J Neurophysiol 114: 2912–2922, 2015.
First published September 16, 2015; doi:10.1152/jn.00385.2015.

Automatic reconstruction of physiological gestures used in a model of birdsong production

Santiago Boari,[1] Yonatan Sanz Perl,[1] Ana Amador,[1] Daniel Margoliash,[2] and Gabriel B. Mindlin[1]
[1]*Department of Physics, FCEN, University of Buenos Aires and IFIBA, CONICET, Buenos Aires, Argentina; and*
[2]*Department of Organismal Biology and Anatomy, University of Chicago, Chicago, Illinois*

Submitted 21 April 2015; accepted in final form 15 September 2015

Boari S, Perl YS, Amador A, Margoliash D, Mindlin GB. Automatic reconstruction of physiological gestures used in a model of birdsong production. *J Neurophysiol* 114: 2912–2922, 2015. First published September 16, 2015; doi:10.1152/jn.00385.2015.—Highly coordinated learned behaviors are key to understanding neural processes integrating the body and the environment. Birdsong production is a widely studied example of such behavior in which numerous thoracic muscles control respiratory inspiration and expiration: the muscles of the syrinx control syringeal membrane tension, while upper vocal tract morphology controls resonances that modulate the vocal system output. All these muscles have to be coordinated in precise sequences to generate the elaborate vocalizations that characterize an individual's song. Previously we used a low-dimensional description of the biomechanics of birdsong production to investigate the associated neural codes, an approach that complements traditional spectrographic analysis. The prior study used algorithmic yet manual procedures to model singing behavior. In the present work, we present an automatic procedure to extract low-dimensional motor gestures that could predict vocal behavior. We recorded zebra finch songs and generated synthetic copies automatically, using a biomechanical model for the vocal apparatus and vocal tract. This dynamical model described song as a sequence of physiological parameters the birds control during singing. To validate this procedure, we recorded electrophysiological activity of the telencephalic nucleus HVC. HVC neurons were highly selective to the auditory presentation of the bird's own song (BOS) and gave similar selective responses to the automatically generated synthetic model of song (AUTO). Our results demonstrate meaningful dimensionality reduction in terms of physiological parameters that individual birds could actually control. Furthermore, this methodology can be extended to other vocal systems to study fine motor control.

dynamical systems; vocal learning; bird's own song; peripheral vocal production model; modeling software

BIRDSONG PRODUCTION REQUIRES the exquisite coordination of multiple sensory and motor systems. The motor commands generated in the songbird's central nervous system (CNS) are translated to the periphery into muscular activity driving the syrinx (the avian vocal organ) and the respiratory system. Singing-related information modulates auditory responses in the forebrain (Coleman et al. 2007; Keller and Hahnloser 2009), and auditory feedback potentially has multiple pathways into the forebrain song system (Bauer et al. 2008; Cardin and Schmidt 2004; Mandelblat-Cerf et al. 2014). Information from the brain stem respiratory nuclei ascends bilaterally (Reinke and Wild 1998; Striedter and Vu 1998) and ultimately reaches forebrain song system nuclei. This suggests that bottom-up information may regulate top-down control of singing more strongly than has been traditionally conceived (Schmidt et al. 2004). This new perspective on birdsong production places it in the context of principles of neuromechanical control that have been elucidated for many systems (Nishikawa et al. 2007).

How to conceptualize information processing in the song system given this wealth of pathways and influences has been challenging. Is there a potentially simplifying organizing principle? The common physical phenomenon involved in the production of all birdsong sounds is the modulation of the airflow passing through the syringeal labia (Mindlin and Laje 2005). Songbirds accomplish this by generating self-sustained oscillations of the labia, presumably by coordinated dynamical recruitment of the complex musculature of the syrinx. This acts in coordination with dynamical respiratory patterns (Goller and Suthers 1996a, 1996b; Suthers et al. 1999). Finally, upper vocal tract filters strongly modulate the sound emitted by the syrinx. The study of the physical mechanisms behind birdsong production has included the generation of synthetic songs with simple models (Gardner et al. 2001; Laje et al. 2002), the direct measurement of the parameters proposed to be controlling the syrinx (Mindlin et al. 2003), and the study of the correlation between the conjectured time-dependent parameters needed to synthesize songs and the recorded parameters used in the generation of the song (Perl et al. 2011). This line of research opens the possibility that a low-dimensional model of birdsong production capable of producing realistic sounds would provide, through its time-dependent parameters, a proxy for the output of the forebrain song system during singing.

Recent results have assessed the biological significance of such low-dimensional models by measuring neuronal responses to songs generated by a biomechanical model of song production (Amador et al. 2013). The songs of that study represented synthetic copies of the bird's own song (BOS). The strategy was using a replay phenomenon found in several species of songbirds. Neurons in the forebrain's song system nucleus HVC are highly selective to auditory presentation (playback) of BOS (Doupe and Konishi 1991; Margoliash 1986). Similar phasic/tonic patterns of activity are observed for individual neurons while the bird is singing and for BOS playback (Dave and Margoliash 2000; Prather et al. 2008). In zebra finches (*Taeniopygia guttata*), this replay phenomenon is highly state dependent: the neurons respond to auditory renditions of BOS when the bird is asleep or anesthetized, but the auditory response is highly diminished when the bird is awake (Cardin and Schmidt 2003; Dave et al. 1998; Rauske et al. 2003; Schmidt and Konishi 1998). These neurons are deemed

Address for reprint requests and other correspondence: A. Amador, Intendente Guiraldes 2160, Pabellon 1, Dept. Fisica, Ciudad Universitaria, Buenos Aires (1428), Argentina (e-mail: anita@df.uba.ar).

selective on the basis that they give weaker or no responses to other auditory stimuli, including white noise, simple tones, songs of conspecifics, and spectrally or temporally modified versions of BOS (Margoliash 1983). With this replay phenomenon, it was observed in sleeping zebra finches that robust responses of neurons in the telencephalic song system nucleus HVC were elicited by auditory presentations of BOS and its synthetic copy (SYN) but very weak or no responses to the songs of conspecific birds (CON) or to BOS played in reverse (REV) (Amador et al. 2013). This synthesis was driven by a small number of time-dependent parameters emulating physiological instructions that the bird can control during the song production process.

Beyond the validation of a low-dimensional description for the biomechanics of birdsong production, the small number of time-dependent parameters required to generate the synthesis motivated the introduction of the concept of motor "gestures" in birdsong production: a sequence of trajectories occurring near bifurcations in a parameter space of subsyringeal air sac pressure and syringeal membrane tension. An intriguing result was that the timing of the gesture trajectory extrema (GTEs) gave insight into the neural code in HVC (Amador et al. 2013). The possibility of interpreting activity at different areas of the CNS in terms of parameters directly controlling behavior opens the possibility of addressing unresolved problems of motor coding and sensorimotor integration.

An important aspect of this program is to have an automatic procedure to reconstruct trajectories in parameter space that can drive the production model to synthesize realistic songs. In Amador et al. (2013) the authors made algorithmic choices manually in the song reconstructions. The songs were reconstructed prior to the electrophysiological experiments, revealing SYN as one of the stimuli effectively stimulating HVC neurons and thus demonstrating the validity of the manual approach. It remains to be determined, however, what solution space yields effective SYN stimuli. Manual fitting might in principle be biased toward some regimes of the solution space and/or be biased against other regimes of possible solutions. Furthermore, the fitting procedure is complex and requires a level of bioacoustics knowledge that makes it challenging to adopt. Finally, many of the songs studied in Amador et al. (2013) were relatively simple, containing harmonic stacks (almost constant fundamental frequency and high spectral content) and syllables with well-defined fundamental frequencies, preferentially with long notes. These syllables are optimal for distinguishing between the GTE representations presented in Amador et al. (2013) and the "clock" representation previously hypothesized (Fee et al. 2004; Hahnloser et al. 2002). It would be valuable, however, to evaluate the effectiveness of the production model with a larger number and more complex song types better representing the range of variation in zebra finch song. Thus, if an algorithmic and automatic procedure was shown to have general utility, this would facilitate further research into coding of birdsong vocal production.

Here we develop an algorithmic and automatic method to reconstruct the trajectories in parameter space that are required to synthesize realistic copies of zebra finch song. Importantly, we confirm the validity of the synthetic songs by comparing the activity of highly selective HVC neurons in response to BOS and AUTO. Having an automatic procedure opens the possibility of exploring additional simplifications that might still support realistic song synthesis.

MATERIALS AND METHODS

Automatic Reconstruction of Gestures

The computational model for birdsong production consists of equations describing the dynamics for the separation between the syringeal labia. These are obtained by writing Newton's equations for each labium, which is assumed to be an elastic tissue capable of both lateral displacements as well as sustaining longitudinal waves (for a detailed description see Amador and Mindlin 2008; Gardner et al. 2001; Mindlin and Laje 2005; Perl et al. 2011). The assumed kinematics enables computation of the pressure between the labia, an essential step in the derivation of the physical equations of motion, which ultimately consists of autosustained oscillations. When the equations are integrated for different values of the control parameters (tension of the syringeal labium and the air sac pressure), regions of the parameter space presenting qualitatively similar dynamics can be identified (Amador and Mindlin 2008). The boundaries between those regions are called bifurcation lines. It is possible to further simplify the original system of equations, which emerges from physical considerations, to a simpler set of "normal form" equations (Guckenheimer and Holmes 1997) capable of displaying the same dynamical regimes (Sitt et al. 2010). This simplified dynamical system is our basic computational model and reads

$$\frac{\mathrm{d}x}{\mathrm{d}t} = y$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\alpha\gamma^2 - \beta\gamma^2 x - \gamma^2 x^3 - \gamma x^2 y + \gamma^2 x^2 - \gamma x y \qquad (1)$$

where x represents the position of the syringeal labia, therefore describing the dynamics of one phonating source. In previous work, we have shown that this model is capable of producing realistic copies of songs of several songbird species, including zebra finches (Sitt et al. 2008), canaries (*Serinus canaria*) (Gardner et al. 2001), Chingolo sparrow (*Zonotrichia capensis*) (Laje et al. 2002), and others. In the normal form equations (*Eq. 1*) this is achieved by varying γ (a scaling, static parameter) and the time-dependent, dimensionless parameters α and β, which are related to the air sac pressure and syringeal labial tension, respectively (Amador et al. 2013; Perl et al. 2011). Although this is a simplified version of the sound source [songbird syrinx has two phonating sources whose tension (β) can be independently controlled], many sounds are generated using either one side of the syrinx or both synchronized, and hence the model would be appropriate for reproducing these sounds.

The reason why this simple dynamical system can account for many of the sounds present in zebra finch songs is the existence of different classes of bifurcations that this system presents: qualitatively different ways in which stationary solutions (representing labia movements) transition from stability toward oscillating solutions (representing oscillating labia that periodically obstruct the airflow). The Hopf bifurcation is a mechanism by which oscillations are born with infinitesimally small amplitudes and a well-defined, nonzero frequency (Guckenheimer and Holmes 1997; Strogatz 1994). The periodic obstruction of the flow that occurs in that process has little harmonic content, and therefore the resulting sound is almost tonal. The other mechanism by which oscillations are born occurs for smaller values of the parameter β and is called saddle node in limit cycle (SNILC) bifurcation. In this process, an oscillation is born with substantial amplitude and zero frequency. The oscillations generated in this way have high harmonic content, and therefore correspond to rough sounds. For a detailed comparison between sounds generated through different bifurcations see Figs. 3 and 5 in Amador and

Mindlin (2008). Crossing a SNILC bifurcation, the farther away from the bifurcating curve in parameter space, the larger the frequency of the oscillations and the more tonal the sound gets. Therefore, the dynamical process in which an oscillatory regime is generated will determine the relationship between the harmonic content of the signal and its frequency. The sound source model used here has been extensively characterized, e.g., showing that oscillations born in a SNILC bifurcation give rise to sounds whose harmonic (spectral) content and fundamental frequency obey specific relations (Sitt et al. 2008). This relationship was empirically observed in the vocalizations of zebra finches containing low-frequency sounds (fundamental frequency $<$ 1.2 kHz). In this way, a preliminary step for our reconstruction procedure consists of generating a database where, for different values of the parameters (α, β) we list the fundamental frequency (ω) and the spectral content index (SCI), which is the centroid of the spectrum normalized to the fundamental frequency (see Sitt et al. 2008). The database is generated by integrating the equations of our complete model [nonlinear sound source and linear filters (Perl et al. 2011)] for each point (α,β) in parameter space. Then, the generated solutions are spectrally analyzed. In this work, we used a value of $\alpha = 0.15$ to represent "on" states (i.e., solutions corresponding to phonation). The range of β values explored for solutions with different values of fundamental frequency and spectral content was $0.002 < \beta < 2.99$. This range of α and β values allowed us to synthesize sounds with fundamental frequencies between 413 Hz and 6,780 Hz, which correspond to the range of fundamental frequencies found in zebra finch song.

We worked with sound files sampled at 44,100 Hz. We performed a sliding-window Gabor filtering on the time series data by defining segments of 1,024 samples centered at each point. For each segment, a Gabor filter was applied ($\sigma = 220$ samples). This filtering eliminates sharp edges in the sliding-window spectrographic analysis, achieved by a fast Fourier transform to each filtered segment. For segments corresponding to phonation, clear maxima above a threshold were obtained in the power spectrum. Each segment qualifying as phonation had at least one frequency for which the square root of its power was larger than 12,000 (arbitrary units). This corresponds to ~5% of the power at the fundamental frequency of a typical zebra finch vocalization. If the segment qualified, the first peak between 400 Hz and 8,000 Hz was identified as the fundamental frequency. The procedure was repeated for every sample in a phonating interval. Autocorrelation function (ACF)-based pitch detection algorithms (PDAs) were tested, including YIN (De Cheveigne and Kawahara 2002; Tchernichovski et al. 2000). These methods efficiently capture fast frequency sweeps but tend to present many artifacts as frequency jumps in segments with slower, subtler amplitude modulation. Since we consider continuity to be an important feature in the motor gesture hypothesis, first peak from power spectrum was chosen over ACF-PDAs because it allowed us to extract a continuous time trace of the fundamental frequency for each phonating segment.

Once a fundamental frequency was computed from the sound segment, the β value associated with the closest fundamental frequency was retrieved from the database. Since the variation of SCI along isofrequencies in the range of α values where SNILC bifurcations take place is small, we chose in this work $\alpha_{on} = 0.15$, $\alpha_{off} = -0.15$, leaving frequency control to the variations of β. This allows us to simplify the task of reconstructing time-dependent parameters leading to realistic synthesis of song. The pair ($\alpha = 0.15$,β) computed at every sample of a phonating interval provides the time-dependent parameters of the equations ruling the behavior of the labia. This constitutes a computational simplification with respect to previous studies, in which both (α,β) were chosen to simultaneously minimize the differences between the SCI and fundamental frequency of the synthetic sounds and the recorded sound segments. In this new reconstruction strategy, β represents the tension, while the pressure is approximated by a series of "on" and "off" values that place the dynamical system within the oscillatory or stationary regimes, respectively. Note that although we present a computational simplification for the parameters of the sound source oscillations, in physical terms we are not reducing the dimensionality compared with previous descriptions of the problem (Amador et al. 2013; Perl et al. 2011). As shown below, a proxy for the pressure is still needed in order to synthesize realistic songs.

To synthesize sound, one has to emulate a time-dependent airflow. This requires the geometrical variable $x(t)$ computed with the dynamical system (*Eq. 1*) and the average velocity of air through the lumen. The sound envelope was used as a proxy for the average velocity since the sound envelope will monotonically follow the air sac pressure, which is a monotonic function of the average velocity. We computed the envelope by integrating a first-order differential equation driven by the rectified sound trace:

$$\frac{d\alpha_{env}}{dt} = -\frac{1}{\tau}\alpha_{env} + \left|\text{sound}(t)\right| \quad (2)$$

with $\tau = 1$ ms.

Following the results of Perl et al. (2011) and Amador et al. (2013), the synthesis also includes two successive filters: a closed-open tube representing the trachea and a Helmholtz resonator accounting for the oropharyngeal-esophageal cavity (Fletcher et al. 2006; Riede et al. 2006). The effect of the tracheal tube is to differentially amplify frequency components around 2,500 Hz and 7,500 Hz of the signal emerging from the syrinx. The effect of the Helmholtz resonator is to amplify frequency components around 4,000 Hz. Operationally, the pressure $P_i(t)$ at the input of the trachea and the pressure fluctuations transmitted to the resonator $P_t(t)$ were computed as

$$\begin{aligned} P_i(t) &= \alpha_{env}(t)x(t) - rP_i(t-T) \\ P_t(t) &= (1-r)P_i(t-0.5T) \end{aligned} \quad (3)$$

where T stands for the time it takes to a sound wave to traverse the trachea back and forth once, and r stands for the reflection coefficient of the wave at the interface between the trachea and oropharyngeal-esophageal cavity (OEC). In our simulations $r = 0.1$, and $T \cong 0.2$ ms.

Note that in order to synthesize realistic songs we need the envelope of the original sounds as a proxy of the air sac pressure (*Eq. 3*). Thus the computation of sound source parameters by fitting the parameters of a nonlinear model is simplified in this algorithmic procedure, but the dimensionality of the physical processes remains unchanged compared with previous approaches (Amador et al. 2013; Perl et al. 2011). The pressure is estimated from the sound itself when the sound envelope is computed.

The final filter, a Helmholtz resonator with losses, was computed by solving the linear system of differential equations representing its equivalent circuit, namely,

$$\begin{aligned} \frac{di_1}{dt} &= \Omega_1 \\ \frac{d\Omega_1}{dt} &= ai_1 + b\Omega_1 + ci_3 + d\frac{dP_t}{dt} + eP_t \\ \frac{di_3}{dt} &= f\Omega_1 + gi_3 + hP_t \end{aligned} \quad (4)$$

where the sound output is proportional to i_3 and P_t is the input to the Helmholtz resonator (the pressure fluctuations transmitted from the trachea). The parameters, in arbitrary units, are $a = -540 \times 10^6$, $b = -7{,}800$, $c = 1.8 \times 10^8$, $d = 1.2 \times 10^{-2}$, $e = 7.2 \times 10^{-1}$, $f = -0.83 \times 10^{-2}$, $g = -5 \times 10^2$, and $h = 10^{-4}$. In prior studies, songs synthesized with a simpler resonance model of the OEC failed to drive HVC neurons (see Supplemental Fig. 1 in Amador et al. 2013).

The final element required for the synthesis of artificial songs is noise (Perl et al. 2011). Absence of noise leads to synthetic sounds eliciting only weak responses in sleeping or anesthetized birds (see

Amador et al. 2013, Supplemental Fig. 1). The noise was added to the parameter representing the labial tension (β), compatible with electromyography (EMG) recordings of syringeal muscles (G. B. Mindlin, personal observations). In the simulations of this work, Gaussian noise was added to β, with amplitude three orders of magnitude smaller than the total β range.

The codes for implementing this procedure can be downloaded from http://www.lsd.df.uba.ar.

Automatic Reconstruction of GTEs

The automated analysis approach developed here can be extended to approximate the timing of GTEs through the analysis of the sound envelope. The key observation is that the transitions between time intervals with qualitative different behaviors in the fundamental frequency are reflected as local minima in the sound envelope. These instances are often computationally more robust than more subtle changes in fundamental frequencies occurring in transitions between some tension gestures. For example, in the transition between an exponential decay and a harmonic stack, the derivatives at the transition point can take similar values, making the identification of the gesture transition very difficult. Thus here we will define the computational steps that we followed to reconstruct these significant time instances (GTEs) from the sound envelope. In RESULTS, we describe the application of the method to different songs.

Computing GTEs

A second, "GTE identification" algorithm was developed. Sound was first Hilbert transformed into $s(t)$. This time trace was then integrated by a one-dimensional linear dynamical system:

$$\frac{dz}{dt} = -\frac{1}{\tau}z + |s(t)| \quad (5)$$

with $\tau = 1$ ms. A Savitzky-Golay filter (Press et al. 2007) was then applied (np = 513, nr = nl = 256, 4th order of the smoothing polynomial). Finally, the obtained time trace was normalized into $n(t)$, our envelope (normalization with respect to the absolute maximum of the envelope). A five-point stencil derivation of the signal was computed and further filtered with a Savitzky-Golay filter (same parameters as above) to obtain $d(t)$, representing a smoothed derivative function of the original song.

The computation of syllable beginnings and ends involved detecting the time instances at which the normalized signal $n(t)$ exceeded a threshold set to 0.025 [as $n(t)$ is a normalized signal, this corresponds to 2.5% of the maximum value]. Intrasyllabic maxima and minima were computed, inspecting the changes in sign of the smoothed derivative $d(t)$. After calculating all the minima, we extracted the significant minima, comparing the value $n(t)$ of the minima with the adjacent maxima, and evaluating that the amplitude ratio between minima and maxima was lower than a factor $\mu_1 < 1$. This avoids having small fluctuations interfering with the identification of the minima of the envelope that correspond to qualitative changes in the sound. A similar criterion was adopted to identify significant maxima, defined as the maxima with an amplitude ratio between the maxima and adjacent minima being higher than a factor $\mu_2 > 1$. With these definitions, we selected as candidates for GTEs the starts and ends of syllables, significant minima (indicating the instances when gesture transitions within a syllable take place), and significant maxima (as proxies of pressure maxima). We chose $(\mu_1, \mu_2) = (0.8, 2.6)$, which minimized the distance between manual GTEs and algorithmically computed ones for four birds previously analyzed in the literature (Amador et al. 2013).

The codes for implementing this procedure can be downloaded from http://www.lsd.df.uba.ar.

Subjects

Experiments were performed on adult male zebra finches in accordance with a protocol approved by the University of Buenos Aires (FCEN-UBA) Institutional Animal Care and Use Committee (C.I.C.U.A.L.).

Stimuli

Before each experiment, songs were recorded with a directional microphone from a male zebra finch individually housed in a sound-isolation chamber. For each bird, the song motif and most common bout length were determined. Stimuli were crafted by repeating a recording of the bird's stereotyped song motif three times consecutively. Intermotif times were adjusted such that the crafted stimulus had a similar temporal pattern, as did a typical song bout. This procedure eliminated variability across motif renditions in a stimulus.

Stimuli presented included the following: *1*) BOS; *2*) BOS played in reverse (REV), in which the temporal structure of individual syllables and the global syllable order were reversed but overall spectrum was the same as the BOS; *3*) song from a conspecific adult male (CON); and *4*) synthetic copy generated by an automatic reconstruction of BOS (AUTO). Twenty presentations of each stimulus were presented, one stimulus every 10 s, with random choice of stimuli between presentations.

Surgeries

Preparatory surgeries were conducted 2 days before the days of experiments. Animals were anesthetized with isoflurane (Baxter Healthcare). The birds were head-fixed in a stereotaxic device, and lidocaine ointment (2.5% Denver Farma) was applied to the scalp, after which the scalp was dissected along the midline. A stainless steel post was then attached to the caudal part of the bird's skull with dental cement and cyanoacrylate.

On the days of experiments, an animal was anesthetized with 20% urethane (60–100 μl total; Sigma, St. Louis, MO) administered into the pectoral muscle in 20- to 30-μl aliquots at 30-min intervals. The bird was placed in a sound-attenuating chamber, and its head was immobilized via the mounted post. Small craniotomies were made over HVC following stereotaxic coordinates, and the dura was opened with an insect pin. Recording microelectrodes were lowered into HVC with a single-axis hydraulic micromanipulator (Narishige MO-10).

Electrophysiological Recordings

Recordings were made with single-channel tungsten microelectrodes (3–5 MΩ, Microprobes). HVC neurons were identified by stereotaxic coordinates, firing rates, and the selective response to BOS. Signals were recorded with a data acquisition board (National Instruments DAQ PCI-6251) interfaced with MATLAB (The MathWorks, Natick, MA). Data acquisition onset was synchronized with stimulus presentation onset (delay $<$ 0.03 ms), achieved by the use of the RTSI bus on the DAQ board (proprietary NI bus). Sampling frequency was set at 20 kHz, and data were band-pass filtered between 300 Hz and 5,000 Hz during acquisition.

Data Analysis and Processing

Spike sorting. Spike detection and sorting were implemented with the software wave_clus (Quiroga et al. 2004), which allows the automatic extraction of the different spike features in the data with a wavelet transform and an automatic classification of the data in different clusters by using superparamagnetic clustering. After the automatic extraction of spikes was performed, we checked the software performance by inspecting spike shape and amplitude, ISI distributions, and separation of the clusters along the wavelet coeffi-

cient space. Typically, one to three neurons were isolated per recording site. Sites where the recordings degraded before the entire stimulus repertoire was presented were discarded.

PSTH analysis. After single units were isolated from the spike records, the trials corresponding to a given stimulus were used to calculate raster plots and poststimulus time histograms (PSTHs) with 10-ms bins. Bin values were converted to spike rates (spikes/s), yielding an average of the time-dependent firing rate of the unit. Mean spike rates were calculated over the duration of the stimuli. The variance of the PSTH for each stimulus across all trials was also calculated, with the rationale that a high variance is consistent with a well-defined neural activity pattern of excitation and inhibition. All measurements were corrected for the spontaneous firing rate of the neuron, by subtracting a spontaneous rate (or variance) measured over an interval of duration equal to the duration of the stimulus. Finally, we normalized the response strengths to the response to BOS, as follows:

$$\mathrm{RRS}_{\frac{X}{\mathrm{BOS}}} = \frac{M_X - M_\mathrm{S}}{M_\mathrm{BOS} - M_\mathrm{S}} \quad (6)$$

$$\mathrm{RRV}_{\frac{X}{\mathrm{BOS}}} = \frac{\sigma_X - \sigma_\mathrm{S}}{\sigma_\mathrm{BOS} - \sigma_\mathrm{S}} \quad (7)$$

where RRS is the relative response strength, RRV is the relative response variability, X stands for either AUTO, REV, or CON, S stands for the spontaneous activity, and M indicates mean and σ variance.

Correlations. Pearson's correlation coefficients were computed to assess the temporal course similarity between the neural responses to BOS with respect to AUTO, CON, and REV motifs. We calculated each stimulus' average motif response by time-shifting the neural response to all subsequent motifs presented in each protocol and aligning them with the first. To achieve this with high temporal resolution, for each protocol we constructed a 10-ms sliding window histogram sampled at the sound signal's sampling frequency. This allowed us to precisely align the neural responses between different motifs. Finally, to construct the stimulus' motif response for each protocol, we computed the mean between all responses and smoothed it with a Savitzky-Golay filter. Pearson's correlation coefficients were computed for each protocol with these smooth histograms of the neural data. When computing correlations between responses to motifs of different lengths (BOS-CON correlation), we padded the shorter signal with a segment of spontaneous activity from the smooth histogram.

RESULTS

Automatically Synthesized Songs

The automatic procedure transformed a recorded song into a series of instructions, which were then capable of driving the dynamical system (*Eq. 1*) to synthesize realistic sounds. Figure 1 displays an example. Note the similarities between a recorded zebra finch song (Fig. 1*A*; sound trace, *top*, sonogram, *bottom*) with the AUTO song (Fig. 1*B*). The power of the dynamical model is reflected in the different spectral contents of sounds of different frequencies. In other words, fundamental frequency and SCI (see MATERIALS AND METHODS) of the sound traces did not require a delicate simultaneous control of several parameters.

Close inspection of the spectrographs shows that for some syllables the estimated fundamental frequencies of the synthesized sounds tend to be lower than the fundamental frequencies in the natural songs. This is an effect of the finite grid that is used to generate the database containing, for different values of

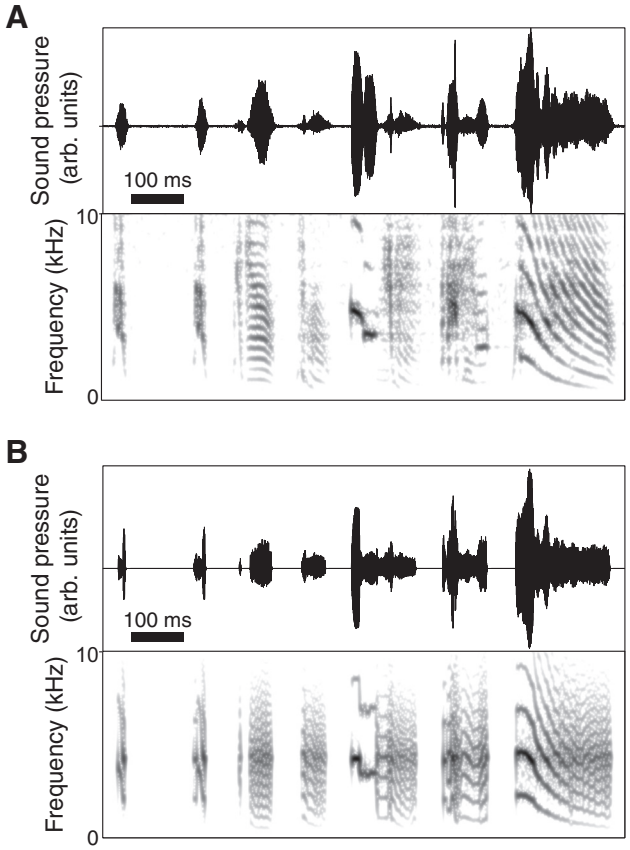

Fig. 1. Song recording and synthesis. Recorded zebra finch song (*A*) and automatically synthesized song (*B*). The sound signal is shown at *top* and its corresponding sonogram at *bottom*. The automatically reconstructed motor instructions that drive the model produce a synthetic copy capturing syllables' fundamental frequency and spectral content.

β, the fundamental frequencies and spectral content indexes. For example, compare the third and the final syllables of the natural song, both of which have clear harmonic structure (Fig. 1*A*), with the corresponding syllables in AUTO (Fig. 1*B*), which have lower fundamental frequencies. The automatic procedure also does not fully capture the most complex, aperiodic sounds. The timing of transitions between distinct vocal gestures, however, is faithfully reproduced.

Automatic reconstruction of the time-dependent parameters was performed for the songs of five adult male zebra finches. This resulted in exploration of a wide range of stereotypical zebra finch song syllables (see Figs. 1, 2*E*, 4*E*, and 5). The range of fundamental frequencies explored with these songs (from ~400 Hz to 6.7 kHz) is represented in the model by differences in the parameter time traces that generate them. The automated procedure is robust in the range of zebra finch vocalizations explored, meaning that it is capable of reproducing both the fundamental frequency and the SCI of each sound segment with the reconstruction performed from the recorded sound signal.

Electrophysiological Experiments

Given the selectively of HVC neurons for BOS over conspecific songs (Margoliash 1986; Margoliash and Konishi 1985) and the sensitivity of HVC to slight changes in the acoustic parameters of BOS (Margoliash 1983; Theunissen and

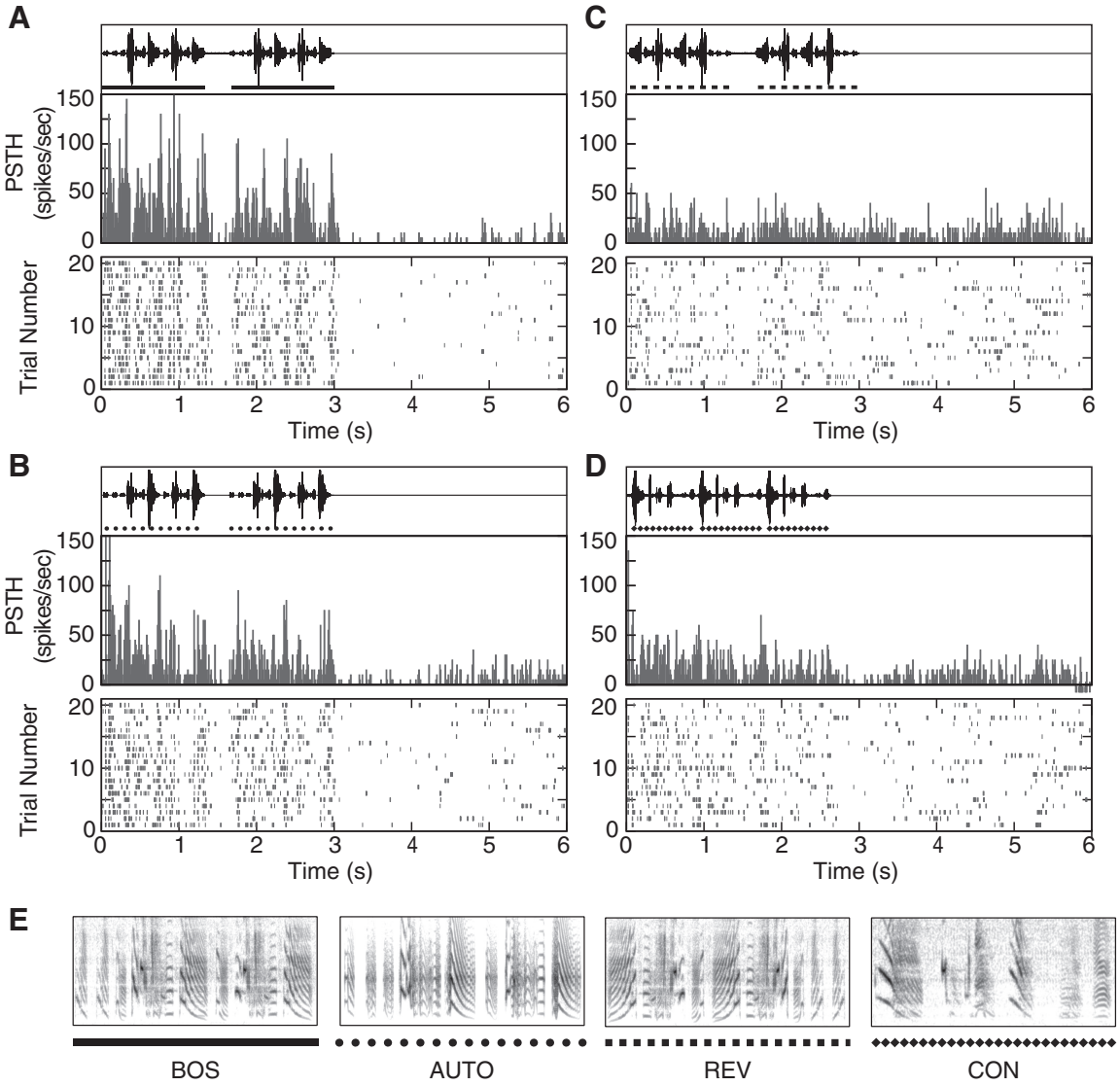

Fig. 2. Neural selectivity experiment. Electrophysiological recordings of the response activity of an HVC single unit to auditory presentations of different stimuli. The protocol consisted of 20 randomly organized auditory presentations of each stimulus. Neural activity was processed through a spike sorting algorithm (wave_clus). *A–D*: raster plots of the 20 trials (*bottom*), poststimulus time histograms (PSTHs; *middle*) and the sound signal of each stimulus presented (*top*). *E*: sonograms for a song motif of each of the presented stimuli. The symbol patterns at *bottom* are the same as in *A–D*. The activity elicited as response to the bird's own song (BOS; *A*) presents a well-defined pattern of excitatory and inhibitory activity; and *B* shows a similar activity pattern elicited by the automatically synthesized song (AUTO). *C* and *D* represent the weak activity in response to presentations of the reverse song (REV) and the song of an adult male conspecific (CON).

Doupe 1998), a strong standard to test the model is whether synthetic songs can drive HVC neurons. Previously, songs synthesized with expert supervision (SYN) were capable of eliciting HVC-selective activity (Amador et al. 2013). Here we assess whether similar results are obtained with the fully automated version of the synthesized song (AUTO).

Twenty selective HVC neurons were isolated in five birds, arising from one to five recording sites per bird. A neuron was considered as BOS selective when it presented a statistically significant response to BOS compared with spontaneous activity (comparing mean spike rates, $P < 0.05$, paired *t*-test) and had stronger responses to BOS than to REV and CON. For the 20 neurons thus selected, the average mean response CON/BOS was 0.20 ± 0.22. This is comparable to the average mean response CON/BOS (0.16 ± 0.34), similarly analyzed, from a larger sample of HVC single units recorded in anesthetized zebra finches in a previous study (Margoliash et al. 1994). Thus the two data sets reflect similar underlying parent distributions.

Anesthetized birds were presented with BOS, REV, AUTO, and the song of a conspecific adult male (CON). Exemplar data for one neuron are shown in Fig. 2. The responses to BOS and AUTO are much stronger than the responses to REV and CON. Moreover, there is a clear temporal pattern of suppression and excitation that is consistent across all trials in BOS and all trials in AUTO and is similar comparing BOS and AUTO. This result is of particular interest taking into account the simplifications of the model and the BOS selectivity of the neurons.

Similar results were obtained for all 20 neurons. All 20 neurons showed tonic excitation in response to BOS and were probably HVC interneurons (Hahnloser et al. 2002). To quantify the response, we computed the PSTH and considered two features from them: the mean activity during stimulus presen-

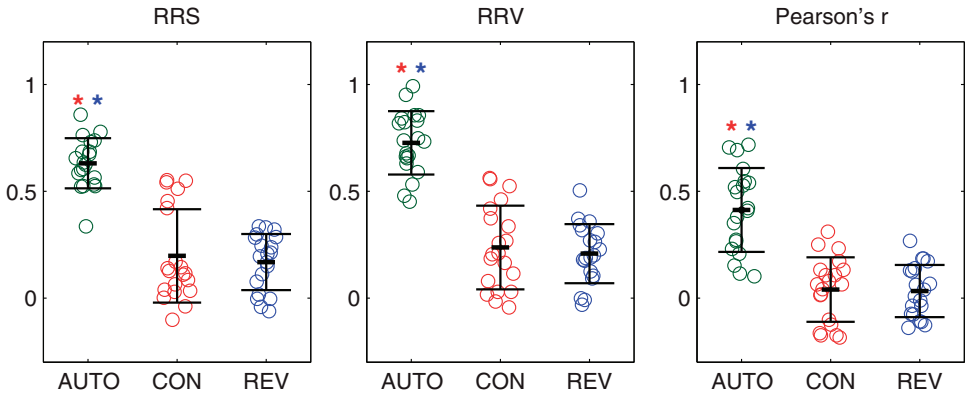

Fig. 3. Results in terms of response strength and variability. Results for the set of experiments performed ($n = 5$ birds, $N = 20$ HVC single units measured). Thick bars represent the mean value, and error bars represent ±SD of the data. RRS: relative response strength ($RRS_{X/BOS}$; see *Eq. 6*): mean $RRS_{AUTO/BOS} = (0.63 \pm 0.12)$, which is significantly larger than the mean of $RRS_{CON/BOS}$ (0.20 ± 0.22) and $RRS_{REV/BOS}$ (0.17 ± 0.13). RRV: relative response variability ($RRV_{X/BOS;}$ see *Eq. 7*): mean $RRV_{AUTO/BOS} = 0.73 \pm 0.15$, which is significantly larger than the mean of $RRV_{CON/BOS}$ (0.24 ± 0.20) and $RRV_{REV/BOS}$ (0.21 ± 0.14). Measurements are corrected in all cases by the spontaneous firing of each unit. Pearson's *r*: pairwise correlation coefficients: mean $r_{BOS\ vs.\ AUTO} = 0.41 \pm 0.20$, which is significantly larger than mean $r_{BOS\ vs.\ CON} = 0.04 \pm 0.15$ and mean $r_{BOS\ vs.\ REV} = 0.03 \pm 0.12$. *Paired *t*-test ($P < 0.001$).

tation and its variance (see MATERIALS AND METHODS). The first quantity reflects whether the overall response is larger during the presentation of the stimulus (mean firing rate across a BOS presentation can get as high as 70 spikes/s) than during spontaneous firing (0.5–26 spikes/s in a window with the same length as BOS). We then computed the RRS of this quantity compared with BOS (see MATERIALS AND METHODS).

The variance reflects whether there is a pattern formed by inhibition and excitation across the stimulus presentation, while RRV measures the magnitude against BOS. Figure 3 shows the results of all the experiments ($n = 5$ birds, $N = 20$ single units), displaying the calculation for each neural response, the mean, and error bars indicating ±SD. Since these are BOS-relative measurements, a value of 1 means the same level of response as BOS and a value of 0 would mean no significant activity in response to that particular stimulus.

In terms of both response strength and response variability, the artificial songs synthesized with automatically reconstructed parameters elicited a significantly larger level of activity than the reverse song and the song of a conspecific adult male (paired *t*-test, $P < 0.01$). It is noteworthy that the $RRS_{AUTO/BOS}$ mean value is 0.63 ± 0.12, significantly larger than the mean value of both $RRS_{CON/BOS}$ and $RRS_{REV/BOS}$, which are 0.20 ± 0.22 and 0.17 ± 0.13, respectively. For RRV we have similar results: the $RRV_{AUTO/BOS}$ mean value is 0.73 ± 0.15, while the $RRV_{CON/BOS}$ mean value is 0.24 ± 0.20 and the $RRV_{REV/BOS}$ mean value is 0.21 ± 0.14.

Features of the pattern of excitatory-inhibitory responses to BOS are also observed in responses to AUTO (Fig. 2). The presence of response patterns is revealed as a high mean value for the $RRV_{AUTO/BOS}$. In other words, AUTO was not only capable of producing an increase in the firing activity of HVC neurons but also did so with a pattern of excitation and inhibition.

To quantify the similarity to the BOS response pattern, we computed the pairwise linear correlation coefficients (Pearson's *r*) between neural responses to BOS with respect to AUTO, CON, and REV. Each neural response was averaged across each stimulus' motif presentations within a protocol and smoothed with a Savitzky-Golay filter (see MATERIALS AND METHODS). The resultant correlation coefficients are AUTO vs. BOS 0.41 ± 0.20, REV vs. BOS 0.03 ± 0.12, and CON vs. BOS 0.04 ± 0.15, and the distributions (with ±1 SD error bars) are shown in Fig. 3, *right*. These results show that there is a shared temporal response pattern to the AUTO and BOS stimuli presented, with the AUTO vs. BOS correlation significantly higher than REV vs. BOS and CON vs. BOS correlations (paired *t*-test, $P < 0.001$). It remains to be seen whether the correlation between AUTO and BOS would be greater in sleeping or singing birds. In summary, these results show that AUTO is able to elicit significant responses in HVC single units by increasing the firing rate of the cell (RRS analysis), that it does so with the presence of a certain pattern of excitation and inhibition in the unit (RRV analysis), and that these response patterns, while not identical, share some of the temporal course features with BOS-elicited patterns (Pearson's correlation analysis).

Figure 4 illustrates an unexpected result: an automatically reconstructed synthetic song capable of eliciting a response that, at specific temporal instances, is stronger than that elicited by BOS (see PSTH at ~125 spikes/s). Since selectivity involves measuring the response strength across the stimulus, even in this case the total response to BOS is larger than the response to AUTO ($RRS_{AUTO/BOS} = 0.53$ and $RRV_{AUTO/BOS} = 0.83$). Yet the reported higher local response can be appreciated in Fig. 4 as a sharpening of the PSTH around specific instances of the song motifs. At those precise times, higher values are reached, and a clear excitatory-inhibitory pattern in the raster plots emerges. Since the highly precise responses take place after particularly pronounced inhibitory valleys, they might be the result of inhibitory rebounds. In the higher-order auditory nucleus caudomedial mesopallium (CMM) in starlings, stimuli comprising isolated notes of a motif can elicit excitatory responses that are not observed when the entire motif is presented, and deleting individual notes from a motif can release excitatory response peaks that are not otherwise present in the unmanipulated motif (Meliza et al. 2010). If similar processes are at work here, then the unanticipated excitatory peaks might result from incomplete modeling of preceding notes causing a release from inhibition. CMM projects to the caudomedial lobule (CLM), a distinct region

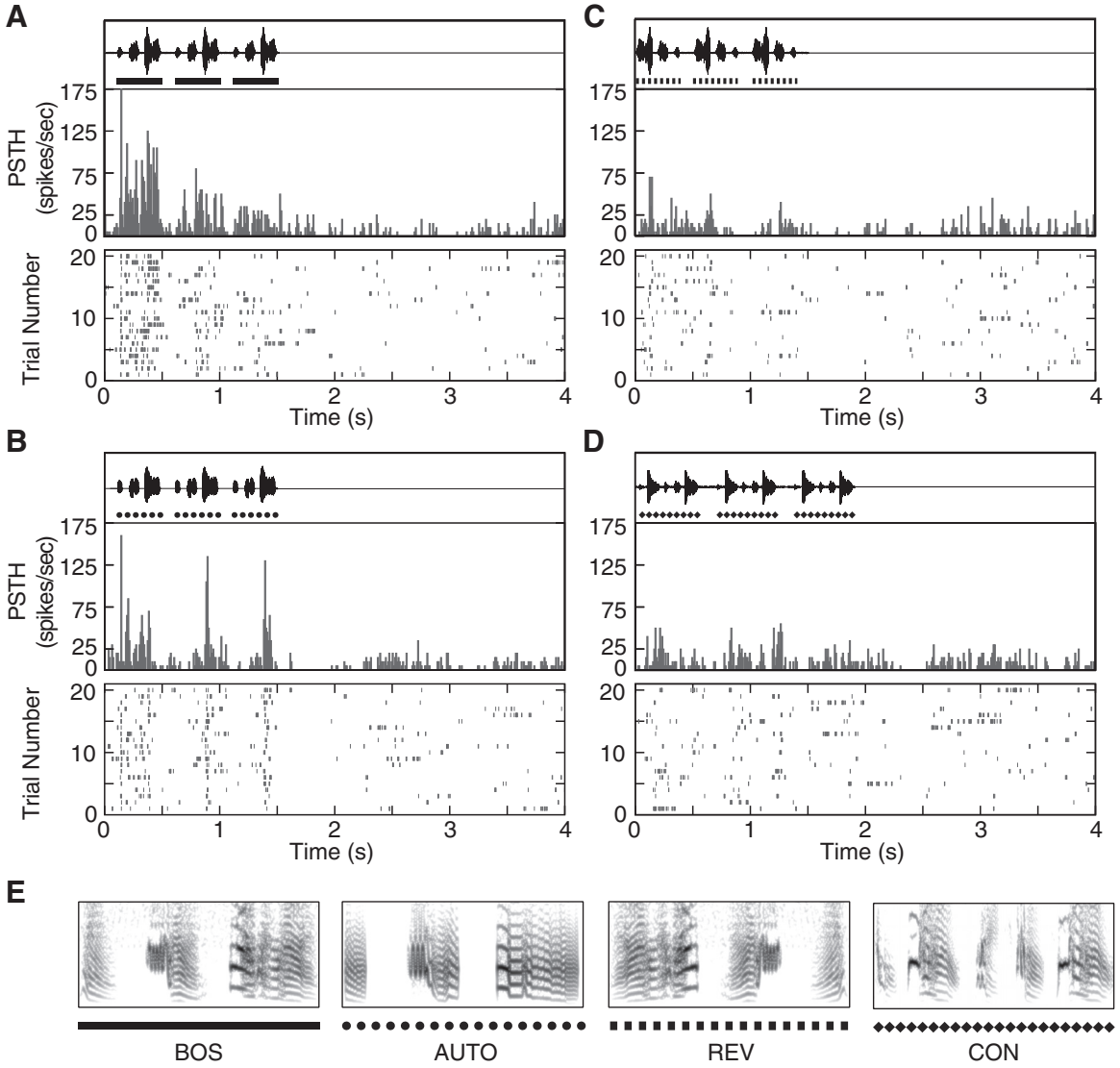

Fig. 4. Neuron presenting higher response to AUTO than to BOS. Same layout and data processing as in Fig. 2 but this time showing the recordings and results for a particular case: a selective HVC neuron that presented locally sharper responses (in terms of the excitatory-inhibitory pattern) to AUTO than to BOS.

of which projects to HVC (Akutagawa and Konishi 2010; Bauer et al. 2008).

Automatic Reconstruction of GTEs

To test the algorithmic procedure of parsing the song into gestures (identifying the times of GTEs), we applied the GTE identification algorithm (see MATERIALS AND METHODS) to the songs of four of the birds analyzed in Amador et al. (2013). Figure 5 displays the result for the song of one of the birds. The sound trace and its spectrogram are illustrated in the top two panels in Fig. 5, while the reconstructed sound envelope is shown in the bottom panel. The dotted vertical lines across the middle and bottom panels in Fig. 5 are located at the time instances selected by running the algorithm described in MATERIALS AND METHODS (automatic identification of GTEs); the square dots indicate the GTEs as identified using a manual reconstruction of gestures (Amador et al. 2013). A total of 25 of the 26 manually computed GTEs in this example were closer than 5 ms of their closest automatically obtained set of 30 GTEs. The largest source of variance comparing the two approaches was that the automated reconstruction tended to identify clusters of closely spaced GTEs in some limited regions of song while this was uncommon with manual reconstruction (Fig. 5). The average time distance between the manually and automatically obtained GTEs for the example shown in Fig. 5 was 1.9 ± 2.8 ms (mean $\pm$ SD). Analogously, for the other three birds analyzed the mean time differences were 2.6 ± 3.0 ms, 1.1 ± 1.3 ms, and 3.5 ± 3.8 ms. These results show that automatically extracted GTEs are in good agreement with previously defined GTEs in Amador et al. (2013).

The reliability of the automatic method was further explored by reconstructing the GTEs from several song renditions for one of the birds in the present study. The results are illustrated in Fig. 6. In Fig. 6*A*, one sound trace and its sonogram are displayed at top and the envelope is shown at bottom. The automatic GTEs for different song renditions are displayed in a raster plot (Fig. 6*B*). We used syllable onsets and offsets as the reference frame to evaluate the timing of GTEs across renditions. We performed a linear time scaling on each syllable

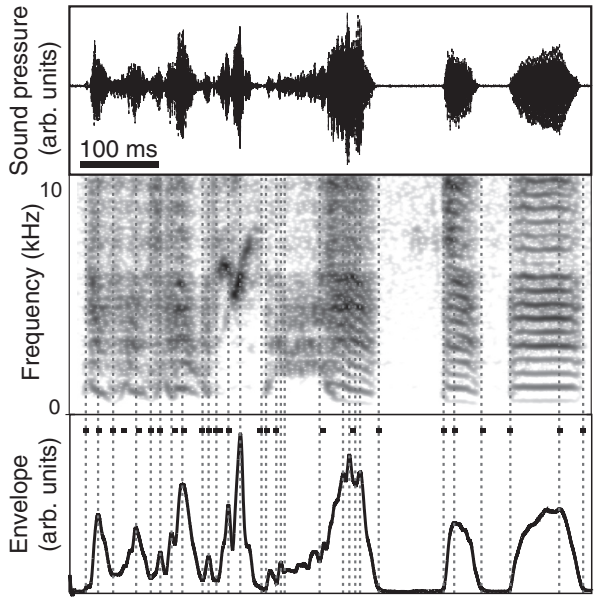

Fig. 5. Automatic reconstruction of gesture trajectory extrema (GTEs). Sound pressure (*top*) and sonogram (*middle*) of a zebra finch song. GTEs were extracted with a manual method (square dots) and by an automatic procedure (dashed lines, see MATERIALS AND METHODS). Automatically extracted GTEs are obtained by finding significant maxima and significant minima of the smoothed envelope of the sound wave (*bottom*), in addition to syllable onsets and offsets. The procedure to manually extract GTEs is explained in Amador et al. (2013).

and then quantified the temporal jitter for each successive GTE within a syllable as the standard deviation for all the presentations of that syllable. Excluding syllable onsets and offsets, for the remaining 10 GTEs, the average dispersion was 3.4 $\pm$ 2.28 ms, with 9 of 10 GTEs showing a dispersion of $<$5.5 ms, with values ranging from 0.84 to 5.48 ms. This result gives additional confidence that the new automated detection of GTEs described here is a robust method.

DISCUSSION

We have tested over the years the hypothesis that low-dimensional dynamics was capable of providing a succinct yet biologically meaningful description of singing behavior (Amador et al. 2013; Gardner et al. 2001; Mindlin et al. 2003; Perl et al. 2011). Taking this approach has resulted in models that reduce the description of singing behavior to a low number of parameters. Since the syrinx is a nonlinear phonating device that transduces these parameters into sound, even simple parameter time traces can capture rich information about the acoustics. It is for this reason that these parameter time traces might be thought of as a coordinate system to interpret the neural coding involved in birdsong production. The demonstration here of the validity of an automated procedure to model song in a low-dimensional framework furthers this program, allowing us to rule out subjective evaluations of parameter features that may end up increasing the dimensionality of the reconstruction. Typically, sound analysis is performed in terms of acoustic features such as fundamental frequency, amplitude, and some measure of the entropy (Sober et al. 2008; Tchernichovski et al. 2000). A description in terms of motor coordinates complements the analysis and has some elements in common: the activity of some syringeal muscles is transduced into fundamental frequency, and in some species such as the zebra finch for which the fundamental frequency is reasonably lower than resonances of the passive vocal tract the sound amplitude is a good proxy for the air sac pressure. On the other hand, as has been extensively reported, the physical mechanisms involved (and in particular, the bifurcations involved in the onset of the labial oscillations) link several acoustic features together, as fundamental frequency and spectral content. This results in simple instructions being transduced into sounds with a high degree of internal logic (for example, a precise spectral content for each sound of a given fundamental frequency). Similarly, the shape of the air sac pressure pattern determines the nature of the attack of the sound (by conditioning the sound envelope), and therefore its timbre (a most elusive acoustic feature, difficult to express in terms of simple acoustic measures).

The automatic reconstruction was bounded to reduce the number of parameters to be fitted. We used the normal form of a SNILC bifurcation as the nonlinear dynamical system emulating the labial dynamics in the syrinx. The reconstruction of the fundamental frequency gave rise, through a lookup table, to time-dependent parameters of the normal form. The integration of these equations with the reconstructed parameters generated sounds with the desired spectral content. For filtering the sounds, we used a closed-open tube connected to a Helmholtz resonator of fixed parameters. Noise was added to the reconstructed parameter (Amador et al. 2013). This process is fully automatic and requires as input only the recorded sound signal for each bird. Note that simplicity in the reconstruction of gestures from the model does not imply an additional reduction of the amount of information needed to synthesize realistic sounds: the air sac pressure was not computed by fitting parameters in the dynamical model of the labia, but a proxy was computed from the sound.

The synthetic sounds generated in this way were capable of eliciting responses in HVC neurons selective to BOS. This is remarkable, given the dramatic fall of response reported in previous works as soon as the parameters of a successful physical model were slightly changed (Amador et al. 2013). In terms of both mean activity as well as variance, neurons responded significantly to the algorithmically generated sounds that we developed. We chose variance as a relevant response parameter to investigate because it reflects a coding strategy that stands out from random activity. Correlating the response to BOS and automatically reconstructed sounds is more restrictive: it would likely require that we reproduce all the relevant features of BOS and properly weight them (see Meliza et al. 2010). But even so, we have shown that responses to automatically reconstructed songs (AUTO) have a significantly higher correlation coefficient to BOS than either REV or CON, meaning that AUTO- and BOS-elicited responses share some features of their temporal time course pattern of excitation and inhibition. As an example of strong responses to synthetic sound, highly structured and yet poorly correlated with the response to BOS, we report one case where in certain regions of song the response to our synthetic sound was larger than that to BOS.

It is never clear, a priori, whether the physics paradigm of dramatically reducing the complexity of a problem will give rise to a pertinent description of it. When dealing with biological systems, this issue becomes particularly acute, as the

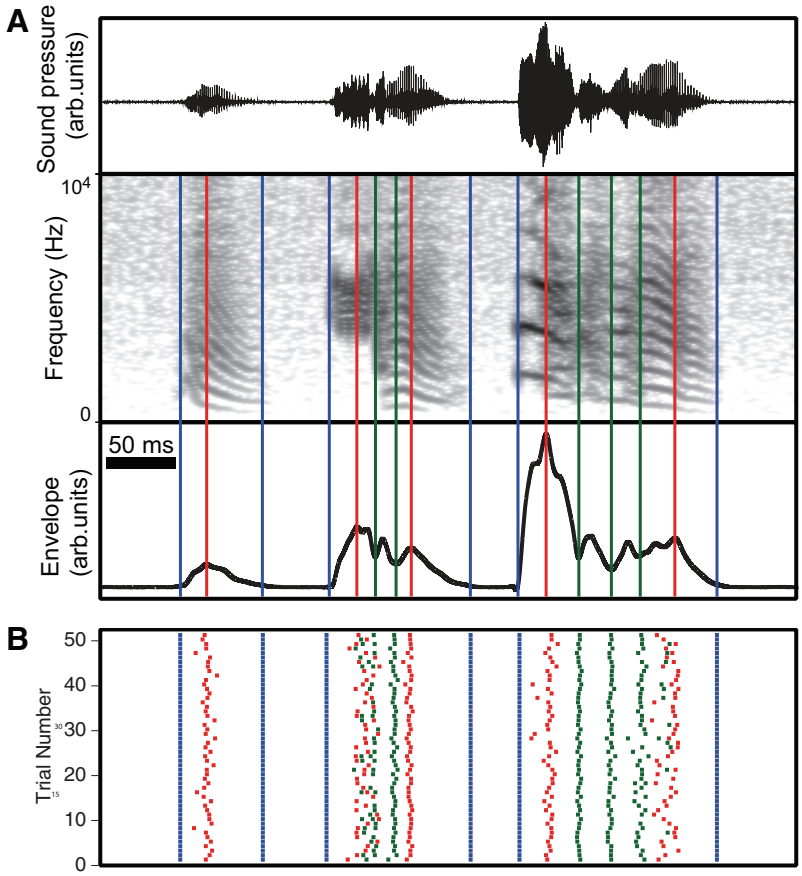

Fig. 6. Reliability of GTE automatic extraction. The GTEs from 51 song renditions for 1 bird were calculated. *A*: sound trace, sonogram, and envelope of 1 example song. *B*: automatic reconstruction of GTEs for different song renditions. GTE times were linearly scaled to a fixed syllable duration. This takes advantage of the highly reliable estimates of syllable onsets and offsets as a reference frame to study the robustness of GTE timing within a syllable. The dispersion for all the GTEs (excluding onsets and offsets) across all the renditions was 3.4 ± 2.28 ms.

systems generally express an enormous number of degrees of freedom. In this work, we analyzed birdsong under this paradigm and found that an automatic reconstruction of parameters of a low-dimensional model can give rise to songs that are biologically relevant in the sense that they can elicit responses from neurons highly selective to BOS. Whether we use the reconstructed parameters as motor coordinates to interpret neural coding or use them to drive bio-prosthetic devices, simplified models usually anticipate deeper understanding.

In machine learning and statistics, dimensionality reduction has been used as a process of reducing the number of variables under consideration. The data transformation can be linear (as in the widely used method of principal component analysis) or nonlinear, but in general when this method is applied to biomechanical problems the variables extracted are hard to relate to the physiology. The method presented here constitutes a nonlinear dimensionality reduction that is derived from physical principles studying the biomechanics of the peripheral system. In this way, the variables can be more easily related with physiological parameters the subject can control. Moreover, since we are not fitting the parameters of a physical system but those of a normal form, the number of parameters is reduced to a minimum expression. This means that even if many physiological instructions participate synergistically in the control of a given feature, we reduce the parameter search to find a minimal number of instructions from which the feature can be controlled. In our example, both air sac pressure and syringeal tension are involved in the control of the fundamental frequency, yet in this work we show that it is possible to generate pertinent synthetic songs by reconstructing the parameters of a normal form with only one fitted parameter [$\beta(t)$].

Given the biomechanical similarities between the sound sources of vocalizations in songbirds and mammals (Amador and Margoliash 2013; Riede and Goller 2010), the method presented here can be used to generate synthetic vocalizations in a wide variety of species. Recently, our model for the vocal source was used to reproduce macaque monkeys' calls, correlating motor cortical activity with estimated time-varying parameters [$\alpha(t)$, $\beta(t)$] (Fukushima et al. 2014), showing that this framework might be useful for decoding motor cortical activity with a nonlinear dynamical model during vocal production. Reducing the high dimensionality of produced sounds could be advantageous for applications of brain-machine interface for speech production.

More generally, the methodology of analyzing the biomechanics of the peripheral system in terms of its physical properties and generating nonlinear dynamical models to reproduce behavior can also be extended to other systems, paving the road from the CNS to the control of the peripheral devices in charge of executing the actions implied in behavior.

ACKNOWLEDGMENTS

Discussions with M. Long and V. Katlowitz concerning automatic GTE extraction are acknowledged.

GRANTS

This work describes research partially funded by CONICET, ANCyT, UBA, and National Institute on Deafness and Other Communication Disorders Grant RO1-DC-012859.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

Author contributions: S.B. and A.A. performed experiments; S.B., Y.S.P., A.A., and G.B.M. analyzed data; S.B., Y.S.P., A.A., D.M., and G.B.M. interpreted results of experiments; S.B. and A.A. prepared figures; S.B., A.A., D.M., and G.B.M. drafted manuscript; S.B., A.A., D.M., and G.B.M. edited and revised manuscript; S.B., Y.S.P., A.A., D.M., and G.B.M. approved final version of manuscript; A.A. and G.B.M. conception and design of research.

REFERENCES

Akutagawa E, Konishi M. New brain pathways found in the vocal control system of a songbird. *J Comp Neurol* 518: 3086–3100, 2010.

Amador A, Margoliash D. A mechanism for frequency modulation in songbirds shared with humans. *J Neurosci* 33: 11136–11144, 2013.

Amador A, Mindlin GB. Beyond harmonic sounds in a simple model for birdsong production. *Chaos* 18: 043123, 2008.

Amador A, Perl YS, Mindlin GB, Margoliash D. Elemental gesture dynamics are encoded by song premotor cortical neurons. *Nature* 495: 59–64, 2013.

Bauer EE, Coleman MJ, Roberts TF, Roy A, Prather JF, Mooney R. A synaptic basis for auditory-vocal integration in the songbird. *J Neurosci* 28: 1509–1522, 2008.

Cardin JA, Schmidt MF. Song system auditory responses are stable and highly tuned during sedation, rapidly modulated and unselective during wakefulness, and suppressed by arousal. *J Neurophysiol* 90: 2884–2899, 2003.

Cardin JA, Schmidt MF. Auditory responses in multiple sensorimotor song system nuclei are co-modulated by behavioral state. *J Neurophysiol* 91: 2148–2163, 2004.

Coleman MJ, Roy A, Wild JM, Mooney R. Thalamic gating of auditory responses in telencephalic song control nuclei. *J Neurosci* 27: 10024–10036, 2007.

Dave AS, Margoliash D. Song replay during sleep and computational rules for sensorimotor vocal learning. *Science* 290: 812–816, 2000.

Dave AS, Yu AC, Margoliash D. Behavioral state modulation of auditory activity in a vocal motor system. *Science* 282: 2250–2254, 1998.

De Cheveigne A, Kawahara H. YIN, a fundamental frequency estimator for speech and music. *J Acoust Soc Am* 111: 1917–1930, 2002.

Doupe AJ, Konishi M. Song-selective auditory circuits in the vocal control system of the zebra finch. *Proc Natl Acad Sci USA* 88: 11339–11343, 1991.

Fee MS, Kozhevnikov AA, Hahnloser RH. Neural mechanisms of vocal sequence generation in the songbird. *Ann NY Acad Sci* 1016: 153–170, 2004.

Fletcher NH, Riede T, Suthers RA. Model for vocalization by a bird with distensible vocal cavity and open beak. *J Acoust Soc Am* 119: 1005–1011, 2006.

Fukushima M, Saunders RC, Fujii N, Averbeck BB, Mishkin M. Modeling vocalization with ECoG cortical activity recorded during vocal production in the macaque monkey. *Conf Proc IEEE Eng Med Biol Soc* 2014: 6794–6797, 2014.

Gardner T, Cecchi G, Magnasco M, Laje R, Mindlin GB. Simple motor gestures for birdsongs. *Phys Rev Lett* 8720: 208101, 2001.

Goller F, Suthers RA. Role of syringeal muscles in controlling the phonology of bird song. *J Neurophysiol* 76: 287–300, 1996a.

Goller F, Suthers RA. Role of syringeal muscles in gating airflow and sound production in singing brown thrashers. *J Neurophysiol* 75: 867–876, 1996b.

Guckenheimer J, Holmes P. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. New York: Springer, 1997.

Hahnloser RH, Kozhevnikov AA, Fee MS. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419: 65–70, 2002.

Keller GB, Hahnloser RH. Neural processing of auditory feedback during vocal practice in a songbird. *Nature* 457: 187–190, 2009.

Laje R, Gardner TJ, Mindlin GB. Neuromuscular control of vocalizations in birdsong: a model. *Phys Rev E Stat Nonlin Soft Matter Phys* 65: 051921, 2002.

Mandelblat-Cerf Y, Las L, Denisenko N, Fee MS. A role for descending auditory cortical projections in songbird vocal learning. *eLife* 3: e02152, 2014.

Margoliash D. Acoustic parameters underlying the responses of song-specific neurons in the white-crowned sparrow. *J Neurosci* 3: 1039–1057, 1983.

Margoliash D. Preference for autogenous song by auditory neurons in a song system nucleus of the white-crowned sparrow. *J Neurosci* 6: 1643–1661, 1986.

Margoliash D, Fortune ES, Sutter ML, Yu AC, Wrenhardin BD, Dave A. Distributed representation in the song system of oscines: evolutionary implications and functional consequences. *Brain Behav Evol* 44: 247–264, 1994.

Margoliash D, Konishi M. Auditory representation of autogenous song in the song system of white-crowned sparrows. *Proc Natl Acad Sci USA* 82: 5997–6000, 1985.

Meliza CD, Chi Z, Margoliash D. Representations of conspecific song by starling secondary forebrain auditory neurons: toward a hierarchical framework. *J Neurophysiol* 103: 1195–1208, 2010.

Mindlin GB, Gardner TJ, Goller F, Suthers R. Experimental support for a model of birdsong production. *Phys Rev E* 68: 41908, 2003.

Mindlin GB, Laje R. *The Physics of Birdsong*. Berlin: Springer, 2005.

Nishikawa K, Biewener AA, Aerts P, Ahn AN, Chiel HJ, Daley MA, Daniel TL, Full RJ, Hale ME, Hedrick TL. Neuromechanics: an integrative approach for understanding motor control. *Integr Comp Biol* 47: 16–54, 2007.

Perl YS, Arneodo EM, Amador A, Goller F, Mindlin GB. Reconstruction of physiological instructions from Zebra finch song. *Phys Rev E Stat Nonlin Soft Matter Phys* 84: 051909, 2011.

Prather JF, Peters S, Nowicki S, Mooney R. Precise auditory-vocal mirroring in neurons for learned vocal communication. *Nature* 451: 305–310, 2008.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes: The Art of Scientific Computing*. Cambridge, UK: Cambridge Univ. Press, 2007.

Quiroga RQ, Nadasdy Z, Ben-Shaul Y. Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput* 16: 1661–1687, 2004.

Rauske PL, Shea SD, Margoliash D. State and neuronal class-dependent reconfiguration in the avian song system. *J Neurophysiol* 89: 1688–1701, 2003.

Reinke H, Wild J. Identification and connections of inspiratory premotor neurons in songbirds and budgerigar. *J Comp Neurol* 391: 147–163, 1998.

Riede T, Goller F. Peripheral mechanisms for vocal production in birds—differences and similarities to human speech and singing. *Brain Lang* 115: 69–80, 2010.

Riede T, Suthers RA, Fletcher NH, Blevins WE. Songbirds tune their vocal tract to the fundamental frequency of their song. *Proc Natl Acad Sci USA* 103: 5543–5548, 2006.

Schmidt MF, Ashmore RC, Vu ET. Bilateral control and interhemispheric coordination in the avian song motor system. *Ann NY Acad Sci* 1016: 171–186, 2004.

Schmidt MF, Konishi M. Gating of auditory responses in the vocal control system of awake songbirds. *Nat Neurosci* 1: 513–518, 1998.

Sitt JD, Amador A, Goller F, Mindlin GB. Dynamical origin of spectrally rich vocalizations in birdsong. *Phys Rev E* 78: 011905, 2008.

Sitt JD, Arneodo EM, Goller F, Mindlin GB. Physiologically driven avian vocal synthesizer. *Phys Rev E* 81: 31927, 2010.

Sober SJ, Wohlgemuth MJ, Brainard MS. Central contributions to acoustic variation in birdsong. *J Neurosci* 28: 10370–10379, 2008.

Striedter G, Vu E. Bilateral feedback projections to the forebrain in the premotor network for singing in zebra finches. *J Neurobiol* 34: 27–40, 1998.

Strogatz SH. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Cambridge, MA: Perseus, 1994.

Suthers RA, Goller F, Pytte C. The neuromuscular control of birdsong. *Philos Trans R Soc Lond B Biol Sci* 354: 927–939, 1999.

Tchernichovski O, Nottebohm F, Ho CE, Pesaran B, Mitra PP. A procedure for an automated measurement of song similarity. *Anim Behav* 59: 1167–1176, 2000.

Theunissen FE, Doupe AJ. Temporal and spectral sensitivity of complex auditory neurons in the nucleus HVc of male zebra finches. *J Neurosci* 18: 3786–3802, 1998.