

# Neural Networks for Scene Analysis

---

**DeLiang Wang**

*The Ohio State University*

# Outline of Tutorial

---

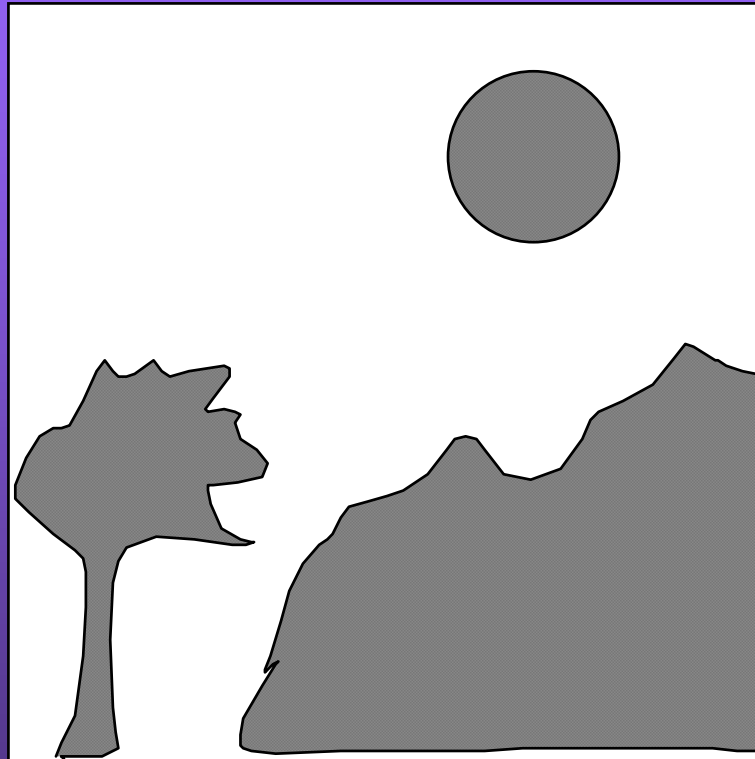
- **Introduction**
  - Scene analysis problem
  - Traditional neural network methods
- **Oscillatory Correlation Theory**
  - LEGION network
- **Visual Scene Analysis**
- **Auditory Scene Analysis**
- **Summary**

# Scene Analysis Problem



# Simplified Scenario

---



# Boltzmann Machine Approach

---

- **Figure-ground segregation**
  - Binary figure units and edge units
  - Local, symmetric, fixed connections with both excitatory and inhibitory connections
- **Two kinds of input**
  - Bottom-up: location and orientation of line elements
  - Top-down: visual attention to provide bias
- **Simulated annealing results in selection of a figural object and its boundary**
- **Tested on small, synthetic images**
- **See Hinton & Sejnowski (1987)**

# Feature-Boundary-Feature (FBF) Model

---

- **Consists of two subsystems**
  - Feature contour system detects local features and performs diffusion within a region
  - Boundary contour system detects local edges and performs contour completion
- **Labeling by a fill-in process**
  - It spreads a region label until reaching a boundary
  - Problematic with either too few or too many labels
- **Tested on simple images**
- **See Grossberg & Wyse (1991)**

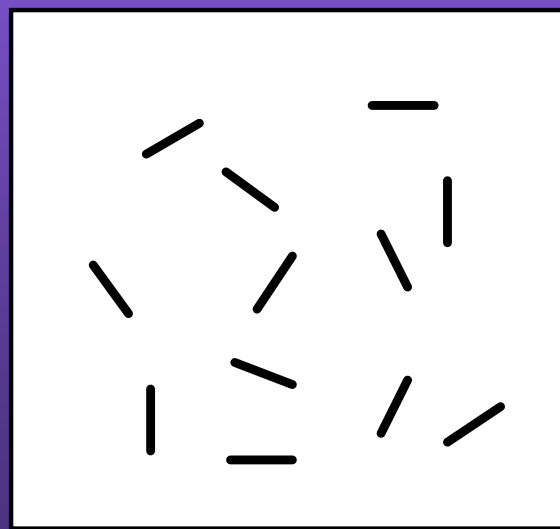
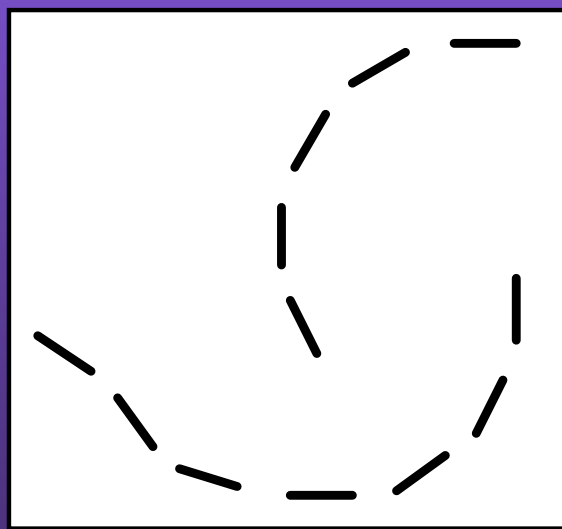
# Classification-based Approach

---

- **Segmentation viewed as classification**
  - Training followed by classification
  - Classification is labeling
- **Various NN training algorithms can be used**
  - Self-organizing maps (Koh et al'95; Alirezaie et al'97)
  - MLP (Alirezaie et al'97)
- **Tested on real images with good results**

# Limitations of Classification-based Approach

- Classification is based on local information only while scene segmentation requires image context
- Example: same set of line elements arranged in different ways produces different organizations





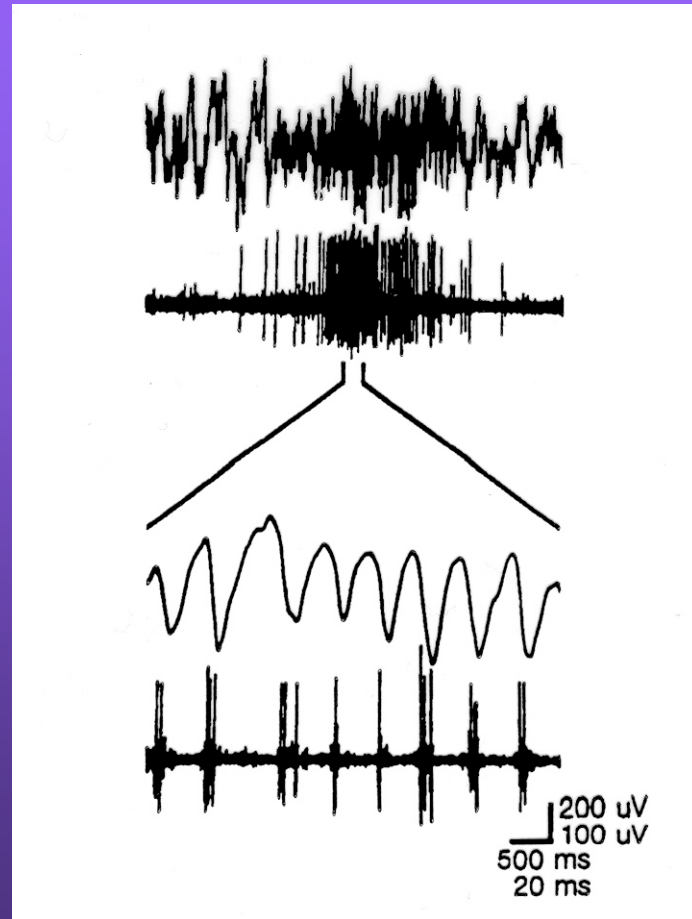
# Temporal Correlation Theory

---

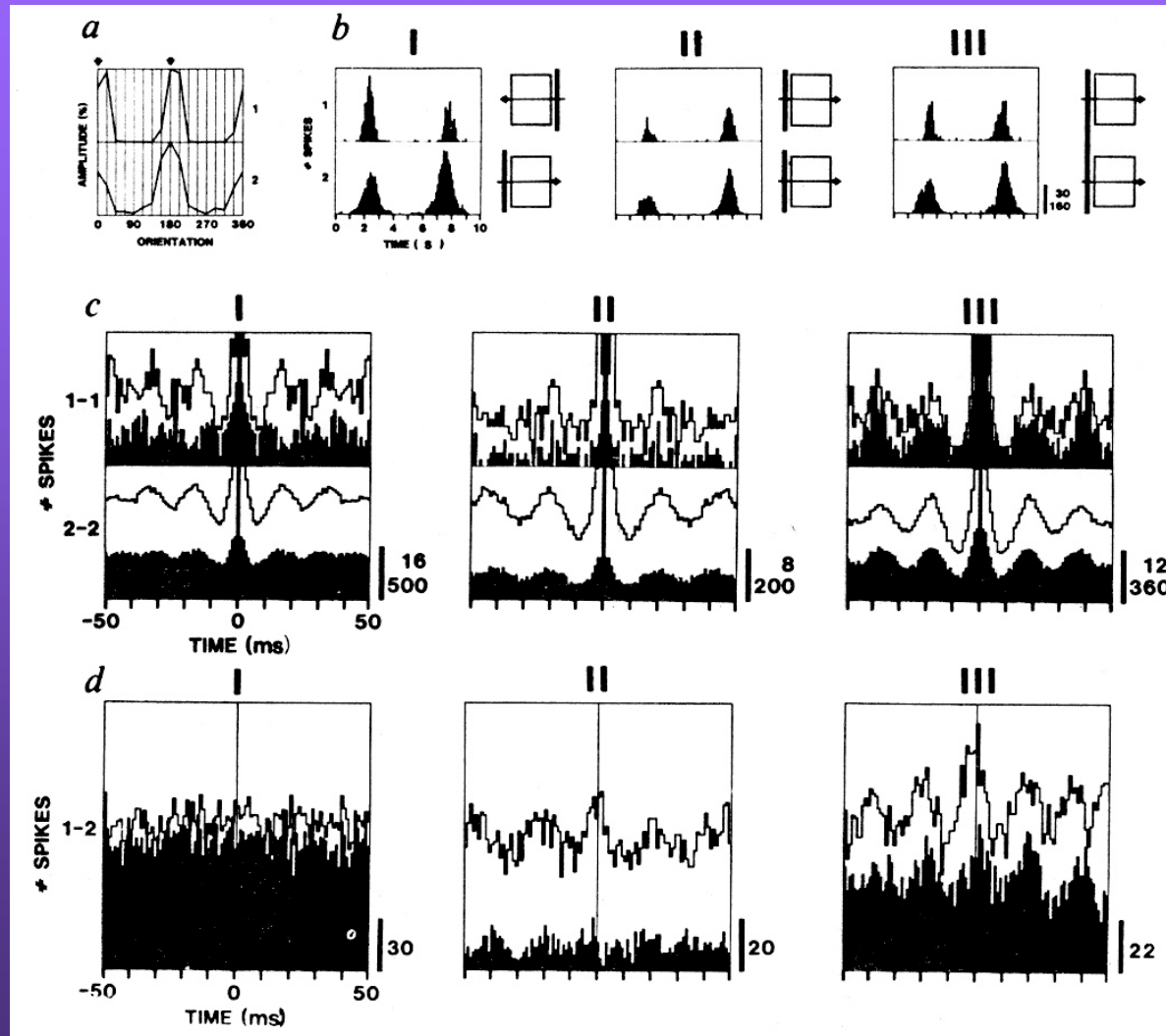
- **Feature binding is a fundamental problemn**
  - In neuroscience
  - In perception
- **Temporal correlation as a representation**
  - Extra degree of freedom
  - A plausible mechanism
- **See von der Malsburg'81, Milner'74, Abeles'82**

# Neurophysiological Evidence

- Gray & Singer (1989)

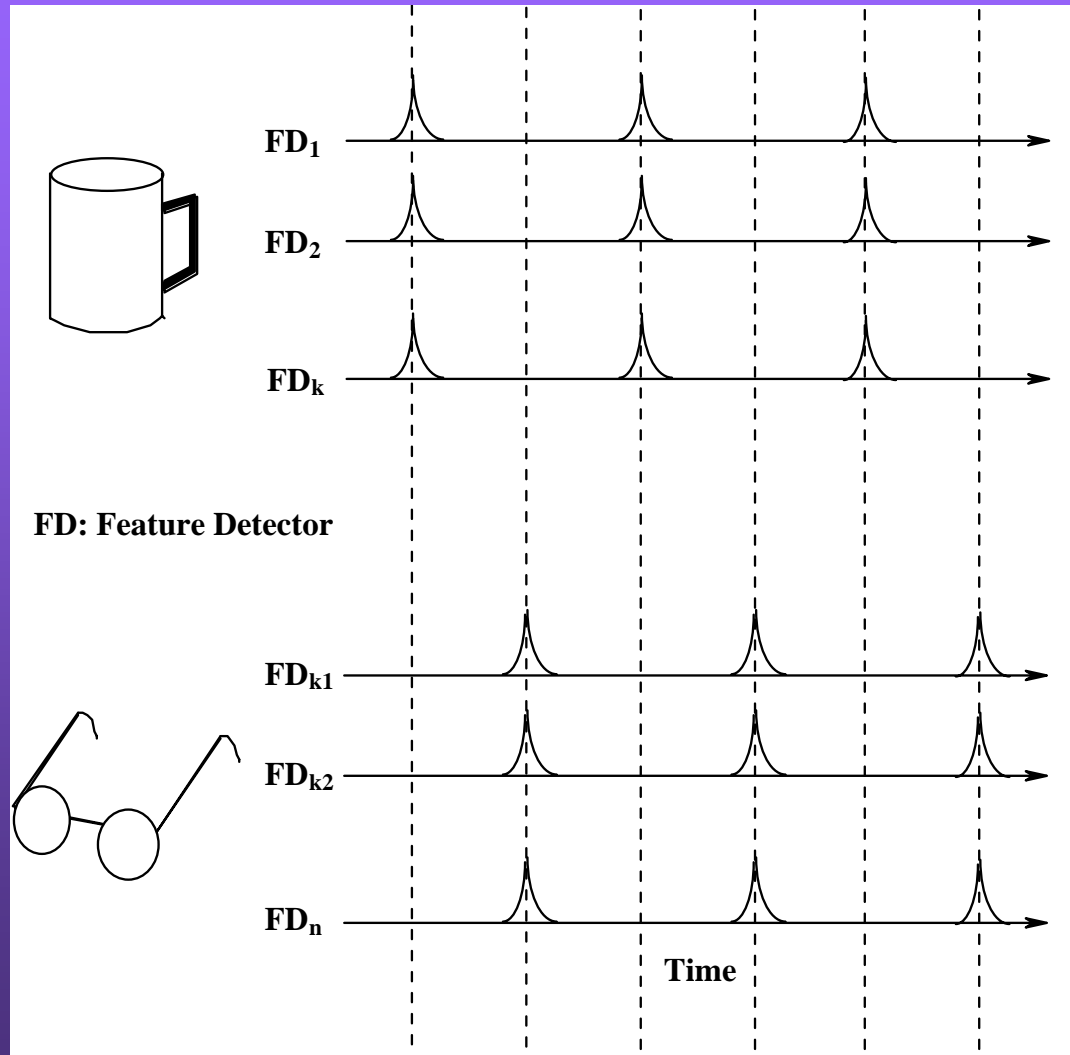


# Neurophysiological Evidence - continued



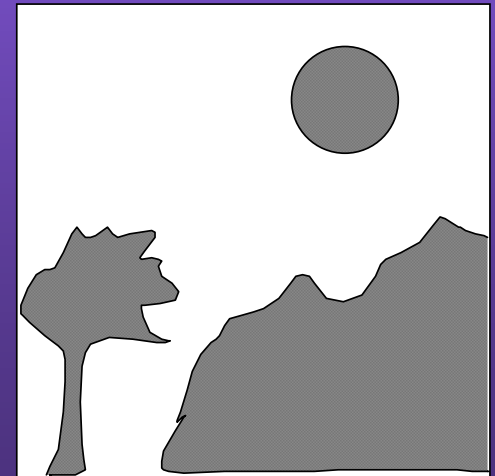
- Gray et al. (1989) WCCI'02 Tutorial (Wang)

# Oscillatory Correlation Theory



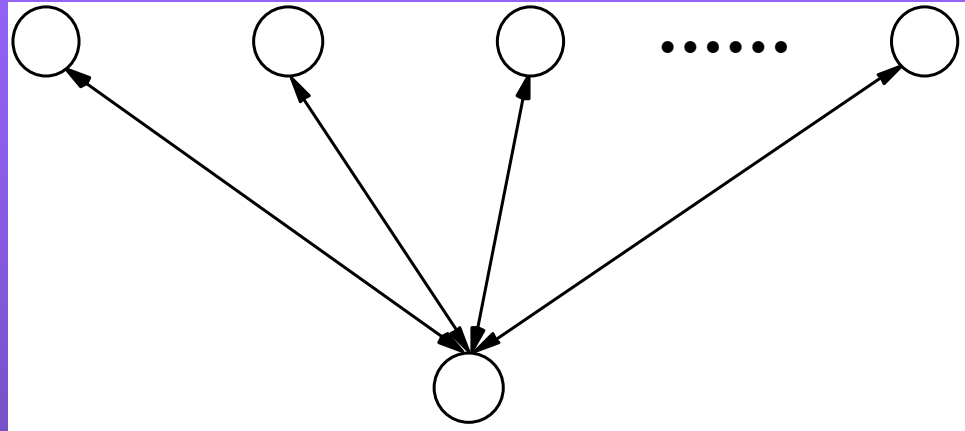
# Computational Requirements for Oscillatory Correlation

- **Need to synchronize locally coupled oscillator population**
  - Extensive literature in theoretical physics and applied mathematics
- **Need to desynchronize different populations, when facing multiple objects**
- **Critically, the above functions *must* be achieved very rapidly**

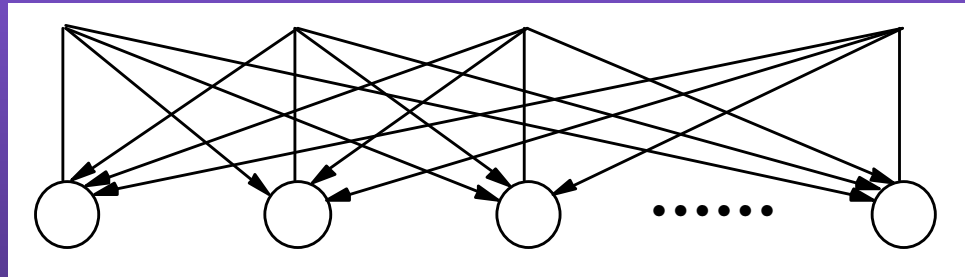


# Three Possible Ways to Reach Synchrony

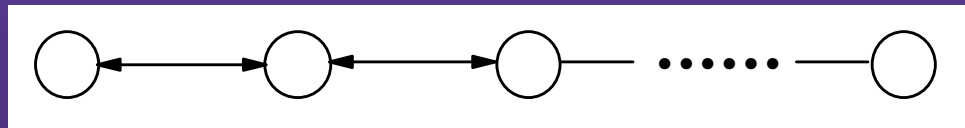
- Comparator model



- Globally connected

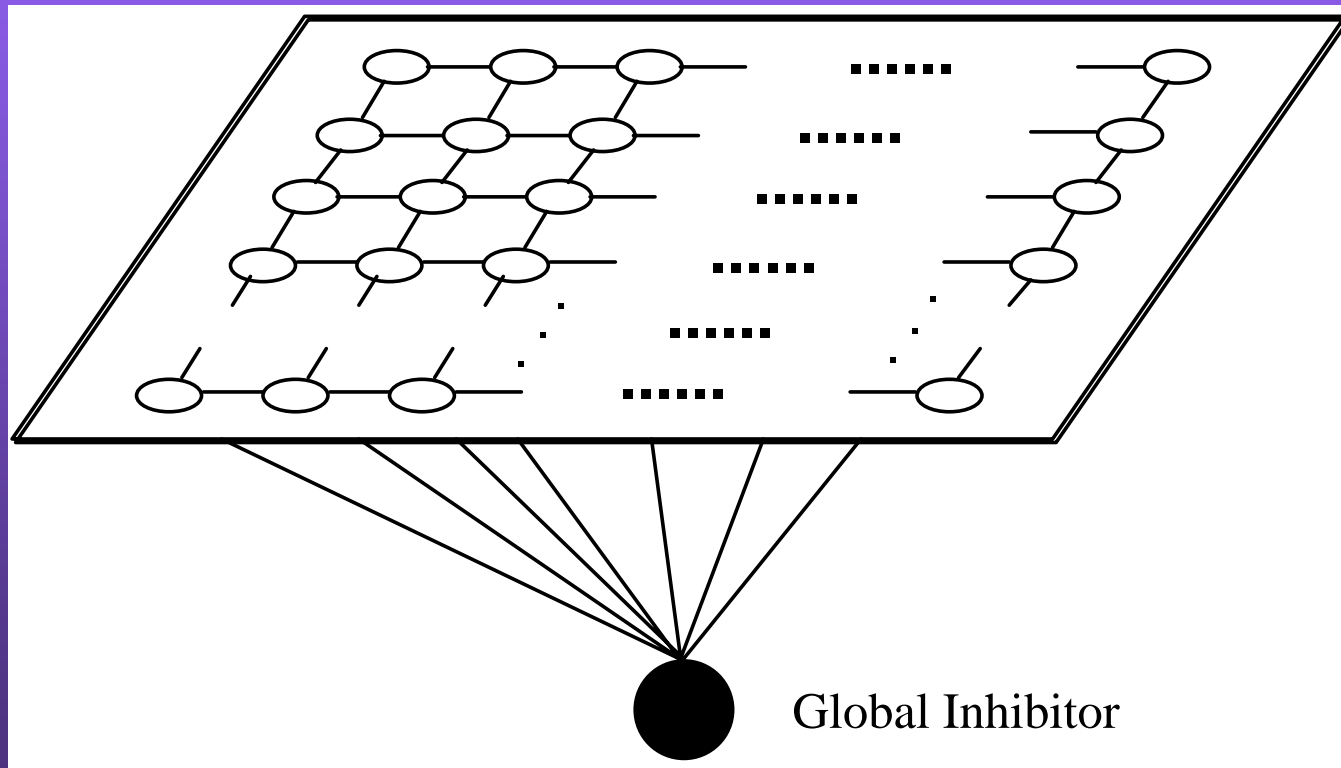


- Locally connected



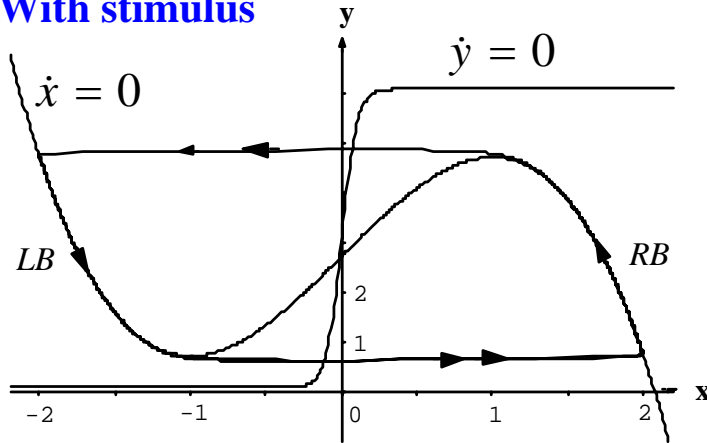
# LEGION Architecture

- **LEGION** - **L**ocally **E**xcitatory **G**lobally **I**nhibitory **O**scillator **N**etwork (Terman & Wang'95)

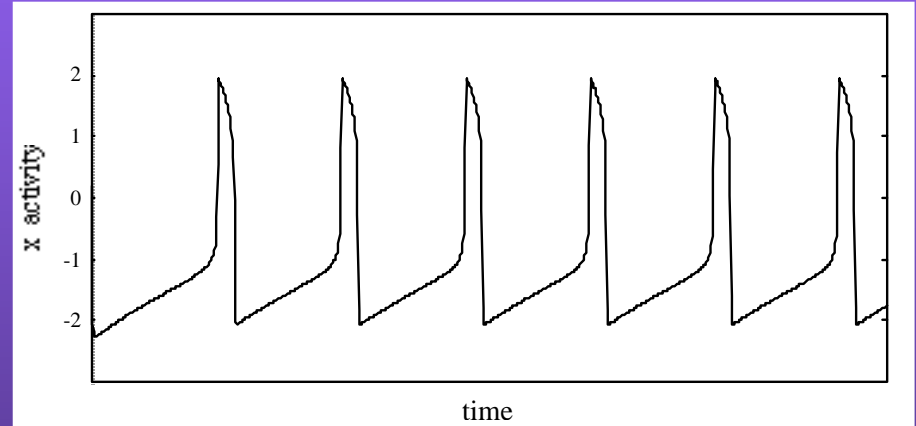
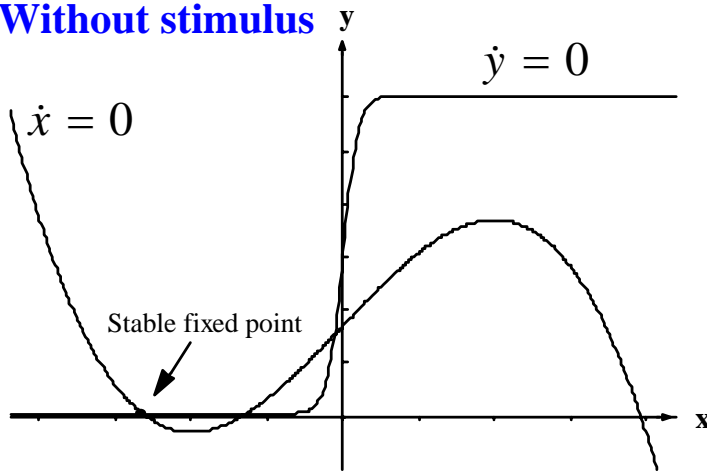


# Single Relaxation Oscillator

**With stimulus**



**Without stimulus**



**Typical  $x$  trace (membrane potential)**



# Model of a Relaxation Oscillator

---

- **Model definition**

$$\begin{aligned}\dot{x}_i &= 3x_i - x_i^3 + 2 - y_i + I_i + S_i + \rho \\ \dot{y}_i &= \varepsilon(\alpha(1 + \tanh(x_i / \beta)) - y_i)\end{aligned}$$

- **Coupling between oscillators**

$$\begin{aligned}S_i &= \sum_{k \in N(i)} W_{ik} S_{\infty}(x_k, \theta_x) - W_z S_{\infty}(z, \theta_z) \\ S_{\infty}(v, \theta) &= \frac{1}{1 + \exp[-K(v - \theta)]}\end{aligned}$$

- Where  $N(i)$  is the set of neighboring oscillators that connect to oscillator  $i$

# Model of a Relaxation Oscillator - continued

---

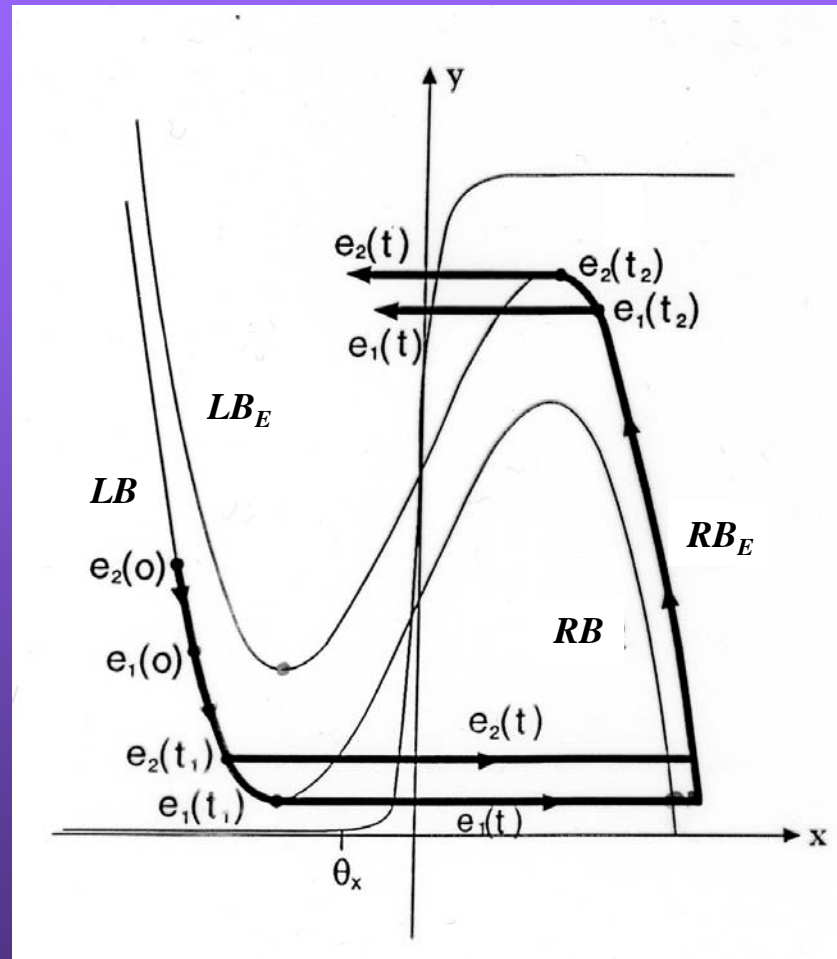
- **Global inhibitor**

$$\dot{z} = \phi(\sigma_{\infty} - z)$$

where  $\sigma_{\infty} = 0$  if  $x_i < \theta_z$  for every oscillator,  
and  $\sigma_{\infty} = 1$  otherwise

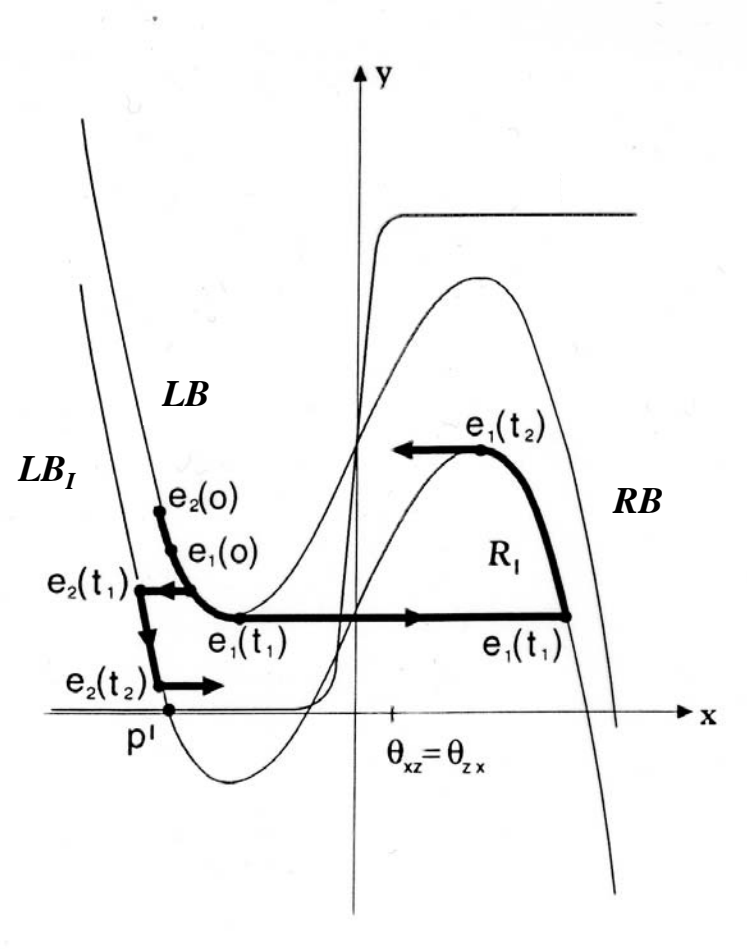
- **Dynamic normalization of incoming weights to an oscillator (Wang'95)**

# Fast Threshold Modulation for Two Oscillators



Somers & Kopell'93

WCCI'02 Tutorial (Wang)



# Summary of Analytical Results

---

- **Definitions:** A *pattern* is a connected region, and a *block* a subset of oscillators stimulated by a given pattern. The following results are established for  $\varepsilon > 0$  sufficiently small
- **Theorem 1. (*Synchronization*).** The parameters of the system can be chosen so that all of the oscillators in a block synchronize. Moreover, the rate of synchronization is exponential

## Summary of Analytical Results - continued

---

- **Theorem 2.** (*Multiple patterns*) If at the beginning all the oscillators of the same block synchronize with each other and the distance between any two oscillators belonging to two different blocks is greater than some constant, then:
  - Synchronization within each block is maintained
  - The ordering of the activation among different blocks is fixed
  - Given a certain period of time, at least one block is in its active phase
  - At most one block is in its active phase at any time

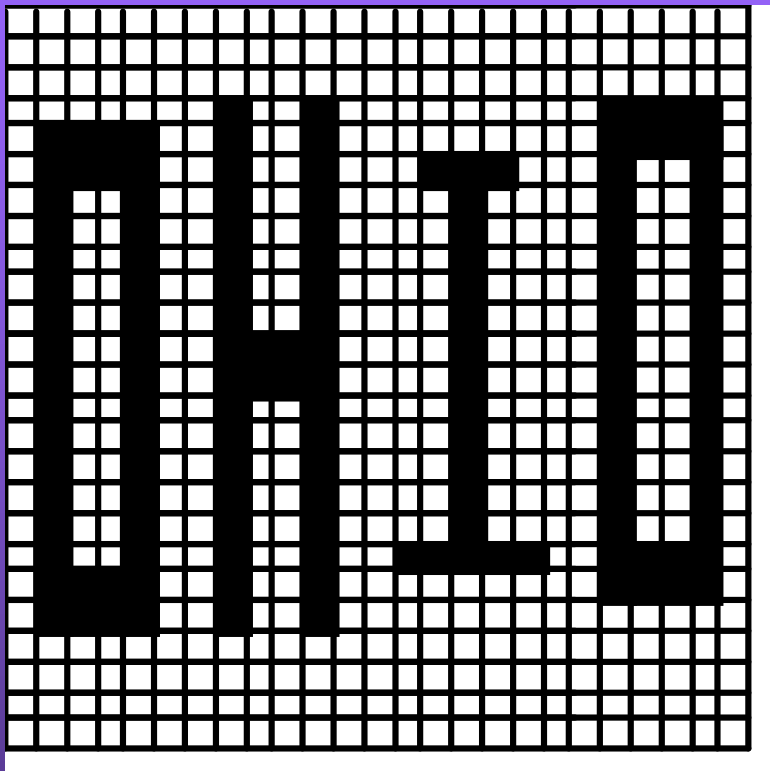
## Summary of Analytical Results - continued

---

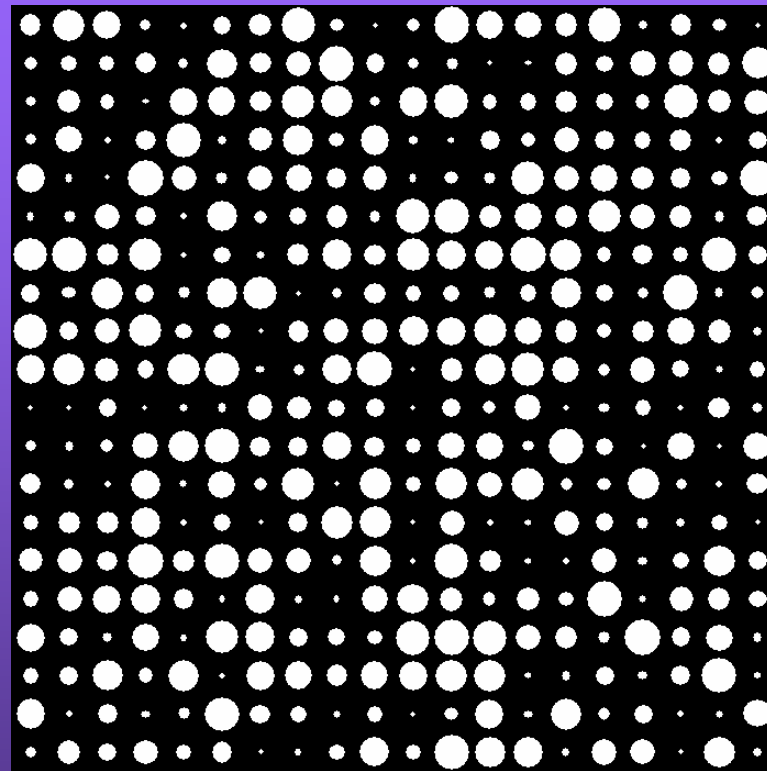
- **Theorem 3.** (*Desynchronization*) If at the beginning all the oscillators of the system lie not too far away from each other, then the condition of Theorem 2 will be satisfied after some time. Moreover, the time it takes to satisfy the condition is no greater than  $N$  cycles, where  $N$  is the number of patterns.
- The entire mechanism is called *Selective Gating* (Terman & Wang'95)

# LEGION Example: Demo

---

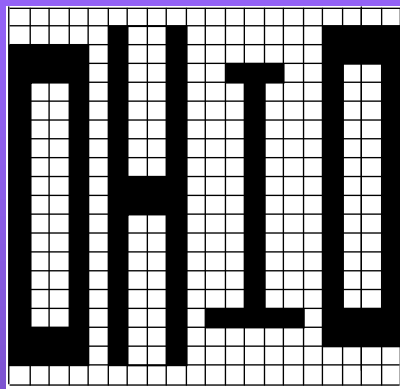


Input image

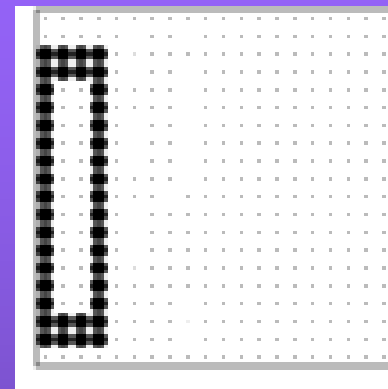
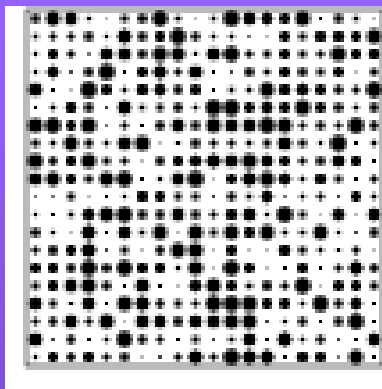




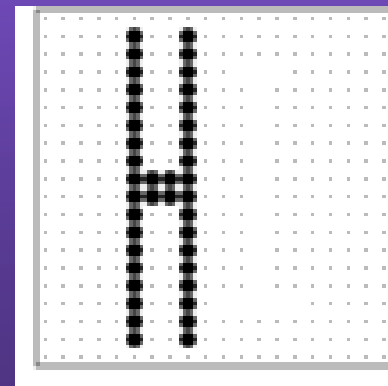
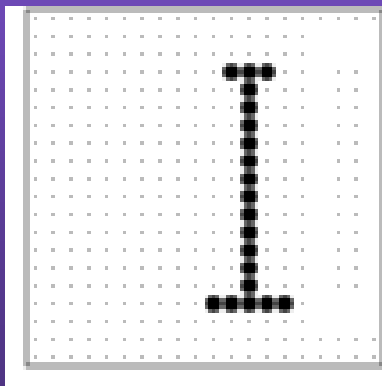
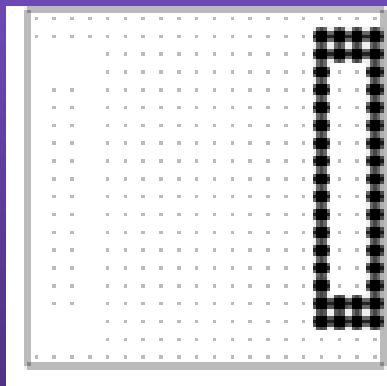
# LEGION Example



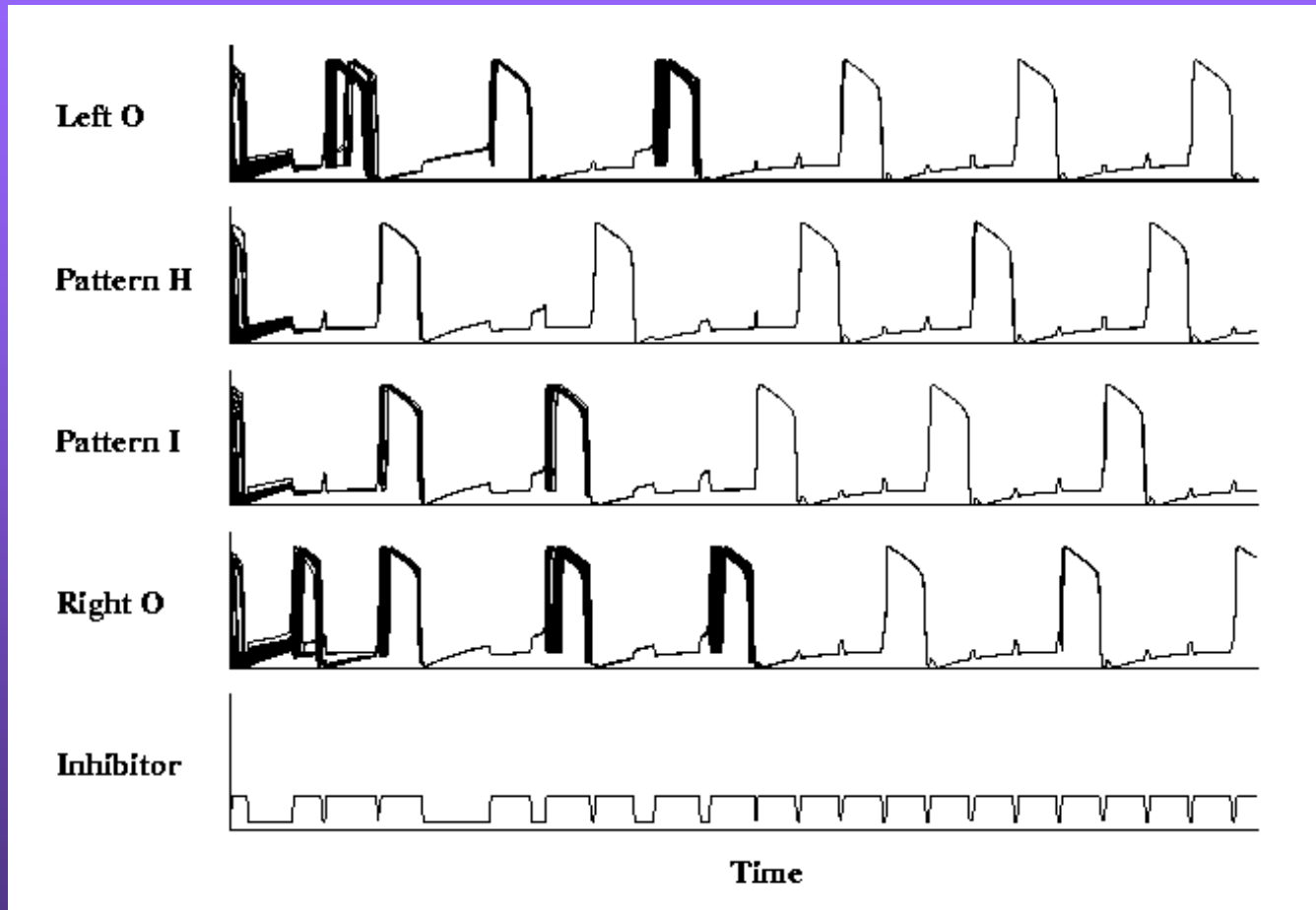
Input image



Successive snapshots

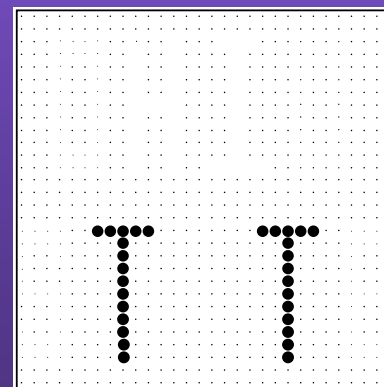
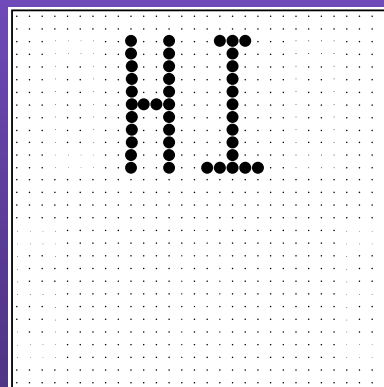
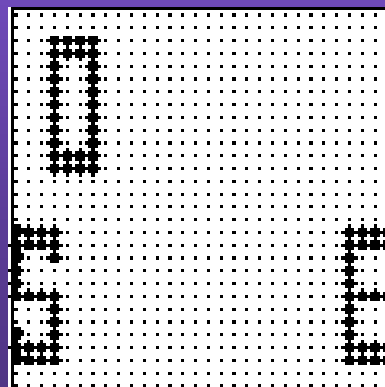
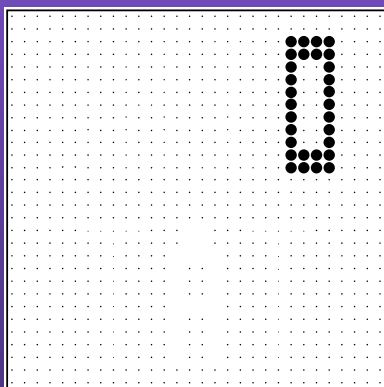
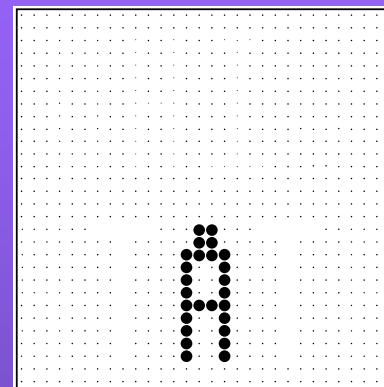
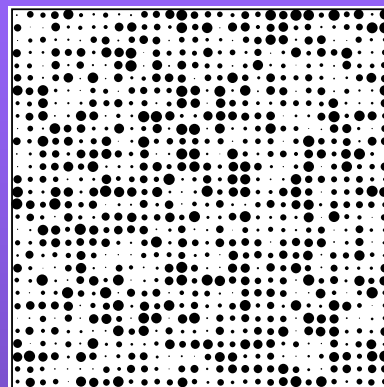
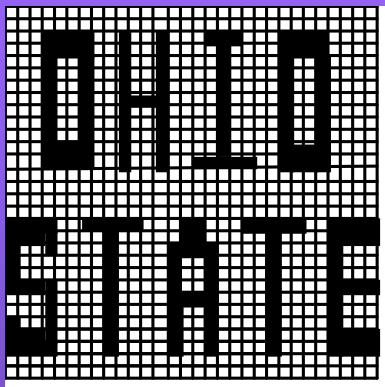


# LEGION Example - Temporal Traces

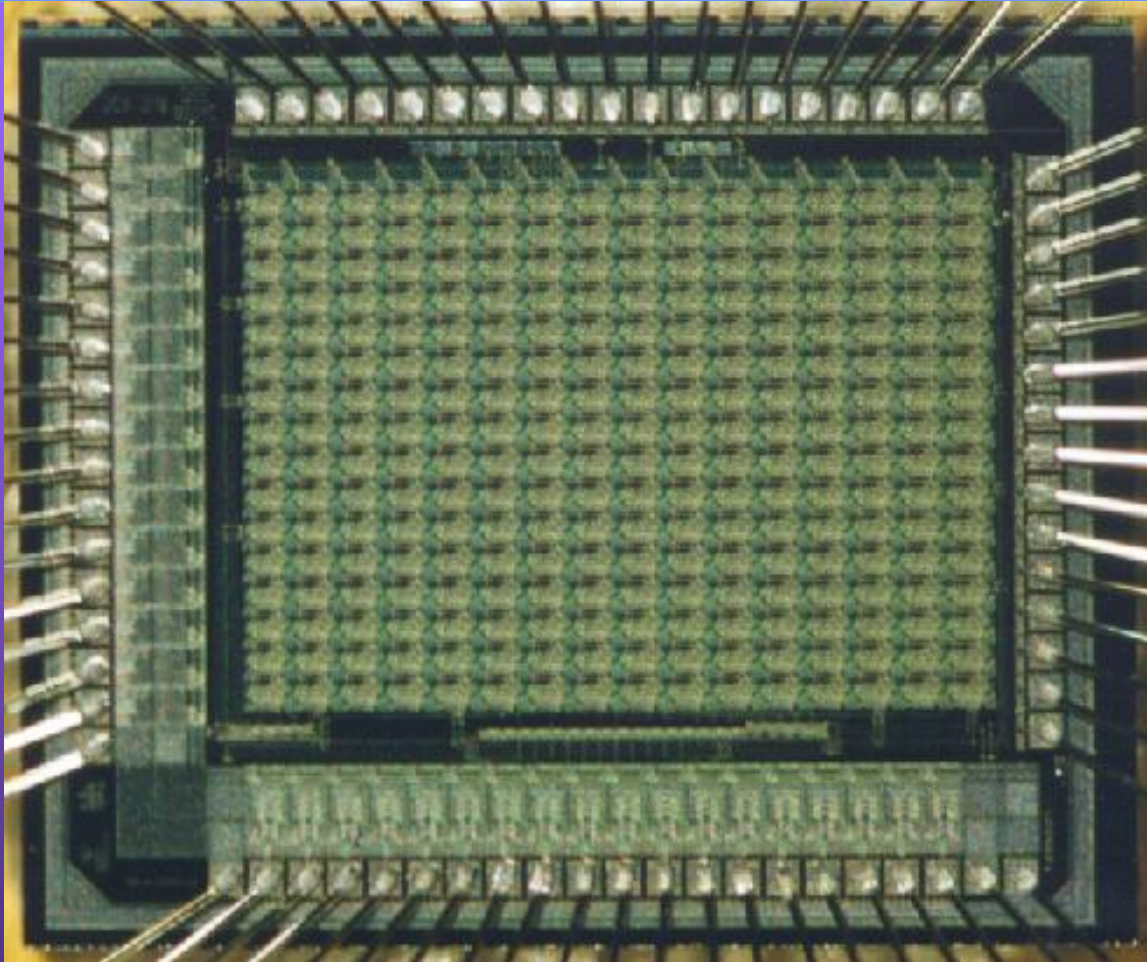


# LEGION Example: Segmentation Capacity

Input  
image



# LEGION on a Chip



The chip area is  $6.7 \text{ mm}^2$  (Core  $3 \text{ mm}^2$ ) and implements a  $16 \times 16$  LEGION network (Cosp'00)

# Oscillatory Correlation Approach to Image Segmentation

---

- **Feature extraction first takes place**
  - An image feature can be pixel intensity, depth, local image patch, texture element, optic flow, etc.
- **Connection weights between two neighboring oscillators are set to be proportional to feature similarity**
- **Global inhibitor controls granularity of segmentation**
  - Larger inhibition results in more and smaller regions
- **Segments pop out from LEGION in time**

# Image Segmentation Example

---



Input image



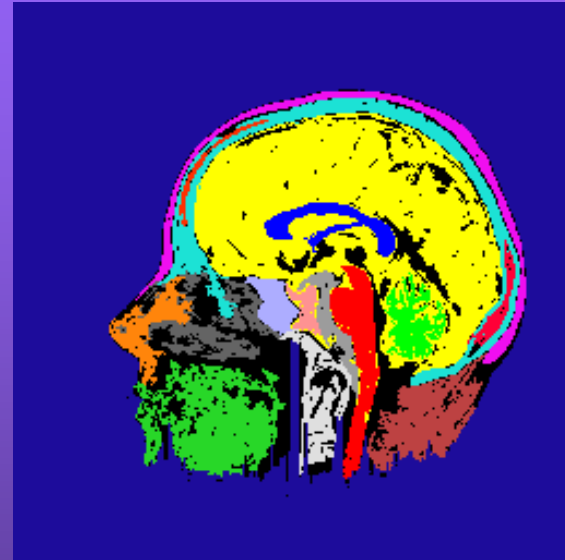
Segmentation result

# Image Segmentation Example

---

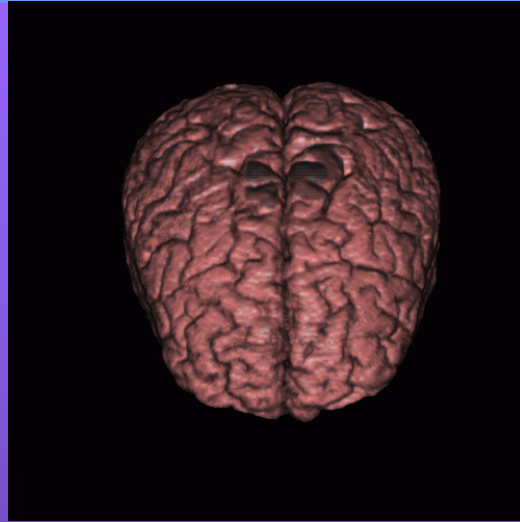
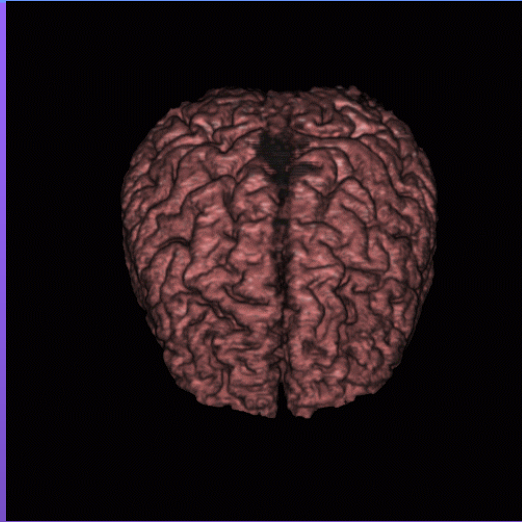


Input image



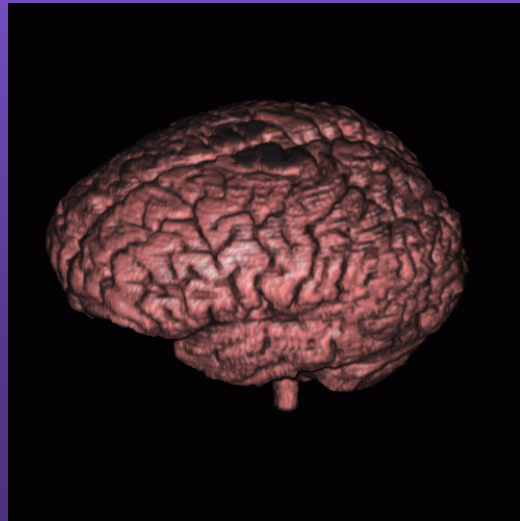
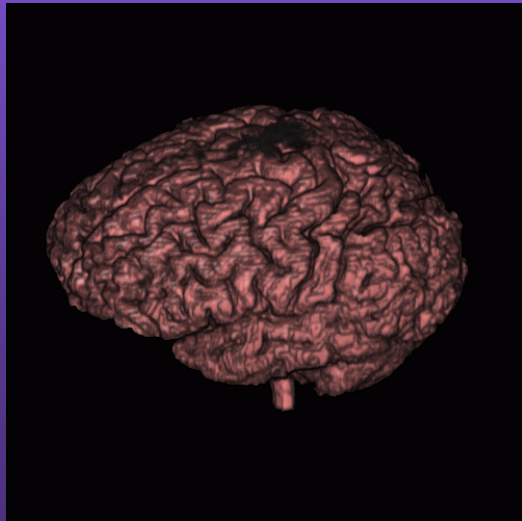
Segmentation result

# 3D MRI Image Segmentation



**Left:** LEGION  
results

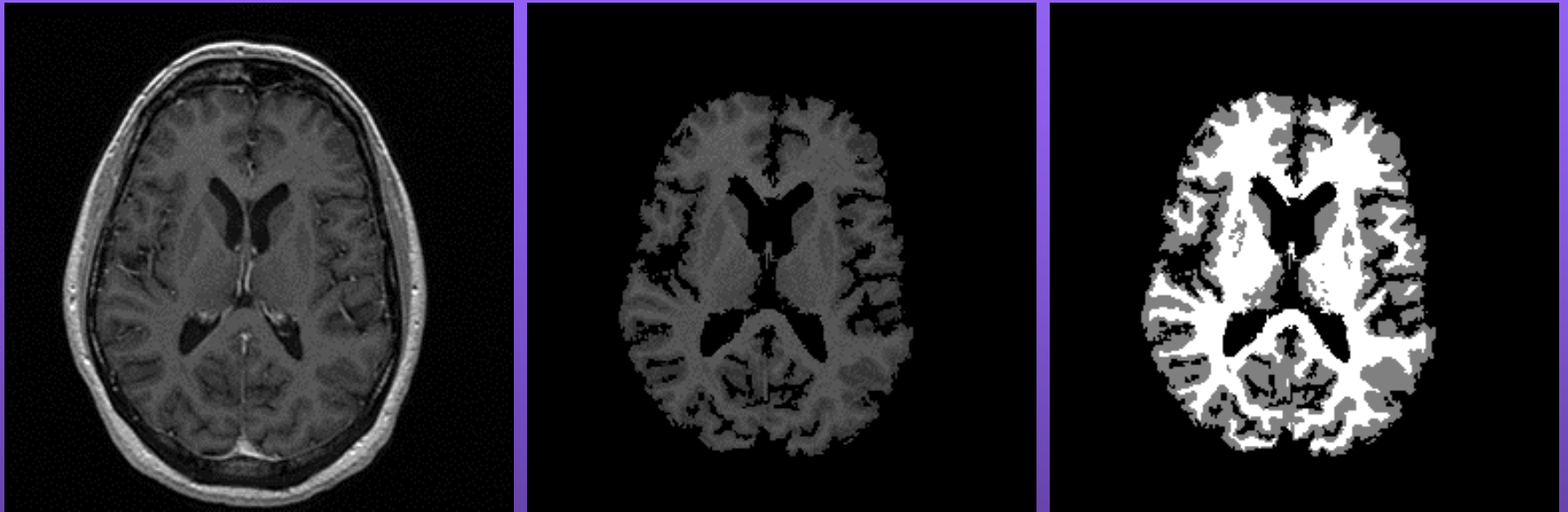
**Right:** Manual  
results





# 3D MRI Image Segmentation: a 2D View

---

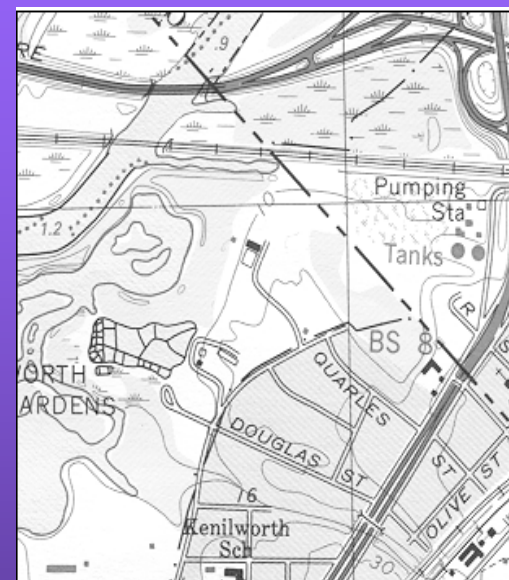


**Left:** Input

**Middle:** Segmentation results

**Right:** Further segmentation into white and gray matter

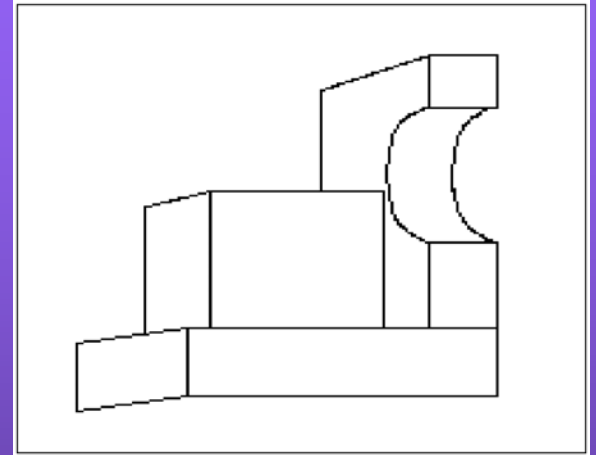
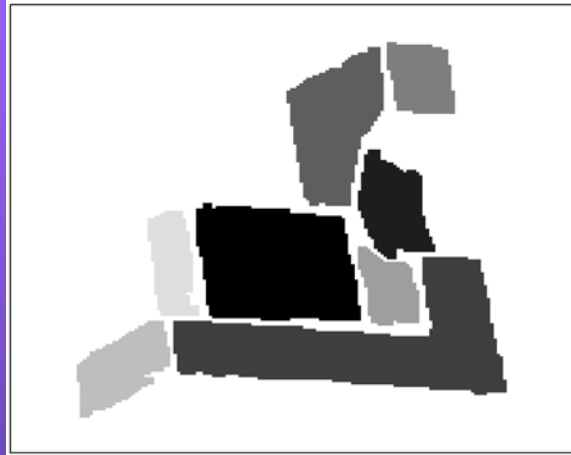
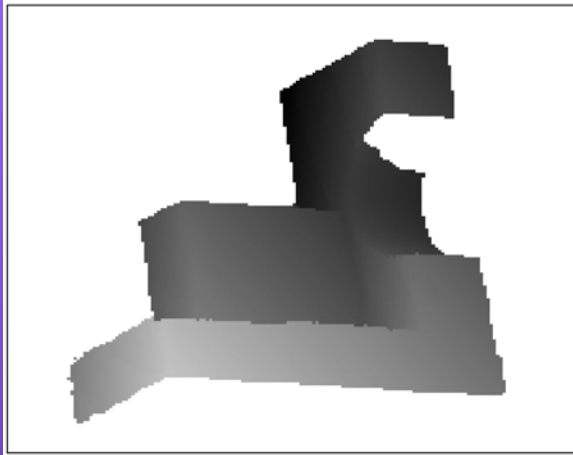
# Aerial Image Analysis



## Extraction of hydrographic objects

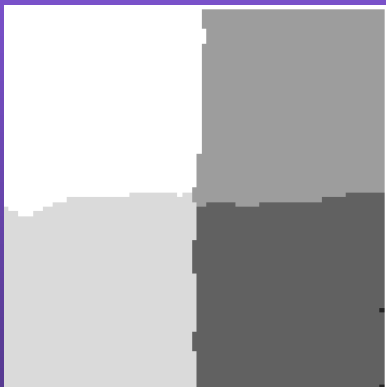
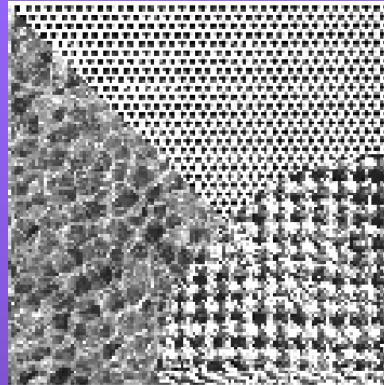
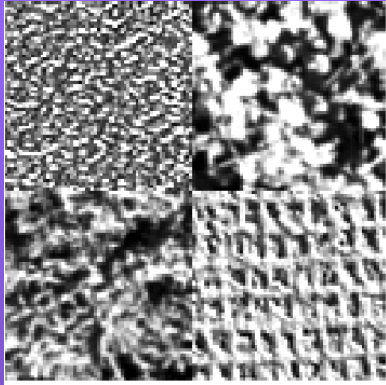
# Range Image Analysis

---



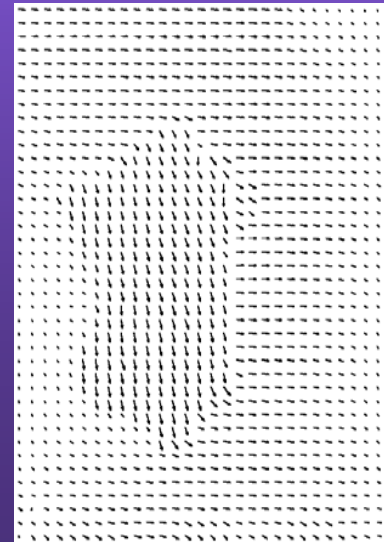
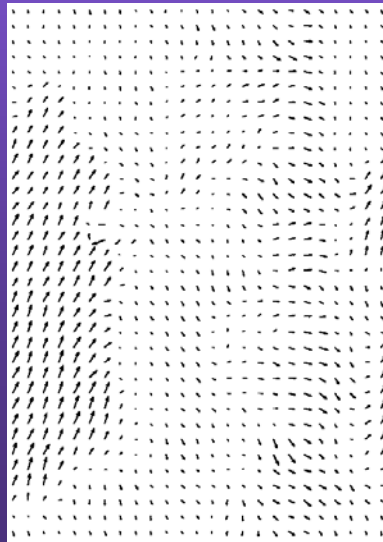
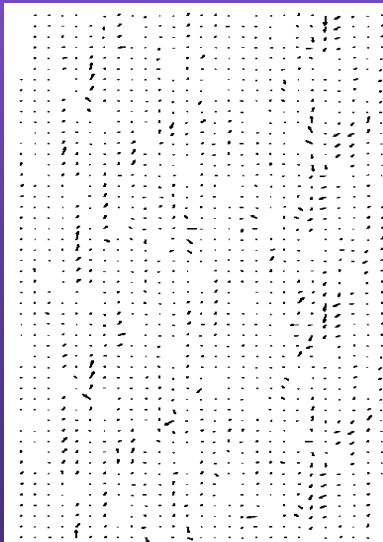
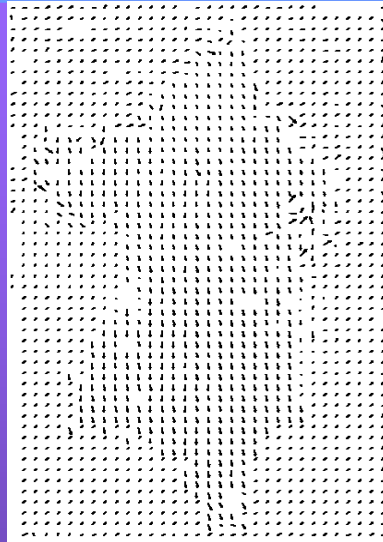
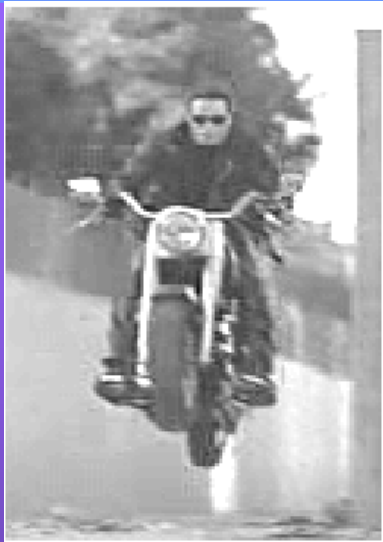
**Left:** input; **Center:** segmentation result; **Right:** actual 3D model

# Texture Segmentation



**Upper:** input; **Lower:** segmentation result

# Motion Analysis and Comparisons



# Auditory Scene Analysis (Bregman'90)

---

- **Listeners are able to parse the complex mixture of sounds arriving at the ears in order to retrieve a mental representation of each sound source**
- **Auditory scene analysis (ASA) takes place in two conceptual stages:**
  - **Segmentation.** Decompose the acoustic signal into ‘sensory elements’ (segments)
  - **Grouping.** Combine segments into groups, such that segments in the same group are likely to have arisen from the same environmental source

## ASA Problem - continued

---

- **The grouping process involves two mechanisms:**
  - **Primitive grouping.** Innate data-driven mechanisms, consistent with those described by the Gestalt psychologists for visual perception (proximity, similarity, common fate, good continuation etc.)
  - **Schema-driven grouping.** Application of learned knowledge about speech, music and other environmental sounds

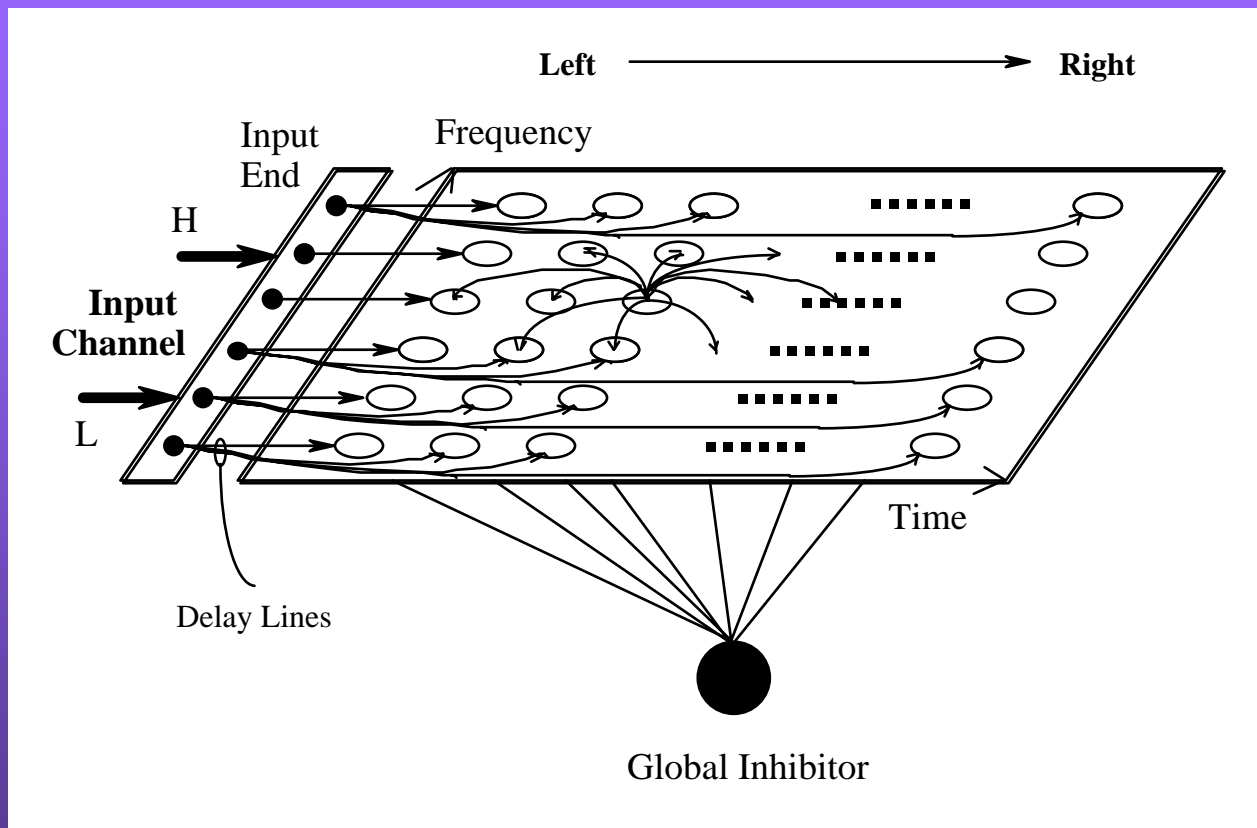
# Binding Problem in Audition

---

- **Information about acoustic features (pitch, spectral shape, interaural differences, AM, FM) is extracted in distributed areas of the auditory system**
- **How are these features combined to form a whole?**



# LEGION Architecture for Stream Segregation



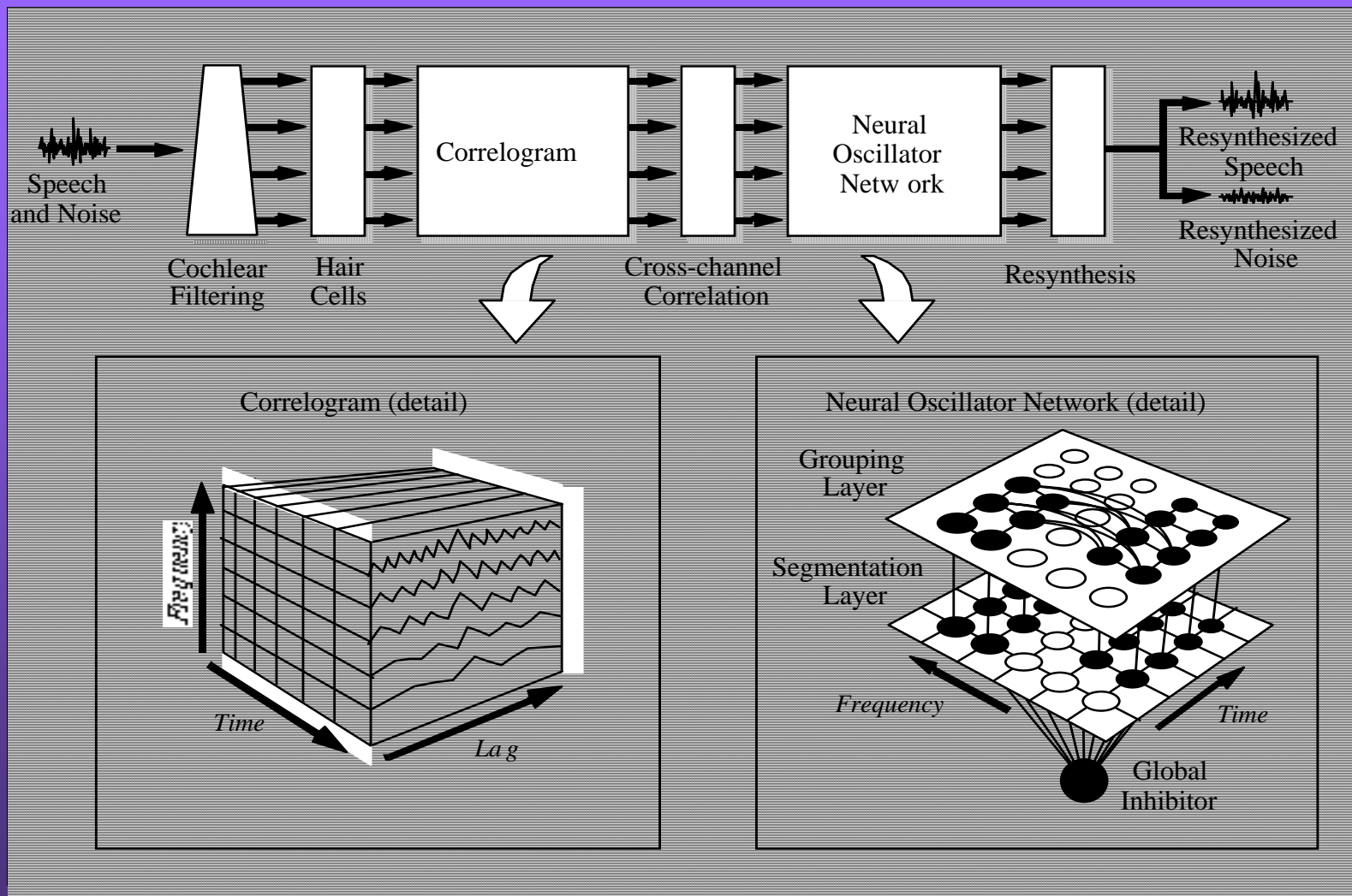
- See Wang'96

# Computational Auditory Scene Analysis

---

- **The ASA problem and the binding problem are closely related; the oscillatory correlation framework can address both issues**
- **Previous work also suggests that:**
  - **Representation** of the auditory scene is a key issue
  - **Temporal continuity** is important (although it is ignored in most frame-based sound processing algorithms)
  - **Fundamental frequency (F0)** is a strong cue for grouping

# A Multi-stage Model for CASA



# Auditory Periphery Model

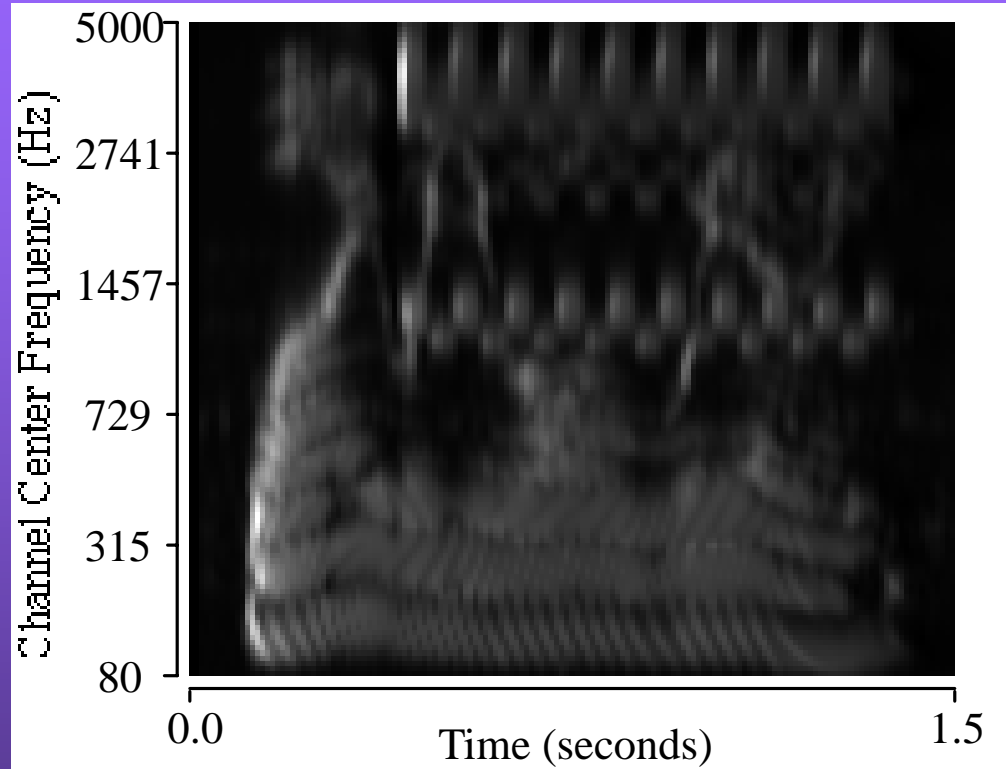
---

- A bank of gammatone filters

$$g_i(t) = t^{n-1} \exp(-2\pi b_i t) \cos(2\pi f_i t + \phi_i) H(t)$$

- $n$ : filter order (fourth-order is used)
  - $b$ : bandwidth
  - $H$ : Heaviside function
- Meddis hair cell model converts gammatone output to neural firing

# Auditory Periphery - Example



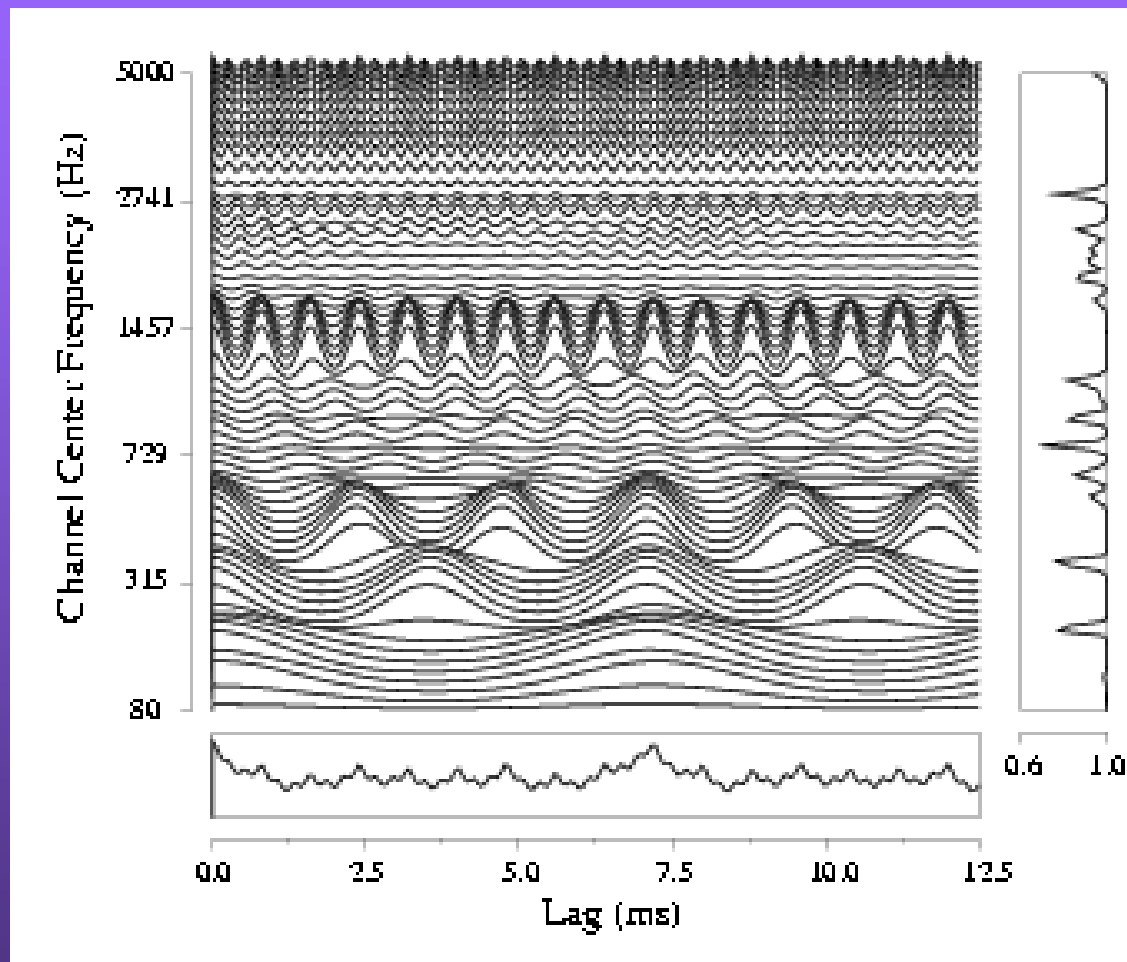
- Hair cell response to utterance: “Why were you all weary?” mixed with phone ringing
- 128 filter channels arranged in ERB

# Mid-level Auditory Representations

---

- Mid-level representations form the basis for segment formation and subsequent grouping processes
- *Correlogram* extracts periodicity information from simulated auditory nerve firing patterns
- *Summary correlogram* can be used to identify F0
- *Cross-correlation* between adjacent correlogram channels identifies regions that are excited by the same frequency component

# Mid-level Representations - Example



**Correlogram and cross-correlation for the speech/telephone mixture**

# Oscillator Network: Segmentation Layer

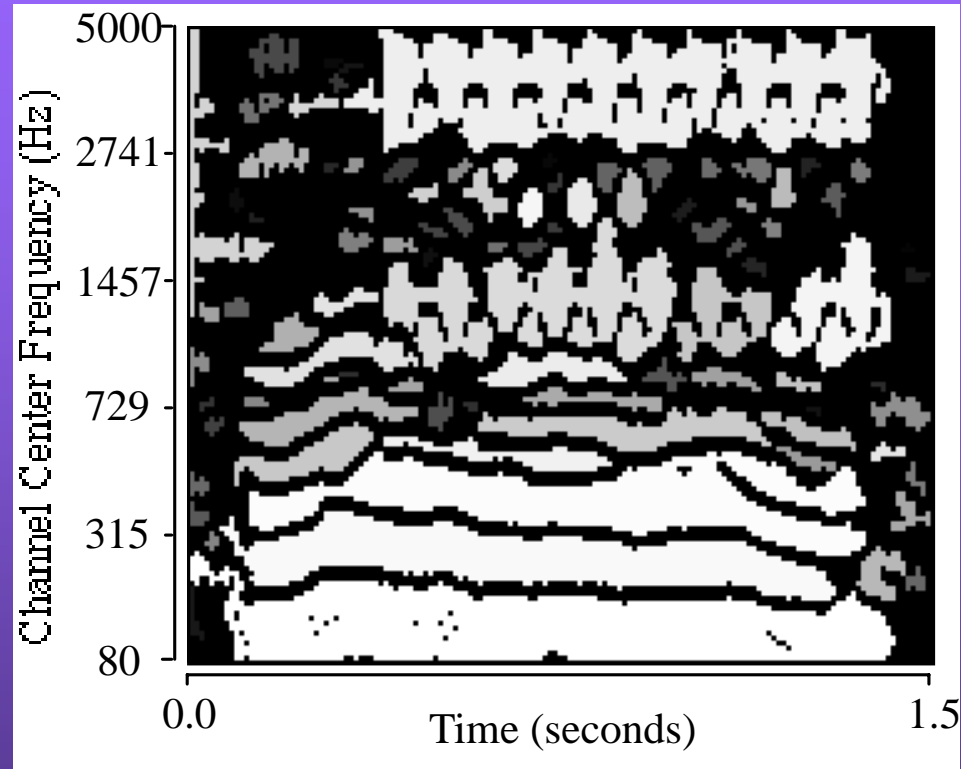
---

- **Horizontal weights are unity, vertical weights are unity if correlation exceeds threshold, otherwise 0**
- **Oscillators receive input if energy in corresponding channel exceeds a threshold**
- **All oscillators are connected to a global inhibitor, which ensures that different segments are desynchronized from one another**
- **A LEGION network**



# Segmentation Layer - Example

---



- Output of the segmentation layer to the speech/telephone mixture

# Oscillator Network: Grouping Layer

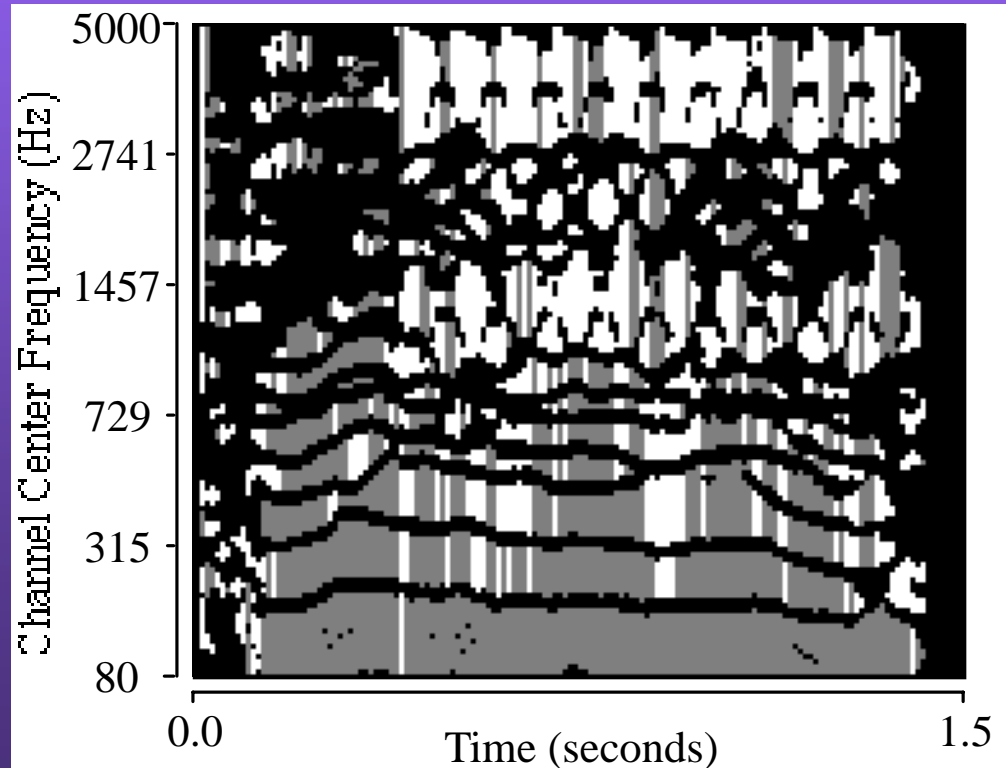
---

- The second layer is a two-dimensional oscillator network without global inhibition, which embodies the grouping stage of ASA
- Oscillators in the second layer only receive input if the corresponding oscillator in the first layer is stimulated
- At each time frame, an F0 estimate from the summary correlogram is used to classify channels into two categories; those that are consistent with the F0, and those that are not

## Grouping Layer - continued

- Enforce a rule that all channels of the same time frame within each segment must have the same F0 category as the majority of channels

Result of the speech  
telephone example

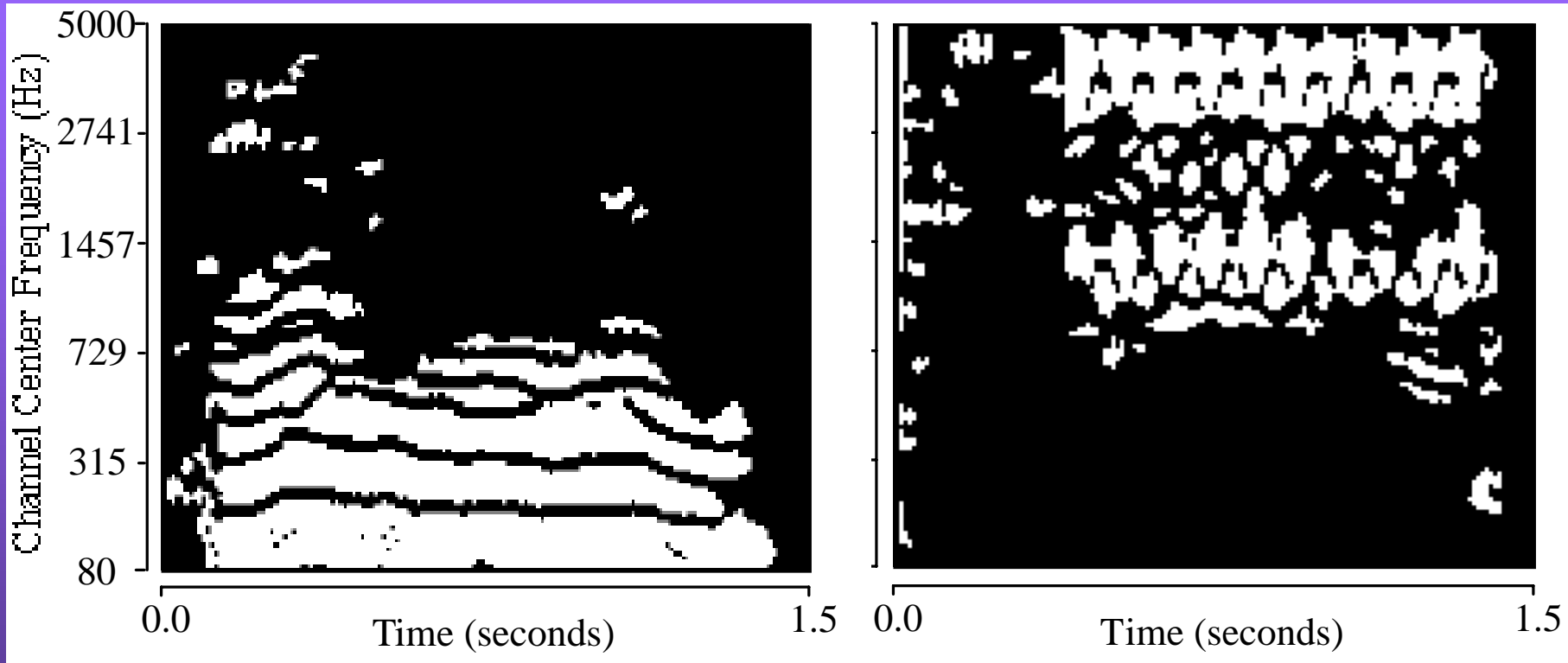



## Grouping Layer - continued

---

- **Grouping is limited to the time window of the longest segment**
- **There are horizontal connections between oscillators in the same segment**
- **Vertical connections are formed between pairs of channels within each time frame; mutual excitation if the channels belong to the same F0 category, otherwise mutual inhibition**

# Grouping Layer - Example



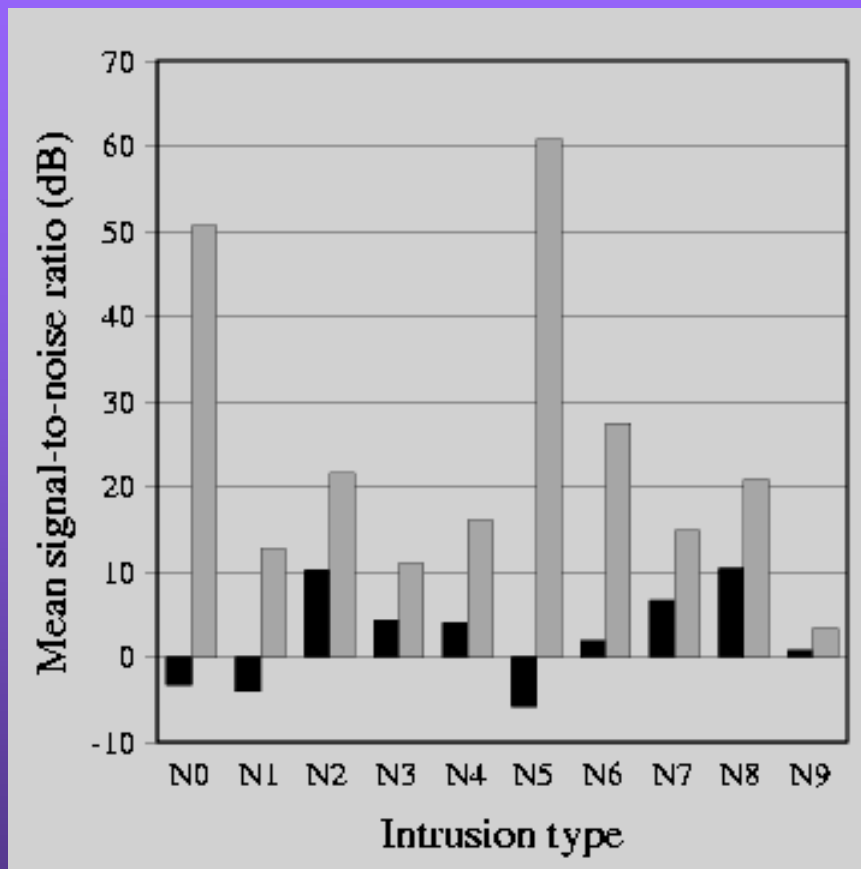
- Two streams emerge from the group layer
  - Foreground: left  (original mixture )
  - Background: right

# Evaluation

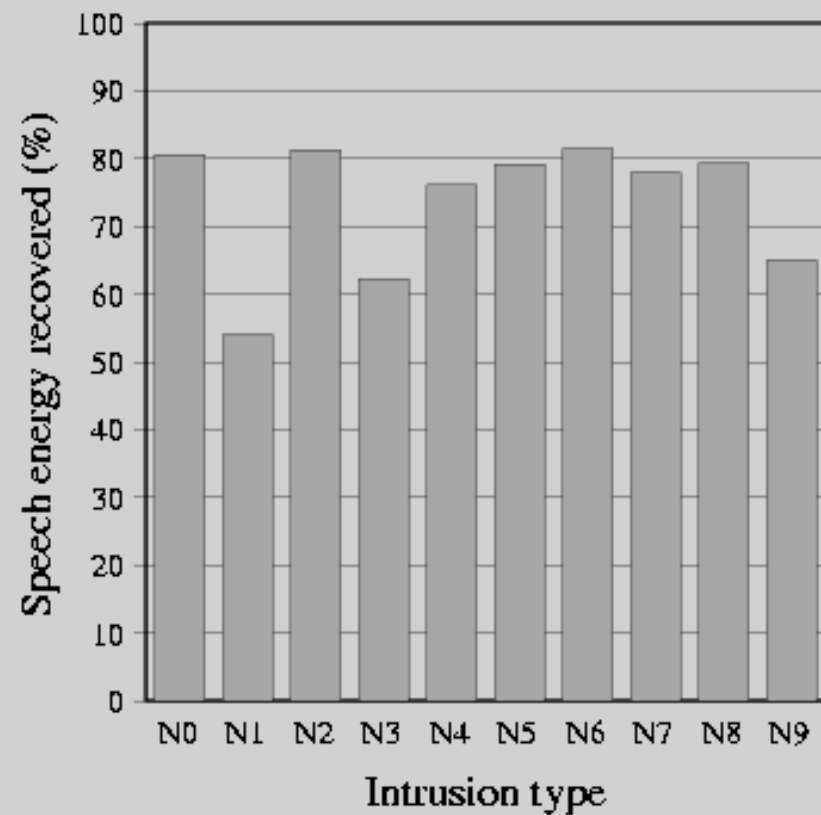
---

- **Evaluated on a corpus of 100 mixtures (Cooke'93): 10 voiced utterances x 10 noise intrusions**
  - Noise intrusions have a large variety
- **Resynthesis pathway allows estimation of SNR after segregation; improvement in SNR after processing for each noise condition**

# Results of Evaluation



**Changes in SNR**



**Speech energy retained**

# An Extended Model for Speech Segregation

---

- **Pitch estimation of target speech**
- **Differential processing for low and high frequency channels**

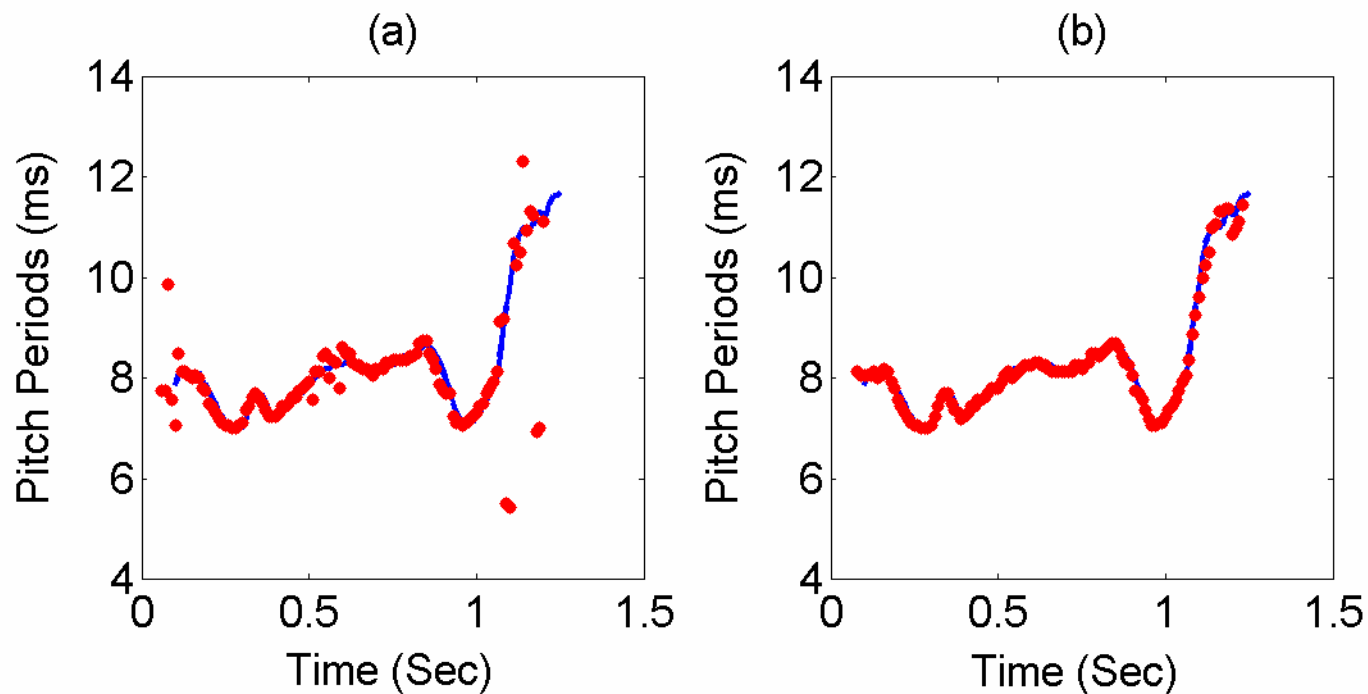


# Extended Model for Speech Segregation

---

- **Estimation of Target Pitch Contour**
  - Use the foreground stream from the Wang-Brown model as the basis
  - Label pitch points as reliable or unreliable depending on the agreement between estimated pitch and its basis of support at a particular time frame
  - Further interpolation between reliable pitch trajectories

# Pitch Tracking - Example



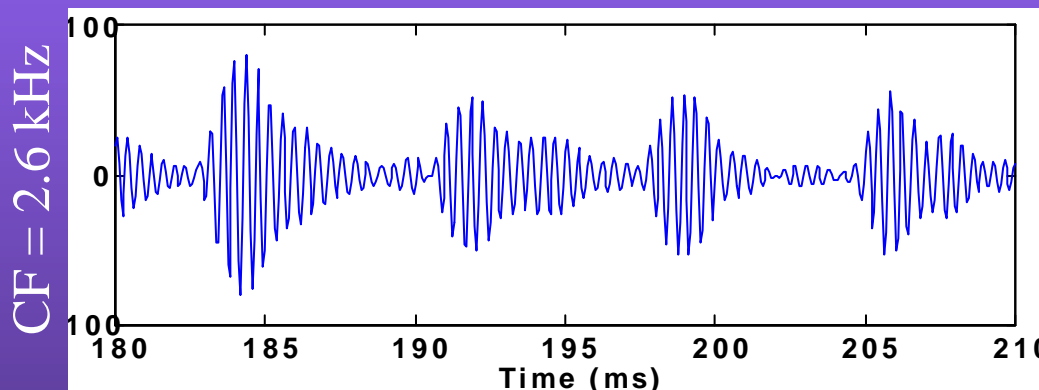
**Left: Global pitch (Line: pitch track of clean speech)**

**Right: Estimated target pitch**

# Amplitude Modulation

- **Differential Processing of High-frequency Channels**

- Motivation: Amplitude modulation (AM) due to beats or combinational tones



*Input:* Single  
Male Utterance

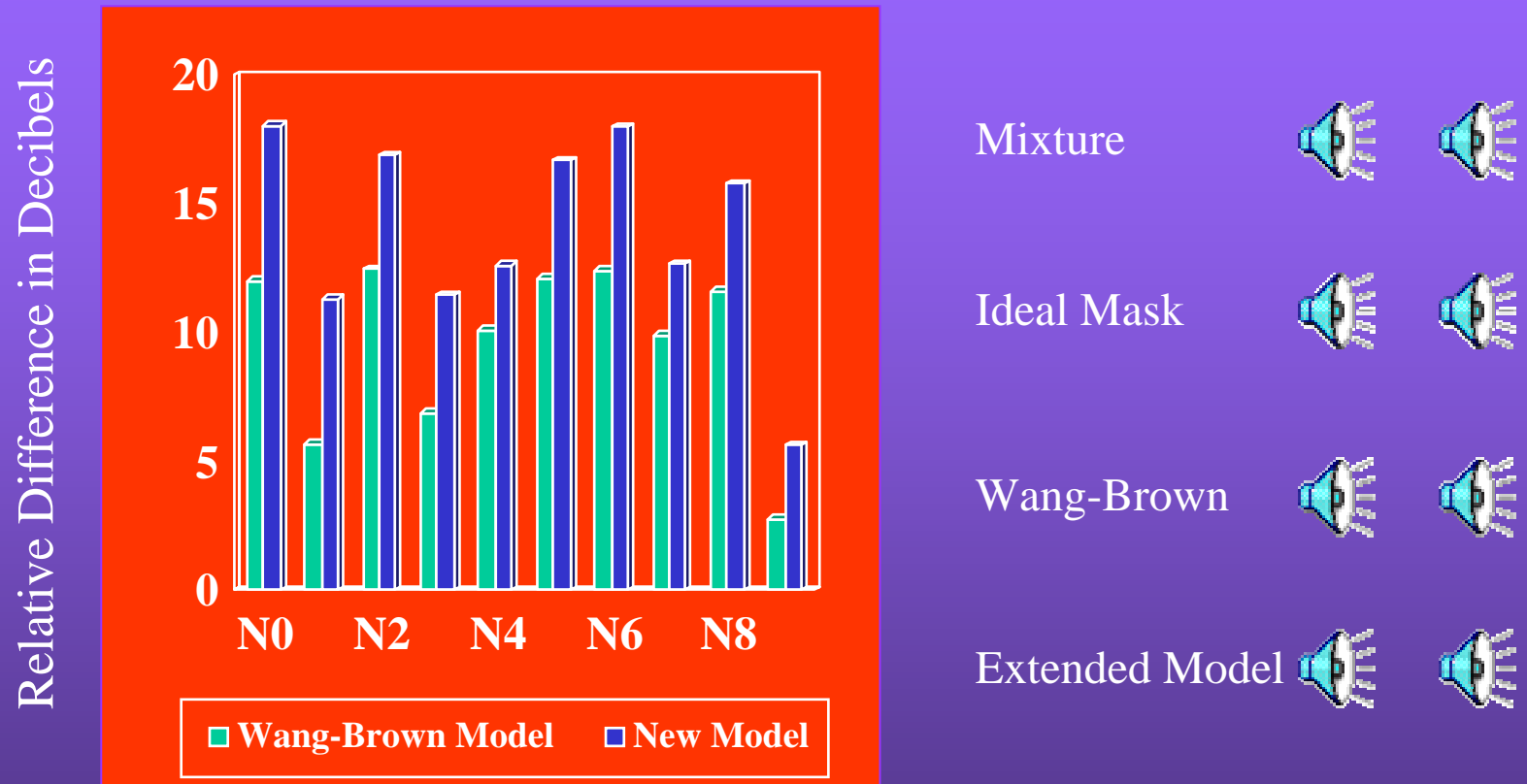
- Calculate envelopes for high-frequency channels ( $> 1$  kHz)
- Further computation based on extracted envelopes, not filter responses

## Extended Model - continued

---

- **Evaluation Based on Ideal Binary Masks**
  - Motivation: Auditory masking - stronger signal masks weaker one within a critical band
  - Further motivation: Ideal binary masks give excellent listening experience and ASR performance
  - Thus, ideal binary masks are used as ground truth for evaluation

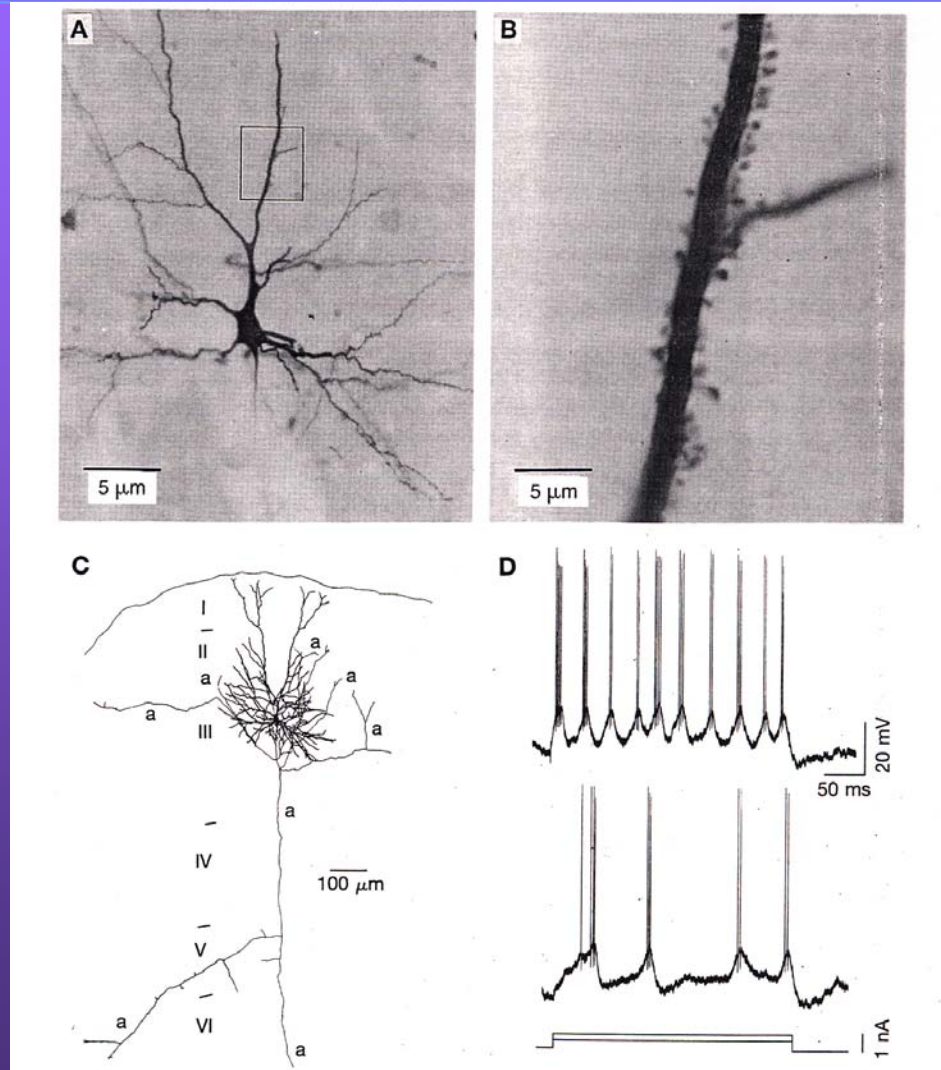
# New Results



- The extended model yields significant improvement

# Back to Biology

Gray & McCormick '96



# Summary

---

- **Survey of different approaches to scene analysis**
- **Emphasis on oscillatory correlation approach**
  - Rigorous analysis of the selective gating mechanism and LEGION dynamics
  - Both synchronous oscillations and the structure of the model are neurally plausible
  - An effective method for scene segmentation and a theory for perceptual organization
- **Both visual and auditory scene analysis are addressed with key applications illustrated**