

**Eric Larson, Cyrus P. Billimoria and Kamal Sen**

*J Neurophysiol* 101:323-331, 2009. First published Nov 5, 2008; doi:10.1152/jn.90664.2008

**You might find this additional information useful...**

---

This article cites 23 articles, 10 of which you can access free at:

<http://jn.physiology.org/cgi/content/full/101/1/323#BIBL>

This article has been cited by 1 other HighWire hosted article:

**Analyzing Variability in Neural Responses to Complex Natural Sounds in the Awake Songbird**

G. D. Grana, C. P. Billimoria and K. Sen

*J Neurophysiol*, June 1, 2009; 101 (6): 3147-3157.

[\[Abstract\]](#) [\[Full Text\]](#) [\[PDF\]](#)

Updated information and services including high-resolution figures, can be found at:

<http://jn.physiology.org/cgi/content/full/101/1/323>

Additional material and information about *Journal of Neurophysiology* can be found at:

<http://www.the-aps.org/publications/jn>

---

This information is current as of August 7, 2010 .

# A Biologically Plausible Computational Model for Auditory Object Recognition

Eric Larson,<sup>1,2</sup> Cyrus P. Billimoria,<sup>1,2</sup> and Kamal Sen<sup>1,2</sup>

<sup>1</sup>Hearing Research Center and Center for Biodynamics, and <sup>2</sup>Department of Biomedical Engineering, Boston University, Boston, Massachusetts

Submitted 11 June 2008; accepted in final form 29 October 2008

**Larson E, Billimoria CP, Sen K.** A biologically plausible computational model for auditory object recognition. *J Neurophysiol* 101: 323–331, 2009. First published November 5, 2008; doi:10.1152/jn.90664.2008. Object recognition is a task of fundamental importance for sensory systems. Although this problem has been intensively investigated in the visual system, relatively little is known about the recognition of complex auditory objects. Recent work has shown that spike trains from individual sensory neurons can be used to discriminate between and recognize stimuli. Multiple groups have developed spike similarity or dissimilarity metrics to quantify the differences between spike trains. Using a nearest-neighbor approach the spike similarity metrics can be used to classify the stimuli into groups used to evoke the spike trains. The nearest prototype spike train to the tested spike train can then be used to identify the stimulus. However, how biological circuits might perform such computations remains unclear. Elucidating this question would facilitate the experimental search for such circuits in biological systems, as well as the design of artificial circuits that can perform such computations. Here we present a biologically plausible model for discrimination inspired by a spike distance metric using a network of integrate-and-fire model neurons coupled to a decision network. We then apply this model to the birdsong system in the context of song discrimination and recognition. We show that the model circuit is effective at recognizing individual songs, based on experimental input data from field L, the avian primary auditory cortex analog. We also compare the performance and robustness of this model to two alternative models of song discrimination: a model based on coincidence detection and a model based on firing rate.

## INTRODUCTION

An important goal in neuroscience is to understand the biophysical basis for tasks performed by the brain. Theoretical neuroscience can play an important role in clarifying computations associated with tasks and specifying biologically plausible models for implementing the computations. Such a theoretical approach involves relating three different levels of organization—behavior, computation, and biophysics—and requires addressing two fundamental questions connecting these levels: What are the computations associated with specific tasks performed by the brain? How might such computations be implemented by biophysical mechanisms? Previous work has addressed these questions in the context of specific computations. An example of such a computation is multiplicative gain modulation, thought to be associated with important tasks, such as attention modulation while viewing objects and coordinate transformations while reaching for objects

(Salinas and Sejnowski 2001). This computation might be mediated by biophysical mechanisms such as strong recurrent connectivity (Salinas and Abbott 1996), changes in input synchrony (Salinas and Sejnowski 2000), or balanced excitation and inhibition (Chance et al. 2002). Another example is a multiplicative computation associated with a visual neuron sensitive to looming (Gabbiani et al. 2002). These examples motivate the use of a similar approach for other important behaviors mediated by the brain.

An important task performed by humans and many animals is object recognition. This problem has been studied intensively in the visual system (Logothetis and Sheinberg 1996; Riesenhuber and Poggio 2000). In contrast relatively little is known about the recognition of auditory objects. A class of sounds that are of particular importance for human recognition is complex vocal communication sounds, such as speech. The computations associated with such a task and their underlying biophysical mechanisms remain poorly understood. The combined knowledge of vocal communication behavior and underlying neural circuitry makes the songbird an attractive model system for investigating this problem (Doupe and Kuhl 1999). Previous work has outlined a computational method for song recognition using a classification scheme based on a spike distance metric (Machens et al. 2003; Narayan et al. 2006; Wang et al. 2007). However, whether and how neural mechanisms could perform such computations remain unclear. Here, we propose neural circuits that could implement such computations. Specifying such circuits is important for several reasons. First, they provide biophysically plausible implementations of computations underlying song recognition. Second, they make it possible to search experimentally for such circuits in areas where such computations might be performed. Third, they allow the design of artificial electronic circuits that perform similar computations. Such circuits might be used for developing artificial systems that perform complex computations such as sound recognition in a manner that mimics the brain. In this study, we investigate three different neural circuits that implement song recognition, comparing both performance and robustness. Finally, we discuss experimental tests for distinguishing between these alternative models for song recognition.

Address for reprint requests and other correspondence: K. Sen, Hearing Research Center, Department of Biomedical Engineering, Center for Biodynamics, Program in Mathematical and Computational Neuroscience, Boston University, 44 Cummington Street, Boston, MA 02215 (E-mail: kamalsen@bu.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.

## METHODS

*Neural data and analysis*

The van Rossum spike distance metric (van Rossum 2001) gives a measure of the dissimilarity between two spike trains  $a$  and  $b$  by convolving each spike train with a kernel and computing the squared Euclidean distance between them; in our case we use a decaying exponential kernel as

$$\frac{1}{\tau} \int [(a * e^{-t/\tau}) - (b * e^{-t/\tau})]^2 dt \quad (1)$$

Changing the exponential time constant  $\tau$  changes the timescale of analysis. When  $\tau$  is on the order of 1 ms, the van Rossum distance is in a timing regime, acting as a coincidence detector where distance is based on the number of noncoincident spikes. When  $\tau$  is on the order of 1 s, the van Rossum distance is in a rate regime where distance is based on the difference in overall spike rate. When  $\tau$  is between 1 ms and 1 s, the van Rossum distance is in an intermediate rate–time regime where the relative timing and rate of spiking both contribute to the distance.

This metric has previously been implemented in an analytical method for discriminating spike trains using neural data from zebra finch field L, the avian auditory cortex analog (Narayan et al. 2006). Neural responses to 20 distinct conspecific songs over 10 trials were recorded and principal component analysis on spike waveforms was then used to isolate 30 single- or multiunit neural recordings. All procedures were in strict accordance with the National Institutes of Health guidelines as approved by the Boston University Charles River Campus Institutional Animal Care and Use Committee.

The biologically plausible circuits proposed here use this same data set to classify spike trains. Since the songs used to evoke spike trains vary in length from 820 ms to over 4 s, all spike trains are truncated to be 820 ms long so that differences in the song length cannot be taken advantage of to improve discrimination performance. To test discrimination on each of the 30 recordings, the 200 response spike trains are categorized as one of the 20 songs using model discrimination circuits (see following text) or a van Rossum analytical method similar to that used in Narayan et al. (2006). In the analytical van Rossum discrimination method, the dissimilarity between one of the 200 spike trains and 20 “template” spike trains evoked in response to each of the 20 songs at random trial numbers is calculated. The input spike train is categorized as one of the 20 songs based on the minimum dissimilarity between it and those 20 template spike trains. In both the analytical and model discrimination methods, the number of correct categorizations divided by the total number of categorizations (200) yields the percentage correct discrimination for each neuron across all songs and trials.

*Integrate-and-fire model*

All simulations use standard integrate-and-fire (IAF) model neurons as described by Dayan and Abbott (2001). IAF neurons with synaptic inputs have kinetics described by the differential equation

$$\tau_m \frac{dV}{dt} = (E_L - V) + R_m I_e + R_m g_{se} P_{se} (E_{se} - V) + R_m g_{si} P_{si} (E_{si} - V) \quad (2)$$

with the time-varying voltage of the cell  $V$ , the membrane time constant of the cell  $\tau_m$ , the leak or steady-state voltage  $E_L$ , the membrane resistance  $R_m$ , and time-varying injected current  $I_e$ . Parameters for the excitatory and inhibitory synaptic inputs of the cell are the conductances  $g_{se}$  and  $g_{si}$ , time-varying presynaptic inputs  $P_{se}$  and  $P_{si}$ , and synaptic potentials  $E_{se}$  and  $E_{si}$ , respectively, where  $E_{se} > V_{th}$  to drive the cell potential higher and  $E_{si} < E_L$  to drive it lower. IAF

differential equations are numerically integrated and whenever  $V$  is greater than the activation threshold  $V_{th}$ ,  $V$  rises instantaneously to the action potential voltage  $V_{ap}$  and at the next time step drops to the reset voltage  $V_{re}$ .

An absolute refractory period of 2 ms is imposed on all model cells. After convolving the input spike trains with a decaying exponential kernel with time constant  $\tau_e$ , multiple excitatory inputs are summed to produce  $P_{se}$  (and inhibitory inputs summed to produce  $P_{si}$ ) as

$$P_{se} = \sum_j \text{input}_j(t) * e^{-t/\tau_e} \quad (3)$$

All model cells have values  $E_L = -70$  mV,  $V_{th} = -55$  mV,  $V_{ap} = 0$  mV,  $V_{re} = -80$  mV,  $E_{se} = 0$  mV, and  $E_{si} = -90$  mV chosen to be in the physiologically plausible range. Parameters  $R_m g_{se}$ ,  $R_m g_{si}$ ,  $\tau_m$ , and  $\tau_e$  vary across models. For cells that receive inputs only from synapses,  $I_e$  is set to zero and all cells have Gaussian noise ( $\mu = 0$  mV,  $\sigma = 1.5$  mV) injected at each integration time step.

*Discrimination model*

In the van Rossum-like (vR) circuit, two IAF neurons  $D_1$  and  $D_2$  receive synaptic inputs from spike trains  $a$  and  $b$  convolved with decaying exponentials.  $D_1$  is excited by  $a$  and inhibited by  $b$ , whereas  $D_2$  is inhibited by  $a$  and excited by  $b$ . The outputs from  $D_1$  and  $D_2$  inhibit a tonically firing output cell  $S$ , effectively completing a bridge-rectification-like procedure with inversion. The output firing rate of the  $S$  cell is used as the similarity between  $a$  and  $b$ . Instead of the squared difference used in the vR metric, we chose to implement the simpler absolute value operation because using squared differences as inputs did not significantly improve discrimination performance and squaring is more complicated to implement in a biologically plausible circuit (Koch 2004). We chose to invert the dissimilarities to yield similarities because biologically plausible classification using a maximum-type operation was simpler than classification using a minimum-type operation.

Sets of these template-based similarity calculations are then used to discriminate spike trains using a decision network comprising a winner-takes-all circuit.

*Decision network*

We implemented, modified, and extended a decision network based on a model originally proposed by Wang (2002) and Wong and Wang (2006). The initial biophysically realistic model by Wang feeds the outputs of two receptive fields to two competing, mutually inhibiting neuronal populations consisting of thousands of units. The two overall population firing rates of these populations are then thresholded to make a binary stimulus classification. The circuit is designed to use biologically inspired methods to perform integration of information over the course of approximately 1 s—an order of magnitude longer than the longest time constant in the model (of 100-ms *N*-methyl-D-aspartate [NMDA] receptors)—which is accomplished via slow recurrent excitation. In the subsequent work by Wong and Wang, the dynamics of this architecture were simplified using a two-variable model. This simpler reduced model captures the essential features of the original biophysical model, so we chose to implement the reduced model. We extended the reduced model from accepting 2 inputs to  $N$  inputs by linearly combining inhibitory effects of the other  $N - 1$  competing populations on population  $i$  as

$$x_i = J_s S_i - J_D \sum_{j=1 \text{ to } N} (1 - \delta_{i,j}) S_j + I_0 + I_i + I_{\text{noise},i} \quad (4)$$

$$r_i = (Ax_i - B) / [1 - \exp[-D(Ax_i - B)]] \quad (5)$$

$$\frac{dS_i}{dt} = \frac{-S_i}{\tau_s} + (1 - S_i)\gamma r_i \quad (6)$$

where  $r_i$  is the population firing rate,  $x_i$  is an intermediate activity term,  $S_i$  describes the slow NMDA gating, and  $\delta[i, j]$  is the Kronecker delta function. The constants  $A = 270 \text{ (VnC)}^{-1}$ ,  $B = 108 \text{ Hz}$ ,  $D = 0.1540 \text{ s}$ ,  $\gamma = 0.641$ , the network classification threshold  $\theta = 15 \text{ Hz}$ , and the time-varying, filtered noise term  $I_{\text{noise},i}$  are unchanged in our implementation. Three parameters are changed from the original Wong and Wang model to optimize performance: the base current  $I_0$ ; the cross-population inhibition coefficient  $J_D$ ; and the intrapopulation excitation coefficient  $J_S$ .

The input currents  $I_i$  (in Eq. 4) for the decision network are the vR circuit outputs, in the form of  $S$  neuron outputs convolved with 100-ms decaying exponentials. The inputs in the original Wong and Wang model had an average root-mean-square (RMS) amplitude of 0.0156 nA, so our input currents were scaled down by a factor of 442 and 61 for the vR and coincidence detection models, respectively, to match the average RMS input current levels across neurons. When combined with template-based vR model calculations, this modified model accomplishes the dual task of integrating information over the entire stimulus and providing a nonlinear winner-takes-all classification of the input as one of the 20 input songs. Although the songs are truncated to 820 ms to prevent the use of song duration information in classification, the decision network is allowed to run beyond this time to make a decision when necessary.

### Alternative models of discrimination

For comparison to the vR model, we developed simple coincidence detection (CD) and rate detection (RD) models that are also used to classify stimuli. The CD model consists of a single neuron per trained song, which calculates the similarity between the trained song and the input song by counting coincident spikes between the memory spike train and sensory response to the input song. The same decision network described previously is then used to classify songs. The RD circuit uses  $N$  classifier neurons to classify the  $N$  groups of spike trains based on overall firing rate. After assigning each of the  $N$  neurons a range of spike rates, each classifier neuron is parameter-tuned to fire only when the input spike rate exceeds the lower bound of the spike rate range for its assigned group. Each of the  $N$  classifier neurons has back-propagating inhibition that inhibits the firing of all neurons tuned to rates below its turn-on rate and each is coupled to an auxiliary neuron via mutual excitation such that one action potential leads to repeated firing. All  $N$  classifier neurons are fed the input spike train and the one that is firing at the end of the stimulus assigns the spike train to its group. The decision time—when the  $N$  neurons were examined to determine which was firing—is fixed at 820 ms, the length of the shortest song.

To maximize the performance of the rate discrimination model on the neural recordings, an algorithm was developed to search the space of possible rate decision boundary sets. By iteratively moving bound locations, while using the fact that moving one decision boundary affects classification of only two spike train groups, boundaries are placed to maximize the resulting discrimination performance score. In the field L neural recordings, spike rate distributions of responses to the 20 songs have large variances and significantly overlap, with spike rates ranging from 0 to 90 Hz ( $\sigma = 25 \text{ Hz}$ ) and an average trail-to-trial SD of 4.8 Hz. Thus rate-based discrimination is generally poor. To test the RD classification scheme on rate-discriminable data, an artificial data set of 20 groups of 10,000 spike trains is generated with normally distributed spike rates. The mean spike rates span 10 to 200 Hz in 10-Hz increments to represent the 20 groups, each with 2 Hz SD, making the ideal classification decision thresholds the midpoints between distribution means. Histograms of the generated distributions are shown in Fig. 4B.

### Model optimization

The resistance–conductance products  $R_m g_{si}$  and  $R_m g_{se}$ , the membrane time constant  $\tau_m$ , and the exponential kernel time constant  $\tau_e$  are optimized independently for the three models using the data set from zebra finch field L. A grid search algorithm over the parameter space maximizes the mean percentage correct discrimination performance across all songs, trials, and neurons. All parameters for each model are fixed across neurons (except  $R_m g_{se}$  for the rate detection circuit and two parameters for the decision network; see following text). Tuning and setting optimal parameters for each neuron independently increases the percentage correct discrimination of the CD and vR models an average across neurons of 3.1% ( $\sigma = 4.3\%$ ) and 4.1% ( $\sigma = 1.9\%$ ) compared with fixed tuning, respectively.

Optimal parameters for the vR model are approximately  $R_m g_{se} = 6.0$ ,  $R_m g_{si} = 30.7$ ,  $\tau_e = 10 \text{ ms}$ , and  $\tau_m = 42 \text{ ms}$  for the difference calculating  $D_1$  and  $D_2$  neurons. The exponential time constant of 10 ms agrees with the analytical van Rossum value of  $\tau = 13 \text{ ms}$  found by Narayan et al. (2006). To make the inverting neuron  $S$  tonically fire, it is fed a constant current, with optimal parameters  $R_m I_e = 102 \text{ mV}$ ,  $R_m g_{si} = 0.72$ ,  $\tau_e = 38 \text{ ms}$ , and  $\tau_m = 20 \text{ ms}$ . Optimal parameters for the CD model similarity neuron are  $R_m g_{se} = 0.3$ ,  $\tau_e = 1 \text{ ms}$ , and  $\tau_m = 2 \text{ ms}$ . Optimal parameters for the RD model rate-thresholding classifier neurons are  $\tau_e = 2 \text{ ms}$  and  $\tau_m = 10 \text{ s}$ . The input strength  $R_m g_{se}$  parameter in the RD model determines the spike rate above which each rate-thresholding neuron would begin to fire, so these values are set for each song's thresholding neuron in each neuron's data set individually. The input strength  $R_m g_{se}$  generally takes on values in the range 0.03 to 0.2. Additionally, the auxiliary neuron excitatory strength  $R_m g_{se} = 2,700$  to cause rapid, repeated firing once threshold is reached and the back-propagating inhibition strength  $R_m g_{si} = 2,000$  to stop firing in neurons tuned to lower rates.

Three parameters of the decision network network are optimized. The base input current level  $I_0 = 0.67 \text{ nA}$  is fixed across neurons and models.  $J_S$ , the intrapopulation excitation level, and  $J_D$ , the cross-population inhibition level, are tuned individually for each neuron in both the CD and vR models because fixing these two parameters across all neurons causes an average performance decrease in each model of 3.5 and 5.7%, respectively. Optimal  $J_S$  values fall in the range 0.0001–0.40 nA, whereas optimal  $J_D$  parameters fall in the range 0.3–150 nA. Even with this per-neuron decision network optimization, substituting a perfect-integrating winner-takes-all maximum operation increases the performance of both the CD and vR models. The CD model achieves a mean percentage correct of 34.1%–6.5% better than with the decision network—outperforming the analytical CD method on 20 neurons for an overall average 2.8% better. The vR model achieves a mean percentage correct of 50.3%–9.5% better than with the decision network—outperforming the analytical method on 22 of the 30 neurons for an overall average 2.5% better.

### Spike train analysis and statistics

To help characterize the outputs of the models, spike trains are analyzed in terms of overall spike rate, sparseness, and reliability. The sparseness of spike trains is calculated by examining spike counts  $r_i$  in  $N$  poststimulus time histogram bins (bin width 10 ms) as described by Vijne and Gallant (2000):  $S = [1 - (\sum r_i/N)^2 / (\sum r_i^2/N)] / [1 - (1/N)]$ . Using the method described by Schrieber et al. (2003), neuronal reliability is calculated by averaging normalized inner products of output spike trains (filtered with a Gaussian kernel, zero mean,  $\tau = 10 \text{ ms}$ )  $s_i$  as  $R_{\text{corr}} = 2/[N(N-1)] \sum_{i=1}^{N-1} \sum_{j=i+1}^N \langle s_i, s_j \rangle / (\|s_i\| \|s_j\|)$ . Correlation coefficients were calculated using the Pearson product moment test. One-way ANOVA was used to determine performance differences between models.

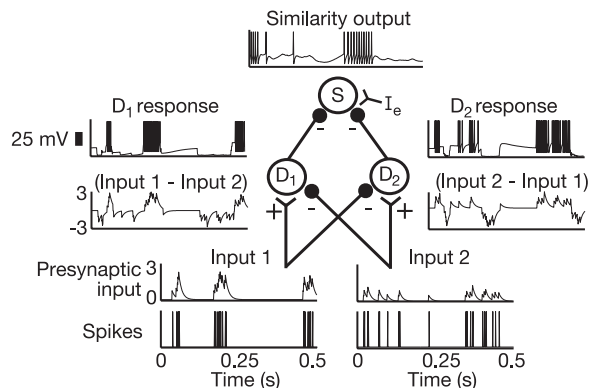


FIG. 1. The van Rossum-like (vR) circuit encodes a van Rossum-like distance between 2 input spike trains via a bridge-rectification scheme. The spike trains *Input 1* and *Input 2* are taken by *D*<sub>1</sub> as excitatory and inhibitory and by *D*<sub>2</sub> as inhibitory and excitatory, respectively. Each *D* neuron calculates either the upper or lower half of  $|Input 1 - Input 2|$ . Neuron *S* uses the external excitatory input *I*<sub>e</sub> to invert the sum of the 2 inhibitory inputs from *D*<sub>1</sub> and *D*<sub>2</sub>, yielding the total similarity.

RESULTS

*A biologically plausible model for song recognition*

Our first model, the vR model, was inspired by the van Rossum spike distance metric (van Rossum 2001). The vR model uses three IAF neurons to calculate the dissimilarity between two spike trains *a* and *b*. It first performs a bridge rectification- or absolute value-like procedure on the difference between the two spike trains through cells *D*<sub>1</sub> and *D*<sub>2</sub> and then it integrates and inverts the result using the *S* cell to output a total similarity. Figure 1 illustrates the similarity calculation performed by this circuit. Both the analytical van Rossum distance and this circuit calculate the distance between two spike trains by convolving with a kernel, taking the difference, rectifying the result, and integrating, but this circuit uses an absolute value-like operation instead of a squaring operation to rectify the signal, coupled with an inversion process to calculate a similarity instead of dissimilarity (see METHODS).

This vR-inspired dissimilarity model was extended to categorize spike trains using a decision network consisting of a

template-matching winner-takes-all circuit (see METHODS). Given an input spike train and a set of memory spike trains, the dissimilarity is calculated between the input and each memory. The memory with the minimum dissimilarity from the input spike train is chosen as the match using the decision network. See Fig. 2, *A* and *B* for an illustration and example of this classification decision scheme. Figure 2*B* shows the integration over time of the decision network. The firing rates of the network settle to around 1 Hz before the stimulus onset at 0 ms, after which, in this example, the correct song's population firing rate climbs to the decision threshold at about 410 ms. The high firing rate of the winning population then suppresses the firing of the other decision networks, creating a winner-takes-all scenario.

The decision network is allowed to take longer than the stimulus duration of 820 ms to settle on a decision. A histogram of the decision times of the vR decision network for all songs, trials, and neurons is plotted in Fig. 2*C*. For the vR and CD networks, decisions occur after 820 ms 9.2 and 2.8% of the time, respectively; no decision is reached by 1,640 ms 0.77 and 0.62% of the time, respectively; and two songs are chosen (two population responses cross the 15-Hz threshold, taken to be an incorrect response) 0 and 0.01% of the time, respectively.

To validate the vR model as an adequate representation of the analytical method, the performances of the analytical van Rossum method and the model vR circuit were compared. In Fig. 2*D* the vR model discrimination performance is plotted against the van Rossum analytical method performance in the rate-time scheme (exponential kernel time constant  $\tau = 13$  ms). The model discriminated better than the analytical method for 3 of the 30 neurons, and the analytical van Rossum and model vR performances across neurons were significantly correlated with  $R = 0.96$  ( $P < 0.001$ ).

*Alternative models*

For comparison with the vR model, simple coincidence detection and rate detection models were devised. Compared with the vR model, the CD model uses finer timescales. It

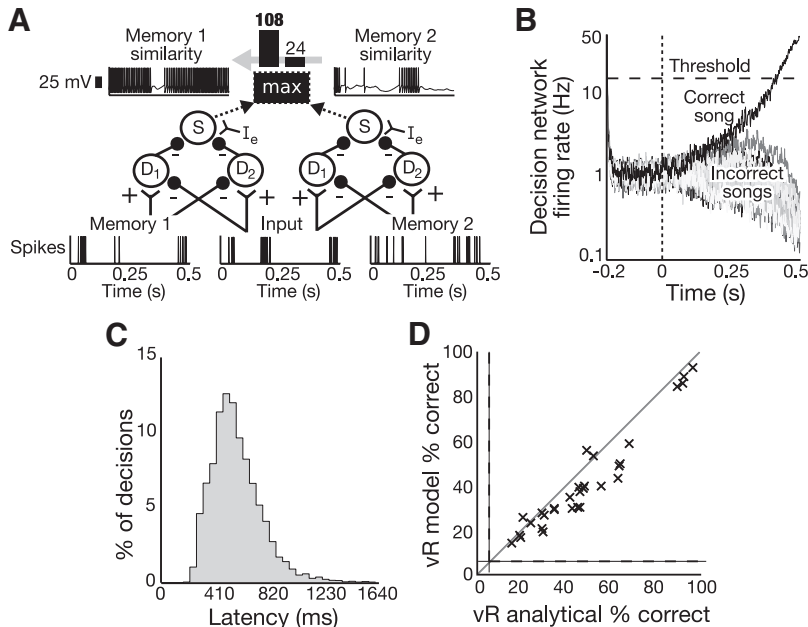


FIG. 2. The vR discrimination model performs on par with discrimination using the analytical van Rossum discrimination method in a rate-time regime. *A*: 2 vR-like distance calculations are performed between the *Input* spike train and the *Memory 1* and *Memory 2* spike trains. The output rate is greater for the *Memory 1* similarity than that for the *Memory 2* similarity, so *Input* is classified as song 1. *B*: the decision network thresholds 20 different responses to the similarity calculations to classify inputs. The time course of the classification of the *Input* spike train, with the correct song in black and the 19 incorrect songs in grayscale, is shown with the decision network integrating information from the start time of 0 ms beyond the classification time around 410 ms. *C*: a histogram of decision network latencies across all songs, trials, and neurons shows that most decisions are made before the 820-ms input cutoff. *D*: performance classifying spike train recordings from 30 units (x's) in zebra finch field L in response to 20 different songs for the vR model is plotted against the analytical van Rossum performance (with  $\tau = 11$  ms). Equal performance is the gray line; chance performance levels are indicated by the dashed lines.

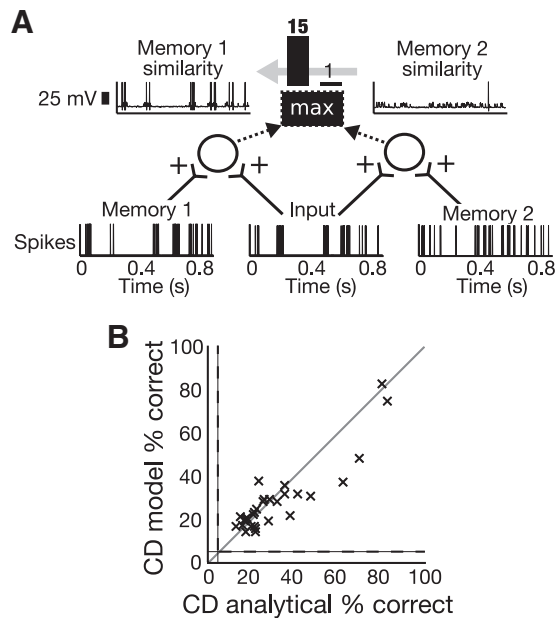


FIG. 3. A discrimination model that uses coincidence detection (CD) performs similarly to the analytical van Rossum method in a timing regime. *A*: 2 CD-based similarity calculations are performed for an *Input* spike train and spike trains *Memory 1* and *Memory 2*. The output rate is greater for the *Memory 1* similarity than that for the *Memory 2* similarity, so the *Input* is classified as song 1. *B*: performance classifying 20 songs using spike train recordings from 30 different units (x's) in zebra finch field L. CD model discrimination is plotted against performance of analytical van Rossum method (with  $\tau = 2$  ms). Equal performance is the gray line; chance performance levels are indicated by the dashed lines.

calculates a measure of similarity between input and memory spike trains by counting the number of simultaneous spikes. A single neuron was tuned to fire only when two excitatory synaptic inputs fired simultaneously, allowing the output rate of the neuron to yield a measure of spike train similarity. A winner-takes-all template-matching scheme was again used for song discrimination, whereby similarities between the input train and each song memory are calculated and the memory spike train with the maximum similarity to the input is chosen as the match using the decision network. Figure 3*A* shows an example of the CD discrimination scheme.

The quality of the CD model was investigated by comparing the analytical van Rossum method with a short time constant ( $\tau = 2$  ms) to the performance of the CD model. Figure 3*B* shows the CD model discrimination performance versus that of the van Rossum analytical method in the timing regime. The CD model performed better than the analytical method for 14 of the 30 neurons and the performances were significantly correlated with  $R = 0.90$  ( $P < 0.001$ ).

We next examined a rate detection model. To classify  $N$  groups of spike trains based on spike rate, each group of spike trains is assigned a nonoverlapping continuous range of spike rates. In the classification scheme, any input spike train that falls within a group's spike rate range is assigned to that group. The RD circuit has  $N$  classifier neurons for the  $N$  groups of spike trains; each neuron fires only when the input spike rate exceeds the lower bound of the spike rate range for its assigned group and has back-propagating inhibition to turn off all neurons tuned to rates below its turn-on rate. By using long membrane time constants, these  $N$  classifier neurons act as

spike-count-thresholding neurons. Each of these  $N$  neurons is also coupled to its own auxiliary neuron via mutual excitation, such that a single firing leads to sustained activity. The one neuron of the  $N$  classifier neurons that is firing at the end of the stimulus—which is always the neuron with the highest activation threshold that fired in response to the input—assigns the input to its group. See Fig. 4*A* for an illustration of this categorizing circuit.

To examine this rate detection model, we looked at the relationship between the performance of the RD model and the van Rossum analytical method. Figure 4*C* shows the RD model discrimination performance versus that of the analytical method in the rate scheme (time constant  $\tau = 10$  s). The performance of the RD scheme and analytical method were significantly correlated with  $R = 0.46$  ( $P < 0.02$ ).

The field L data were not easily rate discriminable because analytical and model rate discrimination never exceeded 35% accuracy. To obtain rate-discriminable data, 20 groups of artificial spike trains were generated (see METHODS) with spike rate distributions shown in grayscale in Fig. 4*B*. Analytical

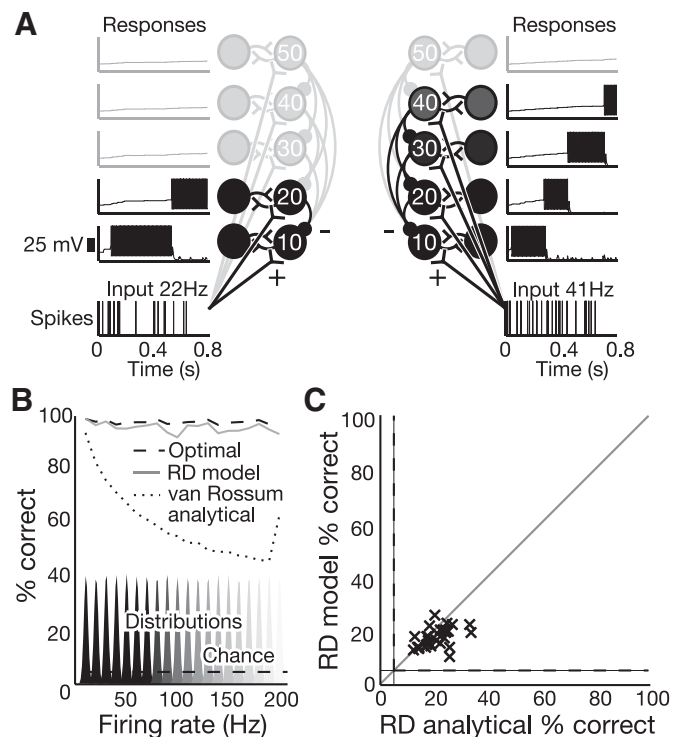


FIG. 4. *A*: rate detection (RD) inhibitory network accurately discriminates spike train groups with minimally overlapping spike rate distributions. The RD model (responses shown for 2 sample inputs in *A*) consists of one neuron for each of the 20 spike rate groups that fires only above a certain rate threshold, thereafter inhibiting the firing of all neurons tuned to lower rates, and a mutually exciting auxiliary cell that leads to sustained firing after a single spike; spike trains are categorized based on which characteristic neuron is firing at the end of the stimulus. *B*: performance discriminating 20 groups of 10,000 artificial spike trains generated with normally distributed mean rates linearly spaced from 10 to 200 Hz ( $\sigma = 2$  Hz, distributions in grayscale). The RD model (solid gray) performs close to analysis using optimal decision thresholds (dashed black) and outperforms the analytical van Rossum (with  $\tau = 10$  s, dotted line). *C*: the RD model performs worse than the analytical van Rossum method in a rate scheme on field L data. RD model performance classifying spike train recordings from 30 units (x's) in zebra finch field L in response to 20 different songs is plotted against the performance of analytical vR methods ( $\tau = 10$  s), with equal performance on the gray line and chance performance on the dashed lines.

performance classifying spike trains based on optimal theoretical rate thresholds lies directly above the RD model performance using neurons tuned to those thresholds (Fig. 4B). At around 96% correct, the RD model proposed here matched the accuracy of analytical methods, with the difference in performance never exceeding 5.3% across all 20 spike train groups. The performance of the analytical van Rossum method, on the other hand, decreased as spike rates increased, dropping to >51% below the model performance.

#### Comparing the three models: performance and robustness

The vR model outperforms the CD model, which in turn outperforms the RD model in discriminating the recorded spike trains. Figure 5A shows the percentage correct of classification using each of the three models for each biological neural recording across all 20 songs and 10 trials, as well as the mean percentage correct across 30 neural recordings  $\pm$  1SE. Using one-way ANOVAs, the mean vR percentage correct of 40.8% was significantly better than the mean CD percentage correct of 27.6% ( $P = 0.01$ ), as well as the mean RD percentage correct of 17.3% ( $P < 0.001$ ), and the CD percentage correct was

significantly better than the mean RD percentage correct ( $P < 0.001$ ).

The relative robustness of the models was tested in three ways: by adding Gaussian jitter to individual spikes; by adding Gaussian jitter to the starting time of memory playback; and by adding or removing spikes. In the CD and vR models, template memory spike trains were manipulated and, in the RD model, which did not have memory spike trains, the input trains themselves were manipulated. Spike train jitter and onset jitter were each zero mean with SD that varied from 0 to 160 ms. For spike removal, 0 to 100% of the original number of spikes were deleted and, for spike addition, 0 to 100% of the original number of spikes were spuriously added to the spike trains. The normalized errors of all three models due to spike removal/addition, spike jitter, and onset jitter are shown in Fig. 5, C–E, respectively. The normalized error is given by one minus the percentage correct under modification divided by the percentage correct in the base case (zero jitter, no addition/deletion).

The vR model normalized performance decrease was less than that of the CD model due to jitter, whereas the RD model was hardly affected. Using one-way ANOVAs, the normalized error due to jitter for the vR model was significantly less than that of CD across all nonzero jitters  $\leq 40$  ms ( $P < 0.026$  for all) and RD normalized error was significantly less than vR normalized error across all jitters  $\geq 6$  ms ( $P < 0.001$  for all). It should be noted that although the normalized error was significantly less for the RD model, the raw percentage correct under jitter of the RD model was never significantly better than that of the vR model.

The relative levels of performance error across models due to onset jitter were very similar to those due to spike jitter. Using one-way ANOVAs, the normalized error due to jitter for the vR model was significantly less than that of CD across all nonzero jitters  $\leq 40$  ms ( $P < 0.027$  for all) and RD normalized error was significantly less than vR normalized error across all jitters  $\geq 6$  ms ( $P < 0.039$  for all). It should be noted that although the normalized error is significantly less for the RD model, the raw percentage correct under jitter of the RD model was never significantly better than that of the vR model. Performance of the vR model was significantly different from the base case (no onset jitter) for jitter levels  $\geq 10$  ms ( $P < 0.029$ ).

The RD model was less robust to spike addition and deletion than the CD and vR models. The normalized error due to spike addition/deletion of the RD model was significantly worse than that of either CD or vR across all spike deletions  $>5\%$  ( $P < 0.001$  for all), except for 100% deletion, and across all spike additions  $>5\%$  ( $P < 0.001$  for all). The normalized error was significantly less for CD than that for vR for spike additions  $\geq 52\%$  ( $P < 0.023$  for all) and all spike deletions  $>32\%$  ( $P < 0.012$ ). It should be noted that, whereas the normalized error of the CD model was significantly better than that of the vR model for many additions/deletions, the raw performance of the CD model was never significantly better than that of the vR model.

To help characterize the inputs and outputs of the vR model, input and output spike trains were examined in terms of spike time reliability, overall firing rate, and sparseness. The output of the vR model at the summation level  $S$  was split into two categories: one for responses to input spike trains that should be matched to the template train and the other for responses to

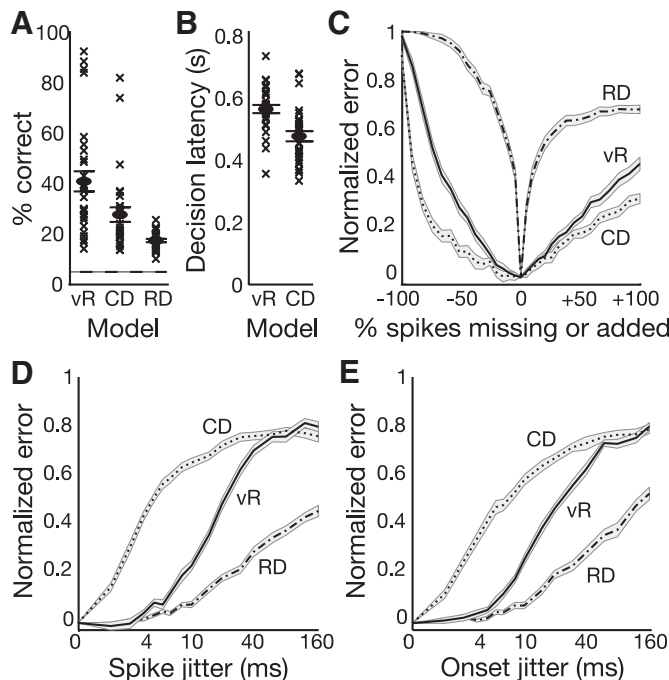


FIG. 5. The vR model outperforms the CD and the RD models and is robust to spike addition, moderate spike deletion, and jitter. *A*: model performance in discriminating neural recordings from 30 zebra finch field L sites in response to 20 different songs. For all 3 models, percentage correct is plotted for each neuron (x's) with the mean across neurons  $\pm$  SE. *B*: the average latency (time between stimulus onset and classification decision) of each neuron's decision network (x's) is plotted for the CD and vR discrimination models. *C–E*: the normalized error  $\pm$  SE across all neurons is (*C*) plotted against the percentage of spikes removed or added from the playback bank; (*D*) plotted against applied jitter (adding Gaussian jitter to each spike time individually); or (*E*) plotted against applied onset jitter (uniformly translating all spikes in a memory spike train by a normally distributed random amount). The normalized error is given by the percentage correct for a given case divided by the percentage correct in the zero jitter, no spike deletions/additions case. Performance of the vR and CD models decrease  $<25\%$  for 45% spike deletion, whereas the RD model performance rapidly degrades. For both spike jitter and onset jitter, the vR model greatly outperforms CD, whereas the RD model is unaffected by small amounts of jitter.

input spike trains that should not be matched to the template train. The reliability, firing rate, and sparseness of the output  $S$  cell in response to correct inputs were all significantly correlated with performance ( $R = 0.57$ ,  $R = 0.60$ ,  $R = -0.64$ , respectively,  $P < 0.001$  each), whereas only the sparseness of responses to incorrect inputs were significantly correlated with performance ( $R = 0.60$ ,  $P < 0.001$ ). The reliability and sparseness of the input data set itself were also positively correlated with performance ( $R = 0.48$ ,  $P = 0.008$ ;  $R = 0.42$ ,  $P = 0.02$ ).

## DISCUSSION

The ability to discriminate and recognize objects is a fundamentally important function of the brain. Previous studies in the visual system have provided insights into some of the computations associated with this important problem (Logothetis and Sheinberg 1996; Riesenhuber and Poggio 2000). However, relatively little work has been done in the auditory system. It is hard to extrapolate the knowledge gained from visual studies to the auditory system due to the many differences between these modalities. Perhaps most important, most of the visual studies used static images to probe recognition, whereas, in audition, the temporal dimension is critical for the perception of auditory objects, such as words. To our knowledge this study is the first to propose a biologically plausible computational model for auditory object recognition. For concreteness we focused on the songbird system and used real experimental data as input to our model. However, the structure of the model is general and may be implemented in a wide variety of auditory recognition tasks including the recognition of speech in humans. Furthermore, this model could be used to discriminate and recognize spike trains from other sensory modalities, where spike timing and the pattern of spikes are important. For example, this model could be used to discriminate between spike trains in responses to video in the visual system. As proposed, this model does not, however, deal with certain challenges posed by variability in stimuli. For example, the model does not contain mechanisms for dealing with variations in playback speed or intensity. As proposed, the model would likely only demonstrate intensity-invariant object recognition when fed field L inputs, which themselves showed intensity invariance, such as recordings from Billimoria et al. (2008).

Our computational models for auditory object recognition were inspired by a spike distance metric: the van Rossum metric. The basic idea is to use a measure of distance to cluster objects, a fundamental concept in pattern recognition. Although the van Rossum metric is somewhat biologically plausible, it does not specify the neural mechanisms underlying the distance computation. This makes it difficult to search experimentally for van Rossum metric-like calculations or build synthetic dissimilarity circuits. The vR model proposed here gives us an explicit, biologically plausible vR circuit layout for use in discrimination and recognition tasks.

Mazurek et al. (2003) used an analytical network as a winner-takes-all firing rate difference detection circuit. In their network, the firing rates of two spike trains  $a$  and  $b$  are compared by calculating  $\int (a - b)dt$  and  $\int (b - a)dt$  and thresholding both results. The first integrated difference to reach the threshold determines categorization as  $a$ -like or  $b$ -like. The vR circuit proposed here functions differently in

three important ways: 1) using IAF neurons instead of analytical firing rate subtraction rectifies the differences  $a - b$  and  $b - a$ ; 2) in the vR model the importance of the firing rate versus the precise spike timing of  $a$  and  $b$  can be tuned via time constants, instead of dealing exclusively with overall rate; and 3) summing together the rectified differences  $a - b$  and  $b - a$  (and inverting them) provides a measure of overall *similarity* instead of providing a difference in overall activity levels. This explicit connection to dissimilarity effectively makes the vR model a pattern-recognition circuit. The effectiveness of this pattern recognition versus the effectiveness of rate detection can be important, as seen in the performance improvement of the vR model over the rate detection model on the field L data set.

### *Model performance, robustness, and plausibility*

The vR model circuit accomplishes vR metric-like calculations using basic units of three neurons. In a template-matching winner-takes-all discrimination scheme with input data from field L, the rate-time-based vR model outperforms both the RD-based and CD-based discrimination methods. vR model circuit performance is highly correlated with the performance of the analytical van Rossum method, suggesting that the circuit is truly performing van Rossum-like calculations. Additionally, the vR model outperformed both the CD and RD models by significant amounts, which agrees with the findings of Narayan et al. (2006) that methods using intermediate rate-time information in spike trains outperform those using timing or rate information exclusively.

When the decision network was replaced with a perfect-maximum-selecting operation, we found that the vR model actually outperformed the analytical van Rossum discrimination method by a small amount. This slight increase in performance is due to additional tunable parameters in the model. Tuning the excitatory and inhibitory synaptic strengths in difference-calculating neurons enhances performance because the relative levels of excitation and inhibition—as well as the overall synaptic conductance levels in relation to the firing threshold—determine the amount of spiking that occurs in response to spike rate-time differences. In the analytical method, all rate-time differences contribute to the dissimilarity calculation; in the model, appropriate synaptic weights cause lower-level rate-time differences to be disregarded in the spike counts of the output cell  $S$  because the subthreshold potentials they cause do not make the difference-calculating  $D_1$  and  $D_2$  neurons spike. This effective denoising capability suggests that biological or artificial implementations of this vR circuit, if combined with a good upstream decision mechanism, can perform as well as the analytical van Rossum method through parameter tuning.

The performance differences between the RD model and the analytical van Rossum method in a rate regime are likely due to three factors. First, the analytical van Rossum method, even with the long time constant  $\tau = 10$  s, still gives weight to the timing of individual spikes because the duration of time that the spike counts are different determines the contribution to the dissimilarity. This helps the analytical method on the field L data set where timing is relatively precise, but hurts it on the artificial data set where spike timing is random. Second, the

model thresholds the single input train based on the statistics of a large group of spike trains instead of comparing two randomly selected spike trains, as the analytical method does. This comparatively hurts the model on the field L data set because the spike rate distributions for songs overlap significantly, but it helps the model on the artificial set because the group statistics are a good predictor of individual spike train rates. Third, as seen with the artificial data set, the analytical van Rossum performance degraded with increasing overall firing rate, although the field L data set had an average response firing rate across neurons, songs, and trials of only 25 Hz. Since the field L data set had such relatively low firing rates, this degradation did not occur, allowing the analytical vR model to outperform the RD model.

In addition to requiring only three cells per similarity calculation, the vR similarity-calculating circuit functions well for a broad range of cell parameters. The vR model IAF neurons had cell membrane parameters  $E_L$ ,  $V_{th}$ ,  $V_{ap}$ ,  $V_{re}$ ,  $E_{se}$ , and  $E_{si}$  taken to be near canonical values in the physiologically plausible range (Dayan and Abbott 2001) and the optimized parameters  $R_m J_e$ ,  $R_m g_{se}$ ,  $R_m g_{si}$ ,  $\tau_m$ , and  $\tau_e$  were all fixed across neurons and stayed in the physiological range. The vR model was also robust to parameter changes, requiring at least a 39% increase or decrease in one of the four tuned parameters (whereas the other three were held fixed) of the difference-calculating cells to cause a 10% normalized performance decrease. This type of parameter robustness or flexibility should lend itself to biological and artificial implementations.

### Experimental testing

Each of the three models discussed here should manifest differently in vivo. A coincidence detection circuit would have sparser outputs with increased activity for one specific input. The rate detection circuit would have little to no activity during most of the stimulus, followed by some rapid firing where timing relative to the activity of other recognition neurons was crucial. The vR circuit would have outputs that were stronger for one specific input. We would expect CD, vR, and RD schemes to manifest with membrane and synaptic time constants on the order of 1 ms, 10 ms, and 1 s, respectively, for optimal performance. To the best of our knowledge, these parameters have not been measured in upstream areas—such as cmM—thought to affect neurons selective for conspecific songs (Gentner and Margoliash 2003) and would make useful future work.

Once a candidate  $S$  cell is identified by its response selectivity, the input stimuli can be modified to help classify it as a vR-circuit neuron. Introducing precisely timed errors in the selected input song should cause precisely timed activity changes in the selective cell. Comparing the responses of these modified to the unmodified stimuli should allow for the observation of activity increases for spike dissimilarity and activity decreases for spike similarity cells.

The recognition model proposed here requires comparing the incoming sensory activity pattern with stored patterns in memory. To perform such a comparison a real system would need to activate a “playback” of auditory patterns stored in memory. One way to achieve this would be for the onset of

auditory activity to trigger the playback. This does not seem implausible, given that onset cues are strong throughout the auditory system. This onset cue could also provide the source of current input used by the  $S$  cells to invert the dissimilarity-based activity of the difference cells  $D_1$  and  $D_2$ . Although the mechanism of memory playback is unknown, playback of activity patterns has been observed in the songbird song production system during sleep (Dave and Margoliash 2000), as well as in other areas such as the hippocampus (Louie and Wilson 2001; Wilson and McNaughton 1994) and visual cortex (Ji and Wilson 2007), also during sleep. Finding playback of neural activity patterns corresponding to memorized songs in awake birds would lend support to this type of recognition model.

### ACKNOWLEDGMENTS

We thank R. Narayan for the neural data, R. Maddox for helpful discussions, and K.-F. Wong and X.-J. Wang for providing the MATLAB code for the decision network model.

### GRANTS

This work was supported by National Institute on Deafness and Other Communication Disorders Grant 1R01 DC-007610-01A1 and is based on work supported under a National Science Foundation Graduate Research Fellowship awarded to E. Larson.

### REFERENCES

- Billimoria CP, Kraus BJ, Narayan R, Maddox RK, Sen K. Invariance and sensitivity to intensity in neural discrimination of natural sounds. *J Neurosci* 28: 6304–6308, 2008.
- Chance FS, Abbott LF, Reyes AD. Gain modulation from background synaptic input. *Neuron* 35: 773–782, 2002.
- Dave AS, Margoliash D. Song replay during sleep and computational rules for sensorimotor vocal learning. *Science* 290: 812–816, 2000.
- Dayan P, Abbott LF. *Theoretical Neuroscience*. London: MIT Press, 2001.
- Doupe AJ, Kuhl PK. Birdsong and human speech: common themes and mechanisms. *Annu Rev Neurosci* 22: 567–631, 1999.
- Gabbiani F, Krapp HG, Koch C, Laurent G. Multiplicative computation in a visual neuron sensitive to looming. *Nature* 420: 320–324, 2002.
- Gentner TQ, Margoliash D. Neuronal populations and single cells representing learned auditory objects. *Nature* 424: 669–674, 2003.
- Ji D, Wilson MA. Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat Neurosci* 1: 13–15, 2007.
- Koch C. *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford Univ. Press, 2004.
- Logothetis NK, Sheinberg DL. Visual object recognition. *Annu Rev Neurosci* 19: 577–621, 1996.
- Louie K, Wilson MA. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron* 29: 145–156, 2001.
- Machens CK, Schutze H, Franz A, Kolesnikova O, Stemmler MB, Ronacher B, Herz AV. Single auditory neurons rapidly discriminate conspecific communication signals. *Nat Neurosci* 6: 341–342, 2003.
- Mazurek ME, Roitman JD, Ditterich J, Shadlen MN. A role for neural integrators in perceptual decision making. *Cereb Cortex* 13: 1257–1269, 2003.
- Narayan R, Grana G, Sen K. Distinct timescales in cortical discrimination of natural sounds in songbirds. *J Neurophysiol* 96: 252–258, 2006.
- Riesenhuber M, Poggio T. Models of object recognition. *Nat Neurosci* 3, Suppl.: 1199–1204, 2000.
- Salinas E, Abbott LF. A model of multiplicative neural responses in parietal cortex. *Proc Natl Acad Sci USA* 93: 11956–11961, 1996.
- Salinas E, Sejnowski TJ. Impact of correlated synaptic input on output firing rate and variability in simple neuronal models. *J Neurosci* 20: 6193–6209, 2000.
- Salinas E, Sejnowski TJ. Gain modulation in the central nervous system: where behavior, neurophysiology, and computation meet. *Neuroscientist* 7: 430–440, 2001.
- Schreiber S, Fellous JM, Whitmer D, Tiesinga P, Sejnowski TJ. A new correlation-based measure of spike timing reliability. *Neurocomputing* 52–54: 925–931, 2003.

- van Rossum MC.** A novel spike distance. *Neural Comput* 13: 751–763, 2001.
- Vijne WE, Gallant JL.** Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287: 1273–1276, 2000.
- Wang L, Narayan R, Grana G, Shamir M, Sen K.** Cortical discrimination of complex natural stimuli: can single neurons match behavior? *J Neurosci* 27: 582–589, 2007.
- Wang X-J.** Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* 36: 955–968, 2002.
- Wilson MA, McNaughton BL.** Reactivation of hippocampal ensemble memories during sleep. *Science* 5172: 603–604, 1994.
- Wong K-F, Wang X-J.** A recurrent network mechanism of time integration in perceptual decisions. *J Neurosci* 26: 1314–1328, 2006.