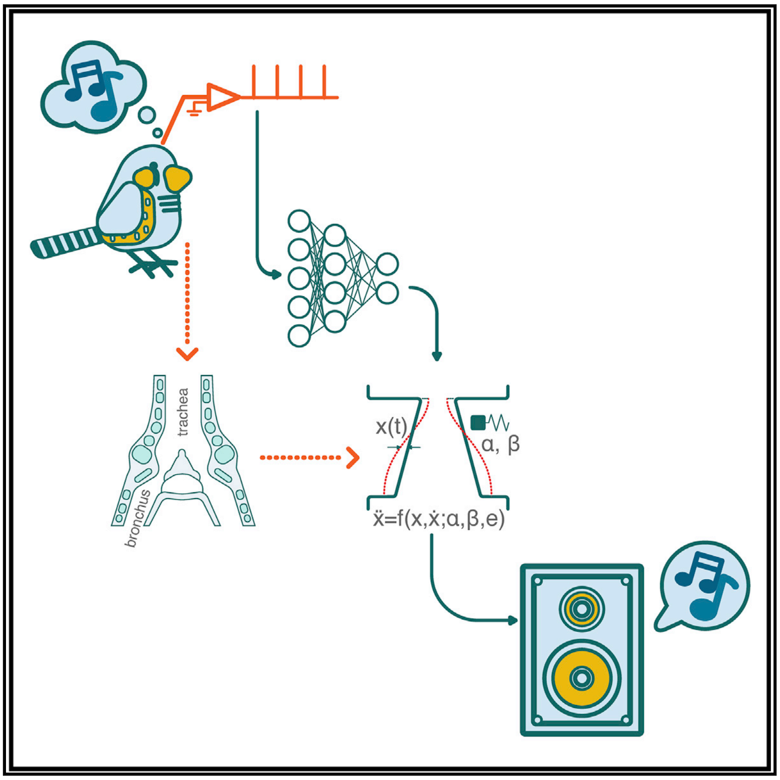
Report

Current Biology

Neurally driven synthesis of learned, complex vocalizations

Graphical abstract



Highlights

- Songbirds sing rich, complex, learned songs
- A model of their vocal organ can synthesize song with few control parameters
- This allows neurally driven song synthesis via a simple neural network
- Brain machine interfaces can be enhanced by understanding the biomechanics

Authors

Ezequiel M. Arneodo, Shukai Chen, Daril E. Brown II, Vikash Gilja, Timothy Q. Gentner

Correspondence

tgentner@ucsd.edu

In brief

Songbirds, like humans, need to control a sophisticated vocal organ to produce rich vocal sequences. Arneodo et. al. use knowledge of the biomechanics of the vocal organ and the structure of the vocal sequence to synthesize birdsong from recorded premotor neural activity.

Arneodo et al., 2021, Current Biology *31*, 3419–3425
August 9, 2021 © 2021 Elsevier Inc.
https://doi.org/10.1016/j.cub.2021.05.035

Current Biology

CellPress

Report

Neurally driven synthesis of learned, complex vocalizations

Ezequiel M. Arneodo,[1,2,3] Shukai Chen,[4] Daril E. Brown II,[5] Vikash Gilja,[5] and Timothy Q. Gentner[1,2,6,7,8,*]
[1]Biocircuits Institute, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[2]Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[3]IFLP-CONICET, Departamento de Física, Universidad Nacional de La Plata, CC 67, La Plata 1900, Argentina
[4]Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[5]Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[6]Kavli Institute for Brain and Mind, 9500 Gilman Drive, La Jolla, CA 92093, USA
[7]Neurobiology Section, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[8]Lead contact
*Correspondence: tgentner@ucsd.edu
https://doi.org/10.1016/j.cub.2021.05.035

SUMMARY

Brain machine interfaces (BMIs) hold promise to restore impaired motor function and serve as powerful tools to study learned motor skill. While limb-based motor prosthetic systems have leveraged nonhuman primates as an important animal model,[1–4] speech prostheses lack a similar animal model and are more limited in terms of neural interface technology, brain coverage, and behavioral study design.[5–7] Songbirds are an attractive model for learned complex vocal behavior. Birdsong shares a number of unique similarities with human speech,[8–10] and its study has yielded general insight into multiple mechanisms and circuits behind learning, execution, and maintenance of vocal motor skill.[11–18] In addition, the biomechanics of song production bear similarity to those of humans and some nonhuman primates.[19–23] Here, we demonstrate a vocal synthesizer for birdsong, realized by mapping neural population activity recorded from electrode arrays implanted in the premotor nucleus HVC onto low-dimensional compressed representations of song, using simple computational methods that are implementable in real time. Using a generative biomechanical model of the vocal organ (syrinx) as the low-dimensional target for these mappings allows for the synthesis of vocalizations that match the bird's own song. These results provide proof of concept that high-dimensional, complex natural behaviors can be directly synthesized from ongoing neural activity. This may inspire similar approaches to prosthetics in other species by exploiting knowledge of the peripheral systems and the temporal structure of their output.

RESULTS AND DISCUSSION

We describe two methods for synthesizing realistic vocal signals from neural activity recorded in a premotor nucleus of zebra finches (*Taeniopygia guttata*). Each method exploits a different trait of the vocal-motor process. First, we leverage understanding of the biomechanics of birdsong production. We employ a biomechanical model of the vocal organ that captures much of the spectro-temporal complexity of song in a low-dimensional parameter space.[24] This dimensionality reduction, compared to the full time-frequency representation of song, enables training of a shallow feedforward neural network (FFNN) that maps neural activity onto the model parameters. As a second synthesis method, we capitalize on predictive components in the temporal covariance between neural activity and song, which can be learned by a recurrent, long-short-term memory neural network (LSTM)[25] trained directly on frequency domain representations (spectrograms) of the vocal output.

Neuronal input for each synthesis comes from the sensory-motor nucleus HVC, where neurons generate high-level commands that drive the production of learned song. Adult male zebra finches (*Taeniopygia guttata*) sing individually stereotyped motifs comprising a sequence of 3–10 syllables. Activity in multiple HVC neuronal subtypes is modulated during singing: projection neurons targeting area X and RA ($HVC_{x/RA}$) evince short, precise, sparse activity bursts during the motif,[15,17,26–30] while inhibitory interneurons (HVC_I) display more tonic activity during singing.[14,29,31,32] To obtain ensemble HVC activity and vocal output, we implanted 16- or 32-channel Si probes in male, adult (>120-day-old) zebra finches and recorded extracellular voltages simultaneously while each bird sang (n = 4 birds, 70–120 vocal motifs per session). Neural recordings were sorted automatically using Kilosort and manually curated to exclude noise.[33] Non-noise clusters were classified as single- or multi-unit activity (SUA or MUA) based on the number of refractory period violations and putatively as projection or interneurons based on the sparseness of the activity during singing. The recordings were dominated by MUA clusters (n = 88) and HVC interneurons (HVC_I; n = 29), with relatively few putative projection neurons ($HVC_{X/RA}$; n = 15). Example song-aligned neural activity histograms are shown in Figure 1A. Example rasters with the numbers of clusters per bird are shown in Figure S1.

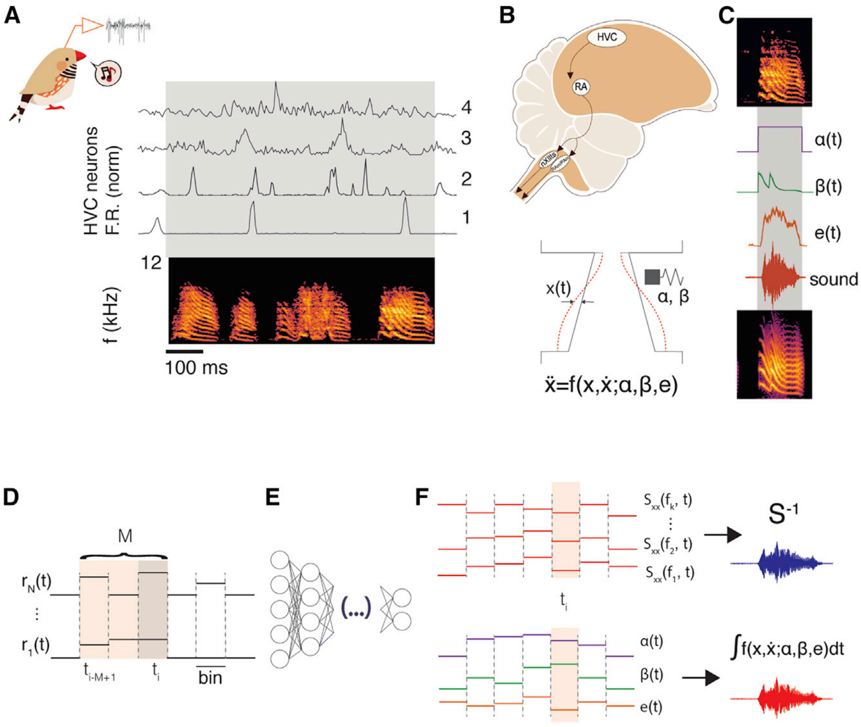


Figure 1. A neural-network-based decoder to synthesize birdsong from premotor neural activity

(A) Neural activity is collected from awake-singing animals. Sorted, extracellularly recorded single- and multi-units show different degrees of singing-related sparseness, robustness, and spiking precision (4 example clusters; top traces: normalized mean firing rate over 70 repetitions of the bird's motif; below: spectrogram of the motif; see also Figure S1).

(B) Downstream of HVC, the posterior motor pathway leads into nuclei that control the muscles driving the sound production (nXII and RAm/PAm).[34] Syringeal and respiratory muscles act coordinately to modulate the flow of air through sets of labia and produce sound.[35] The complex labial motion is captured by the equations of a nonlinear oscillator;[23] parameters that define acoustic properties of the sounds are surrogates of the activities of syringeal and respiratory muscles.[36]

(C) To reproduce a particular vocalization (top) from the biomechanical model, we fit the parameters (middle {α(t), β(t), e(t)}) such that, upon integration, the synthetic song (bottom) matches the pitch and spectral richness.

(D) The input of the neural network is an array with the values of a set of neural features (spike counts of sorted units/multi-units) over a window of M previous time steps.

(E) The hidden layer(s) of the network are composed either by a densely connected layer (FFNN) or two layers of LSTM cells.

(F) When training or reconstructing directly the spectral features of the song, the output of the network is a vector of powers across a range of frequency bands at a given time; the generated spectral slices are then inverted to produce synthetic song (top). When training or reconstructing via the biomechanical model, the output of the network at a given time is a 3-dimensional vector of parameters (as illustrated in C); the equations of the model are then integrated with these values to produce synthetic song (bottom). Illustrations were taken from Arneodo.[37]

Biomechanically meaningful compression enhances neurally driven synthesis

Synthesizing a complex motor sequence from neural activity requires mapping between two high-dimensional representations. To reduce the dimensionality of the problem, we leveraged a biomechanical model of the avian vocal organ that transforms neural activity to vocal output. The model accounts for the syrinx and the vocal tract.[24,27,38] The syrinx contains labial folds that oscillate when induced by the sub-syringeal air sac pressure and modulate the airflow to produce sound (Figure 1B).[35] The dynamics of the labia can be modeled after the motion equations of a nonlinear oscillator, in which the features of the sounds produced are determined by two time-varying parameters,[23,24,36] representing physiological motor instructions (the sub-syringeal pressure and the activity of the muscles that tense the labia).[24] In its simplest form, the syrinx model is computable in real time to produce synthetic vocalizations.[38] We model the vocal tract (the trachea, the oropharyngeal-esophageal cavity, and the beak) as a passive acoustic filter that determines species-specific spectral traits, such as the timbre.[24,27,39]

To synthesize song from neural activity via the biomechanical model, we first fit the parameters of the model to produce a synthetic version of each vocalization.[24,27,38] We searched for the parameters that produce, upon integration of the equations of the model, the closest match in pitch, spectral richness, and amplitude of the target vocalization. This effectively compresses each segment of a bird's own song (BOS) into a time series in a 3D parameter space, which generates a corresponding segment of synthetic song (SYN) (Figure 1C).[36,38] For each session, we randomly select 60% of the motifs for training, split each motif into 5-ms bins, and train a one-hidden-layer FFNN to predict the biomechanical model parameters corresponding to each bin independently from the neural activity in a 50-ms, immediately preceding time window. The neural activity was represented by the average firing rate of each cluster, split into 1-ms bins. To avoid introducing temporal correlations, we randomized the order in which each pair of neural activity window and target model parameters was presented to the network. After training, we predict the values of the biomechanical model parameters corresponding to a test set of neural activity and integrate the differential equations of the model to produce each bin of neurally driven synthetic song. This yields synthetic vocalizations that sound similar to the bird's own. An example motif from each bird is illustrated in Figure 2 (and Audio S1, S2, S3, and S4).

In contrast, implementing a FFNN to directly predict the spectrotemporal features of a song results in a low-quality synthesis. We trained a similar network as before but with the spectral components of the song, as represented by the power across 64 frequency bands, as the targets. Examples of songs synthesized in this way for each bird (Figure 3; Audio S1, S2, S3, and S4) show how the FFNN fails to produce well-defined harmonic stacks that are typical of the zebra finch song and to faithfully reproduce vocal onsets and offsets.

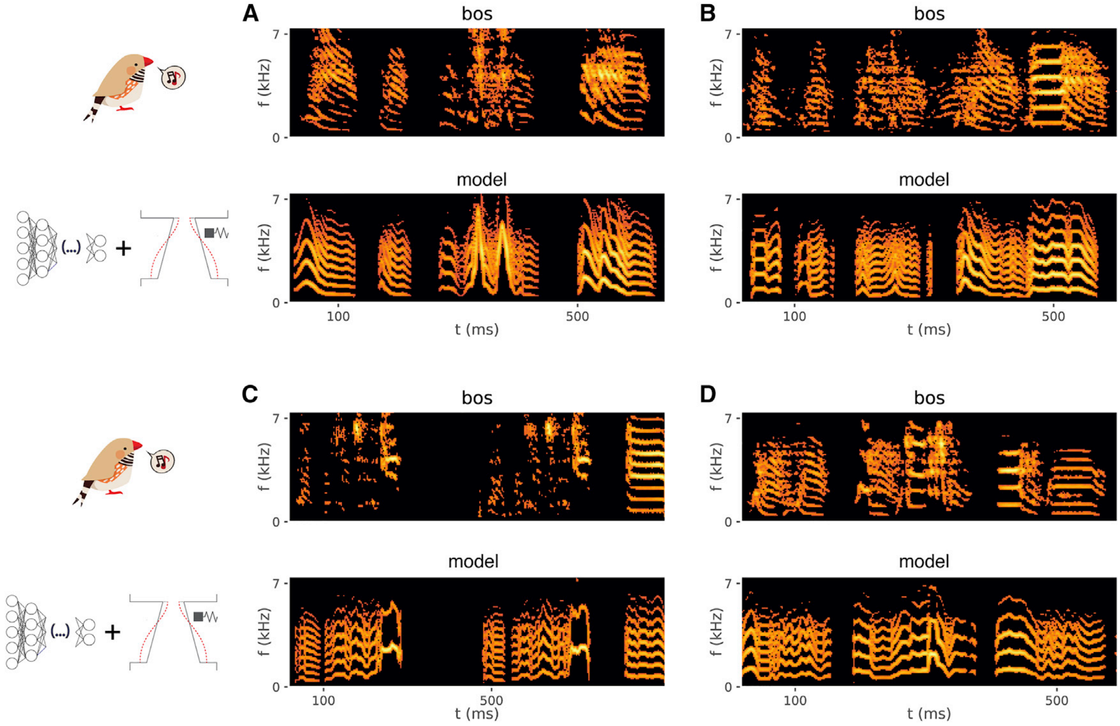


Figure 2. Song synthesized from premotor neural activity via a biomechanical model of the vocal organ is similar to the recorded bird's own song

Spectrogram of one or two instances of a bird's motif (BOS; upper) and corresponding song generated by inferring the biomechanical model parameters from neural activity using a shallow FFNN and integrating the model, for four different birds (z007, z017, z020, and z028, respectively; see also Audio S1, S2, S3, and S4, respectively, and Figures S2 and S4).

The differing capacities for the FFNN to predict model parameters compared to spectro-temporal coefficients (Figures 2, 3, and 4) suggest that reducing the dimensionality of the behavior enhances prediction. To confirm, we trained the FFNN to reproduce a different "compression" of the behavior, namely the first 3 principal components (PCs) of the spectrogram. The performance at predicting the values of the 3 PCs from the neural activity is similar to that at predicting the biomechanical model parameters (Figure S4A). The advantage of the latter is in their generative capacity to produce songs more similar to the BOS.

The biophysical model can be integrated in real time to produce synthetic vocalizations online.[38] This motivated us to skip the spike-sorting step and use a representation of the neural activity that requires cheaper computation and no manual curation. Instead of representing the neural activity with clustered spikes, we trained the FFNN to predict the biomechanical model parameters directly from supra-threshold events in each electrode signal.[1,3,40–42] The results (Figures 4, S2, and S3) suggest that it is plausible to replace the spike sorting by a computationally cheaper representation of the neural activity without significant deterioration of the synthesis.

Exploiting temporal structure

The failure of the FFNN to accurately predict the spectral coefficients of a bird's motif may reflect the inability of this model to capture more complex temporal dynamics across response clusters that precede specific vocalizations. To capture these dynamics, we trained an LSTM[25] to predict the spectral components of the song (64 frequency bands) directly from the preceding 50 ms of neural activity, using the same input and output data as described in the previous section. Unlike the FFNN, the LSTM yields a neurally driven song synthesis that sounds similar to the intended bird's own song (Figure 3; Audio S1, S2, S3, and S4).

Because zebra finch songs are highly stereotyped across renditions from the same singer, we wondered whether the LSTM might be capturing a trivial correspondence between two synchronized signals (the neuronal activity and the song). To rule out this possibility, we independently shuffled the spectral coefficients in each time slice of the spectrograms. This created a novel stereotyped "song," in which the explicit relationship between neural activity and any given acoustic feature varied across time, but the temporal correlation between the average neural activity and the waveform envelope was unchanged. After normalizing the durations of all instances of the motif in a bird's own song through dynamic time warping,[43] we permuted the coefficients within each spectral time slice using the same pseudo-random mask for each motif. The quality of the LSTM synthesis dropped significantly for the pseudo-random stereotyped songs (Figure S4C), indicating a non-trivial relationship between temporal dynamics of HVC population responses and the spectro-temporal characteristics of natural song.

To assess the relative similarity of all our syntheses to BOS quantitatively, we compared the spectrograms of each synthesized and target motif pair using two different metrics. We

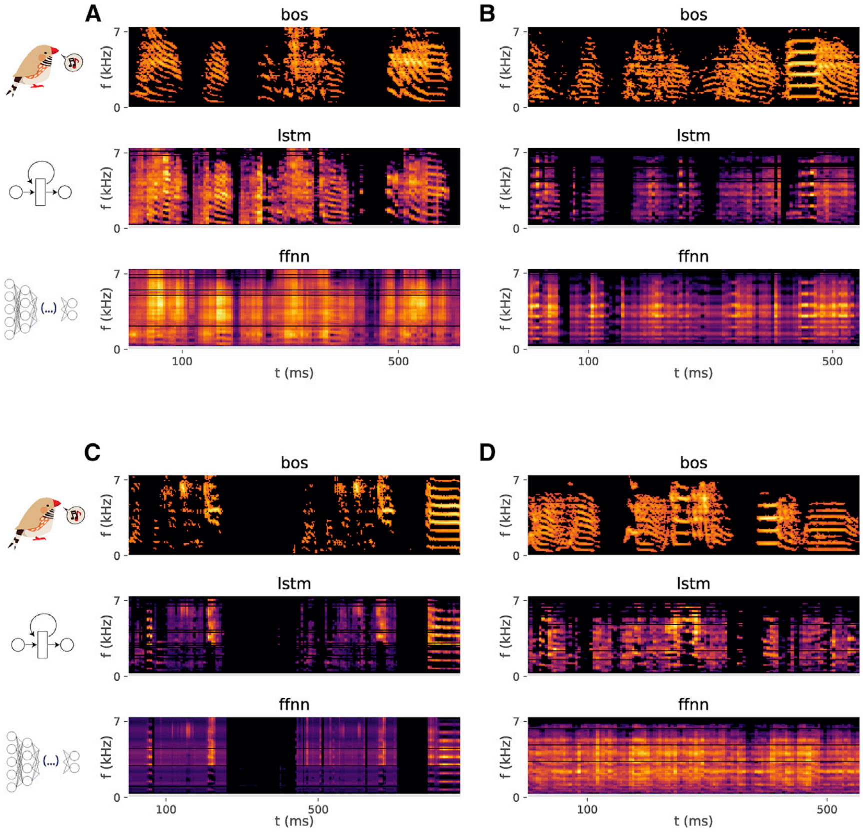


Figure 3. Direct spectrogram synthesis from neural activity

Spectrogram of a bird's motif (BOS; upper), song generated by inferring the spectrogram directly from neural activity, after training an LSTM (middle), and song generated by feeding neural activity and spectrograms to a FFNN of similar architecture to the one used for Figure 2 (bird z007, z017, z020, and z028, respectively, for each panel. See also Audio S1, S2, S3, and S4, respectively, and Figure S4).

computed the mean distance between spectrograms as the earth mover's distance (d_{EMD})[44] between each pair of corresponding spectral slices, averaged across time. Intuitively, d_{EMD} measures the work required to transform one distribution of spectral power onto another and is sensitive to differences in spectral richness (harmonic stacks versus pure tones or broadband energy distribution) and vocalization onset or offset timing. Figure 4A displays a summary of pairwise mean distances $< d_{EMD} >$ aggregated for all birds (Figure S3 shows data grouped by bird). We show the distance between each motif of BOS and the corresponding synthesis achieved with each different strategy. For reference, we compute the distance between all pairs of motifs of BOS ($bos_i - bos_j$) and the distances between pairs of BOS and conspecific birds' motifs ($bos_i - con_j$). We also computed the mean spectral correlation $< \rho >$ between each synthetic motif and its target across the span of each motif. Results are presented in the same manner in Figures 4B and S3. Consistent with the intuition from Figures 2 and 3, both measures show that an FFNN trained to directly predict the spectrogram yields poorer song synthesis compared to the LSTM and the biomechanical model-aided network, even when trained with supra-threshold neural events.

We show that it is possible to synthesize a rich vocal behavior from neural ensemble activity recorded in singing songbirds, a well-established animal model for vocal communication. Similarity of the synthesis with respect to the bird's intended vocalization is significantly enhanced by either compressing behavior into a low-dimensional parameter space or by exploiting the spectro-temporal correlation structure of song by the synthesis algorithm. Our results provide insight into how BMIs for complex behavior may be enhanced through detailed understanding of the underlying biomechanics of motor control and the statistical structure of the target behavior.

Introducing a biomechanical model of the vocal organ enables dimensionality reduction and generativity. Compressing the behavior into a few dimensions enables the use of cheap computations that can be implemented in real time[38] with relatively small training sets. As a compressive model, the "latent" space provided by the time-varying parameters of the biomechanical model is attractive in that it is a proxy for muscular and respiratory activities.[24] In principle, however, other low-dimensional representations of behavior should also be reconstructable from neural activity. Indeed, the 3 strongest PCs of the spectrogram can be predicted by a simple feedforward network almost as well as the parameters of the biomechanical model (Figures S4A and S4B), although these PCs yield poorer synthetic song (Figure S4A) compared to the biomechanical model (Figures 4A and S4B). As a generative model, song synthesized by the biomechanical model is similar to the BOS (Figure 2) and can evoke responses in neural circuits that are highly selective to the BOS.[27,45–47] It remains to be tested whether the biomechanical model-generated song will be sufficient for sensory feedback in closed-loop experiments where neurally driven synthesis replaces BOS[38] and whether the synthetic vocalizations can drive naturalistic responses in females, given the species' ability to differentiate songs based on fine differences in spectro-temporal structure.[39,48,49] Perhaps more detailed models are necessary to fulfill certain functions (BOS replacement and successful courtship). It is also likely that a similar approach can be translated to other species and motor behaviors that admit a low-dimensional, generative representation. This could be the case when the biomechanics are sufficiently understood[6,22,41,50–56] or when there are enough examples of the behavior to enable data-driven dimensionality reductions.[6,7,57]

The limited repertoire of male zebra finches might suggest that a direct synthesis could be achieved by relatively simple means. Yet the FFNN trained to predict spectral coefficients, which, because of its loss function, is close to a regularized nonlinear regression, yielded poor-quality songs compared to all other methods (Figures 4 and S3). The reason for this is not entirely clear, but it may reflect

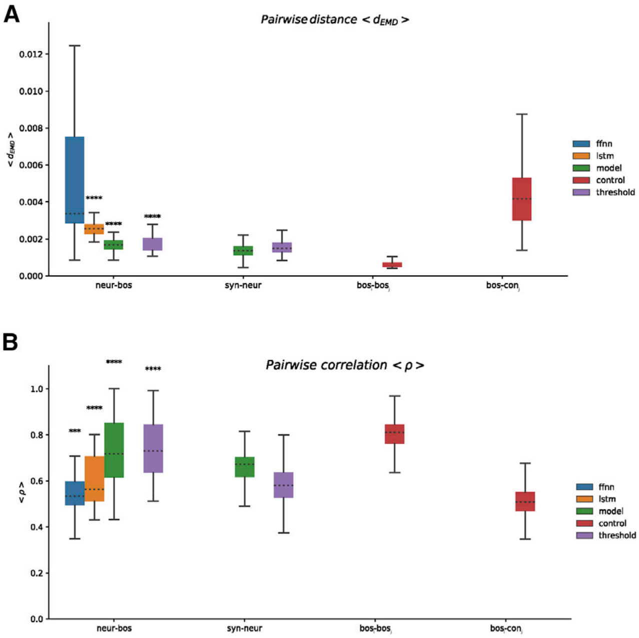


Figure 4. Performance comparisons aggregated for all birds

(A) Boxplots showing time-averaged earth mover's distance ($< d_{EMD} >$; lower = better). Neur-bos, distance between each pair of synthesized or target spectrograms in the test set. Ffnn and lstm indicate training directly with spectrogram via a FFNN and LSTM using sorted spikes, model indicates training or synthesis via the biomechanical model of the vocal organ using sorted spikes, and threshold indicates the same training or testing as in model, albeit using supra-threshold activity instead of sorted spikes. Syn-neur indicates the distance between the synthetic instance of each motif in the testing set (the one produced when fitting the parameters of the biomechanical model for a given motif) and the one synthesized from neural activity. Bos_i-bos_j indicates distance between each pair of motifs of BOS. Bos_i-con_j indicates distance between pairs of bos and songs from a pool of conspecific birds.

(B) Boxplots showing time-averaged spectral correlation ($< \rho >$; higher = better); same pairs as in (A) (***$p < 0.001$; ****$p < 0.0001$; Mann-Whitney U test, one sided against bos_i-con_j). Performance for each bird is shown in Figure S3.

the neuronal subtype compositions of our datasets. Unfortunately, the relatively low yield of well-isolated single units in our recordings (Figure S1) prevents us from examining the contributions of HVC interneurons and projection neurons directly. Most clusters in our datasets were MUA, which are likely dominated by interneuron activity.[29] Although the precise function of HVC interneuron activity is not fully understood, it plays a prominent role in multiple models of HVC[29,31,32] sequence generation, and fluctuations in the average firing rates of interneurons are closely timed to bursting in projection neurons.[15] While one should not interpret our results to support the presence of an explicit "motor code" in HVC ensemble activity, employing a recurrent network that captures the temporal structure of our neural population activity, rather than a FFNN, nonetheless yields synthetic songs that are much closer matches to the birds own song.

We have demonstrated a BMI for a complex communication signal, using computation blocks that are implementable in real time in an established animal model for production and learning of complex vocal behavior. The strength of our approach lies in the ability to find a low-dimensional parameterization of the behavior in a manner that it can be driven with the activities recorded from relatively small samples (by tens) of neurons. Doing so with recordings from the superficially located nucleus HVC enables accessibility by less invasive micro-electrode arrays, capable of resolving not only LFP, which has been shown suitable for BMI,[41] but also SUA and MUA.[1,3,40–42,58,59] This provides a novel tool for probing the neural circuits underlying the production, acquisition, and maintenance of vocal communication signals and unlocks access to new models and experiments directed at understanding how neuronal activity is transformed into natural action and how peripheral effectors shape the neural basis of action.[22,54] Our approach also provides a proving ground for vocal prosthetic strategies. While birdsong differs in important ways from human speech, the two vocal systems have many similarities, including features of the sequential organization and strategies for their acquisition,[60,61] analogies in neuronal organization and function,[10,12] genetic bases,[9] and physical mechanisms of sound production.[19,23] The experimental accessibility, relatively advanced understanding of the neural and peripheral systems, and status as a well-developed model for vocal production and learning make songbirds an attractive animal model to advance speech BMI, much like the nonhuman primate model for motor BMI.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Subjects
  - ○ Ethical note
- METHOD DETAILS
  - ○ Neural and audio recordings
  - ○ Electrode implant
  - ○ Dataset preparation
  - ○ Biomechanical model of the vocal organ
  - ○ Song waveform generation
  - ○ Spectrum shuffle mask
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Performance Evaluation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.05.035.

ACKNOWLEDGMENTS

We thank K. Perks, B. Theilman, M. Thielk, and T. Sainburg for critical discussions and/or input on the manuscript. We acknowledge illustrations and artwork by Ondina Arneodo and Mariano Chena. This work was supported by the US National Institutes of Health (grant R01DC018446), the Kavli Institute for the Brain and Mind (IRG no. 2016-004), the Office of Naval Research (MURI N00014-13-1-0205), the US National Science Foundation (Graduate Research Fellowship Grant No. DGE-1650112 to D.E.B.), the University of California—Historically Black Colleges and Universities Initiative (D.E.B.), and a Pew Latin American Fellowship in the Biomedical Sciences (E.M.A.). The Tesla K40, Titan Z, and Titan X Pascal used for this research were donated by the NVIDIA Corporation.

AUTHOR CONTRIBUTIONS

E.M.A., S.C., D.E.B., V.G., and T.Q.G. participated in the design of the experiment and the writing of the manuscript. E.M.A. built the experimental setup and collected the data. E.M.A., S.C., and D.E.B. analyzed the data. T.Q.G. supervised the study.

DECLARATION OF INTERESTS

V.G. is a compensated consultant of Paradromics, Inc., a brain-computer interface company.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an under-represented minority in science. One or more of the authors of this paper received support from a program designed to increase minority representation in science.

Received: August 28, 2019
Revised: April 3, 2021
Accepted: May 18, 2021
Published: June 16, 2021

REFERENCES

1. Gilja, V., Nuyujukian, P., Chestek, C.A., Cunningham, J.P., Yu, B.M., Fan, J.M., Churchland, M.M., Kaufman, M.T., Kao, J.C., Ryu, S.I., and Shenoy, K.V. (2012). A high-performance neural prosthesis enabled by control algorithm design. Nat. Neurosci. *15*, 1752–1757.
2. Velliste, M., Perel, S., Spalding, M.C., Whitford, A.S., and Schwartz, A.B. (2008). Cortical control of a prosthetic arm for self-feeding. Nature *453*, 1098–1101.
3. Gilja, V., Pandarinath, C., Blabe, C.H., Nuyujukian, P., Simeral, J.D., Sarma, A.A., Sorice, B.L., Perge, J.A., Jarosiewicz, B., Hochberg, L.R., et al. (2015). Clinical translation of a high-performance neural prosthesis. Nat. Med. *21*, 1142–1145.
4. Hochberg, L.R., Bacher, D., Jarosiewicz, B., Masse, N.Y., Simeral, J.D., Vogel, J., Haddadin, S., Liu, J., Cash, S.S., van der Smagt, P., and Donoghue, J.P. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature *485*, 372–375.
5. Akbari, H., Khalighinejad, B., Herrero, J.L., Mehta, A.D., and Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. Sci. Rep. *9*, 874.
6. Chartier, J., Anumanchipalli, G.K., Johnson, K., and Chang, E.F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron *98*, 1042–1054.e4.
7. Anumanchipalli, G.K., Chartier, J., and Chang, E.F. (2019). Speech synthesis from neural decoding of spoken sentences. Nature *568*, 493–498.
8. Doupe, A.J., and Kuhl, P.K. (1999). Birdsong and human speech: common themes and mechanisms. Annu. Rev. Neurosci. *22*, 567–631.
9. Pfenning, A.R., Hara, E., Whitney, O., Rivas, M.V., Wang, R., Roulhac, P.L., Howard, J.T., Wirthlin, M., Lovell, P.V., Ganapathy, G., et al. (2014). Convergent transcriptional specializations in the brains of humans and song-learning birds. Science *346*, 1256846.
10. Jarvis, E.D. (2019). Evolution of vocal learning and spoken language. Science *366*, 50–54.
11. Brainard, M.S., and Doupe, A.J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. Nat. Rev. Neurosci. *1*, 31–40.
12. Long, M.A., Katlowitz, K.A., Svirsky, M.A., Clary, R.C., Byun, T.M., Majaj, N., Oya, H., Howard, M.A., 3rd, and Greenlee, J.D.W. (2016). Functional segregation of cortical regions underlying speech timing and articulation. Neuron *89*, 1187–1193.
13. Gadagkar, V., Puzerey, P.A., Chen, R., Baird-Daniel, E., Farhang, A.R., and Goldberg, J.H. (2016). Dopamine neurons encode performance error in singing birds. Science *354*, 1278–1282.
14. Vallentin, D., Kosche, G., Lipkind, D., and Long, M.A. (2016). Neural circuits. Inhibition protects acquired song segments during vocal learning in zebra finches. Science *351*, 267–271.
15. Lynch, G.F., Okubo, T.S., Hanuschkin, A., Hahnloser, R.H.R., and Fee, M.S. (2016). Rhythmic continuous-time coding in the songbird analog of vocal motor cortex. Neuron *90*, 877–892.
16. Roberts, T.F., Hisey, E., Tanaka, M., Kearney, M.G., Chattree, G., Yang, C.F., Shah, N.M., and Mooney, R. (2017). Identification of a motor-to-auditory pathway important for vocal learning. Nat. Neurosci. *20*, 978–986.
17. Fee, M.S., Kozhevnikov, A.A., and Hahnloser, R.H.R. (2004). Neural mechanisms of vocal sequence generation in the songbird. Ann. N Y Acad. Sci. *1016*, 153–170.
18. Giret, N., Kornfeld, J., Ganguli, S., and Hahnloser, R.H.R. (2014). Evidence for a causal inverse model in an avian cortico-basal ganglia circuit. Proc. Natl. Acad. Sci. USA *111*, 6063–6068.
19. Elemans, C.P.H., Rasmussen, J.H., Herbst, C.T., Düring, D.N., Zollinger, S.A., Brumm, H., Srivastava, K., Svane, N., Ding, M., Larsen, O.N., et al. (2015). Universal mechanisms of sound production and control in birds and mammals. Nat. Commun. *6*, 8978.
20. Srivastava, K.H., Holmes, C.M., Vellema, M., Pack, A.R., Elemans, C.P.H., Nemenman, I., and Sober, S.J. (2017). Motor control by precisely timed spike patterns. Proc. Natl. Acad. Sci. USA *114*, 1171–1176.
21. Goller, F., and Cooper, B.G. (2004). Peripheral motor dynamics of song production in the zebra finch. Ann. N Y Acad. Sci. *1016*, 130–152.
22. Takahashi, D.Y., Fenley, A.R., Teramoto, Y., Narayanan, D.Z., Borjon, J.I., Holmes, P., and Ghazanfar, A.A. (2015). Language development. The developmental dynamics of marmoset monkey vocal production. Science *349*, 734–738.
23. Titze, I.R. (1988). The physics of small-amplitude oscillation of the vocal folds. J. Acoust. Soc. Am. *83*, 1536–1552.
24. Perl, Y.S., Arneodo, E.M., Amador, A., Goller, F., and Mindlin, G.B. (2011). Reconstruction of physiological instructions from zebra finch song. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *84*, 051909.
25. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.
26. Leonardo, A., and Fee, M.S. (2005). Ensemble coding of vocal control in birdsong. J. Neurosci. *25*, 652–661.
27. Amador, A., Perl, Y.S., Mindlin, G.B., and Margoliash, D. (2013). Elemental gesture dynamics are encoded by song premotor cortical neurons. Nature *495*, 59–64.
28. Picardo, M.A., Merel, J., Katlowitz, K.A., Vallentin, D., Okobi, D.E., Benezra, S.E., Clary, R.C., Pnevmatikakis, E.A., Paninski, L., and Long, M.A. (2016). Population-level representation of a temporal sequence underlying song production in the zebra finch. Neuron *90*, 866–876.
29. Liberti, W.A., 3rd, Markowitz, J.E., Perkins, L.N., Liberti, D.C., Leman, D.P., Guitchounts, G., Velho, T., Kotton, D.N., Lois, C., and Gardner, T.J. (2016). Unstable neurons underlie a stable learned behavior. Nat. Neurosci. *19*, 1665–1671.
30. Hahnloser, R.H.R., Kozhevnikov, A.A., and Fee, M.S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature *419*, 65–70.
31. Yu, A.C., and Margoliash, D. (1996). Temporal hierarchical control of singing in birds. Science *273*, 1871–1875.
32. Kozhevnikov, A.A., and Fee, M.S. (2007). Singing-related activity of identified HVC neurons in the zebra finch. J. Neurophysiol. *97*, 4271–4283.
33. Pachitariu, M., Steinmetz, N.A., Kadir, S.N., Carandini, M., and Harris, K.D. (2016). Fast and accurate spike sorting of high-channel count probes with KiloSort. In Advances in Neural Information Processing Systems 29, D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett, eds. (Curran Associates), pp. 4448–4456.

34. Wild, J.M. (1997). Neural pathways for the control of birdsong production. J. Neurobiol. *33*, 653–670.

35. Goller, F., and Larsen, O.N. (1997). A new mechanism of sound generation in songbirds. Proc. Natl. Acad. Sci. USA *94*, 14787–14791.

36. Perl, Y.S., Arneodo, E.M., Amador, A., and Mindlin, G.B. (2012). Nonlinear dynamics and the synthesis of zebra finch song. Int. J. Bifurcat. Chaos *22*, 1250235.

37. Arneodo, O. (2021). Zebra finch electrophysiology illustrations. Figshare. https://doi.org/10.6084/m9.figshare.14577156.v1.

38. 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.

39. Lohr, B., and Dooling, R.J. (1998). Detection of changes in timbre and harmonicity in complex sounds by zebra finches (Taeniopygia guttata) and budgerigars (Melopsittacus undulatus). J. Comp. Psychol. *112*, 36–47.

40. Christie, B.P., Tat, D.M., Irwin, Z.T., Gilja, V., Nuyujukian, P., Foster, J.D., Ryu, S.I., Shenoy, K.V., Thompson, D.E., and Chestek, C.A. (2015). Comparison of spike sorting and thresholding of voltage waveforms for intracortical brain-machine interface performance. J. Neural Eng. *12*, 016009.

41. Brincat, S., Jia, N., Salazar-Gómez, A.F., Panko, M., Miller, E., and Guenther, F. (2013). Which Neural Signals are Optimal for Brain-Computer Interface Control? (Verlag der Technischen Universität Graz).

42. Fraser, G.W., Chase, S.M., Whitford, A., and Schwartz, A.B. (2009). Control of a brain-computer interface without spike sorting. J. Neural Eng. *6*, 055004.

43. 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.

44. Peleg, S., Werman, M., and Rom, H. (1989). A unified approach to the change of resolution: space and gray-level. IEEE Trans. Pattern Anal. Mach. Intell. *11*, 739–742.

45. Margoliash, D., and Fortune, E.S. (1992). Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. J. Neurosci. *12*, 4309–4326.

46. Doupe, A.J., and Konishi, M. (1991). Song-selective auditory circuits in the vocal control system of the zebra finch. Proc. Natl. Acad. Sci. USA *88*, 11339–11343.

47. Dave, A.S., and Margoliash, D. (2000). Song replay during sleep and computational rules for sensorimotor vocal learning. Science *290*, 812–816.

48. Prior, N.H., Smith, E., Lawson, S., Ball, G.F., and Dooling, R.J. (2018). Acoustic fine structure may encode biologically relevant information for zebra finches. Sci. Rep. *8*, 6212.

49. Paul, A., McLendon, H., Rally, V., Sakata, J.T., and Woolley, S.C. (2021). Behavioral discrimination and time-series phenotyping of birdsong performance. PLoS Comput. Biol. *17*, e1008820.

50. Assaneo, M.F., Trevisan, M.A., and Mindlin, G.B. (2013). Discrete motor coordinates for vowel production. PLoS ONE *8*, e80373.

51. Tankus, A., Fried, I., and Shoham, S. (2012). Structured neuronal encoding and decoding of human speech features. Nat. Commun. *3*, 1015.

52. d'Avella, A., Saltiel, P., and Bizzi, E. (2003). Combinations of muscle synergies in the construction of a natural motor behavior. Nat. Neurosci. *6*, 300–308.

53. Wiltschko, A.B., Johnson, M.J., Iurilli, G., Peterson, R.E., Katon, J.M., Pashkovski, S.L., Abraira, V.E., Adams, R.P., and Datta, S.R. (2015). Mapping sub-second structure in mouse behavior. Neuron *88*, 1121–1135.

54. Zhang, Y.S., and Ghazanfar, A.A. (2018). Vocal development through morphological computation. PLoS Biol. *16*, e2003933.

55. Tytell, E.D., Holmes, P., and Cohen, A.H. (2011). Spikes alone do not behavior make: why neuroscience needs biomechanics. Curr. Opin. Neurobiol. *21*, 816–822.

56. Guenther, F.H., and Brumberg, J.S. (2011). Brain-machine interfaces for real-time speech synthesis. Annu Int Conf IEEE Eng Med Biol Soc *2011*, 5360–5363.

57. Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507.

58. Khodagholy, D., Gelinas, J.N., Thesen, T., Doyle, W., Devinsky, O., Malliaras, G.G., and Buzsáki, G. (2015). NeuroGrid: recording action potentials from the surface of the brain. Nat. Neurosci. *18*, 310–315.

59. Hermiz, J., Hossain, L., Arneodo, E.M., Ganji, M., Rogers, N., Vahidi, N., Halgren, E., Gentner, T.Q., Dayeh, S.A., and Gilja, V. (2020). Stimulus driven single unit activity from micro-electrocorticography. Front. Neurosci. *14*, 55.

60. Lipkind, D., Zai, A.T., Hanuschkin, A., Marcus, G.F., Tchernichovski, O., and Hahnloser, R.H.R. (2017). Songbirds work around computational complexity by learning song vocabulary independently of sequence. Nat. Commun. *8*, 1247.

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

62. Arneodo, E., and Gentner, T.Q. (2021). Chronic recordings in HVC with silicon probe arrays in singing zebra finches. Figshare.

63. Yamahachi, H., Zai, A.T., Tachibana, R.O., Stepien, A.E., Rodrigues, D.I., Cavé-Lopez, S., Lorenz, C., Arneodo, E.M., Giret, N., and Hahnloser, R.H.R. (2020). Undirected singing rate as a non-invasive tool for welfare monitoring in isolated male zebra finches. PLoS ONE *15*, e0236333.

64. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.; SciPy 1.0 Contributors (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.

65. Sitt, J.D., Amador, A., Goller, F., and Mindlin, G.B. (2008). Dynamical origin of spectrally rich vocalizations in birdsong. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *78*, 011905.

66. Riede, T., Suthers, R.A., Fletcher, N.H., and Blevins, W.E. (2006). Songbirds tune their vocal tract to the fundamental frequency of their song. Proc. Natl. Acad. Sci. USA *103*, 5543–5548.

67. Riede, T., Schilling, N., and Goller, F. (2013). The acoustic effect of vocal tract adjustments in zebra finches. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. *199*, 57–69.

68. Kinsler, L.E., Frey, A.R., Coppens, A.B., and Sanders, J.V. (1999). Fundamentals of Acoustics, Fourth Edition (Wiley).

69. Boari, S., Perl, Y.S., Amador, A., Margoliash, D., and Mindlin, G.B. (2015). Automatic reconstruction of physiological gestures used in a model of birdsong production. J. Neurophysiol. *114*, 2912–2922.

70. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

71. Kingma, D.P., and Ba, J. (2014). Adam: a method for stochastic optimization. arXiv, arXiv:1412.6980v9. https://arxiv.org/abs/1412.6980.

72. Griffin, D., and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust. Speech Sig. Proc. *32*, 236–243.

73. McFee, B., Lostanlen, V., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., Zalkow, F., Malek, A., Lee, K., et al. (2020). librosa/librosa: 0.8.0. Zenodo.

74. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

75. 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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Original data	This paper, Figshare	https://doi.org/10.6084/m9.figshare.14502198
Software and algorithms		
Code for acquisition, pipeline, analysis written in C/Python.	This paper	https://github.com/zekearneodo/swissknife
Code for analysis written in Python.	This paper	https://github.com/kaichensh/curr_bio_2021
Other		
Printable Microdrive and signal conditioning hardware designs.	This paper	https://github.com/singingfinch/bernardo

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed and will be fulfilled by the lead contact, Timothy Q. Gentner (tgentner@ucsd.edu).

Materials availability

Printable hardware and electronic designs developed for this work are available in the following github repository: https://github.com/singingfinch/bernardo.

Data and code availability

Data generated in this study have been deposited to https://doi.org/10.6084/m9.figshare.14502198.

Code for data acquisition, processing pipeline and analysis developed for this work is available in the following github repositories: https://github.com/zekearneodo/swissknife; https://github.com/kaichensh/curr_bio_2021.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects

Electrophysiology data was collected from n = 4 adult (> 120 dph) male zebra finches.[62] Birds were individually housed for the entire duration of the experiment and kept on a 14:10h light:dark cycle. The birds were not used in any other experiments.

Ethical note

All procedures were approved by the Institutional Animal Care and Use Committee of the University of California (protocol number S15027).

METHOD DETAILS

Neural and audio recordings

We used 4-shank, 16/32 site Si-Probes (Neuronexus A4x2-tet-5mm-150-200-121 -PEDOT coated, bird z007-; Buzsaki32 -pedot coated, bird z028-; A4x1-tet-3mm-150-121 -birds z017 & z020-). We mounted the probes on an in-house designed, printable microdrive and implanted them targeting nucleus HVC. Audio was registered with a microphone (Earthworks M30) connected to a preamplifier (ART Tube MP). Extracellular voltages and pre-amplified audio were amplified and digitized at 30kHz using an Intan RHD2000 acquisition system, Open ephys and custom software. Ref and gnd were shorted together via a 0 Ohm resistor on the headstage (Intan RHD2132/RHD2116).

Electrode implant

Animals were anesthetized with a gaseous mixture of Isoflurane/oxygen (1%–2.5%, 0.7 lpm) and placed in a stereotaxic frame. Analgesia was provided by means of a 2mg/kg dose of carprofen (Rimadyl) administered I.M. The scalp was partially removed and the upper layer of the skull over the y-sinus was uncovered. The probe was attached to the shaft of a microdrive of our design (https://github.com/singingfinch/bernardo/tree/master/hardware/printable_microdrive) which was printed in-house using a B9 Creator printer and the BR-9 resin. A craniotomy site was open 2400 μm lateral to the y-sinus (right/left hemispheres). The dura was removed, and the electrode array was lowered to a 300-500 μm depth. The opening was then covered with artificial dura (DOWSIL 3-4680 Silicone Gel Kit) and the microdrive was cemented to the skull using dental cement (C&B Metabond). A reference wire was made with a 0.5 mm segment of platinum-iridium wire (0.002'') soldered to a silver wire lead and inserted between the dura and the skull in through a craniotomy roughly 3mm medial (contralateral to the hemisphere where the electrode was inserted) and 1.5 mm anterior to the y-sinus. The reference electrode was cemented to the skull and the silver lead was soldered to the ref and gnd leads of the Neuronexus probe. Most of the open area, including the electrode and the microdrive, was covered with a chamber designed and 3D printed in house, which was cemented to the skull. The skin incision was sutured and glued to the chamber with superglue. The mass of the probe, microdrive and protective chamber was 1.2-1.4g. Upon returning to a single-housing cage, a weight reliever mechanism was set up: an end of a segment of thin nylon wire (fishing line) was attached to an ad hoc pin in the chamber; the other end routed through a set of pulleys and attached to a counterweight mass of ~1g.[63]

Dataset preparation

Song detection

A template matching filter written in python was used to find putative instances of the motif, and then curated manually to rule out false positives.

Spike sorting

Spikes were detected and sorted using Kilosort; details of the procedure can be found in Pachitariu et al.[33] The number of clusters was initialized to 32/64 (twice the number of channels of the probe) and the algorithm was allowed to automatically merge similar clusters. In post hoc curation, we removed the clusters that were visibly noise (as per the waveform) and labeled units as putatively SUA/MUA depending on whether the fraction of refractory period (2ms) violations was below/above 3% respectively.

Single Unit type classification

SUA clusters were classified as putatively representing sparse firing projection neurons or tonically firing interneurons, based on their base firing rate and their bursting behavior. We labeled a SUA cluster a putative projection neuron if its mean, spontaneous firing rate was below 5Hz and it produced at most 4 bursts with a frequency of 100 Hz or higher during the motif.[32]

Supra-threshold event detection

We wrote scripts in Python to detect spiking events in each channel. First, the RMS of each channel was estimated using a running window, over a period of time that ranged from minutes to an hour. Then, events that deviated in absolute value more than a number of RMS (2.5-5.5) were detected using the package *peakutils* (min_distance = 0.5ms).

Neural activity features

With all clusters spike-sorted or supra-threshold events, we extracted spike counts within each motif and collapsed them into 1ms (30 samples at 30,000 samples/second) time bins.

Spectral features

When training the networks with spectral features, the target at each time step was a vector containing a spectrogram slice (in log power scale). We generated the spectral slices using the spectrogram function of the signal module in the scipy package.[64] We used 5ms windows (150 samples) and kept the 64 first bands above 300 Hz.

Biomechanical model of the vocal organ

Model

A model of the zebra finch vocal organ has been previously introduced and it is explained in detail in Perl et al.[36] and Arneodo et al.[38] This model considers mainly a sound source and a vocal tract that further shapes the acoustics of the vocalizations.

The source (syrinx) comprises two sets of tissues or labia that can oscillate induced by the sub-syringeal pressure and modulate the airflow to produce sound.[35] The motion of the labia is represented as a surface wave propagating in the direction of the airflow, that can be described in terms of the lateral displacement of the midpoint of the tissue.[23] Its mathematical form is the motion equation of a nonlinear oscillator in which two parameters that determine the acoustic features of the solutions are controlled by the bird: the sub-syringeal air sac pressure and the stiffness of the restitution (through the activity of syringeal muscles). In order to integrate the model in real time, a set of equations was found that is computationally less expensive yet capable of displaying topologically equivalent sets of solutions as the parameters are varied:[65]

$$\begin{cases} \dfrac{dx}{dt} = y \\ \dfrac{dy}{dt} = \gamma^2\alpha + \gamma^2\beta x + \gamma^2 x^2 - \gamma^2 x^3 - \gamma xy - \gamma x^2 y \end{cases}$$

where x represents the departure of the midpoint position of the oscillating labia, γ is a time scaling factor, and the parameters α and β are functions of the air sac pressure and the activity of the ventral syringeal muscle, respectively.

The upper vocal tract further shapes the sound produced by the source, determining spectral properties such as the timbre. We used a model for the vocal that includes a tube, accounting for the trachea, followed by a Helmholtz resonator, accounting for the oropharyngeal-esophageal cavity (OEC)[66,67] (see Figure 1A in Arneodo et al.[38]). The pressure at the input of the tube that represents the trachea is $P_i(t) = ax(t) - r\,x(t-\tau)$, where $ax(t)$ is the contribution to the fluctuations by the modulated airflow, r is the reflection coefficient at the opposing end of the tube of length L and $\tau = 2L/c$, with c the sound velocity. The pressure fluctuations at the output of the trachea force the air at the glottis, approximated by the neck of the Helmholtz resonator that represents the OEC. The mass of air at the glottis, forced into the cavity, is subject to a restitution force exerted by the larger mass of air in it.

In acoustics, it is common to write an analog electric computational model to describe a system of filters. The acoustic pressure is represented by an electric potential and the volume flow by the electric current.[68] In this framework, short constrictions are inductors, and cavities (smaller than the wavelengths) are well represented by capacitors. The equations for the equivalent circuit of the post-tracheal part of the vocal tract, (see Figure 1B in Arneodo et al.[38]) read:

$$\begin{cases} \dfrac{di}{dt} & = & \Omega_1, \\ \dfrac{d\Omega_1}{dt} & = & -\dfrac{1}{L_g C_h} i_1 - R_h \left(\dfrac{1}{L_b} + \dfrac{1}{L_g} \right) \Omega_1 + \\ & & + i_3 \left(\dfrac{1}{L_g C_h} - \dfrac{R_b R_h}{L_b L_g} \right) + \dfrac{1}{L_g} \dfrac{dV_{ext}}{dt} + \dfrac{R_h}{L_g L_b} V_{ext}, \\ \dfrac{di_3}{dt} & = & -\dfrac{L_g}{L_b} \Omega_1 - \dfrac{R_b}{L_b} i_3 + \dfrac{1}{L_b} V_{ext}, \end{cases}$$

where the electric components relate to geometric parameters of acoustic elements, and are described in detail in Perl et al.[36] and Arneodo et al.[38] The pressure fluctuations at the glottal end of the trachea relate linearly to the electric tension V_{ext} driving the circuit. Following the same scheme, the electrical potential at the resistor standing for the beak $V_b = i_3 R_b$ is the analog of the pressure fluctuations at the output of the beak. In our model, this quantity is the sound radiated by the vocal organ.

Parameter fitting

In order to fit the parameter series that will lead to reconstruction of the song, we perform a procedure similar to that previously described.[27,36] Timescale parameter is set to a value of 23,500; α is set to -0.15 during vocalization and 0.15 otherwise, and β is set in order to minimize the distance in the (pitch, spectral content) space between the synthesized and the recorded song segments;[36] the envelope (e(t) in the main text) is obtained by rectifying and smoothing the recorded waveform; the parameters of the vocal tract were fixed, in the same values as in Perl et al.[36] In order to extract the pitch of the song, we follow a modification of the automatic procedure presented in Boari et al.,[69] and we add a layer of manual curation. When integrating the model, we apply the extracted envelope (e(t)) as an extra multiplicative factor when computing $ax(t)$, since it recovers the amplitude fluctuations that were discarded when reducing the model to its normal form and driving it with the bi-valued parameter α. The parameters accounting for the geometry of the vocal tract are constants and are set to the same values as in Perl et al.[24]

Neural network training

Neural network -based decoders were implemented in python 3.6, using Tensorflow 2.0 and Keras. They were run on Ubuntu 16.04 and 18.04 PCs equipped with NVidia GPUs (Tesla k40c, Titan Xp, and Titan X Pascal). CUDA version was 10.2.

LSTM network architecture

The network has 2 layers of LSTM cells, with Nx5 cells in the first layer and N in the second, where N is the number of clusters in the neural data. The output layer has as many relu units as the target space (64 for the spectrogram bands). The input of the network is a Nx50 array that contains the spike count of each cluster, in each of 50, 1ms bins preceding the output bin. The output of the network is a 1x64 array containing the spectral bands corresponding to a 5ms bin. Both LSTM layers utilized 20% dropout and 0.001 L2 regularization during training to prevent overfitting.[70]

Feed-forward Network architecture

The network has 1 dense hidden layer of Nx25 relu units, where N is the number of clusters in the neural data. The output layer has as many relu units as the target space (p = 64 for the spectrogram bands, p = 3 for the biomechanical model parameters). The input of the network is a Nx50 vector that contains the spike count of each cluster, in each of 50, 1ms bins preceding the output bin. The output of the network is a 1xP array containing the spectral bands (p = 64) or the biomechanical model parameters (p = 3) corresponding to a 5ms bin. The hidden layer utilized 20% dropout and 0.001 L2 regularization during training

Training procedure

We utilized a gradient-based optimizer (Adam/rmsprop[71]) and mean square error (MSE) as a loss function for LSTM/FFNN. We used 40% of all the motifs for testing and the rest motifs for training. We made 3 passes using non-overlapping motifs as a testing set, in order to have as many decoded examples as the number of motifs in the session. In each pass, all of the neural-activity/decoder-target pairs (one per bin) were fed in random order to the network, both when training and decoding. We reserved 10% of the training set as a validation subset for early stopping, where the training session would be stopped if validation loss failed to decrease within 5/10 training epochs. Figure 4 shows the results of this motif -based training averaged across all birds. As an alternative training method, we masked a fraction of each motif (roughly 3.3%), trained on the complement, then generated the song corresponding to the masked fraction. We repeated this piece-wise procedure tiling the whole motif, and generated entire motifs using segments of data that were novel to the decoder. Figure S3 shows the results of both the motif-wise and piece-wise training for individual birds.

Song waveform generation

Spectrogram inversion

We used LSEE-STFTM algorithm to invert spectrograms back to audio waves,[72] as implemented in the librosa python package.[73] The algorithm iteratively estimates a signal from the short-time Fourier transform magnitude (STFTM), through minimizing the mean square error between the short-time Fourier transform (STFT) and the estimated STFT, and subsequently performs STFT on the estimated signal, the magnitude of which will be passed on to the next iteration.

Within each iteration, a signal was approximated using the equation below:

$$x(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n)\, y_w(mS,\, n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}$$

where $x(n)$ denotes the estimated signal; $w(n)$ denotes the analysis window used in STFT. The variable S is a positive integer, representing the sampling rate of the STFT. Here, $y_w(mS,\, n)$ is the target signal corresponding to $Y_w(mS,\, n)$, which denotes the target STFTM, in our case spectrogram powers. To calculate in each iteration, we used a sinusoidal window:[72]

$$w_s(n) = \frac{2w_r(n)}{\sqrt{4a^2+2b^2}}\left[a + b\cos\left(\frac{2\pi n}{L} + \phi\right)\right]$$

where L represents the length of the window. Here, $w_r(n)$ is a rectangular window with an amplitude of $\sqrt{S/L}$ within $0 \le n < L$ and zero anywhere outside. A modified Hamming window can be obtained by setting $a = .54$, $b = -.46$, $\phi = \frac{\pi}{L}$. After obtaining an $x(n)$ value within each iteration, the STFT of $x(n)$was calculated, which was used in place of $Y_w(mS,\, n)$in the next iteration. The squared error between the target STFTM and the estimated STFTM is proven to decrease in each iteration of the algorithm.

Biomechanical model integration

Once the model parameters are predicted by the decoder, they are re-sampled and fed to an ordinary differential equation integrator. Resampling to 30 Khz is performed (with cubic interpolation). A fourth order runge-kutta ODE integrator (custom coded) integrates then the equations of the model with a time step of (900 KHz)$^{-1}$.

Synthesis through principal components

PC decomposition was made using the PCA module in the scikit-learn package.[74] We obtained the principal component decomposition of all the spectrograms of all the motifs sung by each bird during the length of the experiment. 512 frequency bins were used for the spectrograms, which were concatenated and projected onto the N principal components $T_{N\,=}$ S W_N, where S is the 512-dimension spectrum time series and W_N the transformation matrix (the matrix of the L eigenvectors of SS^T with the largest eigenvalues) via the pca.fit and pca.fit_transform methods. For reconstruction from principal components, the inverse transformation was applied by means of the pca.inverse_transform method.

Spectrum shuffle mask

Time warping

We adopted a simplified version of Dynamic Time Warping (DTW[75]) specific to zebra finch songs. Instead of segmenting the song into different syllables and matching each syllable to different syllable templates, we took advantage of the stereotypical nature of zebra finch songs and directly computed minimal distance matrices (D) between each song-level spectrogram and a spectrogram template. Starting at the first slice of each spectrogram,

$$D(i,\, j) = d(i,\, j) + min\{D(i-1,\, j)\ iff\ w_j(l-1) \neq w_j(l-2),\ D(i-1,\, j-1),\ D(i-1,\, j-2)\}$$

Where i indexes the time frames of the input pattern, j indexes the time frames of a single template, l indexes the ordered steps along a specific path. d(i, j) is the local distance between slice i and slice j. w_j(l) denotes the specific step at l in the space of j. Once a distance matrix D was calculated, we determined an optimal path with the lowest cumulative distance between the input and the template, and proceeded to stretch, delete or keep each input slice, depending on the path.

Masking

We applied a random yet consistent shuffling mask, P, to our entire warped spectrogram repertoire so that spectral consistency across time is disrupted while the temporal pattern within each motif remains. For the i-th spectrogram slice in each warped song, we shuffled all 64 spectral elements using the same shuffling pattern P_i. Treating all spectrograms with the same shuffling mask P enabled us to determine whether our model is decoding the spectral information within birdsongs or recreating the same pattern regardless of spectral consistency across time. In our shuffling training session, we used the shuffled spectrograms as output.

Reordering mask

After training, we tested our model on novel neural data, the target of which were also shuffled spectrograms. In order to visually compare our model's performance with and without shuffling, we reordered the reconstructed shuffled spectrograms. We achieved this by applying a reordering mask, R, that traces and reverses all the shuffling done through the aforementioned shuffling mask P. For any spectrogram S, $R(P(S)) = S$.

QUANTIFICATION AND STATISTICAL ANALYSIS

Performance Evaluation

Root Mean Square Error (RMSE)

We used RMSE between each pair of original and predicted spectrogram magnitude as a metric to evaluate the performance of our models.

Spectral correlation

To obtain the spectral correlation across time for a pair of spectrograms, we first computed the pearson correlation coefficient between each corresponding pair of spectral slices that conform the two spectrograms (via the function pearsonr from the stats module of the scipy python package[64]). Then, we obtained the time-averaged value across the span of the motif.

Earth mover's distance

To obtain the distance across time for a pair of spectrograms, we computed the earth mover's distance (d_{EMD}) or Wasserstein metric between each pair of spectral slices that conform the two spectrograms (via the function wasserstein_distance from the stats module of the scipy python package[64]). Prior comparison, each spectral slice was normalized such that the total area under the slice be 1; for silences, a value of 1 was assigned to the first bin of the spectrogram. Then, we obtained the time-averaged value across the span of the motif.

Spectrogram Normalization

In order to account for variations among motifs from different birds, we normalized spectrograms for each bird so that the collection of original spectrograms for each bird had a maximum power of 1 and minimum power of 0:

$$\widehat{p_i} = \frac{p_i - p_{max}}{p_{max} - p_{min}}$$

Where p_i is the power of a point on either an original spectrogram or a predicted spectrogram before normalization, while $\widehat{p_i}$ is the normalized power of the corresponding point. p_{max} denotes the maximum power of the entire set of original spectrograms, while p_{min} represents the minimum power of the entire set of original spectrograms. With such normalization, we were able to account for variations among motifs from different birds while keeping the variations within motifs from the same bird.

Pairwise performance comparisons

We performed comparisons among and between different sets of songbirds (displayed in Figure 4 boxplots for instance). *BOS-BOS:* comparisons provide a baseline of the variability of the bird's own motifs during the session: comparison across each pair of motifs. *neur-bos:* comparison across each pair of natural motifs and it's corresponding one decoded from neural activity.

In order to provide an extra control reference, we also computed spectrogram comparisons against a set of 47 motifs from conspecific birds (other zebra finches; about half of them from our colony and half from other colonies). This produced the sets: *BOS-CON:* comparisons across each BOS motif and all of the conspecific (CON) motifs.