BIRDSONG SEQUENCE EXHIBITS LONG CONTEXT DEPENDENCY COMPARABLE TO HUMAN LANGUAGE SYNTAX

A PREPRINT

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

[1]Primate Research Institute, Kyoto University, JAPAN

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

November 12, 2020

ABSTRACT

Context dependency is a key feature in sequential structures of human language, which requires reference between words far apart in produced sequence. Assessing how long the past context has 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 show that birdsongs have a long context dependency comparable to grammatical structure in human language. We newly estimated the context dependency in birdsongs in a scalable way using a neural-network-based language model whose accessible context length is sufficiently long. Quantitative comparison with the parallel analysis of English sentences revealed that the context dependency in the birdsong was much shorter than that in the sentence, but was comparable to the grammatical structure when semantic factors were removed. Our findings are in accordance with the previous generalization in comparative studies that birdsong is more homologous to human language syntax than the entirety of human language including semantics.

Keywords 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 live (Friston, 2003, 2010; Friston and Stephan, 2007). 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 (Harris, 1945; Chomsky, 1957; Larson, 2017; Khandelwal et al., 2018; Dai et al., 2019).

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* (Hosino and Okanoya, 2000;

*Corresponding Author: rtachi@gmail.com

Okanoya, 2004). Previous studies have suggested that those birdsongs exhibit non-trivially long dependency on previous outputs (Katahira et al., 2011; Warren et al., 2012; Markowitz et al., 2013). Complex sequential patterns of syllables have been discussed in comparison with human language syntax from the viewpoint of formal linguistics (Okanoya, 2004; Berwick et al., 2011, 2012; Berwick and Chomsky, 2016). Neurological studies also revealed homological network structures for the vocal production, recognition, and learning of songbirds and humans (Kuypers, 1958; Wild et al., 1997; Prather et al., 2008). 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 (Katahira et al., 2011; Markowitz et al., 2013). However, the reported lengths of context dependency were 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, the use of a limited language model is problematic for comparative studies because human languages are not modeled precisely by a Markov process (Chomsky, 1956; Rabin and Scott, 1959).

The present study aimed to assess the context dependency in songs of Bengalese finches (*Lonchura striata* var. *domestica*) using modern techniques for the natural language processing. Recent advancements in the machine learning field, particularly in artificial neural networks, provide powerful language models (Vaswani et al., 2017; Devlin et al., 2018; Dai et al., 2019), which are suitable for analyses of birdsong data, and can potentially refer to 200–900 syllables from the past when the data include such long dependency (Khandelwal et al., 2018; Dai et al., 2019). 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 allowed us to assess the sequential property of birdsongs in the same wat as human language data. Moreover, it it preferable to have a common set of syllable categories, which is shared among classifications for all birds, to represent general patterns in the sequences. 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 (Khandelwal et al., 2018; Dai et al., 2019), 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. 3) behaves as a simulator of birdsong production, which exploits the longest context among currently available models (Vaswani et al., 2017; Devlin et al., 2018; Dai et al., 2019). The proposed method is data-agnostic and, thus, enabled a direct comparison between Bengalese finches' songs and human language sentences.

Here, we demonstrate that the context dependency in Bengalese finch's song is much shorter than in English sentences, but is slightly longer than and more comparable to the dependency in purely syntactic representation of the sentences, where words were replaced with grammatical category labels (such as noun and verb) to remove semantic information. These findings corroborate the idea that birdsong sequences are more homologous to human language syntax than the entirety of human language including semantics (Berwick et al., 2011; Gibson and Tallerman, 2012; Miyagawa et al., 2013) and provide a new piece of evidence for the hypothesis that human language modules, such as syntax and semantics, evolved from different precursors that are shared with other animals (e.g., birdsongs and alarm calls respectively; Okanoya, 2007; Okanoya and Merker, 2007; Miyagawa et al., 2013, 2014; Nóbrega and Miyagawa, 2015).

Results

Unsupervised, individual-invariant classification of song syllables

The context dependency analysis requires discrete representations, or "labels", of song syllables. 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 (Kingma and Welling, 2014; Coffey et al., 2019; Goffinet et al., 2019; Sainburg et al., 2019b). We extended this approach and newly proposed an end-to-end unsupervised clustering method named ABCD-VAE, which utilizes the attention-based categorical sampling with the Dirichlet prior. This method automatically classifies syllables into an unspecified number of statistically optimal categories. It also allowed us to exploit the speaker-normalization technique developed for unsupervised learning of human language from speech recordings (van den Oord et al., 2017; Chorowski et al., 2019; Dunbar et al., 2019; Tjandra et al., 2019), yielding syllable classification modulo individual variation.

We used a dataset of Bengalese finches' songs that was originally recorded for previous studies. 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, and fed them 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

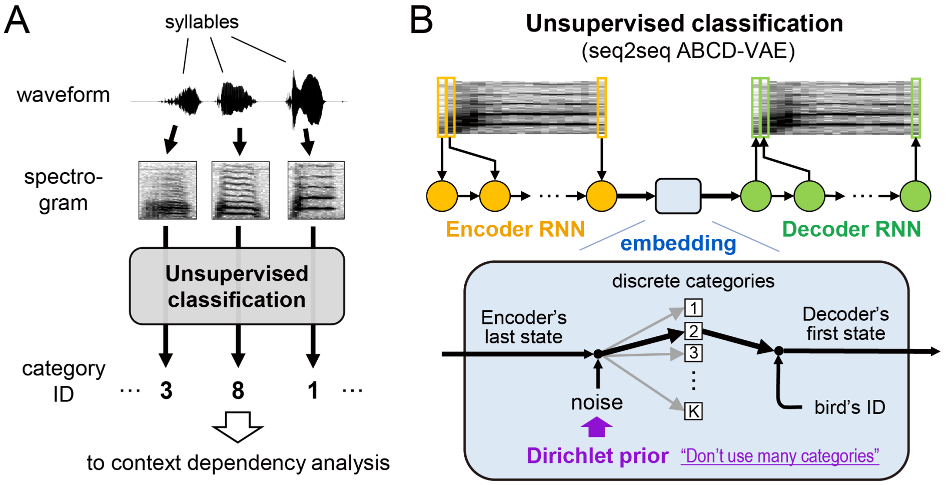

Figure 1. Schematic diagram of newly proposed syllable classification. (A) Each sound waveform segment was converted into the time-frequency representations (spectrograms), and was assigned to one of syllable categories by the unsupervised classification. (B) The unsupervised classification was implemented 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.

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 as a statistical optimum owing to the Dirichlet prior (Bishop, 2006; O'Donnell, 2015; Little, 2019).

As a result, the classifier detected 39 syllable categories in total for all the birds (Fig. 2). Syllables that exhibit similar acoustic patterns tended to be classified into the same category across different birds (Fig. 2A). Almost all birds produced not all but a part of syllable categories in their songs (Fig. 2B). The syllable repertoire of each bird covered 26 to 38 categories (34.78 ± 3.19). Conversely, each category consisted of syllables produced by 8 to 18 birds (16.05 ± 2.52). The detected categories appeared to align with major differences in the spectrotemporal pattern (Fig. 2C).

Quantitative evaluation of syllable classification

We assessed the reliability of the detected classification by its alignment with manual annotations by a human expert (see Tachibana et al., 2014). We scored the alignment using two metrics. One was Cohen's Kappa coefficient (Cohen, 1960), which has been used to evaluate syllable classifications in previous studies (Katahira et al., 2011; Tachibana et al., 2014). 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. To get rid of the force-alignment and any other post-processing, we also evaluated the classification using a more recently developed metric called V-measure (Roseberg and Hirschberg, 2007). The two evaluation metrics showed that the unsupervised classification was mostly consistent with manual annotations assigned by a human expert (Table 1; see also Fig 2D); Even the lowest Kappa coefficient reached the level of "almost perfect agreement" (Landis and Koch, 1977) and the lowest V-measure score among the birds was significantly greater than the chance level ($p < 0.0001$). Hence, our unsupervised clustering of syllables is as reliable as the manual classification by the expert.

To evaluate the individual-invariance of the model-predicted classification, we also measured the identifiability of each individual bird from the category of a syllable it uttered, fitting the conditional categorical distribution to 90% of the syllables by the maximum likelihood criterion and then evaluating the prediction accuracy on the other 10%. As a baseline, we also measured the individual predictability from continuous-valued features of syllables extracted by the canonical VAE (Kingma and Welling, 2014; Coffey et al., 2019; Goffinet et al., 2019; Sainburg et al., 2019b, see Method for details). The feature-to-individual classifier for this baseline was implemented by a feed-forward neural

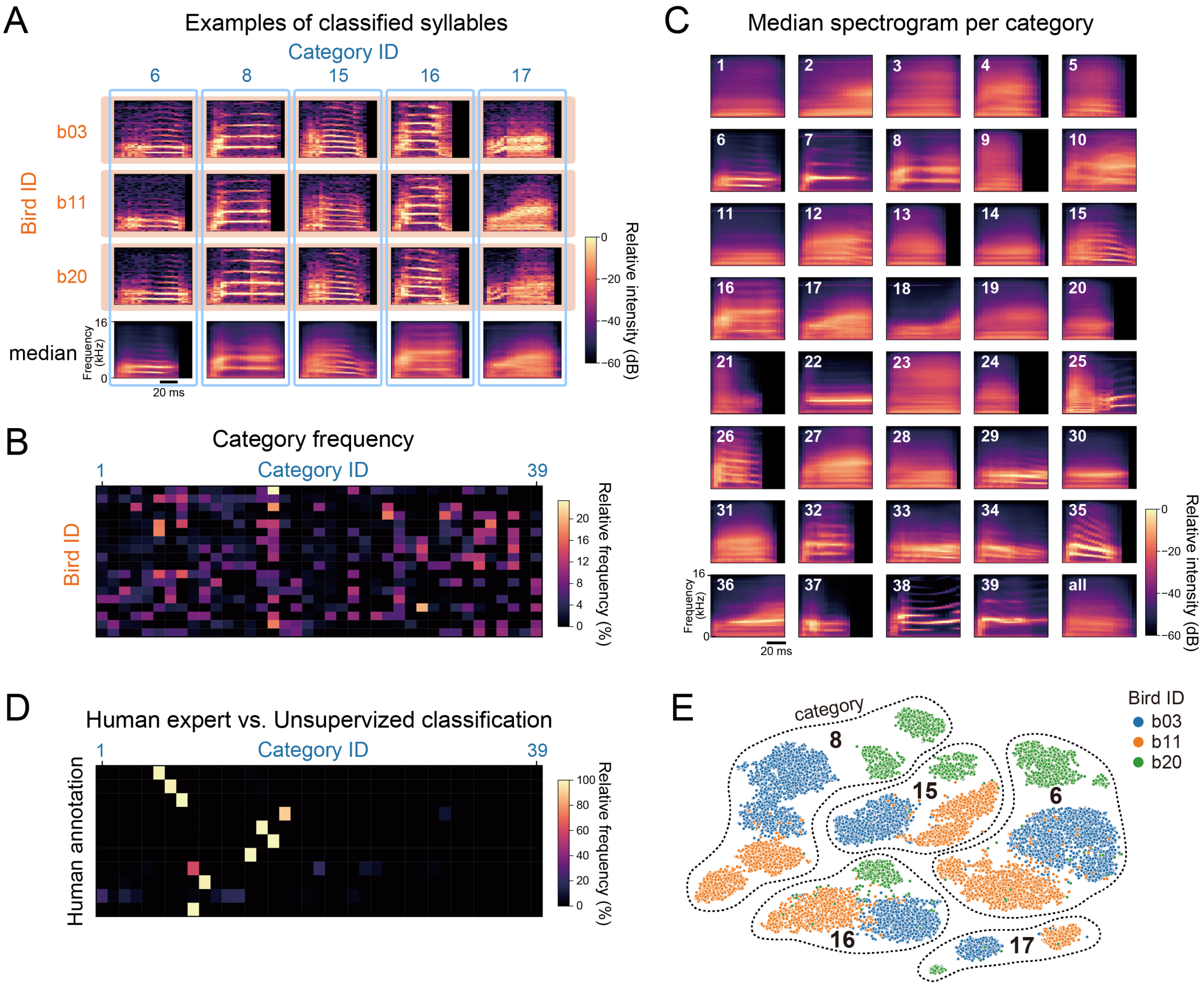

Figure 2. Clustering results of Bengalese finch syllables based on the ABCD-VAE. (A) Syllable spectrograms and their classification across individuals. Syllables in each of the first to third rows (yellow box) were sampled from the same individual. Each column (blue frame) corresponds to the syllable categories assigned by the ABCD-VAE. The bottom row provides the median spectrogram of each category over all the 39 individuals. The examples had the greatest classification probability (> 0.999) among the syllables of the same individual and category. (B) Relative frequency of syllable categories (columns) per individual (rows). (C) Median spectrogram of each syllable category predicted by the ABCD-VAE. (D) Relative frequency of syllable categories (columns) per label manually annotated by a human expert. Only data from a single individual (b03) were presented because the manual annotations were not shared across individuals. (E) 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.

Table 1. Scores of the clustering by the ABCD-VAE. Cohen's kappa coefficient and V-measure evaluated the alignment of the clustering by the ABCD-VAE to manual annotations by a human expert. Since the manual annotation was not shared across individuals, we scored each individual separately and report the median, maximum, and minimum over the individuals. Even the minimum V-measure score was statistically significantly greater than the random baseline (obtained from 10,000 samples of shuffled manual annotation) and the minimum kappa coefficient exhibited so-called "almost perfect agreement". The individual predictability scored the amount of individuality included in the syllable categories yielded by the ABCD-VAE. The score was defined by the accuracy of predictions of the individual uttering a test syllable (10% of the entire data) based on the greatest probability assigned by a classifier fitted to the other 90% of the syllables by the maximum likelihood optimization. The individual predictability from the ABCD-VAE categories was notably smaller than that from the continuous representation obtained via the canonical VAE (the embedding-to-individual classifier was implemented by a feed-forward neural network with a single hidden layer), evidencing the individual-invariance of the ABCD-VAE categories.

Metric		Score	Note
Cohen's kappa (vs. human annotation)	Median	0.9376	
	Max	0.9929	
	Min	0.8101	"Almost perfect agreement"
V-measure (vs. human annotation)	Median over individuals	0.7985	
	Max	0.8879	
	Min	0.6527	$p < 0.0001$
Individual predictability		0.2670	$\ll$ 0.8662 of the canonical VAE

network with a single hidden layer. The individual predictability from the discrete syllable categories was notably smaller than that from the continuous-valued features (Table 1). Thus, the proposed clustering is considered to have ignored individual variations (and other minor differences) visible in the syllable embeddings obtained via the canonical continuous-valued VAE (see also Fig. 2E).

Birdsong sequence more context-dependent than English syntax

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 become 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. 3A). Thus, the context dependency can be quantified as the maximum length of the truncated contexts where the difference is statistically detectable (Khandelwal et al., 2018; Dai et al., 2019). For the context-dependent prediction, we employed the Transformer language model (Vaswani et al., 2017; Devlin et al., 2018; Dai et al., 2019).

Each sequence included syllables that form a continuous song performance, or "bout". We obtained a total of 9,139 bouts, and used 9,039 of them to train the Transformer. The remaining 100 bouts 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 bout. We compared the model predictions between the full-context ("Full", Fig. 3A) and the truncated-context ("Truncated") conditions. Then, the context dependency was quantified by a statistical measure of the effective context length (Khandelwal et al., 2018; Dai et al., 2019), which is the maximum length of the truncated context wherein the mean prediction difference between the two contexts was significantly greater than the canonical 1% threshold in perplexity (at 0.05 level of significance estimated from 10,000 bootstrapped samples; Khalighinejad et al., 2017). For comparison, we performed the same analysis on English sentence datasets (12,327 training sentences and 2,006 test sentences; Silveira et al., 2014) in two different forms. One of them represented the words by their lemma (i.e., original word forms without grammatical inflection; e.g., '*fixed*' was represented as '*fix*'). The other form contained only the grammatical information by replacing words with the part-of-speech (PoS) tags such as nouns and verbs (Perfors et al., 2011a). This process made the analyzed context dependencies free of semantic factors such as co-occurrences of topic-specific words at distance (e.g., '*salt*' and '*sugar*' co-occur in cooking recipes irrespective of a grammatical relation).

The statistically effective context length (SECL) of the Bengalese finch song was eight (pink line in Fig. 3B). In other words, restricting available contexts to eight or less preceding syllables significantly decreased the prediction accuracy comparing with the full-context baseline, while the difference became marginal when nine or more syllables were included in the truncated context. This number is lower than the SECL of the English sentence data, which was ten or

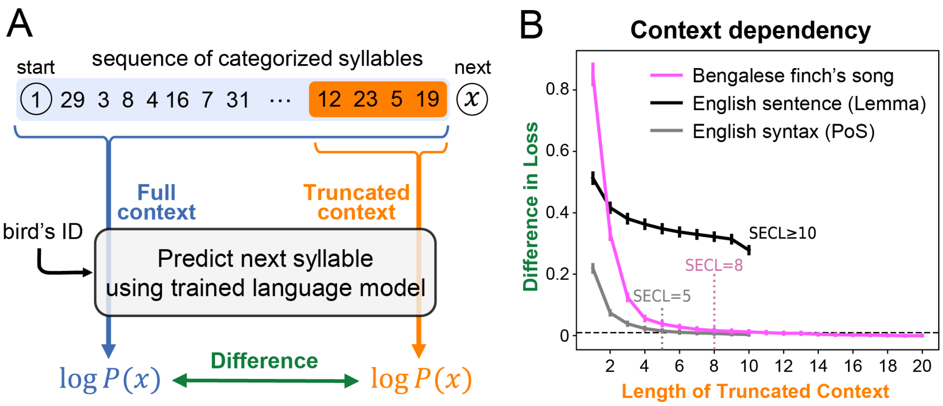

Figure 3. (A) Schematic diagram of the evaluation metric. Predictive probability of each categorized syllable (denoted by x) was computed using the trained language model, conditioned on the full and truncated contexts consisting of preceding syllables (highlighted in blue and orange, respectively). The logarithmic difference of the two predictive probabilities was evaluated, and SECL was defined by the maximum length of the truncated context wherein the prediction difference is statistically significantly greater than a canonical threshold. (B) The differences in the mean loss (negative log probability) between the truncated- and full-context predictions. 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.

greater (black line, achieved the upper bound). On the other hand, the SECL in English decreased to five when we replaced words with the PoS tags and removed the semantic factors (gray line). Hence, the context dependency in Bengalese finch songs is more comparable to that in the English syntax than in the full English including semantics.

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 song bouts. We addressed this question by fitting a state-of-the-art language model, Transformer, to the bouts, 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, particularly for the comparison to human language in text. Various studies have observed how fundamental the classification of voice elements is to animal vocalization (Payne and McVay, 1971; Seyfarth et al., 1980; Hosino and Okanoya, 2000; Kojima, 2003; Suzuki et al., 2006; Kakishita et al., 2007; Markowitz et al., 2013; Kershenbaum et al., 2016; Sainburg et al., 2019a, but see Katahira et al., 2011; Morita and Koda, 2019; Sainburg et al., 2019b for categorization-free approaches).

Our syllable clustering is based on the AVCD-VAE and features the following advantages over previous approaches. First, the ABCD-VAE works in a completely unsupervised fashion. The system finds the statistically optimal classification of syllables instead of generalizing manual labeling of syllables by human annotators (as opposed to Tachibana et al., 2014). Thus, the obtained results are more objective and reproducible (cf. Janik, 1999). Second, the ABCD-VAE detects the statistically optimal number of syllable categories rather than pushing syllables into a pre-specified number of classes (as opposed to Jang et al., 2017; van den Oord et al., 2017; Chorowski et al., 2019). 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 (van den Oord et al., 2017; Chorowski et al., 2019; Tjandra et al., 2019). Finally, the end-to-end clustering by the ABCD-VAE is more optimal than the previous two-step approach—acoustic feature extraction followed by clustering—because the feature extractors are not optimized for clustering and the clustering algorithms are often blind to the optimization objective of the feature extractors (Coffey et al., 2019; Goffinet et al., 2019; Sainburg et al., 2019b). Chorowski et al. (2019) also showed that a similar end-to-end clustering is better at finding speaker-invariant categories in human speech than the two-step approach.

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 (Payne and McVay, 1971; Katahira et al., 2011; Tachibana et al., 2014). Similarly, previous VAE analyses of animal voice often used a convolutional neural network that processed spectrograms as images of a fixed size (Coffey et al., 2019; Goffinet et al., 2019). By contrast, the present study adopted a RNN (specifically, a version called the long short-term memory, abbreviated as LSTM Hochreiter and Schmidhuber, 1997) 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 (Maass et al., 2002; Natschläger et al., 2003; Jaeger and Haas, 2004). Future studies may replace the LSTM in the ABCD-VAE with a reservoir computer to build a more biologically plausible model of vocal learning (cf. Dehaene et al., 1987). Similarly, we may filter some frequency bands in the input sound spectra to simulate the auditory perception of the target animal (cf. the Mel-frequency cepstral coefficients, MFCCs, are used in human speech analysis; Chung et al., 2016; Chorowski et al., 2019; Tjandra et al., 2019), and/or adopt more anatomically/bio-acoustically realistic articulatory systems for the decoder module (cf. Wang et al., 2020, implemented the source-filter model of vocalization based on an artificial neural network). 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 (Figure 2). 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 (see Figure 2E and the supporting information S1.4; Coffey et al., 2019; Goffinet et al., 2019; Sainburg et al., 2019b). 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 (Katahira et al., 2011; Markowitz et al., 2013; Tachibana et al., 2014; Kershenbaum et al., 2016; Sainburg et al., 2019b).

The end-to-end classification by the ABCD-VAE can be applied to a broad range of studies on animal vocalization, including cases where sequential organization of voice units is not at issue. The limitations of the proposed method are the prerequisite for appropriate voice segmentation as it operates on predefined time series of sound spectra, and a single category is assigned to each time series. Although birdsongs often exhibit clear pauses and researchers use them to define syllable boundaries, appropriate voice segmentation is not necessarily clear for other animals (Kershenbaum et al., 2016; Sainburg et al., 2019b), including human speech (Chiu et al., 2017; Dunbar et al., 2017, 2019; Rao et al., 2017). 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 MFCC) and merge contiguous classmate frames to define a syllable-like span (Chorowski et al., 2019; Tjandra et al., 2019).

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 eight 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. (2011) showed that there are at least two dependent context lengths. 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 by Markowitz et al. (2013), 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 (Katz, 1987; Kneser and Ney, 1995; Bengio et al., 2001, 2003; Goldwater et al., 2006; Teh, 2006). 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 (Dai et al., 2019). Thus, the dependency length reported in this study is not likely to be upper-bounded by the model limitations and provides a more precise estimation of the real dependency length in a birdsong than previous studies. The long context dependency in Bengalese finch songs is also evidenced by experimental studies. Warren et al. (2012) reported that several pairs of syllable categories had different transitional probability 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. They also 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 (12.17 syllables on average with the SD of 11.30 according to our own bout data, excluding consecutive repetitions); therefore, the dependency on the previous occurrence cannot be captured by memorizing just one or two previously uttered syllable(s).

Our study also found that Bengalese finch songs are more comparable to human language syntax than to the entirety of human language including semantics. This was demonstrated by our analysis of English sentences represented by sequences of lemmas and PoS categories. While the lemma-represented English sentences exhibited long context dependency beyond ten words as reported in previous studies (Khandelwal et al., 2018; Dai et al., 2019), the dependency length decreased to five—below the Bengalese finch result—when the PoS representation was used and semantic information was removed from the sentences. The gap between the two versions of English suggests that the major factor of long-distance dependencies in human language is the semantics, not the syntax. This is consistent with previous studies reporting that human language syntax prefers shorter dependency (Gibson, 1998; Futrell et al., 2015). Moreover, comparative studies between birdsong and human language often argue the former's lack of semantic function (Berwick et al., 2011, 2012; Gibson and Tallerman, 2012; Miyagawa et al., 2013, 2014), without referential variations seen in alarm calls (Seyfarth et al., 1980; Ouattara et al., 2009; Suzuki et al., 2016). This claim led to the hypothesis that human language syntax and semantics evolved from different precursors—sequence-generating system, such as animal song, and information-carrying system such as alarm calls—which were integrated to shape the entirety of human language (Okanoya, 2007; Okanoya and Merker, 2007; Miyagawa et al., 2013, 2014; Nóbrega and Miyagawa, 2015). Our findings are in accordance with this view, providing a novel relative similarity between birdsong and human language syntax compared to the whole linguistic system. Note that this kind of direct comparative study of human language and animal song was not feasible until flexible language models based on neural networks became available.

The reported context dependency on eight previous syllables also has an implication for possible models of Bengalese finch 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 (Katz, 1987; Kneser and Ney, 1995; Hosino and Okanoya, 2000; Goldwater et al., 2006; Teh, 2006). This approach, however, requires an exponential number of contexts to be represented in the model. In the worst case, the number of possible contexts is $39^8 = 5,352,009,260,481$ when there are 39 syllable types and the context length is eight as detected in this study. 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 (cf. Morita and Koda, 2020). Katahira et al. (2011) assert that the song syntax of the Bengalese finch can be better described with a lower-order hidden Markov model (Rabiner, 1989; Beal et al., 2002, HMM;) 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 (Perfors et al., 2011b) and animal voice sequences (Morita and Koda, 2019) than sequential models like HMM. Another compression possibility is to represent consecutive repetitions of the same syllable categories differently from transitions between heterogeneous syllables (cf. Kershenbaum et al., 2014). 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 (Nishikawa et al., 2008): 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 (Bengio et al., 2001, 2003; Ryeu et al., 2001; Tsuda, 2001; Maass et al., 2002; Jaeger and Haas, 2004; Nishikawa and Okanoya, 2006).

While this study discussed context dependency in the context of memory durability required for generating/processing birdsongs (cf. Katahira et al., 2011; Warren et al., 2012; Markowitz et al., 2013), there are different definitions of context dependency designed for different research purposes. Sainburg et al. (2019a) studied the *mutual information* between birdsong syllables—including Bengalese finch ones—appearing at each discrete distance. Following a study on human language by Lin and Tegmark (2017), Sainburg et al. analyzed patterns in the decay of mutual information to diagnose the generative model behind the birdsong data, instead of addressing the question about memory. Importantly, their mutual information analysis cannot replace our model-based analysis to assess the memory-oriented context dependency: Mutual information is a pairwise metric of probabilistic dependence between two tokens (e.g., words in human languages, syllables in birdsongs), and thus, everything in the middle is ignored. To see the problem, suppose that some tokens reflect the individuality of the speaker (see Figure S3.1a in the supporting information; section S3.1 also provides a more concrete, mathematical example of this problematic situation, and S3.2 introduces other examples that demonstrate difficulties in the mutual information analysis). Two occurrences of speaker-encoding tokens are dependent on each other regardless of their distance if the other tokens between the two are ignored, and this pairwise dependence is what mutual information accounts for. It should be clear now that such pairwise dependence does not necessarily match the agent-oriented concept of context dependency as the only thing relevant to the song recognition task (or speaker identification in this toy example) is the most recent occurrence of the correlating tokens. By contrast, our language modeling approach captured the agent-oriented concept of context dependency as desired. Dependency on a token in the past is detected if the prediction of upcoming tokens becomes notably more difficult by limiting the available context to the more recent tokens (Figure S3.1b; Khandelwal et al., 2018; Dai et al., 2019). In other words, reference to a token in the distant past is considered unnecessary if the same information (e.g., speaker identity) is available from more recent tokens. Therefore, the present study complements, rather than repeats/replaces, the mutual information analysis and findings from it.

We conclude the present paper by noting that the analysis of context dependency via neural language modeling is not limited to Bengalese finch's song. Since neural networks are universal approximators and potentially fit to any kind of data (Cybenko, 1989; Hornik, 1991; Jin et al., 1995; Maass et al., 2002; Lu et al., 2017), the same analytical method is applicable to other animals' voice sequences (Payne and McVay, 1971; Suzuki et al., 2006; Markowitz et al., 2013; Morita and Koda, 2019). Moreover, the analysis of context dependency can also be performed in principle on other sequential behavioral data besides vocalization, including dance (Frith and Beehler, 1998; Scholes, 2006, 2008) and gestures (van Lawick-Goodall, 1968; de Waal, 1988; Tanner and Byrne, 1996; Liebal et al., 2006). Hence, our method provides a crossmodal research paradigm for inquiry into the effect of past behavioral records on future decision making.

Materials & Methods

Recording and segmentation of Bengalese finch's song

We used the same recordings of Bengalese finch songs that were originally reported in our earlier studies Tachibana et al. (2014, 2015). The data were collected from 18 adult males (>140 days after hatching), each isolated in a birdcage placed inside a soundproof chamber. 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 (see Tachibana et al. (2014, 2015) for more information about the recording).

Song syllables were segmented from the continuous recordings using the thresholding algorithm proposed in the previous studies (Tachibana et al., 2014, 2015). We defined a sequence of the syllables as a bout if every two adjacent syllables in the sequence were spaced at most 500 msec apart. These segmentation processes yielded 465,310 syllables and 9,139 bouts in total (≈ 10.79 hours).

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 (but see Bellman and Kalaba, 1959; Levenshtein, 1966; Morita and O'Donnell, To appear, for alternative approaches). 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 (Katahira et al., 2011; Tachibana et al., 2014; Mielke and Zuberbühler, 2013; Morita and Koda, 2019). In this study, we took a non-parametric approach based on a sequence-to-sequence (seq2seq) autoencoder (Bowman et al., 2016; Chung et al., 2016; Zhao et al., 2017; Sainburg et al., 2019b). 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 Figure 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 (Hochreiter and Schmidhuber, 1997, the encoder was bidirectional; Schuster and Paliwal, 1997).

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 (VAE; Kingma and Welling, 2014; Bowman et al., 2016; Zhao et al., 2017, see also Coffey et al., 2019; Goffinet et al., 2019; Sainburg et al., 2019b for its applications to animal vocalization). 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 Figuref 2E 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 (Figure 1B). The ABCD-VAE married discrete autoencoding techniques (Jang et al., 2017; van den Oord et al., 2017; Chorowski et al., 2019) and the Bayesian clustering popular in computational linguistics and cognitive science (e.g., Anderson, 1990; Kurihara and Sato, 2004, 2006; Teh et al., 2006; Kemp et al., 2007; Goldwater et al., 2009; Feldman et al., 2013; Kamper et al., 2017; Morita and O'Donnell, To appear). 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 is optimized for clustering, aiming at optimal 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 (Figure 1B), and thus individual-specific patterns need not be encoded in the discrete features (van den Oord et al., 2017; Chorowski et al., 2019; Tjandra et al., 2019, this is also the framework adopted in the ZeroSpeech 2019, a competition on unsupervised learning of spoken human languages; Dunbar et al., 2019).

iii Thanks to the Dirichlet prior, the ABCD-VAE can detect the optimal number of categories on its own (under an arbitrarily specified upper bound; Bishop, 2006; O'Donnell, 2015; Little, 2019). This is the major update from the previous discrete VAEs that eat up all the categories available (Jang et al., 2017; van den Oord et al., 2017; Chorowski et al., 2019).

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; cf. Mikolov et al., 2013; Deng et al., 2018).

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.

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 (Cohen, 1960) and V-measure (Roseberg and Hirschberg, 2007). 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 (Katahira et al., 2011; Tachibana et al., 2014). 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

Table 2. The size of the training and test data used in the neural language modeling of Bengalese finch songs and the English language. The "SECL" portion of the test syllables was used to estimate the SECL.

Data type	Usage	# of bouts/sentences	# of syllables/words	
			Total	SECL
Bengalese finch	Training	9,039	458,753	—
	Test	100	6,557	4,657
English	Training	12,327	179,456	—
	Test	2,006	21,759	8,833

to the most common human-annotated label to compute Cohen's Kappa (following Katahira et al., 2011). On the other hand, the second metric, V-measure, can score alignment between any pair of classifications, even with different numbers of categories. V-measure is defined based on two desiderata: (i) Each of the predicted clusters should only contain members of a single ground truth class (homogeneity); (ii) The members of each ground truth class should be clustered into the same category (completeness). The two metrics are defined on a scale of 0 (worst) to 1 (best), and their harmonic mean yields the V-measure.

Language modeling

After the clustering of the syllables, each bout, $\mathbf{x} := (x_1, \dots, x_T)$, 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 (Vaswani et al., 2017; Al-Rfou et al., 2018; Dai et al., 2019). We trained the language model on 9,039 bouts, containing 458,753 syllables (Table 2). These training data were defined by the complement of the 100 test bouts that were selected in the following way so that they were long enough (i) and at least one bout per individual singer was included (ii):

i The bouts containing 20 or more syllables were selected as the candidates.

ii For each of the 18 finches, one bout was uniformly randomly sampled among those uttered by that finch.

iii The other 82 bouts were uniformly randomly sampled from the remaining candidates.

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

Measuring context dependencies

After training the language model, we estimated how much of the context $x_1, \dots, x_{t-1}$ 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 $x_{t-L}, \dots, x_{t-1}$ 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 (Figure 3A). This context length L is called the *effective context length* (ECL) of the trained language model (Khandelwal et al., 2018).

One potential problem with the ECL estimation using the Bengalese finch 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. 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 20 or more syllables (including `<sos>`) in the test bouts were used for the analysis (4.657 syllables in total, Table 2).

English data

For comparison, we also estimated the SECL of the language model trained on English data. The data were constructed from the Universal Dependencies English Web Treebank (the training and test portions; Silveira et al., 2014). The database consists of textual English sentences and each word is annotated with the lemma and PoS category. We constructed two versions of training and test data using these lemma and PoS representations of the words: Words may exhibit correlation with one another due to their semantics (e.g., same topic) when they are coded as the lemma. By contrast, the PoS representation of words removes such semantic information, and allowed us to assess the purely syntactic dependencies among the words (cf. Perfors et al., 2011b). Note that this semantics-free data may serve as a more appropriate baseline for the study of birdsongs, whose variation is considered not to encode different meanings (Okanoya, 2007; Okanoya and Merker, 2007; Berwick et al., 2011, 2012; Gibson and Tallerman, 2012; Miyagawa et al., 2013, 2014) unlike alarm calls (Seyfarth et al., 1980; Ouattara et al., 2009; Suzuki et al., 2016).

The words that were preceded by ten or more tokens (including `<sos>`) in the test data sentences were used to estimate the SECL. Accordingly, the upper bound on the SECL (=10) was lower than in the analysis of the Bengalese finch data (=20). The reason for the different settings is that the English sentences were shorter than the Bengalese finch bouts: The quartiles of the bout lengths were 22, 44, and 68, while those of the sentence lengths were 7, 14, and 22 (where both the training and test data were included).

Acknowledgments

This study was funded by MEXT Grant-in-aid for Scientific Research on Innovative Areas #4903 (Evolinguistic; JP17H06380), JSPS Grant-in-Aid for Scientific Research C (JP19KT0023), and the JST Core Research for Evolutional Science and Technology 17941861 (#JPMJCR17A4).

References

Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L. (2018). Character-level language modeling with deeper self-attention.

Anderson, J. R. (1990). *The adaptive character of thought*. Studies in cognition. L. Erlbaum Associates, Hillsdale, NJ.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002). The infinite hidden Markov model. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 577–584. MIT Press.

Bellman, R. and Kalaba, R. (1959). On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9.

Bengio, Y., Ducharme, R., and Vincent, P. (2001). A neural probabilistic language model. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 932–938. MIT Press.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Berwick, R., Beckers, G., Okanoya, K., and Bolhuis, J. (2012). A bird's eye view of human language evolution. *Frontiers in Evolutionary Neuroscience*, 4:5.

Berwick, R. C. and Chomsky, N. (2016). *Why Only Us: Language and Evolution*. MIT Press.

Berwick, R. C., Okanoya, K., Beckers, G. J., and Bolhuis, J. J. (2011). Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Science*, 15(3):113–121.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*.

Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. (2017). State-of-the-art speech recognition with sequence-to-sequence models.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2:113 – 124.

Chomsky, N. (1957). *Syntactic Structures*. Mouton and Co., The Hague.

Chorowski, J., Weiss, R. J., Bengio, S., and van den Oord, A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2041–2053.

Chung, Y.-A., Wu, C.-C., Shen, C.-H., yi Lee, H., and Lee, L.-S. (2016). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *INTERSPEECH*, pages 765–769.

Coffey, K. R., Marx, R. G., and Neumaier, J. F. (2019). DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, 44(5):859–868.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context.

de Waal, F. B. (1988). The communicative repertoire of captive bonobos (Pan Paniscus), compared to that of chimpanzees. *Behaviour*, 106(3-4):183–251.

Dehaene, S., Changeux, J. P., and Nadal, J. P. (1987). Neural networks that learn temporal sequences by selection. *Proceedings of the National Academy of Sciences*, 84(9):2727–2731.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2018). Arcface: Additive angular margin loss for deep face recognition.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., Besacier, L., Sakti, S., and Dupoux, E. (2019). The Zero Resource Speech Challenge 2019: TTS without T. In *Proceedings of Interspeech 2019*, pages 1088–1092.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330.

Feldman, N. H., Goldwater, S., Griffiths, T. L., and Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751–778.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352. Neuroinformatics.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127–138.

Friston, K. J. and Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3):417–458.

Frith, C. B. and Beehler, B. M. (1998). *The Birds of Paradise: Paradisaeidae*. Bird Families of the World. Oxford University Press, Oxford.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Gibson, K. R. and Tallerman, M. (2012). *The Oxford Handbook of Language Evolution*. Oxford University Press.

Goffinet, J., Mooney, R., and Pearson, J. (2019). Inferring low-dimensional latent descriptions of animal vocalizations. *bioRxiv*.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 459–466, Cambridge, MA. MIT Press.

Goldwater, S., L Griffiths, T., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54.

Harris, Z. S. (1945). Discontinuous morphemes. *Language*, 21(3):121–127.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Hosino, T. and Okanoya, K. (2000). Lesion of a higher-order song nucleus disrupts phrase level complexity in bengalese finches. *Neuroreport*, 11(10):2091–2095.

Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80.

Jang, E., Gu, S., and Poole, B. (2017). Categorical reparameterization with Gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Janik, V. M. (1999). Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods. *Animal Behaviour*, 57(1):133–143.

Jin, L., Gupta, M. M., and Nikiforuk, P. N. (1995). Universal approximation using dynamic recurrent neural networks: discrete-time version. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 1, pages 403–408.

Kakishita, Y., Sasahara, K., Nishino, T., Takahasi, M., and Okanoya, K. (2007). Pattern extraction improves automata-based syntax analysis in songbirds. *Lecture Notes in Artificial Inteligence*, 4828:320–332.

Kamper, H., Jansen, A., and Goldwater, S. (2017). A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174.

Katahira, K., Suzuki, K., Okanoya, K., and Okada, M. (2011). Complex sequencing rules of birdsong can be explained by simple hidden Markov processes. *PLoS ONE*, 6(9):1–9.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401.

Kemp, C., Perfors, A., and Tenenbaum, J. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–321.

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., Bohn, K., Cao, Y., Carter, G., Cäsar, C., Coen, M., DeRuiter, S. L., Doyle, L., Edelman, S., Ferrer-i Cancho, R., Freeberg, T. M., Garland, E. C., Gustison, M., Harley, H. E., Huetz, C., Hughes, M., Hyland Bruno, J., Ilany, A., Jin, D. Z., Johnson, M., Ju, C., Karnowski, J., Lohr, B., Manser, M. B., McCowan, B., Mercado, E., Narins, P. M., Piel, A., Rice, M., Salmi, R., Sasahara, K., Sayigh, L., Shiu, Y., Taylor, C., Vallejo, E. E., Waller, S., and Zamora-Gutierrez, V. (2016). Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 91(1):13–52.

Kershenbaum, A., Bowles, A. E., Freeberg, T. M., Jin, D. Z., Lameira, A. R., and Bohn, K. (2014). Animal vocal sequences: not the Markov chains we thought they were. *Proceedings of the Royal Society of London B: Biological Sciences*, 281(1792).

Khalighinejad, B., Cruzatto da Silva, G., and Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, 37(8):2176–2185.

Khandelwal, U., He, H., Qi, P., and Jurafsky, D. (2018). 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)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. The International Conference on Learning Representations (ICLR) 2014.

Kneser, R. and Ney, H. (1995). Improved backing-off for n-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal*, volume 1, pages 181–184.

Kojima, S. (2003). *A Search for the Origin of Human Speech: Auditory and Vocal Functions of Chimpanzee*. Trans Pacific Press and Kyoto University Press, Rosanna, Melbourne; Kyoto.

Kurihara, K. and Sato, T. (2004). An application of the variational Bayesian approach to probabilistic context-free grammars. In *International Joint Conference on Natural Language Processing Workshop Beyond Shallow Analyses*.

Kurihara, K. and Sato, T. (2006). Variational Bayesian grammar induction for natural language. In Sakakibara, Y., Kobayashi, S., Sato, K., Nishino, T., and Tomita, E., editors, *Grammatical Inference: Algorithms and Applications: 8th International Colloquium, ICGI 2006, Tokyo, Japan, September 20-22, 2006. Proceedings*, pages 84–96. Springer Berlin Heidelberg, Berlin, Heidelberg.

Kuypers, H. G. J. M. (1958). Corticobulbar connexions to the pons and lower brain-stem in man: an anatomical study. *Brain*, 81(3):364–388.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Larson, B. (2017). Long distance dependencies. Oxford Bibliographies.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Liebal, K., Pika, S., and Tomasello, M. (2006). Gestural communication of orangutans (Pongo pygmaeus). *Gesture*, 6(1):1–38.

Lin, H. W. and Tegmark, M. (2017). Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299.

Little, M. A. (2019). *Machine Learning for Signal Processing: Data Science, Algorithms, and Computational Statistics*. Oxford University Press.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6231–6239. Curran Associates, Inc.

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560.

Markowitz, J. E., Ivie, E., Kligler, L., and Gardner, T. J. (2013). Long-range order in canary song. *PLOS Computational Biology*, 9(5):1–12.

Mielke, A. and Zuberbühler, K. (2013). A method for automated individual, species and call type recognition in free-ranging animals. *Animal Behaviour*, 86(2):475–482.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Miyagawa, S., Berwick, R., and Okanoya, K. (2013). The emergence of hierarchical structure in human language. *Frontiers in Psychology*, 4:71.

Miyagawa, S., Ojima, S., Berwick, R. C., and Okanoya, K. (2014). The integration hypothesis of human language evolution and the nature of contemporary languages. *Frontiers in Psychology*, 5:564.

Morita, T. and Koda, H. (2019). Superregular grammars do not provide additional explanatory power but allow for a compact analysis of animal song. *Royal Society Open Science*, 6(7):190139. Preprinted in arXiv:1811.02507.

Morita, T. and Koda, H. (2020). Difficulties in analysing animal song under formal language theory framework: comparison with metric-based model evaluation. *Royal Society Open Science*, 7(2):192069.

Morita, T. and O'Donnell, T. J. (To appear). Statistical evidence for learnable lexical subclasses in Japanese. *Linguistic Inquiry*. Accepted with major revisions.

Natschläger, T., Markram, H., and Maass, W. (2003). Computer models and analysis tools for neural microcircuits. In Kötter, R., editor, *Neuroscience Databases: A Practical Guide*, pages 123–138. Springer US, Boston, MA.

Nishikawa, J., Okada, M., and Okanoya, K. (2008). Population coding of song element sequence in the Bengalese finch hvc. *European Journal of Neuroscience*, 27(12):3273–3283.

Nishikawa, J. and Okanoya, K. (2006). Dynamical neural representation of song syntax in bengalese finch: a model study. *Ornithological Science*, 5(1):95–103.

Nóbrega, V. A. and Miyagawa, S. (2015). The precedence of syntax in the rapid emergence of human language in evolution as defined by the integration hypothesis. *Frontiers in Psychology*, 6:271.

O'Donnell, T. J. (2015). *Productivity and reuse in language : a theory of linguistic computation and storage.* MIT Press, Cambridge, MA; London, England.

Okanoya, K. (2004). Song syntax in Bengalese finches: proximate and ultimate analyses. *Advances in the Study of Behavior*, 34:297–345.

Okanoya, K. (2007). Language evolution and an emergent property. *Current Opinion in Neurobiology*, 17(2):271–276. Cognitive neuroscience.

Okanoya, K. and Merker, B. (2007). Neural substrates for string-context mutual segmentation: A path to human language. In Lyon, C., Nehaniv, C. L., and Cangelosi, A., editors, *Emergence of Communication and Language*, pages 421–434. Springer London, London.

Ouattara, K., Lemasson, A., and Zuberbühler, K. (2009). Campbell's monkeys use affixation to alter call meaning. *PLOS ONE*, 4(11):1–7.

Payne, R. S. and McVay, S. (1971). Songs of humpback whales. *Science*, 173(3997):585–597.

Perfors, A., B Tenenbaum, J., L Griffiths, T., and Xu, F. (2011a). A tutorial introduction to bayesian models of cognitive development. *Cognition*, 120:302–321.

Perfors, A., Tenenbaum, J. B., and Regier, T. (2011b). The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Prather, J. F., Peters, S., Nowicki, S., and Mooney, R. (2008). Precise auditory–vocal mirroring in neurons for learned vocal communication. *Nature*, 451(7176):305–310.

Rabin, M. O. and Scott, D. (1959). Finite automata and their decision problems. *IBM Journal of Research and Development*, 3(2):114–125.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

Rao, K., Sak, H., and Prabhavalkar, R. (2017). Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, pages 193–199.

Roseberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure.

Ryeu, J. K., Aihara, K., and Tsuda, I. (2001). Fractal encoding in a chaotic neural network. *Phys. Rev. E*, 64:046202.

Sainburg, T., Theilman, B., Thielk, M., and Gentner, T. Q. (2019a). Parallels in the sequential organization of birdsong and human speech. *Nature Communications*, 10(3636).

Sainburg, T., Thielk, M., and Gentner, T. Q. (2019b). Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv*.

Scholes, E. I. (2006). Courtship Ethology of Carola's Parotia (Parotia Carolae). *The Auk*, 123(4):967–990.

Scholes, E. I. (2008). Evolution of the courtship phenotype in the bird of paradise genus Parotia (Aves: Paradisaeidae): homology, phylogeny, and modularity. *Biological Journal of the Linnean Society*, 94(3):491–504.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Seyfarth, R., Cheney, D., and Marler, P. (1980). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, 210(4471):801–803.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Suzuki, R., Buck, J. R., and Tyack, P. L. (2006). Information entropy of humpback whale songs. *The Journal of the Acoustical Society of America*, 119(3):1849–1866.

Suzuki, T. N., Wheatcroft, D., and Griesser, M. (2016). Experimental evidence for compositional syntax in bird calls. *Nature Communications*, 7(1):10986.

Tachibana, R. O., Koumura, T., and Okanoya, K. (2015). Variability in the temporal parameters in the song of the bengalese finch (*Lonchura striata* var. *domestica*). *Journal of Comparative Physiology A*, 201(12):1157–1168.

Tachibana, R. O., Oosugi, N., and Okanoya, K. (2014). Semi-automatic classification of birdsong elements using a linear support vector machine. *PLOS ONE*, 9(3):1–8.

Tanner, J. E. and Byrne, R. W. (1996). Representation of action through iconic gesture in a captive lowland gorilla. *Current Anthropology*, 37(1):162–173.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 985–992, Stroudsburg, PA, USA. Association for Computational Linguistics.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Tjandra, A., Sisman, B., Zhang, M., Sakti, S., Li, H., and Nakamura, S. (2019). VQVAE unsupervised unit discovery and multi-scale Code2Spec inverter for Zerospeech Challenge 2019.

Tsuda, I. (2001). Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behavioral and Brain Sciences*, 24(5):793–810.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. (2017). Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6306–6315. Curran Associates, Inc.

van Lawick-Goodall, J. (1968). The behaviour of free-living chimpanzees in the Gombe stream reserve. *Animal Behaviour Monographs*, 1:161–311.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, X., Takaki, S., and Yamagishi, J. (2020). Neural source-filter waveform models for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:402–415.

Warren, T. L., Charlesworth, J. D., Tumer, E. C., and Brainard, M. S. (2012). Variable sequencing is actively maintained in a well learned motor skill. *Journal of neuroscience*, 32(44):15414–15425.

Wild, J., Li, D., and Eagleton, C. (1997). 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*, 377(3):392–413.

Zhao, T., Zhao, R., and Eskenazi, M. (2017). 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)*, pages 654–664. Association for Computational Linguistics.