MEASURING CONTEXT DEPENDENCY IN BIRDSONG USING ARTIFICIAL NEURAL NETWORKS

A PREPRINT

Takashi Morita[1,2] Hiroki Koda[2] Kazuo Okanoya[3] Ryosuke O. Tachibana[3*]

[1]SANKEN, Osaka University, JAPAN

[2]Primate Research Institute, Kyoto University, JAPAN

[3]Center for Evolutionary Cognitive Sciences, Graduate School of Arts and Sciences, the University of Tokyo, JAPAN

November 25, 2021

ABSTRACT

Context dependency is a key feature in sequential structures of human language, which requires reference between words far apart in the produced sequence. Assessing how long the past context has an effect on the current status provides crucial information to understand the mechanism for complex sequential behaviors. Birdsongs serve as a representative model for studying the context dependency in sequential signals produced by non-human animals, while previous reports were upper-bounded by methodological limitations. Here, we newly estimated the context dependency in birdsongs in a more scalable way using a modern neural-network-based language model whose accessible context length is sufficiently long. The detected context dependency was beyond the order of traditional Markovian models of birdsong, but was consistent with previous experimental investigations. We also studied the relation between the assumed/auto-detected vocabulary size of birdsong (i.e., fine- vs. coarse-grained syllable classifications) and the context dependency. It turned out that the larger vocabulary (or the more fine-grained classification) is assumed, the shorter context dependency is detected.

***K*eywords** birdsong, context dependency, Bengalese finch, language modeling, discrete variational autoencoder, unsupervised clustering, individual normalization

Introduction

Making behavioral decisions based on past information is a crucial task in the life of humans and animals [1, 2]. Thus, it is an important inquiry in biology how far past events have an effect on animal behaviors. Such past records are not limited to observations of external environments, but also include behavioral history of oneself. A typical example is human language production; The appropriate choice of words to utter depends on previously uttered words/sentences. For example, we can tell whether '*was*' or '*were*' is the grammatical option after a sentence '*The photographs that were taken in the cafe and sent to Mary* ___' only if we keep track of the previous words sufficiently long, at least up to '*photographs*', and successfully recognize the two closer nouns (*cafe* and *station*) as modifiers rather than the main subject. Similarly, semantically plausible words are selected based on the topic of preceding sentences, as exemplified by the appropriateness of *olive* over *cotton* after "sugar" and "salt" are used in the same speech/document. Such dependence on the production history is called context dependency and is considered a characteristic property of human languages [3, 4, 5, 6].

*Corresponding Author: rtachi@gmail.com

Birdsongs serve as a representative case study of context dependency in sequential signals produced by non-human animals. Their songs are sound sequences that consist of brief vocal elements, or *syllables* [7, 8]. Previous studies have suggested that those birdsongs exhibit non-trivially long dependency on previous outputs [9, 10, 11]. Complex sequential patterns of syllables have been discussed in comparison with human language syntax from the viewpoint of formal linguistics [8, 12]. Neurological studies also revealed homological network structures for the vocal production, recognition, and learning of songbirds and humans [13, 14, 15]. In this line, assessing whether birdsongs exhibit long context dependency is an important instance in the comparative studies, and several previous studies have addressed this inquiry using computational methods [16, 9, 11, 17, 18]. However, the reported lengths of context dependency were often measured using a limited language model (Markov/n-gram model) that was only able to access a few recent syllables in the context. Thus, it is unclear if those numbers were real dependency lengths in the birdsongs or merely model limitations. Moreover, there is accumulating evidence that birdsong sequencing is not precisely modeled by a Markov process [16, 17].

The present study aimed to assess the context dependency in songs of Bengalese finches (*Lonchura striata* var. *domestica*) using modern techniques for natural language processing. Recent advancements in the machine learning field, particularly in artificial neural networks, provide powerful language models [19, 6], which can simulate various time series data without hypothesizing any particular generative process behind them. The neural network-based models also have a capacity to effectively use information in 200–900 syllables from the past (when the data include such long dependency) [5, 6], and thus, the proposed analysis no longer suffers from the model limitations in the previous studies. We performed the context dependency analysis in two steps: unsupervised classification of song syllables and context-dependent modeling of the classified syllable sequence. The classification enabled flexible modeling of statistical ambiguity among upcoming syllables, which are not necessarily similar to one another in acoustics. Moreover, it is preferable to have a common set of syllable categories, which is shared among classifications for all birds, to represent general patterns in the sequences and also to provide the language model with as big data as possible. Conventional classification methods depending on manual labeling by human experts could spoil such generality due to arbitrariness in integrating the category sets across different birds. To satisfy these requirements, we employed a novel, end-to-end, unsupervised clustering method ("seq2seq ABCD-VAE", see Fig. 1). Then, we assessed the context dependency in sequences of the classified syllables by measuring the effective context length [5, 6], which represents how much portion of the song production history impacts on the prediction performance of a language model. The language model we used ("Transformer", see Fig. 5) behaves as a simulator of birdsong production, which exploits the longest context among currently available models [19, 6].

Results

Unsupervised, individual-invariant classification of Bengalese finch syllables

We first converted birdsong syllables into discrete representations, or "labels". When predicting an upcoming syllable from previous outputs, probable candidates can have non-similar acoustic profiles. For example, "bag" and "beg" in English are similar to each other in terms of phonology but have different syntactic and semantic distributions, belonging to different grammatical categories (noun and verb, respectively). An appropriate language model must assign a more similar probability to syntactically/semantically similar words like "bag" and "wallet" than acoustically similar ones like "bag" and "beg". Likewise, it is desirable to perform the context dependency analysis of birdsong based on a flexible model of sequence processing so that it can handle ambiguity about possible upcoming syllables that do not necessarily resemble one another from acoustic perspectives. Categorizing continuous-valued signals and predicting the assigned discrete labels based on a categorical distribution is a simple but effective way of achieving such flexible models, especially when paired with deep neural networks [20, 21, 22]. Syllable classification has also been adopted widely in previous studies of birdsong syntax [7, 23, 11, 18].

Recent studies have explored fully unsupervised classification of animal vocalization based on acoustic features extracted by an artificial neural network, called variational autoencoder or VAE [24, 25, 26]. We extended this approach and proposed a new end-to-end unsupervised clustering method named ABCD-VAE, which utilizes the attention-based categorical sampling with the Dirichlet prior (see also [27]). This method automatically classifies syllables into an unspecified number of categories in a statistically principled way. It also allowed us to exploit the speaker-normalization technique developed for unsupervised learning of human language from speech recordings [28, 29], yielding syllable classification modulo individual variation. Having common syllable categories across different individuals helps us build a unified model of syllable sequence processing. Individual-invariant classification of syllables is also crucial for deep learning-based analysis that requires a substantial amount of data; i.e., it is hard to collect sufficient data for training separate models on each individual.

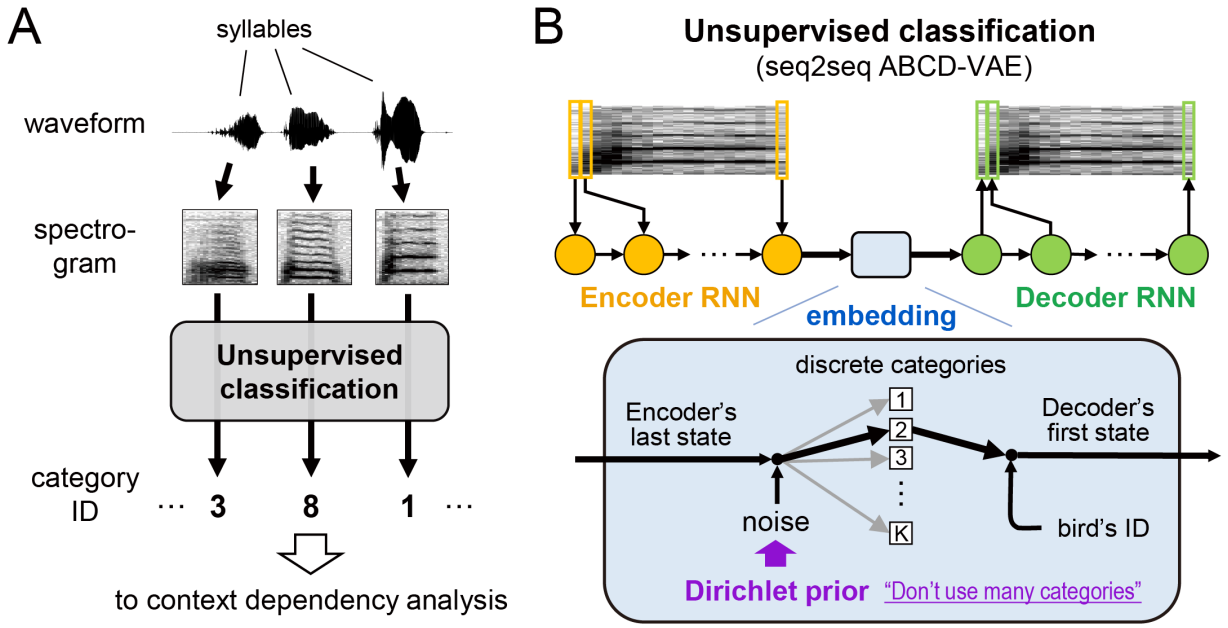

Figure 1. Schematic diagram of newly proposed syllable classification. (A) Each sound waveform segment was converted into the time-frequency representations (spectrograms), and was assigned to one of syllable categories by the unsupervised classification. (B) The unsupervised classification was implemented as a sequence-to-sequence version of the variational autoencoder, consisting of the attention-based categorical sampling with the Dirichlet prior ("seq2seq ABCD-VAE"). The ABCD-VAE encoded syllables into discrete categories between the encoder and the decoder. A statistically optimal number of categories was detected under an arbitrarily specified upper bound thanks to the Dirichlet prior. The identity of the syllable-uttering individual was informed to the decoder besides the syllable categories; Accordingly, individual-specific patterns need not have been encoded in the discrete syllable representation.

We used a dataset of Bengalese finches' songs that was originally recorded for previous studies [30, 31]. Song syllables in the recorded waveform data were detected and segmented by amplitude thresholding. We collected 465,310 syllables in total from 18 adult male birds. Some of these syllables were broken off at the beginning/end of recordings. We filtered out these incomplete syllables, and fed the other 461,994 syllables to the unsupervised classifier (Fig. 1A). The classifier consisted of two concatenated recurrent neural networks (RNNs, see Fig. 1B). We jointly trained the entire network such that the first RNN represented the entirety of each input syllable in its internal state ("encoding" Fig. 1B) and the second RNN restored the original syllable from the internal representation as precisely as possible ("decoding"). The encoded representation of the syllable was mapped to a categorical space ("embedding") before the decoding process. The number of syllable categories was automatically detected owing to the Dirichlet prior [32].

As a result, the classifier detected 37 syllable categories in total for all the birds (Fig. 2B). Syllables that exhibited similar acoustic patterns tended to be classified into the same category across different birds (Fig. 2A). All birds produced not all but a part of syllable categories in their songs (Fig. 2C). The syllable repertoire of each bird covered 24 to 36 categories (32.39±3.35). The detected syllable vocabulary size was greater than the number of annotation labels used by a human expert (5–14; provided in [30]; see also [33] for parallel results based on different clustering methods). Conversely, each category consisted of syllables produced by 7 to 18 birds (15.76±2.91). The detected categories appeared to align with major differences in the spectrotemporal pattern (Fig. 2B).

Quantitative evaluation of syllable classification for Bengalese finch

Speaker-invariant clustering of birdsong syllables should meet at least two desiderata: (i) the resulting classification must keep consistency with the conventional bird-specific classification (i.e., clustered syllables must belong to the same bird-specific class), and (ii) the discovered syllable categories should be anonymized. Regarding (i), we evaluated the alignment of the detected classification with manual annotations by a human expert [30]. We scored the alignment using two metrics. One was Cohen's kappa coefficient [34], which has been used to evaluate syllable classifications

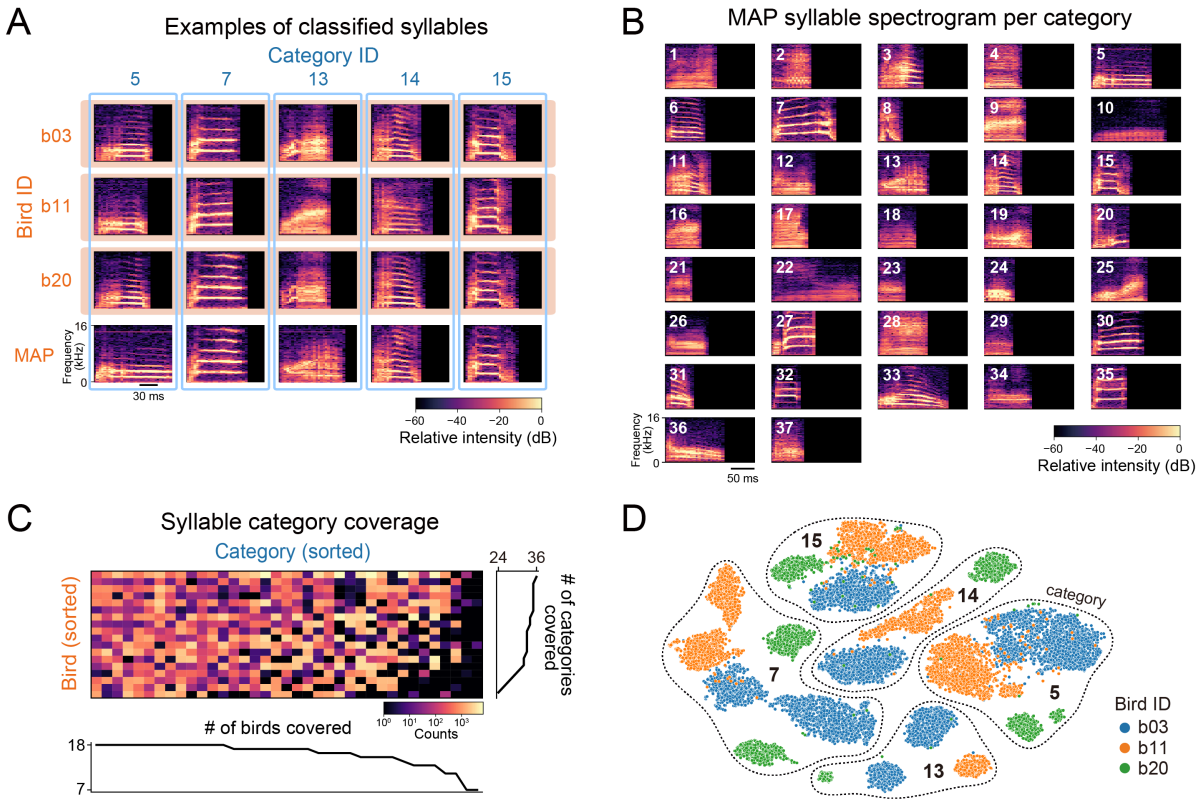

Figure 2. Clustering results of Bengalese finch syllables based on the ABCD-VAE. (A) Syllable spectrograms and their classification across individuals. Syllables in each of the first to third rows (orange box) were sampled from the same individual. Each column (blue frame) corresponds to the syllable categories assigned by the ABCD-VAE. The bottom row provides the spectrogram of each category with the greatest classification probability (MAP: maximum-a-posteriori) over all the individuals. The individual-specific examples also had the greatest classification probability (> 0.999) among the syllables of the same individual and category. (B) Spectrogram of the MAP syllable in each category. (C) Syllable counts per individual bird (rows) and category (columns). The number of non-zero entries is also reported in the line plots. (D) Comparison between syllable embeddings by the canonical continuous-valued VAE with the Gaussian noise (scatter points) and classification by the ABCD-VAE (grouped by the dotted lines). The continuous representation originally had 16 dimensions and was embedded into the 2-dimensional space by t-SNE. The continuous embeddings included notable individual variations represented by colors, whereas the ABCD-VAE classification ignored these individual variations.

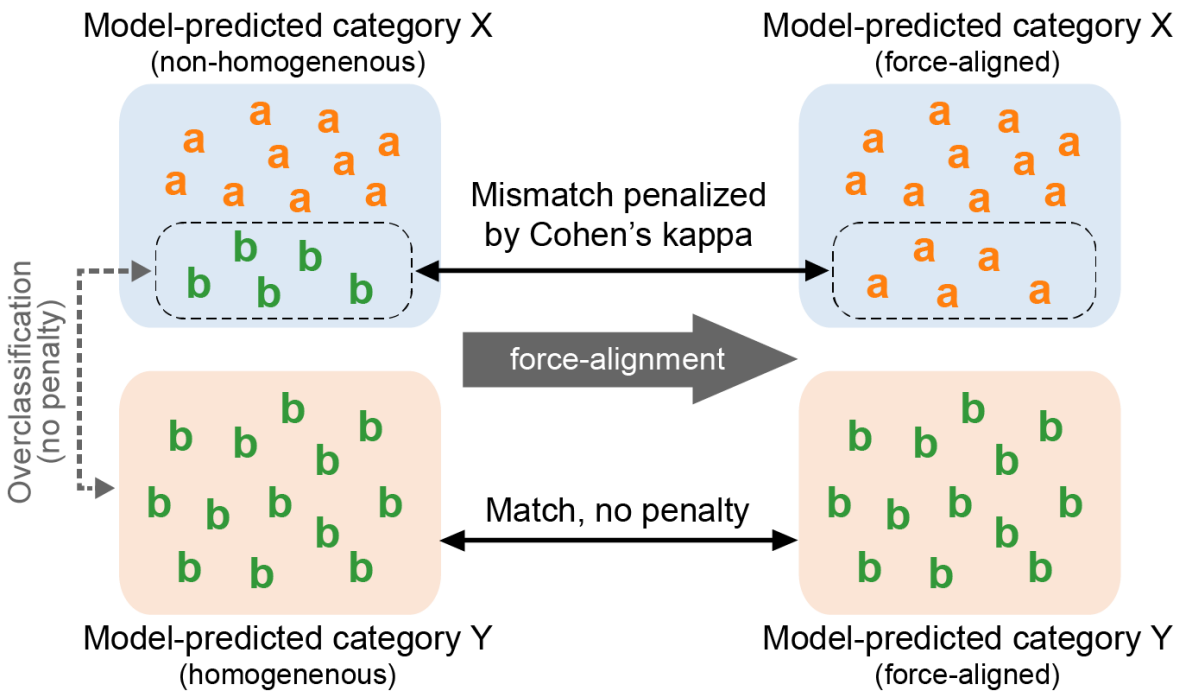

Figure 3. Illustration of clustering evaluation by Cohen's kappa coefficient and homogeneity. The left hand-side is the original clustering results, wherein the letters "a" and "b" represent manual annotation of individual data points, and the rounded rectangles covering the data points (labeled as "X" and "Y") represent model-predicted categories. The right hand-side is the result of force-alignment used for computing Cohen's kappa coefficient, which replaces minority annotation labels with the majority within the model-predicted category. The model-predicted category "X" contains "a" and "b", and thus, is not homogeneous. It is also penalized by Cohen's kappa; "X" is force-aligned with the majority annotation, "a", by relabeling all its data points as "a", and Cohen's kappa penalizes the replacement of the original "b" annotations with "a". On the other hand, the other model-predicted category, Y, is homogeneous because it only contains "b" data. There is no penalty in Cohen's kappa evaluation because the force-alignment does not change the original annotations. Note that "b" data are split into "X" and "Y" but neither Cohen's kappa nor homogeneity penalizes this overclassification.

in previous studies [9, 30]. A problem with this metric is that it requires two classifications to use the same set of categories while our model predictions and human annotations had different numbers of categories and, thus, we needed to force-align each of the model-predicted categories to the most common human-annotated label to use the metric [9]. For example, suppose that the model classified 300 syllables into a category named "X". If 200 of the syllables in "X" are labeled as "a" by the human annotator and the other 100 are labeled as "b", then all the syllables in "X" received "a" as their force-aligned label of model predictions (Fig. 3). This force-alignment makes the 100 syllables misaligned with their original label "b". Thus, the force-alignment scores uniformity of syllables within the model-predicted categories regarding the manual annotations. To get rid of the force-alignment and any other post-processing, we also evaluated the classification using a more recently developed metric called homogeneity [35]. The homogeneity checks whether the category-mate syllables according to the ABCD-VAE were annotated with the same manual label (see the Method for its mathematical definition). Note that homogeneity does not penalize overclassification (see the supporting information S1.5 for additional evaluation that takes overclassification into account). For example, suppose that the ABCD-VAE classified 300 syllables into a category named "X" and another 300 into "Y". The homogeneity is maximized even if all the 300 syllables in "X" are labeled "a" and all the 300 in "Y" are also labeled as "a". This is because all the category-mate syllables receive the same label. Instead, the homogeneity penalizes label mismatches within the model-detected categories, as in the case where 200 of the "X" syllables are labeled "a" and the other 100 are labeled "b" (Fig. 3). Thus, the homogeneity is considered a unified version of Cohen's kappa plus force-alignment.

To assess fulfillment of the second desideratum for ideal clustering (ii), we quantified the speaker-normalization effect of the ABCD-VAE by measuring the perplexity of speaker identification. We built a simple speaker identification model

based on a syllable category uttered by the target bird, fitting the conditional categorical distribution to 90% of all the syllables by the maximum likelihood criterion and then evaluating the prediction probabilities on the other 10%. The prediction probabilities of the test data were averaged in the log scale (= entropy) and then exponentiated to yield the perplexity. Intuitively, the perplexity tells the expected number of birds among whom we have to guess by chance to identify the target speaker even after the information about the syllable category uttered by the target bird is provided. Thus, greater perplexity is an index of successful speaker-normalization.

We compared the performance of the ABCD-VAE with baseline scores provided by the combination of the canonical, continuous-valued VAE (which we call Gauss-VAE) [24, 25, 26] and the Gaussian mixture model (GMM) [36, 32, 37]. This baseline model can be seen as a non-end-to-end version of our clustering method, having distinct optimizations for feature extraction and clustering. The Gauss-VAE was trained on the same datasets and by the same procedure as the ABCD-VAE. On the other hand, the GMM was trained in several ways. First, we built both bird-specific and common models: the former consisted of multiple models, each trained on data collected from a single individual bird, whereas the latter was a single model trained on the entire data collected from all the birds. The bird-specific clusterings provide "topline" scores because the gold-standard annotations by the human expert were also defined in a bird-specific way, and hence, they do not suffer from individual variations included in the Gauss-VAE features. On the other hand, the all-birds-together classifications tell us how much degree of difficulties exist in the clustering without end-to-end optimization or speaker normalization and, thus, serve as a baseline. Another kind of variation in the GMMs we tested was the number of syllable categories. We tested three ways of determining the number: (i) equals to the results from automatic detection by the ABCD-VAE, (ii) equals to the manual annotations by the human expert, and (iii) automatically detected from the distribution of syllable features defined by the Gauss-VAE. (i) and (ii) were obtained by specifying the number of mixture components of the GMM and training the GMM by the maximum likelihood criterion. On the other hand, (iii) was implemented by Bayesian estimation of active mixture components under the Dirichlet distribution prior [32].

As a result, the ABCD-VAE achieved a greater kappa coefficient on average than the baseline models without subject-specific training (Table 1). Moreover, the comparison of the worst-bird scores ("min" in the table) showed that the ABCD-VAE was more robust than the topline models that were optimized to each bird separately. The ABCD-VAE achieved "almost perfect agreements" with the human expert ($\kappa > 0.8$) for sixteen of the eighteen birds and "substantial" agreements ($0.6 < \kappa \geq 0.8$) for the other two [38]. Similarly, the ABCD-VAE outperformed the baseline classifications in the average and worst-bird homogeneity scores. This result was also competitive with the topline models, especially regarding the worst-bird score. These results suggest that the syllable categories discovered by the ABCD-VAE kept consistency with the conventional subject-specific classifications, while the consistency was lost in the other all-birds-together classifications without speaker-normalization. In the meantime, the ABCD-VAE scored the greatest individual perplexity, indicating that the discovered syllable categories were more anonymized and individual-invariant than the baselines (see also Fig. 2D).

Unsupervised classification of zebra finch syllables

To further assess the effectiveness/limitations of the ABCD-VAE, the same clustering was performed on zebra finch syllables (*Taeniopygia guttata*). We collected 237,610 syllables from 20 adult male zebra finches. Again, the data included incomplete syllables that were broken off at the beginning/end of the syllables, and after filtering out those incomplete syllables, we fed the remaining 231,792 to the ABCD-VAE.

Speaker-normalized classification of zebra finch syllables was not as successful (or interpretable) as that of Bengalese finch syllables. While the syllables were classified into 17 categories in total (8 to 14 categories covered by a single bird, mean$\pm$SD:11.2$\pm$1.77), most of the classifications were not confident; 10 out of the 17 detected categories had a low mean classification probability under 30% whereas all but two categories of Bengalese finch syllables had a mean classification probability over 75% (Fig. 4C). Syllables with seemingly major spectral differences were force-aligned across individuals (Fig. 4A). Specifically, syllables consisting of multiple segments with distinct spectral patterns (or notes) seem to lack correspondents in different birds' repertoire (e.g., Category 14 and 16).

Quantitative evaluation also indicates that the speaker-normalized clustering of zebra finch syllables by the ABCD-VAE was not as well-aligned with bird-specific human annotations as that of Bengalese finch (Table 2). While the topline bird-specific models scored about 0.9 of Cohen's kappa coefficient and homogeneity, the scores of the ABCD-VAE stayed around 0.7. Nevertheless, it is of note that the ABCD-VAE outperformed the baseline all-birds-together models, except the one that automatically detected the number of categories (and achieved the upper bound at 128). This auto-detection model achieved high Cohen's kappa and homogeneity by specializing its categories to individual birds (i.e., by resorting to individual-specific classifications); as a result, the model scored a low individual perplexity, indicating that each individual was almost completely identifiable from the model-predicted category of a syllable. By contrast, the ABCD-VAE only used 17 categories and the high individual perplexity indicates that those categories were

Table 1. Quantitative evaluation of the clustering by the ABCD- vs. Gauss-VAE for Bengalese finch syllables. Cohen's kappa coefficient and homogeneity evaluated the alignment of the discovered clusters with manual annotations by a human expert. These scores for each individual bird were computed separately and their mean, maximum, and minimum over the individuals were reported since the manual annotation was not shared across individuals (see Method). Additionally, the perplexity of individual identification scored the amount of individuality included in the syllable categories yielded by the ABCD- and Gauss-VAE. The best scores are in boldface (results under the all-birds-together and bird-specific settings were ranked separately).

Method	# of clusters (source)	Cohen's Kappa mean [min,max]	Homogeneity mean [min,max]	Speaker Perplexity
ABCD-VAE	37	**0.8990** [**0.7740**, **0.9929**]	**0.9084** [**0.7635**, 0.9868]	**8.0434**
Gauss-VAE + GMM (All-Birds-Together)	37 (ABCD-VAE)	0.7446 [0.5956, 0.8912]	0.7844 [0.6004, 0.9086]	4.0783
	14 (manual)	0.6057 [0.4250, 0.8972]	0.6718 [0.5053, 0.8536]	6.7212
	≥128 (auto-detected)	0.8475 [0.5725, 0.9911]	0.8773 [0.6666, **0.9869**]	1.7112
Gauss-VAE + GMM (Bird-Specific)	37 (ABCD-VAE)	0.9304 [0.6619, 0.9906]	0.9292 [0.6479, 0.9893]	—
	5–14 (manual)	0.7888 [0.5012, 0.9328]	0.8090 [0.4732, 0.9254]	—
	50–109 (auto-detected)	**0.9516** [0.7629, 0.9982]	**0.9505** [**0.7687**, **0.9962**]	—

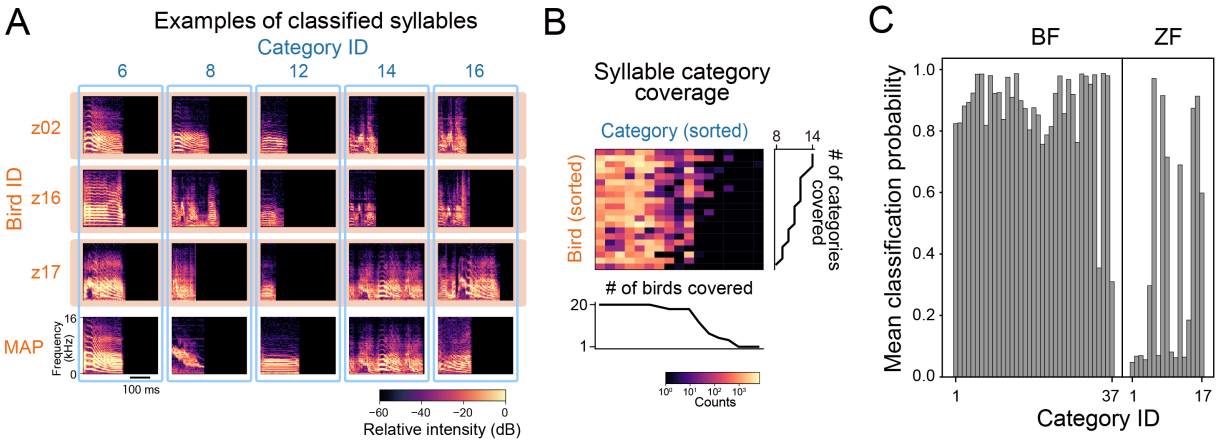

Figure 4. Clustering results of zebra finch syllables based on the ABCD-VAE. (A) Syllable spectrograms and their classification across individuals. Syllables in each of the first to third rows (orange box) were sampled from the same individual. Each column (blue frame) corresponds to the syllable categories assigned by the ABCD-VAE. The bottom row provides the spectrogram of each category with the greatest classification probability (MAP: maximum-a-posteriori) over all the individuals. The individual-specific examples had a top-5 classification probability among the syllables of the same individual and category. (B) Syllable counts per individual bird (rows) and category (columns). The number of non-zero entries is also reported in the line plots. (C) Mean classification probability of Bengalese finch (left) and zebra finch (right) syllables per category.

Table 2. Quantitative evaluation of the clustering by the ABCD- vs. Gauss-VAE for zebra finch syllables. Cohen's kappa coefficient and homogeneity evaluated the alignment of the discovered clusters with manual annotations by a human expert. These scores for each individual bird were computed separately and their mean, maximum, and minimum over the individuals were reported since the manual annotation was not shared across individuals (see Method). Additionally, the perplexity of individual identification scored the amount of individuality included in the syllable categories yielded by the ABCD- and Gauss-VAE. The best scores are in boldface (results under the all-birds-together and bird-specific settings were ranked separately).

Method	# of clusters (source)	Cohen's Kappa mean [min,max]	Homogeneity mean [min,max]	Speaker Perplexity
ABCD-VAE	17	0.7097 [0.4413, 0.9288]	0.6793 [0.4972, 0.8718]	**12.2834**
Gauss-VAE + GMM (All-Birds-Together)	17 (ABCD-VAE)	0.6012 [0.2845, 0.9274]	0.6177 [0.3030, 0.8942]	4.3094
	13 (manual)	0.6102 [0.0401, 0.9741]	0.6315 [0.0433, 0.9609]	5.7021
	$\geq$128 (auto-detected)	**0.8938** [**0.6843**, **0.9915**]	**0.9016** [**0.7643**, **0.9894**]	1.3092
Gauss-VAE + GMM (Bird-Specific)	17 (ABCD-VAE)	0.9579 [0.8847, 0.9938]	0.9545 [0.8828, 0.9905]	—
	4–13 (manual)	0.8762 [0.7915, 0.9744]	0.8623 [0.7056, 0.9607]	—
	18–47 (auto-detected)	**0.9812** [**0.9360**, **1.0000**]	**0.9782** [**0.9274**, **1.0000**]	—

anonymized. Looking at each individual bird, the ABCD-VAE yielded "almost perfect agreement" with the manual annotations ($\kappa > 0.8$) for seven of the twenty birds, "substantial" agreement ($0.6 < \kappa \geq 0.8$) for other seven, and "moderate agreement" for the remaining six ($0.4 < \kappa \geq 0.6$).

Analysis of context dependency

The classification described above provided us sequences of categorically represented syllables. To assess the context dependency in the sequence, we then measured differences between syllables predicted from full-length contexts and truncated contexts. This difference becomes large as the length of the truncated context gets shorter and contains less information. And, the difference should increase if the original sequence has a longer context dependency (Fig. 5A). Thus, the context dependency can be quantified as the minimum length of the truncated contexts where the difference becomes undetectable [5, 6]. For the context-dependent prediction, we employed the Transformer language model [19, 6]. Transformer is known to capture long-distance dependency more easily than RNNs since it can directly refer to any data point in the past at any time while RNNs can only indirectly access past information through their internal memory [39, 19]. There is also accumulating evidence that Transformer successfully represents latent structures behind data, such as hierarchies of human language sentences [19, 40, 41].

Each sequence included syllables from a single recording. In this section, we only report the analysis of Bengalese finch songs; since the classification of zebra finches' syllables was not as reliable as Bengalese finches', we report the analysis of zebra finch song as supplementary results in S3. We obtained a total of 7,879 sequences of Bengalese finch syllables (each containing 8–338 syllables, 59.06 syllables on average), and used 7,779 of them to train the Transformer (see Table 3). The remaining 100 sequences were used to score its predictive performance from which the dependency was calculated. The model predictions were provided of the log conditional probability of the test syllables (x) given the preceding ones in the same sequence. We compared the model predictions between the full-context ("Full", Fig. 5A) and the truncated-context ("Truncated") conditions. Then, the context dependency was quantified by a statistical measure of the effective context length [5, 6], which is the minimum length of the truncated context wherein the mean prediction difference between the two contexts was not significantly greater than the canonical 1% threshold in perplexity [42].

To see the relation between the number of syllable categories and context dependency, we also performed the same analysis based on more coarse/fine-grained syllable classifications into 10 to 80, 160, and 320 categories. These

Table 3. The size of the training and test data used in the neural language modeling of Bengalese finch songs. The "SECL" portion of the test syllables was used to estimate the SECL. The numbers of syllables in parentheses report the incomplete syllables that were broken off at the start/end of recordings, which were labeled with a distinct symbol.

Species	Usage	# of sequences	# of syllables	
			Total	SECL
Bengalese Finch	Training (incomplete)	7,779	458,992 (3,275)	—
	Test (incomplete)	100	6,557 (41)	4,657 (36)
Zebra Finch	Training (incomplete)	11,722	234,674 (5,763)	—
	Test (incomplete)	100	2,936 (55)	1,536 (49)

classifications were derived from the k-means clustering on the L2-normalized feature vectors of syllables given by the ABCD-VAE.

The statistically effective context length (SECL) of the Bengalese finch song was eight based on the 37 syllable categories that were automatically detected by the ABCD-VAE (Fig. 5B). In other words, restricting available contexts to seven or fewer preceding syllables significantly decreased the prediction accuracy compared with the full-context baseline, while the difference became marginal when eight or more syllables were included in the truncated context.

When syllables were classified into more fine-grained categories, the difference between the model predictions based on the truncated and full contexts became smaller (Fig. 5B; $p < 0.001$ according to the linear regression of the loss difference on the number of syllable categories and the length of truncated contexts, both in the log scale). That is, the context dependency traded off with the number of syllable categories. When 160 or 320 categories were assumed, the SECL of the Bengalese finch songs decreased to 5.

Discussion

This study assessed the context dependency in Bengalese finch's song to investigate how long individual birds must remember their previous vocal outputs to generate well-formed songs. We addressed this question by fitting a state-of-the-art language model, Transformer, to the syllable sequences, and evaluating the decline in the model's performance upon truncation of the context. We also proposed an end-to-end clustering method of Bengalese finch syllables, the ABCD-VAE, to obtain discrete inputs for the language model. In the section below, we discuss the results of this syllable clustering and then move to consider context dependency.

Clustering of syllables

The clustering of syllables into discrete categories played an essential role in our analysis of context dependency in Bengalese finch songs. Various studies have observed how fundamental the classification of voice elements is to animal vocalization [43, 44, 7, 11, 45, 18].

Our syllable clustering is based on the ABCD-VAE [27] and features the following advantages over previous approaches. First, the ABCD-VAE works in a completely unsupervised fashion. The system finds a classification of syllables from scratch instead of generalizing manual labeling of syllables by human annotators [30]. Thus, the obtained results are more objective and reproducible [46]. Second, the ABCD-VAE automatically detects the number of syllable categories in a statistically grounded way (following the Bayesian optimality under the Dirichlet prior) rather than pushing syllables into a pre-specified number of classes [47, 28, 29]. This update is of particular importance when we know little about the ground truth classification—as in the cases of animal song studies—and need a more non-parametric analysis. Third, the ABCD-VAE adopted the speaker-normalization technique used for human speech analysis and finds individual-invariant categories of syllables [28, 29]. Finally, the end-to-end clustering by the ABCD-VAE is more statistically principled than the previous two-step approach—acoustic feature extraction followed by clustering—because the distinct feature extractors are not optimized for clustering and the clustering algorithms are often blind to the optimization objective of the feature extractors [25, 26]. We consider that such a mismatch led the combination of Gauss-VAE and GMM to

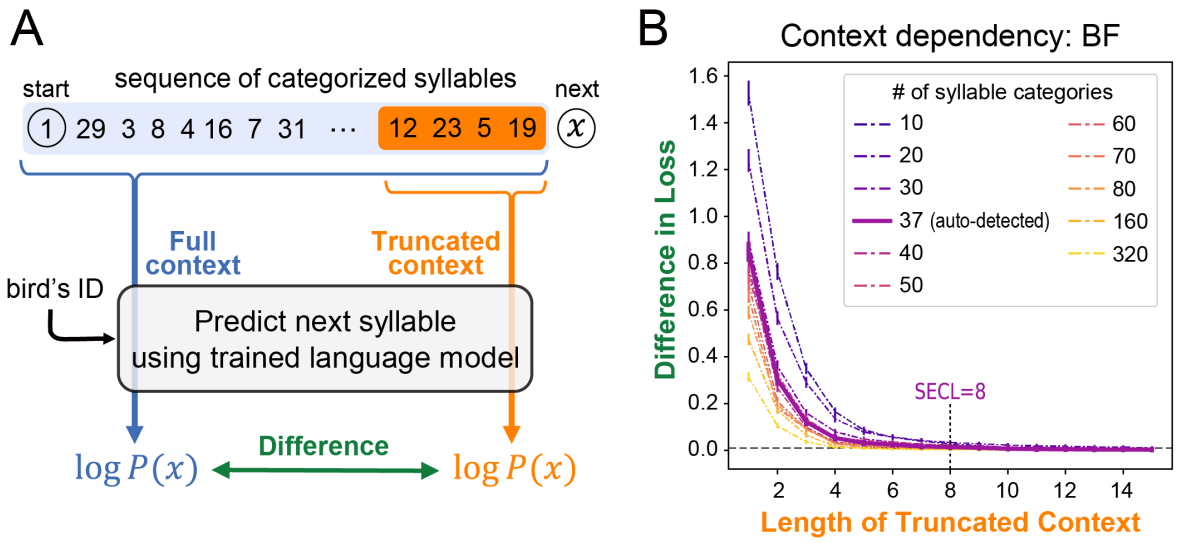

Figure 5. (A) Schematic diagram of the evaluation metric. Predictive probability of each categorized syllable (denoted by x) was computed using the trained language model, conditioned on the full and truncated contexts consisting of preceding syllables (highlighted in blue and orange, respectively). The logarithmic difference of the two predictive probabilities was evaluated, and SECL was defined by the minimum length of the truncated context wherein the prediction difference is not statistically significantly greater than a canonical threshold. (B) The differences in the mean loss (negative log probability) between the truncated- and full-context predictions of Bengalese finch songs. The x-axis corresponds to the length of the truncated context. The error bars show the 90% confidence intervals estimated from 10,000 bootstrapped samples. The loss difference is statistically significant if the lower side of the intervals are above the threshold indicated by the horizontal dashed line.

detect greater numbers of syllable categories than the ABCD-VAE and manual annotations, even when the clustering was specialized for each individual bird and not disturbed by individual variations (see Table 1). Chorowski et al. [29] also showed that a similar end-to-end clustering is better at finding speaker-invariant categories in human speech than the two-step approach.

We acknowledge that discrete representation of data is not the only way of removing individual variations; previous studies have also explored individual normalization on continuous-valued features using deep neural networks. Variational fair autoencoders (VFAE), for example, use speaker embeddings as background information of VAE (in both the encoder and decoder while the ABCD-VAE only fed the speaker information to the decoder) [48]. As the authors note, however, the use of background information does not completely remove individual variations in the extracted features because continuous-valued features can distinguish infinitely many patterns (in principle) and do not have a strong bottleneck effect like discrete categories, making V(F)AE lose motivation to remove individual variations from the features (see also our supporting information S1.4). Accordingly, VFAE has another learning objective that minimizes distances between feature vectors averaged within each speaker. More recently, researchers started to use adversarial training to remove individual and other undesirable variations [49]. In adversarial training, an additional classifier module is installed in the model, and that classifier attempts to *identify* the individual from the corresponding feature representation. The rest of the model is trained to *deceive* the individual classifier into misclassification by anonymizing the encoded features. Both VFAE and adversarial training are compatible with the ABCD-VAE and future studies may combine these methods to achieve stronger speaker-normalization effects. Note, however, that those normalization techniques would not yield speaker-invariant categories if there are no such categories; different individuals may exhibit completely different syllable repertoires and force alignment across individuals can be inappropriate in such cases. Specifically, we suspect that simply adopting other normalization methods would not lead to a more reliable classification of zebra finch syllables modulo speaker variations, unless we find more appropriate segmentation.

It should be noted that the classical manual classification of animal voice was often based on *visual* inspection on the waveforms and/or spectrograms rather than auditory inspection [43, 9, 30]. Similarly, previous VAE analyses of animal voice often used a convolutional neural network that processed spectrograms as images of a fixed size [25, 26]. By

contrast, the present study adopted a RNN [50] to process syllable spectra frame by frame as time series data. Owing to the lack of ground truth as well as empirical limitations on experimental validation, it is difficult to adjudicate on the best neural network architecture for auto-encoding Bengalese finch syllables and other animals' voice. Nevertheless, RNN deserves close attention as a neural/cognitive model of vocal learning. There is a version of RNN called *reservoir computer* that has been developed to model computations in cortical microcircuits [51, 52]. Future studies may replace the LSTM in the ABCD-VAE with a reservoir computer to build a more biologically plausible model of vocal learning [53]. Similarly, we may filter some frequency bands in the input sound spectra to simulate the auditory perception of the target animal [29], and/or adopt more anatomically/bio-acoustically realistic articulatory systems for the decoder module [54]. Such Embodied VAEs would allow constructive investigation of vocal learning beyond mere acoustic analysis.

A visual inspection of classification results shows that the ABCD-VAE can discover individual-invariant categories of the Bengalese finch syllables (Fig. 2), which was also supported by their alignment with human annotations and low individuality in the classified syllables (Table 1). This speaker-normalization effect is remarkable because the syllables exhibit notable individual variations in the continuous feature space mapped into by the canonical VAE and cross-individual clustering is difficult there [25, 26, 55]. Previous studies on Bengalese finch and other songbirds often assigned distinct sets of categories to syllables of different individuals, presumably because of similar individual variations in the feature space they adopted [9, 11, 30, 45].

By contrast, speaker-normalized clustering of zebra finch syllables was less successful, as evidenced by the lower classification probability (Fig. 4B) and consistency with speaker-specific manual annotations (Table 2) than that of Bengalese finch syllables. A visual inspection of category-mate syllables across individuals suggests that one major challenge for finding individual-invariant categories is the complex syllables that exhibit multiple elements, or 'notes', without clear silent intervals (gaps; Fig. 4A). Such complex syllables may be better analyzed by segmenting them into smaller vocal units [56, 57, 58, 12, 59], and the prerequisite for appropriate voice segmentation is a major limitation of the proposed method because the unclarity of segment boundaries in low-level acoustic spaces is a common problem in analyses of vocalization, especially of mammals' vocalization [45], including human speech [60, 61]. A possible solution to this problem (in accordance with our end-to-end clustering) is to categorize sounds frame by frame (e.g., by spectrum and MFCCs) and merge contiguous classmate frames to define a syllable-like span [29, 27, 62, 63].

Context dependency

According to our analysis of context dependency, Bengalese finches are expected to keep track of up to eight previously uttered syllables—not just one or two—during their singing. This is evidenced by the relatively poor performance of the song simulator conditioned on the truncated context of one to seven syllables compared to the full-context condition. Our findings add a new piece of evidence for long context dependency in Bengalese finch songs found in previous studies. Katahira et al. [9] showed that the dependent context length was at least two. They compared the first order and second order Markov models, which can only access the one and two preceding syllable(s), respectively, and found significant differences between them. A similar analysis was performed on canary songs (note, however, that chunks of homogeneous syllable repeats, called phrases, were used as song units rather than individual syllables) by Markowitz et al. [11], with an extended Markovian order (up to seventh). The framework in these studies cannot scale up to assess longer context dependency owing to the empirical difficulty of training higher-order Markov models [64, 65]. By contrast, the present study exploited a state-of-the-art neural language model (Transformer) that can effectively combine information from much longer contexts than previous Markovian models and potentially refer up to 900 tokens [6]. Thus, the dependency length reported in this study is less likely to be upper-bounded by the model limitations and provides a more precise estimation (or at least a tighter lower-bound) of the real dependency length in a birdsong than previous studies.

The long context dependency on eight previous syllables in Bengalese finch songs is also evidenced by experimental studies. Bouchard and Brainard [66] found that activities of Bengalese finches' HVC neurons in response to listening to a syllable x_t encoded the probability of the preceding syllable sequence $x_{t-L}, \ldots, x_{t-1}$ (i.e., context) given x_t, or $p(x_{t-L}, \ldots, x_{t-1} \mid x_t)$. They reported that the length L of the context encoded by HVC neurons (that exhibited strong activities to the bird's own song) reached 7–10 syllables, which is consistent with the dependency length of eight syllables estimated in the present study. Warren et al. [10] also provided evidence for long context dependency from a behavioral experiment. They reported that several pairs of syllable categories of Bengalese finch songs had different transitional probabilities depending on whether or not the same transition pattern occurred in the previous opportunity. In other words, $\mathbb{P}(B \mid AB \ldots A_) \neq \mathbb{P}(B \mid AC \ldots A_)$ where A, B, C are distinct syllable categories, the dots represent intervening syllables of an arbitrary length ($\not\ni A$), and the underline indicates the position of B whose probability is measured. Moreover, they found that the probability of such history-dependent transition patterns is harder to modify through reinforcement learning than that of more locally dependent transitions. These results are

consistent with our findings. It often takes more than two transitions for syllables to recur (12.24 syllables on average with the SD of 11.02 according to our own Bengalese finch data, excluding consecutive repetitions); therefore, the dependency on the previous occurrence cannot be captured by memorizing just one or two previously uttered syllable(s).

There is also a previous study that suggests a longer context dependency in Bengalese finch songs than estimated in this study (i.e., $\gg$8). Sainburg et al. [18] studied the mutual information between birdsong syllables—including Bengalese finch ones—appearing at each discrete distance. They analyzed patterns in the decay of mutual information to diagnose the generative model behind the birdsong data, and reported that birdsongs were best modeled by a combination of a hierarchical model that is often adopted for human language sentences and a Markov process: subsequences of the songs were generated from a Markov process and those subsequences were structured into a hierarchy. Mutual information decayed exponentially in the local Markov domain, but the decay slowed down and followed the power-law as the inter-syllable distance became large. Sainburg et al. estimated that this switch in the decay pattern occurred when the inter-syllable distance was around 24 syllables. This estimated length was substantially longer than our estimated context dependency on eight syllables. The difference between the two results might be attributed to several factors. First, the long-distance mutual information may not be useful for the specific task of predicting upcoming syllables that defined the context dependency here and in the previous studies based on language modeling. It is possible that all the information necessary for the task is available locally while the mutual information does not asymptote in the local domain (see S4 for concrete examples). Another possible factor responsible for the longer context dependency detected by Sainburg et al. is that their primary analysis was based on long-sequence data concatenating syllables recorded in a single day (amounting to 2,693–34,588 syllables, 11,985.56 on average, manually annotated with 16–26 labels per individual). Importantly, they also showed that the bimodality of mutual information decay in the Bengalese finch song became less clear when the analysis was performed on bouts (consisting of 8–398 syllables, 80.98 on average). Since our data was more akin to the latter, potential long dependency in the hierarchical domain might be too weak to be detected in the language modeling-based analysis.

We also found that the greater number of syllable categories is assumed, the shorter context length becomes sufficient to predict upcoming syllables. We attribute this result to the minor acoustic variations among syllables that are ignored as a noise in the standard clustering or manual classification but encoded in the fine-grained classifications. When predicting upcoming syllables based on the fine-grained categories, the model has to identify the minor acoustic variations encoded by the categories. And it has been reported that category-mate syllables (defined by manual annotations) exhibited systematically different acoustic variations depending on the type of surrounding syllables [67]. Thus, the identification of fine-grained categories would improve by referring to the local context, rather than syllables far apart from the prediction target. This increases the importance of the local context compared to predictions of more coarse-grained categories.

The reported context dependency on previous syllables also has an implication for possible models of birdsong syntax. Feasible models should be able to represent the long context efficiently. For example, the simplest and traditional model of the birdsong and voice sequences of other animals—including human language before the deep learning era—is the n-gram model, which exhaustively represents all the possible contexts of length $n-1$ as distinct conditions [64, 65, 7]. This approach, however, requires an exponential number of contexts to be represented in the model. In the worst case, the number of possible contexts in Bengalese finch songs is $37^8 = 3,512,479,453,921$ when there are 37 syllable types and the context length is eight as detected in this study. While the effective context length can be shortened if birds had a larger vocabulary size, the number of logically possible contexts remains huge (e.g., $160^5 = 104,857,600,000$). Such an exhaustive representation is not only hard to store and learn—for both real birds and simulators—but also uninterpretable to researchers. Thus, a more efficient representation of the context syllables is required [68]. Katahira et al. [9] assert that the song syntax of the Bengalese finch can be better described with a lower-order hidden Markov model [69] than the n-gram model. Moreover, hierarchical language models used in computational linguistics (e.g., probabilistic context-free grammar) are known to allow a more compact description of human language [70] and animal voice sequences [71] than sequential models like HMM. Another compression possibility is to represent consecutive repetitions of the same syllable categories differently from transitions between heterogeneous syllables [16, 17] (see also [72] for neurological evidence for different treatments of heterosyllabic transitions and homosyllabic repetitions). This idea is essentially equivalent to the run length encoding of digital signals (e.g., `AAABBCDDEEEEE` can be represented as `3A2B1C2D5E` where the numbers count the repetitions of the following letter) and is effective for data including many repetitions like Bengalese finch's song. For the actual implementation in birds' brains, the long contexts can be represented in a distributed way [73]: Activation patterns of neuronal ensemble can encode a larger amount of information than the simple sum of information representable by individual neurons, as demonstrated by the achievements of artificial neural networks [51, 52, 74].

We conclude the present paper by noting that the analysis of context dependency via neural language modeling is not limited to Bengalese/zebra finch's song. Since neural networks are universal approximators and potentially fit to any kind of data [75, 76], the same analytical method is applicable to other animals' voice sequences [43, 11, 71], given

reasonable segmentation and classification of sequence components like syllables. Moreover, the analysis of context dependency can also be performed in principle on other sequential behavioral data besides vocalization, including dance [77, 78] and gestures [79, 80]. Hence, our method provides a crossmodal research paradigm for inquiry into the effect of past behavioral records on future decision making.

Materials and methods

Recording and preprocessing

We used the same recordings of Bengalese finch songs that were originally reported in our earlier studies [30, 31]. The data were collected from 18 Bengalese finches, each isolated in a birdcage placed inside a soundproof chamber. All the birds were adult males (>140 days after hatching). All but two birds were obtained from commercial breeders, and the other two birds (bird ID: b10 and b20) were raised in laboratory cages. Note that one bird (b20) was a son of another (b03), and learned its song from the father bird. No other birds had any explicit family relationship. The microphone (Audio-Technica PRO35) was installed above the birdcages. The output of the microphone was amplified using a mixer (Mackie 402-VLZ3) and digitized through an audio interface (Roland UA-1010/UA-55) at 16-bits with a sampling rate of 44.1 kHz. The recordings were then down-sampled to 32 kHz [30, 31]. Recording process was automatically started upon detection of vocalization and terminated when no voice was detected for 500–1000 msec (the threshold was adjusted for individual birds). Thus, the resulting recordings roughly corresponded to bout-level sequences, and we used them as the sequence unit for the analysis of context dependency.

An additional dataset for song recordings of 20 zebra finches was kindly provided by Prof. Kazuhiro Wada (Hokkaido University). The recording was performed in the same procedure as previously reported [81, 82].

Song syllables were segmented from the continuous recordings using the thresholding algorithm proposed in the previous studies [30, 31]. The original waveforms were first bandpass-filtered at 1–8 kHz. Then, we obtained their amplitude envelope via full-wave rectification and lowpass-filtered it at 200 Hz. Syllable onsets and offsets were detected by thresholding this amplitude envelope at a predefined level, which was set at 6–10 SD above the mean of the background noise level (the exact coefficient of the SD was adjusted for individual birds). The mean and SD of background noise were estimated from the sound level histogram . Sound segments detected from this thresholding algorithm were sometimes too close to their neighbors (typically separated by a <5 msec interval), and such coalescent segments were reidentified as a single syllable, by lower-bounding possible inter-syllable gaps at 3–13 msec for Bengalese finches and 3–10 msec for zebra finches (both adjusted for individual birds). Finally, extremely short sound segments were discarded as noise, by setting a lower bound on possible syllable durations at 10–30 ms for Bengalese finches and 5–30 msec for zebra finches (adjusted for individual birds). These segmentation processes yielded 465,310 Bengalese finch syllables (≈ 10.79 hours) and 237,610 zebra finch syllables ($\approx$7.72 hours) in total.

Clustering of syllables

To perform an analysis parallel to the discrete human language data, we classified the segmented syllables into discrete categories in an unsupervised way. Specifically, we used an end-to-end clustering method, named the seq2seq ABCD-VAE, that combined (i) neural network-based extraction of syllable features and (ii) Bayesian classification, both of which worked in an unsupervised way (i.e., without top-down selection of acoustic features or manual classification of the syllables). This section provides an overview of our method, with a brief, high-level introduction to the two components. Interested readers are referred to S1 in the supporting information, where we provide more detailed information. One of the challenges to clustering syllables is their variable duration as many of the existing clustering methods require their input to be a fixed-dimensional vector. Thus, it is convenient to represent the syllables in such a format [83, 84]. Previous studies on animal vocalization often used acoustic features like syllable duration, mean pitch, spectral entropy/shape (centroid, skewness, etc.), mean spectrum/cepstrum, and/or Mel-frequency cepstral coefficients at some representative points for the fixed-dimensional representation [9, 30, 71]. In this study, we took a non-parametric approach based on a sequence-to-sequence (seq2seq) autoencoder [85]. The seq2seq autoencoder is a RNN that first reads the whole spectral sequence of an input syllable frame by frame (*encoding*; the spectral sequence was obtained by the short-term Fourier transform with the 8 msec Hanning window and 4 msec stride), and then reconstructs the input spectra (*decoding*; see the schematic diagram of the system provided in the upper half of Fig. 1B). Improving the precision of this reconstruction is the training objective of the seq2seq autoencoder. For successful reconstruction, the RNN must store the information about the entire syllable in its internal state—represented by a fixed-dimensional vector—when it transitions from the encoding phase to the decoding phase. And this internal state of the RNN served

as the fixed-dimensional representation of the syllables. We implemented the encoder and decoder RNNs by the LSTM [50].

One problem with the auto-encoded features of the syllables is that the encoder does not guarantee their interpretability. The only thing the encoder is required to do is push the information of the entire syllables into fixed-dimensional vectors, and the RNN decoder is so flexible that it can map two neighboring points in the feature space to completely different sounds. A widely adopted solution to this problem is to introduce Gaussian noise to the features, turning the network into the *variational* autoencoder [24, 85, 86]. Abstracting away from the mathematical details, the Gaussian noise prevents the encoder from representing two dissimilar syllables close to each other. Otherwise, the noisy representation of the two syllables will overlap and the decoder cannot reconstruct appropriate sounds for each.

The Gaussian VAE represents the syllables as real-valued vectors of an arbitrary dimension, and researchers need to apply a clustering method to these vectors in order to obtain discrete categories. This two-step analysis has several problems:

- i The VAE is not trained for the sake of clustering, and the entire distribution of the encoded features may not be friendly to existing clustering methods.
- ii The encoded features often include individual differences and do not exhibit inter-individually clusterable distribution (see Fig. 2D and the supporting information S1.4).

To solve these problems, this study adopted the ABCD-VAE, which encoded data into discrete categories with a categorical noise under the Dirichlet prior, and performed end-to-end clustering of syllables within the VAE (Fig. 1B). The ABCD-VAE married discrete autoencoding techniques [47, 28, 29] and the Bayesian clustering popular in computational linguistics and cognitive science [36, 37]. It has the following advantages over the Gaussian VAE + independent clustering (whose indices, except iii, correspond to the problems with the Gaussian VAE listed above):

- i Unlike the Gaussian VAE, the ABCD-VAE includes clustering in its learning objective, aiming at statistically grounded discrete encoding of the syllables.
- ii The ABCD-VAE can exploit a speaker-normalization technique that has proven effective for discrete VAEs: The "Speaker Info." is fed directly to the decoder (Fig. 1B), and thus individual-specific patterns need not be encoded in the discrete features [28, 29].
- iii Thanks to the Dirichlet prior, the ABCD-VAE can detect the statistically grounded number of categories on its own [32]. This is the major update from the previous discrete VAEs that eat up all the categories available [47, 28, 29].

Note that the ABCD-VAE can still measure the similarity/distance between two syllables by the cosine similarity of their latent representation immediately before the computation of the classification probability (i.e., logits).

The original category indices assigned by the ABCD-VAE were arbitrarily picked up from 128 possible integers and not contiguous. Accordingly, the category indices reported in this paper were renumbered for better visualization.

Other clustering methods

Clustering results of the ABCD-VAE were evaluated in comparison with baselines and toplines provided by the combination of feature extraction by the Gaussian VAE [24, 25, 26] and clustering on the VAE features by GMM [36, 32, 37]. The number K of GMM clusters was either predetermined or auto-detected. The former fit K multivariate Gaussian distributions by the expectation maximization algorithm while the latter was implemented by Bayesian inference with the Dirichlet distribution prior, approximated by mean-field variational inference. Since a single run of the expectation maximization and variational inference only achieved a local optimum, the best among 100 runs with random initialization was adopted as the clustering results. We used the scikit-learn implementation of GMMs (`GaussianMixture` and `BayesianGaussianMixture`) [87]. The default parameter values were used unless otherwise specified above.

In the analysis of context dependency, we obtained fine-/coarse-grained classifications of syllables based on the features extracted immediately before the computation of classification logits by the ABCD-VAE. The ABCD-VAE computes the classification probability based on the inner-product of those features and the reference vector of each category. Thus, we can compute the similarity among syllables by their cosine in the feature space, and accordingly, we applied k-means clustering on the L2-normalized features. We again adopted the scikit-learn implementation of k-means clustering [87].

Evaluation metrics of syllable clustering

The syllable classification yielded by the ABCD-VAE was evaluated by its alignment with manual annotation by a human expert. We used two metrics to score the alignment: Cohen's kappa coefficient [34] and homogeneity [35]. Cohen's Kappa coefficient is a normalized index for the agreement rate between two classifications, and has been used to evaluate syllable classifications in previous studies [9, 30]. One drawback of using this metric is that it only works when the two classifications use the same set of categories. This requirement was not met in our case, as the model predicted classification and human annotation had different numbers of categories, and we needed to force-align each of the model-predicted categories to the most common human-annotated label to compute Cohen's kappa [9]. On the other hand, the second metric, homogeneity, can score alignment between any pair of classifications, even with different numbers of categories. Homogeneity is defined based on the desideratum that each of the predicted clusters should only contain members of a single ground truth class. Mathematically, violation of this desideratum is quantified by the conditional entropy of the distribution of ground truth classes $\mathcal{C}$ given the predicted clusters $\mathcal{K}$:

$$\text{homogeneity}(\mathcal{C}, \mathcal{K}) := \begin{cases} 1 & H(\mathcal{C}) = 1 \\ 1 - \frac{H(\mathcal{C}|\mathcal{K})}{H(\mathcal{C})} & \text{Otherwise} \end{cases} \tag{1}$$

$$H(\mathcal{C} \mid \mathcal{K}) := -\sum_{k \in \mathcal{K}} \sum_{c \in \mathcal{C}} \frac{|c \cap k|}{N} \log \frac{|c \cap k|}{|k|} \tag{2}$$

$$H(\mathcal{C}) := -\sum_{c \in \mathcal{C}} \frac{|c|}{N} \log \frac{|c|}{N} \tag{3}$$

where N denotes the total number of data points, and $|c \cap k|$ is the number of data that belong to the ground truth class c and the model-predicted category k. The non-conditional entropy $H(\mathcal{C})$ normalizes the homogeneity so that it ranges between 0 and 1. As we noted in the Result section, homogeneity does not penalize overclassification, so it is often combined with another evaluation metric for scoring overclassification, called completeness, and constitutes a more comprehensive metric named V-measure [35]. We report the completeness and V-measure scores of the syllable clustering results in the supporting information S1.5.

Language modeling

After the clustering of the syllables, each sequence, $\mathbf{x} := (x_1, \ldots, x_T)$, was represented as a sequence of discrete symbols, x_t. We performed the analysis of context dependency on these discrete data.

The analysis of context dependency made use of a neural language model based on the current state-of-the-art architecture, Transformer [19, 6]. We trained the language model on 7,779 sequences of Bengalese finch syllables (amounting to 458,753 syllables in total; see Table 3). These training data were defined by the complement of the 100 test sequences that were selected in the following way so that they were long enough (i) and at least one sequence per individual singer was included (ii):

- i The sequences containing 15 or more syllables were selected as the candidates.
- ii For each of the 18 Bengalese finches, one sequence was uniformly randomly sampled among the candidates uttered by that finch.
- iii The other 82/80 sequences were uniformly randomly sampled from the remaining candidates.

The training objective was to estimate the probability of the whole sequences $\mathbf{x}$ conditioned on the information about the individual s uttering $\mathbf{x}$: That is, $\mathbb{P}(\mathbf{x} \mid s)$. Thanks to the background information s, the model did not need to infer the singer on its own. Hence, the estimated context dependency did not comprise the correlation among syllables with individuality, which would not count as a major factor especially from a generative point of view.

The joint probability, $\mathbb{P}(\mathbf{x} \mid s)$, was factorized as $\mathbb{P}(\mathbf{x} \mid s) = \prod_{t=1}^{T} \mathbb{P}(x_t \mid x_1, \ldots, x_{t-1}, s)$, and, the model took a form of the left-to-right processor, predicting each syllable x_t conditioned on the preceding context `<sos>`, $x_1, \ldots, x_{t-1}$, where `<sos>` stands for the special category marking the start of the sequence. See the supporting information S2 for details on the model parameters and training procedure.

While the VAE training excluded incompletely recorded syllables positioned at the beginning/end of recordings, we included them in the language modeling by assigning them with a distinct category. This corresponds to the replacement of non-frequent words with the "unk(nown)" label in natural language processing.

Measuring context dependencies

After training the language model, we estimated how much of the context $x_1, \ldots, x_{t-1}$ was used effectively for the model to predict the upcoming syllable x_t in the test data. Specifically, we wanted to know the longest length L of the truncated context $x_{t-L}, \ldots, x_{t-1}$ such that the prediction of x_t conditioned on the truncated context was worse (with at least 1% greater perplexity) than the prediction based on the full context (Fig. 5A). This context length L is called the *effective context length* (ECL) of the trained language model [5].

One potential problem with the ECL estimation using the birdsong data was that the test data was much smaller in size than the human language corpora used in the previous study. In other words, the perplexity, from which the ECL was estimated, was more likely to be affected by sampling error. To obtain a more reliable result, we bootstrapped the test data (10,000 samples) and used the five percentile of the bootstrapped differences between the truncated and full context predictions. Note that the bootstrapping was performed *after* the predictive probability of the test syllables was computed, so there was no perturbation in the available contexts or any other factors affecting the language model. We call this bootstrapped version of ECL the *statistically effective context length* (SECL). It is more appropriate to estimate the SECL by evaluating the same set of syllables across different lengths of the truncated contexts. Accordingly, only those that were preceded by 15 or more syllables (including `<sos>`) in the test sequences were used for the analysis (4,918 syllables of Bengalese finches; see Table 3).

Supporting information

S1 Details on Syllable Clustering by VAE

S2 Details on the Transformer Language Model

S3 Analysis of Context Dependency in Zebra Finch Song

S4 Detailed Comparison with the Mutual Information Analysis

S5 Detailed Comparison with the Markovian Analysis of Context Dependency

Acknowledgments

This study was supported by MEXT Grant-in-aid for Scientific Research on Innovative Areas #4903 (Evolinguistic; JP17H06380) to HK and KO, JSPS Grant-in-Aid for Scientific Research (JP19KT0023, JP21H03781) to ROT, and for Early-Career Scientists (21K17805) to TM, the JST Core Research for Evolutional Science and Technology 17941861 (JPMJCR17A4) to HK and ACT-X 21454934 (JPMJAX21AN) to TM, and the Mitsubishi Foundation Research Grants in the Natural Sciences (202111014) to TM. We deeply thank Prof. Kazuhiro Wada in Hokkaido University for providing the zebra finch recording dataset. We also gratefully acknowledge the support of the Academic Center for Computing and Media Studies, Kyoto University, regarding the use of their supercomputer system.

References

[1] Friston K. Learning and inference in the brain. Neural Networks. 2003;16(9):1325–1352. doi:10.1016/j.neunet.2003.06.005.

[2] Friston K. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience. 2010;11:127–138.

[3] Chomsky N. Syntactic Structures. The Hague: Mouton and Co.; 1957.

[4] Larson B. Long Distance Dependencies; 2017. Oxford Bibliographies.

[5] Khandelwal U, He H, Qi P, Jurafsky D. Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 284–294.

[6] Dai Z, Yang Z, Yang Y, Carbonell J, Le Q, Salakhutdinov R. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 2978–2988.

[7] Hosino T, Okanoya K. Lesion of a higher-order song nucleus disrupts phrase level complexity in Bengalese finches. Neuroreport. 2000;11(10):2091–2095.

[8] Okanoya K. Song syntax in Bengalese finches: proximate and ultimate analyses. Advances in the Study of Behavior. 2004;34:297–345.

[9] Katahira K, Suzuki K, Okanoya K, Okada M. Complex Sequencing Rules of Birdsong Can be Explained by Simple Hidden Markov Processes. PLoS ONE. 2011;6(9):1–9. doi:10.1371/journal.pone.0024516.

[10] Warren TL, Charlesworth JD, Tumer EC, Brainard MS. Variable sequencing is actively maintained in a well learned motor skill. Journal of neuroscience. 2012;32(44):15414–15425. doi:10.1523/JNEUROSCI.1254-12.2012.

[11] Markowitz JE, Ivie E, Kligler L, Gardner TJ. Long-range Order in Canary Song. PLOS Computational Biology. 2013;9(5):1–12. doi:10.1371/journal.pcbi.1003052.

[12] Berwick RC, Okanoya K, Beckers GJL, Bolhuis JJ. Songs to syntax: the linguistics of birdsong. Trends in Cognitive Science. 2011;15(3):113–121.

[13] Kuypers HGJM. Corticobulbar connexions to the pons and lower brain-stem in man: an anatomical study. Brain. 1958;81(3):364–388. doi:10.1093/brain/81.3.364.

[14] Wild JM, Li D, Eagleton C. Projections of the dorsomedial nucleus of the intercollicular complex (DM) in relation to respiratory-vocal nuclei in the brainstem of pigeon (*Columba livia*) and zebra finch (*Taeniopygia guttata*). Journal of Comparative Neurology. 1997;377(3):392–413. doi:10.1002/(SICI)1096-9861(19970120)377:3<392::AID-CNE7>3.0.CO;2-Y.

[15] Prather JF, Peters S, Nowicki S, Mooney R. Precise auditory–vocal mirroring in neurons for learned vocal communication. Nature. 2008;451(7176):305–310. doi:10.1038/nature06492.

[16] Jin DZ, Kozhevnikov AA. A Compact Statistical Model of the Song Syntax in Bengalese Finch. PLOS Computational Biology. 2011;7(3):1–19. doi:10.1371/journal.pcbi.1001108.

[17] Kershenbaum A, Bowles AE, Freeberg TM, Jin DZ, Lameira AR, Bohn K. Animal vocal sequences: not the Markov chains we thought they were. Proceedings of the Royal Society of London B: Biological Sciences. 2014;281(1792). doi:10.1098/rspb.2014.1370.

[18] Sainburg T, Theilman B, Thielk M, Gentner TQ. Parallels in the sequential organization of birdsong and human speech. Nature Communications. 2019;10(3636). doi:10.1038/s41467-019-11605-y.

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 5998–6008.

[20] van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel Recurrent Neural Networks; 2016.

[21] van den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al.. WaveNet: A Generative Model for Raw Audio; 2016.

[22] Dhariwal P, Jun H, Payne C, Kim JW, Radford A, Sutskever I. Jukebox: A Generative Model for Music; 2020.

[23] Okanoya K. Language evolution and an emergent property. Current Opinion in Neurobiology. 2007;17(2):271–276. doi:10.1016/j.conb.2007.03.011.

[24] Kingma DP, Welling M. Auto-Encoding Variational Bayes; 2014. The International Conference on Learning Representations (ICLR) 2014.

[25] Coffey KR, Marx RG, Neumaier JF. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology. 2019;44(5):859–868. doi:10.1038/s41386-018-0303-6.

[26] Goffinet J, Mooney R, Pearson J. Inferring low-dimensional latent descriptions of animal vocalizations. bioRxiv. 2019;doi:10.1101/811661.

[27] Morita T, Koda H. Exploring TTS without T Using Biologically/Psychologically Motivated Neural Network Modules (ZeroSpeech 2020). In: Proceedings of Interspeech 2020; 2020. p. 4856–4860.

[28] van den Oord A, Vinyals O, Kavukcuoglu K. Neural Discrete Representation Learning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 6306–6315.

[29] Chorowski J, Weiss RJ, Bengio S, van den Oord A. Unsupervised Speech Representation Learning Using WaveNet Autoencoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2019;27(12):2041–2053. doi:10.1109/TASLP.2019.2938863.

[30] Tachibana RO, Oosugi N, Okanoya K. Semi-Automatic Classification of Birdsong Elements Using a Linear Support Vector Machine. PLOS ONE. 2014;9(3):1–8. doi:10.1371/journal.pone.0092584.

[31] Tachibana RO, Koumura T, Okanoya K. Variability in the temporal parameters in the song of the Bengalese finch (*Lonchura striata* var. *domestica*). Journal of Comparative Physiology A. 2015;201(12):1157–1168. doi:10.1007/s00359-015-1046-z.

[32] Bishop CM. Pattern recognition and machine learning. Information science and statistics. New York: Springer; 2006.

[33] Sainburg T, Thielk M, Gentner TQ. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. PLOS Computational Biology. 2020;16(10):1–48. doi:10.1371/journal.pcbi.1008228.

[34] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37–46. doi:10.1177/001316446002000104.

[35] Roseberg A, Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Association for Computational Linguistics; 2007. p. 410–420.

[36] Anderson JR. The adaptive character of thought. Studies in cognition. Hillsdale, NJ: L. Erlbaum Associates; 1990.

[37] Feldman NH, Goldwater S, Griffiths TL, Morgan JL. A Role for the Developing Lexicon in Phonetic Category Acquisition. Psychological Review. 2013;120(4):751–778.

[38] Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159–174.

[39] Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies. In: Kolen JF, Kremer SC, editors. A Field Guide to Dynamical Recurrent Networks. Wiley-IEEE Press; 2001. p. 237–243.

[40] Abnar S, Zuidema W. Quantifying Attention Flow in Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 4190–4197.

[41] Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O. Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences. 2020;117(48):30046–30054. doi:10.1073/pnas.1907367117.

[42] Khalighinejad B, Cruzatto da Silva G, Mesgarani N. Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. Journal of Neuroscience. 2017;37(8):2176–2185. doi:10.1523/JNEUROSCI.2383-16.2017.

[43] Payne RS, McVay S. Songs of Humpback Whales. Science. 1971;173(3997):585–597. doi:10.1126/science.173.3997.585.

[44] Seyfarth R, Cheney D, Marler P. Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. Science. 1980;210(4471):801–803. doi:10.1126/science.7433999.

[45] Kershenbaum A, Blumstein DT, Roch MA, Akçay Ç, Backus G, Bee MA, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. Biological Reviews. 2016;91(1):13–52. doi:10.1111/brv.12160.

[46] Janik VM. Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods. Animal Behaviour. 1999;57(1):133–143. doi:10.1006/anbe.1998.0923.

[47] Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax. In: Proceedings of the 5th International Conference on Learning Representations (ICLR); 2017.

[48] Louizos C, Swersky K, Li Y, Welling M, Zemel RS. The Variational Fair Autoencoder. In: Bengio Y, LeCun Y, editors. Proceedings of the 4th International Conference on Learning Representations (ICLR); 2016.

[49] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-Adversarial Training of Neural Networks. Journal of Machine Learning Research. 2016;17(59):1–35.

[50] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation. 1997;9(8):1735–1780. doi:10.1162/neco.1997.9.8.1735.

[51] Maass W, Natschläger T, Markram H. Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. Neural Computation. 2002;14(11):2531–2560. doi:10.1162/089976602760407955.

[52] Jaeger H, Haas H. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. Science. 2004;304(5667):78–80. doi:10.1126/science.1091277.

[53] Dehaene S, Changeux JP, Nadal JP. Neural networks that learn temporal sequences by selection. Proceedings of the National Academy of Sciences. 1987;84(9):2727–2731. doi:10.1073/pnas.84.9.2727.

[54] Wang X, Takaki S, Yamagishi J. Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020;28:402–415. doi:10.1109/TASLP.2019.2956145.

[55] Sainburg T, Thielk M, Gentner TQ. Latent space visualization, characterization, and generation of diverse vocal communication signals. bioRxiv. 2019;doi:10.1101/870311.

[56] Cynx J. Experimental determination of a unit of song production in the zebra finch (Taeniopygia guttata). Journal of Comparative Psychology. 1990;104(1):3–10. doi:10.1037/0735-7036.104.1.3.

[57] Williams H. Models for song learning in the zebra finch: fathers or others? Animal Behaviour. 1990;39(4):745–757. doi:10.1016/S0003-3472(05)80386-0.

[58] Williams H, Staples K. Syllable chunking in zebra finch (Taeniopygia guttata) song. Journal of Comparative Psychology. 1992;106(3):278–286. doi:10.1037/0735-7036.106.3.278.

[59] Lachlan RF, van Heijningen CAA, ter Haar SM, ten Cate C. Zebra Finch Song Phonology and Syntactical Structure across Populations and Continents—A Computational Comparison. Frontiers in Psychology. 2016;7:980. doi:10.3389/fpsyg.2016.00980.

[60] Chiu CC, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, et al.. State-of-the-art Speech Recognition With Sequence-to-Sequence Models; 2017.

[61] Dunbar E, Cao XN, Benjumea J, Karadayi J, Bernard M, Besacier L, et al. The Zero Resource Speech Challenge 2017. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2017. p. 323–330.

[62] van Niekerk B, Nortje L, Kamper H. Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge. In: Proceedings of Interspeech 2020; 2020. p. 4836–4840.

[63] Baevski A, Hsu WN, Conneau A, Auli M. Unsupervised Speech Recognition; 2021.

[64] Katz SM. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1987;35(3):400–401.

[65] Kneser R, Ney H. Improved Backing-off for N-gram Language Modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal. vol. 1; 1995. p. 181–184.

[66] Bouchard KE, Brainard MS. Neural Encoding and Integration of Learned Probabilistic Sequences in Avian Sensory-Motor Circuitry. Journal of Neuroscience. 2013;33(45):17710–17723. doi:10.1523/JNEUROSCI.2181-13.2013.

[67] Wohlgemuth MJ, Sober SJ, Brainard MS. Linked Control of Syllable Sequence and Phonology in Birdsong. Journal of Neuroscience. 2010;30(39):12936–12949. doi:10.1523/JNEUROSCI.2690-10.2010.

[68] Morita T, Koda H. Difficulties in analysing animal song under formal language theory framework: comparison with metric-based model evaluation. Royal Society Open Science. 2020;7(2):192069. doi:10.1098/rsos.192069.

[69] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77:257–286.

[70] Perfors A, Tenenbaum JB, Regier T. The learnability of abstract syntactic principles. Cognition. 2011;118(3):306–338. doi:10.1016/j.cognition.2010.11.001.

[71] Morita T, Koda H. Superregular grammars do not provide additional explanatory power but allow for a compact analysis of animal song. Royal Society Open Science. 2019;6(7):190139. doi:10.1098/rsos.190139.

[72] Fujimoto H, Hasegawa T, Watanabe D. Neural Coding of Syntactic Structure in Learned Vocalizations in the Songbird. Journal of Neuroscience. 2011;31(27):10023–10033. doi:10.1523/JNEUROSCI.1606-11.2011.

[73] Nishikawa J, Okada M, Okanoya K. Population coding of song element sequence in the Bengalese finch HVC. European Journal of Neuroscience. 2008;27(12):3273–3283. doi:10.1111/j.1460-9568.2008.06291.x.

[74] Nishikawa J, Okanoya K. Dynamical neural representation of song syntax in Bengalese Finch: a model study. Ornithological Science. 2006;5(1):95–103. doi:10.2326/osj.5.95.

[75] Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems. 1989;2(4):303–314. doi:10.1007/BF02551274.

[76] Jin L, Gupta MM, Nikiforuk PN. Universal approximation using dynamic recurrent neural networks: discrete-time version. In: Proceedings of ICNN'95 - International Conference on Neural Networks. vol. 1; 1995. p. 403–408.

[77] Frith CB, Beehler BM. The Birds of Paradise: Paradisaeidae. Bird Families of the World. Oxford: Oxford University Press; 1998.

[78] Scholes EI. Courtship Ethology of Carola's Parotia (Parotia Carolae). The Auk. 2006;123(4):967–990. doi:10.1093/auk/123.4.967.

[79] van Lawick-Goodall J. The Behaviour of Free-living Chimpanzees in the Gombe Stream Reserve. Animal Behaviour Monographs. 1968;1:161–311. doi:10.1016/S0066-1856(68)80003-2.

[80] Tanner JE, Byrne RW. Representation of Action Through Iconic Gesture in a Captive Lowland Gorilla. Current Anthropology. 1996;37(1):162–173. doi:10.1086/204484.

[81] Mori C, Wada K. Audition-Independent Vocal Crystallization Associated with Intrinsic Developmental Gene Expression Dynamics. Journal of Neuroscience. 2015;35(3):878–889. doi:10.1523/JNEUROSCI.1804-14.2015.

[82] Hayase S, Wada K. Singing activity-driven Arc expression associated with vocal acoustic plasticity in juvenile songbird. European Journal of Neuroscience. 2018;48(2):1728–1742. doi:10.1111/ejn.14057.

[83] Bellman R, Kalaba R. On adaptive control processes. IRE Transactions on Automatic Control. 1959;4(2):1–9. doi:10.1109/TAC.1959.1104847.

[84] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady. 1966;10(8):707–710.

[85] Bowman SR, Vilnis L, Vinyals O, Dai A, Jozefowicz R, Bengio S. Generating Sentences from a Continuous Space. In: Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning; 2016.

[86] Zhao T, Zhao R, Eskenazi M. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics; 2017. p. 654–664.

[87] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.