. (A) Sonogram of Bengalese finch (BF09) with syllable labels annotated by three human experts. Labeling was done based on visual inspection of sonogram and syllables with similar spectrogram given same syllable. (B) Bigram automaton representation (transition diagram) of syllable sequences obtained from same song set as (A). Ellipses represent one syllable and arrows with values represent transitional probabilities. Rare transitions with probabilities <0.01 are omitted. (C) POMM representation of same sequences as (B). Syllables that have significant higher-order dependency on preceding syllables (colored states in (B)) are divided into distinct states depending on preceding syllables (context).
doi:10.1371/journal.pone.0024516.g001

Results

We analyzed the songs of 16 normal adult male Bengalese finches (See Methods for details.) An example of the sonogram (sound spectrogram) of a Bengalese finch song is shown in Figure 1A. The Bengalese finch song consists of acoustically continuous segments, called "song elements" or "syllables" (in this paper, we used the term "syllable") which are separated by silent intervals. Bengalese finch songs are often analyzed by assigning a label to acoustically similar syllables, usually based on visual inspection on the sonogram. Following this approach, we first analyzed the statistical properties of the syllable label strings. We then directly analyzed the acoustic features using statistical models and compared the results to those of an analysis on manual annotated labels.

Higher-order context dependency in syllable sequences of Bengalese finch song

We show that the song syllable strings annotated by three human experts in analysis of birdsong have higher-order context dependency. The three experts labeled based on visual inspection on sonogram. We cross checked by computing Fleiss's κ coefficient [12], which measures the degree of agreement among more than two annotators (see Methods). As a result, the κ-coefficients were 0.972 ± 0.028 (mean $\pm$ SD) for the 16 birds, and all within the range of "Almost perfect agreement", indicating annotation by the three experts was reliable. Hereafter, we use the labeling results by only one of the labeling experts.

We conducted a hypothesis test for each syllable to verify whether the preceding syllables of the syllable being tested affects the occurrence probability of the next syllable (see Methods). We found more than one significant second-order dependency in all 16 birds. When we restricted the analysis to non-repeated syllables, significant syllables were found in the songs of 11 birds. In total, there were 33 significant syllables (21 for non-repetitive syllables) of 72 candidate syllables. An example is shown in Figure 1. In this song, the syllables labeled "b" are preceded by either "a" or "c", and are followed by "a", "c", or "d" (Fig. 1B). If syllable "c" precedes syllable "b", the transition probability from "b" to "d" is 0.99, but if we do not care about the preceding syllable of "b", the transition probabilities to syllables "a", "d", and "c" are 0.13, 0.55 and 0.31, respectively. There was a significant difference between these two probability distributions ($\chi^2(2)=511.99, p<10^{-5}$), indicating that preceding syllables "a" and "c" had a significant effect on the transition probabilities from syllable "b".

This second-order context-dependency can be visually captured by splitting the syllables into distinct states depending on the preceding states. Such representation, in which different states are allowed to emit the same syllable, is regarded as a model called the partially observable Markov model (POMM) [11], thus we call this the POMM representation. For example, the state corresponding to "b" in Fig. 1B is divided into states $b(\leftarrow a)$ and $b(\leftarrow c)$ depending on the preceding syllables (a or c). From the transition diagram where the first-order history dependency was assumed (Fig. 1B), it may seem that transition from syllable "b" to "a", "d", and "c" are random, but with the POMM representation (Fig. 1C),

we can capture the tendency that if "a" precedes, "b" is followed by the syllables "a" or "c", but if "c" precedes, "b" is followed by the syllable "d" almost deterministically. In addition, from the POMM representation, we can see that the syllable sequences "bcbd" and "jcd" are sung in chunks. Taken together, we conclude that the sequencing rules of the Bengalese finch song have higher order Markov dependency, and cannot be described using a simple Markov process, where states and syllables have one-to-one mapping. Nevertheless, a transition diagram in which a simple Markov process is implicitly assumed has been often used for analyzing Bengalese finch songs because of its simplicity [7,13]. We need to be careful when interpreting such representation if we derive the information about variability of the syllable sequence. Even if branching points are found in the diagram, it does not necessarily imply that the following syllable is variable (or stochastic): it may be a stereotyped given more than one previous syllables.

Hidden Markov model analysis on acoustic features

Next we searched for a suitable statistical description of the Bengalese finch song directly from acoustic feature data extracted from the audio-signal of the song. We used HMMs [14], which have been widely used for time-series data modeling, including human speech recognition and also birdsong annotation (but used differently from the present study) [15]. In HMMs, the observed data are assumed to be generated from probabilistic distributions (here, we use a single Gaussian distribution) associated with hidden states, which are usually assumed to be generated from a first-order Markov process. We extend the HMMs to incorporate second-order transition dynamics of hidden states, in accordance with the above results (see Methods section). In addition, we also include the "zeroth-order HMM", which has the same structure as the first-order HMM but without a transition matrix (i.e., context dependency). This model exactly corresponds to the GMM, a statistical model used for data clustering. HMMs with higher than second-order are in principle possible to construct. However, because of their computational cost, which increases exponentially with order, we did not examine them. Furthermore, as can be expected from the following results, such higher-order HMMs would not produce better results than with the first-order HMM. In addition to the order in hidden Markov processes, there is a degree of freedom in the model structure, that is, the number of hidden states, denoted by K. Based on the Bayesian model selection technique (see Methods section) and cross-validation, we explored the best model for describing the Bengalese finch song within a set of our models.

Model comparison

Figures 2A and B compare the lower bound on the log marginal likelihood, which is a model selection criterion (see Methods section), among various hidden states (K) and orders of Markov processes. The model that gives the largest lower bound is regarded as the most appropriate for the given data. This criterion automatically embodies a *Bayesian Occam's razor* [16,17]: a model with many parameters are given a larger penalty than one with fewer parameters. Thus, the simplest model, which can sufficiently describe the given data set, is selected. We see that for a representative bird (Figure 2A) and the average over all birds (Figure 2B), the second-order HMMs showed a larger bound when a small number of states (K) were given. However, for a large number of states, the first-order HMMs gave the largest bound. Similar results were obtained for almost all the song data from other birds we analyzed. The exceptions were for the songs in which no significant second-order context dependency (excluding the repetitive syllables) was observed. Two-way ANOVA with factors K and the model order revealed that there were main effects of both factors ($F(15,225) = 295.25$, $p < 10^{-5}$ for K; $F(2,30) = 1123.3$, $p < 10^{-5}$ for model order) and interaction ($F(30,450) = 68.559$ $p < 10^{-5}$). The results of post hoc comparison between the first- and second- order models for each K (with Bonferronni corrections) are plotted in Figure 2B. We see that the lower bounds for the first-order HMM (with $K > 14$) were significantly larger than those for second-order HMM. These results with the approximate Bayesian model comparison suggest that the first-order HMM would suffice when sufficient hidden-states are available. Here, "sufficient hidden-states" is related to the number of syllable types in song and degree of second-order history dependency. The number of syllable types for each bird derived from manual annotation was 9.75 ± 2.77 (mean $\pm$ SD). We see that if the number of states is larger than this number, the first order HMM reaches the performance of the second order model (Figure 2B, D; see Table S1 for for detailed data of individual bird). We will interpret these results in more detail in the Discussion section.

To evaluate how well the models describe the statistics of song acoustic features more directly, we computed the predictive performance of the models based on cross-validated log-likelihood on test data (that were not used for model training). The test data consisting of ten bouts for each bird were constructed from the song recorded from the same bird on the same date with the training data. The results for a representative bird and all birds are shown in Figures 2C and D, respectively. The number of states and model order significantly affected the predictive performance ($F(15,225) = 301$, $p < 10^{-5}$ and $F(2,30) = 1107.2$, $p < 10^{-5}$, respectively). Also, there was significant interaction between the number of states and model order ($F(30,450) = 68.861$ $p < 10^{-5}$). The post hoc comparison between the first- and second- order models for each K revealed that the second-order model yielded significantly better performance than the first-order model if K was small, and for $K > 10$ there was no significant difference between the two models (Figure 2D). Thus, we cannot claim that the first-order HMM is superior in predictive performance, but it is at least comparable to the second-order HMM.

Comparison with manual annotation

Next we discuss how each model annotates given song syllables. We first compare them with the manual annotations described above (For details in computing the model annotation, see Methods section). We evaluated the agreement between the annotations of the models and of human experts by computing Cohen's kappa coefficient, which measures the degree of agreement between two annotators (see Methods section). Figure 3A shows the results for all songs. With a sufficient number of hidden states, the average performances (thick lines) of the first- and second-order HMMs reached the region of "almost perfect agreement", while the GMM saturated in the region of "substantial agreement". Thus, the syllable sequences obtained from the first- and second-order HMMs were in almost perfect agreement with those obtained from manual labeling, while GMM did not provide a comparable result.

Kappa coefficients for each model-order (with K selected using the lower bound) were 0.781 ± 0.103 (mean $\pm$ SD) (range from 0.566-0.905) for the zeroth-order model (GMM), 0.911 ± 0.055 (range from 0.790-0.978) for the first-order HMM, and 0.873 ± 0.090 (range from 0.688-0.973) for the second-order HMM. There were significant differences between GMM and the-first order HMM ($p < 0.001$), and between GMM and the second-order HMM ($p < 0.001$). While the mean performances of the-first order

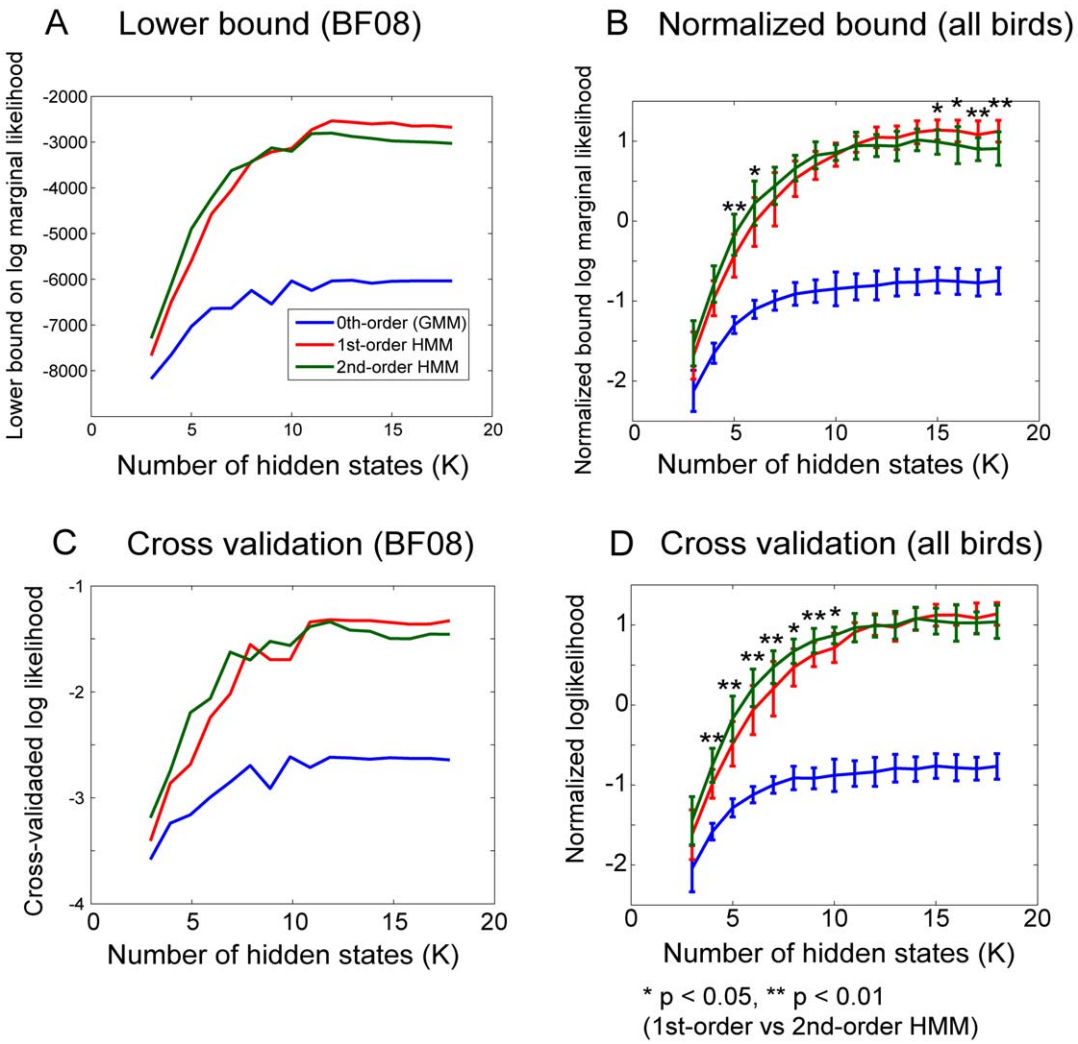

Figure 2. Comparison of statistical models with various states K and model orders on acoustic features of Bengalese finch song. (A-B) Plot of lower bound on marginal log likelihood. Larger this bound, the more appropriate model is for representing given data. For both cases, first-order HMM gave largest bound provided there was sufficient number of states available. (C-D) Cross validated log-likelihood on test data sets obtained from same bird on same date but ten different bouts from those used for training model. (A,C): representative bird (BF08). (B, D): average over all birds on normalized value. Error bars indicate standard deviation.
doi:10.1371/journal.pone.0024516.g002

HMM were slightly better than those of the second-order HMM, there was no significant difference between them. The differences between GMM (zeroth-order HMM) and the two HMMs suggest that taking into account the context information improves stable labeling. Human experts may use such context information.

Figure 3B compares manual labeling and the labeling of two of models, first-order HMM and GMM, in which the number of hidden states K is selected based on the model selection criteria described in the Methods section. We observed that the two models tended to divide syllables into larger letters. For example, a human labels the first repeated introductory notes using only "a", but the two models label them using two states "a1" and "a2". This may reflect the fact that our method is more sensitive to differences in acoustic features than humans. An important difference between the zeroth-order model and first-order HMM is in the sequence "bcbd" by manual labeling. As shown in Figure 1, the subsequent syllable after the first "b" and the second "b" depends on the previous syllable (whether "a" or "c"), and these two syllables "b" are divided into distinct states in the POMM representation (Figure 1C). The first-order HMM obtained the following representation: it divided the syllable "b" into the states "b1" and "b2", while the zeroth-order model, the GMM, did not (red rectangle in Figure 3B). This difference is due to context dependency. As we can see in the red rectangle in Figure 3C, the distributions of "b1" and "b2" largely overlap; thus, indistinguishable without using information of the preceding syllables. For the other songs we analyzed, similar properties were often observed: of the 54 syllables where significant second-order dependency was found in manual annotation-based analysis, the first-order HMM divided 30 syllables into distinct states according to the preceding syllables, while the GMM did so for 17 syllables and the second-order HMM did for 23. As a recent study showed, the contexts affect the acoustic properties [13]. Thus, even the GMM, which does not incorporate pre-state dependency, aligned the different states for the same syllables solely on the differences in acoustic features. However, the difference between the GMM and HMMs suggests that HMMs tend to align different syllables depending on the context, not solely on the acoustic features.

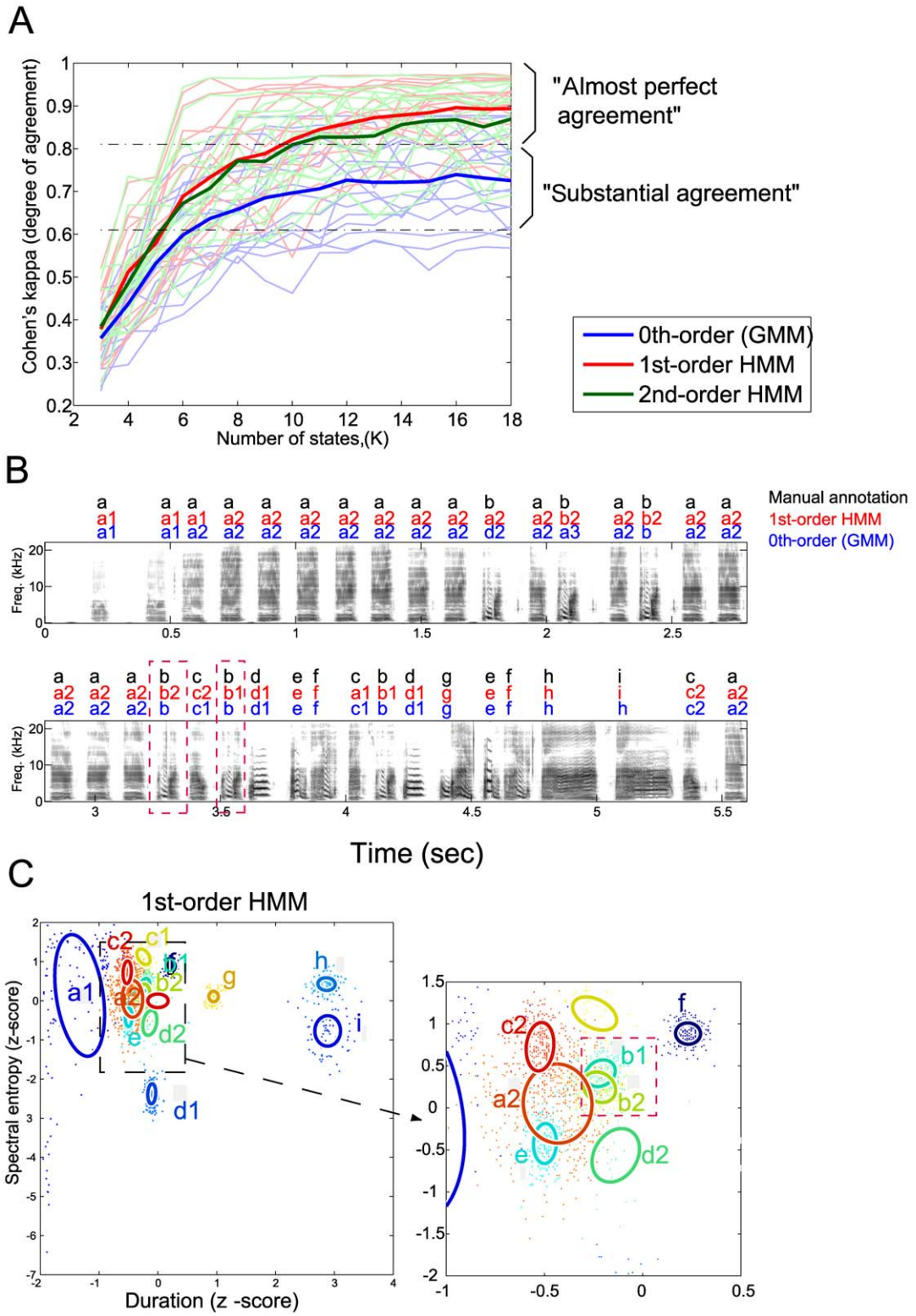

Figure 3. Comparison of annotation results using various models and human experts. (A) Kappa's coefficients, which are measure for agreement between model annotations and manual annotation by human experts, are functions of number of states, K. Thin lines represent results for individual birds, and thick lines represent average for each model order. (B) Example of annotations for song of BF09. Black labels represent manual annotations done by visual inspection of sonograms. Red and blue labels are labeled using Gaussian mixture model (zeroth-order model) with $K = 13$ and first-order HMM with $K = 18$, respectively. Number of states (K) that gives the highest lower-bound on log marginal likelihood were used. (C) Example of model fitting results on sound feature space (duration and spectral entropy) for same song as (B). Results from first-order HMM. Ellipses represent contour of Gaussian distribution of each state, and letters indicate syllable aligned to state.
doi:10.1371/journal.pone.0024516.g003

Discussion

We explored the statistical properties of complex sequencing rules of the Bengalese finch song. To achieve this, we analyzed history dependencies in the syllable sequences annotated by human experts. Then we applied statistical models to acoustic feature data. We discuss the implications of our results and possible neural implementation.

First-order HMM is sufficient for producing higher-order Markov sequences

We have seen that in all songs we examined, the first-order HMMs showed comparable or superior performance (i.e., gave a larger lower bound, better agreement with manual annotation, and cross-validated prediction error) compared with the second-order HMMs when there were enough hidden states compared to the number of syllable types, whereas when the number of states was small, the second-order HMMs performed better. These results suggest that (1) when the number of hidden states (labels) is small, considering the higher-order context dependency among hidden states, it leads to a better explanation of the birdsong data, but (2) when we use enough hidden states, the first-order HMMs become sufficient for explaining the data. Such models give better descriptions of data by using a smaller parameter set than higher-order HMMs.

These findings perhaps are counterintuitive given that our first observation was that some syllables in the Bengalese finch song sequences have dependence on at least two previous syllables. We interpreted our results as follows. Even if the hidden state sequences of the first-order HMMs have only first-order dependency, the emitting syllable acoustic features can have higher-order dependency. This can occur when the different hidden states (States 2 and 4 in Figure 4A) have similar emission distributions (corresponding syllable "b" in Figure 4A). Although the hidden state sequence is a first-order Markov sequence ("12345…"), the emitted syllables can have second-order dependency ("abcbd…"). This representation was that the first-order HMM was indeed attained through an automatic parameter fitting process. This representation is possible when the first-order HMM can use redundant states to assign the distinct states to the syllables with similar acoustic features. Here, "redundant" refers to the situation when the number of states is larger than the number of syllable types.

By adopting this representation, we can avoid exponential growth in the number of parameters when second-order context dependency is sparse. Let us consider an extreme case where all syllables truly depend on two previous syllables (i.e., all components of the transition matrix from the previous two states to the next state are mutually independent). For this case, if we use first-order HMMs to represent the data in the above manner, we need K^2 hidden states. This requires a $K^2 \times K^2$ transition matrix, which is larger than that of second-order HMMs ($K^2 \times K$). In addition, first-order HMMs have parameters for K^2 Gaussian components, whereas second-order HMMs have those for only K components. If two different models can represent the true data distribution, the model with fewer parameters gives a larger marginal log likelihood and better predictive performance. Thus, in this case, a second-order HMM will be selected. On the other hand, if most syllables depend only on one last syllable, and some portion of the syllable depends on the two previous syllables, a first-order HMM can represent the statistical structure of the sequence and will be selected. Our results suggest that the statistical structure of the Bengalese finch song is close to the latter.

Possible neural implementation

These results motivate us to discuss neural implementation of this statistical model structure. In songbirds, two nuclei are mainly related to generating songs: the HVC (proper name) and the robust nucleus of the archistriatum (RA). Neural activity in the HVC appears to encode sequential information [18,19], while RA encodes the acoustic structure of individual song syllables [18,20,21]. The HVC projects to the RA, while the RA projects to the nuclei that control the syrinx and respiration muscles. The sequential pattern in birdsongs is assumed generated in a feedforward chain of RA-projecting neurons in the HVC [19,22,23]. A theoretical study has shown that such a feedforward-chain mechanism can be extended to generate stochastic branching sequences that obey a first-order Markov process [11]. Whether it can be extended to a higher-order Markov process is unknown. Our results imply that a feedforward chain mechanism

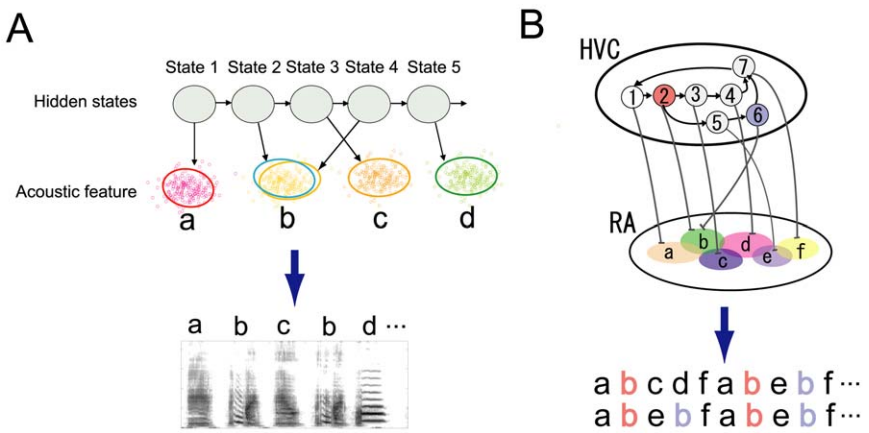

Figure 4. Schematic diagram representing how first-order HMM generates sequences that have higher-order context dependency and its neural implementation. (A) key point is that different states (States 2 and 4) can generate similar acoustic feature space ("b"). This mechanism allows observed song sequences to have higher-order context dependency, even if hidden state sequences are generated from simple Markov process. (B) Schematic of proposed model for neural implementation of Bengalese finch song syntax. Each circle in HVC represents neuron group consisting of feedforward chain of RA-projection neurons. Each group in HVC plays role in generating particular syllables through neuron group in RA. Groups 2 and 6 in HVC project to same RA neuron group that generates song syllable "b", as do States 2 and 4 in (A).
doi:10.1371/journal.pone.0024516.g004

that obeys the first Markov process would suffice and be suitable to explain the song syntax of the Bengalese finch. We propose a mechanism that relates our statistical model and the neural circuits. First, we relate a group of feedforward chains to the hidden state of a first-order HMM. Figure 4B illustrates this model. The circles in the HVC represent the neuron groups consisting of a feedforward chain of RA-projecting neurons. The arrows represent the firing order of the chain. The firing order is a first-order Markov process that includes stochastic transitions from group 2 to group 3 or group 5. We assume that different groups (groups 2 and 6) generate the same syllables, "b", by having a similar projection pattern to the RA. Due to this mechanism, even though the firing order of the HVC(RA) neurons in the HVC is a first-order Markov process, the observable syllable sequences obeys higher-order Markov processes, as observed in our song analysis with the first-order HMM. Although this mechanisms have been predicted in our own theoretical work [9] and another study [11], our present method automatically extracted the same representation of sequence from real acoustic data of birdsong, strengthening the plausibility of this mechanism.

Related work

Jin and Kozhevnikov recently proposed a statistical model for describing the sequencing rules of Bengalese finch songs independently of us [24]. While the mechanism behind their model is similar to ours in the sense that many-to-one mapping from states to syllables allow the model to generate sequences with higher-order history dependency, the present study differs from theirs in several respects. First, Jin and Kozhevnikov adopted two-stage methods for extracting syllable sequences from audio signals: they first assigned syllable labels for the audio signals of segmented syllables by using an automatic clustering method and then they constructed syntax models for describing the sequences of syllable labels. On the other hand, our method using HMMs directly extracts sequence structure from audio features, thus, the results are not directly affected by the property of syllable labeling. Instead, the results of our method depend on the selection of auditory features and emission probability models for each state. As the HMMs in the present study can represent sequential variability and acoustical variability in a common statistical model, dissociating these variabilities is relatively easy. The sequential variability is represented by the entropy of state transition, while acoustical variability is represented by the entropy of emission probability distribution for each state. It may be possible to associate these two kinds of variability with corresponding neural variability. The former may be related with the activity of HVC and the later the activity of RA. Second, Jin and Kozhevnikov did not compare the annotation obtained by their automatic method with manual annotation made by visual inspection. Although manual annotation involves arbitrariness, it has been used in many studies on birdsong. A novel point of the present study is to compare the model-based annotation and manual annotation while checking the reliability of the manual annotation by cross-checking between three annotators. Third, as Jin and Kozhevnikov analyzed the songs from only two Bengalese finches; whether their results are common properties for Bengalese finches is questionable. In the present study, we analyzed the songs from 16 Bengalese finches and found that some of them have no second-order dependency excepting in the repetitive syllables. Fourth, Jin and Kozhevnikov did not consider the models with state transition that obey higher-order Markov processes: state transitions in their models obey a first order Markov process (except for the effect of adaptation). Thus, the possibility that the model with higher-order Markov process could describe the song sequences better than their models was not discussed. In contrast, we included the model with second-order Markov processes and showed that the many-to-one mapping with first-order Markov process is at least comparable to the second-order model. These differences of the present study do not produce results contradicting those of Jin and Kozhevnikov. The results rather strengthen the claims also made in Jin and Kozhevnikov. What was missing in our model, though incorporated in Jin and Kozhevnikov's, is the effect of adaptation. This serves to describe the statistics of repetitive syllables. In the present study, we did not focus on repetitive syllables. Incorporating the effect of adaptation in our model would improve the descriptive power for the Bengalese finch songs.

Future work

We showed that the first order HMM gives annotations that were close to manual annotations compared to a standard clustering technique (GMM), which does not use context information. This result suggests that our method with an ordinary HMM can be used as a convenient tool for annotating the Bengalese finch song, instead of time-consuming manual annotation. We applied our method to the songs of only adult healthy Bengalese finches. Investigating the developmental change of the Bengalese finch song or those developed with abnormal conditions such as isolated from song tutors, or with lesions in the song-related nucleus, may be for future study. Such studies will give valuable insight into how complex sequencing rules are formed through learning. We also applied HMMs to neural activities recorded from HVC in an anesthetized condition for extracting neural state transitions in a previous study [25]. It would be interesting to see the relationship between the state transitions extracted from activities in HVC of singing bird using such methods, with the state transitions extracted from its song as we did in the present study.

Methods

Recording

We analyzed undirected songs (songs in the absence of a female) of 16 adult male Bengalese finches (labeled as BF01-BF16) ranging 133-163 days of age. They were raised in colonies at the RIKEN Brain Science Institute. Before recording, each bird was moved to a sound proof room and isolated from the other birds. Songs were recorded for 24 hours using a microphone placed in the room. All experimental procedures and housing conditions were approved by the Animal Care and Use Committee at RIKEN (approval ID: H22-2-217).

Sound feature extraction

To extract acoustic features from each syllable, we used Sound Analysis Pro (SA+) software [26], which is a widely used tool for quantifying song features in birdsongs ([27] and references therein). We used three representative features: syllable duration, mean pitch, and mean Wiener (spectral) entropy. We applied a feature batch module in SA+ for extracting the acoustic features from wave format audio files. We then randomly picked and analyzed thirty song bouts for each bird from all recordings. For cross-validation, ten additional bouts were also randomly picked.

Evaluation of agreement of annotations

Annotation analysis. To evaluate the agreement between manual annotations by different human annotators and between manual and model annotation, we used Fleiss's κ coefficient and Cohen's κ coefficient, respectively [12]. They are statistical measures of inter-annotator agreement for categorical items.

They are more robust measures than simple percent agreement calculation since they take into account agreement occurring by chance. Cohen's kappa measures agreement between two raters, while Fleiss' kappa does when there are more than two raters. For both measures, if κ-coefficients fall in the range of 0.81-1.00, the result is interpreted as "Almost perfect agreement". For the range of 0.61-0.80 - "Substantial agreement", 0.41-0.60 - "Moderate agreement", 0.21-0.40 - "Fair agreement", 0.0-0.20 - "Slight agreement", and <0 - "Poor agreement".

Evaluation of second-order context dependency

To evaluate second-order context dependency, we conducted a hypothesis test for each syllable to verify whether the preceding syllable of the syllable being tested affects the occurrence probability of the next syllable. In particular, we seek the syllable that has both more than one preceding syllable and more than one subsequent syllable. We then tested whether the probabilities of transitions from the syllable depend on the preceding syllable by doing a χ^2 test of goodness of fit between the probability distributions that ignore the preceding syllables and those conditioned on the most frequent preceding syllable [11]. We interpret the syllable having second-order context dependency if $p \leq 0.05/n$, where n denotes the number of candidate syllables (the Bonferroni correction).

Higher-order hidden Markov models

We consider a second-order HMM, whose directed graphical model is shown in Figure S1. At first glance, it may seem difficult to apply a forward-backward algorithm, which is the standard algorithm for inferring hidden state sequences, to this model. However, if we combine two succeeding states into one *context states*, we can transform this graphical model into one with the same form as the first-order HMM, as shown in Figure S1. We introduced a dummy state denoted as d_1 for the beginning of the sequences. If we use an m-th order HMM, $m-1$ dummy states $(d_1,...,d_{m-1})$ are required. For the second-order HMM, there are $K+K^2$ context states, including ones that contain dummy states. In general, there are $\sum_{l=1}^{m-1} K^l + K^m$ context states. For each context state, the number of transition targets is K. Hence, the transition matrix has $(\sum_{l=1}^{m-1} K^l + K^m) \times K$ elements.

Parameter fitting

To train HMMs using given acoustic feature data, we used the Variational Bayes (VB) method [28,29,17]. The VB method has been widely used as an approximation of the Bayseian method for statistical models that have hidden variables. The VB method approximates true Bayesian posterior distributions with a factorized distribution using an iterative algorithm similar to the expectation maximization (EM) algorithm. For the limit of a large number of samples, the results of the VB coincides with those of the EM algorithm. We used VB because of the following two advantages: (1) its low computational cost, which is comparable to the EM algorithm, and (2) it can select an appropriate model based on the model-selection criterion computed in a model learning process. Full Bayesian approaches based on a sampling technique give a more accurate model-selection criterion, but their high computational cost is unfavorable for our purpose (especially for the second order HMMs, which have a large number of parameters).

The VB algorithm for the GMM is detailed in [17], while those for the first-order HMMs are detailed in [29]. We derived the VB algorithm for the second-order HMMs for the first time, but we only have to change the transition matrix from the algorithm for the first-order HMMs, which is a straightforward extension as described above.

Model selection

We denote the model index M, which refers to the number of states K and order of Markov process m. By using Bayes theorem, the posterior of the model index given data X is given by

$$p(M|X) = \frac{p(X|M)p(M)}{p(X)}. \quad (1)$$

We naturally assume that $p(M)$ is the uniform distribution, i.e., we have no a priori assumption of the model structure. Then, $p(M|X) \propto p(X|M)$, hence the model gives the highest posterior probability that corresponds to the one that gives the highest marginal likelihood $p(X|M)$. Ideally, the marginal likelihood $p(X|M)$ is obtained by marginalization over hidden variable sets (denoted as Y) and parameter sets (denoted as θ) as

$$p(X|M) = \sum_Y \int d\theta p(X,Y|\theta,M)p(\theta|M). \quad (2)$$

However, this marginalization procedure is infeasible. Therefore, we used a lower bound on the log marginal likelihood $\log p(X|M)$ instead. The variational free energy F, which is computed in a model learning process, gives an upper bound on $-\log p(X|M)$. In other words, the log marginal likelihood $\log p(X|M)$ is lower bounded by negative variational free energy $-F$. To emphasize the statistical meanings, we call $-F$ the lower bound on the log marginal likelihood.

Computing model annotation

We assumed that each hidden state in the models we used represents one syllable. The models asigned the label that corresponds to the state that gave the highest posterior probability of generating the acoustic features for each syllable (see Methods). The posterior probabilities were computed using the Baum-Welch algorithm [14]. We then aligned a syllable label to each state so that the aligned labels were the most frequently labeled syllables by human experts in the syllable set that the model state aligned. We allowed more than one state to share the same syllable (many-to-one mapping from states to a syllable).

Supporting Information

Figure S1 Graphical model representation for second-order HMM describing how parameter estimation for second-order HMM can be done. (A) Naive graphical model for second-order HMM. In this graph, we introduce a node (represented as a circle) for each random variable. For each conditional distribution, we add arrows to the graph from the nodes corresponding to the variables on which the distribution is conditioned. (B) Another representation of second-order HMM using context states that combine two states.
(TIF)

Table S1 Statistics of songs recorded from individual birds.
(DOC)

Acknowledgments

We are grateful to Olga Feher (RIKEN, BSI) for checking our manuscript and giving us useful advices. We also thank the anonymous reviewer for helpful suggestions.

Author Contributions

Conceived and designed the experiments: KK KO MO. Performed the experiments: KS. Analyzed the data: KK KS. Wrote the paper: KK MO.

References

1. Doupe AJ, Kuhl PK (1999) Birdsong and human speech: common themes and mechanisms. Annual Review of Neuroscience 22: 567–631.
2. Honda E, Okanoya K (1999) Acoustical and syntactical comparisons between songs of the whitebacked munia (Lonchura striata) and its domesticated strain, the Bengalese finch (Lonchura striata var. domestica). Zoological Science 16: 319–326.
3. Okanoya K (2006) The Bengalese finch: a window on the behavioral neurobiology of birdsong syntax. Annals of the New York Academy of Sciences 1016: 724–735.
4. Okanoya K, Yamaguchi A (1998) Adult Bengalese finches (Lonchura striata var. domestica) require real-time auditory feedback to produce normal song syntax. Journal of Neurobiology 33: 343–356.
5. Hosino T, Okanoya K (2000) Lesion of a higher-order song nucleus disrupts phrase level complexity in Bengalese finches. Neuroreport 11: 2091.
6. Okanoya K (2004) Song syntax in Bengalese finches: proximate and ultimate analyses. Advances in the Study of Behavior 34: 297–346.
7. Sakata JT, Brainard MS (2006) Real-time contributions of auditory feedback to avian vocal motor control. Journal of Neuroscience 26: 9619.
8. Sakata JT, Brainard MS (2009) Social context rapidly modulates the influence of auditory feedback on avian vocal motor control. Journal of Neurophysiology 102: 2485.
9. Katahira K, Okanoya K, Okada M (2007) A neural network model for generating complex birdsong syntax. Biological Cybernetics 97: 441–448.
10. Yamashita Y, Takahasi M, Okumura T, Ikebuchi M, Yamada H, et al. (2008) Developmental learning of complex syntactical song in the Bengalese finch: A neural network model. Neural Networks 21: 1224–1231.
11. Jin DZ (2009) Generating variable birdsong syllable sequences with branching chain networks in avian premotor nucleus HVC. Physical Review E 80: 51902.
12. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychological Bulletin 76: 378–382.
13. Wohlgemuth MJ, Sober SJ, Brainard MS (2010) Linked Control of Syllable Sequence and Phonology in Birdsong. Journal of Neuroscience 30: 12936.
14. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications inspeech recognition. Proceedings of the IEEE 77: 257–286.
15. Kogan JA, Margoliash D (1998) Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. The Journal of the Acoustical Society of America 103: 2185.
16. MacKay D (1992) Bayesian interpolation. Neural Computation 4: 415–447.
17. Bishop C (2006) Pattern recognition and machine learning. Springer: New York, .
18. Yu AC, Margoliash D (1996) Temporal hierarchical control of singing in birds. Science 273: 1871.
19. Hahnloser RHR, Kozhevnikov AA, Fee MS (2002) An ultra-sparse code underliesthe generation of neural sequences in a songbird. Nature 419: 65–70.
20. Leonardo A, Fee MS (2005) Ensemble coding of vocal control in birdsong. Journal of Neuroscience 25: 652.
21. Sober SJ, Wohlgemuth MJ, Brainard MS (2008) Central contributions to acoustic variation in birdsong. Journal of Neuroscience 28: 10370.
22. 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.
23. Long MA, Fee MS (2008) Using temperature to analyse temporal dynamics in the songbird motor pathway. Nature 456: 189–194.
24. Jin DZ, Kozhevnikov AA (2011) A compact statistical model of the song syntax in bengalese finch. PLoS Computational Biology 7: e1001108.
25. Katahira K, Nishikawa J, Okanoya K, Okada M (2010) Extracting state transition dynamics from multiple spike trains using hidden markov models with correlated poisson distribution. Neural Computation 22: 2369–2389.
26. Tchernichovski O, Nottebohm F, Ho CE, Pesaran B, Mitra PP (2000) A procedure for an automated measurement of song similarity. Animal Behaviour 59: 1167–1176.
27. Wu W, Thompson JA, Bertram R, Johnson F (2008) A statistical method for quantifying songbird phonology and syntax. Journal of Neuroscience Methods 174: 147–154.
28. Attias H (1999) Inferring parameters and structure of latent variable models by variational bayes. In: Proceedings of the Proceedings of the Fifteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-99). San FranciscoCA, Morgan Kaufmann. pp 21–30.
29. Beal M (2003) Variational algorithms for approximate Bayesian inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.