J Neurophysiol 89: 2697–2706, 2003;
10.1152/jn.00801.2002.

Model of Song Selectivity and Sequence Generation in Area HVc of the Songbird

Patrick J. Drew and L. F. Abbott

Volen Center for Complex Systems and Department of Biology, Brandeis University, Waltham, Massachusetts 02454-9110

Submitted 12 September 2002; accepted in final form 13 January 2003

Drew, Patrick J. and L. F. Abbott. Model of song selectivity and sequence generation in area HVc of the songbird. *J Neurophysiol* 89: 2697–2706, 2003;; 10.1152/jn.00801.2002. In songbirds, nucleus HVc plays a key role in the generation of the syllable sequences that make up a song. Auditory responses of neurons in HVc are selective for single syllables and for combinations of syllables occurring in temporal sequences corresponding to those in the bird's own song. We present a model of HVc that produces syllable- and temporal-combination-selective responses on the basis of input from recorded bird songs filtered through spectral temporal receptive fields similar to those measured in field L, a primary auditory area. Normalization of the field L outputs, similar to that proposed in models of visual processing, plays an important role in the generation of syllable-selective responses in the model. For temporal-combination-selective responses, *N*-methyl-D-aspartate (NMDA) conductances provide a memory that allows inhibitory neurons to gate responses to a final syllable in a sequence on the basis of responses to earlier syllables. When the same network that produces temporal-combination-selective responses is excited by a nonspecific timing signal, it generates a similar pattern of output as it does in response to auditory song input. Thus the same model network can perform both sensory and motor functions.

INTRODUCTION

Neural circuits can generate and respond to temporal sequences that last much longer than the integration time constants of single neurons. Selectivity for temporal sequences requires a memory mechanism for storing information over the duration of a sequence, as well as a mechanism that allows this stored information to gate responses. Modeling studies of temporal-sequence selectivity can be used to explore possible mechanisms by testing their viability and suggesting measurable experimental consequences through which they can be confirmed or invalidated (Buonomano and Karmarkar 2002; Troyer and Doupe 2000a,b). The selectivity of neurons in the bird song system, a set of interconnected nuclei devoted to song learning, production, and recognition, provides an excellent system on which to base such studies (Doupe and Konishi 1991; Doupe and Kuhl 1999; Konishi 1985; Margoliash 1997). Many neurons in these nuclei respond selectively to complex temporal auditory sequences within the bird's own song. Song-selectivity includes responses to specific individual syllables within a song (Lewicki 1996; Margoliash 1983; Margoliash and Fortune 1994) and to combinations of syllables presented in a specific temporal order (Lewicki 1996; Lewicki and Arthur 1996; Lewicki and Konishi 1995; Margoliash and Fortune 1994). Here, we construct a model of syllable- and temporal-combination-selective neurons and show that the resulting circuit can generate as well as respond selectively to specific temporal sequences.

Song-selective responses, as well as other auditory responses, occur in many of the nuclei associated with the song system in birds (Doupe and Kuhl 1999; Konishi 1985; Margoliash 1997). Field L, which receives direct input from the thalamic auditory nucleus ovoidalis, is roughly the analog in the bird of mammalian primary auditory cortex. Neurons in field L have been measured and characterized in terms of spectral temporal receptive fields (STRFs) (Sen et al. 2001; Theunissen et al. 2000), which provide a concise way of simulating their responses. In our model, responses generated in this way provide the feedforward input to second-stage neurons that are selective for either syllables within recorded birdsongs or temporal combinations of these syllables by virtue of their network interactions. We think of these song-selective neurons as being located in area HVc (high vocal center), a region where neural responses are strongly song selective (Lewicki and Arthur 1996; Margoliash 1983; Margoliash and Fortune 1994). Thus our model consists of two stages: an input stage based on frequency, but not song-selective field L responses, and an output stage generating responses similar to those of song-selective units in HVc.

Not surprisingly, the situation in songbirds is considerably more complex. First, although we assume a direct projection from field L to HVc, field L neurons may project to a neighboring structure, the HVc shelf, rather than directly to HVc (Fortune and Margoliash 1995; Kelley and Nottebohm 1979; Margoliash 1997; Vates et al. 1996). Second, HVc receives input from a number of other areas, including the medial magnocellular nucleus of the archistriatim (mMAN), the thalamic nucleus uvaeformis (Uva), the nucleus interfacialis (NIf), and the hyperstriatum ventrale (cHV) (Nottebohm et al. 1982; Vates et al. 1996, 1997). Song-selective activity in mMAN appears to follow that in HVc (Vates et al. 1997), and Uva appears to be associated with motor rather than sensory processing of song (Margoliash 1997; Williams and Vicario 1993). However, NIf, in particular, is a potential source of sensory input to HVc (Coleman and Mooney 2002). Further-

Address for reprint requests: P. Drew, Volen Center for Complex Systems and Dept. of Biology, Brandeis Univ., Waltham, MA 02454-9110 (E-mail: drew@brandeis.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

more, responses in both NIf and cHV can be song-selective, as are those of some neurons in field L, but to a lesser extent than those in HVc (Janata and Margoliash 1999; Lewicki and Arthur 1996; Sen at al. 2001). In light of the complicated interconnectivity of the song system, our model should be viewed as a set of general frequency-selective neurons providing input to neurons that generate song-selective responses through network interactions. Although we refer to these as field L and HVc stages, the exact location of the input and song-selective neurons is somewhat ambiguous. Furthermore, we are using a two-stage network to approximate a system in which song-selectivity arises progressively over a number of different areas.

There are at least two classes of excitatory neurons in HVc (Dutar and Perkel 1998; Mooney 2000). Neurons that project to area X (X projecting) are hyperpolarized during song playback but fire in response to specific portions of the song. Neurons that project to the robust nucleus of the archistriatum (RA projecting) are generally depolarized during song playback and also fire at specific points within the song. Our model applies to the RA-projecting neurons in HVc. HVc is a motor structure that, in addition to its sensory responses, plays an important role in song production (Hahnloser et al. 2002; Margoliash 1997; Vu et al. 1994). The network we construct to reproduce song-selective sensory responses in HVc can also generate similar sequences of activity in response to a general timing signal, which might represent input from Uva. Thus like HVc, our model can act as both a sensory and a motor network.

METHODS

The model consists of two stages; a field L stage that is modeled using linear filters and a normalization operation, and an HVc stage where neurons are modeled as integrate-and-fire units. For the syllable-selective examples, we used a single integrate-and-fire neuron, and for temporal-combination selectivity we used a network of integrate-and-fire units. All simulations were implemented using MatLab.

Filtering, normalization, and weights

Recorded songs provided input to the model in the form of a spectrogram, $s(t, f)$. This was passed through a set of linear filters, representing the action of neurons early in the auditory pathway. In these linear filters, an STRF function, $F_i(\tau, f)$, determines how the magnitude of the spectrogram at frequency f, a time period τ in the past, affects the output of unit i. To implement this, STRF outputs, $x_i(t)$, were generated by integrating the product of the song spectrogram times the STRF filter, $F_i(\tau, f)$

$$x_i(t) = \int_0^\infty d\tau \int_0^\infty df s(t-\tau, f) F_i(\tau, f) \quad (1)$$

We modeled the STRF filter on the data of Theunissen et al. (2000) and Sen et al. (2001) as having a Gaussian frequency profile about a preferred frequency with width f_{width} and a time profile described by a gamma function displaced by a latency τ_0. Thus the STRF was modeled as

$$F_i(\tau, f) = \left[\alpha^5(\tau-\tau_0)^5 \exp(-\alpha(\tau-\tau_0)) \exp\left(-\frac{(f-f_i)^2}{2f_{\text{width}}^2}\right)\right]_+ \quad (2)$$

where $\alpha = 3/\text{ms}$, $\tau_0 = 0$ or 8 ms, $f_{\text{width}} = 100$ Hz, f_i is the preferred frequency of the STRF, and $[z]_+ = z$ for $z > 0$ and is zero otherwise. The values of f_i were evenly spaced every 125 Hz between 0 and 8,000 Hz. The STRFs were divided into two banks, one with $\tau_0 = 0$ ms and one with $\tau_0 = 8$ ms. This helped in the detection of frequency sweeps, which are important components of many syllables. Even with the longer delay, the STRFs carried no information from further back than 70 ms, which is not long enough to overlap with more than one syllable. For convenience, we assign the label i that specifies particular STRFs in order of their preferred frequencies.

To represent the effects of saturation and suppression by surrounding units, STRF outputs were normalized by making the transformation

$$x_i(t) \to \frac{x_i(t)}{\epsilon + \sqrt{\sum_j x_j^2(t)}} \quad (3)$$

Note that as $x_i \to \infty$, this expression approaches a finite limit, and that, because of the sum over j, other STRF filters ($j \neq i$) with large outputs will suppress the response of unit i. Here $\epsilon = 0.05$ is a parameter that controls where the response begins to saturate (see Fig. 2). The firing rate of field L unit i is taken to be proportional to a rectified version (to eliminate negative firing rates) of the normalized output of the corresponding field L filter

$$r_i(t) = \beta[x_i(t)]_+ \quad (4)$$

where β is a constant (see *Model neuron* for its value).

Model neuron

All neurons in the HVc stage of the model were modeled as leaky integrate-and-fire units for which the membrane potential V is described by the equation

$$\tau_m \frac{dV}{dt} = V_{\text{rest}} - V + g_{\text{AHP}}(t)(E_{\text{AHP}} - V) + g_{\text{ex}}(t)(E_{\text{ex}} - V) + g_{\text{in}}(t)(E_{\text{in}} - V). \quad (5)$$

The conductances g_{AHP}, g_{ex}, and g_{in} are divided by the leak conductance, making them dimensionless. We set the effective membrane time constant $\tau_m = 20$ ms for excitatory neurons and $\tau_m = 10$ ms for inhibitory neurons. The resting potential is $V_{\text{rest}} = -70$ mV, and the synaptic reversal potentials are $E_{\text{ex}} = 0$ mV and $E_{\text{in}} = -70$ mV for excitation and inhibition, respectively. In addition, $E_{\text{AHP}} = -70$ mV. Action potentials are generated whenever V reaches a threshold potential of -50 mV, after which the membrane potential is reset to -70 mV.

For the syllable-selective units shown in Figs. 3 and 4, the excitatory conductance g_{ex} is the sum of a syllable-selective term, $g_{\text{syllable}}(t)$, (*Eq. 6*) and a nonselective background input. The inhibitory conductance g_{in} consists solely of a nonselective background. The background inputs are generated by Poisson spike trains (representing the summed input from many afferents) with rates of 1,500 Hz for excitation and 1,000 Hz for inhibition. Each time a spike arrives, the corresponding synaptic conductance (g_{ex} or g_{in}) is increased by 0.1. After that, this contribution decays exponentially with a time constant of 2 ms for excitation and 10 ms for inhibition.

The syllable-selective excitatory conductance, g_{syllable}, is computed by summing the firing rates of the N presynaptic field L units multiplied by appropriate synaptic weights, w_i

$$g_{\text{syllable}}(t) = \gamma \sum_{i=1}^{N} w_i r_i(t) \quad (6)$$

where $\beta\gamma$ (results of the model only depend on a multiplicative combination of these 2 parameters) is between 0.5 and 2, depending on the syllable. The less variability in the peak frequencies within the syllable, the smaller γ needed to be. Syllable-selectivity was conferred by choosing the synaptic weights on the basis of the field L responses at a particular time t_{syllable} during the syllable being selected for. As discussed in the text, weights were chosen to select for local maxima

in the field L responses at a particular time $t_{syllable}$ during the syllable using the following rule: if $r_i(t_{syllable}) > r_{i-1}(t_{syllable})$ and $r_i(t_{syllable}) > r_{i+1}(t_{syllable})$

$$w_{i-1} = r_{i-1}(t_{syllable}), \quad w_i = r_i(t_{syllable}), \quad \text{and} \quad w_{i+1} = r_{i+1}(t_{syllable}) \tag{7}$$

with the understanding that the STRFs are labeled in order of their preferred frequencies. Otherwise $w_i = 0$. To provide a uniform scale, the weights are also normalized

$$w_i \rightarrow \frac{w_i}{\sqrt{\sum_j w_j^2}} \tag{8}$$

Choosing the right time point in the syllable is important for accurate syllable recognition. Changing $t_{syllable}$ a few milliseconds either way can sometimes impair recognition, because the acoustic characteristics of a syllable can change rapidly.

The after-hyperpolarizing potential (AHP) conductance, which is included in all the excitatory model neurons, is incremented by 0.8 every time the neuron fires an action potential, has an absolute maximum of twice the resting membrane conductance, and otherwise decays exponentially with a time constant of 100 ms. No AHP was used for Fig. 4, *C* and *D,* to eliminate confounding effects of repetitive stimulation. This did not affect the volume dependence being illustrated in the figure. The AHP has little effect on the syllable and temporal-combination selectivity of the model, but it plays a key role in the generation of temporal sequences.

Network model

The network model (Fig. 5) used for Figs. 6 and 7, consists of 120 neurons, 30 of each type: A-selective excitatory neurons, which receive field L input tuned to syllable A; A-selective inhibitory neurons, which receive tonic excitation and are also driven by the A-selective excitatory neurons; AB-selective excitatory neurons, which receive field L input tuned to syllable B; and B-suppressing inhibitory neurons, which receive tonic excitation, are suppressed by the A-selective inhibitory neurons and, in turn, suppress the AB-selective excitatory neurons. All the neurons in the network model receive the nonselective background excitatory and inhibitory inputs described above. The A-selective and AB-selective excitatory units receive syllable-selective excitatory input, as described by *Eq. 6*. In place of this syllable-selective input, the A-selective inhibitory neurons receive a constant excitatory conductance of 0.4 during song playback, and the B-suppressing inhibitory neurons receive a constant excitatory conductance of 0.5. Neurons in the network model are coupled to each other through AMPA, GABA, and *N*-methyl-D-aspartate (NMDA) synapses. For AMPA and GABA synapses, g_{ex} and g_{in} are incremented by the amounts listed in the table below (for the different synaptic connections of the model) when a presynaptic action potential arrives. For recurrent synapses in the network model, saturation of individual synapses at high-input rates was also implemented to prevent runaway excitation. These conductance changes then decay exponentially with the same time constants given above, 2 ms for excitation and 10 ms for inhibition.

NMDA conductances were added to g_{ex} in the following way (Wang 1999). When a presynaptic action potential activates an NMDA synapse in the model, a variable s_1 is incremented by 1, $s_1 \rightarrow s_1 + 1$. Otherwise, s_1 decays exponentially with a time constant of 2 ms. From s_1, a second variable s_2 is computed from the equation

$$\tau_2 \frac{ds_2}{dt} = \tau_2 s_1(1 - s_2) - s_2 \tag{9}$$

with $\tau_2 = 120$ ms. This implements both the finite rise and decay times of the NMDA conductance and the saturation of the conductance at high input rates. The NMDA contribution to g_{ex} is the appropriate number given in the table below times $s_2/[1 + \exp(-0.062V)/3.57)]$. The denominator, with the membrane potential V taken to be in millivolts, describes the well-known voltage dependence of the NMDA conductance.

The strengths for all the synapses of the network model are shown in Table 1 (in Table 1, A-selective excitatory neurons are listed as A, A-selective inhibitory neurons as Ai, AB-selective excitatory neurons as AB, and B-suppressing inhibitory neurons as Bi). The columns of Table 1 correspond to the presynaptic neuron and the rows to the postsynaptic neuron. The numbers below are for the network model shown in Figs. 6 and 7. There are no autapses. Small changes in these conductances were required for the network to respond to other syllable sequences.

TABLE 1.

	A AMPA	Ai GABA	AB AMPA	Bi GABA	A NMDA	AB NMDA
A	0.125	—	—	—	0.05	—
Ai	—	—	—	—	0.175	—
AB	—	—	0.125	0.15	—	0.05
Bi	—	0.065	—	—	—	—

To generate the motor pattern in Fig. 7, the A- and AB-selective neurons are injected simultaneously with excitatory conductances of approximately 0.55 and 0.8, respectively, for 10-ms pulses separated by 75–100 ms. The A-selective and B-suppressing inhibitory neurons receive constant background excitatory conductances of 0.4 and 0.65. For both sequence recognition and generation, the strengths of the background conductances did not require precise tuning. A relatively wide range of parameters produced qualitatively similar results, although, for sequence generation, it helped to keep the background conductance to the B-suppressing inhibitory neurons high to prevent the AB-selective neurons from responding to the first timing pulse. For the ABC-generating network shown in Fig. 7*C*, two additional inhibitory populations (analogous to and having the same parameters as the A-selective inhibitory and B-suppressing inhibitory neurons) and an ABC-selective population of neurons (analogous to the AB-selective neurons) were added to the network model. In addition, the time constant of the AHP was increased to 200 ms for all the excitatory neurons in the network simulations shown in Fig. 7, *B* and *C*.

The network model was robust to approximately 10% variations in its synaptic conductances. As parameters were varied away from optimal, the model degraded gracefully without uncontrollable excess levels of activity. Generally, generation of the correct sequence degraded first when parameters were adjusted, followed by the ability of the network to respond selectively to the sequence. The parameters controlling syllable recognition could be varied by even larger amounts, depending on the syllable being detected, before a syllable-selective neuron stopped responding or responded nonselectively.

RESULTS

As mentioned in the introduction, we are interested in modeling two kinds of song-selective responses: syllable selective and temporal-combination selective. We begin by constructing syllable-selective units, which form the basis for the temporal-combination selectivity discussed later. In both cases, the input to the model consists of spectrograms from recorded songs (kindly supplied by M. Kao, K. Sen, and A. Doupe). These are processed through an array of STRFs modeled after those in field L and then normalized to reproduce saturation and surround-suppression effects within the field L stage of the model. Syllable- and temporal-combination selectivity arises in the HVc network of the model through a combination of feedforward and recurrent circuitry. The field L and HVc stages are modeled in quite different ways. The field L stage is modeled

descriptively as a set of firing rates generated by STRFs without a specific biophysical representation. This is because we are not exploring in this study how field L responses arise. The HVc stage, on the other hand, is modeled as a network of spiking model neurons (integrate-and-fire neurons) receiving and interacting through realistic synaptic conductances. This more biophysical representation allows us to explore specific cellular, synaptic, and circuit mechanisms that can produce syllable- and temporal-combination selectivity.

Field L stage

The first stage of our model is an array of STRFs based on the simplest ones found in field L (Sen et al. 2001; Theunissen et al. 2000). The STRFs act as filters on the spectrograms of recorded songs, producing an output that provides a measure of the amplitude of the spectrogram over a particular time and frequency range. Specifically, the output of a given STRF at a given time is proportional to the integral of the spectrogram amplitude times a Gaussian-shaped frequency profile approximately 200 Hz wide that extends backward $\leq$70 ms prior to that time. The center of the Gaussian frequency profile defines the preferred frequency of the STRF. The preferred frequencies of different STRFs are evenly spaced every 125 Hz between 0 and 8,000 Hz, giving full overlapping coverage of all frequencies in that range. Each STRF is convolved with the song spectrogram (Fig. 1*A*), producing a set of filter outputs (Fig. 1*B*). The STRF-generated outputs are ordered with respect to their preferred frequencies in Fig. 1*B*, which makes the output resemble an approximate duplicate of the song spectrogram seen in Fig. 1*A*.

It is difficult to generate song-specific responses directly from the outputs of the field L STRFs. This is because loud syllables generate larger responses than soft syllables, as seen in Fig. 1*B*, and these large responses can overwhelm the selectivity of downstream units for softer syllables. For this reason, we assume that the field L responses are normalized in a manner similar to what has been suggested for responses in areas of the mammalian visual system (Heeger 1992; Simoncelli and Heeger 1998). The normalization operation reproduces saturation effects at high stimulus intensities and also allows high activity in some field L units to suppress all of them. Specifically, if we think of the full array of STRF outputs as being represented by a vector, the normalization procedure consists of dividing this vector, at each point in time, by a factor that is a linear function of its length (see METHODS). The responses from the array of STRFs after normalization are shown in Fig. 1*C*. It is clear from this figure that the different syllables now produce responses of more equivalent magnitudes than in Fig. 1*B*.

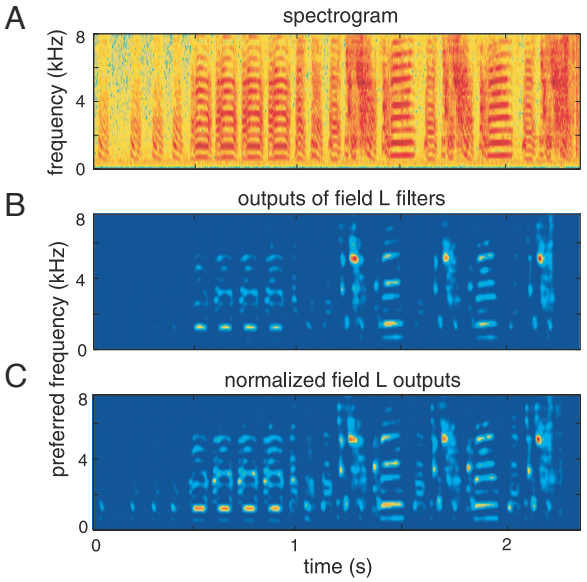

FIG. 1. Steps leading to the output of the field L stage of the model. *A–C*: horizontal axis represents time and the color represents the amplitude, with blue the lowest and red the highest. *A*: vertical axis is frequency. *B* and *C*: vertical axis is preferred frequency of the corresponding spectral temporal receptive fields (STRFs). *A*: song spectrogram. *B*: outputs of the field L STRF filters applied to the spectrogram. *C*: STRF outputs after normalization.

The effect of normalization on field L responses is quantified in Fig. 2. Here the magnitude of the full field L output (the length of the field L output vector) is plotted as a function of the magnitude of its input before normalization (the length of the output vector of the field L filters). The curve in Fig. 2 illustrates the effect of the normalization operation, which causes the initial linear rise to change to a slower increase for higher sound intensities. Figure 2 also shows the range over which typical song syllables drive the field L units and also the range for inter-syllable periods. The scale of the normalization effect has been chosen so that responses to syllables are near the saturation region, whereas responses between syllables are well below saturation.

In summary, the field L stage of our model consists of an array of STRFs that act as linear filters on song spectrograms to produce a set of outputs selective for different preferred frequencies. At each time, this set of outputs is normalized, representing saturation and "surround-suppression" effects. The firing rates that represent the output of the field L stage of the model are then proportional to half-wave rectified versions of these normalized STRF outputs.

Syllable selectivity

As mentioned previously, the syllable- and temporal-combination-selective units in our model are integrate-and-fire neurons driven by the field L outputs. Individual syllable-selective units receive excitatory synaptic conductances proportional to a weighted sum of the firing rates of the field L units. Syllable selectivity arises from an appropriate choice of the weights in this sum, with each weight corresponding to a unitary synaptic conductance.

We tried a number of schemes for determining the optimal weights for generating syllable-selective responses. The scheme that worked best was to set most of the weights to zero and to reserve the small number of nonzero synaptic weights for the peak STRF responses. In other words, we use a sparse representation of the song syllables to drive syllable-selective neurons, which is somewhat analogous to edge detection in visual object recognition. Specifically, a time was chosen within the middle of the syllable to be detected, and peak frequencies were identified by finding field L units that fired more rapidly than their neighbors with the next higher or next lower preferred frequencies. Weights for the three units around each peak were then set proportional to their firing rates, while all other weights were set to zero (see METHODS). This proce-

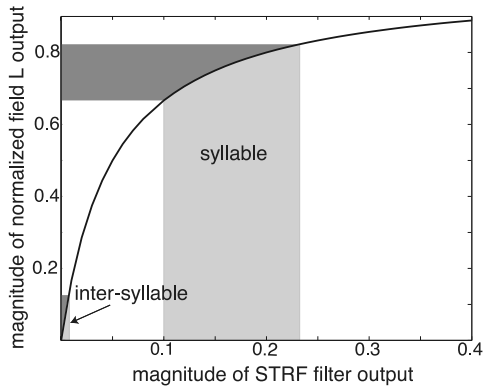

FIG. 2. Effect of normalization. Length of the vector of field L outputs after normalization (magnitude of normalized field L output) plotted as a function of its value before normalization (magnitude of STRF filter output). Typical ranges for syllable stimuli and for intervals between syllables are shown with shading. Inter-syllable inputs fall on the lower portion of the linear part of the curve, whereas syllable inputs fall near the saturating region.

dure generates weight values that are similar to those obtained by setting weights proportional to the rectified difference between the firing rate of the presynaptic field L unit during the selected syllable and its mean firing rate. With this approach, weights could be set on the basis of a single example of the syllable being selected. Even though weights were selected from a single example, selectivity generalized well across other instances, such as repetitions of the selected syllable or appearances of the syllable in a different song.

When presented with different recordings of vocalizations, our syllable-selective neurons fired strongly when syllables similar to the example syllable were played and weakly or not at all to other syllables (Fig. 3, *A*, *B*, and *D*). Normalization within the field L stage of the model plays a critical role in syllable selectivity. Without normalization, model HVc cells become selective to syllables primarily on the basis of their loudness, rather than their spectral characteristics. In the song appearing in Fig. 3, *C* and *D*, loud syllables occur between the two instances of the syllable marked A. Without field L normalization (Fig. 3*C*), an HVc unit set to be selective for syllable A responds more strongly to these loud syllables than to A, a problem that is significantly ameliorated when normalization is included (Fig. 3*D*).

The selectivity for a specific syllable was retained in the presence of noise, although the response decreased in magnitude as the level of noise increased (Fig. 4*A*). In the absence of field L normalization, the syllable-selective unit lost selectivity in the presence of noise and began to respond to the noise rather than to the syllable (Fig. 4*B*). In general, the selectivity of the model was robust when noise, either artificial (white noise) or natural (sounds of other birds) was added to the song, and when noise was introduced through the background, stimulus-independent synaptic input (see METHODS). For example, the model retained a reasonable amount of selectivity when we increased the variance of the synaptic input threefold (data not shown).

Normalization also allowed syllable-selective responses to

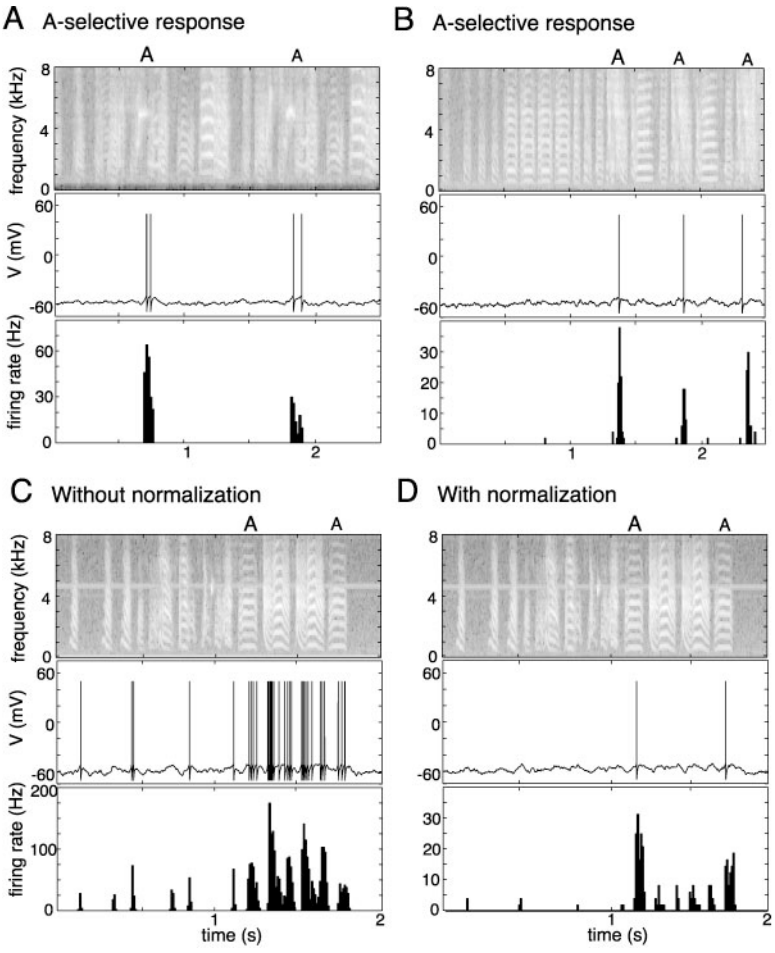

FIG. 3. Syllable selectivity. In all panels, the *top plot* is a song spectrogram, the *middle plot* is a sample voltage trace of a syllable-selective unit, and the *bottom trace* is a histogram of firing rates over repeated runs. Selected syllable is denoted by the letter A, with the larger font indicating the instance of the syllable used to set the synaptic weights in the model. *A*: syllable-selective response. Weights were set using the 1st occurrence of the syllable, but this produced a response selective for both instances within the song. *B*: another example of selectivity using a different bird's song. *C*: response of the model without field L normalization. When the output from the field L stage is not normalized, the syllable-selective neurons respond to the louder syllables, rather than to the one that its weights are selective for. *D*: response of the model to the same song as in *C* with normalization of field L outputs. Responses to loud syllables other than A are suppressed.

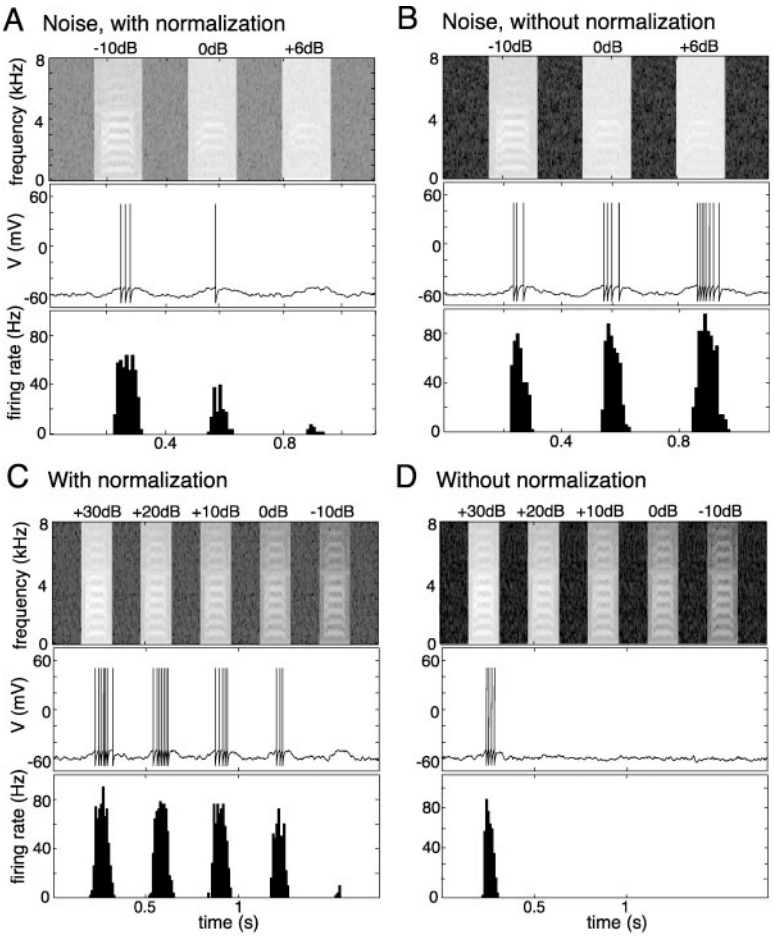

FIG. 4. Effects of noise and volume on syllable selectivity. In all panels, the *top plot* is a spectrogram of the sound input to the field L filters, the *middle plot* is a sample voltage trace of a syllable-selective unit, and the *bottom trace* is a histogram of firing rates over repeated runs. The same syllable is used in each repetition in all panels. In *A* and *B*, the dB labels indicate the level of added noise. In *C* and *D*, the dB labels indicate the volume of the syllable playback. *A*: selected syllable presented along with white noise of increasing amplitude. Number of spikes elicited drops as the amount of background noise increases. *B*: without normalization, the same sequence as in *A* evokes responses than increase with noise due to a loss of response selectivity and increasing response to the noise input. *C*: selectivity remains relatively constant over a range of syllable volumes, although it is minimal for the lowest syllable volume shown (−10 dB). *D*: without normalization, selectively is strongly affected by volume.

persist over a wide range of syllable volumes. In the example of Fig. 4*C*, responses of approximately equal magnitude were retained over a 30-dB range, a feature that was lost when normalization was removed (Fig. 4*D*). For Fig. 4*D*, we adjusted the magnitude of the synaptic conductance carrying the syllable-selective drive from the field L outputs to the model HVc neuron so that the response at +30 dB without normalization matched that with normalization shown in Fig. 4*C*. In this case, no response appeared at any lower levels of song playback. If this adjustment was not made, the model generated unrealistically high firing rates at high stimulus volumes when normalization was removed.

Not all syllables were recognizable by our model. It was easiest to select for syllables with power tightly concentrated at one or a few frequencies, such as whistles and harmonic stacks. Syllables with broadly distributed power generated weaker responses and more false positive responses to incorrect syllables. This is at least partially due to our choice of field L STRFs, because these respond particularly well to harmonic stacks and pure tones. More complicated STRFs, selective for specific frequency sweeps or other features, could provide a better basis set for other types of sounds. Our judgments concerning the accuracy of the model depend on our subjective definition of what constitutes a syllable. Usually this was easy to determine, but in a few noisy cases it was not entirely clear. Of course, what we define and what the bird perceives as distinct syllables may not be the same.

Temporal-combination selectivity

The critical feature that must be added to expand and extend syllable-selectivity to temporal-combination selectivity is a memory trace of the sequence being selected that can gate the response. Figure 5 shows a schematic of the network we used to generate temporal-combination-selective responses. It consists of two subnetworks of excitatory neurons that, by themselves, would be selective for two different syllables labeled A and B. Both of these use the same syllable-selectivity mecha-

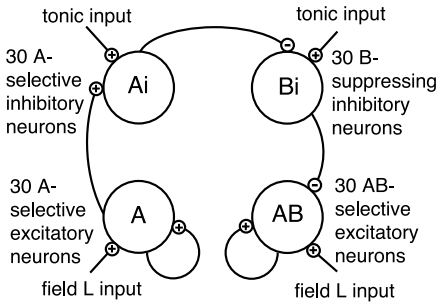

FIG. 5. Schematic of the network for temporal-combination selectivity. Each circle represents a group of neurons (A: A-selective excitatory; AB, AB-selective excitatory; Ai, A-selective inhibitory; Bi, B-suppressing inhibitory). Synapses denoted by pluses and minuses are excitatory and inhibitory. The A- and AB-selective excitatory neurons receive A- and B-selective input from the field L stage. Both sets of inhibitory neurons receive tonic excitation throughout song playback. The A to Ai synapses have a strong *N*-methyl-D-aspartate (NMDA) component.

nism as the neurons discussed in the previous section but, in addition, they have excitatory recurrent connections that amplify their responses. We term these two groups of excitatory neurons A-selective and AB-selective, the latter because the neurons that receive B-selective input from field L end up, in the full network, selective for the temporal sequence AB.

Similar to the proposal of Lewicki and Konishi (1995), temporal-combination selectivity arises from the connections of the A- and AB-selective excitatory neurons to inhibitory neurons. During song playback, the inhibitory neurons receive a constant excitatory synaptic input that, by itself, would keep them active during the song, as is seen experimentally (Mooney 2000). We imagine this input to be the result of pooled excitatory drive from neurons responding to different syllables within the song. In addition to this constant drive, a subset of inhibitory neurons, which we call A-selective inhibitory neurons, receives drive from the A-selective excitatory neurons, carried by NMDA conductances. This excitatory drive retains the memory that syllable A has occurred because of the long time constant (120 ms) of the NMDA conductance. This is reflected in an increased firing rate of the A-selective inhibitory neurons that can last up to several hundred milliseconds after syllable A is presented (Fig. 6). The duration of this effect is longer than the decay time constant of the NMDA conductance because significant excitation remains even if only a fraction of the NMDA conductance is activated.

The A-selective inhibitory neurons inhibit another set of inhibitory neurons, called B-suppressing neurons, which in turn inhibit the AB-selective excitatory neurons. The B-suppressing inhibitory neurons fire persistently at a high enough rate to suppress the response of the AB-selective neurons, except when they are shut off by the A-selective inhibitory neurons for several hundred milliseconds after syllable A occurs. When the persistent inhibition of the B-suppressing neurons is temporarily removed through the action of A-selective inhibitory neurons, the neurons of the AB-selective network respond selectively to syllable B. However, this occurs only if syllable A precedes B, thereby making the neurons AB selective.

When a song containing the sequence AB is presented, the A-selective neurons respond to syllable A, and the AB-selective neurons respond to syllable B (Fig. 6*A*), but only when it is presented after A (Fig. 6*B*). The temporal-combination-selective neurons are specific for the sequence AB, not for an arbitrary syllable followed by B (Fig. 6*C*). Finally, when the interval between inputs to the A-selective neurons and to the AB-selective neurons is increased, the average number of spikes falls off as the response of the B-suppressing interneurons recover from inhibition (Fig. 6*D*). This time course is controlled by the time constant and strength of the NMDA current to the A-selective interneurons, as well as the relative strength of the tonic background input.

It is possible to chain together circuits like this to achieve selectivity to longer sequences such as ABC. If the AB-selective neurons have NMDA-mediated connections to another inhibitory population of neurons, they can behave in a manner similar to the A-selective neurons, gating the response to a subsequent syllable, making a group of ABC-selective neurons. In this way, selectivity for a sequence of arbitrary length can arise (see Fig. 7*C*).

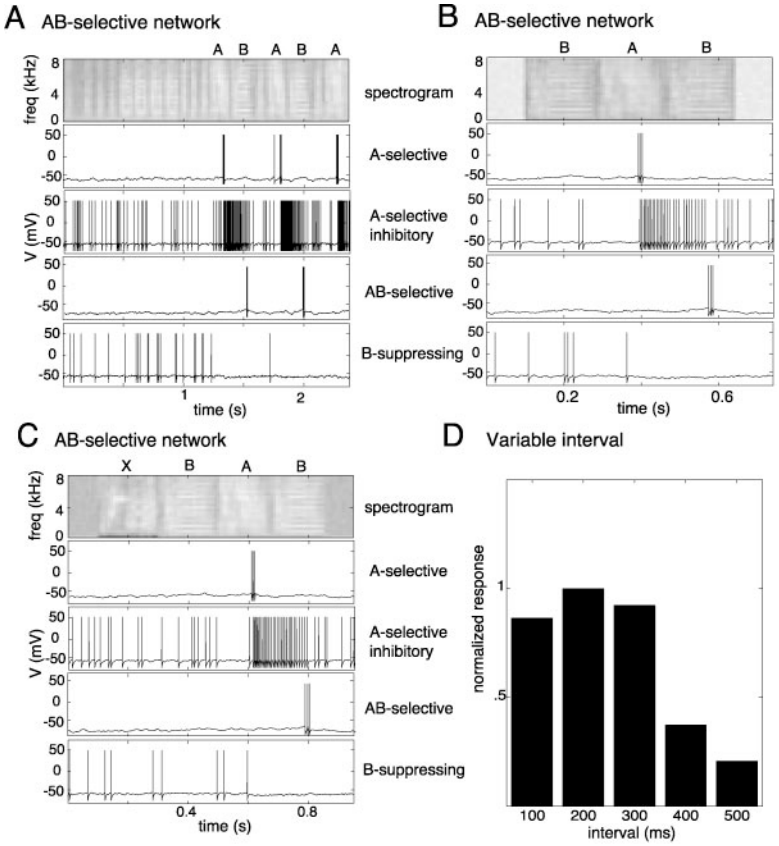

FIG. 6. Temporal-combination-selective responses. In *A–C*, the *top plot* is a song spectrogram, and the other plots, from *top* to *bottom*, are the membrane potentials of an A-selective excitatory neuron, an A-selective inhibitory neuron, an AB-selective excitatory neuron, and a B-suppressing inhibitory neuron. *A*: syllable A evokes a response in the A-selective excitatory and inhibitory neurons, inhibiting the B-suppressing inhibitory neuron, which permits the AB-selective neuron to fire. *B*: response to the sequence AB but not to BA. *C*: response to the sequence AB but not to XB, where X is a different syllable than A. *D*: relative responses of the AB-selective units for different delays between the 2 syllables. Temporal-combination responses survive ≤500-ms separations. In this example, the conductance generated by field L outputs in response to recorded songs were replaced by equivalent conductances pulses representing syllables A and B for us to consider different time delays between these syllables.

Sequence generation

In addition to exhibiting sensory responses, HVc is a motor structure participating in song production as a motor pattern generator (Hahnloser et al. 2002; Margoliash 1997; Vu et al. 1994). The network we have constructed to model temporal-combination selectivity has a particular sequence of syllables built into its circuitry, so it seems reasonable that it too might be capable of generating motor patterns representing the same sequences that it responds to when working in sensory mode. To test this idea, we removed the auditory input from the network model, and replaced the syllable-specific drive to its A-selective and AB-selective neurons with a generic timing signal. This timing signal took the form of periodic excitatory conductance pulses delivered to the A-selective and AB-selective neurons with approximately the same amplitude as the syllable-selective conductances they receive when the network is operating in sensory mode. However, a crucial difference is that the timing pulses do not distinguish between syllables. Thus we have replaced syllable-selective drive to these neurons with a uniform signal the serves only to generate and clock their responses but not to select between them.

We found that, when driven by such a generic timing signal, the same network that gives rise to responses selective for a particular sequence can also generate them. Specifically, we simultaneously stimulated the A- and AB-selective neurons of the network with identical excitatory conductance pulses while the inhibitory neurons received constant input (Fig. 7*A*). The model HVc network produced a similar pattern of activity in response to this generic timing signal as it did for actual auditory song input. The sequencing of responses, A then B, arises from the circuitry of the network by the mechanisms discussed in the previous section. In other words, during the first pulse, A-selective neurons respond, but the AB-selective neurons do not fire because they are inhibited by the B-suppressing interneurons. However, the firing of the B-suppressing neurons is terminated by the activity of the A-selective inhibitory neurons and, on the second pulse, the AB-selective neurons fire. The A-selective neurons do not fire in response to the second timing pulse, although they receive it with the same strength as the first timing pulse, due to the presence of an AHP (see METHODS). There is evidence for such a conductance in RA-projecting neurons from measurements in slice experiments (Dutar et al. 1998).

When a series of pulses is used, the network generates the sequence ABABAB. . . The repetition of the sequence occurs because the time between three timing pulses is sufficient for the A-selective neurons to recover from the AHP (Fig. 7*B*). This motor output is similar to the motif repetition often seen in zebra finch songs. The motif can be generated at a variety of rates (≥25% faster or slower than what is seen in Fig. 7*B*) by increasing or decreasing the repetition rate of the timing pulses. Sometimes, especially for rapid repetition rates, individual units may skip a cycle of the motif because then have not recovered sufficiently from the previous AHP. However, unless then entire population synchronizes these skips, some units will always respond on any given cycle.

The AHP, which suppresses responses for a short period of time following activation, is critical for preventing repeated firing of a single unit to every timing pulse. If the motifs being generated are too long, unwanted repetitions will occur. To see if somewhat longer motifs could be generated, we constructed a network with additional units excited and inhibited by input from a third syllable, C. In other words, populations of C-selective inhibitory, C-suppressing inhibitory, and ABC-selective excitatory units were added to the network in the manner discussed at the end of the previous section (also see METHODS). In addition, the duration of the AHP conductance was increased to provide longer suppression of repeated responses (see METHODS). The result was a network that generated the sequence ABC when stimulated by a nonspecific timing pulse (Fig. 7*C*). Although such a three-component motif can be generated, additional suppression mechanisms would have to be included to generated longer motifs to avoid unrealistically long AHP times.

DISCUSSION

The proposed model of syllable selectivity makes several testable predictions. Because of the syllable-selective weights,

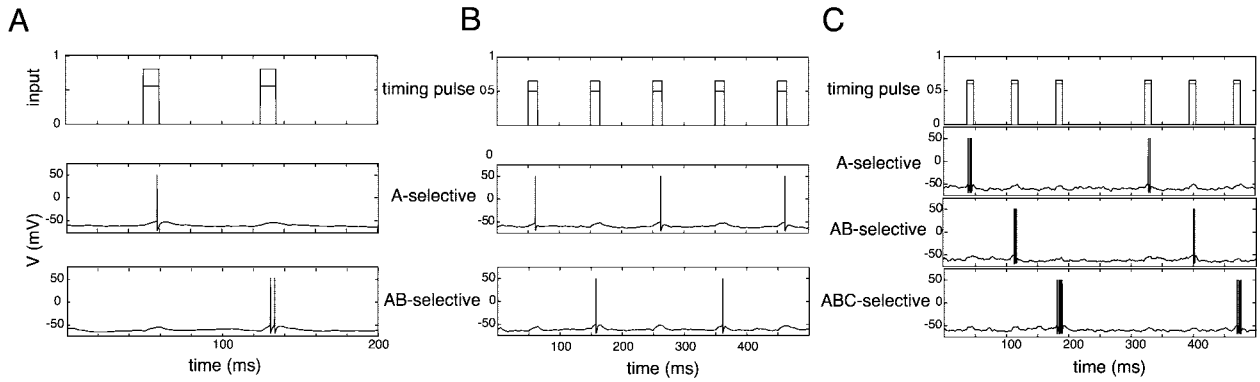

FIG. 7. Network acting as a motor pattern generator. The *top traces* in *A* and *B* are the conductance timing pulses (in units of the resting membrane conductance) to the A- and AB-selective neurons (pulses to the AB-selective neurons are slightly larger than to the A-selective neurons). Other traces are the membrane potentials of an A-selective and an AB-selective excitatory neuron. *A*: pair of excitatory timing pulses to both sets of neurons generates a response in the A-selective neuron followed by a response in the AB-selective neuron. Thus the same temporal sequence that evokes temporal-combination-selective responses in the network is generated by the nonspecific timing pulse input. *B*: a series of timing pulses to the A- and AB-selective populations results in the motor sequence ABABA being produced, similar to motif repetitions found in real songs. *C*: example of an expanded network (see METHODS) generating a 3-syllable motif (ABC). In this panel, the *top plot* shows the pulses to the A- (smaller pulses), AB-, and ABC-selective units (larger pulses), and the other 3 panels show the responses of such units.

individual tone components of a syllable (e.g., a harmonic from a stack) should depolarize a syllable-selective neuron when presented alone, whereas presentation of sound components not in the syllable should not. The normalization step in the model causes syllable-selective responses to remain constant over a range of volumes, something that can be checked experimentally. The weights in the model are not determined by any dynamic learning rule, but they could possibly be generated by the type of winner-take-all rules used in feature-selecting networks (e.g., Hertz et al. 1991). Although little is known about the actual synaptic connections involved in the real circuit, the sparse connectivity used in our model does not seem unreasonable.

The sparse representation of song syllables we have used preserves the salient features of a syllable, its auditory "edges," but discards the rest of the signal, which is more likely to be corrupted by background noise and is more variable from syllable rendition to rendition. The nature of the syllable and the background noise level affect the optimal sparseness of the representation. Sparser representations are best for syllables with power concentrated at a few frequencies or in a noisy background.

As mentioned previously, the proposed model of temporal-sequence selectivity is related to a suggestions of Lewicki and Konishi (1995) that slow inhibitory conductances (elicited by B input and terminated by A input) could sum with excitatory input to generate temporal-combination selectivity. Our model shows that this general mechanism can work using realistic inputs, conductances, and spiking neurons. Furthermore, it indicates that such a model can also generate motor sequences as well as sensory responses.

Temporal-sequence selectivity requires that earlier elements in a sequence gate the response to later elements. In our model, the necessary memory is stored in NMDA conductances. Such a conductance is ideal for this purpose because it activates quickly, allowing fast responses to subsequent syllables, but inactivates slowly retaining the memory of the previous syllable. Metabotropic receptors might be an alternative to NMDA receptors for this purpose, but they have the disadvantage of activating slowly.

In a preliminary version of this work, we constructed a model in which the memory component required for temporal-selective responses arose from reverberating network activity (Drew and Abbott 2002). However, recent recordings suggest that inhibitory neurons play a more prominent role than was assumed in this earlier model (Mooney 2000 and private communication), so we have not considered this possibility here.

The mechanism of response gating that produced temporal-combination selectivity in our model was inhibition of the B-suppressing neurons through prolonged, NMDA-mediated, excitatory drive to the A-selective inhibitory neurons. The mechanism proposed by Lewicki and Konishi (1995) had the prolonged effect of syllable A maintained by slow inhibitory synapses (such as $GABA_B$ conductances) from the A-selective inhibitory neurons to the B-suppressing neurons. We find this approach less favorable because of the observation that inhibitory neurons in HVc exhibit sustained activity throughout the song (Mooney 2000). For generic parameter values in this model, the build up of slow inhibition due to the sustained activity of inhibitory neurons, as seen in Figs. 6 and 7, shuts down the B-suppressing responses independent of whether syllable A occurs, resulting in a loss of temporal-combination selectivity. This can be avoided by adjusting the strength of the slow inhibition onto B-suppressing neurons so that only the A-selective response, and not the sustained level of inhibition, is sufficient to eliminate B-suppressing activity. However, because of the required degree of parameter tuning, the resulting model is less robust than the model we have considered.

Another alternative mechanism is to have NMDA-mediated connections from the A-selective neurons to the AB-selective neurons. This can generate the required selectivity if neither this input nor the B-selective input alone is sufficient to elicit spiking, but their sum is suprathreshold. We studied such a mechanism but found that it produced more variable responses and required more precise parameter tuning than the scheme involving disinhibition of AB-selective units. Another problem with the alternative model is that synaptic parameters that allow the network to detect temporal sequences did not lead to the generation of a motor pattern in response to a nonspecific timing input. Instead, the direct excitatory connection from the A- to AB-selective neurons caused both sets of neurons to fire nearly simultaneously, rather than in sequence.

The model of temporal-combination-selective units we presented predicts that the conductance of an AB-selective neuron should decrease after syllable A is presented due to the removal of B-suppressing inhibition. Furthermore, the time course for the ability of syllable A to affect the response to a subsequent syllable B, as a function of the time interval between these syllables, should match roughly the decay time of the NMDA conductance. A few examples of combination-selective (but not necessarily temporal-combination-selective) neurons showing modulations of response as the gap between syllables was changed suggest this as a possibility (Margoliash 1983; Margoliash and Fortune 1994), but further measurements are needed to test this prediction fully.

In its sequence-generation mode, the model provides a general mechanism for producing structured patterns of activity from generic timing signals. There is evidence that the temporal structuring of song, analogous to our timing pulses, comes from Uva (Vu et al. 1994; Williams and Vicario 1993) or regions below it. The spacing of the syllables generated by the model in its motor mode can be controlled by varying the frequency of the pulses that drive it. In general terms, the model supports the idea that sensory and motor structures, and their mechanisms, need not be thought of as separate entities. In some cases, constructing a network to fill a sensory role may unavoidably lead to a network that can provide motor function as well.

We thank A. Doupe, M. Kao, K. Sen, and the rest of the Brainard and Doupe Laboratories for exceptionally valuable advice and comments and for supplying some of the birdsong recordings we used. We also thank R. Mooney, M. Rosen, and J. Peelle for helpful comments and advice.

This research was supported by the National Science Foundation (IBN-9817194 and IGERT-9972756).

REFERENCES

Buonomano DV and Karmarkar UR. How do we tell time? *Neuroscientist* 8: 42–51, 2002.

Coleman MJ and Mooney R. Source of auditory input to the songbird nucleus HVc revealed by pairwise recordings in NIf and HVc. *Soc Neurosci* Prog. No. 588.4, 2002.

Drew PJ and Abbott LF. Modeling temporal combination selective neurons of the songbird. In: *Computational Neuroscience Trends in Research 2002*, edited by Bower J. Amsterdam: Elsevier, 2002, p. 789–794.
Doupe AJ and Konishi M. Song selective auditory circuits in the vocal control system of the zebra finch. *Proc Nat Acad Sci USA* 88: 11339–11343, 1991.
Doupe AJ and Kuhl PK. Birdsong and human speech: common themes and mechanisms. *Annu Rev Neurosci* 22: 567–631, 1999.
Dutar P, Vu M, and Perkel DJ. Multiple cell types distinguished by physiological, pharmacological and anatomic properties in nucleus HVc of the adult zebra finch. *J Neurophysiol* 80: 1828–1838, 1998.
Fortune ES and Margoliash D. Parallel pathways and convergence onto HVc and adjacent neostriatum of adult zebra finches (*Taeniopygia guttata*). *J Comp Neurol* 360: 413–441, 1995.
Hahnloser RHR, Kozhevnikov AA, and Fee MS. An ultra-sparse code underlies the generation of neural sequences in the songbird. *Nature* 419: 65–70, 2002.
Heeger DJ. Normalization of cell responses in cat striate cortex. *Vis Neurosci* 9: 181–197, 1992.
Hertz J, Krogh A, and Palmer RG. *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison-Wesley, 1991.
Janata P and Margoliash D. Gradual emergence of song selectivity in sensorimotor structures of the male zebra finch song system. *J Neurosci* 19: 5108–5118, 1999.
Katz LC and Gurney ME. Auditory responses in the zebra finch's motor system for song. *Brain Res* 211: 192–197, 1981.
Kelley DB and Nottebohm F. Projections of a telencephalic auditory nucleus—field L—in the canary. *J Comp Neurol* 183: 455–470, 1979.
Konishi M. Birdsong: from behavior to neuron. *Annu Rev Neurosci* 8: 125–170, 1985.
Lewicki MS. Intracellular characterization of song-specific neurons in the zebra finch auditory forebrain. *J Neurosci* 16: 5855–5863, 1996.
Lewicki MS and Arthur BJ. Hierarchical organization of auditory temporal context sensitivity, *J Neurosci* 16: 6987–6998, 1996.
Lewicki MS and Konishi M. Mechanisms underlying the sensitivity of songbird forebrain neurons to temporal order. *Proc Nat Acad Sci USA* 92: 5582–5586, 1995.
Margoliash D. Acoustic parameter underlying the response of song-specific neurons in the white-crowned sparrow. *J Neurosci* 3: 1039–1057, 1983.
Margoliash D. Functional organization of forebrain pathways for song production and perception. *J Neurobiol* 33: 671–693, 1997.
Margoliash D and Fortune ES. Temporal and harmonic combination-sensitive neurons in the song-system of white-crowned sparrows. *J Neurosci* 12: 4309–4326, 1994.
Mooney R. Different subthreshold mechanisms underlie song selectivity in identified HVc neurons of the zebra finch. *J Neurosci* 20: 5420–5436, 2000.
Nottebohm F, Kelley DB, and Paton JA. Connections of vocal control nuclei in the canary telencephalon. *J Comp Neurol* 207: 344–357, 1982.
Sen K, Theunissen FE, and Doupe AJ. Feature analysis of natural sounds in the songbird auditory forebrain. *J Neurophysiol* 86: 1445–1458, 2001.
Schmidt MF and Konishi M. Gating of auditory responses in the vocal control system of awake songbirds. *Nature Neurosci* 1: 513–518, 1998.
Simoncelli EP and Heeger DJ. A model of neuronal responses in visual area MT. *Vision Res* 38: 743–761, 1998.
Theunissen FE, Sen K, and Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J Neurosci* 20: 2315–2331, 2000.
Troyer TW and Doupe AJ. An associational model of birdsong sensorimotor learning I. Efference copy and the learning of song syllables. *J Neurophysiol* 84: 1204–1223, 2000a.
Troyer TW and Doupe AJ. An associational model of birdsong sensorimotor learning II. Temporal hierarchies and the learning of song sequence. *J Neurophysiol* 84: 1224–1239, 2000b.
Vates GE, Broome BM, Mello CV, and Nottebohm F. Auditory pathways of caudal telencephalon and their relation to the song system of adult male zebra finches. *J Comp Neurol* 366: 613–642, 1997.
Vates GE, Vicario DS, and Nottebohm F. Reafferent thalamo-"cortical" loops in the song system of oscine songbirds. *J Comp Neurol* 380: 275–290, 1996.
Vu ET, Mazurek ME, and Kuo YC. Identification of a forebrain motor programming network for the learned song of zebra finches. *J Neurosci* 11: 6924–6934, 1994.
Wang XJ. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J Neurosci* 19: 9587–9603, 1999.
Williams H and Vicario DS. Temporal patterning of song production: participation of nucleus uvaeformis of the thalamus. *J Neurobiol* 24: 903–912, 1993.