

# Quantitative approaches to the study of neural coding

©Bertrand Delgutte, 2003-2005

# The Problem

- Neural and behavioral data have different forms that are not easily comparable
  - In some cases, data from different species and states (awake vs. anesthetized) are being compared
  - Even when neural and behavioral data have similar forms (e.g. a threshold), they both depend on arbitrary criteria
- Both neural and behavioral responses are probabilistic: They are not identical for different repetitions of the same stimulus. Therefore, they provide a noisy (imprecise) representation of the stimulus.
- Are there quantitative tools to describe the precision of stimulus representation in neural and behavioral data?

# Questions that can be addressed by quantitative methods

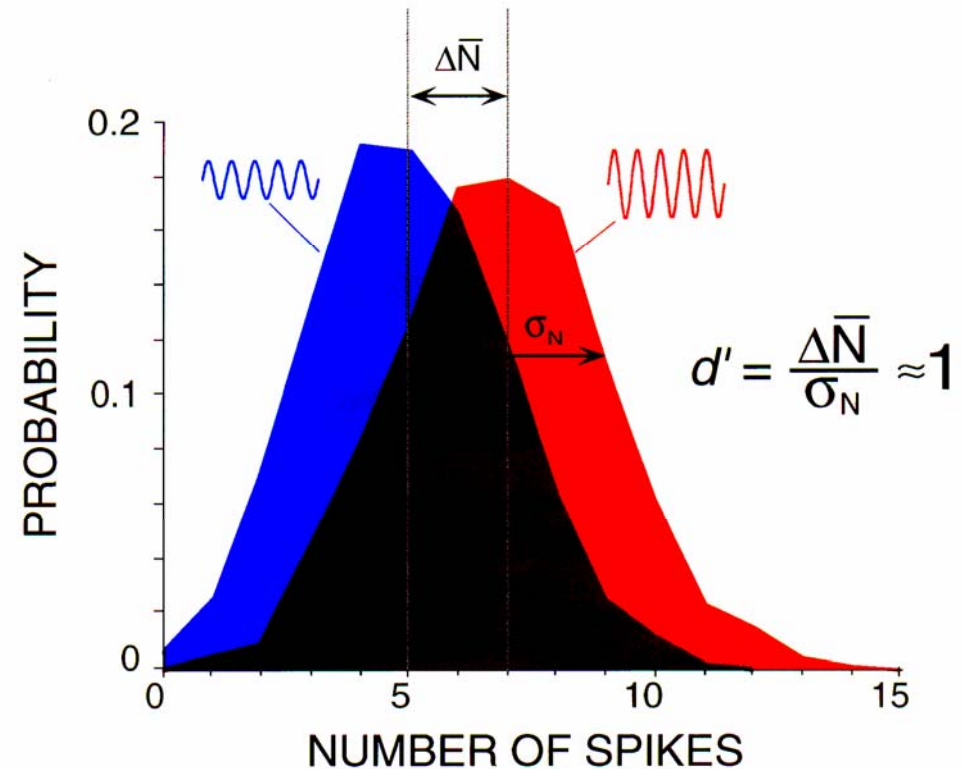
- Does a particular class of neurons provide sufficient information to account for performance in a particular behavioral task?
- Does a particular neural code (e.g. rate-place vs. temporal) provide sufficient information for the task?
- Which stimulus feature is a particular neuron or neural population most sensitive to? Are there “feature detector” neurons?
- Is there stimulus information in the correlated firings of groups of neurons?
- How efficient is the neural code?

# Outline

- Signal detection theory (a.k.a. ideal observer analysis)
  - Single neuron
  - Combined performance for neural population
- Shannon information theory

# Neural variability limits detection performance

- When a sound stimulus is presented repeatedly, the number of spikes recorded from an auditory neuron differs on each trial. The blue and red surfaces are model spike count distributions for two pure tones differing in intensity by 3 dB.
- The overlap in the spike count distributions limits our accuracy in identifying which of the two stimuli was presented based on the responses of this neuron.
- A measure of the separation between the two distributions (and therefore of the neuron's ability to discriminate) is the *discriminability index*  $d'$ , which is the difference in means of the two distributions divided by their standard deviation.
- The *just noticeable difference* (JND) or *difference limen* (DL) is often taken as the intensity increment for which  $d' = 1$ . This criterion corresponds to 76% correct in a two-interval, two-alternative psychophysical experiment.



[Delgutte \(unpublished\)](#)

# Conditional Probability and Bayes' Rule

Conditional Probability:

$$P(S | R) = P(S, R) / P(R)$$

Statistically Independent Events:

$$P(S, R) = P(S) P(R)$$

$$P(S | R) = P(S)$$

Bayes' Rule:

$$P(S | R) = \frac{P(R | S) P(S)}{P(R)}$$

# Bayesian Optimal Decision and the Likelihood Ratio

## The Problem:

- Choose between two alternatives (stimuli)  $S_0$  and  $S_1$  with prior probabilities  $P(S_0)$  and  $P(S_1)$  given the observation (neural response)  $R$  so as to minimize the probability of error.
- Conditional probabilities (stimulus-response relationships)  $P(R|S_0)$  and  $P(R|S_1)$  are known.

## Bayes' Optimal Decision Rule:

- Given a specific value of  $R$ , choose the alternative which maximizes the posterior probability:

$$P(S_1 | R) \underset{S_0}{\overset{S_1}{>}} P(S_0 | R)$$

## Equivalent Likelihood Ratio Test (LRT):

LRT separates priors (under experimenter control) from neuron's stimulus-response characteristic.

$$LR = \frac{P(R | S_1)}{P(R | S_0)} \underset{S_0}{\overset{S_1}{>}} \frac{P(S_0)}{P(S_1)}$$

# Properties of the Likelihood Ratio

$$LR = \frac{P(R | S1)}{P(R | S0)} > \frac{P(S0)}{P(S1)}$$

- Separates computation (left side) from biases and costs (prior probabilities, right)
- If the conditional probabilities  $P(R|S_i)$  are either Poisson or Gaussian with equal variances, LRT reduces to comparing the response  $R$  with a threshold  $\gamma$ :

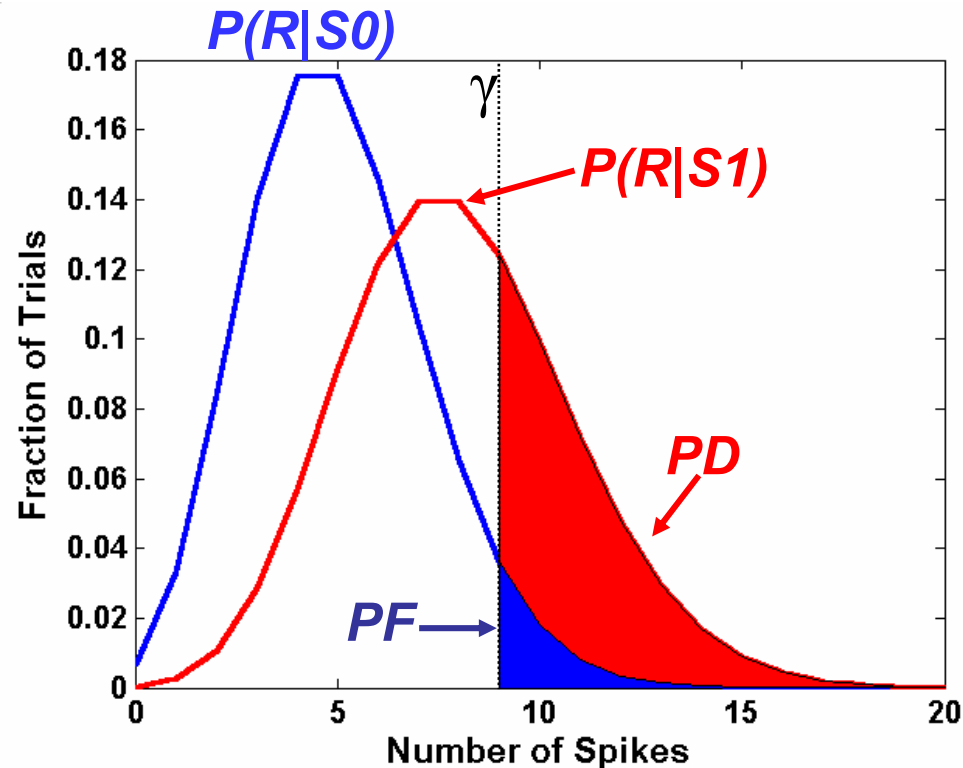
$$R > \gamma$$
$$R < \gamma$$

- LRT works not only for scalar observations (e.g. a spike count from a single neuron), but also for multidimensional observations (e.g. temporal discharge patterns and ensembles of neurons).
- Decision rule invariant to monotonic transformation (e.g. logarithm)
- For *independent* observations (e.g. simultaneous observations from multiple neurons or multiple observation intervals from same neuron),  $\log LR$  is additive:

$$\log LR(R1, R2, \dots, RN) = \sum_{i=1}^N \log LR(Ri)$$



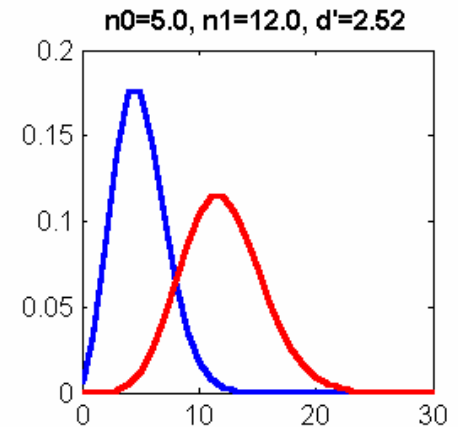
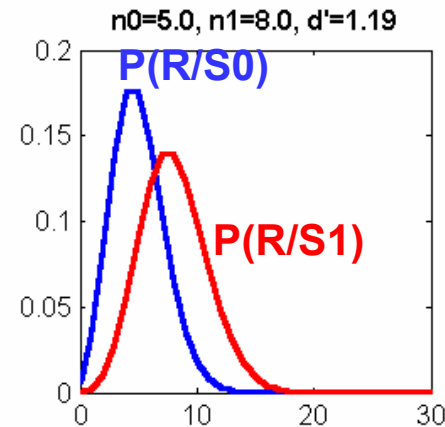
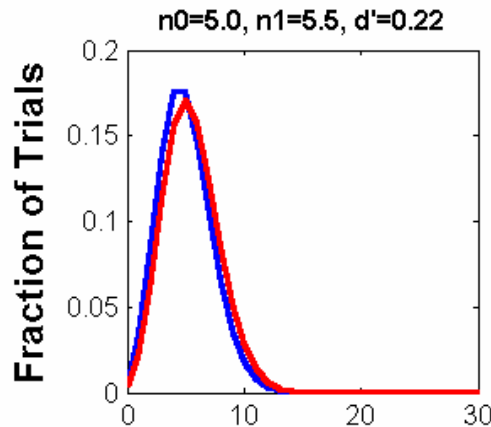
# Detection and False Alarm Probabilities



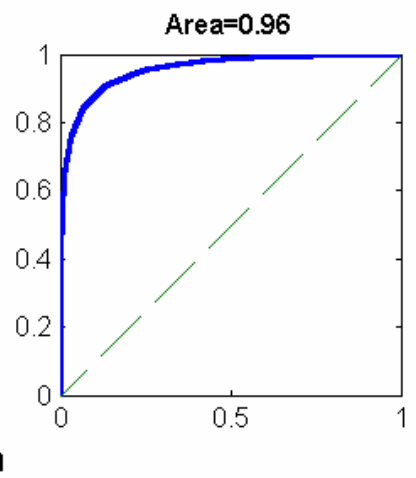
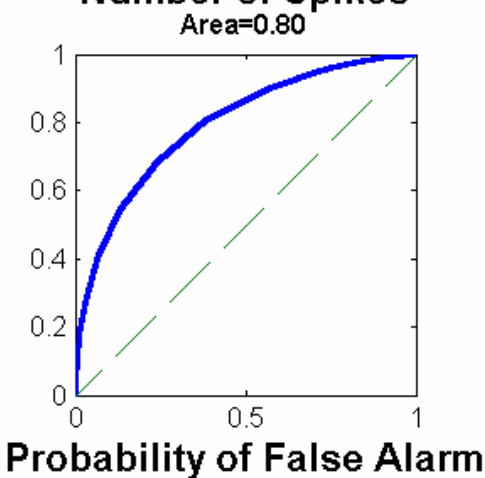
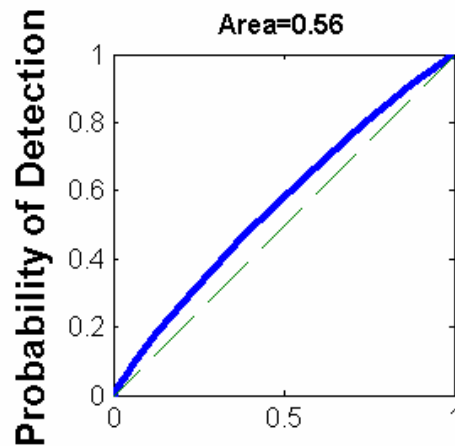
- The neural response (in general LR) is compared to a criterion to make a decision about which stimulus was presented
- Two types of errors: “misses” and “false alarms”
- $PF$  is the probability of false alarm
- $PD$  is the probability of detection;  $PM = 1 - PD$  is the probability of a miss

# Receiver Operating Characteristic (ROC)

Conditional Probabilities (Poisson)



ROC Curves



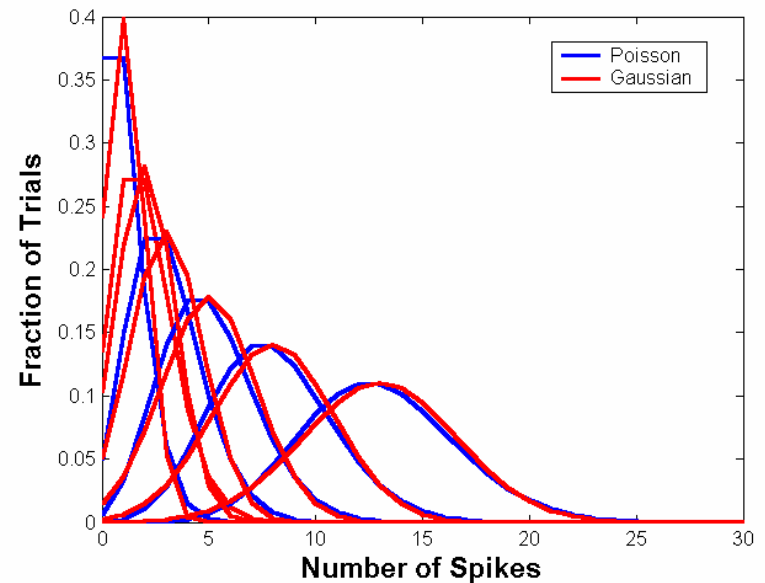
- The area under the ROC curve gives a distribution-free measure of performance

# $d'$ as a measure of performance

- If the conditional probabilities  $P(R|S0)$  and  $P(R|S1)$  are Gaussian with equal variances, then  $d' = \Delta r/\sigma$  completely determines the performance (the ROC curve):

$$PC_{2I} = 1/\sqrt{2\pi} \int_{-d'/\sqrt{2}}^{\infty} e^{-z^2/2} dz$$

- Many probability distributions (including Poisson) approach a Gaussian when the mean response becomes moderately large.
- If so, performance in non-Gaussian cases can be approximated by  $d'$ . If, as in the Poisson case, the variances are unequal under the two alternatives, they can be averaged.
- $d'$  is a poor measure of performance when the number of spikes is very small or when the two stimuli are widely separated



# Visual discrimination of random-dot patterns by monkeys and cortical neurons

Figures removed due to copyright reasons.

Please see:

Figures from Newsome, W. T., K. H. Britten, C. D. Salzman, and J. A. Movshon. "Neuronal mechanisms of motion perception. Cold Spring Harbor Symp." *Quant Biol* 55 (1990): 697-705.

- Single-unit recordings from Area MT (middle temporal) of awake macaques
- Single-unit responses and behavioral performance were recorded simultaneously
- Performance of typical neuron matches behavioral performance

# Forward masking in the auditory nerve?

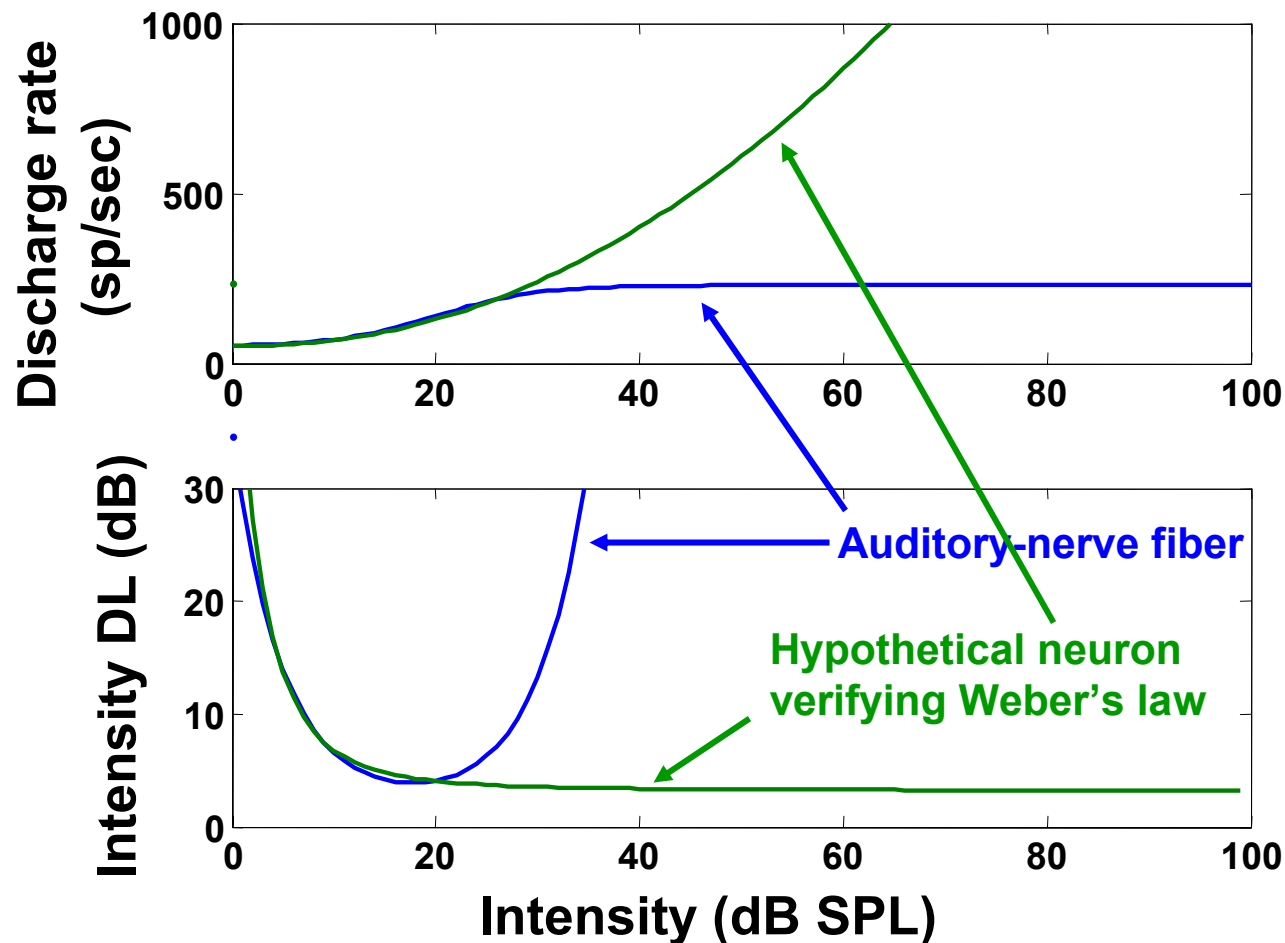
Figures removed due to copyright reasons.

Please see:

Relkin, E. M., and C. W. Turner. "A reexamination of forward masking in the auditory nerve." *J Acoust Soc Am* 84, no. 2 (Aug 1988): 584-91.

Psychophysical masked thresholds grow much faster with masker level than neural thresholds. Maximum masking can reach 30-50 dB.

# Performance based on spike count from single AN fiber severely deviates from Weber's law



Poisson statistics assumed

# The “lower envelope principle”

Somatosensory: Detection of  
sinusoidal vibration

Auditory: Pure tone detection

Figure removed due to copyright  
reasons.

Figure removed due to copyright reasons.

Please see:

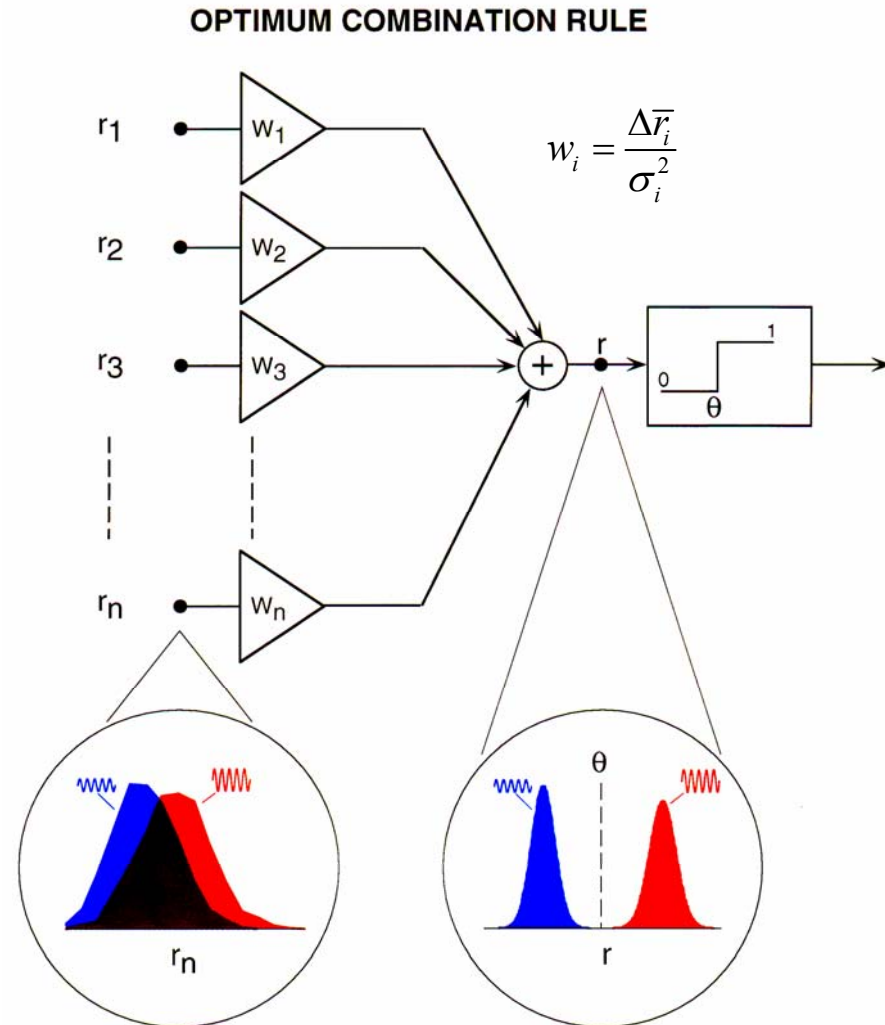
Delgutte, B. “Physiological models for  
basic auditory percepts.” *Auditory  
Computation*. Edited by H. Hawkins, and  
T. McMullen. New York, NY: Springer-  
Verlag, 1996, pp. 157-220.

Mountcastle (1972)

Psychophysical performance is determined by the most  
sensitive neuron in the population.

# Optimal pooling of information across neurons

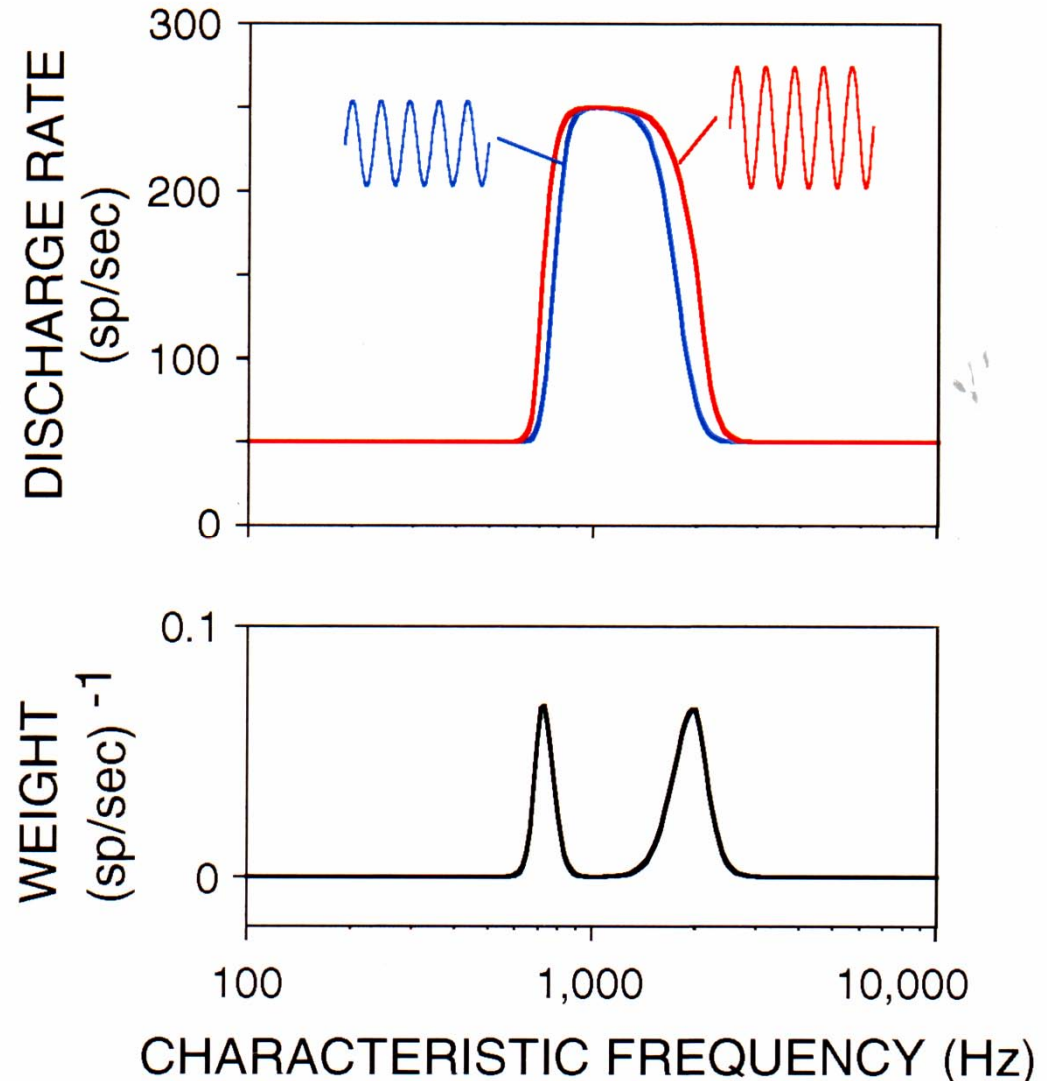
- Performance in discriminating two stimuli can be improved by combining information across neurons. Specifically, if the spike counts from  $N$  neurons are either Poisson or Gaussian and statistically-independent, the optimum combination rule is to form a weighted sum of the spike counts. The discriminability index for this optimum combination is given by  $d^2 = \sum_i d_i'^2$ .
- The structure of the optimum detector is identical to that of a single-layer perceptron in artificial neural networks. The weights can be interpreted as the strengths of synaptic connections and the threshold device as the threshold for all-or-none spike discharges in a postsynaptic neuron.
- When responses of the different neurons are not statistically independent, benefits of combining information may be better or worse than in independent case depending on the nature of the correlations.



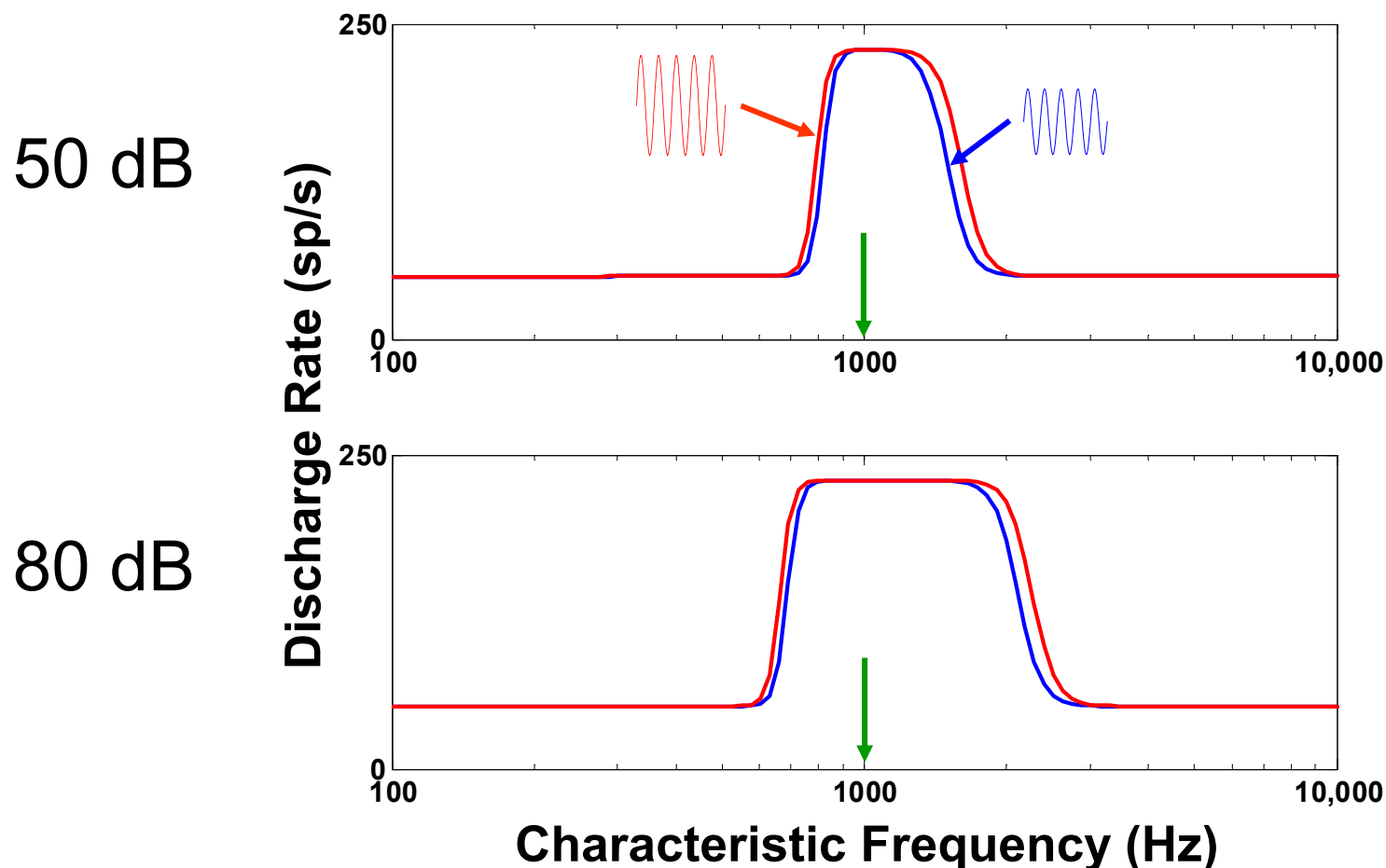


# Rate-Place Model for Intensity Discrimination (Siebert)

- Siebert (1968) developed the first model for predicting the performance of a sensory system based on the activity in primary neurons. His model incorporated all the key features of AN activity known at the time, including logarithmic cochlear frequency map, cochlear tuning, saturating rate-level functions and Poisson discharge statistics.
- The model predicts constant performance in pure-tone intensity discrimination over a wide range of intensities (Weber's law) by relying on the unsaturated fibers on the skirts of the cochlear excitation pattern.
- Psychophysical experiments (Viemeister, 1983) have since ruled out this model because listeners show good performance even when masking noise restricts information to a narrow band around the tone frequency.



Spread of excitation to remote cochlear places makes model predict Weber's Law...



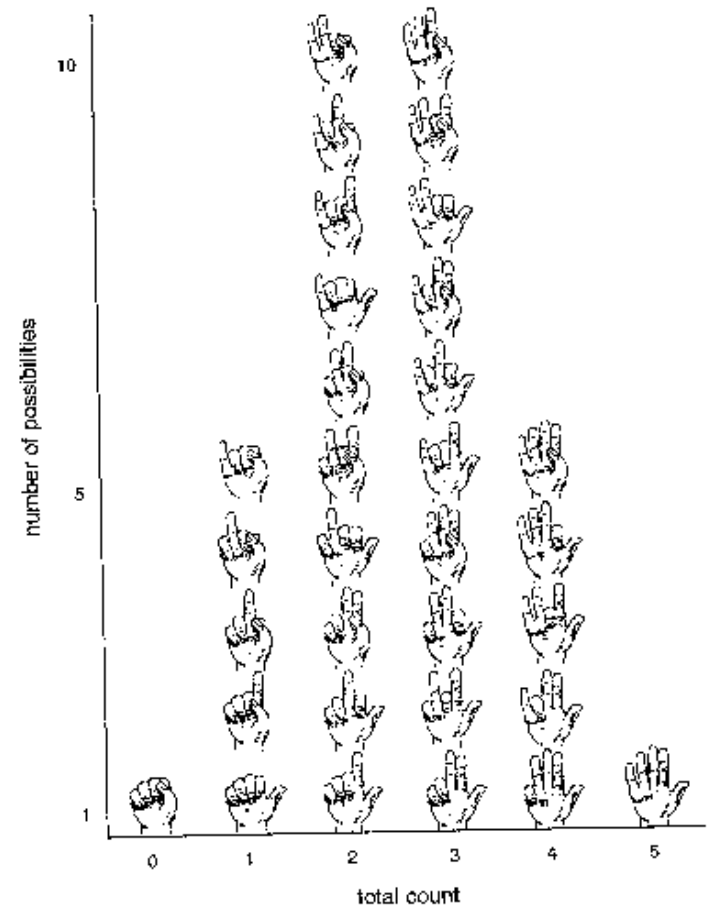
...but performance severely degrades in **band-reject noise**.

# Rate vs. Temporal Codes

**Figure 3.3**

Encoding numbers in a digital code. Probably the most popular code used by people ordering drinks is one in which only the total number of raised digits carries the information. In this code, one hand can carry  $\log_2(6) \approx 2.58$  bits of information (0 included, but not appreciated by any waiter). If we imagine the fingers to be time bins in a discretization of the spike train, with finger up (down) denoting the presence (absence) of a spike, then this conventional “bar code” is equivalent to a “rate code”—only the total number of spikes in the five bins, and not their temporal sequence, carries information. But, as the figure makes clear, if we keep track of “timing” and allow the position of each finger to carry information, then one hand can convey  $2^5 = 32$  distinct messages, or 5 bits of information. This finger code has a greater capacity for carrying information, but the bar code is more robust as the message is, for example, invariant to being viewed in a mirror. This robustness derives from the redundancy of the code, since one number may be represented by several combinations of finger positions. One could also imagine neural codes in which particular patterns of spikes—represented here as particular finger configurations—are endowed with special significance.

- Temporal code can convey more information than rate code
- Distinction between rate and temporal codes not that easy:
  - Rate has to change when new stimulus occurs. How fast can it change and still be called “rate code”?
  - Poisson process completely specified by instantaneous firing rate. If instantaneous rate tracks stimulus waveform, is it a rate or a temporal code?



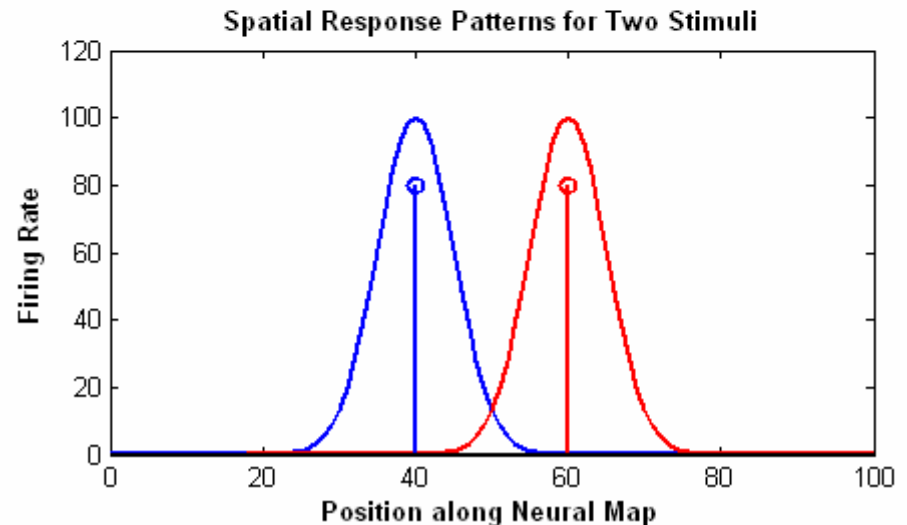
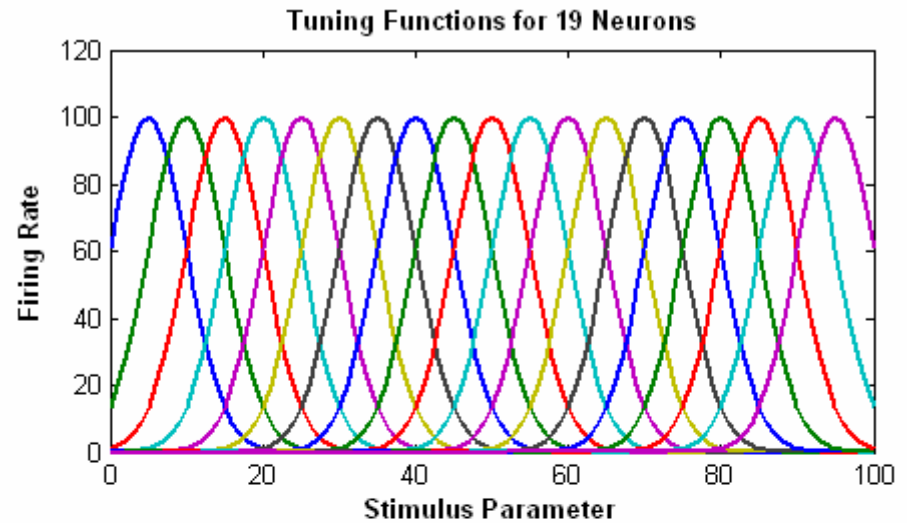
From:

Rieke, Fred, David Warland, Rob de Ruyter vanSteveninck, and William Bialek. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press/Bradford Books, 1997 (c).

Used with permission.

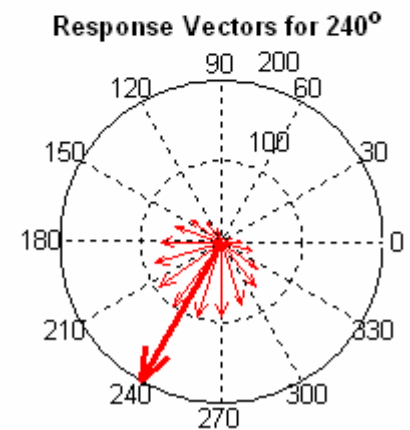
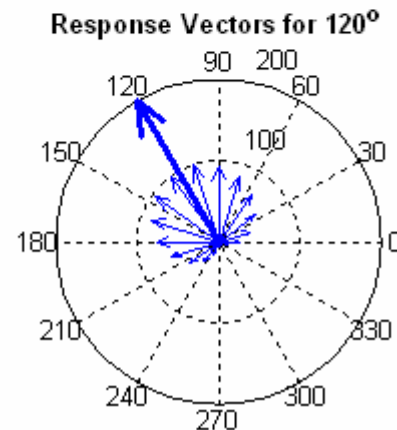
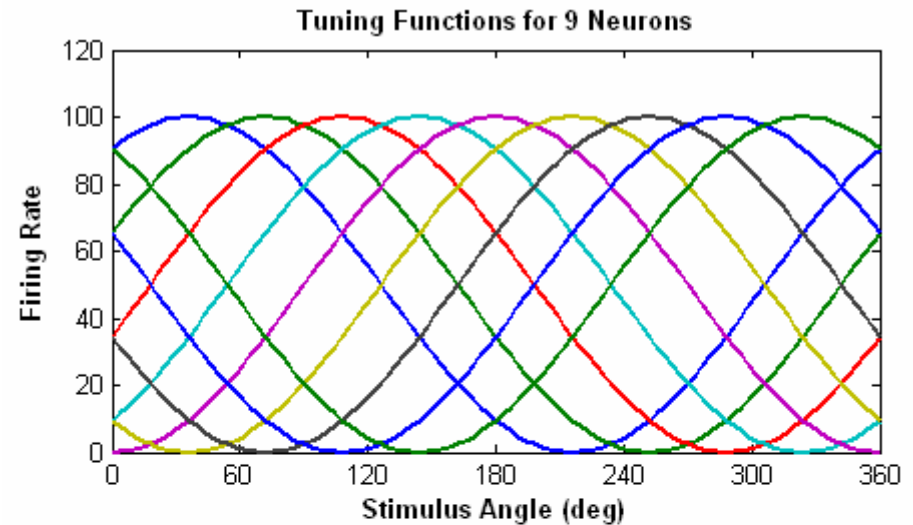
# Rate-Place Code

- Narrow tuning functions
- Neurons are topographically organized to form map of stimulus parameter
- Each stimulus activates a restricted population of neurons
- Stimulus estimated from location of maximum activity along map
- Examples:
  - Retinal map of space
  - Tonotopic map
  - ITD map in barn owl (Jeffress model)?



# A distributed code: The population vector (Georgopoulos, 1986)

- Broad tuning functions around a preferred direction
- Topographic map not necessary
- Each stimulus activates entire neural population to a degree
- Each neuron's response can be seen as a vector, with angle equal to the neuron's best direction and magnitude equal to firing rate
- Vector sum across neural population accurately points to the stimulus direction
- Example: Direction of arm movement in motor cortex.
- Could apply to any angular dimension, e.g. sound azimuth.



# Entropy of a Random Variable

Definition: 
$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x)$$

- Entropy is expressed in *bits* (binary choices)
- For a uniform probability distribution ( $p(x) = 1/N$  for all  $x$ ), the entropy is the logarithm (base 2) of the number  $N$  of possible values of  $X$ . All other distributions have a lower entropy.
- The entropy represents the number of independent binary choices required to uniquely specify the value of a random variable, i.e. it quantifies the “uncertainty” about that variable. It is also the average length of a binary string of 1’s and 0’s required to encode the variable.

# Entropy rates in sounds

For a continuous signal source, the *entropy rate* (in bits/s) is

$$R = \lim_{T \rightarrow \infty} 1/T H(T)$$

- CD-quality audio: 44,100 Hz X 16 bits X 2 channels = 1.4 Mb/s
- MP3: 32-128 kb/s
- Telephone: 30 kb/s
- Vocoder: 2.4-9.6 kb/s
- Nonsense speech: 8 phonemes/sec X  $\log_2(32)$  phonemes = 40 b/s (overestimate because not equiprobable)
- This lecture: 2 bits/3600s = 0.0006 b/s
- What about neuron firings?

# Entropy rate of spike trains

## Spike Times (binomial distribution, McKay & McCulloch 1952)

Figures removed due to copyright reasons.

Please see:

Figures from Rieke F., D. A. Bodnar, W. Bialek. "Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents." *Proc Biol Sci* 262, no. 1365 (December 22, 1995): 259-65. (Courtesy of The Royal Society. Used with permission.)

$$R_T \approx r \log_2 \left( \frac{e}{r \Delta t} \right)$$

$r$  is the average firing rate,  $\Delta t$  the temporal resolution

*Example:*  $H_T = 950$  bits/s (3.2 bits/spike)

for  $r = 300$  spikes/s,  $\Delta t = 1$  ms

## Spike Counts (exponential distribution - optimal)

$$R_R = [\log_2(1 + \bar{n}) + \bar{n} \log_2(1 + 1/\bar{n})] / T$$

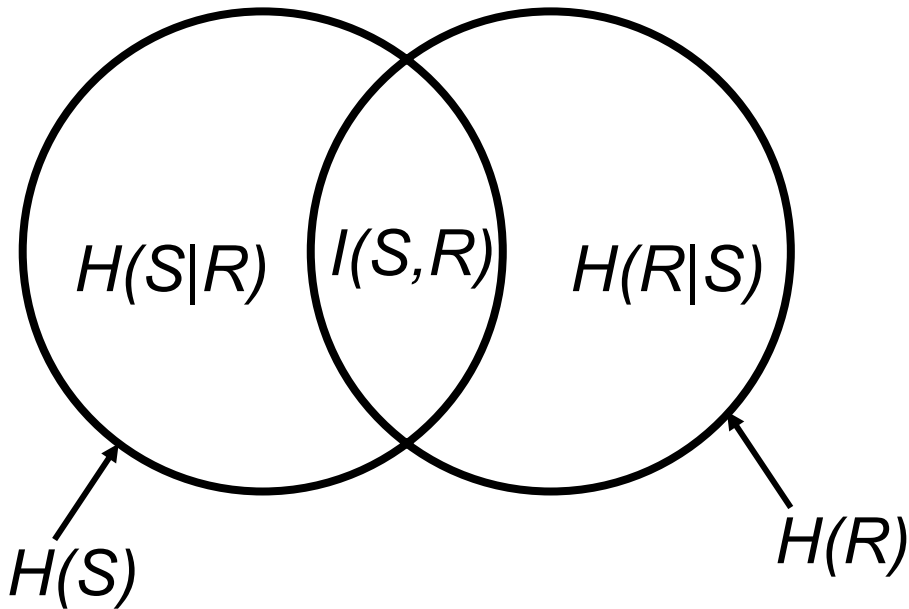
$n = rT$  is the mean spike count,  $T$  is the recording time

*Example:*  $H_R = 26$  bits/s (0.1 bits/spike) for  $r = 300$  spikes/s,  $T = 300$  ms

The entropy rate of a spike train gives an upper bound on the information about the stimulus that can be transmitted by the spike train.



# Conditional Entropy and Mutual Information



Entropy and Information.	
$p(s,r) = p(s r) \cdot p(r)$	Bayes' theorem
$H(S) = -\sum_i p(s_i) \log_2 p(s_i)$	Entropy of S
$H(R,S) = -\sum_i \sum_j p(s_i, r_j) \log_2 p(s_i, r_j)$	Joint entropy of R and S
$H(R S) = -\sum_j p(s_j) \sum_i p(r_i s_j) \log_2 p(r_i s_j)$	Conditional entropy of R given S or neuronal noise
$H(S R) = -\sum_j p(r_j) \sum_i p(s_i r_j) \log_2 p(s_i r_j)$	Conditional entropy of S given R or stimulus equivocation
Equivalent forms for average information:	
$I(R, S) = H(R) - H(R S)$	
$I(R, S) = H(S) - H(S R)$	
$I(R, S) = H(R) + H(S) - H(R, S)$	

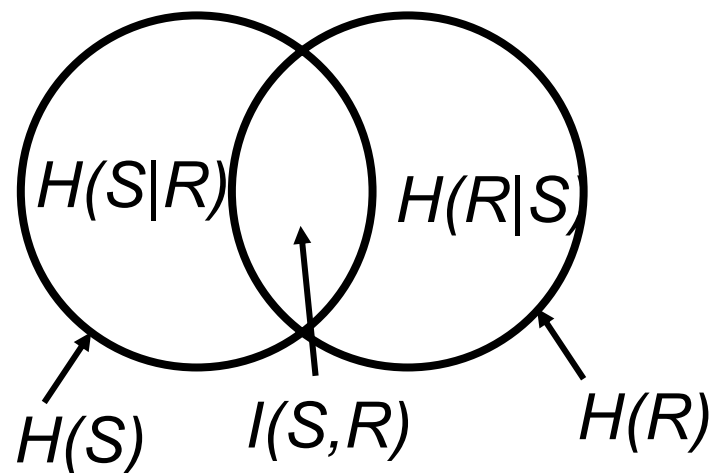
- The entropy  $H(S)$  represents the uncertainty about the stimulus in the absence of any other information
- The *conditional entropy*  $H(S|R)$  represents the remaining stimulus uncertainty after the neural response has been measured
- $I(S,R) = H(S) - H(S|R)$  is the *mutual information* between S and R; it represents the reduction in uncertainty achieved by measuring R
- If S and R are statistically independent, then  $I(S,R)=0$  and  $H(S,R)=H(S)+H(R)$
- By symmetry,  $I(S,R) = I(R,S) = H(R) - H(R|S)$

# Direct method for computing mutual information between stimulus and neural response

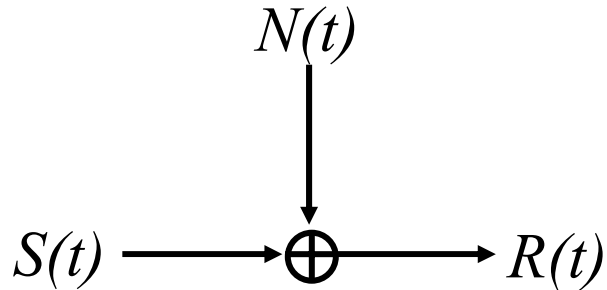
- The most obvious method for measuring  $I(S,R)$  would be to subtract  $H(S|R)$  from  $H(S)$ . But  $H(S|R)$  is hard to measure because it requires estimating the stimulus from neural response.
- Trick: By symmetry,  $I(S,R)$  is also  $H(R) - H(R|S)$ .  $H(R|S)$  is the entropy of the part of the neural response that is NOT predictable from the stimulus, i.e. the noise in the response.

## Method:

- Present stimulus set  $S$  with probability  $P(S)$ . Measure neural response to many presentations of each stimulus.
- Estimate  $P(R)$  for entire stimulus set and compute  $H(R)$  from  $P(R)$ .
- To estimate the noise in the response, average response to all presentations of same stimulus, and subtract average from response to each trial. Compute  $H(R|S)$  from the estimated noise distribution.
- Subtract  $H(R|S)$  from  $H(R)$ .
- Assumptions about noise distribution (e.g. Gaussian) can be made to simplify estimation of probabilities.
- In practice, this method becomes prohibitive in its data requirements for large dimensionality, e.g. when computing information for population of neurons.



# Classic Example (Shannon): Information Rate of Gaussian Channel



Both the stimulus  $S(t)$  and the additive noise  $N(t)$  have Gaussian probability distributions which are fully characterized by the power spectra  $S(f)$  and  $N(f)$ .

$$R_{INFO} = \int_0^W \log_2 \left( 1 + \frac{S(f)}{N(f)} \right) df$$

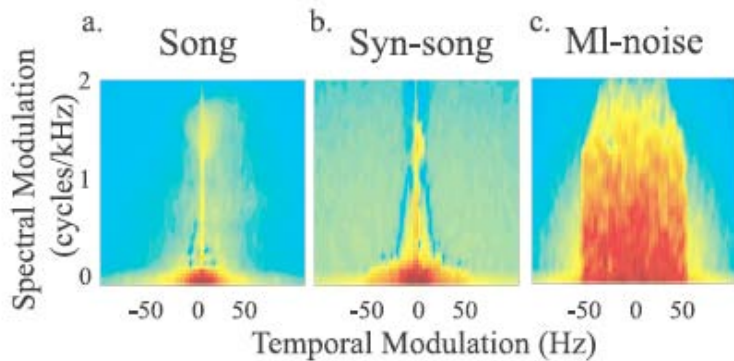
Information rate is entirely determined by available bandwidth  $W$  and signal-to-noise ratio  $S/N$ .

Telephone channel:  $S/N=30$  dB,  
 $W=3000$  Hz  $\Rightarrow I = 30$  kb/s

Figures removed due to copyright reasons.

Please see:  
Borst, A., and F. E. Theunissen.  
“Information theory and neural coding.” Figure 2 in *Nat Neurosci* 2, no. 11 (Nov 1999): 947-57.

# Natural and synthetic songs are coded more efficiently than noise in zebra finch auditory neurons

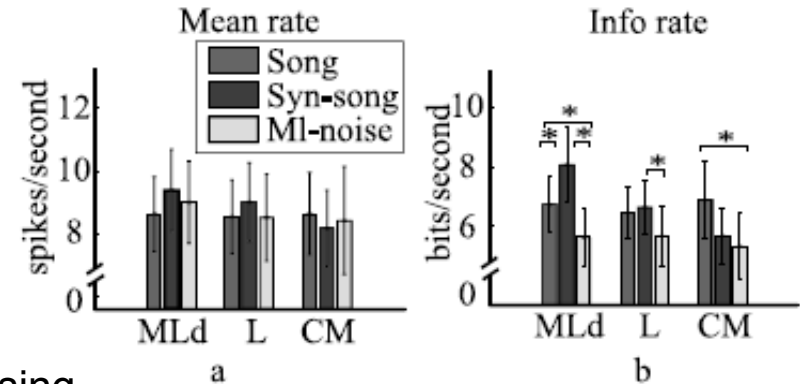
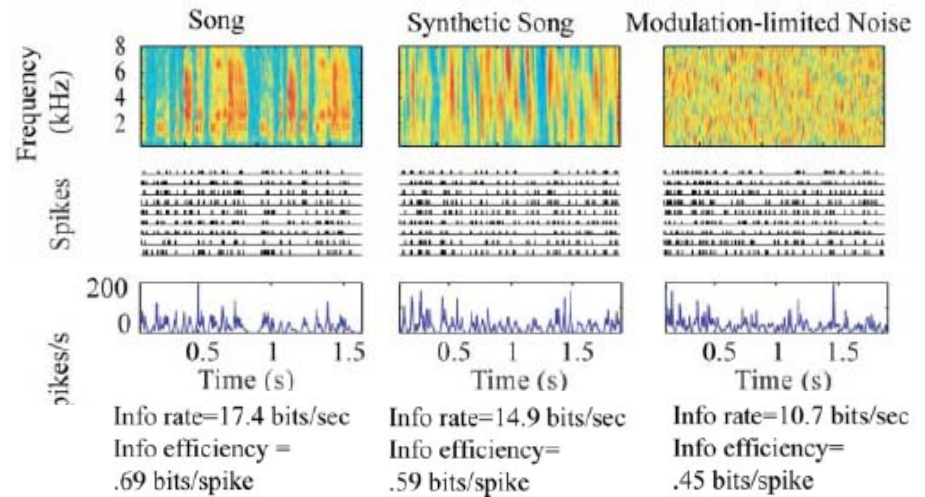


## 3 types of stimuli:

- Song: 20 natural songs from male zebra finch
- Syn-song: Mimics spectral and temporal modulations in song
- ML-noise: Noise with limited range of modulations

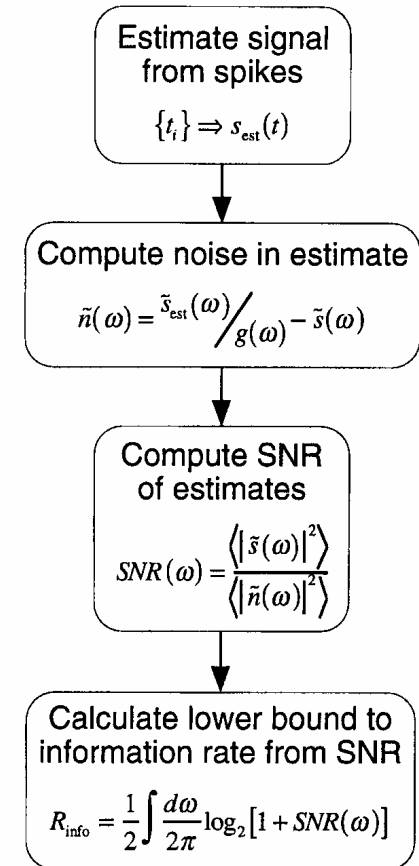
## Neural recording from 3 locations:

- MLd: Auditory midbrain
- Field L: Primary cortex
- CM: Secondary cortex involved in song processing



# Using the stimulus reconstruction method to estimate stimulus information in spike train

- Computing  $I(S,R)$  directly from the definition often requires too much data.
- Data processing theorem: If  $Z=f(R)$ , then  $I(S,Z) \leq I(S,R)$ .  
 “Data processing cannot increase information”.
  - Special case: If  $\hat{S}$  is an estimate of  $S$  based on neural response  $R$ , then  $I(S,\hat{S}) \leq I(S,R)$
- Method:
  - Use Gaussian stimulus  $S(t)$  and compute linear estimate  $\hat{S}(t)$  from neural response.
  - Define noise as  $N(t)=S(t)-\hat{S}(t)$ , assumed to be Gaussian.
  - Compute power spectra of  $N(t)$  and  $\hat{S}(t)$ .
  - Use formula for information rate of Gaussian channel to estimate  $R_{INFO}$ .
- The method is meant to give a lower bound on the information rate. The better the stimulus estimate and the Gaussian assumption, the more accurate the bound.



From:

Rieke, Fred, David Warland, Rob de Ruyter vanSteveninck, and William Bialek. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press/Bradford Books, 1997 (c).  
 Used with permission.

# Linear reconstruction of a stimulus from the spike train (Bialek et al., 1991)

- Given a stimulus waveform  $s(t)$  and a spike train  $r(t)$ , what is the linear filter  $h(t)$  which gives the least-squares estimate of  $s(t)$  from  $r(t)$ ?

$$\hat{s}(t) = h(t) * r(t) = \sum_i h(t - t_i)$$

Find  $h(t)$  to minimize  $E = \int_0^T (s(t) - \hat{s}(t))^2 dt$

- The solution is given (in the frequency domain) by the *Wiener filter*:

$$H(f) = \frac{S_{rs}(f)}{S_r(f)}$$

- $H(f)$  is the Fourier transform of  $h(t)$ , *i.e.* the filter frequency response,  $S_r(f)$  is the power spectrum of the spike train, and  $S_{rs}(f)$  is the cross-spectrum between  $r(t)$  and  $s(t)$

# Example of linear stimulus reconstruction from a spike train

Figures removed due to copyright reasons.

Please see:

Borst, A., and F. E. Theunissen. "Information theory and neural coding." Figure 4 in *Nat Neurosci* 2, no. 11 (Nov 1999): 947-57.

- Fly visual system ("H1 neuron"), moving grating stimulus
- Reconstruction resembles lowpass filtered stimulus

# Bullfrog advertisement call

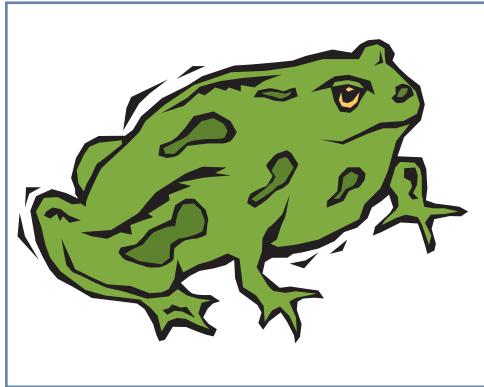
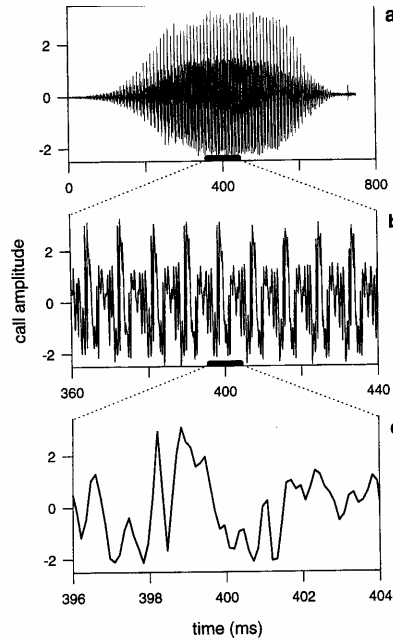
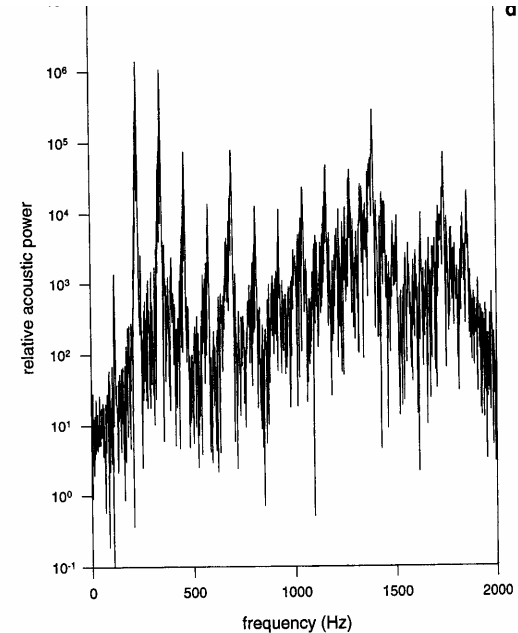


Figure by MIT OCW.

## Waveform



## Power Spectrum



Male bullfrogs produce an *advertisement call* to convey information about location and breeding readiness to both sexes.

From:

Rieke, Fred, David Warland, Rob de Ruyter vanSteveninck, and William Bialek. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press/Bradford Books, 1997 (c). Used with permission.



# Information rates in bullfrog primary auditory fibers are higher for call than for noise

Figures removed due to copyright reasons.

Please see:

Figures from Rieke F., D. A. Bodnar, W. Bialek. "Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents." *Proc Biol Sci* 262, no. 1365 (December 22, 1995): 259-65. (Courtesy of The Royal Society. Used with permission.)

- Linear stimulus reconstruction method was used to estimate (a lower bound on) information rate for both broadband Gaussian noise and call-shaped noise.
- Information rates are higher for call-shaped noise (40-250 bits/s) than for broadband noise (10-50 bits/s).

# Coding efficiency in bullfrog primary auditory fibers

$$\text{Coding Efficiency} = \frac{\text{Information Rate}}{\text{Entropy Rate in Spike Train}}$$

- The coding efficiency is a measure of the fraction of the information contained in the spike train that is used to code the stimulus.
- Coding efficiency ranged from 0.05 to 0.2 for the broadband stimulus, 0.35 to 0.9 for the call-shaped noise.
- These high coding efficiencies were achieved despite the poor quality of the stimulus reconstructions.

Figures removed due to copyright reasons.

Please see:

Figures from Rieke F., D. A. Bodnar, W. Bialek. "Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents." *Proc Biol Sci* 262, no. 1365

(December 22, 1995): 259-65. (Courtesy of The Royal Society.

Used with permission.)

# Information rates and coding efficiencies in sensory neurons

Neural System and Species	Information Rate (bits/spike)	Information Rate (bits/s)	Coding Efficiency	Best Timing Precision and Variance : Mean
<b>Constant stimulus</b>				
Cat retinal ganglion cells	~0.04-0.10	0.4-0.8	-	-
Primary visual cortex (VI) of rhesus monkeys	-	0.62	-	-
Middle temporal area (MT) of rhesus monkeys	0.025	0.89 ± 0.29	-	~1.3
Inferior temporal area (IT) of rhesus monkeys	~0.13	0.9	-	-
Hippocampus of rhesus monkeys	~0.18	0.32 (max = 1.2)	-	-
<b>Variable stimulus: reconstruction method</b>				
HI motion-sensitive neuron of a fly	0.75	64	30%	-
Frog auditory afferents	0.66	23	11%	~2 ms
Vibratory receptors of the bullfrog sacculus	2.6	155	50-60%	-
Cricket mechanoreceptors	0.6-3.2	75-294	50-60%	~0.4 ms
Salamander retinal ganglion cells	1.9	3.7 (up to 10 for a population of >10 cells)	26% (>79% for >10 cells)	-
The MT of anesthetized rhesus monkeys	~0.65	6.7 (max = 12.3)	-	2-4 ms
The MT of alert rhesus monkeys	0.6	5.5	<30%	-
<b>Variable stimulus: direct method</b>				
HI motion-sensitive neuron of a fly	2.43	80	50%	1.5-3 ms; <0.1
Salamander and rabbit retinal ganglion cells	3.7	16.3	59%	≥ 0.70 ms; ≥ 0.05
The MT of alert rhesus monkeys	1.5	12 (max = 29)	Up to 45%	<2 ms; ~1.4

Table by MIT OCW.

# Information Theory Pro's and Con's

## Pros

- Does not assume any particular neural code
- Can be used to identify the stimulus features best encoded by neurons or to compare effectiveness of different putative neural codes
- One number summarizes how well stimulus set is coded in neural response

## Cons

- Information estimate depends on stimulus set. Stimulus probabilities in environment hard to specify.
- Does not specify how to read out the code: the code might be unreadable by the rest of the nervous system.
- For all but simplest examples, estimation of mutual information requires huge amount of data. Methods that try to circumvent data limitations (e.g. stimulus reconstruction) make additional assumptions (e.g. linearity or Gaussian distribution) that are not always valid.

Slide 25:

Borst, A., and F. E. Theunissen. "Information theory and neural coding." *Nat Neurosci* 2, no. 11 (Nov 1999): 947-57.

Slide 35:

Buracas, G. T., and T. D. Albright. "Gauging sensory representations in the brain." *Trends Neurosci* 22, no. 7 (Jul 1999): 303-9.