frontiers
in Behavioral Neuroscience

REVIEW
published: 20 December 2021
doi: 10.3389/fnbeh.2021.811737

Toward a Computational Neuroethology of Vocal Communication: From Bioacoustics to Neurophysiology, Emerging Tools and Future Directions

Tim Sainburg [1,2]* *and Timothy Q. Gentner* [1,3,4,5]

[1] Department of Psychology, University of California, San Diego, La Jolla, CA, United States, [2] Center for Academic Research & Training in Anthropogeny, University of California, San Diego, La Jolla, CA, United States, [3] Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA, United States, [4] Neurobiology Section, Division of Biological Sciences, University of California, San Diego, La Jolla, CA, United States, [5] Kavli Institute for Brain and Mind, University of California, San Diego, La Jolla, CA, United States

OPEN ACCESS

Edited by:
Susanna Pietropaolo,
UMR5287 Institut de Neurosciences Cognitives et Intégratives d'Aquitaine (INCIA), France

Reviewed by:
Andrea M. Simmons,
Brown University, United States
Andrew Bass,
Cornell University, United States

***Correspondence:**
Tim Sainburg
tsainburg@ucsd.edu

Specialty section:
This article was submitted to Individual and Social Behaviors, a section of the journal Frontiers in Behavioral Neuroscience

Received: *09 November 2021*
Accepted: *29 November 2021*
Published: *20 December 2021*

Citation:
Sainburg T and Gentner TQ (2021) Toward a Computational Neuroethology of Vocal Communication: From Bioacoustics to Neurophysiology, Emerging Tools and Future Directions. Front. Behav. Neurosci. 15:811737. doi: 10.3389/fnbeh.2021.811737

Recently developed methods in computational neuroethology have enabled increasingly detailed and comprehensive quantification of animal movements and behavioral kinematics. Vocal communication behavior is well poised for application of similar large-scale quantification methods in the service of physiological and ethological studies. This review describes emerging techniques that can be applied to acoustic and vocal communication signals with the goal of enabling study beyond a small number of model species. We review a range of modern computational methods for bioacoustics, signal processing, and brain-behavior mapping. Along with a discussion of recent advances and techniques, we include challenges and broader goals in establishing a framework for the computational neuroethology of vocal communication.

Keywords: neuroethology, computational neuroethology, denoising, vocalization, UMAP, sequence model, morph

1. INTRODUCTION

Over the past several years emerging methods have enabled biologists to capture and quantify ethological data in ways that yield new insights into the structure and organization of behavior. These methods capitalize on two advances: the ability to record and annotate very-large behavioral datasets, and the use of new computational tools to reveal structure within and between these datasets. The ethological and neuro-ethological study of animal communication has a long history, and its future stands to benefit greatly from these new methods. Here, we discuss this emerging set of tools available to the animal communication researcher. We contextualize these computational methods within the emerging field of computational ethology more broadly and discuss how these tools can be applied in behavior and neurophysiology.

Many of the challenges that exist in the computational neuroethology of vocal behavior are neither new nor unique and parallel those in other areas of human and animal behavior. For example, the algorithmic discovery of vocal units and sequential organization in animal communication parallels the zero-speech challenge in language acquisition: given limited sensory information, can we build a system that discovers subwords, words, and sequential and syntactic organization present in speech (Versteegh et al., 2015). In animal communication the challenge is similar: can we infer vocal segment boundaries, categories, and temporal organization from the physical and temporal characteristics of the signal. The computational neuroethology of vocal communication also parallels the emerging field of motion sequencing and the mapping behavioral kinematics, where new technologies allowing researchers to map postures and behavioral kinematics have facilitated new understandings of behavioral dynamics across scales (Anderson and Perona, 2014; Berman, 2018; Brown and De Bivort, 2018; Christin et al., 2019; Datta et al., 2019; Pereira et al., 2019). It is the goal of computational neuroethology to not only develop an understanding of the organization of behaviors, but also the neural and cognitive mechanisms that facilitate behavior. This review synthesizes work from several fields including bioacoustics, systems neuroscience, and computational neuroethology to discuss emerging methodologies and frameworks which span these fields and are available to vocal communication researchers.

The review begins with considerations in bioacoustics and signal processing and then shifts to a consideration of acoustic structure, sequential organization, and eventually to mapping the acoustic and sequential structure of vocal communication to neurophysiology correlates of behavior and perception. Throughout our review of current approaches, we relay ongoing challenges, discuss future directions, and attempt to give practical advice on vocal analyses.

2. SIGNAL PROCESSING AND DENOISING

Recorded sounds typically contain a mixture of both relevant and irrelevant components. Computational ethology often relies on modeling structure in data without making assumptions about the relevant features. Thus it is often important to remove irrelevant features (i.e., background noise) prior to analysis. Ones operationalization of noise can vary based upon the end goal of the analysis. A simple example is band-pass filtering: because vocalizations typically occur in a confined frequency range, it is reasonable to consider signal outside of that range noise and filter it away. When a recording contains vocalizations from two animals, a songbird with song in a high-frequency range, and heterospecific calls in a low-frequency range, if the subject of interest is the songbird, a simple high-pass filter can be applied to attenuate the non-target calls. When frequency ranges overlap between signal and noise, however, the problem of noise reduction becomes more difficult.

2.1. Noise Reduction

Determining what constitutes noise in recordings is non-trivial and impacts what type of noise reduction algorithm can and should be used. In a systematic review of noise reduction methods in bio-acoustics, Xie et al. (2020) outline six classes of noise reduction algorithms used for bio-acoustics: (1) Optimal FIR filter (e.g., Kim et al., 2000), (2) spectral subtraction (e.g., Boll, 1979; Kiapuchinski et al., 2012; Sainburg et al., 2020b), (3) minimum-mean square error short-time spectral amplitude estimator (MMSE-STSA; e.g., Ephraim and Malah, 1984; Alonso et al., 2017; Brown et al., 2017) (4) wavelet based denoising (e.g., Ren et al., 2008; Priyadarshani et al., 2016) (5) image processing based noise reduction, and (6) deep learning based noised reduction. These noise reduction algorithms can be broadly divided into two categories: stationary and non-stationary noise reduction (**Figure 1A**). Stationary noise reduction acts on noise that is stationary in intensity and spectral shape over time, such as the constant hum of electronics. Non-stationary noise reduction targets background noise that is non-stationary and can fluctuate in time, like the on-and-off presence of a plane flying overhead (**Figure 1B**). Stationary noise reduction algorithms operationalize noise as stationary signals, for example, the constant hum from a nearby electronic device in a laboratory setting, or insect noise in a field setting.

One approach to stationary noise reduction is spectral gating, a spectral-subtraction algorithm (e.g., Kiapuchinski et al., 2012; Sainburg et al., 2020b). The general notion is that for each frequency component of the signal, any time-frequency component below a threshold is discarded as noise. Spectral gating computes the mean and standard deviation of each frequency channel of a Short-Time Fourier Transform (STFT) of a signal (e.g., a spectrogram) and optionally a noise clip. A threshold, or gate, for each frequency component is then set at some level above the mean (e.g., three standard deviations). This threshold determines whether a time-frequency component in the spectrogram is considered signal or noise. The spectrogram is then masked based upon this threshold and inverted (with an inverse STFT) back into the time domain.

2.2. Non-stationary Noise Reduction

While stationary noise reduction algorithms can operationalize noise as any stationary acoustic signal, non-stationary algorithms vary in how they determine what is signal and what is noise. Non-stationary noise can be more challenging to remove because it can be difficult to algorithmically define the difference between signal and noise. Because the hum of a computer in the background of a lab-recording is stationary, it can be defined as noise and can be readily removed. A bird hopping around its cage can produce time-varying sounds in the same frequency range as song, making it especially pernicious.

One approach for determining the boundary between signal and non-stationary noise is to determine the timescale on which the signal acts and treat anything outside of that timescale as noise. For example, zebra finch motifs are generally between 0.5 and 1.5 s long repeated one to four times (Bruno and Tchernichovski, 2019). Any acoustic event that is outside of that time range could be considered noise. Spectral gating can be extended to non-stationary noise reduction by computing a variable gate based upon the current estimate of background noise. In the Python package noisereduce (Sainburg, 2019), this background estimate is computed using a time-smoothed spectrogram (using a forward and backward IIR filter) on a timescale parameterized by the expected signal length, an approach motivated by the Per-Channel Energy Normalization algorithm (outlined in Section 3). An example of this is given in **Figure 1**, where stationary and non-stationary spectral gating noise reduction is applied to birdsong with an airplane noise occurring in the background of the middle of the recording. Because the airplane noise is non-stationary, The stationary

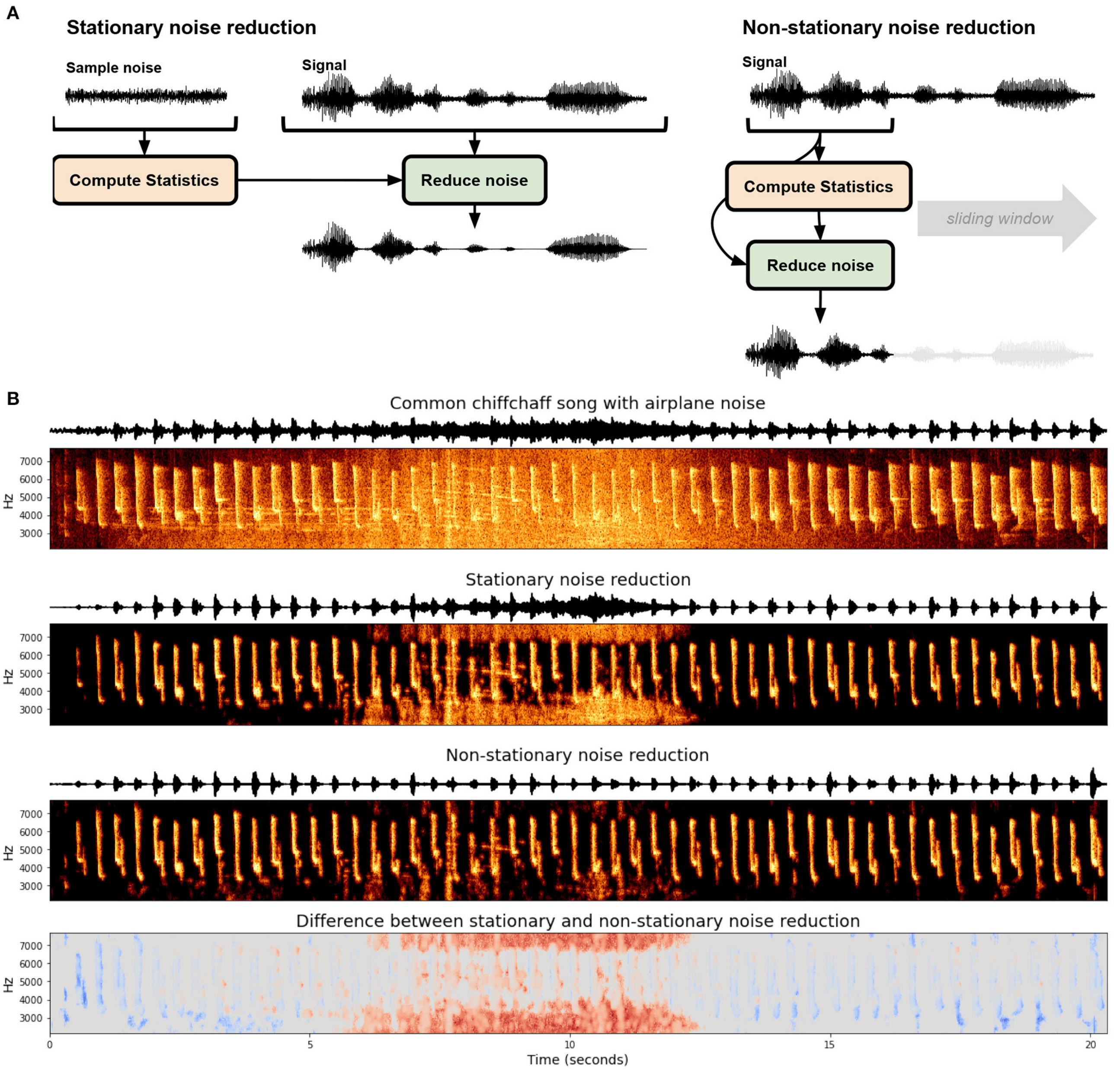

FIGURE 1 | Stationary and non-stationary spectral gating noise reduction. **(A)** An overview of each algorithm. Stationary noise reduction typically takes in an explicit noise signal to calculate statistics and performs noise reduction over the entire signal uniformly. Non-stationary noise reduction dynamically estimates and reduces noise concurrently. **(B)** Stationary and non-stationary spectral gating noise reduction using the noisereduce Python package (Sainburg, 2019) applied to a Common chiffchaff (*Phylloscopus collybita*) song (Stowell et al., 2019) with an airplane noise in the background. The bottom frame depicts the difference between the two algorithms.

approach fails in two ways relative to the non-stationary approach: the airplane noise is not fully successfully gated at its peak in the middle of the recording (shown as red in the bottom panel) and weaker syllables of song are treated as noise and reduced in the beginning and end of the clip (shown in blue in the bottom panel). Advantages of non-stationary noise reduction are not unique to acoustic noise: when we know the timescale of a signal we can use the same non-stationary principles to remove noise occurring at different timescales. For example in the continuous recording of neural data, action potentials occur within the range of 1 ms. Events occurring over tens or hundreds of milliseconds can therefore be treated as noise.

2.3. Reducing Noise With Deep Learning

A promising future avenue for noise reduction is in explicitly training machine learning algorithms to mask or remove noise, as is done in speech enhancement and segregation (Wang and Chen, 2018). At present, however, deep learning based noise reduction has not been utilized directly in bio-acoustics (Xie et al., 2020). Xie et al. (2020) attribute this to a lack of utility when using denoising in some applications of deep learning-based bio-acoustics detection (Kong et al., 2017). The utility of noise reduction exists beyond classification tasks, however. For example, computing spectral features and acoustic similarity between vocalizations can be susceptible to background noise. Recent work by Stowell et al. (2019) shows that manipulating datasets by superimposing background environment noise on vocal datasets can reduce confounds and improve identification across recording conditions. Similar approaches could be used to remove noise. For example, spectral gating could be extended with neural networks by training a neural network to learn a mask to gate away background noise and recover the lower-noise spectrogram, as has been done in speech enhancement applications (Wang and Chen, 2018; Lee and Kim, 2020).

It is also important to consider what information is being removed by pre-processing techniques such as denoising. Pre-processing methods throw away potentially valuable information that will influence downstream analyses. De-noising vocal data without careful consideration can remove lower amplitude syllables of birdsong or infrequent vocalizations outside of the expected frequency range.

3. SIGNAL REPRESENTATION

An important consideration in any analysis pipeline is how to represent the data that goes in. Animal vocalizations are typically recorded using one or more microphones at a sampling rate that can capture the full spectral range of the vocalization. Performing analyses directly upon recorded waveforms is not always optimal for capturing informative structure in vocal data, however. Waveforms are high-dimensional representations of audio that can make it difficult for learning algorithms to capture time-frequency structure in vocalizations. Spectro-temporal representations can be both lower-dimensional, and more explicitly capture complex time-frequency relationships in vocalizations.

Spectrograms are, at present, the most common form of vocalization representation, both for visualization and as input to learning algorithms, both in bio-acoustics and speech. When representing vocal data with a spectrogram, the parameters used to compute the spectrogram can have an important influence on the performance of the algorithm (Elie and Theunissen, 2016; Knight et al., 2020). The most important parameterization of spectrograms is the trade-off between temporal and frequency resolution when computing a spectrogram, a result of the Heisenburg Uncertainty Principle (Gardner and Magnasco, 2006; Moca et al., 2021). For example, three spectrograms are shown in **Figures 2A–C** with different windows used to compute the Short-Time Fourier Transform. The first has an intermediate-sized window with intermediate time and frequency resolution (**Figure 2A**), the second uses a short window with high time-resolution and low frequency-resolution (**Figure 2B**), and the third uses a long window with high frequency-resolution but low time-resolution (**Figure 2C**).

A number of approaches exist to improve time and frequency resolution. Time-frequency reassigned spectrograms attempt to improve time-frequency resolution using additional information from the phase spectrum (**Figure 2D**) (Fulop and Fitz, 2006; Gardner and Magnasco, 2006; Xiao and Flandrin, 2007). Wavelet transforms (**Figure 2E**) have more recently been used in representing animal vocalizations (Main and Thornton, 2015; Priyadarshani et al., 2016, 2020; Hsu et al., 2018), and allows multi-scaled emphasis on time vs. frequency, for example emphasizing frequency resolution at lower frequencies and time-resolution at higher frequencies, intuitively because an uncertainty of 50 Hz is more relevant at 500 Hz than at 5,000 Hz. Most recently, the superlet (**Figure 2F**) enables time-frequency super-resolution by geometrically combining sets of wavelets with increasing constrained bandwidths (Moca et al., 2021).

There are also several variants of spectrograms and time-frequency representations that differentially emphasize time-frequency information. For example, log-scaling spectrograms in frequency emphasizes lower frequency ranges over higher frequency ranges, which parallels both the cochlea and perception (Eldredge et al., 1981). Mel-scaling (**Figure 2G**), is a form of log-scaling fit to fit human perception (Stevens et al., 1937), though the specific scaling range relative to human perception are imperfect (Greenwood, 1997). Mel-Frequency Cepstral Coefficients (MFCCs; **Figure 2H**) additionally compute the Discrete Cosine Transform on the Mel-spectrogram, and were, until recently, commonly used for speech recognition because they are generally robust to noise and emphasize the frequency range of speech (**Figure 2H**) (Muda et al., 2010). Another model, directly relevant to physiology, is the Cochleagram (Brown and Cooke, 1994; Feather et al., 2019; Rahman et al., 2020). Cochleagrams mimic the cochlea by using a filter bank associated with points on the basilar membrane to mimic an impulse response **Figure 2I**).

A new approach that has shown much promise in bio-acoustics is Per-Channel Energy Normalization (PCEN; **Figure 2J**; Wang et al., 2017; Lostanlen et al., 2018). Lostanlen et al. (2018) identify three advantages of PCEN: (1) temporal integration, (2) adaptive gain control, and (3) dynamic gain compression. Temporal integration estimates the background noise at each frequency band. Adaptive gain control then adapts the gain of the spectral representation. Finally, dynamic range compression adaptively shifts the range of low and high amplitude components of the signal. Adaptive gain control is ubiquitous to mammalian auditory processing and is also often used in cochleagrams (Rahman et al., 2020). PCEN has been shown to aid in enhancing animal vocalizations relative to background noise across distances from the microphone (Lostanlen et al., 2019a) and reduce biases in bio-acoustics background settings such as dawn vs. dusk (Lostanlen et al., 2019b).

Descriptive basis-features features can also be used to represent vocalizations for downstream analyses. One challenge with using basis-features for vocal analysis is in determining

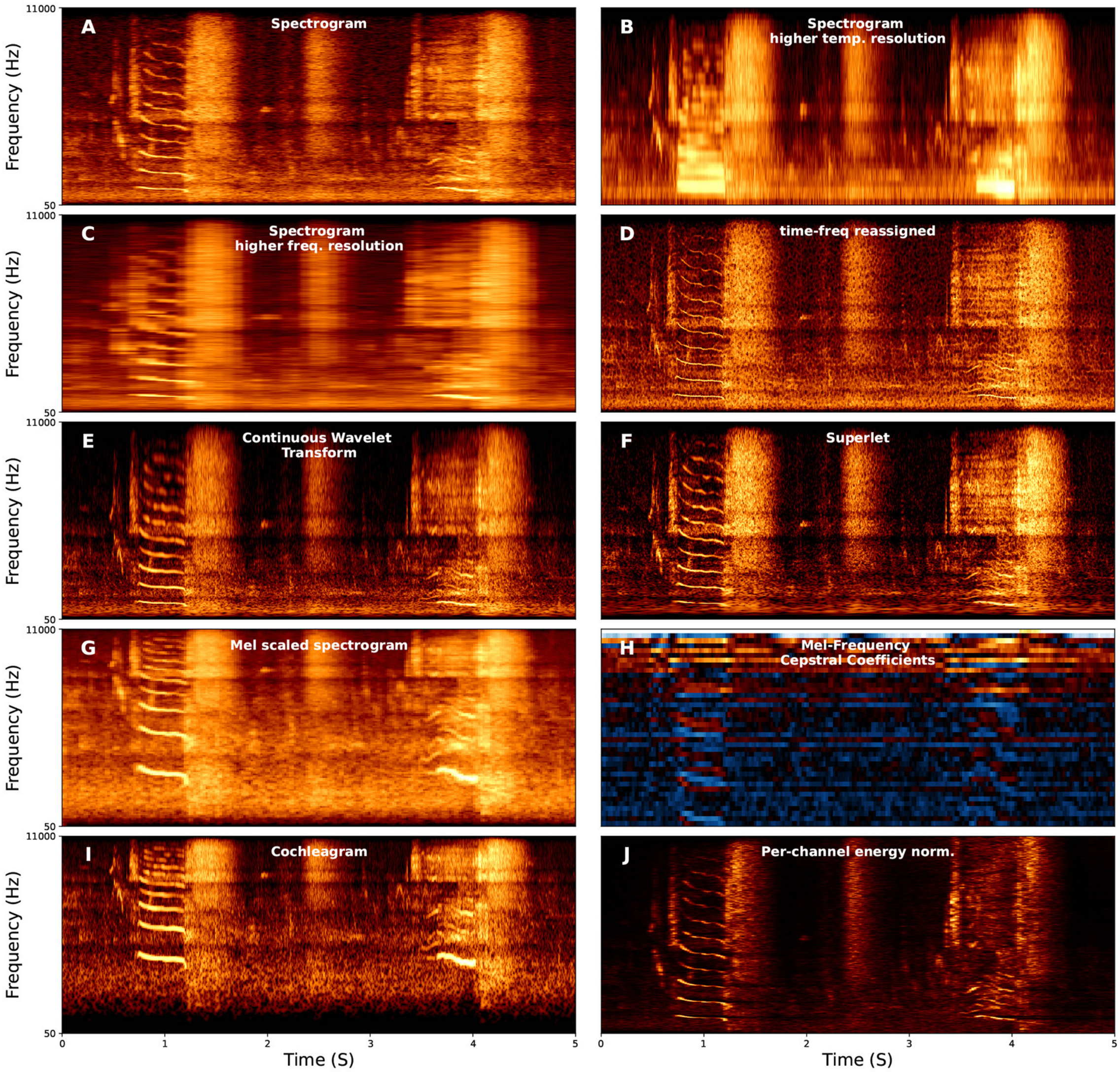

FIGURE 2 | Examples of several different spectral representations of a five-second red deer (*Cervus elaphus*) vocalization. For each axis, the *x*-axis corresponds to time, and the *y*-axis corresponds to frequency. The y-axis corresponds to frequency and is linearly spaced in **(A–F,J)** between 50 and 11,000 Hz and log-spaced in the same range for **(G–I)**. **(F)** Continuous Wavelet Transform using the Morlet (i.e., Gabor) wavelet.

what basis-features are relevant (Tchernichovski et al., 2000; Elie and Theunissen, 2016). Very few species have been rigorously examined to determine what acoustic features distinguish vocal units (Elie and Theunissen, 2016; Kershenbaum et al., 2016). Swamp sparrow notes, for example, are relatively simple vocalizations and can be well-described using just the length of the note, the peak frequency at the start of the note, and the peak frequency at the end of the note (Clark et al., 1987). One approach to determining what features are relevant in a vocal signal is to train classifiers to predict behaviorally-relevant information, such as individual identity, age, or the activity the animal is engaged on a full set to basis features, and retain those features which are highly informative (Elie and Theunissen, 2016, 2018).

4. IDENTIFYING, SEGMENTING, AND LABELING VOCALIZATIONS

Vocalization data can be recorded in a number of different settings, ranging from single individuals in well-controlled and acoustically isolated lab settings to multi-individual and multi-species recordings taken next to a busy highway. When vocalizations are produced by isolated, single individuals, segmenting out vocalizations can often be performed simply by thresholding the vocal envelope and assuming any detected noise events that match the statistics of the vocalizing animal (e.g., frequency and length of vocalization) are vocalizations (Tchernichovski et al., 2000). More complex environments and species with more complex vocal structure require more complex solutions (Priyadarshani et al., 2018).

Experimental paradigms in neuroethology differ from bioacoustics in that environmental sounds can usually be controlled, but are still faced with the challenge of often being made in colony settings with multiple vocalizing individuals or individuals who make non-vocal sounds such as interaction with a living space. Regardless of context, recent advances in machine learning algorithms for passive monitoring of acoustic environments allow for real-time labeling of species and individuals in noisy environments.

Automatic vocalization annotation can be broken down into three related tasks: identification, segmentation, labeling. Identifying refers to what animal is vocalizing and at what times and frequency channels. Segmentation refers to the segmentation of vocalizations into their constituent units, labeling then refers to grouping units into discrete element categories. A spectrogram outlining all three tasks is given in **Figure 3A**. Two individuals in the target species, Australian pied butcherbird *Cracticus nigrogularis* are vocalizing over top of background noise from another, unidentified, species of songbird, as well as an unidentified species of insect. Each bird's song can be divided into segmental units (notes) which can be further categorized into discrete element categories ("A," "B," "C," ...). In such a dataset, labeling challenges occur over multiple levels: identifying the species, identifying the individual, segmenting vocal units, and labeling vocal units into discrete categories. Some algorithms perform only one of these steps at a time, while others perform all three.

4.1. Detecting Species and Individuals

To detect species in continuous bio-acoustic data, several open-source tools and datasets have recently been made available for passive acoustic monitoring. A summary of many of these software and their features are given in Priyadarshani et al. (2018, Table 4). Over the past few years machine learning competitions challenging researchers to produce species recognition algorithms have motivated an increasing number of open-source approaches to bioacoustic sound recognition (e.g., Lasseck, 2013; Murcia and Paniagua, 2013; Goëau et al., 2014; Stowell et al., 2019). The same tools can be applied to differentiating between individuals in the same recording environment (e.g., Adi et al., 2010; Mielke and Zuberbühler, 2013). Most recent approaches rely on deep neural networks to detect vocalizations in noisy environments (e.g., Stowell et al., 2019; Cohen et al., 2020a). Current neural networks generally rely on some combination of convolutional filters in the temporal-frequency space of spectrograms (Convolutional Neural Networks or CNNs, **Figure 3B**) and temporal-recurrence (Recurrent Neural Networks, or RNNs, **Figure 3C**). Convolutional filters in the time-frequency space of spectrograms allow neural networks to learn complex spectro-temporal features used to classify sounds (**Figure 3B**). Temporal recurrence allows neural networks to learn sequential and temporal relationships that unfold over long time delays (**Figure 3C**). In combination, recurrent and convolutional architectures allow complex, non-linear spectrotemporal features that occur over arbitrary timescales to be captured by neural network architectures.

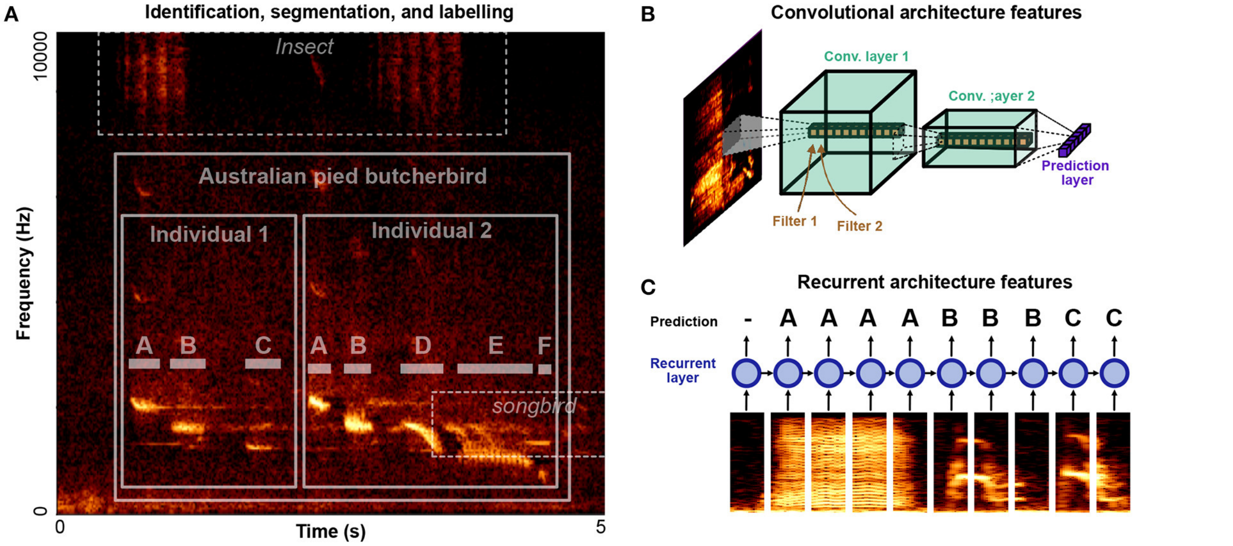

FIGURE 3 | Levels of organization and architectural design features for identifying, segmenting, and labeling vocalizations. **(A)** An example clip of Australian pied butcherbird song (Janney et al., 2016) is shown containing two male butcherbirds, alongside background noise containing insect noise and another bird's song. A classification task exists at several levels: identifying the target species, differentiating individuals, detecting note boundaries, and classifying notes. **(B)** A cartoon diagram of a convolutional neural network architecture applied to a spectrogram. Convolutional filters are applied in time-frequency space. Deeper layers have larger spectrotemporal receptive fields and learn more complex filters. **(C)** A cartoon diagram of a recurrent neural network applied to a song spectrogram. Spectral slices are input to recurrent layers in the network (depicted as a circle) which are recurrent in time, allowing information to be integrated over time.

4.2. Segmenting and Labeling Vocal Units

Beyond identifying individuals and species, many analyses of vocal communication rely on the temporal segmentation and categorization of vocalizations into discrete units. Unlike identifying species or individuals, where an objective measure exists of what animal produced a vocalization, the segmental units that comprise animal vocalizations are less well-defined. In comparison to human language, where linguistic units are determined based on their functional role, substantially less is known about the function each vocal unit plays in most species' communication, or even what should define the beginning and ending of a vocal unit (Kershenbaum et al., 2016; Mizuhara and Okanoya, 2020). Analyses of most animals, therefore, rely on easily discernible physical features of vocalizations. For example in songbirds, songs are typically segmented at different hierarchical levels, though no strict definition of these levels of organization are agreed upon by all researchers. Common units of birdsong are notes, corresponding to abrupt changes in frequency, syllables, defined by periods of silence surrounding continuous vocalizations, motifs, stereotyped repetitive combinations of acoustic elements, and phrases, series of stereotyped or commonly associated syllables. Despite the ubiquity with which these terms are used, most vocal units have not been validated in terms of the species' own perceptual system, and those that do, like the Bengalese finch 'syllable' (Mizuhara and Okanoya, 2020) call into question the commonly assumed role they play in communication. It is therefore ideal, but not always feasible, to validate decisions about vocal units based upon perceptual, physiological, or functional roles those vocal units play in the animal's communication (Suzuki et al., 2006; Kershenbaum et al., 2016). Still, most analyses of animal communications rely on human perceptual decisions at some level, whether it is to label discrete classes of birdsong phrases, or determine the representational space upon which an "unsupervised" learning algorithm will discretize units (discussed in Section 5).

When vocal units are defined and vocal classes are chosen, machine learning algorithms can be used to systematize and vastly speed up the classification and segmentation of vocal units. Most commonly, supervised recognition algorithms are used, where the algorithm explicitly learns to algorithmically map acoustic data to the researcher's labeling scheme. Over the past decades, vocalization labeling algorithms have paralleled those used in other acoustic domains, such as speech and music recognition. At present, tools rely on deep neural networks. The field of deep learning has changed rapidly over the past decade, with different architectures of neural networks quickly outperforming the previous architectures (Nassif et al., 2019). Prior to deep learning, automated birdsong element recognition relied on algorithms such as Hidden Markov Models (Kogan and Margoliash, 1998), support vector machines (Tachibana et al., 2014), template matching (Anderson et al., 1996), or k-Nearest-Neighbors labeling (Nicholson, 2016), following alongside contemporary speech recognition algorithms. Like sound event detection, current approaches tend to rely on recurrent and convolutional neural network architectures. TweetyNet (Cohen et al., 2020a), for example, uses a recurrent and convolutional architecture to capture complex spectro-temporal patterns over long timescales. Future advances in neural network architectures will likely continue to follow those in speech recognition, for example, using transformer architectures (Karita et al., 2019) as well as semi-supervised and unsupervised pre-training methods such as wav2vec (Schneider et al., 2019). One important divergence between speech recognition and animal vocalization classification is the reliance upon data availability, however. An ideal animal vocalization classifier works well on very small amounts of labeled data, requiring less experimenter time, whereas speech recognition systems tend to have an abundance of data available (though speech recognition for low-resource languages may be an area to watch).

A second approach to labeling vocalizations is to actively involve the experimenter in the algorithm via human-in-the-loop labeling (e.g., Wimmer et al., 2010; Kim and Pardo, 2018). Human-in-the-loop algorithms rely on a combination of supervised and unsupervised learning. Supervised learning comprises learning algorithms that are trained with labeled data, such as classification tasks. Unsupervised learning refers to algorithms that do not require supervised labels, such as dimensionality reduction. Human-in-the-loop algorithms leverage both, by proposing an initial coarse segmentation and/or labeling of the dataset through unsupervised learning, which the human then partially revises (e.g., merging or splitting putative classes of vocalizations) via a graphical user interface (GUI). The revised data is then re-processed by the algorithm and sent back to the user to revise, until the experimenter is content with the resulting labeled dataset. Using a combination of human expertise and machine processing enables quicker labeling of large bio-acoustics data with minimal human effort. A further discussion of unsupervised algorithms is discussed below in Section 5.

5. EXTRACTING RELATIONAL STRUCTURE AND CLUSTERING

Classifying vocal elements into discrete categories (e.g., "A," "B," "C," ...) is for many analyses a necessary abstraction that enables the analysis of recurring events. At the same time, this symbolic abstraction ignores acoustic relationships both within discrete element categories and between them. For example, in **Figure 3**, are the syllables of birdsong **Figure 3A** more similar to the syllables **Figure 3B** or the syllables **Figure 3C**? Determining the relatedness (or distance) between vocalizations can enable the quantification of how vocalizations change over time (Mets and Brainard, 2018; Kollmorgen et al., 2020), how vocal repertoires differ across individuals and species (Miller, 1979; Sainburg et al., 2020b), and map and visualize broad structure present in vocal repertoires (Sainburg et al., 2020b; Goffinet et al., 2021).

5.1. Operationalizing Relatedness

Given a dataset of vocalizations segmented into discrete units, relatedness is a measure quantifying the similarity of vocalizations relative to one another. The basis for operationalizing relatedness can utilize physical properties of signals, perceptual judgments, or behavioral and physiological responses to the signal. Most commonly, the relationships between vocal elements are computed on either spectrotemporal representations or on the basis of descriptive features of the vocalization, such as frequency modulation, fundamental frequency, and vocal envelope (Miller, 1979; Sainburg et al., 2020b; Goffinet et al., 2021).

How different aspects of the vocalization should weigh into a measure of similarity is non-trivial. No metric for similarity is objectively correct, even when metrics are derived purely from objective physical features. For example, what is the relative importance of a vocalization's duration vs. fundamental frequency in determining similarity? One ground truth metric for an algorithms judgement of similarity is its relationship with human's perceptual judgment of similarity (Tchernichovski et al., 2000), though there is no guarantee that these measures reflect the animal's own perception and physiology (Dooling and Prior, 2017). An ideal measure of similarity could be derived through careful experimentation gleaning the animal's own judgment of similarity (Kershenbaum et al., 2016), but in most cases, this task would be unfeasible and time-consuming. Even when performed carefully, perception varies from animal to animal, based upon experience (Lachlan et al., 2010).

In addition, when vocal features are continuous, accounting for differences in duration and temporal alignment requires consideration. Approaches vary from averaging over time (Elie and Theunissen, 2016), pooling using attention mechanisms (Morfi et al., 2021), using dynamic time warping (Kogan and Margoliash, 1998), and zero-padding (Sainburg et al., 2020b). Similarly, at least some animals rely on spectral shape rather than absolute pitch when recognizing acoustic objects (Bregman et al., 2016). A recent approach accounting for variability in frequency is dynamic frequency warping (Somervuo, 2019). Striking a balance between spectrotemporal tolerance and absolutely discounting spectrotemporal alignment can have substantial impact on the final measure of similarity.

5.2. Learning a Similarity Space

Once a metric for similarity is determined, that distance can be used to infer a structured representation of the relationships in a repertoire as a whole, providing a new representational space with which to quantify vocalizations.

Perhaps the most intuitive and pervading example of a learned embedding space for vocal similarity is multi-dimensional scaling [MDS, e.g., (Miller, 1979; Dooling et al., 1987; Morfi et al., 2021)]. Multi-dimensional scaling takes a graph of pairwise similarity measures between each vocalization in the dataset and attempts to find an embedding that best preserves the similarity structure of that graph. As the number of vocal elements in a dataset gets larger, however, the number of pairwise distances between vocal elements increases exponentially. This is computationally an issue because computing 10,000 pairwise distances between 100 elements is computationally feasible, but 10,000,000,000 pairwise distances between 100,000 elements is not.

Trying to preserve the pairwise relationships between every element in a dataset can also over-emphasize irrelevant relationships in vocal data. For example, if a bird's vocal repertoire comprises 10 motifs classes all produced with the same frequency, the vast majority of pairwise distance relationships computed (90%) will be between class, while only 10% of pairwise relationships computed will be within class. In many cases, both in animal communication and in dimensionality reduction more broadly, there is utility in weighing relationships between similar vocal elements more highly than relationships between less similar vocalizations. This contrast is defined in the dimensionality reduction literature as the emphasis of local vs. global structure (De Silva and Tenenbaum, 2002). Algorithms that attempt to preserve every pairwise relationship are called global dimensionality reduction algorithms, while algorithms that emphasize capturing relationships only to nearby points in dataspace (more similar vocalizations) are called local dimensionality reduction algorithms. In many vocalization datasets, emphasizing local over global structure better preserves categorical structure such as individual and call identity (Sainburg et al., 2020c; Goffinet et al., 2021; Morfi et al., 2021). A visual demonstration contrasting local and global structure preservation is given in **Figure 4A**. While global embedding algorithms like MDS attempt to preserve every pair-wise relationship, local algorithms preserve only local (e.g., nearest-neighbor) relationships, capturing more within-cluster structure. In **Figures 4B–G** an example is given with macaque coo calls, in which a local structure-preserving algorithm (UMAP, described below) more clearly pulls apart clusters corresponding to individual identity than MDS.

At present, the two dominant local dimensionality reduction algorithms are UMAP and t-SNE. UMAP and t-SNE differ in several important ways beyond the scope of this paper, but their key intuition and the steps underlying the algorithms remain similar: first, compute a (nearest-neighbors) graph of pairwise relationships between nearest neighbors in the original dataset (e.g., using Euclidean distance or an arbitrary similarity metric) then, embed that graph into an embedding space via gradient descent (Sainburg et al., 2021). UMAP, in particular, has been shown to capture complex structure in vocal repertoires such as differences in vocal dialect, vocal stereotypy, vocal element categories, inter-species similarity, and individual identity, in contrast to classic methods like MDS and PCA (Goffinet et al., 2021; Morfi et al., 2021; Sainburg et al., 2021).

One challenge with graph-based dimensionality reduction algorithms like MDS, UMAP and t-SNE is that they are non-parametric dimensionality reduction algorithms, meaning they do not learn the relationship between input data (e.g., a spectrogram of the vocalization) and their embeddings. Learning a parametric relationship between vocalizations and their embeddings allows a fast mapping between data and embedding, i.e., for applications that necessitate real-time feedback such as brain-machine interfacing.

The most common parametric dimensionality reduction algorithm is PCA, where a linear transform is learned between data and an embedding space. Similarly, neural networks such as autoencoders can be used to learn a set of basis features which can be complex and non-linear (Kohlsdorf et al., 2020; Sainburg et al., 2020c; Goffinet et al., 2021; Singh Alvarado et al., 2021). For example, an autoencoder trained on images of

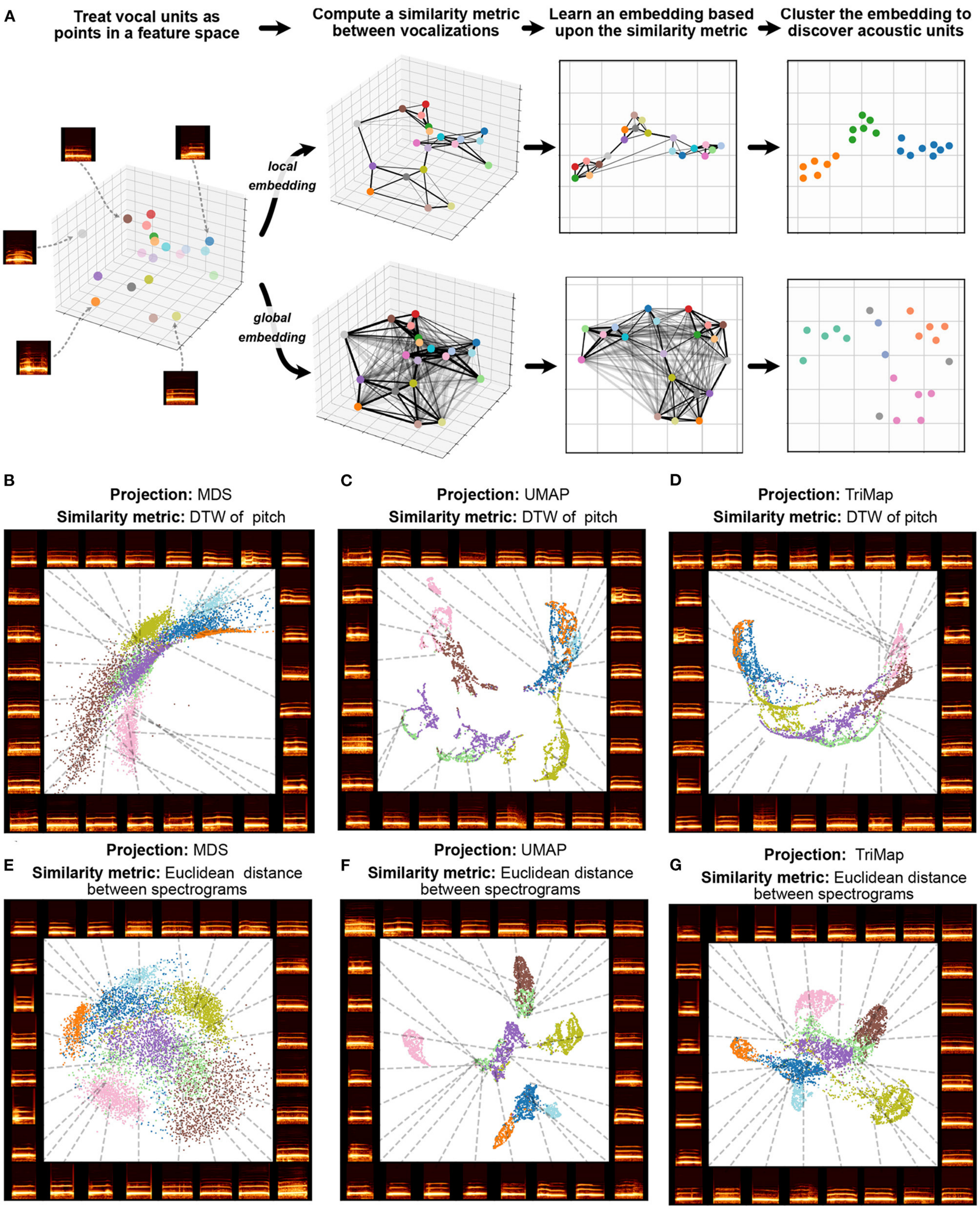

FIGURE 4 | Local and global embeddings. **(A)** The steps outlined in Section 5 exhibit the differences between the relationships preserved in local and global embeddings. **(B–D)** Projections of a dataset of macaque coo calls (Fukushima et al., 2015) using two similarity metrics (Dynamic Time Warping over frequency, and Euclidean distance between spectrograms) and three projection algorithms (Multidimensional Scaling, UMAP, and TriMap) **(E–G)**. Colors represent individual identity.

faces can learn to linearize the presence of glasses or a beard (Radford et al., 2015; Sainburg et al., 2018b, 2021). Autoencoders trained on animal vocalization data can similarly learn complex non-linear relationships in vocal data. In Section 8 we discuss how these complex learned features could be utilized in animal vocalizations to learn acoustic features such animal age, sex, and attractiveness, which can, in principle, be utilized for playback experiments.

A recent extension to UMAP, Parametric UMAP weds the advantages of UMAP with the parametric embedding of neural networks (Sainburg et al., 2021). Parametric UMAP acts by optimizing the UMAP loss function over arbitrary neural networks (e.g., convolutional recurrent networks were used with Cassin's vireo song in Sainburg et al., 2021) which can be balanced with additional losses such as MDS and autoencoding, to preserve additional global structure in UMAP projections. Parametric neural network-based approaches such as Parametric UMAP can also embed data on a similar timescale as PCA, enabling real-time applications, as opposed to non-parametric methods such as UMAP, t-SNE, and MDS.

Another class of neural network based dimensionality reduction algorithms rely on triplet-loss-based similarity preservation. Triplet-based embeddings have been used for birdsong for classification and embedding (Morfi et al., 2021; Renteria et al., 2021). Triplet networks learn an embedding space by sampling three types of vocal units: an anchor, a positive sample that is perceptually similar to the anchor point, and a negative sample that is perceptually distant from a vocal unit. The loss then encourages the positive sample to be pulled to the anchor, and the negative sample to be pushed further away. For example, Morfi et al. (2021) describe a triplet-loss-based network trained to produce vocal embeddings based upon a metric of perceptual distances. Like graph-based dimensionality reduction algorithms, triplet-loss-based embeddings rely on a pre-defined experimenter-determined notion of distance. Morfi et al. suggest a forthcoming animal-defined metric but in-lieu use a descriptive feature-based metric in the software Luscinia (Lachlan, 2007) which is correlated with human perceptual judgments of zebra finch song (Holveck et al., 2008).

5.3. Finding Latent Units Through Clustering

Learned embedding spaces enable the inference of broad structure acoustic structure from the statistics of vocalizations, enabling further downstream discovery of vocal units based upon distributional properties in embedding spaces (Kershenbaum et al., 2016; Sainburg et al., 2020b; Keen et al., 2021). Unsupervised clustering of vocal elements lies in contrast with supervised learning, where class labels are determined by experimenters, as in Section 4. Sainburg et al. (2020b) observe that labels obtained by clustering UMAP embeddings of Cassin's vireo and Bengalese finch syllables are more similar to experimenter labels than clustering PCA projections or spectrograms. Further, these latent projections capture additional acoustic and syntactic structure than the ground truth experimenter labels. In addition to acoustic structure, vocal elements can be clustered on the basis of syntactic organization. For example, incorporating transition information through Partially observable Markov Models (POMMs; Jin and Kozhevnikov, 2011) and Hidden Markov Models (HMMs; Katahira et al., 2011; Sainburg et al., 2020b) into a labeling scheme for birdsong better explains sequential structure than hand-labels or clustering without reference to temporal sequencing. An alternative approach is to perform clustering prior to embedding, directly upon the inferred relational graph (Frasier et al., 2017).

One challenge in unsupervised vocal unit discovery through methods such as UMAP embeddings is their reliance upon pre-defined vocal unit temporal boundaries. Although clustering on latent projections enables an unsupervised extraction of vocal categories from segmental units, latent projections rely on a pre-defined temporal segmentation of acoustic units from the vocal stream. In some species, atomic vocal units can be determined by clearly defined physical features of the signal, like long pauses between syllables, however, even in the case of clearly defined physical features, those units are not necessarily the base units of perception (Mizuhara and Okanoya, 2020). An open issue in vocal analysis is the unsupervised temporal segmentation of vocalizations into elements when clear physical boundaries are not available. This problem parallels both unsupervised speech discovery (i.e., ZeroSpeech), and the challenge of discovering behavior units in other areas of computational neuroethology (e.g., Motion Sequencing). In speech, phonemes are not clearly defined by physical characteristics, thus approaches for segmentation rely upon a combination of temporal and distributional information alongside imposed priors. Ongoing efforts in unsupervised speech segmentation, syllabic unit discovery, and word discovery can motivate parallel approaches in animal communication. In addition, physiological and kinematic measures such as articulation and breathing rate can aid in determining vocal boundaries. In computational neuroethology, new methods in tracking behavioral kinematics provide similar continuous behavioral datasets to those discussed in this paper (e.g., Wiltschko et al., 2015, 2020; Berman et al., 2016; Mathis et al., 2018; Pereira et al., 2019; Dunn et al., 2021; Marshall et al., 2021). For example, MoSeq (Wiltschko et al., 2015, 2020) discovers animal behavioral states using depth camera recordings of animals by fitting the behavioral data to an Autoregressive Hiden Markov Model. They find stereotyped sub-second mouse behavioral states, dubbed syllables, that underlie a syntax of behavior, much like birdsong. Communicative behavior is also not produced solely in the auditory domain. Improving methods for uncovering structure in animal behavior more broadly will facilitate research on the interaction between multi-sensory and multi-modal vocal behavior, like the dances that accompany many bird songs (Williams, 2001).

5.4. Data Augmentation

Another approach that is largely underutilized in bio-acoustic vocal recognition algorithms is data augmentation, an approach that is currently used in most state-of-the-art machine perception applications. In automatic speech recognition, for example, several current state-of-the-art approaches (e.g., Baevski et al., 2020; Gulati et al., 2020) use SpecAugment (Park et al., 2019) in which the classifier learns a policy of various augmentations such as warping and masking frequency channels in time. Lostanlen et al. (2019b) demonstrate the utility of augmenting bio-acoustics datasets with diverse background acoustics to facilitate better generalization across environments and conditions. Augmentation in settings where little labeled data are available has also proven successful on several semi-supervised learning benchmarks (e.g., Berthelot et al., 2019). One difficulty with performing data augmentation with bio-acoustics data, however, is the extent to which slight manipulations can affect the perceptual class that vocalizations fall into (Morfi et al., 2021).

6. INFERRING TEMPORAL AND SEQUENTIAL STRUCTURE

Identifying sequential organization typically relies upon the abstraction of vocalizations into discrete sequences of elements, effectively treating vocal data as corpora from which to perform symbolic analyses. Kershenbaum et al. (2016) identify six classes of models and analyses for analyzing temporal sequences: Markov chains, Hidden Markov Models, Network-based analyses, Formal grammars analyses, and temporal models. Analyses of temporal organization in animal communication has traditionally been largely influenced by Chomsky's hierarchy of formal grammars, with a focus on trying to understand what class of the Chomsky hierarchy animal's behaviors belong within (Hauser et al., 2002; Rohrmeier et al., 2015; Jiang et al., 2018; Morita and Koda, 2019). For example, Markov models, Hidden Markov Models, and Network models are all finite-state models in the Chomsky hierarchy.

6.1. Short-Timescale Organization and Graphical Analysis

Broadly, analyses over vocal organization can be broken down into two classes: analyses over short- and long-distance (i.e., short- and long-timescale) sequential organization. Short-timescale analyses are concerned with relationships between adjacent, or near adjacent elements in a sequence. Markov models, for example, capture short-timescale dynamics of vocal communication. A typical Markov model of birdsong is simply a

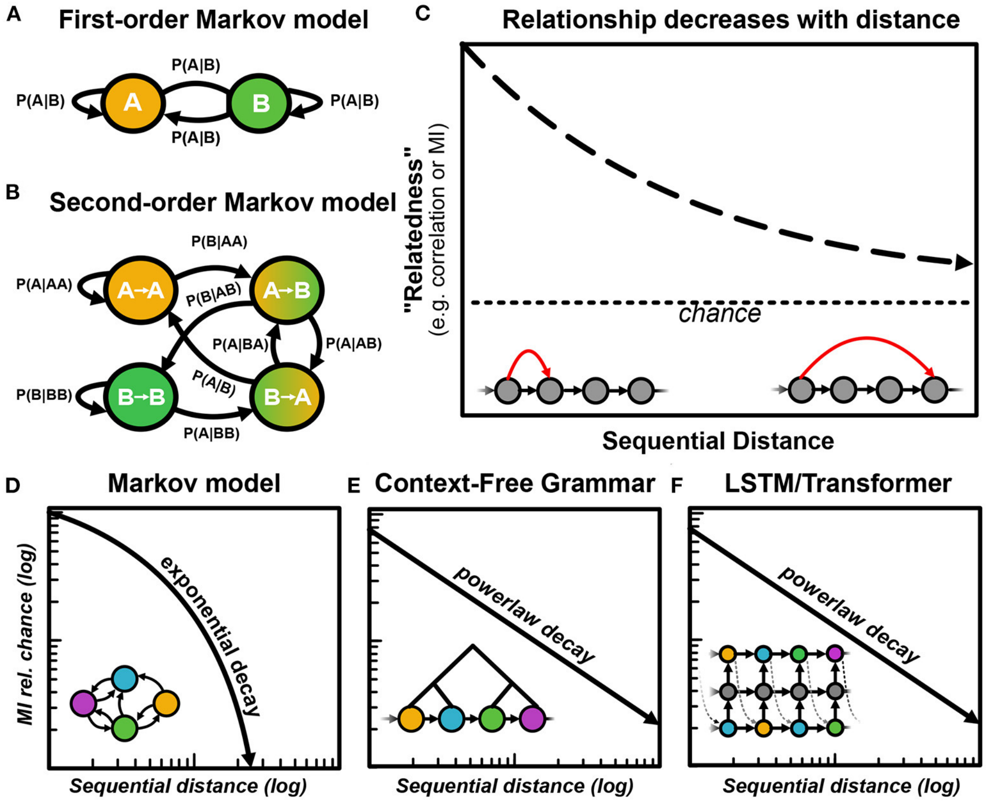

FIGURE 5 | Capturing long and short-range sequential organization with different models. **(A)** An example of a 2-state Markov model, capturing $2^2 = 4$ transitional probabilities between states. **(B)** An example second-order Markov model, capturing $2^3 = 8$ transition probabilities between states. **(C)** A visualization of the general principle that as sequential distances increase, the relatedness between elements (measured through mutual information or correlation functions) decays toward chance. **(D)** Sequences generated by Markov models decay exponentially toward chance. **(E)** Context-free grammars produce sequences that decay following a power law. **(F)** Certain neural network models such as LSTM RNNs and Transformer models produce sequences that also decay following a power law.

transition matrix representing the probability of transitions from each element to each other elements [e.g., $P(B|A)$ **Figure 5A**]. As Markov models increase in order, they become increasingly capable of capturing long-distance relationships, though high-order Markov models are rarely used in practice because of the number of parameters and amount of data needed to compute them (**Figure 5B**). Approaches such as Hidden Markov Models (Katahira et al., 2011) and Probabilistic Suffix Trees (Markowitz et al., 2013; Cohen et al., 2020b) can compute more succinct high-order Markov relationships, though the amount of data needed to capture these deeply contextualized relationships (e.g., $P(F|A, B, C, D)$) is still a limiting factor in capturing long-range organization with Markov models. Short-range relationships are also often captured graphically, treating any transition probability above zero as an edge in the graph. Graphical representations and metrics for vocal sequencing can explain general sequencing characteristics of vocalizations such as network motifs, communities, and clusters (Sasahara et al., 2012; Weiss et al., 2014; Hedley, 2016; Kershenbaum et al., 2016; Patricelli and Hebets, 2016).

6.2. Mutual Information and Long-Timescale Organization

Relationships that extend beyond adjacencies and over longer timescales are called long-range orlong-timescale relationships. For example, how related are two notes within a phrase, two phrases within a bout of song, or two bouts of song sung within a day?

Broadly, elements that are further displaced in a vocalization from one another tend to be less related. When two elements in a sequence are further apart, the relatedness between those two elements tends to be lower. For example, in birdsong, notes within a phrase are more likely to be related than notes separated by multiple phrases. The same is true of most sequential and temporal data: we can better predict what a stock price will look like tomorrow, than in 10 years. As we look further and further out into a sequence, the relatedness between elements will decrease alongside our ability to predict the future, until the relatedness approaches chance (**Figure 5C**). We can capture this relatedness over symbolic sequences using information theory. For example, given a sequence of discrete elements $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow f$, we can estimate the mutual information between pairs of elements at e.g., a distance of 2 elements ($a - c$, $b - d$, $c - e$, and $d - f$) or 3 elements ($a - d$, $b - e$, and $c - f$). As the distance increases between pairs of elements, we expect the relatedness (mutual information) to decay toward chance as a function of sequential distance.

We can estimate the extent to which a signal exhibits long-range relationships by computing how long the mutual information between pairs of elements remains above chance. Such approaches have been used variously across animal vocalization datasets in birds and whales (Suzuki et al., 2006; Sainburg et al., 2019). Similar approaches have also been used to observe long-range structure in animal motion ethology data, such as the long-range structure in *Drosophila* (Berman et al., 2016) motility.

6.3. Inferring Structure From Sequential Relationships

The shape of the decay in relatedness as a function of sequential distance can not only tell us about the timescales that vocal sequences are operating over but can also give indications about the structure underlying sequential organization. For example, sequences generated by Markov processes, such as finite-state grammars decay exponentially (Li, 1990; Lin and Tegmark, 2017) (**Figure 5D**). Intuitively, Markov models are memoryless; each state is dictated only by the set of transition probabilities associated with the previous state. As a result, the relatedness between states decays very quickly. When there are deep latent relationships present in the structure underlying the sequence, relatedness between sequentially disparate elements decays more slowly. For example, Probabilistic Context-Free Grammars can produce power-law relationships in mutual information as a function of sequential distance (Lin and Tegmark, 2017) (**Figure 5E**).

Characterizations of statistical relationships over abstracted discrete units enables comparative analyses across species because these measures make no assumptions about units or temporal organization underlying the signal. Characterizing correlations and information decay has an especially rich history in uncovering long-range structure dating back to Claude Shannon's original work (Shannon, 1951; Li, 1990; Lin and Tegmark, 2017). Language corpora such as speech and written text decay in information following the combination of a power-law over longer distances, and exponential decay over shorter distances, attributed to the finite-state processes underlying phonological organization (Sainburg et al., 2019) and the hierarchical organization underlying language at higher levels of organization such as syntax and discourse (Alvarez-Lacalle et al., 2006; Altmann et al., 2012; Lin and Tegmark, 2017; Sainburg et al., 2019). At the same time, however, young children's speech contains the same long-range information context before complex syntax is present in speech, indicating possible extra-linguistic mechanisms at play dictating these long-range statistical relationships (Sainburg et al., 2020a). Long-range mutual information decay and correlations have also been demonstrated that in animals such as songbirds (Sainburg et al., 2019) and humpback whales (Suzuki et al., 2006), extending over minute- and hour-long timescales. In particular, birdsong exhibits similar exponential short-range and power-law long-range mutual information decay to human speech, indicating potential parallels in the mechanisms governing how patterns of vocalizations are temporally sequenced. Similar observations in non-vocal behavioral sequences (Berman et al., 2016; Sainburg et al., 2020a) also exhibit these long-range sequential organizations, suggesting similarities in latent dynamics that facilitate long-range statistical relationships.

It is tempting to suggest that these parallels suggest shared underlying structure generating mechanisms, such as universals in the hierarchical organization of behavior (e.g., Lashley, 1951; Dawkins, 1976), though we should be wary of making any extended inferences based upon the observation of long-range information decay. For example, we can infer that power-law sequential relationships are produced by non-Markovian mechanisms because the decay is not exponential. However, the set of generative mechanisms that can produce power-law relationships in signals is not understood well enough to attribute the origins of these relationships to, for example, any specific class of formal grammar. Power-law mutual information decay in signals can also be drawn simply from coupling vocal or behavioral 1/f noise found in exogenous environmental signals.

While it is well-acknowledged that many animal vocalizations are organized hierarchically (Dawkins, 1976; Rohrmeier et al., 2015), the implications of that hierarchy in terms of underlying cognitive and physiological mechanisms are not well understood. For example, on very short timescales, birdsong motor sequencing is dictated by a hierarchical cascade of motor programs running originating in the premotor region HVC eventually ending in motor output (Doupe and Kuhl, 1999). Recent physiological evidence shows that these high-level nuclei also contain information about future states displaced from current vocalizations as well (Cohen et al., 2020b), though the mechanisms by which those relationships are learned, maintained, and ultimately dictate behavior are not yet clear.

Although we do not have access to the mechanisms underlying the observed long-distance relationships in vocal and non-vocal behavioral sequences, we do know that many vocal and behavioral sequences cannot be well-captured by Markovian models, thus alternative methods for modeling, characterizing, and forming hypotheses about the long-range organization in behavioral sequences are crucial to furthering our understanding of long-range structure in behavioral sequences. One promising approach is the use of deep neural network models such as RNNs and transformer networks (Tran et al., 2018; Morita et al., 2020). Unlike Markov models, recent neural network models like RNNs and transformer models do capture power-law relationships in sequential data (**Figure 5F**) (Lin and Tegmark, 2017; Shen, 2019). In language, transformer networks, in particular, have changed the landscape of natural language processing by capturing deeply contextual and complex implicit relationships in linguistic sequences. In birdsong, the same approaches show promise (Morita et al., 2020). For example, Morita et al. (2020) train a transformer network on Bengalese finch song and find that it captures long-range dependencies extending well beyond that of a Markov model. Like modeling language sequences, however, neural-network-based approaches suffer from the same issues of being black box and providing little explanatory power over the sequential structure they learn. In addition, the amount of data required to train a model to capture complex sequential dependencies is vast. Although the number of parameters does not increase exponentially with the amount of context the model captures (as in Markov models) state-of-the-art transformer models have billions of parameters requiring training data comprised of billions to trillions of characters (Brown et al., 2020). In language, the dataset size needed to train transformer models scales with the number of parameters in the model to prevent overfitting (Kaplan et al., 2020). When dataset sizes are smaller, LSTM RNNs perform better than more state-of-the-art Transformer language models (Ezen-Can, 2020), though transformers allow you to explicitly specify the length of temporal context allowed in the model, making them an attractive model for controlling context when generating vocal sequences (Morita et al., 2020). Although non-human animal vocalization repertoires are smaller and syntactic organization is less complex than language, birdsong analyses relying on language models will need to address the same challenges.

Neural network-based models also provide the ability to capture temporal dependencies that mutual information and correlation functions do not. Mutual-information-function-based and correlation-based analyses compute relationships between vocal elements as a function of sequential distance, ignoring any temporal relationships between disparate elements. This is both a benefit and a shortcoming of correlation methods. Ignoring intermediary temporal relationships enables the characterization of structure at temporal distances without having to additionally model higher-order combinatorial relationships [e.g., $P(F|A))$ vs. $P(F|A, B, C, D)$]. For the same reason, mutual-information-function-based and correlation-based analyses are coarse descriptions of temporal structure and miss the full temporal dynamics of the signal that neural-network-based models can capture (Morita et al., 2020).

7. SYNTHESIZING VOCALIZATIONS

Although the methods discussed in Section 5 allow us to learn representational spaces of animal vocalizations, providing new ways to infer structure in vocal repertoires, analyses on vocal signals alone lack grounding in animal behavior, perception, and physiology. In this section, we give an overview of methods for synthesizing animal vocalizations as a means to systematically control vocalization stimuli and relate vocal representations to physiology and behavior.

An ideal model for vocal synthesis exhibits several features: (1) it can model the entire vocal repertoire of a species or multiple species, (2) the parameters of the model can be related to physiological properties of the vocalizing species, and (3) the parameters of the model can be explained in terms of understandable features (i.e., it is not a black box algorithm). Throughout this section, we find that current synthesis algorithms have tradeoffs in how they balance aspects of these ideals.

One reason to systematically synthesize animal vocalizations is to probe their perceptual and physiological representations of vocal space, for example, determining how animals categorically perceive the difference between two categories of vocal units (Nelson and Marler, 1989). Traditionally, categorical perception in animals has been studied on the basis of human speech sound stimuli (Sinnott et al., 1976; Kuhl and Miller, 1978; Kuhl and Padden, 1983). Even with speech, however, the features that can be manipulated are limited. Recently methods in machine learning have furthered our ability to manipulate complex non-linear speech features substantially. These same approaches

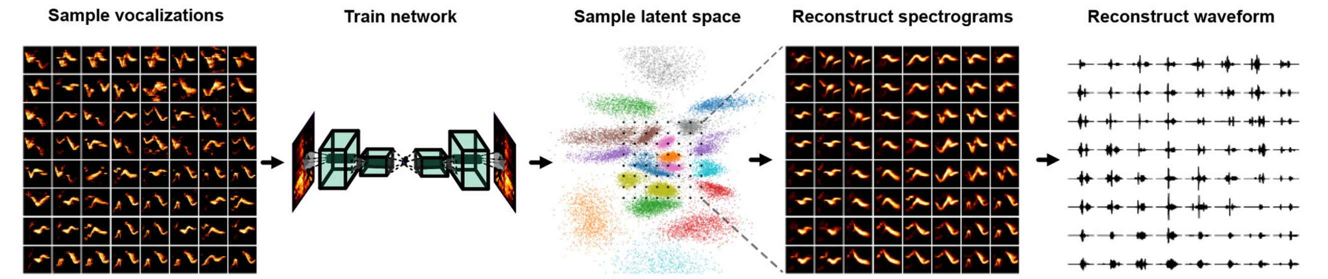

FIGURE 6 | Steps involved in synthesizing vocalization from a Variational Autoencoder (VAE) trained on spectrograms.

can often be applied to animal communication (Anikin, 2019; Sainburg et al., 2020c).

7.1. Source-Filter Models

Source-filter models have their origins in vocoding speech (Dudley, 1939), but have been used in numerous animal vocalization synthesis paradigms (DiMattina and Wang, 2006; Chakladar et al., 2008; Arneodo and Mindlin, 2009; Furuyama et al., 2017). Source-filter models decompose vocalizations into the source of the voice and filters (Kawahara, 2006). For example, the STRAIGHT algorithm (Kawahara et al., 1999; Kawahara, 2006) has been used to morph between macaque monkey vocalizations for investigations of monkey and human perception and physiology related to categorization (Chakladar et al., 2008; Furuyama et al., 2017). STRAIGHT breaks down the macaque vocalization into the fundamental frequency (the source) and its harmonics from higher-resonant or formant frequencies (the filter) (Chakladar et al., 2008). It then uses landmarks based upon these estimated parameters from the two sounds being morphed and interpolates between them to generate the morph stimuli. Furuyama et al. (2017), for example, used this method to parametrically vary generated morphs based on source and filter properties to determine the features macaques use to distinguish between conspecifics. Soundgen (Anikin, 2019) is a recent open-source GUI-based web app for R that is designed to synthesize nonverbal vocalizations using a source-filter model, including animal vocal signals such as birdsong and primate vocalizations. Related source-filter models have been developed to synthesize birdsong based upon underlying physiological mechanisms (Fee et al., 1998; Sitt et al., 2008, 2010; Arneodo and Mindlin, 2009; Arneodo et al., 2012). Recently, Arneodo et al. (2021) demonstrated that synthetic source-filter models can be coupled with neural recordings accurately reconstruct vocalizations from neural data alone. One drawback of source-filter models is the difficulty with which they can be fitted to the diversity of non-human vocalizations that exist. For example, the source-filter models of birdsong described above can well describe the dynamics of zebra finch song, but not the dual-syringeal dynamics of European starling song. Without reference to explicit hypotheses about underlying production mechanism, HMM based source-filter approaches provide one potential solution to this problem birdsong (Bonada et al., 2016).

7.2. Neural Network Models

An alternative approach to synthesizing animal vocalizations is the use of neural-network-based synthesis algorithms. These neural-network-based algorithms can be used to sample directly from the learned representational spaces described in Section 3. A simple example is autoencoder-based synthesis (**Figure 6**) (Sainburg et al., 2018a; Zuidema et al., 2020). Autoencoders can be trained on spectral representations of vocal data, and systematically sampled in the learned latent space to produce new vocalizations. Insofar as the neural network or latent projection can learn to represent the entire vocal repertoire, the entire vocal repertoire can be sampled from. In addition to sampling vocalizations from a latent distribution, vocal features can be manipulated in latent space. Well-defined latent spaces and higher-dimensional latent projections can learn to linearize complex non-linear reltionships in data. For example, in pictures of faces, the presence of a glasses, hair color, and the shape of a person's face can all be manipulated as linear features (Radford et al., 2015; Sainburg et al., 2018b, 2021). With more complex features, such as the attractiveness of a call or the age of the vocalizer, a promising avenue for future research would be to synthesize vocalizations, varying these complex non-linear features for playback studies.

Like most areas of deep learning, substantial progress has been made on the task of audio synthesis in the past few years. Basic methods comprise autoencoders (Engel et al., 2017; Kohlsdorf et al., 2020; Sainburg et al., 2020c), Generative Adversarial Networks (GANd) (Donahue et al., 2018; Engel et al., 2019; Sainburg et al., 2020c; Tjandra et al., 2020; Pagliarini et al., 2021) and autoregressive approaches (Mehri et al., 2016; Oord et al., 2016; Kalchbrenner et al., 2018; Prenger et al., 2019). One advantage of GAN-based models is that their loss is not defined directly by reconstruction loss, resulting in higher-fidelity syntheses (Larsen et al., 2016). Typically, approaches for synthesizing vocalizations based on neural networks rely on treating magnitude spectrogram like an image, training a neural network architecture in the same manner as one would an image, and finally inverting the sampled spectrogram into a waveform (Sainburg et al., 2020b; Zuidema et al., 2020; Pagliarini et al., 2021). When synthesizing vocalizations from neural networks trained on the magnitude spectrogram, the estimation of phase is necessary to invert the spectrogram into a waveform signal for playback. The de-facto algorithm for spectral inversion has been Griffin and Lim (Griffin and Lim, 1984), though several recent approaches have been shown to improve over the Griffen-Lim algorithm recently (Prša and Rajmic, 2017; Masuyama et al., 2019). An alternative to Griffen-Lim inversion is to train neural networks to invert spectrograms either directly in the neural network architecture (Kumar et al., 2019), or perform inversion in a second network (Masuyama et al., 2019). Spectrogram-based audio synthesis can also be sidestepped entirely, training the network directly on waveform (Mehri et al., 2016; Oord et al., 2016; Engel et al., 2017).

7.3. Sound Texture Synthesis

Another approach to sound synthesis is the synthesis of sound texture (Saint-Arnaud and Popat, 1995; McDermott et al., 2009). For example, McDermott et al., (McDermott et al., 2009) propose an approach that relies on computing a set of statistics over stationary elements of sounds, and synthesizing new sounds based upon the computed statistics. By manipulating or interpolating between sound statistics, they synthesize new sound textures. One application, for example, is to manipulate components of sound textures for stimulus playback to determine what sound texture statistics listeners rely upon for recognition (McDermott and Simoncelli, 2011).

7.4. Generating Sequences

A parallel approach to synthesizing vocalizations is to generate vocal sequences from symbolically labeled vocal elements. Synthetic song sequences can be used to understand how animals process and represent temporal and sequential organization. For example, can a songbird differentiate between sounds generated using different underlying models of song syntax? Traditional approaches to song sequence generation rely upon the explicit, hand-crafted, generation of artificial grammars for playback studies. By crafting artificial grammars that differ in underlying structure, such as belonging to different classes of the Chomsky hierarchy (Gentner et al., 2006; Fitch and Friederici, 2012; Kershenbaum et al., 2014), playback studies can be used to determine whether animals can learn these grammars. A number of challenges exist with artificial grammar learning studies, however. One such challenge is the difficulty in crafting sequences that can exclusively be learned by inferring the structure that generated them, for example, making it impossible for the animal to learn by brute-force memorizing every sequence (Fitch and Friederici, 2012). When using artificial grammars, computational and modeling considerations aid in forming hypotheses about how generated grammars can be used. In the context of the neuroethology of vocal communication, these cognitive models can be related to physiological measures (Zuidema et al., 2020). An additional challenge with artificial grammar learning is constructing sequences that are structured in a similar way to natural and behaviorally relevant signals to the animal. For example, artificial grammar studies usually rely on short sequences modeled after human language syntax, rather than the animal's own communication systems. Because the task of generating vocal sequences is performed over symbolic representations of syllables, generating vocal sequences can be performed using the same methods as in text or musical note generation. These approaches can range from generating sequences using Markov models of various orders, to explicitly modeling hierarchical organization in the signal generation algorithm (Roberts et al., 2018).

8. MAPPING VOCAL COMMUNICATION TO PERCEPTION, BEHAVIOR, AND PHYSIOLOGY

The methods discussed here provide a framework to develop a set of constrained spaces from which to understand and model vocal behavior in relation to perception, production, and physiology. Perceptual or relational vocal spaces, such as UMAP projections of spectrograms, provide a low-dimensional space that can be used to infer structure in vocal repertoires. Likewise, symbolic abstractions of vocal behavior to large corpora provides a categorical representation in which vocal behavior is seen as sequential actions on those category sets. In both cases, the methods provide a constrained behavioral representation for physiological analyses.

8.1. Brain-Computer Interfacing

One of the primary challenges facing the field of brain-computer interfacing is scaling up from simple behavioral spaces like moving a cursor on a screen to complex behaviors (Gao and Ganguli, 2015). A clear advantage to the approaches discussed in Section 5 is that we can learn to bring complex vocal behavioral spaces into a compressive low-dimensional behavior spaces, even without a prior model of the structure in that space. For example, Arneodo et al. (2021) find that directly predicting the acoustic structure of zebra finch song from neural data does not perform as well as predicting the parameters of a low-dimensional biophysical model of song production. In the many species in which we do not have access to a biophysical model of vocal production, learned acoustic spaces may be a viable alternative. In contrast, the current state-of-the-art vocal prostheses for speech bypass biophysical models, directly predicting sentences (i.e., symbolic sequences) with the aid of language models (i.e., a sequence model) (Moses et al., 2021). Such methods do not capture important extra-linguistic information such as emotional tone and stress. In future work, a clear pathway forward is to develop BCI models that can both capture symbolic organization aided by sequential models, as well as within-symbol variability in the acoustic signal.

8.2. Vocal Production

Songbirds as a model for systems neuroscience are perhaps best known for the role they play in our understanding of vocal learning (Doupe and Kuhl, 1999). In addition to songbirds, rodent and non-human primate vocal behavior are becoming increasingly prominant models in the neuroscience of vocal production. In non-human primates, recent evidence has suggested some degree of constrained vocal learning in some species (Fischer and Hammerschmidt, 2020). In rodents, recent focus has been placed upon variability and structure mouse in ultrasonic vocalizations (USVs) (Holy and Guo, 2005; Arriaga et al., 2012; Petkov and Jarvis, 2012), singing mice have emerged as a physiological model of turn-taking (Okobi et al., 2019), and the cultural transmission of vocal dialect has been observed in the naked mole rat (Barker et al., 2021). In each of these cases, quantification of how vocalizations vary as well as relationships between vocalizations (either within individual, between conspecifics, or from tutor to pupil) is integral to understanding how we learn to navigate vocal space. For example, Kollmorgen et al. (2020) use nearest-neighbor graphs and t-SNE projections to quantify and visualize the developmental trajectory of zebra finch song during vocal learning. For each syllable, they compute a nearest-neighbors graph based metric termed the "neighborhood production time," which quantifies the developmental time point at which similar (neighboring) syllables were sung. For example, a syllable song on day 45 might have 10 neighbors, sung on days between day 40 and 50, comprising its neighborhood production times. Syllable renditions that are neighbors with predominantly future syllables are deemed anticipations, while syllable repetitions that are neighbors with predominantly past syllables are deemed regressions. They observe that day-by-day, zebra finch songs gradually moves along a constant vocal learning trajectory, but anticipations and regressions differ in how they are consolidated overnight.

The number of neurons we can simultaneously record from physiologically has increased the dimensionality of neural datasets substantially over the past decade, making methods for dimensionality reduction on neural signals such as spike trains increasingly necessary for neural data analysis and opening the door to computational methods that directly link the latent representations of behavioral and neural datasets. Population modeling approaches such as LFADS (Latent Factor Analysis via Dynamical Systems; Pandarinath et al., 2018) reduce large population spiking datasets into low-dimensional trajectories, similar to the approaches discussed here with vocal signals. In the case of LFADS, these embeddings are performed over single trials using a recurrent autoencoder. One promising direction for computational neuroethology is learning the relationship between latent behavioral states and latent physiological states. By developing tools that allow us to learn the relationship between physiological and behavioral representations, we hope to untangle how, for example, movements in behavioral space reflect changes in physiology, and vice-versa. Singh Alvarado et al. (2021) developed a joint encoding model in which they used variational autoencoders to learn a joint latent representation of spectrograms of zebra finch song, and corresponding ensemble neural activity of spiny neurons in songbird basal ganglia (average calcium fluorescence of around 60 ROIs, or putative neurons, per bird). In a series of experiments leading up to this joint mapping, Singh Alvarado et al. demonstrated that Area X spiny neurons are involved in the regulation of vocal variability; exhibiting suppressed activity during female-directed song and enhanced activity during practice. Using the joint vocal-neural latent mapping, they were able to uncover the mapping between specific features of song and variants present in neural ensemble activity. In **Figure 7** we outline several similar maps between behavior, perception, and neural dynamics. Singh Alvarado et al.'s work exhibit that one such latent map, a vocal-motor mapping between motor physiology and vocal behavior, can uncover complex and detailed relationships that traditional methodology cannot. Similar mappings between the physiology, perception, and behavior of sender-receiver dynamics (e.g., **Figure 7**) are also well poised to benefit from emerging latent approaches.

The physiology of vocal syntax is another area poised to benefit from computational ethology. One example is the role of

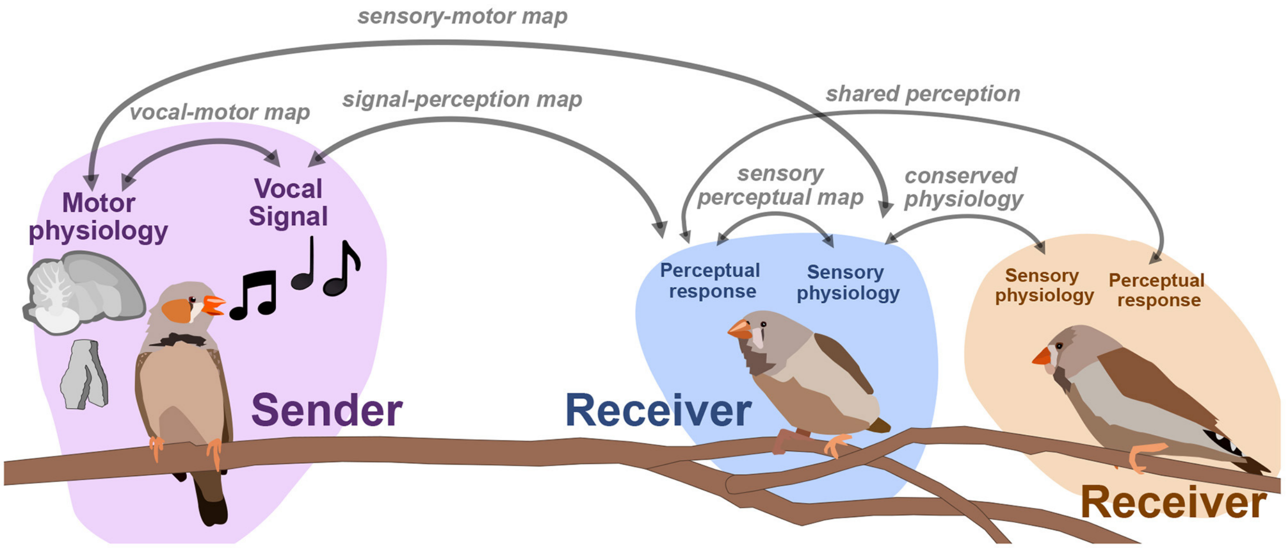

FIGURE 7 | An outline of mappings between perceptual, acoustic, and physiological signals. One focus in sensory and motor neuroscience is to learn the relationships between signals, perception, and physiology.

the songbird premotor nucleus, HVC, in encoding song syntax. Birdsong has a long history of being described sequentially in terms of low-order Markovian transitions between song elements. HVC's role in song syntax, until recently has been described exclusively in terms of these low-order transition statistics (Fujimoto et al., 2011). In a recent example, however, (Cohen et al., 2020b) made use of an automated birdsong labeling paradigm and high-order sequence model to observe 'hidden neural states' encoding sequentially displaced (i.e., high-order) transitions in the premotor nucleus HVC of canaries. To identify non-adjacent dependencies in the song, they used a Prediction Suffix Tree (Markowitz et al., 2013), which can capture high-order Markovian relationships in the song syntax. Prediction Suffix Trees have previously been used to observe long-range dependencies up to the 7th order in canaries (Markowitz et al., 2013). While birds were singing, Cohen et al., used a miniature microscope to image neurons from HVC, a region involved in the songbird vocal motor circuit. They observed that HVC ROIs were locked to individual song-phrases and transitions, and that this phrase locking is modified by non-adjacent context, displaced by several phrases and seconds. As more recent approaches give access to larger datasets enabling the identification of longer-range dependencies in birdsong, it is currently not clear whether we have yet found an upper bound on the sequential displacement of long-range representations of vocal syntax in physiology.

Outside of songbirds and mammals, male *Drosophila* song, although not strictly vocal, is temporally patterned and driven by both environment and internal states (Coen et al., 2014; Calhoun et al., 2019). Calhoun et al. (2019) jointly model and uncover relationships between temporal song structure and interactions with a potential mate. Using a sequential model (an HMM-GLM hybrid) they demonstrate that song patterning is underlined by three hidden sensorimotor states, under which male's song productions differ in their relationship to female behavior. Using optogenetic activation, they were then able to identify neurons involved in switching between these sensorimotor states.

8.3. Vocal Perception

Similar to vocal production, latent and sequential models are promising avenues for better understanding cognitive and physiological underpinnings of vocal perception. In songbirds, primates, and rodents, many foundational studies of auditory categorical perception, perceptual decision making, and their underlying physiology rely upon either relatively simple stimuli such as tones or complex stimuli like human speech (Kuhl and Miller, 1978; Kuhl and Padden, 1983; Russ et al., 2007; ten Cate, 2014; Xin et al., 2019). Categorization in these stimulus spaces are attractive because they are well-characterized and understood. Across species, however, neural responses are often tied to complex and more behaviorally-relevant acoustic phenomena such as recognizing and discriminating between conspecific vocalizations (Bailey et al., 2002; Liu et al., 2019). When the acoustic features underlying vocal repertoires are simple and known, categorical stimuli can be selected directly based upon those features. For example, Lachlan and Nowicki (2015) manipulate a single dimension, the duration of swamp sparrow notes, to determine how notes are categorically perceived in different sequential contexts. In speech, voice onset time (VOT) is a similar single-dimension commonly used for categorical perception paradigms (Liberman et al., 1957). However, it is rarely the case that categorical perception is driven by a single dimension. Thus, building stimuli in more complex feature spaces will be necessary to untangle the relationship between vocal features, perception, and physiology. When biophysical models of vocal structure exist, species relevant stimuli can be generated using biophysical parameters (Arneodo and Mindlin, 2009). When the underlying acoustical structure of a vocal repertoire is more complex and biophysical models of vocalizations have not been defined, neural-network synthesized vocalizations are an attractive alternative. For example, as discussed above, birdsong can be synthesized with neural networks for physiological and perceptual playback studies to determine perceptual similarity between syllables or learn categorical boundaries between song-morphs (Sainburg et al., 2018a; Thielk et al., 2018; Zuidema et al., 2020). By systematically controlling the signal space of a vocal repertoire, we can systematically explore how changes in that space relate to changes in physiology.

Algorithmic approaches are similarly well poised to aid in our understanding of how vocal sequences are maintained and represented. Sequence learning research in human and non-human primates is largely dominated by artificial grammar learning (AGL) research, an umbrella category that comprises several different forms of sequence learning ranging from hierarchically nested tree structures to transitional probabilities (Dehaene et al., 2015). Artificial grammar learning studies aim to determine what structures animals (and humans) are capable of learning, what cognitive mechanisms underlie grammar induction, and what physiological systems underlie those cognitive mechanisms. In the domain of primate sequence learning, neural pathways are generally conserved between humans and non-human primates and involve the ventral regions of cortex (Wilson et al., 2017). Determining an appropriate stimuli set is requisite for developing an AGL paradigm. Latent representations of vocalizations can aid in choosing stimuli from a well-defined stimulus space. For example, when choosing a stimulus set for an A^nB^n grammar, it is desirable depending on the goal of the task to ensure that the constituent vocalizations comprising A and B belong to equidistant or separate clusters in acoustic or perceptual spaces (Zuidema et al., 2020).

While artificial grammar learning has also played a prominent role in birdsong sequence learning (ten Cate and Okanoya, 2012), the structure underlying an animal's own vocal syntax provides an opportunity to study the neural and cognitive underpinnings of a more ethologically-relevant complex sequential structure. Despite the important role vocal syntax production has played in establishing birdsong as a model in systems neuroscience, a surprising gap exists in our knowledge of the physiological circuits underlying how syntactic information is recognized and sequentially integrated when listening to song. Songbird vocal communication contains often very complex syntax that can be structured over long timescales comprising often tens to hundreds of unique, stereotyped vocal units (Cody et al., 2016). Conspecifics pay attention to the structure of that song. Abe and Watanabe (2011) developed a habituation/dishabituation paradigm with Bengalese finches alongside immediate early gene (IEG) expression and lesioning experiments to explore the role of song nuclei on the recognition of grammatical sequences. They found that IEG expression increased when presented with non-conforming/nonpredictive strings in the nuclei LMAN, a basal ganglia output nuclei characterized by recurrent loops that is also involved in vocal learning (Bottjer and Altenau, 2010). Abe and Wantenable then lesioned LMAN and measured song discrimination with their habituation paradigm. They found that discrimination was disturbed in birds where LMAN was lesioned, implicating LMAN in the ability to discriminate syntactic song. How syntactic information is learned, integrated, and maintained in LMAN and associated striatal regions of songbird brain are still open questions.

In contrast to the auditory domain where little is known about syntactic integration, a neural correlate for complex and abstract information integration, NCL, has been well established and characterized in songbird vision with pigeons and corvids (Kröner and Güntürkün, 1999; Güntürkün, 2005). Strong parallels exist between NCL and the primate prefrontal cortex, which is involved in sequence learning. NCL has variously been associated with rule learning (Veit and Nieder, 2013), numerosity (Ditz and Nieder, 2015; Wagener et al., 2018), directed forgetting (Rose and Colombo, 2005; Milmine et al., 2008; Helduser et al., 2013), choice behavior (Kalenscher et al., 2003), working memory (Diekamp et al., 2002; Rinnert et al., 2019), sequence learning (Helduser and Güntürkün, 2012), and reward learning. Anatomically and neurochemically NCL also exhibits strong parallels to the primate prefrontal cortex. NCL is characterized by similar circuitry from auditory and dopaminergic afferents, as well as multi-sensory projections (Kröner and Güntürkün, 1999; von Eugen et al., 2020). Surprisingly, however, an auditory equivalent to the visual working-memory region in NCL has not been found, though they have been observed in the multi-sensory audio-visual integration and association (Moll and Nieder, 2015, 2017). Birdsong is well poised as a signal to be a model of vocal syntax perception, To establish this model, however, it will be imperative to uncover the systems in songbird brain related to working memory and temporal context integration in song. NCL appears to be a likely candidate for processing syntactic vocal signals, though, as yet, this has not been found to be the case.

Although mouse USVs do not appear to contain temporal structure to the same extent as songbirds, mouse USVs are temporally organized (Castellucci et al., 2018) and female mice also show preference for more complex syllables and sequences (Holy and Guo, 2005), making mouse USVs another potential target for the study of syntactic and sequential integration in vocal perception.

9. DISCUSSION

This review covers emerging approaches in the computational neuroethology of vocal communication enabling researchers to engage with large and diverse datasets of vocal signals and to represent them in computationally tractable frameworks.

We started by discussing techniques to process and represent acoustic signals. We then discussed how to parse complex vocal datasets into species, individuals, and discrete vocal elements. Next, we discussed how relational structure can be extracted from vocal signals, how these signals can be clustered in learned latent spaces, and how these latent spaces capture different aspects of the information contained within the underlying signals. We then discussed how temporal structure can be inferred from vocal units, including emerging work on the non-Markovian dynamics underlying vocal behavior. In the next section, we discussed how vocalizations can be synthesized for use in playback experiments that allow an unprecedented degree of control over non-linear and complex vocal feature spaces. Finally, we discussed how these approaches are being applied to the field of neuroethology and emerging frameworks for understanding vocal signals and their underlying physiology.

The methods discussed here provide a promising avenue for a broader, more diverse, and larger-scale neuroethology of vocal communication, than the research practices that have dominated the past several decades, and hold the promise of expanding both the breadth and depth of our understanding. Instead of focusing on a small number of model species, new computational techniques provide a framework for studying vocal behavior across a wide range of animals. While much of vocal neuroethology has recently focused on songbirds and mammals the techniques discussed here are equally applicable to the abundance of other species studied in bioacoustics and behavioral ecology including fish, amphibians, and insects. Even within songbirds, research on vocal learning in songbirds has ignored the majority of species, female birdsong, and most call types (Loo and Cain, 2021). Likewise, because these new computational methodologies can often deal with unstructured data, they enable us to expand beyond simplified, isolated behaviors in controlled environments to more natural or naturalistic behavioral contexts where dynamics involving multi-modal integration and multi-animal social interactions arise. As we capture increasing levels of detail in behavior, our understanding of its sophistication naturally follows. Already, these new computational framework have revealed deep structure in the sequential organization of communication, where large-scale datasets of both symbolic sequences, and latent projections that capture rendition-to-rendition variability, have enabled quantitative analyses of rare (but perhaps meaningful) events, such as long-range syntactic organization. Together, all of these approaches point toward a new framework, in which complex and non-linear behavioral and physiological signals can be represented in compressive and tractable spaces that can capture the complex dynamics and relationships in the increasingly rich datasets available to researchers.

As with any powerful tool, these techniques require careful consideration when put into practice. Broadly, automation and machine learning in data analysis can be fraught with unexpected complications and confounds that may be hard to spot. For example, automating the labeling of large datasets of birdsong syllables can speed up the task of labeling by days, weeks, or months, but can also leave the experimenter with less intuitive knowledge of the animal's vocal repertoire, resulting in a loss of domain knowledge. As we have noted elsewhere (Sainburg et al., 2020b), when domain knowledge is available it should be integrated with one computational approach. Another potential pitfall (and a source of much needed research effort) is in understanding the structure of the latent manifolds that are yielded in many of the described methods. In particular, non-linear latent modeling techniques like UMAP or neural networks can capture complex relationships in vocal data, but interpreting these projections requires an understanding of how data are represented within the geometry of the latent space. For example, UMAP captures primarily local structure in datasets that are present in nearest neighbor graphs, meaning that the relative distances of vocal elements have no explicit relation to the data, as is the case in PCA for example.

Attending to the cautions of computational abstraction, the approaches discussed in this manuscript provide a framework from which to quantify vocal signals that promises to yield important new insights into vocal behavior and neurobiology. These approaches enable neuroethologists to project vocalizations onto low dimensionaland latent manifolds, visualize and quantify the transitional structure and information decay of vocal syntax, and map vocal and neural repertoires into shared neural spaces for functional representation action. As the richness of datasets grow to capture more of the complexities of behavior and physiology, methods and frameworks for modeling and inferring structure in ethological data are increasingly necessary for hypothesis formulation and testing. The methods and frameworks discussed in this review parallel and supplement those in the broader field of computational neuroethology.

AUTHOR CONTRIBUTIONS

This manuscript was written by TS and TG. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a CARTA Fellowship to TS, NIH 5T32MH020002-20 to TS, and 5R01DC018055-02 to TG.

REFERENCES

Abe, K., and Watanabe, D. (2011). Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nat. Neurosci*. 14, 1067–1074. doi: 10.1038/nn.2869

Adi, K., Johnson, M. T., and Osiejuk, T. S. (2010). Acoustic censusing using automatic vocalization classification and identity recognition. *J. Acoust. Soc. Am*. 127, 874–883. doi: 10.1121/1.3273887

Alonso, J. B., Cabrera, J., Shyamnani, R., Travieso, C. M., Bolaños, F., García, A., et al. (2017). Automatic anuran identification using noise removal and audio activity detection. *Expert. Syst. Appl*. 72, 83–92. doi: 10.1016/j.eswa.2016.12.019

Altmann, E. G., Cristadoro, G., and Degli Esposti, M. (2012). On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci. U.S.A*. 109, 11582–11587. doi: 10.1073/pnas.1117723109

Alvarez-Lacalle, E., Dorow, B., Eckmann, J.-P., and Moses, E. (2006). Hierarchical structures induce long-range dynamical correlations in written texts. *Proc. Natl. Acad. Sci. U.S.A*. 103, 7956–7961. doi: 10.1073/pnas.0510673103

Anderson, D. J., and Perona, P. (2014). Toward a science of computational ethology. *Neuron* 84, 18–31. doi: 10.1016/j.neuron.2014.09.005

Anderson, S. E., Dave, A. S., and Margoliash, D. (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. *J. Acoust. Soc. Am*. 100, 1209–1219. doi: 10.1121/1.415968

Anikin, A. (2019). Soundgen: an open-source tool for synthesizing nonverbal vocalizations. *Behav. Res. Methods* 51, 778–792. doi: 10.3758/s13428-018-1095-7

Arneodo, E. M., Chen, S., Brown, I. I., D. E., Gilja, V., and Gentner, T. Q. (2021). Neurally driven synthesis of learned, complex vocalizations. *Curr. Biol*. 31, 3419.e5–3425.e5. doi: 10.1016/j.cub.2021.05.035

Arneodo, E. M., and Mindlin, G. B. (2009). Source-tract coupling in birdsong production. *Phys. Rev. E* 79, 061921. doi: 10.1103/PhysRevE.79.061921

Arneodo, E. M., Perl, Y. S., Goller, F., and Mindlin, G. B. (2012). Prosthetic avian vocal organ controlled by a freely behaving bird based on a low dimensional model of the biomechanical periphery. *PLoS Comput. Biol*. 8:e1002546. doi: 10.1371/journal.pcbi.1002546

Arriaga, G., Zhou, E. P., and Jarvis, E. D. (2012). Of mice, birds, and men: the mouse ultrasonic song system has some features similar to humans and song-learning birds. *PLoS ONE* 7:e46610. doi: 10.1371/journal.pone.0046610

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. *arXiv preprint* arXiv:2006.11477.

Bailey, D. J., Rosebush, J. C., and Wade, J. (2002). The hippocampus and caudomedial neostriatum show selective responsiveness to conspecific song in the female zebra finch. *J. Neurobiol*. 52, 43–51. doi: 10.1002/neu.10070

Barker, A. J., Veviurko, G., Bennett, N. C., Hart, D. W., Mograby, L., and Lewin, G. R. (2021). Cultural transmission of vocal dialect in the naked mole-rat. *Science* 371, 503–507. doi: 10.1126/science.abc6588

Berman, G. J. (2018). Measuring behavior across scales. *BMC Biol*. 16:23. doi: 10.1186/s12915-018-0494-7

Berman, G. J., Bialek, W., and Shaevitz, J. W. (2016). Predictability and hierarchy in drosophila behavior. *Proc. Natl. Acad. Sci. U.S.A*. 113, 11943–11948. doi: 10.1073/pnas.1607601113

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. (2019). Mixmatch: a holistic approach to semi-supervised learning. *arXiv preprint* arXiv:1905.02249.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust*. 27, 113–120. doi: 10.1109/TASSP.1979.1163209

Bonada, J., Lachlan, R., and Blaauw, M. (2016). "Bird song synthesis based on hidden markov models," in *Interspeech 2016; 2016 Sep 08-12* (San Francisco, CA: [Baixas]: ISCA), 2582–2586.

Bottjer, S. W., and Altenau, B. (2010). Parallel pathways for vocal learning in basal ganglia of songbirds. *Nat. Neurosci*. 13, 153–155. doi: 10.1038/nn.2472

Bregman, M. R., Patel, A. D., and Gentner, T. Q. (2016). Songbirds use spectral shape, not pitch, for sound pattern recognition. *Proc. Natl. Acad. Sci. U.S.A*. 113, 1666–1671. doi: 10.1073/pnas.1515380113

Brown, A., Garg, S., and Montgomery, J. (2017). Automatic and efficient denoising of bioacoustics recordings using mmse stsa. *IEEE Access* 6, 5010–5022. doi: 10.1109/ACCESS.2017.2782778

Brown, A. E., and De Bivort, B. (2018). Ethology as a physical science. *Nat. Phys*. 14, 653–657. doi: 10.1038/s41567-018-0093-0

Brown, G. J., and Cooke, M. (1994). Computational auditory scene analysis. *Comput. Speech Lang*. 8, 297–336. doi: 10.1006/csla.1994.1016

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165.

Bruno, J. H., and Tchernichovski, O. (2019). Regularities in zebra finch song beyond the repeated motif. *Behav. Proc.* 163, 53–59. doi: 10.1016/j.beproc.2017.11.001

Calhoun, A. J., Pillow, J. W., and Murthy, M. (2019). Unsupervised identification of the internal states that shape natural behavior. *Nat. Neurosci.* 22, 2040–2049. doi: 10.1038/s41593-019-0533-x

Castellucci, G. A., Calbick, D., and McCormick, D. (2018). The temporal organization of mouse ultrasonic vocalizations. *PLoS ONE* 13:e0199929. doi: 10.1371/journal.pone.0199929

Chakladar, S., Logothetis, N. K., and Petkov, C. I. (2008). Morphing rhesus monkey vocalizations. *J. Neurosci. Methods* 170, 45–55. doi: 10.1016/j.jneumeth.2007.12.023

Christin, S., Hervet, É., and Lecomte, N. (2019). Applications for deep learning in ecology. *Methods Ecol. Evolut.* 10, 1632–1644. doi: 10.1111/2041-210X.13256

Clark, C. W., Marler, P., and Beeman, K. (1987). Quantitative analysis of animal vocal phonology: an application to swamp sparrow song. *Ethology* 76, 101–115. doi: 10.1111/j.1439-0310.1987.tb00676.x

Cody, M. L., Stabler, E., Sanchez Castellanos, H. M., and Taylor, C. E. (2016). Structure, syntax and "small-world" organization in the complex songs of california thrashers (toxostoma redivivum). *Bioacoustics* 25, 41–54. doi: 10.1080/09524622.2015.1089418

Coen, P., Clemens, J., Weinstein, A. J., Pacheco, D. A., Deng, Y., and Murthy, M. (2014). Dynamic sensory cues shape song structure in drosophila. *Nature* 507, 233–237. doi: 10.1038/nature13131

Cohen, Y., Nicholson, D. A., Sanchioni, A., Mallaber, E. K., Skidanova, V., and Gardner, T. J. (2020a). Tweetynet: a neural network that enables high-throughput, automated annotation of birdsong. *bioRxiv*. doi: 10.1101/2020.08.28.272088

Cohen, Y., Shen, J., Semu, D., Leman, D. P., Liberti, W. A., Perkins, L. N., et al. (2020b). Hidden neural states underlie canary song syntax. *Nature* 582, 539–544. doi: 10.1038/s41586-020-2397-3

Datta, S. R., Anderson, D. J., Branson, K., Perona, P., and Leifer, A. (2019). Computational neuroethology: a call to action. *Neuron* 104, 11–24. doi: 10.1016/j.neuron.2019.09.038

Dawkins, R. (1976). Hierarchical organisation: a candidate principle for ethology. *Growing Points Ethol.* 7:54.

De Silva, V., and Tenenbaum, J. B. (2002). "Global versus local methods in nonlinear dimensionality reduction," in *NIPS'02: Proceedings of the 15th International Conference on Neural Information Processing Systems Vol. 15*. (Vancouver, BC: Association for Computing Machinery), 705–712.

Dehaene, S., Meyniel, F., Wacongne, C., Wang, L., and Pallier, C. (2015). The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron* 88, 2–19. doi: 10.1016/j.neuron.2015.09.019

Diekamp, B., Kalt, T., and Güntürkün, O. (2002). Working memory neurons in pigeons. *J. Neurosci.* 22, RC210–RC210. doi: 10.1523/JNEUROSCI.22-04-j0002.2002

DiMattina, C., and Wang, X. (2006). Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *J. Neurophysiol.* 95, 1244–1262. doi: 10.1152/jn.00818.2005

Ditz, H. M., and Nieder, A. (2015). Neurons selective to the number of visual items in the corvid songbird endbrain. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7827–7832. doi: 10.1073/pnas.1504245112

Donahue, C., McAuley, J., and Puckette, M. (2018). Adversarial audio synthesis. *arXiv preprint* arXiv:1802.04208.

Dooling, R. J., Park, T. J., Brown, S. D., Okanoya, K., and Soli, S. D. (1987). Perceptual organization of acoustic stimuli by budgerigars (melopsittacus undulatus): Ii. vocal signals. *J. Compar. Psychol.* 101, 367. doi: 10.1037/0735-7036.101.4.367

Dooling, R. J., and Prior, N. H. (2017). Do we hear what birds hear in birdsong? *Anim. Behav.* 124, 283–289. doi: 10.1016/j.anbehav.2016.10.012

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* 22, 567–631. doi: 10.1146/annurev.neuro.22.1.567

Dudley, H. (1939). Remaking speech. *J. Acoust. Soc. Am.* 11, 169–177. doi: 10.1121/1.1916020

Dunn, T. W., Marshall, J. D., Severson, K. S., Aldarondo, D. E., Hildebrand, D. G., Chettih, S. N., et al. (2021). Geometric deep learning enables 3d kinematic profiling across species and environments. *Nat. Methods* 18, 564–573. doi: 10.1038/s41592-021-01106-6

Eldredge, D. H., Miller, J. D., and Bohne, B. A. (1981). A frequency-position map for the chinchilla cochlea. *J. Acoust. Soc. Am.* 69, 1091–1095. doi: 10.1121/1.385688

Elie, J. E., and Theunissen, F. E. (2016). The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Anim. Cogn.* 19, 285–315. doi: 10.1007/s10071-015-0933-6

Elie, J. E., and Theunissen, F. E. (2018). Zebra finches identify individuals using vocal signatures unique to each call type. *Nat. Commun.* 9, 1–11. doi: 10.1038/s41467-018-06394-9

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., and Roberts, A. (2019). Gansynth: adversarial neural audio synthesis. *arXiv preprint* arXiv:1902.08710.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Norouzi, M., Eck, D., et al. (2017). "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning* (Sydney, NSW: PMLR), 1068–1077.

Ephraim, Y., and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust.* 32, 1109–1121. doi: 10.1109/TASSP.1984.1164453

Ezen-Can, A. (2020). A comparison of lstm and bert for small corpus. *arXiv preprint* arXiv:2009.05451.

Feather, J., Durango, A., Gonzalez, R., and McDermott, J. (2019). "Metamers of neural networks reveal divergence from human perceptual systems," in *Advances in Neural Information Processing Systems*, (Vancouver, BC: Association for Computing Machinery), 10078–10089.

Fee, M. S., Shraiman, B., Pesaran, B., and Mitra, P. P. (1998). The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. *Nature* 395, 67–71. doi: 10.1038/25725

Fischer, J., and Hammerschmidt, K. (2020). Towards a new taxonomy of primate vocal production learning. *Philos. Trans. R. Soc. B* 375, 20190045. doi: 10.1098/rstb.2019.0045

Fitch, W. T., and Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: an overview. *Philos. Trans. R. Soc. B* 367, 1933–1955. doi: 10.1098/rstb.2012.0103

Frasier, K. E., Roch, M. A., Soldevilla, M. S., Wiggins, S. M., Garrison, L. P., and Hildebrand, J. A. (2017). Automated classification of dolphin echolocation click types from the gulf of mexico. *PLoS Comput. Biol.* 13:e1005823. doi: 10.1371/journal.pcbi.1005823

Fujimoto, H., Hasegawa, T., and Watanabe, D. (2011). Neural coding of syntactic structure in learned vocalizations in the songbird. *J. Neurosci.* 31, 10023–10033. doi: 10.1523/JNEUROSCI.1606-11.2011

Fukushima, M., Doyle, A. M., Mullarkey, M. P., Mishkin, M., and Averbeck, B. B. (2015). Distributed acoustic cues for caller identity in macaque vocalization. *R. Soc. Open Sci.* 2, 150432. doi: 10.1098/rsos.150432

Fulop, S. A., and Fitz, K. (2006). Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.* 119, 360–371. doi: 10.1121/1.2133000

Furuyama, T., Kobayasi, K. I., and Riquimaroux, H. (2017). Acoustic characteristics used by japanese macaques for individual discrimination. *J. Exp. Biol.* 220, 3571–3578. doi: 10.1242/jeb.154765

Gao, P., and Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* 32, 148–155. doi: 10.1016/j.conb.2015.04.003

Gardner, T. J., and Magnasco, M. O. (2006). Sparse time-frequency representations. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6094–6099. doi: 10.1073/pnas.0601707103

Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature* 440, 1204–1207. doi: 10.1038/nature04675

Goëau, H., Glotin, H., Vellinga, W.-P., Planqué, R., Rauber, A., and Joly, A. (2014). "Lifeclef bird identification task 2014," in *CLEF: Conference and Labs of the Evaluation Forum, 1180*. (Sheffield: Association for Computing Machinery), 585-597.

Goffinet, J., Brudner, S., Mooney, R., and Pearson, J. (2021). Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *Elife* 10:e67855. doi: 10.7554/eLife.67855

Greenwood, D. D. (1997). The mel scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances and equal frequency ratios. *Hear. Res*. 103, 199–224. doi: 10.1016/S0378-5955(96)00175-X

Griffin, D., and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Trans. Acoust*. 32, 236–243. doi: 10.1109/TASSP.1984.1164317

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., et al. (2020). Conformer: convolution-augmented transformer for speech recognition. *arXiv preprint* arXiv:2005.08100. doi: 10.21437/Interspeech.2020-3015

Güntürkün, O. (2005). The avian 'prefrontal cortex' and cognition. *Curr. Opin. Neurobiol*. 15, 686–693. doi: 10.1016/j.conb.2005.10.003

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298, 1569–1579. doi: 10.1126/science.298.5598.1569

Hedley, R. W. (2016). Composition and sequential organization of song repertoires in cassin's vireo (vireo cassinii). *J. Ornithol*. 157, 13–22. doi: 10.1007/s10336-015-1238-x

Helduser, S., Cheng, S., and Güntürkün, O. (2013). Identification of two forebrain structures that mediate execution of memorized sequences in the pigeon. *J. Neurophysiol*. 109, 958–968. doi: 10.1152/jn.00763.2012

Helduser, S., and Güntürkün, O. (2012). Neural substrates for serial reaction time tasks in pigeons. *Behav. Brain Res*. 230, 132–143. doi: 10.1016/j.bbr.2012.02.013

Holveck, M.-J., Vieira de Castro, A. C., Lachlan, R. F., ten Cate, C., and Riebel, K. (2008). Accuracy of song syntax learning and singing consistency signal early condition in zebra finches. *Behav. Ecol*. 19, 1267–1281. doi: 10.1093/beheco/arn078

Holy, T. E., and Guo, Z. (2005). Ultrasonic songs of male mice. *PLoS Biol*. 3, e386. doi: 10.1371/journal.pbio.0030386

Hsu, S.-B., Lee, C.-H., Chang, P.-C., Han, C.-C., and Fan, K.-C. (2018). Local wavelet acoustic pattern: a novel time-frequency descriptor for birdsong recognition. *IEEE Trans. Multimedia* 20, 3187–3199. doi: 10.1109/TMM.2018.2834866

Janney, E., Taylor, H., Scharff, C., Rothenberg, D., Parra, L. C., and Tchernichovski, O. (2016). Temporal regularity increases with repertoire complexity in the australian pied butcherbird's song. *R. Soc. Open Sci*. 3, 160357. doi: 10.1098/rsos.160357

Jiang, X., Long, T., Cao, W., Li, J., Dehaene, S., and Wang, L. (2018). Production of supra-regular spatial sequences by macaque monkeys. *Curr. Biol*. 28, 1851–1859. doi: 10.1016/j.cub.2018.04.047

Jin, D. Z., and Kozhevnikov, A. A. (2011). A compact statistical model of the song syntax in bengalese finch. *PLoS Comput. Biol*. 7:e1001108. doi: 10.1371/journal.pcbi.1001108

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., et al. (2018). "Efficient neural audio synthesis," in *International Conference on Machine Learning* (Stockholm: PMLR), 2410–2419.

Kalenscher, T., Diekamp, B., and Güntürkün, O. (2003). Neural architecture of choice behaviour in a concurrent interval schedule. *Eur. J. Neurosci*. 18, 2627–2637. doi: 10.1046/j.1460-9568.2003.03006.x

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. *arXiv preprint* arXiv:2001.08361.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., et al. (2019). "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (Singapore: IEEE), 449–456.

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:e24516. doi: 10.1371/journal.pone.0024516

Kawahara, H. (2006). Straight, exploitation of the other aspect of vocoder: perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. Technol*. 27, 349–353. doi: 10.1250/ast.27.349

Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds. *Speech Commun*. 27, 187–207. doi: 10.1016/S0167-6393(98)00085-5

Keen, S. C., Odom, K. J., Webster, M. S., Kohn, G. M., Wright, T. F., and Araya-Salas, M. (2021). A machine learning approach for classifying and quantifying acoustic diversity. *Methods Ecol. Evolut*. 12, 1213–1225. doi: 10.1111/2041-210X.13599

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., et al. (2016). Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biol. Rev*. 91, 13–52. doi: 10.1111/brv.12160

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. *Proc. R. Soc. B Biol. Sci*. 281, 20141370. doi: 10.1098/rspb.2014.1370

Kiapuchinski, D. M., Lima, C. R. E., and Kaestner, C. A. A. (2012). "Spectral noise gate technique applied to birdsong preprocessing on embedded unit," in *2012 IEEE International Symposium on Multimedia* (Irvine, CA: IEEE), 24–27.

Kim, B., and Pardo, B. (2018). A human-in-the-loop system for sound event detection and annotation. *ACM Trans. Interact. Intell. Syst*. 8, 1–23. doi: 10.1145/3214366

Kim, H.-G., Obermayer, K., Bode, M., and Ruwisch, D. (2000). Real-time noise canceling based on spectral minimum detection and diffusive gain factors. *J. Acoust. Soc. Am*. 108, 2484–2484. doi: 10.1121/1.4743167

Knight, E. C., Poo Hernandez, S., Bayne, E. M., Bulitko, V., and Tucker, B. V. (2020). Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics* 29, 337–355. doi: 10.1080/09524622.2019.1606734

Kogan, J. A., and Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: a comparative study. *J. Acoust. Soc. Am*. 103, 2185–2196. doi: 10.1121/1.421364

Kohlsdorf, D., Herzing, D., and Starner, T. (2020). "An auto encoder for audio dolphin communication," in *2020 International Joint Conference on Neural Networks (IJCNN)*, (Glasgow: IEEE), 1–7.

Kollmorgen, S., Hahnloser, R. H., and Mante, V. (2020). Nearest neighbours reveal fast and slow components of motor learning. *Nature* 577, 526–530. doi: 10.1038/s41586-019-1892-x

Kong, Q., Xu, Y., and Plumbley, M. D. (2017). "Joint detection and classification convolutional neural network on weakly labelled bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)* (Kos: IEEE), 1749–1753.

Kröner, S., and Güntürkün, O. (1999). Afferent and efferent connections of the caudolateral neostriatum in the pigeon (columba livia): a retro- and anterograde pathway tracing study. *J. Compar. Neurol*. 407, 228–260. doi: 10.1002/(SICI)1096-9861(19990503)407:2<228::AID-CNE6>3.0.CO;2-2

Kuhl, P. K., and Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic vot stimuli. *J. Acoust. Soc. Am*. 63, 905–917. doi: 10.1121/1.381770

Kuhl, P. K., and Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *J. Acoust. Soc. Am*. 73, 1003–1010. doi: 10.1121/1.389148

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., et al. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint* arXiv:1910.06711.

Lachlan, R. (2007). *Luscinia: A Bioacoustics Analysis Computer Program*. Available online at: http://luscinia.sourceforge.net/page3/page21/page21.html

Lachlan, R., Verhagen, L., Peters, S., and Cate, C., t. (2010). Are there species-universal categories in bird song phonology and syntax? a comparative study of chaffinches (fringilla coelebs), zebra finches (taenopygia guttata), and swamp sparrows (melospiza georgiana). *J. Compar. Psychol*. 124, 92. doi: 10.1037/a0016996

Lachlan, R. F., and Nowicki, S. (2015). Context-dependent categorical perception in a songbird. *Proc. Natl. Acad. Sci. U.S.A*. 112, 1892–1897. doi: 10.1073/pnas.1410844112

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning* (PMLR), 1558–1566.

Lashley, K. S. (1951). *The Problem of Serial Order in Behavior, Vol. 21*. Oxford: Bobbs-Merrill.

Lasseck, M. (2013). "Bird song classification in field recordings: winning solution for nips4b 2013 competition," in *Proceeding of International Symposium Neural Information Scaled for Bioacoustics* (Lake Tahoe, NV: sabiod. org/nips4b, joint to NIPS), 176–181.

Lee, G. W., and Kim, H. K. (2020). Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection. *Appl. Sci.* 10, 3230. doi: 10.3390/app10093230

Li, W. (1990). Mutual information functions versus correlation functions. *J. Stat. Phys*. 60, 823–837. doi: 10.1007/BF01025996

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol*. 54, 358. doi: 10.1037/h0044417

Lin, H. W., and Tegmark, M. (2017). Critical behavior in physics and probabilistic formal languages. *Entropy* 19, 299. doi: 10.3390/e19070299

Liu, S. T., Montes-Lourido, P., Wang, X., and Sadagopan, S. (2019). Optimal features for auditory categorization. *Nat. Commun*. 10, 1–14. doi: 10.1038/s41467-019-09115-y

Loo, Y. Y., and Cain, K. E. (2021). A call to expand avian vocal development research. *Front. Ecol. Evolut*. 9:757972. doi: 10.3389/fevo.2021.757972

Lostanlen, V., Palmer, K., Knight, E., Clark, C., Klinck, H., Farnsworth, A., et al. (2019a). Long-distance detection of bioacoustic events with per-channel energy normalization. *arXiv preprint* arXiv:1911.00417. doi: 10.33682/ts6e-sn53

Lostanlen, V., Salamon, J., Cartwright, M., McFee, B., Farnsworth, A., Kelling, S., et al. (2018). Per-channel energy normalization: why and how. *IEEE Signal Process Lett*. 26, 39–43. doi: 10.1109/LSP.2018.2878620

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (2019b). Robust sound event detection in bioacoustic sensor networks. *PLoS ONE* 14:e0214168. doi: 10.1371/journal.pone.0214168

Main, L., and Thornton, J. (2015). "A cortically-inspired model for bioacoustics recognition," in *International Conference on Neural Information Processing* (Istanbul: Springer), 348–355.

Markowitz, J. E., Ivie, E., Kligler, L., and Gardner, T. J. (2013). Long-range order in canary song. *PLoS Comput. Biol*. 9:e1003052. doi: 10.1371/journal.pcbi.1003052

Marshall, J. D., Aldarondo, D. E., Dunn, T. W., Wang, W. L., Berman, G. J., and Ölveczky, B. P. (2021). Continuous whole-body 3d kinematic recordings across the rodent behavioral repertoire. *Neuron* 109, 420–437. doi: 10.1016/j.neuron.2020.11.016

Masuyama, Y., Yatabe, K., Koizumi, Y., Oikawa, Y., and Harada, N. (2019). "Deep griffin-lim iteration," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (Glasgow: IEEE), 61–65.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., et al. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci*. 21, 1281–1289. doi: 10.1038/s41593-018-0209-y

McDermott, J. H., Oxenham, A. J., and Simoncelli, E. P. (2009). "Sound texture synthesis via filter statistics," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (New Paltz, NY: IEEE), 297–300.

McDermott, J. H., and Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron* 71, 926–940. doi: 10.1016/j.neuron.2011.06.032

Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., et al. (2016). Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint* arXiv:1612.07837.

Mets, D. G., and Brainard, M. S. (2018). An automated approach to the quantitation of vocalizations and vocal learning in the songbird. *PLoS Comput. Biol*. 14:e1006437. doi: 10.1371/journal.pcbi.1006437

Mielke, A., and Zuberbühler, K. (2013). A method for automated individual, species and call type recognition in free-ranging animals. *Anim. Behav*. 86, 475–482. doi: 10.1016/j.anbehav.2013.04.017

Miller, E. H. (1979). An approach to the analysis of graded vocalizations of birds. *Behav. Neural Biol*. 27, 25–38. doi: 10.1016/S0163-1047(79)92732-8

Milmine, M., Watanabe, A., and Colombo, M. (2008). Neural correlates of directed forgetting in the avian prefrontal cortex. *Behav. Neurosci*. 122, 199. doi: 10.1037/0735-7044.122.1.199

Mizuhara, T., and Okanoya, K. (2020). Do songbirds hear songs syllable by syllable? *Behav. Processes* 174:104089. doi: 10.1016/j.beproc.2020.104089

Moca, V. V., Bârzan, H., Nagy-Dăbâcan, A., and Mureşan, R. C. (2021). Time-frequency super-resolution with superlets. *Nat. Commun*. 12, 1–18. doi: 10.1038/s41467-020-20539-9

Moll, F. W., and Nieder, A. (2015). Cross-modal associative mnemonic signals in crow endbrain neurons. *Curr. Biol*. 25, 2196–2201. doi: 10.1016/j.cub.2015.07.013

Moll, F. W., and Nieder, A. (2017). Modality-invariant audio-visual association coding in crow endbrain neurons. *Neurobiol. Learn. Mem*. 137, 65–76. doi: 10.1016/j.nlm.2016.11.011

Morfi, V., Lachlan, R. F., and Stowell, D. (2021). Deep perceptual embeddings for unlabelled animal sound events. *J. Acoust. Soc. Am*. 150, 2–11. doi: 10.1121/10.0005475

Morita, T., and Koda, H. (2019). Superregular grammars do not provide additional explanatory power but allow for a compact analysis of animal song. *R. Soc. Open Sci*. 6, 190139. doi: 10.1098/rsos.190139

Morita, T., Koda, H., Okanoya, K., and Tachibana, R. O. (2020). Birdsong sequence exhibits long context dependency comparable to human language syntax. *bioRxiv*.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., et al. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med*. 385, 217–227. doi: 10.1056/NEJMoa2027540

Muda, L., Begam, M., and Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint* arXiv:1003.4083.

Murcia, R. H., and Paniagua, V. S. (2013). "The ICML 2013 Bird Challenge: Bird identification from continuous audio recordings," in *International Conference on Machine Learning 2013* (Sweden: Citeseer), 26–96.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., and Shaalan, K. (2019). Speech recognition using deep neural networks: a systematic review. *IEEE Access* 7, 19143–19165. doi: 10.1109/ACCESS.2019.2896880

Nelson, D. A., and Marler, P. (1989). Categorical perception of a natural stimulus continuum: birdsong. *Science* 244, 976–978. doi: 10.1126/science.2727689

Nicholson, D. (2016). "Comparison of machine learning methods applied to birdsong element classification," in *Proceedings of the 15th Python in Science Conference* (Austin, TX: Scipy.org), 57–61.

Okobi, D. E., Banerjee, A., Matheson, A. M., Phelps, S. M., and Long, M. A. (2019). Motor cortical control of vocal interaction in neotropical singing mice. *Science* 363, 983–988. doi: 10.1126/science.aau9480

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., et al. (2016). Wavenet: A generative model for raw audio. *arXiv preprint* arXiv:160 9.03499.

Pagliarini, S., Trouvain, N., Lebois, A., and Hinaut, X. (2021). *What does the Canary Say? Low-Dimensional GAN Applied to Birdsong*. [Epub ahead of print].

Pandarinath, C., O'Shea, D. J., Collins, J., Jozefowicz, R., Stavisky, S. D., Kao, J. C., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* 15, 805–815. doi: 10.1038/s41592-018-0109-9

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., et al. (2019). Specaugment: a simple data augmentation method for automatic speech recognition. *arXiv preprint* arXiv:1904.08779. doi: 10.21437/Interspeech.2019-2680

Patricelli, G. L., and Hebets, E. A. (2016). New dimensions in animal communication: the case for complexity. *Curr. Opin. Behav. Sci*. 12, 80–89. doi: 10.1016/j.cobeha.2016.09.011

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., et al. (2019). Fast animal pose estimation using deep neural networks. *Nat. Methods* 16, 117–125. doi: 10.1038/s41592-018-0234-5

Petkov, C. I., and Jarvis, E. (2012). Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Front. Evol. Neurosci*. 4:12. doi: 10.3389/fnevo.2012.00012

Prenger, R., Valle, R., and Catanzaro, B. (2019). "Waveglow: a flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton: IEEE), 3617–3621.

Priyadarshani, N., Marsland, S., and Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: a review. *J. Avian Biol*. 49, jav-01447. doi: 10.1111/jav.01447

Priyadarshani, N., Marsland, S., Castro, I., and Punchihewa, A. (2016). Birdsong denoising using wavelets. *PLoS ONE* 11:e0146790. doi: 10.1371/journal.pone.0146790

Priyadarshani, N., Marsland, S., Juodakis, J., Castro, I., and Listanti, V. (2020). Wavelet filters for automated recognition of birdsong in long-time field recordings. *Methods Ecol. Evolut*. 11, 403–417. doi: 10.1111/2041-210X.13357

Prša, Z., and Rajmic, P. (2017). Toward high-quality real-time signal reconstruction from stft magnitude. *IEEE Signal Process Lett*. 24, 892–896. doi: 10.1109/LSP.2017.2696970

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint* arXiv:1511.06434.

Rahman, M., Willmore, B. D., King, A. J., and Harper, N. S. (2020). Simple transformations capture auditory input to cortex. *Proc. Natl. Acad. Sci. U.S.A*. 117, 28442–28451. doi: 10.1073/pnas.1922033117

Ren, Y., Johnson, M. T., and Tao, J. (2008). Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. *J. Acoust. Soc. Am*. 124, 316–327. doi: 10.1121/1.2932070

Renteria, S., Vallejo, E., and Taylor, C. (2021). Birdsong phrase verification and classification using siamese neural networks. *bioRxiv*. doi: 10.1101/2021.03.16.435625

Rinnert, P., Kirschhock, M. E., and Nieder, A. (2019). Neuronal correlates of spatial working memory in the endbrain of crows. *Curr. Biol*. 29, 2616–2624. doi: 10.1016/j.cub.2019.06.060

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. (2018). "A hierarchical latent vector model for learning long-term structure in music," in *International Conference on Machine Learning* (Stockholm: PMLR), 4364–4373.

Rohrmeier, M., Zuidema, W., Wiggins, G. A., and Scharff, C. (2015). Principles of structure building in music, language and animal song. *Philos. Trans. R. Soc. B* 370, 20140097. doi: 10.1098/rstb.2014.0097

Rose, J., and Colombo, M. (2005). Neural correlates of executive control in the avian brain. *PLoS Biol*. 3:e190. doi: 10.1371/journal.pbio.0030190

Russ, B. E., Lee, Y.-S., and Cohen, Y. E. (2007). Neural and behavioral correlates of auditory categorization. *Hear. Res*. 229, 204–212. doi: 10.1016/j.heares.2006.10.010

Sainburg, T. (2019). timsainb/noisereduce: v1.0 (db94fe2). *Zenodo*. doi: 10.5281/zenodo.3243139

Sainburg, T., Mai, A., and Gentner, T. Q. (2020a). Long-range sequential dependencies precede complex syntactic production in language acquisition. *bioRxiv*. doi: 10.1101/2020.08.19.256792

Sainburg, T., McInnes, L., and Gentner, T. Q. (2021). Parametric umap: learning embeddings with deep neural networks for representation and semi-supervised learning. *Neural Comput*. doi: 10.1162/neco_a_01434

Sainburg, T., Theilman, B., Thielk, M., and Gentner, T. Q. (2019). Parallels in the sequential organization of birdsong and human speech. *Nat. Commun*. 10, 1–11. doi: 10.1038/s41467-019-11605-y

Sainburg, T., Thielk, M., and Gentner, T. (2018a). "Learned context dependent categorical perception in a songbird," in *Conference on Cognitive Computational Neuroscience* (Philadelphia, PA: ccneuro.org), 1-4.

Sainburg, T., Thielk, M., and Gentner, T. Q. (2020b). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Comput. Biol*. 16:e1008228. doi: 10.1371/journal.pcbi.1008228

Sainburg, T., Thielk, M., and Gentner, T. Q. (2020c). Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv* 870311. doi: 10.1101/870311

Sainburg, T., Thielk, M., Theilman, B., Migliori, B., and Gentner, T. (2018b). Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. *arXiv preprint* arXiv:1807.06650.

Saint-Arnaud, N., and Popat, K. (1995). "Analysis and synthesis of sound textures," in *Readings in Computational Auditory Scene Analysis* (CiteSeer). Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.111.586

Sasahara, K., Cody, M. L., Cohen, D., and Taylor, C. E. (2012). Structural design principles of complex bird songs: a network-based approach. *PLoS ONE* 7:e44436. doi: 10.1371/journal.pone.0044436

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint* arXiv:1904.05862. doi: 10.21437/Interspeech.2019-1873

Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell Syst. Techn. J*. 30, 50–64. doi: 10.1002/j.1538-7305.1951.tb01366.x

Shen, H. (2019). Mutual information scaling and expressive power of sequence models. *arXiv preprint* arXiv:1905.04271.

Singh Alvarado, J., Goffinet, J., Michael, V., Liberti, W., Hatfield, J., Gardner, T., et al. (2021). Neural dynamics underlying birdsong practice and performance. *Nature* 599, 635–639. doi: 10.1038/s41586-021-04004-1

Sinnott, J., Beecher, M., Moody, D., and Stebbins, W. (1976). Speech sound discrimination by monkeys and humans. *J. Acoust. Soc. Am*. 60, 687–695. doi: 10.1121/1.381140

Sitt, J., Amador, A., Goller, F., and Mindlin, G. (2008). Dynamical origin of spectrally rich vocalizations in birdsong. *Phys. Rev. E* 78, 011905. doi: 10.1103/PhysRevE.78.011905

Sitt, J. D., Arneodo, E. M., Goller, F., and Mindlin, G. B. (2010). Physiologically driven avian vocal synthesizer. *Phys. Rev. E* 81, 031927. doi: 10.1103/PhysRevE.81.031927

Somervuo, P. (2019). Time-frequency warping of spectrograms applied to bird sound analyses. *Bioacoustics* 28, 257–268. doi: 10.1080/09524622.2018.1431958

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am*. 8, 185–190. doi: 10.1121/1.1915893

Stowell, D., Petrusková, T., Šálek, M., and Linhart, P. (2019). Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *J. R. Soc. Interface* 16, 20180940. doi: 10.1098/rsif.2018.0940

Suzuki, R., Buck, J. R., and Tyack, P. L. (2006). Information entropy of humpback whale songs. *J. Acoust. Soc. Am*. 119, 1849–1866. doi: 10.1121/1.2161827

Tachibana, R. O., Oosugi, N., and Okanoya, K. (2014). Semi-automatic classification of birdsong elements using a linear support vector machine. *PLoS ONE* 9:e92584. doi: 10.1371/journal.pone.0092584

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., and Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. *Anim. Behav*. 59, 1167–1176. doi: 10.1006/anbe.1999.1416

ten Cate, C. (2014). On the phonetic and syntactic processing abilities of birds: from songs to speech and artificial grammars. *Curr. Opin. Neurobiol*. 28, 157–164. doi: 10.1016/j.conb.2014.07.019

ten Cate, C., and Okanoya, K. (2012). Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. *Philos. Trans. R. Soc. B* 367, 1984–1994. doi: 10.1098/rstb.2012.0055

Thielk, M., Sainburg, T., Sharpee, T., and Gentner, T. (2018). "Combining biological and artificial approaches to understand perceptual spaces for categorizing natural acoustic signals," in *Conference on Cognitive Computational Neuroscience* (Philadelphia, PA: ccneuro.org), 1-4.

Tjandra, A., Sakti, S., and Nakamura, S. (2020). Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. *arXiv preprint* arXiv:2005.11676. doi: 10.21437/Interspeech.2020-3033

Tran, K., Bisazza, A., and Monz, C. (2018). The importance of being recurrent for modeling hierarchical structure. *arXiv preprint* arXiv:1803.03585. doi: 10.18653/v1/D18-1503

Veit, L., and Nieder, A. (2013). Abstract rule neurons in the endbrain support intelligent behaviour in corvid songbirds. *Nat. Commun*. 4, 1–11. doi: 10.1038/ncomms3878

Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A.,et al. (2015). "The zero resource speech challenge 2015," in *Sixteenth Annual Conference of the International Speech Communication Association* (Dresden, IEEE), 1-12.

von Eugen, K., Tabrik, S., Güntürkün, O., and Ströckens, F. (2020). A comparative analysis of the dopaminergic innervation of the executive caudal nidopallium in pigeon, chicken, zebra finch, and carrion crow. *J. Compar. Neurol*. 528, 2929–2955. doi: 10.1002/cne.24878

Wagener, L., Loconsole, M., Ditz, H. M., and Nieder, A. (2018). Neurons in the endbrain of numerically naive crows spontaneously encode visual numerosity. *Curr. Biol*. 28, 1090–1094. doi: 10.1016/j.cub.2018.02.023

Wang, D., and Chen, J. (2018). Supervised speech separation based on deep learning: an overview. *IEEE/ACM Trans. Audio Speech Lang. Proc.* 26, 1702–1726. doi: 10.1109/TASLP.2018.2842159

Wang, Y., Getreuer, P., Hughes, T., Lyon, R. F., and Saurous, R. A. (2017). "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (New Orleans, LA: IEEE), 5670–5674.

Weiss, M., Hultsch, H., Adam, I., Scharff, C., and Kipper, S. (2014). The use of network analysis to study complex animal communication systems: a study on nightingale song. *Proc. R. Soc. B. Biol. Sci.* 281, 20140460. doi: 10.1098/rspb.2014.0460

Williams, H. (2001). Choreography of song, dance and beak movements in the zebra finch (taeniopygia guttata). *J. Exp. Biol.* 204, 3497–3506. doi: 10.1242/jeb.204.20.3497

Wilson, B., Marslen-Wilson, W. D., and Petkov, C. I. (2017). Conserved sequence processing in primate frontal cortex. *Trends Neurosci.* 40, 72–82. doi: 10.1016/j.tins.2016.11.004

Wiltschko, A. B., Johnson, M. J., Iurilli, G., Peterson, R. E., Katon, J. M., Pashkovski, S. L., et al. (2015). Mapping sub-second structure in mouse behavior. *Neuron* 88, 1121–1135. doi: 10.1016/j.neuron.2015.11.031

Wiltschko, A. B., Tsukahara, T., Zeine, A., Anyoha, R., Gillis, W. F., Markowitz, J. E., et al. (2020). Revealing the structure of pharmacobehavioral space through motion sequencing. *Nat. Neurosci.* 23, 1433–1443. doi: 10.1038/s41593-020-00706-3

Wimmer, J., Towsey, M., Planitz, B., Roe, P., and Williamson, I. (2010). "Scaling acoustic data analysis through collaboration and automation," in *2010 IEEE Sixth International Conference on e-Science* (Brisbane, QLD: IEEE), 308–315.

Xiao, J., and Flandrin, P. (2007). Multitaper time-frequency reassignment for nonstationary spectrum estimation and chirp enhancement. *IEEE Trans. Signal Proc.* 55, 2851–2860. doi: 10.1109/TSP.2007.893961

Xie, J., Colonna, J. G., and Zhang, J. (2020). Bioacoustic signal denoising: a review. *Artif. Intell. Rev.* 54, 3575–3597. doi: 10.07/s10462-020-09932-4

Xin, Y., Zhong, L., Zhang, Y., Zhou, T., Pan, J., and Xu, N.-L. (2019). Sensory-to-category transformation via dynamic reorganization of ensemble structures in mouse auditory cortex. *Neuron* 103, 909–921. doi: 10.1016/j.neuron.2019.06.004

Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O'Donnell, T. J., Sainburg, T., et al. (2020). Five ways in which computational modeling can help advance cognitive science: Lessons from artificial grammar learning. *Top. Cogn. Sci.* 12, 925–941. doi: 10.1111/tops.12474

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.