




## Distributed representation of discrete sequential vocalization in the Bengalese finch (*Lonchura striata var. domestica*)

Takuya Koumura & Kazuo Okanoya

To cite this article: Takuya Koumura & Kazuo Okanoya (2019): Distributed representation of discrete sequential vocalization in the Bengalese finch (*Lonchura striata var. domestica*), Bioacoustics, DOI: [10.1080/09524622.2019.1607558](https://doi.org/10.1080/09524622.2019.1607558)

To link to this article: <https://doi.org/10.1080/09524622.2019.1607558>

 View supplementary material 

 Published online: 03 May 2019.

 Submit your article to this journal 

 View Crossmark data 



# Distributed representation of discrete sequential vocalization in the Bengalese finch (*Lonchura striata var. domestica*)

Takuya Koumura  and Kazuo Okanoya

Department of Life Sciences, Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan

## ABSTRACT

Bengalese finches learn and produce sequential vocalizations with a complex song syntax. Various models for the song syntax have been proposed, each of which focuses on several different characteristics of the syntax. However, methods to model these multiple characteristics in a single framework have not been well studied. Here, we propose a model that explains three prominent characteristics of the song syntax in Bengalese finches in a single unified framework. First, the generation of a vocal element depends on multiple preceding elements. Second, a song often contains repetitions of a single element type. Third, a song often begins with a special sequence called an introductory sequence. In this study, an effective way was sought to model these three characteristics in the framework of a conditional probability in symbol sequences. The model takes a distributed representation of a preceding sequence as an input, which is defined by a set of decaying values activated when the vocal elements are generated. The proposed model is shown to outperform conventional syntax models in predicting sequences in novel songs. The results suggest that the song syntax of the bird's brain might also be represented by decaying activities of populations of neurons.

## ARTICLE HISTORY

Received 25 January 2019

Accepted 5 April 2019

## KEYWORDS


Sequential vocalization;  
modelling; bird song

## Introduction

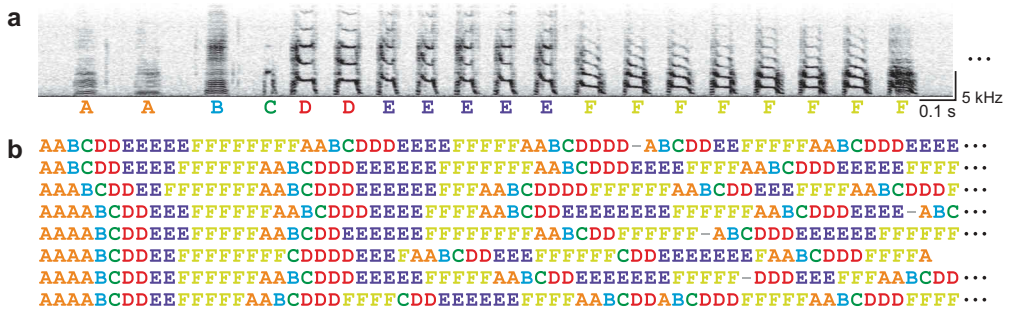
Some animal species, including humans, produce a rapid sequence of vocal elements, which is referred to as a sequential vocalization (Doupe and Kuhl 1999; Kershenbaum et al. 2014). In a sequential vocalization, vocal elements are usually produced with some complex patterns. Finding a good model of such patterns is essential to better understand the functions and mechanisms of sequential vocalization.

Songbirds, especially Bengalese finches, provide powerful behavioural and neural models of sequential vocalization (Okanoya 2004a, 2004b). A song in the Bengalese finch is a sequence of vocal elements called notes (Figure 1(a)), and the patterns of note sequences are called song syntaxes. Birdsong has long been modelled by a probabilistic symbol sequence, in which each symbol corresponds to a single note category (Okanoya 2004b) (Figure 1(b)). A song syntax is modelled by conditional probabilities of symbols

**CONTACT** Kazuo Okanoya  [cokanoya@mail.ecc.u-tokyo.ac.jp](mailto:cokanoya@mail.ecc.u-tokyo.ac.jp)

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/09524622.2019.1607558>.

© 2019 Informa UK Limited, trading as Taylor & Francis Group



**Figure 1.** A song in a Bengalese finch.

(a) An example of an excerpt of a song. Notes at the beginning of the song are shown. The horizontal and the vertical axes show the time and the frequency, respectively. The darker area in the spectrogram indicates a larger power at a particular time and frequency bin. Note categories are shown by the coloured alphabets to the bottom of the spectrogram. A song consists of multiple notes with distinct spectro-temporal patterns. The notes are classified based on the spectro-temporal patterns into a limited number of categories. (b) Examples of note sequences at the beginning of songs. Coloured alphabets indicate the note categories, and the grey hyphens indicate the unclassifiable notes. The dots at the right indicate that the song continues. Although individual sequences differ from one another, most songs contain subsequences ‘AABCDD’ and repetitions of notes ‘E’ and ‘F’. Also, the songs appear to begin with the repetition of ‘A’ followed by ‘BCDD’ followed by the repetition of ‘E’ followed by the repetition of ‘F’.

given the preceding symbol sequences. Previous studies have found several characteristics of the song syntax in Bengalese finches, such as (i) long-range dependencies, (ii) repetitions and (iii) introductory sequences (intros). (i) It is known that the conditional probability of a note category depends on more than one preceding note (Katahira et al. 2011; Yamashita et al. 2011). (ii) A song often contains repetitions of a single note category, with the length of a repetition varying probabilistically (Jin and Kozhevnikov 2011; Wittenbach et al. 2015). (iii) A song usually starts with a sequence with distinct patterns, an intro (Rajan and Doupe 2013). In most birds, an intro is a repetition. Although each of these characteristics has been investigated in multiple independent studies, they have not been modelled very well in a single unified framework.

In this study, we propose a model of song syntax that explains the above three characteristics in a single framework. First, we sought an effective way to model each of the three characteristics in the conventional framework of probabilistic symbol sequences. Two models were made for each of the three characteristics, and the most effective model among them in terms of the likelihood for novel songs was selected. The selected model could explain novel songs effectively, but the framework of a symbolic model itself may not be biologically relevant considering that neural activities that control and represent animal behaviours are not symbolic in nature.

Thus, in the second step of the study, based on the most effective symbolic model, we propose a new model of a song syntax that can predict note categories in novel songs and at the same time is more biologically relevant. The new model is based on conditional probabilities of note categories, as in the symbolic model, but the input is not a symbol sequence but a distributed note representation, that is, a fixed length vector that encodes preceding note sequences. In contrast to the symbolic model, we call the proposed model distributed model. By evaluating the distributed model in terms of the likelihood for a novel song, we demonstrated the behavioural relevance of the distributed model. Because the distributed representation was calculated from

decaying activities of units, each of which encodes a single note category, Bengalese finches might also represent their songs with decaying activities of neural populations.

## Materials and methods

### *Ethics statement*

The experimental procedures were approved by the Institutional Animal Care and Use Committee of the University of Tokyo.

### *Subjects*

Songs in 14 birds were analysed for syntax modelling. All birds were male and older than 120 days post-hatch. Light on and off intervals were 14 h and 10 h, respectively. Food and water were given *ad libitum* before, after, and during song recording.

### *Song recording*

A few of the songs were recorded in the previous study (Koumura and Okanoya 2016). A bird was put in a sound attenuation chamber. After a habituation period of at least two days to the recording environment, sound was recorded for three or four consecutive days using a microphone (PRO 35, Audio-Technica Corporation, Japan), an amplifier (MicTube Duo, Alesis, United States), and an audio interface (OCTA-CAPTURE, Roland, Japan) at 16 bits with a sampling rate of 32 kHz. Light on and off (on for 14 h and off for 10 h, respectively) were controlled with an LED light.

### *Song annotation*

Songs in each bird were individually analysed, because the songs in Bengalese finches are largely different among birds. Notes were detected and classified by supervised machine learning (Koumura and Okanoya 2016). A small number of manually annotated songs were used to train the annotator, which in turn automatically annotated the rest of the songs. All annotations were visually inspected and, if needed, corrected manually. Unclassifiable notes, such as ones that did not appear to belong to any categories or had an intermediate appearance of more than one category, were labelled as 'unclassifiable'. Categories with notes less than 1% of the total number of notes in the songs were labelled as unclassifiable as well.

Note sequences, separated by non-singing calls, with more than seven notes and less than 300-ms silence between notes were extracted as songs. Data in the birds with unclassifiable notes constituting more than 1% of the total number of notes in the songs were discarded. As a result, the data in two birds were discarded.

### *Symbolic syntax model*

The first step of this study was testing an effective way to model symbolic song syntax. The following were tested: (i) whether the length of the dependency is one or multiple

(FO vs. VO models); (ii) whether to model repetitions with a Markov process or with an empirical distribution (MR vs. ER models); (iii) whether to assign different probabilities to intros from non-intros (I vs. NI models). The total number of models to be tested was  $2^3 = 8$  (FO-ER-I, FO-ER-NI, FO-MR-I, FO-MR-NI, VO-ER-I, VO-ER-NI, VO-MR-I and VO-MR-NI). A Markov process was employed as a general framework for the models. In a Markov process, a symbol sequence is modelled by conditional probabilities of symbols depending on the preceding sequence with a finite length. Additionally, in the ER and I models, repetitions and intros were considered as special cases. Here, we describe a brief definition of the models. The detailed algorithms for model construction and hyperparameter setting are in [Appendix 1](#).

The VO-MR-NI model is an ordinary variable-order Markov process. The model is defined by a set of symbols  $S$ , a conditional probability of a symbol  $x \in S$  given a sequence  $X$ ,  $P(x|X)$ , and a set of given sequences  $G$ .  $G$  consists of an empty sequence  $\varepsilon$  (i.e. the sequence of length 0), symbols in  $S$  (i.e. sequences of length 1) and sequences with multiple lengths. The sequences with multiple length  $X$  are included to  $G$  if  $P(x|X)$  is significantly different from  $P(x|X')$  for any  $x$  in  $S$ , where  $X'$  is the longest suffix of  $X$  (e.g. if  $X = ABC$ ,  $X' = BC$ ).  $P(x|X)$  is considered to be different from  $P(x|X')$  if and only if the likelihood ratio  $P(x|X)/P(x|X')$  is larger than a certain threshold (that is, a hyperparameter). In this study, the maximum length of the sequences in  $G$  was 32 for avoiding too large computational costs. If the length of  $X$  is 0,  $P(x|X)$  is defined by the occurrence probability of the note  $x$ ,  $P(x)$ .  $P(x|X)$  is calculated by the occurrence number of the sequence  $Xx$  divided by the number of the sequence  $X$  in all songs.

The FO-MR-NI model is an ordinary first-order Markov process. It is similar to the VO-MR-NI model except that in the FO-MR-NI model  $G$  only contains sequences with lengths of less than 2 (i.e.  $\varepsilon$  and symbols in  $S$ ).

In the I models, the probability of a symbol given intros differs from the probability given non-intros. Specifically, in the VO-MR-I and FO-MR-I models,  $G$  includes  $iX$ , where prefix  $i$  indicates that  $X$  is an intro.  $X$  is defined as an intro if and only if the first note in  $X$  is the first note of the song.

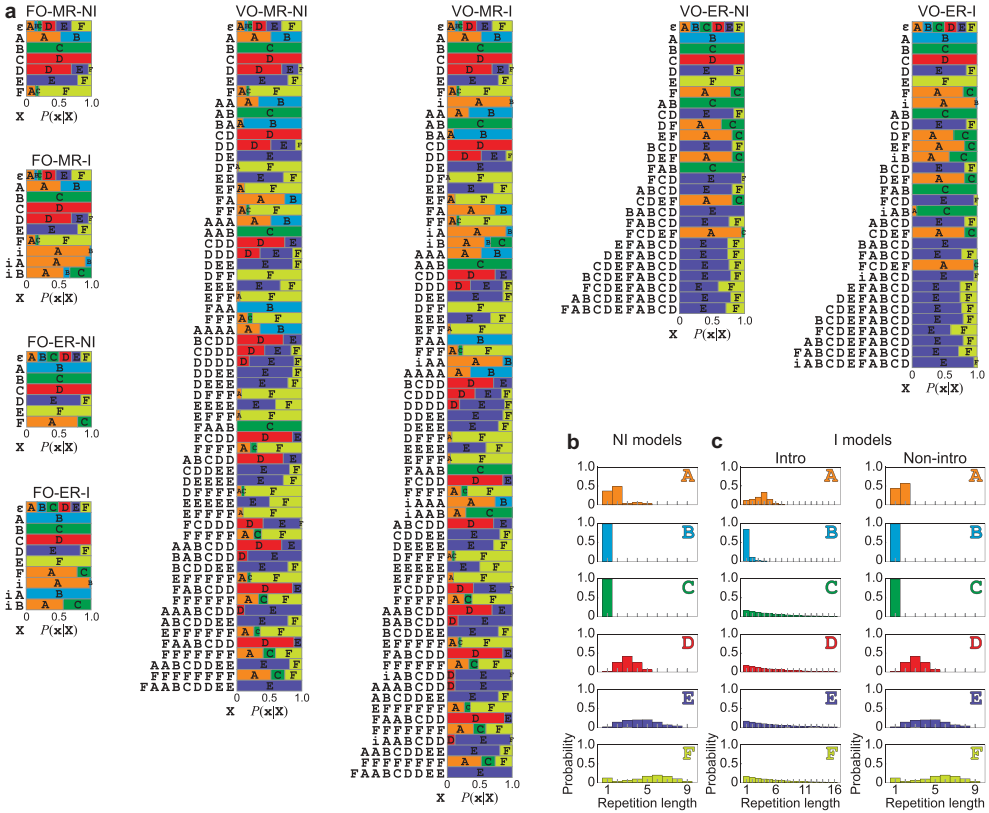
The ER-NI models consist of  $S$ ,  $P(x|X)$ ,  $G$ , and the probability of repetition length for each note category,  $P_x(l)$ , where  $l$  denotes the repetition length.  $P_x(l)$  is calculated by the smoothed empirical distribution of the repetition length. To make repetitions of any length possible, smoothing was done by adding an exponential distribution.

$$P_x(l) = \frac{P'_x(l) + A_x \alpha_x^{l-1}}{1 + \frac{A_x}{1-\alpha_x}}$$

where  $P'_x(l)$  denotes the empirical distribution of the repetition length,  $A_x > 0$  and  $0 < \alpha_x < 1$  are the hyperparameters for smoothing. A symbol without repetition is regarded as  $l = 1$ . In the ER-I models, instead of  $P_x(l)$ , the probability of repetition length of intros  $P_{Ix}(l)$  and that of non-intros  $P_{NIx}(l)$  are calculated. A repetition is considered an intro if and only if the first note in the repetition is the first note of the song.

### **Distributed syntax model**

The distributed syntax model calculates conditional probabilities of note categories from the distributed representation of preceding sequences. The distributed



**Figure 2.** Symbolic syntax models.

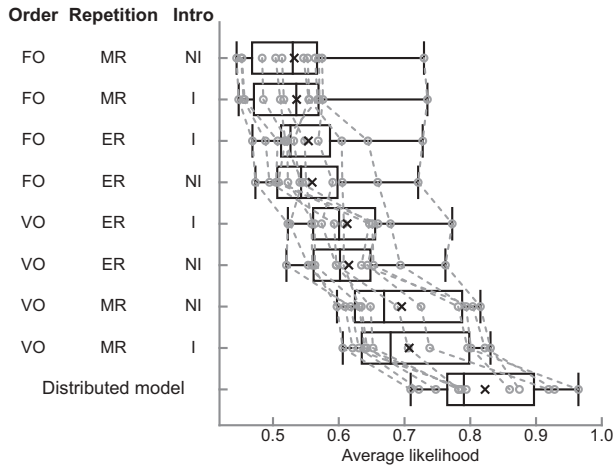
Symbolic syntax models in an example of a bird. (a) probabilities of symbols ‘x’ (horizontal axes) given preceding symbol sequences ‘X’ shown in the left of each panel. The probabilities are shown by the length of the horizontal bars. The name of the condition is shown at the top of each panel. The symbol “i” in an I model indicates an intro, and the symbol “ε” indicates the empty sequence. The conditional probabilities smaller than 0.03 are not shown for better visualization. In the FO models, the length of the given sequences is limited to 1, but not in the VO models. (b) Probabilities of repetition length in the ER-NI models. (c) Probabilities of repetition length in the ER-I models. In the ER-I models, the repetition lengths of intros and non-intros are modelled by different probabilities. Repetition length 1 indicates no repetition. In the ER models, repetitions are reduced to single symbols before applying the conditional probabilities, and the length of the repetition is modelled by an empirical distribution.

representation was defined by a sum of decaying one-hot vectors (with ones at particular note categories) at note onset. Let  $r(t)$  be a distributed representation at time  $t$ , which is a vector of size  $|S|$ , and let each of its elements be  $r_x(t)$ , where  $x$  is a note category.

$$r_x(t) = \sum_{t'_x < t} \left( -\frac{t - t'_x}{\tau_x} \right)$$

where  $t'_x$  is the onset times of the notes with category  $x$ , and  $\tau_x$  is the time constant of the decay for the note category  $x$ . Considering intros, at the beginning of the song  $r_x(0) = 0$  for all  $x$ .

The distributed representation is a given variable of the conditional probability  $y$ . Mapping from  $r(x)$  to the conditional probability is calculated by an artificial neural network  $f$ .



**Figure 3.** Likelihoods for novel songs.

The average likelihood for novel songs in each model type. The average likelihoods in a single bird (grey circles) are connected with grey dashed lines. The box-and-whisker plots show minimums, maximums and quartiles, including all data points. Means among all birds are shown as black crosses. The likelihoods in the VO-MR-I model were larger than in other symbolic models for all birds, and the likelihoods in the distributed model were larger than those of all symbolic models for all birds.

$$y = f(r)$$

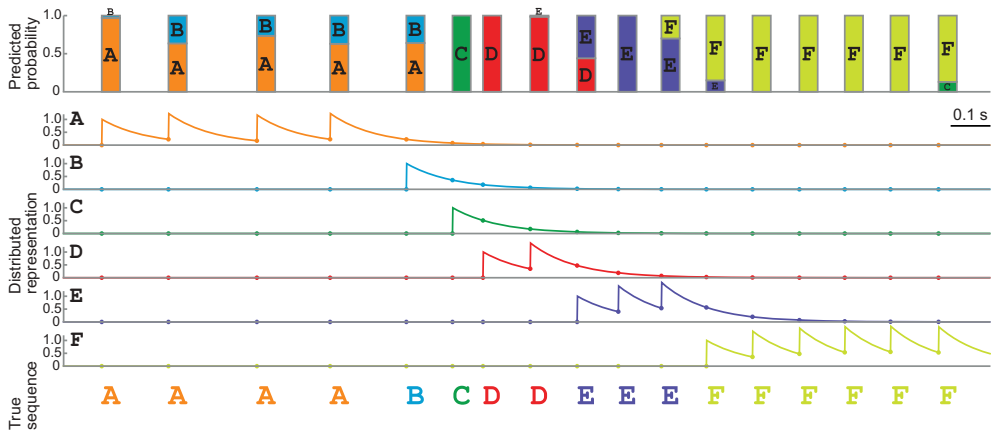
A neural network with a single hidden layer was used with a softmax function as  $f$ . When a conditional probability of a note category was calculated, the current time  $t$  was set to the onset time of the target note. The detailed description of the training procedure for the artificial neural network is in [Appendix 2](#).

### Model evaluation

The eight symbolic models and the distributed model were evaluated by the average likelihood for the validation data (see [Appendix 3](#) for the detailed definition of the average likelihood of each model). Songs were randomly divided into four groups for four-fold cross-validation. Models were trained with songs in the three groups (training data) and evaluated with the remaining group (validation data). This process was repeated for every four combinations of training and validation data. Hyperparameters were optimized to maximize the likelihood in four-fold cross-validation within the training data by dividing the training data further into four groups. It was impossible to conduct leave-one-bird-out cross-validation, because the note categories and the song syntax were largely different among individual birds.

### Statistical test

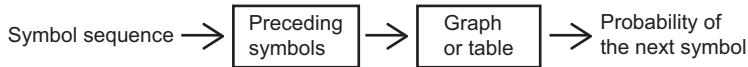
The likelihoods of the models were compared within birds by Wilcoxon signed rank test with Bonferroni correction. The corrected  $p$  values are shown in the results section.



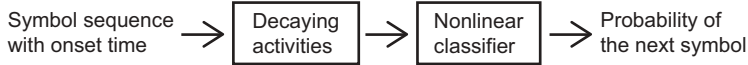
**Figure 4.** Distributed syntax model.

An example of the activities in the distributed syntax model for an excerpt of a song. From top to bottom: predicted probabilities of the next note categories, distributed representation of each note category, and the true note categories. The horizontal axis indicates the time. At the beginning of the song, the values of the representation are 0 for all categories. At the onset of a note, the value for the corresponding unit increases by 1 and decays exponentially. The probabilities of the note categories are calculated using the values obtained just before the note onset, indicated by the dots in the distributed representation.

#### Symbolic model



#### Distributed model



**Figure 5.** Schematic diagrams of the symbolic and distributed models.

A symbolic syntax model takes a symbol sequence as an input and maps it to the probability of the next symbol using a graph or a table of conditional probabilities. In contrast, the distributed model takes a symbol sequence with onset time as an input and represents it with decaying unit activities. The representation is mapped to the probability of the next symbol using a non-linear classifier. Symbolic models have the advantage of being easily and intuitively understood; whereas, the distributed model has the advantage of being directly comparable with neural activities. Both models can be evaluated with observed note sequences using likelihoods, because both of them are generative models of note sequences.

## Results

### *Recorded songs*

Songs in 14 birds were recorded. Songs of each bird were recorded for three or four days. Data in two birds with unclassifiable notes constituting more than 1% of the total notes were discarded. The data in the remaining 12 birds were analysed. The total duration of the songs per day was  $44.5 \pm 22.3$  min (all numbers in this paper with  $\pm$  symbols represent mean  $\pm$  standard deviation). The duration of a song was  $7.5 \pm 2.4$  s. The total number of notes per day and the number of note categories in a bird were  $20,701.8 \pm 10,551.9$  and  $7.8 \pm 3.6$ , respectively. Songs in each bird were analysed independently, because songs were largely different among birds.



### ***Symbolic syntax models***

In the first step of this study, we sought an effective way of symbolically modelling song syntax, using conditional probabilities of symbols given preceding symbol sequences. We focused on the three characteristics of songs in Bengalese finches: (i) long dependency, (ii) repetitions and (iii) intros. The simplest symbolic model using conditional probabilities is a first-order Markov process, in which a probability of a symbol depends only on a preceding symbol. Because of its simplicity, a first-order Markov process has been assumed in many physiological and computational studies, despite that the dependence on only one preceding note has been suggested to be insufficient for songs in Bengalese finches (Katahira et al. 2011; Yamashita et al. 2011). Simply making the dependency longer results in an exponential increase of the total number of sequence patterns the model has to remember, making the model impractical both biologically and computationally. In this study, instead of comparing Markov processes with multiple orders, we compared a first-order Markov process and a variable-order Markov process, in which the length of a given sequence is adaptively determined for each sequence (Ron et al. 1996; Bejerano and Yona 2001; Markowitz et al. 2013). When a note sequence is generated by a variable-order Markov process, each note is sampled from a conditional probability given the preceding symbol sequence. In this process, the conditional probability with the longest given sequence is chosen among those defined in the model. For example, if a model defines conditional probabilities of a symbol given preceding sequences 'A', 'C' and 'BC', the symbol following 'ABC' is sampled from the conditional probability given 'BC' but not that given 'C', because 'BC' is longer than 'C'. When building a model, a conditional probability is defined in the model if the probability is substantially different from that given a shorter sequence. For example, a model defines the conditional probability given 'BC' if it is substantially different from the conditional probability given 'C'. In contrast, the model does not define the conditional probability given 'AC' if it is similar to the conditional probability given 'C'. In this way, a variable-order Markov process efficiently models long dependency by only remembering conditional probabilities that are significantly different from those given shorter sequences. We call these models a first-order (FO) model and a variable-order (VO) model, respectively.

Another important characteristic is a repetition. Some of the previous studies have completely ignored repetitions and replaced them with a single symbol or analysed only the first note in the repetition irrespective of the repetition length (Bouchard and Brainard 2013; Sasahara et al. 2006). In another study, repetitions were modelled with Poisson distributions and non-repetitions with the first-order Markov process, resulting in the model generating similar sequences to those of the actual songs (Kershenbaum et al. 2014). On the other hand, repetitions can be generated in the same neural circuit as non-repetitions if auditory feedback is taken into account, suggesting the same generation mechanism for repetitions and non-repetitions (Wittenbach et al. 2015). In this study, we compared a model that makes a distinction between repetitions and non-repetitions and a model that does not. In the former model, a repetition is reduced to a symbol for the Markov process, and the probability of the repetition length is modelled by an empirical distribution. In the latter model, notes in a repetition are modelled by a Markov process as with the non-repetitions. We call these models an

empirical repetition (ER) and a Markov repetition (MR) models, respectively. In both models, repetitions appearing in intro-like sequences are modelled in the same way as those in non-intro sequences.

The third characteristic we focused on is an intro. Most studies simply exclude intros without detailed description of how they are defined. This may be because it is difficult to determine exactly which note is an intro. In some cases, an intro contains multiple note categories, and, in other cases, an intro-like sequence appears in the middle of a song. Here, we propose simply defining an intro as the sequence at the beginning of the song. In other words, if the first note of a sequence is the first note of the song, the sequence is an intro, and otherwise it is a non-intro. To confirm the effectiveness of this definition, we compared a model that makes a distinction between intros and non-intros, in which conditional probabilities given intros and non-intros are different, and a model that does not. We call the former model an intro (I) model and the latter a no-intro (NI) model.

Each of the three characteristics was modelled in two different ways (FO vs. VO, ER vs. MR and I vs. NI), and combining them resulted in  $2^3 = 8$  models in total (FO-ER-I, FO-ER-NI, FO-MR-I, FO-MR-NI, VO-ER-I, VO-ER-NI, VO-MR-I and VO-MR-NI). Eight models in a bird are shown in [Figure 2](#) as an example. The maximum lengths of the dependency of the VO models were  $13.0 \pm 8.1$  for the VO-ER-I model,  $12.6 \pm 8.0$  for the VO-ER-NI model,  $17.3 \pm 6.0$  for the VO-MR-I model and  $16.9 \pm 5.9$  for the VO-MR-NI model. If the syntax was modelled with a 17th-order Markov process and the number of note categories is 7, for example, the required number of the given sequences would be  $7^{17} \sim 2 \times 10^{14}$ . On the other hand, the numbers of the given sequences (i.e. the numbers of the rows in [Figure 2\(a\)](#)) in the VO models were  $75.7 \pm 62.5$  for the VO-ER-I model,  $69.3 \pm 62.7$  for the VO-ER-NI model,  $184.9 \pm 67.5$  for the VO-MR-I model and  $169.4 \pm 64.8$  for the VO-MR-NI model. This demonstrates the efficiency of the variable-order Markov process. In the I models, the intros (preceded by ‘i’ in [Figure 2\(a\)](#)) were followed by different probabilities from non-intros, suggesting effective modelling of intros. In the MR models, the probability of a note category given the repetition of the category depended on the length of the repetition, indicating the implicit coding of the repetition length by the Markov process (compare  $P(x|D)$ ,  $P(x|DD)$ ,  $P(x|DDD)$  in the VO-MR-NI model, for example). In the ER models, on the other hand, the probabilities of the repetition length are explicitly modelled by the empirical distributions ([Figure 2\(b,c\)](#)).

The eight models were evaluated in terms of a cross-validation likelihood ([Figure 3](#)). The probabilities in each model were estimated from a part of the recorded songs, and the likelihoods for the rest of the songs were calculated. The best model was the VO-MR-I model, (i) the one with the variable order Markov process (ii) with repetitions modelled by the Markov process (iii) that makes a distinction between intros and non-intros. The likelihood of this model was  $0.71 \pm 0.086$  (mean  $\pm$  standard deviation), larger than other models in all birds ( $p = 3.42 \times 10^{-3}$ , Wilcoxon signed-ranked test with Bonferroni correction).

### ***Distributed syntax model***

Model selection demonstrated that generated note categories depended on approximately 10 preceding notes including repetitions, and depended on whether the previous sequences were intros. Although the model effectively captures the characteristics of the song syntax, it may not be natural to assume that the syntax is directly encoded in the brain in the form of a table of conditional probabilities. Also, the number of the given sequences may be too large in analysing experimental data, which are often limited by experimental constraints. Thus, in the second step of this study, we designed a biologically more-relevant syntax model, considering the characteristics of the most effective symbolic model.

Here, we applied an idea developed in the field of natural language processing. Even if there are millions of words with long-range dependencies with variable length, a word is represented with a fixed-length vector (Mikolov et al. 2013a, 2013b). Such a representation is called distributed word representation. In this study, a fixed-length vector was used as the given variable of the conditional probability. Each element of the vector encodes the timing of the preceding notes of the particular category. Specifically, to make the model consistent with the most effective symbolic model, each element of the vector is calculated from the sum of the decaying values activated at note onsets with the particular category (Figure 4). For example, in the song in Figure 4, the note sequences are represented by six time-varying values (i.e. a six-dimensional vector). First, a value corresponding to category A (the time course labelled as A in Figure 4) is activated at the onset of the note. Then, the value gradually decreases until the next arrival of the note with category A (in the case of Figure 4, this happened to be the next note). In this way, we could naturally model (i) long-range dependency and (ii) repetitions. Also, (iii) resetting the values to 0 at the beginning of the song created a distinction between intros and non-intros. Once the given variable is calculated, it is mapped to the conditional probability of the next note categories by a non-linear classifier, such as an artificial neural network (Figure 5). In contrast to the symbolic syntax model using symbolic note representation, we call the new model distributed syntax model using distributed note representation.

Because the output of the model is the conditional probability, the model can be evaluated by a likelihood, as in the conventional symbolic models. The cross-validation likelihood of the distributed model was  $0.82 \pm 0.084$  (mean  $\pm$  standard deviation), larger than the best symbolic model in all birds (Figure 3, bottommost row;  $p = 3.90 \times 10^{-3}$ , Wilcoxon signed-ranked test with Bonferroni correction). This result indicates that the distributed representation combined with a non-linear classifier predicts novel songs better than the conventional model based on symbol sequences. The result demonstrated the behavioural relevance of the distributed model by showing the ability to predict novel songs.

The time constant of the decay was an important parameter that controls the length of the dependency. It was determined by cross-validation within training data. The optimal time constant was  $203.6 \pm 159.7$  ms.

## Discussion

First, to seek an effective way to model song syntax, we evaluated various symbolic models and demonstrated that the length of the dependency should be variable, that repetitions should not be reduced to a single symbol, and that intros should be generated from probabilities different from those of non-intros. Although each of these characteristics has been proposed and demonstrated previously, we believe that testing these characteristics in a single unified framework is an important contribution.

Next, based on these characteristics, we designed a new model that is more biologically relevant and at the same time can generate note sequences. The distributed model predicted the novel songs better than the symbolic models. The result encourages us to expect that distributed representation might also be a good model of the neural representation of song syntax.

### *Effective symbolic model*

The most effective symbolic model had the following characteristics: (i) the probabilities of note categories depend on the variable length of the preceding sequences; (ii) repetitions are modelled by the Markov process as with non-repetitions; (iii) probabilities of note categories given intros and those given non-intros are different.

Although this is the first time that a variable-order Markov process has been applied to song syntax in Bengalese finches, our result is in line with other studies. Long-range dependency of song syntax in Bengalese finches has been indicated in previous studies (Katahira et al. 2011; Yamashita et al. 2011), and a variable-order Markov process has been applied to song syntax in canaries (Markowitz et al. 2013). A variable-order Markov process makes it possible to reduce the number of given sequences by avoiding memorization of the sequence patterns that give similar conditional probabilities with shorter sequences (Ron et al. 1996). Also, the variable-order Markov process does not involve memorization of sequence patterns that do not appear in the songs. This is consistent with the study showing that neurons respond more to frequent patterns than to the patterns that do not appear in the songs (Bouchard and Brainard 2013).

Our result indicates that repetitions are better modelled by Markov processes than by empirical distributions. This is in accordance with the previous computational studies targeting membrane potentials in neural populations (Wittenbach et al. 2015). These studies have shown that repetitions can be generated in the same neural circuit as non-repetitions. Although another study indicated that repetitions should not be modelled by a Markov process (Kershenbaum et al. 2014), this does not contradict the present results, because the other study compared Poisson distributions and a first-order Markov process, which is not suitable for repetitions. That model is similar to the FO-ER model, except that, in the FO-ER model, repetitions were modelled with empirical distributions rather than using Poisson distributions.

A physiological and behavioural study has suggested that intros have a function of preparation for singing, possibly generated with different mechanisms from those of non-intros (Rajan and Doupe 2013). The present result is consistent with this, indicating that intros are better modelled with different probabilities from those of non-intros.

### ***Distributed syntax model***

The proposed distributed syntax model naturally incorporated the characteristics of the best symbolic model. (i) Long-range dependences were modelled by decaying unit activities and the non-linear classifier. The length of the dependency was controlled by the time constant of the decay. (ii) Repetitions and non-repetitions were modelled by the same model framework. (iii) The probabilities given intros were made different from those given non-intros by setting the unit activities to 0 at the beginning of a song.

Conventionally, song-related neural activities are analysed in terms of symbolic representation of note sequences. Spike rates are compared with combinations of two or three note categories (Nishikawa et al. 2008; Fujimoto et al. 2011; Yamashita et al. 2011) or with conditional probabilities with pooled note categories (Bouchard and Brainard 2013). Analysing neural activities for every combination of note categories, including intros and repetitions, is difficult, because the possible combinations increase exponentially with the length of the sequence. The distributed model might be able to provide a good representation for explaining song-related neural activities.

Decaying neural activities in the bird brain have been experimentally demonstrated (Bouchard and Brainard 2016). The model is biologically realizable assuming that there are populations of neurons that encode the occurrence of notes by their decaying activities, which in turn are used for predicting the category of the next note. Approximating biological events by decaying variables has been applied to various phenomena, such as neural activities and biochemical reactions (Honda et al. 2013; Rahman et al. 2018). Our results demonstrate that a similar paradigm is also applicable for explaining animal behaviours.

The design of the distributed model using a fixed-length vector for representing its internal state makes it easy to include other biological components, such as auditory feedback. For example, auditory feedback may be represented with another fixed-length vector and added to the distributed representation of the preceding sequences. Although finding a good way to represent auditory feedback is beyond the scope of this study, it can be modelled in such a way that inclusion of auditory feedback improves the model's performance to predict observed note sequences.

In previous computational studies, a neural circuit for syntax coding has been proposed using the framework of spiking neurons, which explains detailed microscopic mechanisms of syntax coding (Hanuschkin et al. 2011). On the other hand, because the present model is a generative model of note sequences, it can be macroscopically evaluated by the likelihood of actual note sequences. Therefore, these two models do not contradict each other, but they enhance the understanding of song syntax from different points of view.

### ***Time constant***

In the current study, the optimal time constant of the decay was ~200 ms, which approximately corresponds to a duration of several notes. Similar values have been reported in previous studies: 50–200 ms in area X (Kojima and Doupe 2008), 700–1000 ms in HVC (Bouchard and Brainard 2013) and 250 ms in the behavioural experiment

(Okanoya and Dooling 1990). Note that the definitions of the time intervals and the estimation methods are completely different among studies. For instance, 700–1000 ms in Bouchard et al. (2013) (Bouchard and Brainard 2013) is the time for the neural responses to reach their steady state, supposedly larger than the time constant of exponential decay. Nevertheless, it may be interesting that all these studies, including ours, reported a similar order of magnitude.

### ***Limitations and possible applications***

A limitation in this study is that, because the note categories were identified from the sound spectrogram of the whole sequence, the categories could be selected in a sequence-biased manner. For example, the last F in [Figure 1](#) might have been identified as another category if it were not within that repetition of F's. The degree of sequence awareness in category identification can affect the obtained results. If category identification was affected by larger biases of sequence patterns, the resulting note categories can be more easily predicted from the preceding sequences, leading to more stereotyped conditional probabilities and dependence on shorter preceding sequences.

Because sequences in Bengalese finches affect the acoustic structure of a note (Wohlgemuth et al. 2010), it may be better to apply note category identification and syntax estimation iteratively and alternately, instead of fixing the identified categories before estimating a syntax model, as done in this study. Such an application may be an interesting future work.

The proposed model may be applied to quantify changes of sequences within an individual bird. It has been shown that sequence patterns change as a result of brain lesions (Hosino and Okanoya 2000), changes of social contexts (Sakata et al. 2008), external feedback (Warren et al. 2012) and reduced auditory feedback (Okanoya and Yamaguchi 1997). They also change across time as a young adult (Yamashita et al. 2008), in an old individual (James and Sakata 2014), and when singing in a helium atmosphere (Yamada and Okanoya 2003). Such changes may be reflected in the model as changes of the conditional probabilities and/or their given variables (i.e. a set of given sequences in the symbolic model or the decay time constants in the distributed model). For example, sequence patterns become more stereotyped as a bird gets older (Yamashita et al. 2008; James and Sakata 2014) or when a bird sings toward females (Sakata et al. 2008), which may be captured by removing given sequences in the symbolic model, decreasing the decay time constants in the distributed model, and modifying the conditional probabilities in both models. Also, external feedback can modify a conditional probability of note categories given a targeted sequence (Warren et al. 2012). Such modification can be easily reflected in the model by fitting the modelled conditional probabilities to the observed songs.

### ***Data availability***

The data supporting the findings of this study are available within the supplementary materials.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was funded by MEXT/JSPS [#4903-17H06380, #26240019, and #17H01015 to KO] (<https://www.jsps.go.jp/english/index.html>).

## ORCID

Takuya Koumura  <http://orcid.org/0000-0002-8380-9598>

## References

- Bejerano G, Yona G. 2001. Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics*. 17(1):23–43. doi: [10.1093/bioinformatics/bts121](https://doi.org/10.1093/bioinformatics/bts121).
- Bouchard KE, Brainard MS. 2013. Neural encoding and integration of learned probabilistic sequences in avian sensory-motor circuitry. *J Neurosci*. 33(45):17710–17723. doi: [10.1523/JNEUROSCI.2181-13.2013](https://doi.org/10.1523/JNEUROSCI.2181-13.2013).
- Bouchard KE, Brainard MS. 2016. Auditory-induced neural dynamics in sensory-motor circuitry predict learned temporal and sequential statistics of birdsong. *Proc Natl Acad Sci USA*. 113(34):9641–9646. [[accessed 2018 Jul 2]. <http://www.ncbi.nlm.nih.gov/pubmed/27506786>.
- Doupe AJ, Kuhl PK. 1999. BIRDSONG AND HUMAN SPEECH: common themes and mechanisms. *Annu Rev Neurosci*. 22(1):567–631. [accessed 2018 Jan 6]. <http://www.annualreviews.org/doi/10.1146/annurev.neuro.22.1.567>.
- Fujimoto H, Hasegawa T, Watanabe D. 2011. Neural coding of syntactic structure in learned vocalizations in the songbird. *J Neurosci*. 31(27):10023–10033. [[accessed 2013 Mar 6]. <http://www.ncbi.nlm.nih.gov/pubmed/21734294>.
- Hanuschkin A, Diesmann M, Morrison A. 2011. A reafferent and feed-forward model of song syntax generation in the Bengalese finch. *J Comput Neurosci*. 31(3):509–532.
- He K, Zhang X, Ren S, Sun J 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. Washington, DC: IEEE Computer Society. (ICCV '15). p. 1026–1034.
- Honda M, Urakubo H, Koumura T, Kuroda S. 2013. A common framework of signal processing in the induction of cerebellar LTD and cortical STDP. *Neural Netw*. 43:114–124. doi: [10.1016/j.neunet.2013.01.018](https://doi.org/10.1016/j.neunet.2013.01.018).
- Hosino T, Okanoya K. 2000. Lesion of a higher-order song nucleus disrupts phrase level complexity in Bengalese finches. *Neuroreport*. 11(10):2091–2095. doi: [10.1097/00001756-200007140-00007](https://doi.org/10.1097/00001756-200007140-00007).
- James LS, Sakata JT. 2014. Vocal motor changes beyond the sensitive period for song plasticity. *J Neurophysiol*. 112(9):2040–2052. doi: [10.1152/jn.00217.2014](https://doi.org/10.1152/jn.00217.2014).
- Jin DZ, Kozhevnikov AA. 2011. A compact statistical model of the song syntax in Bengalese finch. Friston KJ, editor. *PLoS Comput Biol*. 7(3):e1001108.[accessed 2018 Jan 13]. doi: [10.1371/journal.pcbi.1001108](https://doi.org/10.1371/journal.pcbi.1001108).
- 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. [ACCESSSED 2013 Mar 5]. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3168521&tool=pmcentrez&rendertype=abstract>.

- Kerшенbaum A, Bowles AE, Freeberg TM, Jin DZ, Lameira AR, Bohn K. 2014. Animal vocal sequences: not the markov chains we thought they were. *Proc Biol Sci.* 281:1792. doi: [10.1098/rspb.2014.1370](https://doi.org/10.1098/rspb.2014.1370).
- Kingma D, Ba, J. 2015. Adam: A method for stochastic optimization. In: 3rd International Conference for Learning Representations. San Diego.
- Kojima S, Doupe AJ. 2008. Neural encoding of auditory temporal context in a songbird basal ganglia nucleus, and its independence of birds' song experience. *Eur J Neurosci.* 27(5):1231–1244. doi: [10.1111/j.1460-9568.2008.06083.x](https://doi.org/10.1111/j.1460-9568.2008.06083.x).
- Koumura T, Okanoya K. 2016. Automatic recognition of element classes and boundaries in the birdsong with variable sequences. *PLoS One.* 11(7):e0159188. doi: [10.1371/journal.pone.0159188](https://doi.org/10.1371/journal.pone.0159188).
- Markowitz JE, Ivie E, Kligler L, Gardner TJ. 2013. Long-range order in canary song. *PLoS Comput Biol.* 9(5):e1003052. doi: [10.1371/journal.pcbi.1003052](https://doi.org/10.1371/journal.pcbi.1003052).
- Mikolov T, Chen K, Corrado G, Dean J. 2013a Jan 16. Efficient estimation of word representations in vector space. arXiv. [accessed 2018 Jan 13]. <http://arxiv.org/abs/1301.3781>.
- Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. 2013b Oct 16. Distributed representations of words and phrases and their compositionality. arXiv. [accessed 2018 Jan 13]. <http://arxiv.org/abs/1310.4546>.
- Nishikawa J, Okada M, Okanoya K. 2008. Population coding of song element sequence in the Bengalese finch HVC. *Eur J Neurosci.* 27(12):3273–3283. doi: [10.1111/j.1460-9568.2008.06291.x](https://doi.org/10.1111/j.1460-9568.2008.06291.x).
- Okanoya K. 2004a. The Bengalese finch: a window on the behavioral neurobiology of birdsong syntax. *Ann N Y Acad Sci.* 1016:724–735. [accessed 2013 Mar 6]. <http://www.ncbi.nlm.nih.gov/pubmed/15313802>.
- Okanoya K. 2004b. Song syntax in Bengalese finches: proximate and ultimate analyses. *Adv Study Behav.* 34:297–346.
- Okanoya K, Dooling RJ. 1990. Temporal integration in zebra finches (*Poephila guttata*). *J Acoust Soc Am.* 87(6):2782–2784. [accessed 2018 Jan 13]. <http://asa.scitation.org/doi/10.1121/1.399069>.
- Okanoya K, Yamaguchi A. 1997. Adult Bengalese finches (*lonchura striata* var. *domestica*) require real-time auditory feedback to produce normal song syntax. *J Neurobiol.* 33(4):343–356.
- Rahman M, Willmore DBB, King JA, Harper N. 2018. Exponentially decaying temporal integration in networks can explain much of the dependence of neural responses on stimulus history in primary auditory cortex. In: Proceeding of the 41st Annual MidWinter Meeting, Association for Research in Otolaryngology. San Diego, CA.
- Rajan R, Doupe AJ. 2013. Behavioral and neural signatures of readiness to initiate a learned motor sequence. *Curr Biol.* 23(1):87–93. doi: [10.1016/j.cub.2012.11.040](https://doi.org/10.1016/j.cub.2012.11.040).
- Ron D, Singer Y, Tishby N. 1996. The power of amnesia: learning probabilistic automata with variable memory length. *Mach Learn.* 25(2–3):117–149. doi: [10.1007/BF00114008](https://doi.org/10.1007/BF00114008).
- Sakata JTT, Hampton CM, Brainard MS. 2008. Social modulation of sequence and syllable variability in adult birdsong. *J Neurophysiol.* 99(4):1700–1711. doi: [10.1152/jn.01296.2007](https://doi.org/10.1152/jn.01296.2007).
- Sasahara K, Kakishita Y, Nishino T, Takahasi M, Okanoya K. 2006. A reversible automata approach to modeling birdsongs. In: Computing, 2006. CIC'06. 15th International Conference on. IEEE. Mexico City, Mexico. p. 80–85.
- Warren TL, Charlesworth JD, Tumer EC, Brainard MS. 2012. Variable sequencing is actively maintained in a well learned motor skill. *J Neurosci.* 32(44):15414–15425. doi: [10.1523/JNEUROSCI.1254-12.2012](https://doi.org/10.1523/JNEUROSCI.1254-12.2012).
- Wittenbach JD, Bouchard KE, Brainard MS, Jin DZ. 2015. An adapting auditory-motor feedback loop can contribute to generating vocal repetition. Theunissen FE, editor. *PLoS Comput Biol.* 11(10):e1004471. [accessed 2017 Jul 22]. <http://dx.plos.org/10.1371/journal.pcbi.1004471>.



- Wohlgemuth MJ, Sober SJ, Brainard MS. 2010. Linked control of syllable sequence and phonology in birdsong. *J Neurosci.* 30(39):12936–12949. [accessed 2019 Mar 7]. <http://www.ncbi.nlm.nih.gov/pubmed/20881112>.
- Yamada H, Okanoya K. 2003. Song syntax changes in Bengalese finches singing in a helium atmosphere. *Neuroreport.* 14(13):1725–1729. [accessed 2018 Jan 13]. <http://www.ncbi.nlm.nih.gov/pubmed/14512846>.
- Yamashita Y, Okumura T, Okanoya K, Tani J. 2011. Cooperation of deterministic dynamics and random noise in production of complex syntactical avian song sequences: a neural network model. *Front Comput Neurosci.* 5:18. doi: 10.3389/fncom.2011.00018.
- Yamashita Y, Takahashi M, Okumura T, Ikebuchi M, Yamada H, Suzuki M, Okanoya K, Tani J. 2008. Developmental learning of complex syntactical song in the Bengalese finch: a neural network model. *Neural Netw.* 21(9):1224–1231. [accessed 2019 Mar 7]. <https://www.science-direct.com/science/article/pii/S0893608008000609>.

## Appendices

### Appendix 1. Algorithm for constructing the symbolic models

Inputs: symbol sequences.

Outputs: conditional probabilities of note category  $P(x|X)$ , where  $x$  denotes a note category and  $X$  denotes a note sequence, and the set of given sequences  $G$ . In addition, for the ER-NI models, probability of the repetition length  $P_x(l)$ , where  $l$  denotes the repetition length, or for the ER-I models, probability of the repetition length for intros  $P_{Ix}(l)$  and that for non-intros  $P_{NLx}(l)$ .

Hyperparameters: lower bound of likelihood ratio  $R$ , smoothing constant  $\gamma$ , maximum length  $L$ . In addition, for the ER-NI models, hyperparameters for smoothing the probability of the repetition length:  $A_x$  and  $\alpha_x$ , or for the ER-I models, hyperparameters for smoothing the probability of the repetition length for intros and non-intros:  $A_{Ix}$ ,  $\alpha_{Ix}$ ,  $A_{NLx}$  and  $\alpha_{NLx}$ .

For the ER-NI models, do the following.

Count all repetition length  $C_x(l)$  (e.g. if  $X = \text{AABAB}$ ,  $C_A(1) = 1$ ,  $C_A(2) = 1$ ,  $C_B(1) = 2$ ). Note occurrence without repetition is regarded as  $l = 1$ .

$$P_x(l) = \frac{\frac{C_x(l)}{\sum_{l'} C_x(l')} + A_x \alpha_x^{l-1}}{1 + \frac{A_x}{1 - \alpha_x}}$$

Reduce all the repetitions into single notes (e.g. AABCCD  $\rightarrow$  ABCD).

For the ER-I models, calculate  $P_{Ix}(l)$  and  $P_{NLx}(l)$  for each of the intros and non-intros as for the ER-NI models.

For the I models, attach a symbol ‘i’ at the beginning of each song (e.g. AABCDD  $\rightarrow$  iAABCDD).

Count all combinations of note sequences with length less than or equal to  $L$ , and let the count be denoted by  $C(X)$  (e.g. if  $X = \text{ABABA}$  and  $L = 3$ ,  $C(A) = 3$ ,  $C(B) = 2$ ,  $C(AB) = 2$ ,  $C(BA) = 2$ ,  $C(ABA) = 2$ ,  $C(BAB) = 1$ , and  $C(AA) = C(BB) = C(AAA) = C(AAB) = C(ABB) = C(BAA) = C(BBB) = C(BBA) = 0$ ).

Define  $P(x|X)$  as  $P(x|X) = C(Xx)/C(X)$ , where  $Xx$  is  $X$  followed by  $x$ .

Let the special symbol  $\varepsilon$  be a sequence with length 0. Define  $P(x|\varepsilon)$  as follows:

$$P(x|\varepsilon) = P(x) = \frac{C(x)}{\sum_{x'} C(x')}$$

For each subsequence  $X \in \{X|C(X) > 0\}$ :

If  $\exists x$ ,  $P(x|X)/P(x|X') > R^N$ , add  $X$  to  $G$ , where  $X'$  is a sequence starting from the second symbol of the  $X$ , and  $N$  denotes the number of notes.

For each  $X$  in  $G$ :

For  $i = 2$  to  $L$ , add  $x_2 \dots x_i$  to  $G$  if not contained in  $G$ , where  $x_i$  is the symbol at the position  $i$  in  $X$ .  
Add  $\varepsilon$  to  $G$ .

Replace conditional probabilities  $P(x|X)$  for all  $X$  in  $G$  with the smoothed probability.

$$\text{Smoothed } P(x|X) = \frac{P(x|X) + \gamma}{1 + |S|\gamma},$$

where  $|S|$  is the number of note categories.

Hyperparameters were determined by cross-validation within training data from the following range:  $R \in [1 \times 10^{-6}, 1 \times 10^{-2}]$ ,  $\gamma \in [1 \times 10^{-6}, 1/|S|]$ ,  $A_x \in [1 \times 10^{-3}, 1 \times 10^{-1}]$ ,  $\alpha \in [1.1, 2]$ . For the VO models,  $L = 32$ . For the FO models,  $L = 2$  and  $R = 0$  (i.e.  $G$  contains all sequences with length less than or equals to 2).

## Appendix 2. Training of the classifier of the distributed syntax model

The parameters of the artificial neural network of the distributed syntax model were trained with stochastic gradient descent. Parameters were initialized randomly according to He et al. (2015) (He et al. 2015) and updated for a part of the training data by Adam (Kingma and Ba 2015) until the likelihood stopped increasing for 200 iterations for the other part of the training data.

The values of  $\tau_x$  were assumed to be independent of  $x$  in order to reduce the computational time; thus,  $\tau_x = \tau$  for all  $x$ . The value of  $\tau$  and the number of hidden units were the hyperparameters and determined by cross-validation within training data from the range of [50, 1000] ms for  $\tau$  and [8, 128] for the number of hidden units.

## Appendix 3. Average likelihood

The average likelihood of the MR-NI model is calculated as follows:

$$\text{Average likelihood} = \left( \prod_{k=1}^K \prod_{i=1}^{I_k} P(x_{ki}|X_{ki}) \right)^{\frac{1}{\sum_k I_k}}$$

where  $k$  denotes a song index,  $K$  denotes the number of songs,  $i$  denotes the position of a note in a song,  $I_k$  denotes the number of notes in the song  $k$ , and  $x_{ki}$  denotes the note category at the position  $i$  in the song  $k$ . The likelihood was powered by the inverse of the total number of notes in the data,  $\sum_k I_k$ .  $X_{ki}$  is the longest sequence in  $G$  that ends with  $x_{ki-1}$ . For example, if  $X = \{A, B, C\}$  and  $G = \{\varepsilon, A, B, C, AB\}$ , the average likelihood for a song ABBBC is  $(P(A)P(B|A)P(B|AB)P(B|B)P(C|B))^{1/5}$ .

In the I models, the probabilities given intros are different from the probabilities given non-intros. For example, if  $X = \{A, B, C\}$  and  $G = \{\varepsilon, A, B, C, AB, i, iA\}$ , the average likelihood of VO-MR-I model for a song ABBBC is  $(P_1(A)P_1(B|A)P(B|AB)P(B|B)P(C|B))^{1/5}$ . In this case, the first two probabilities are  $P_1$  because  $G$  contains 'i' and 'iA'.

When calculating the average likelihood of the ER models, all repetitions in the songs were reduced to a single note before applying the conditional probability, which is multiplied by  $P_x(I)$ .

$$\text{Average likelihood} = \left( \prod_{k=1}^K \prod_{i'=1}^{I'_k} P(x_{ki'}|X_{ki'}) P_{x_{ki'}}(I_{ki'}) \right)^{\frac{1}{\sum_k I_k}}$$

where  $I'_k$  and  $i'$  denote the length of and the position at the repetition-reduced song  $k$ . For example, if  $X = \{A, B, C\}$  and  $G = \{\varepsilon, A, B, C, AB\}$ , the average likelihood for a song ABBBC is  $(P(A)P_A(1)P(B|A)P_B(3)P(C|B)P_C(1))^{1/5}$ . The average likelihood of the I-ER models of the above example will be  $(P_1(A)P_{1A}(1)P_1(B|A)P_{NIB}(3)P(C|B)P_{NIC}(1))^{1/5}$ .

The average likelihood of the distributed model is calculated as follows:

$$\text{Average likelihood} = \left( \prod_{k=1}^k \prod_{i=1}^{I_k} y_{ki} \right)^{\frac{1}{\sum_k I_k}}$$

$$y_{ki} = f(r(t_{ki}))$$

where  $t_{ki}$  denotes the onset time of the note at position  $i$  of the song  $k$ .