Temporal and Top-Down Processing in Audition

Malcolm Slaney

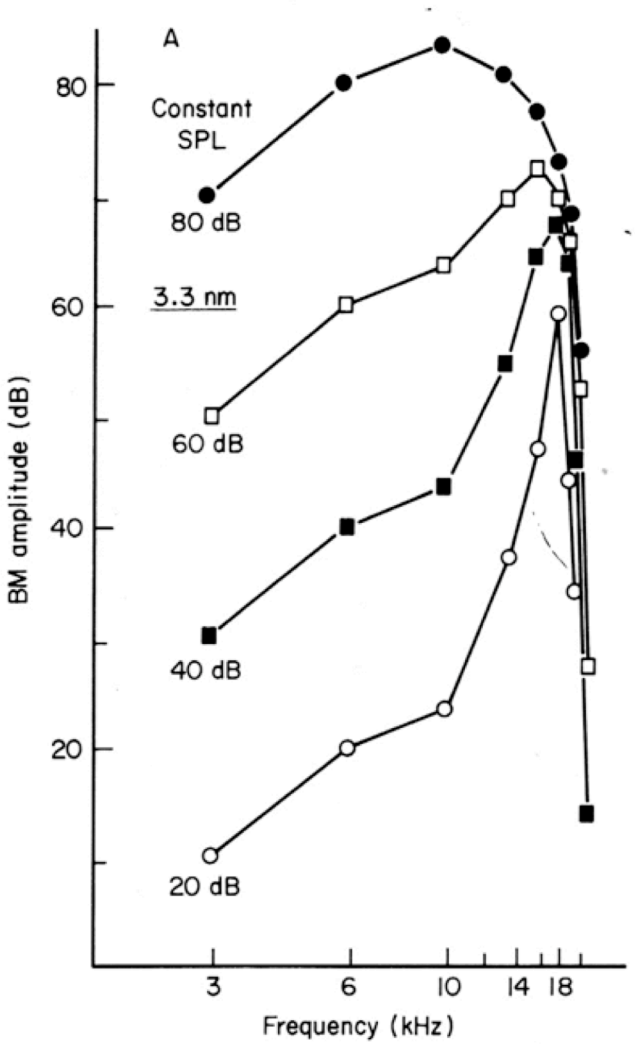
Yahoo! Research

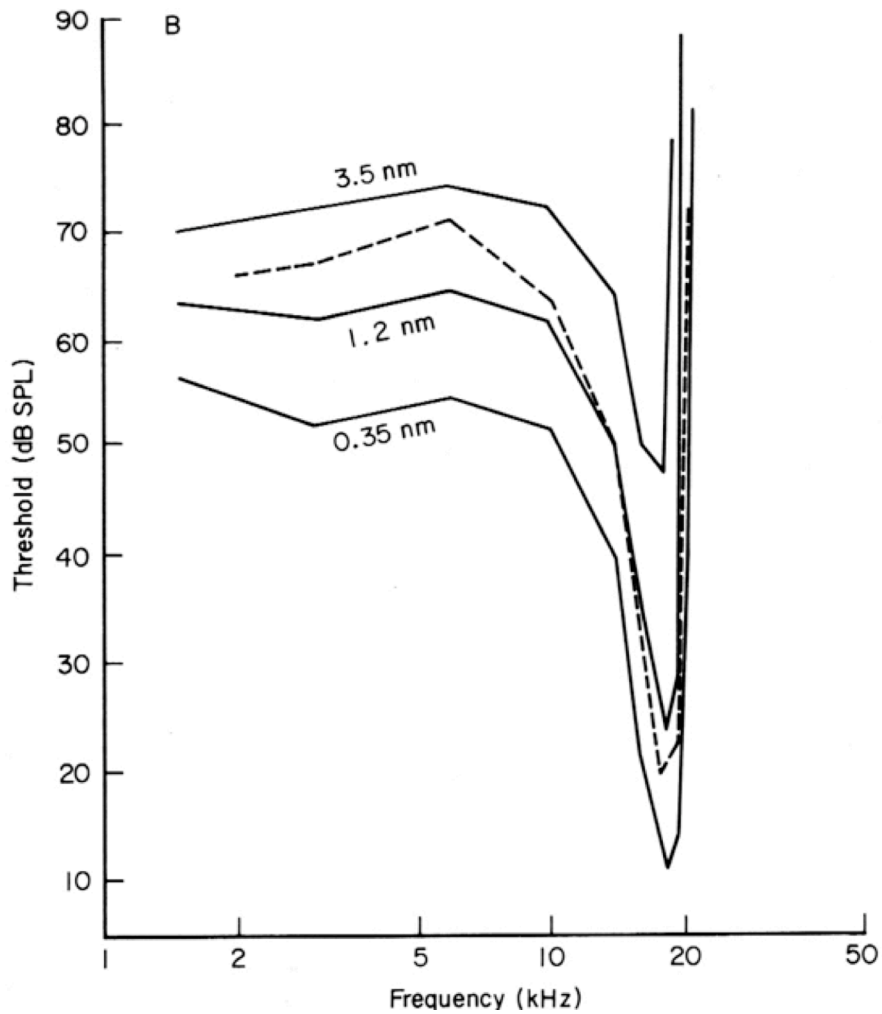
Stanford CCRMA

Outline

- The Non-Linear Cochlea
- Correlogram
- Pitch
- Information Flow
- CASA
- Saliency
- Relational Networks

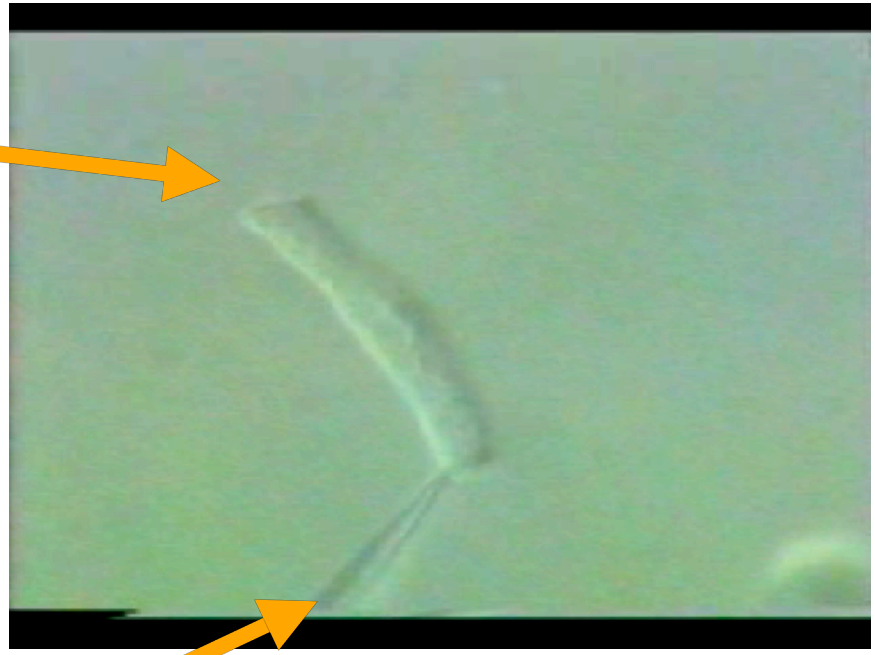
Cochlear Nonlinearity





Hair Cell Motility

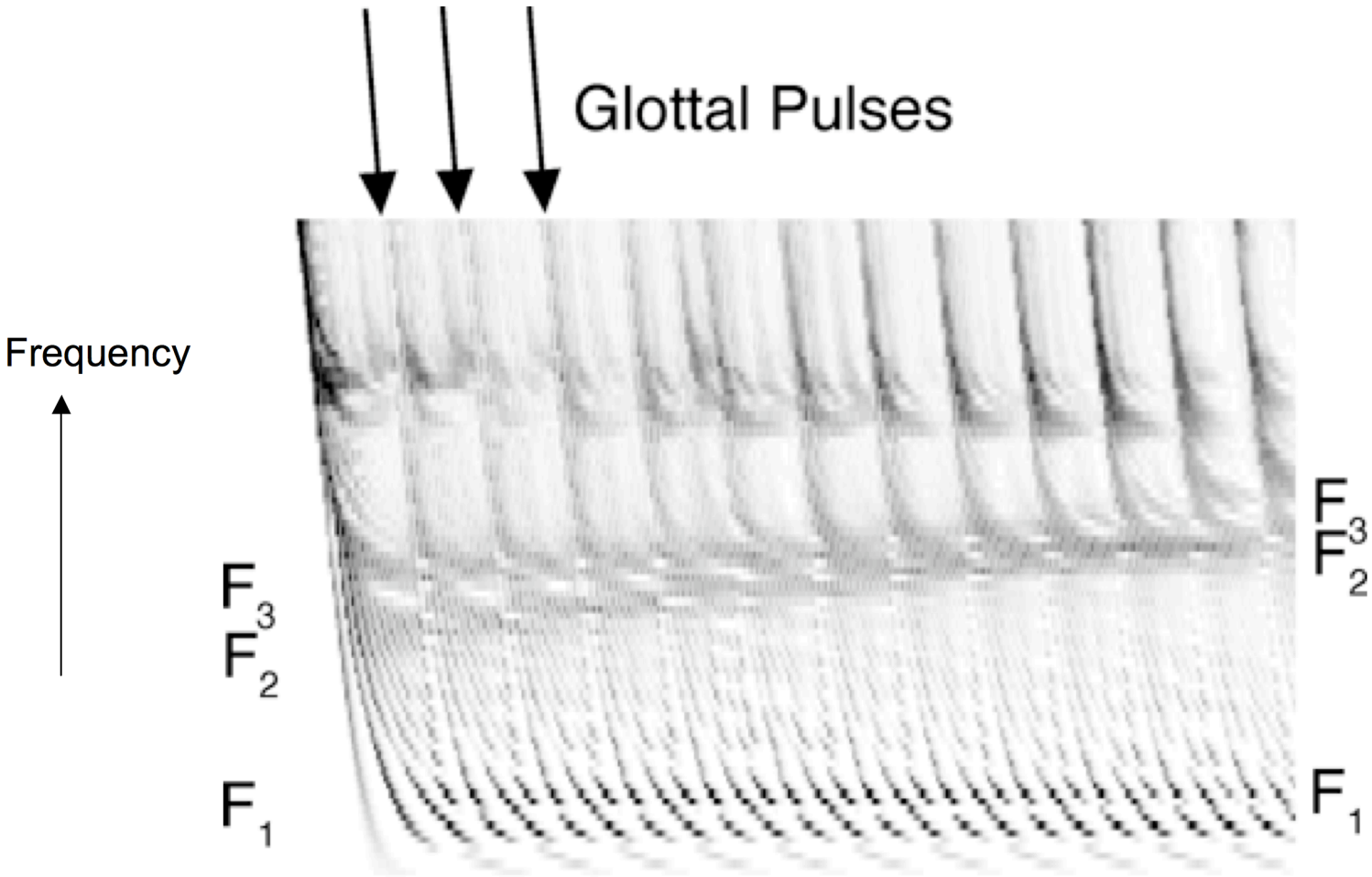
Cillia



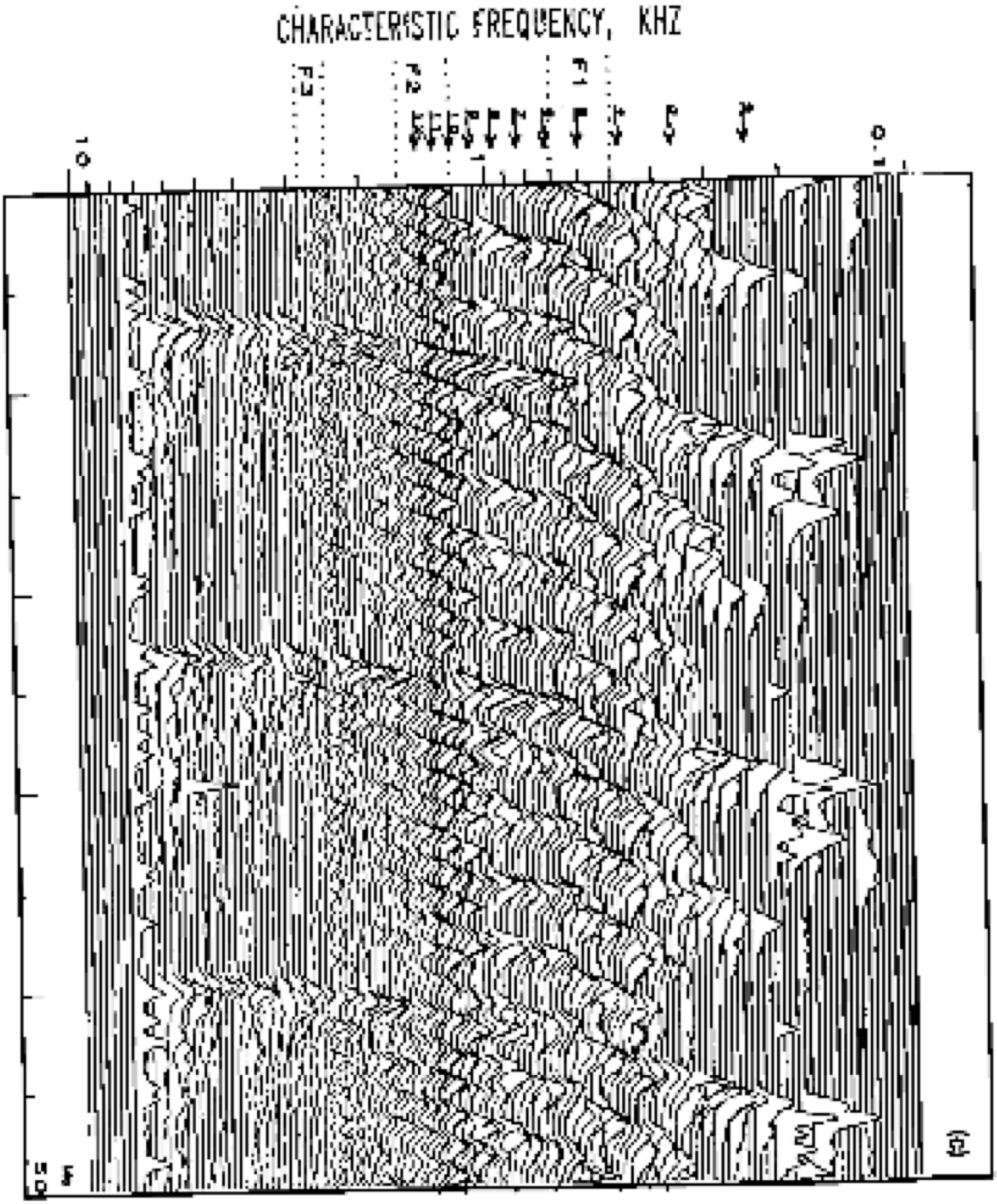
Pipette (changing membrane potential)

By Jonathan Ashmore - http://www.physiol.ucl.ac.uk/ashmore/hairexpl.ht

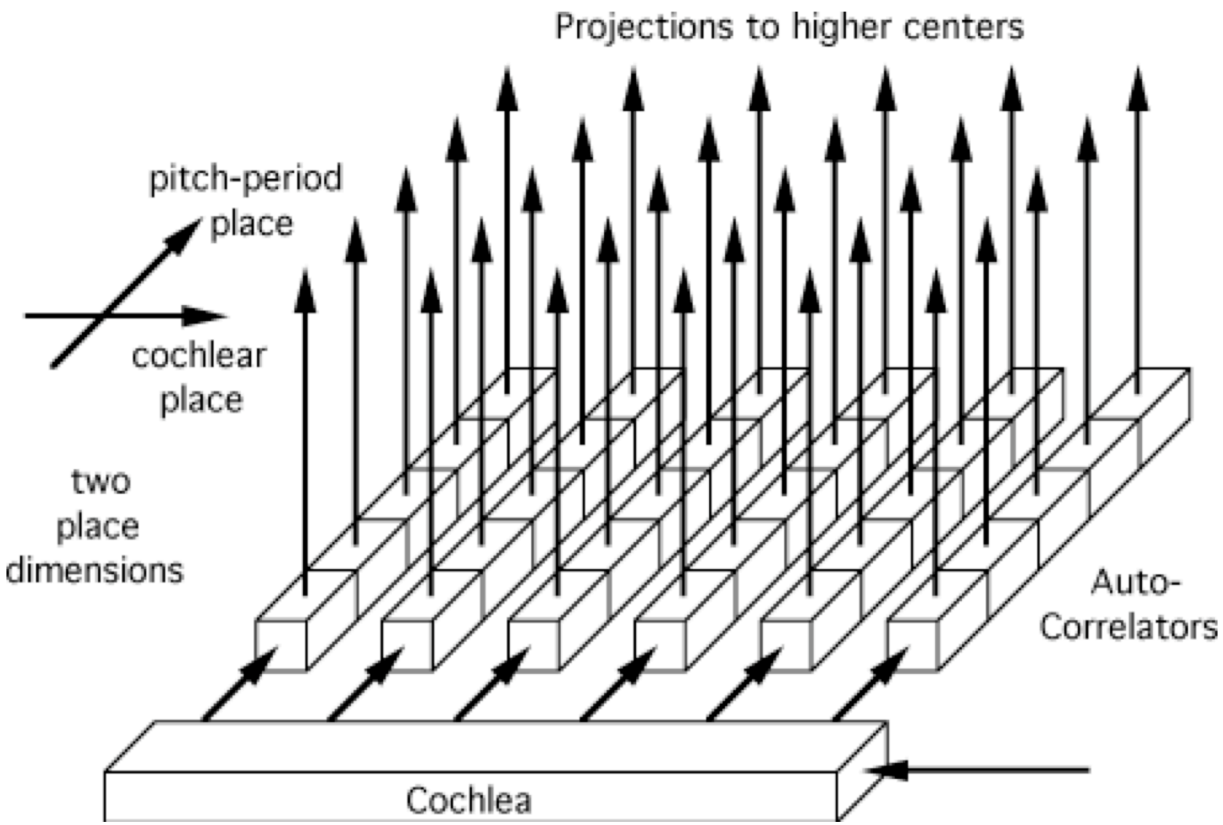
Auditory Nerve Patterns



Cochleagram of “Re”

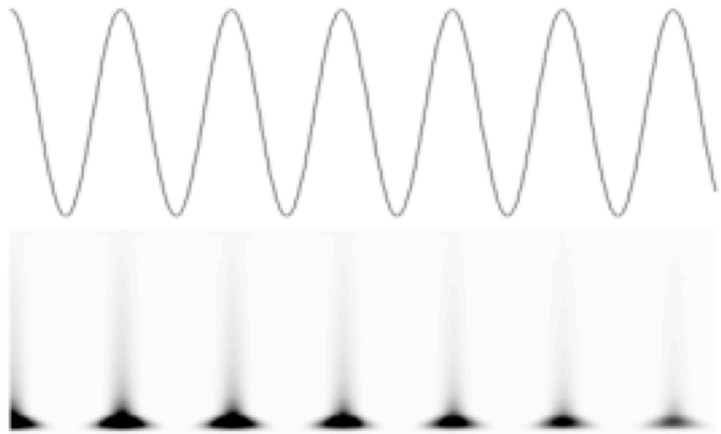

CHARACTERISTIC FREQUENCY, KHZ
10
0.1
50 F

Correlogram Architecture

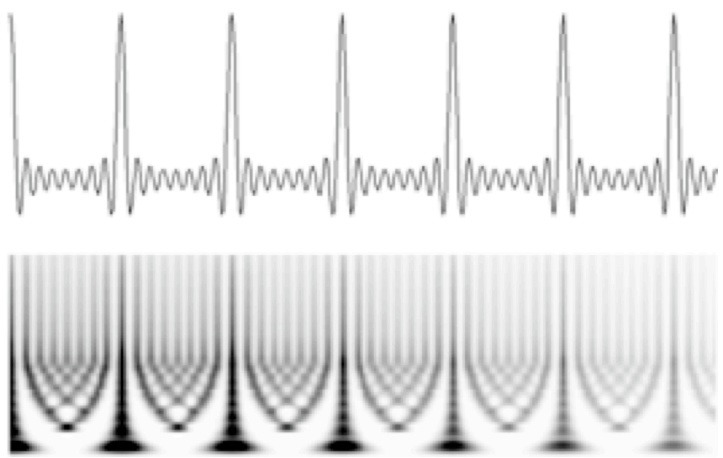


Tone Correlograms

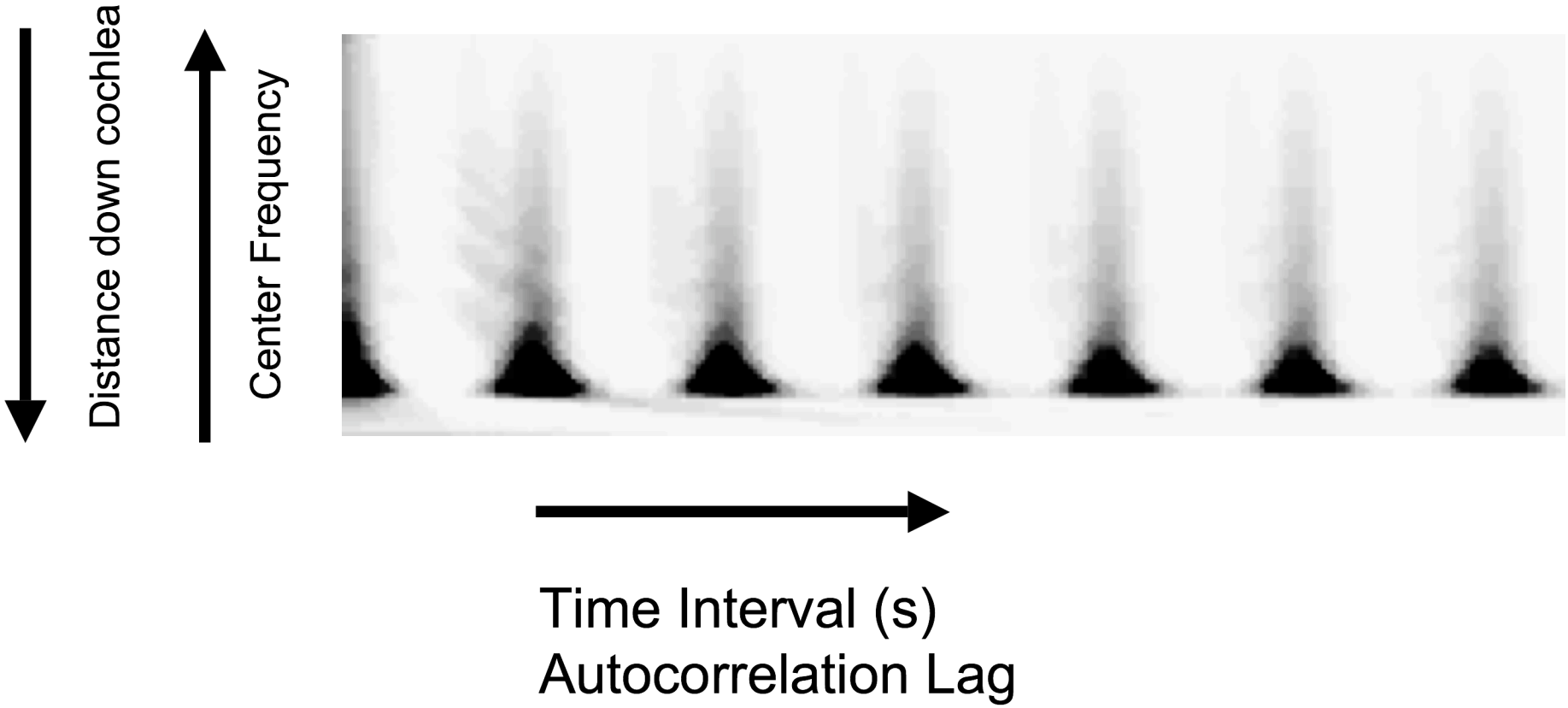
200Hz Tone



Harmonics of 200Hz



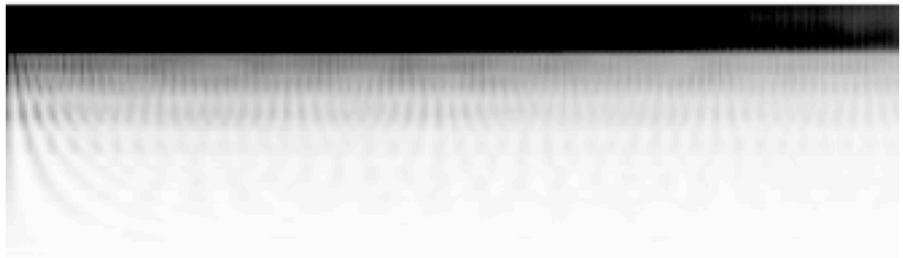
Tone Examples



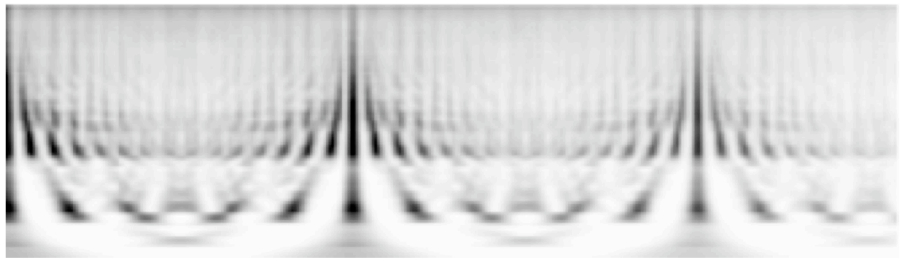
With help from Richard O. Duda

Speech Correlogram

Unvoiced /s/

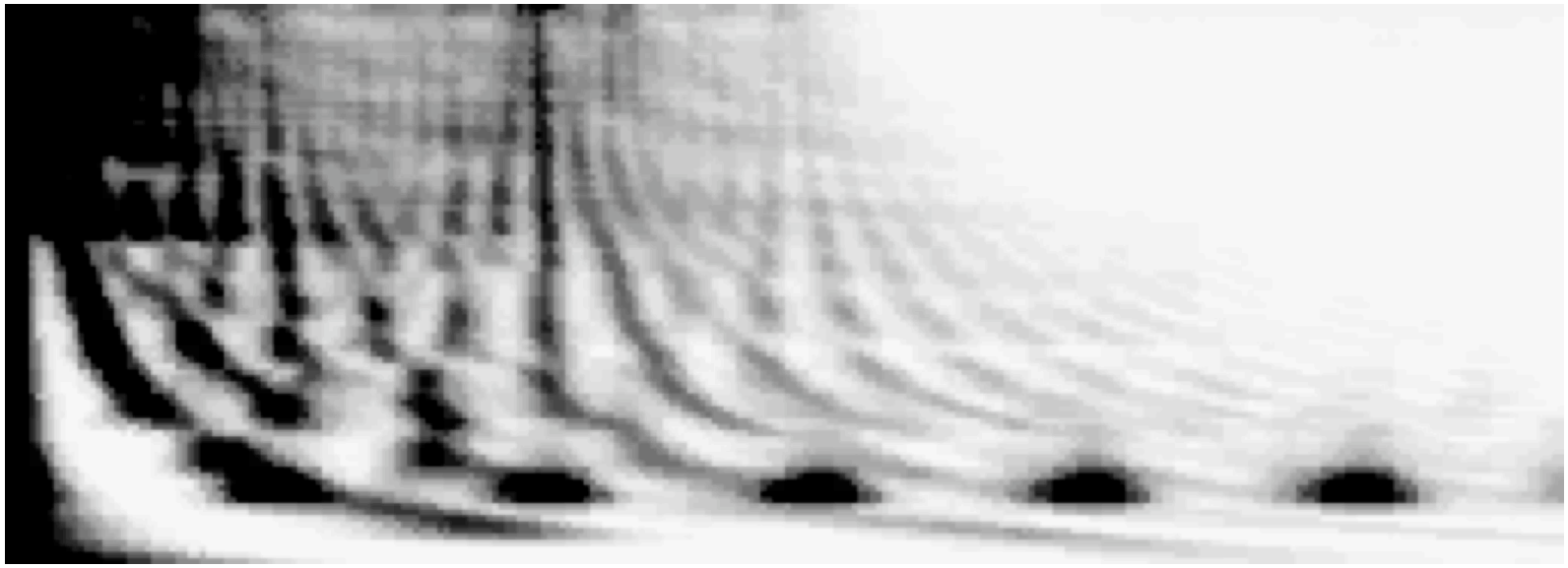


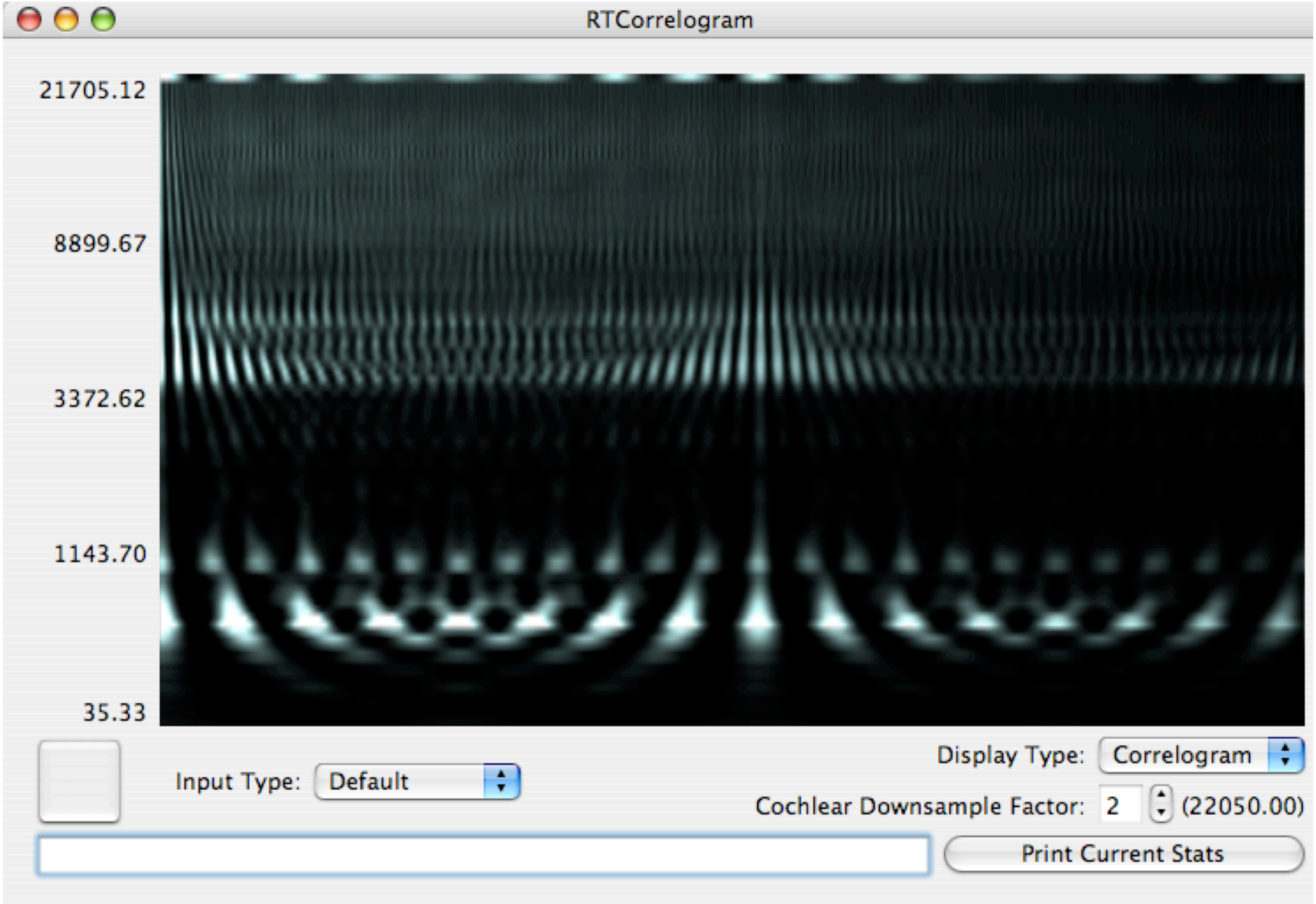
Voiced Vowel



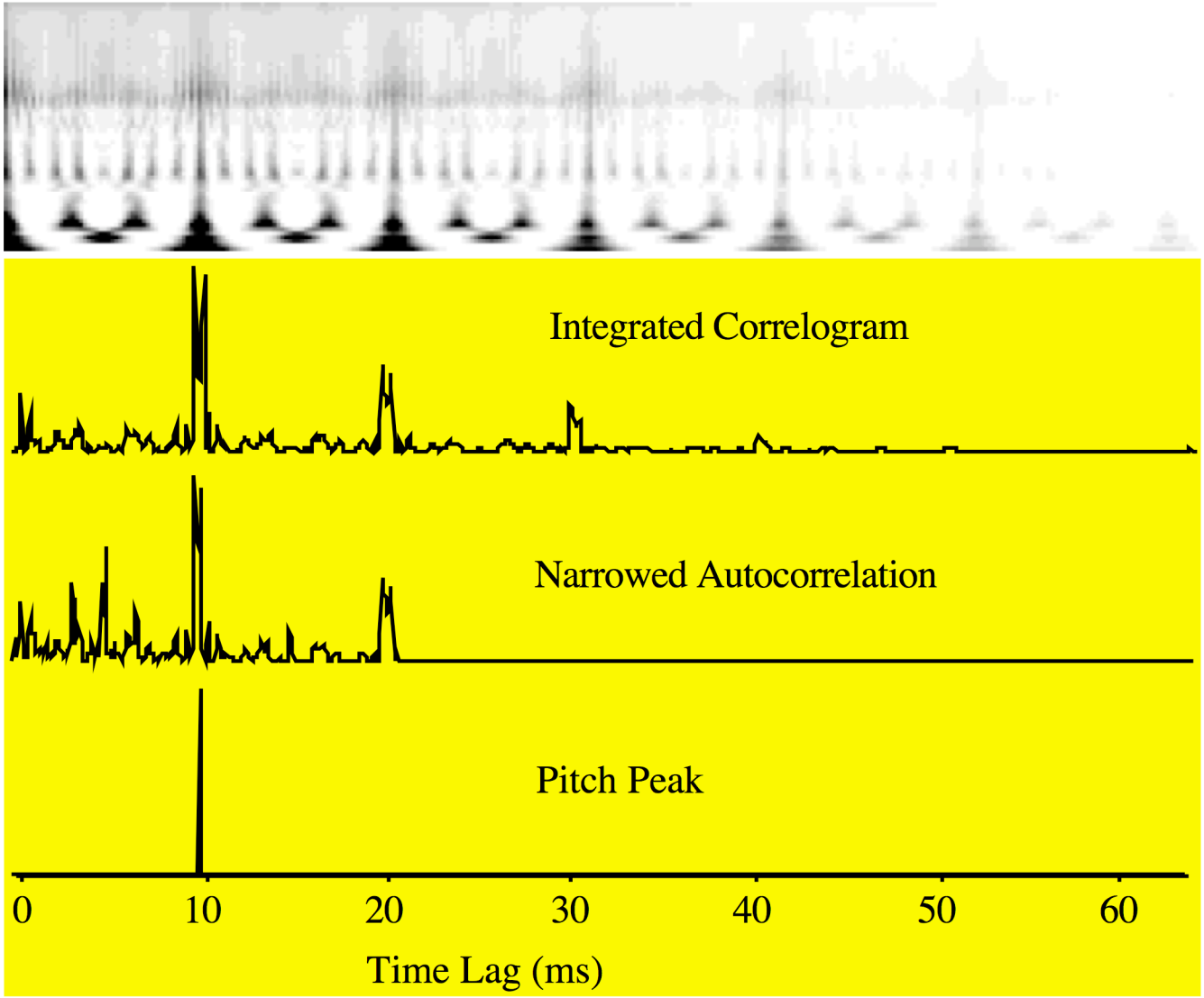
Speech Examples

- Speech Examples
  - Leonardo at Apple

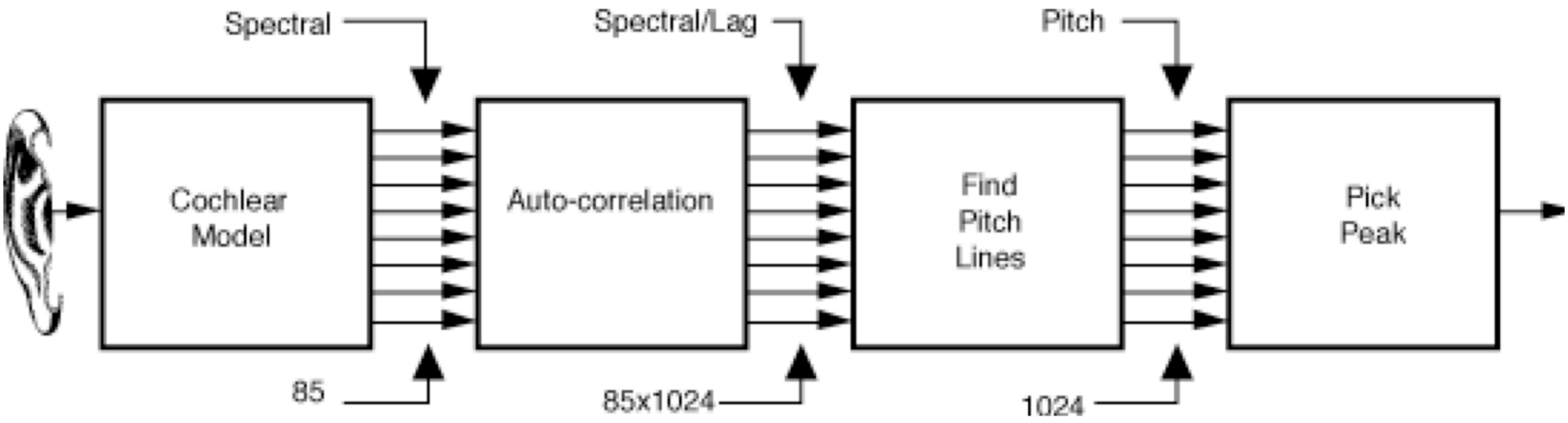



RTCorrelogram
21705.12
8899.67
3372.62
1143.70
35.33
Input Type:
Default
Display Type:
Correlogram
Cochlear Downsample Factor:
2
(22050.00)
Print Current Stats

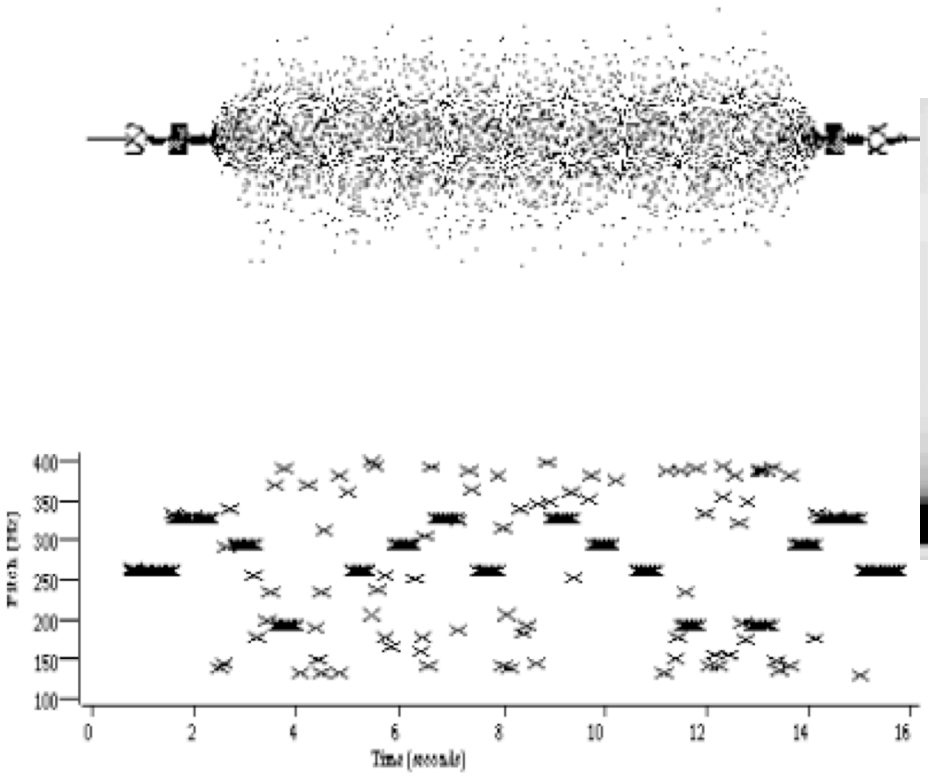
Pitch


Integrated Correlogram
Narrowed Autocorrelation
Pitch Peak
0
10
20
30
40
50
60
Time Lag (ms)

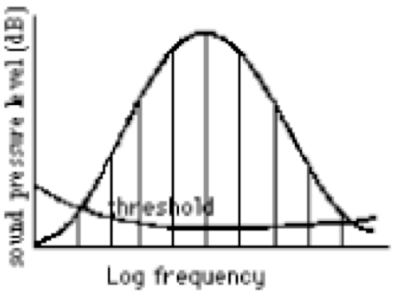
Pitch Architecture

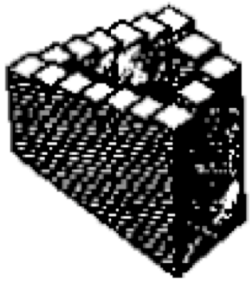


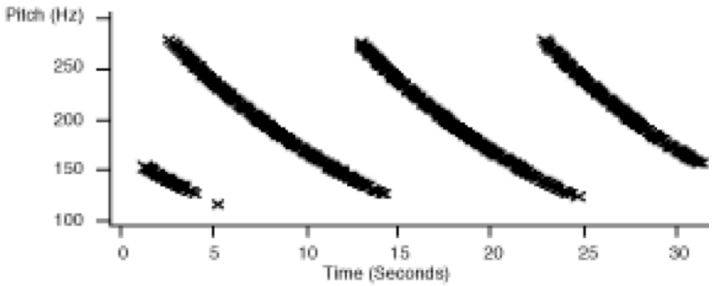
Pitch in Noise



Ambiguity of Pitch

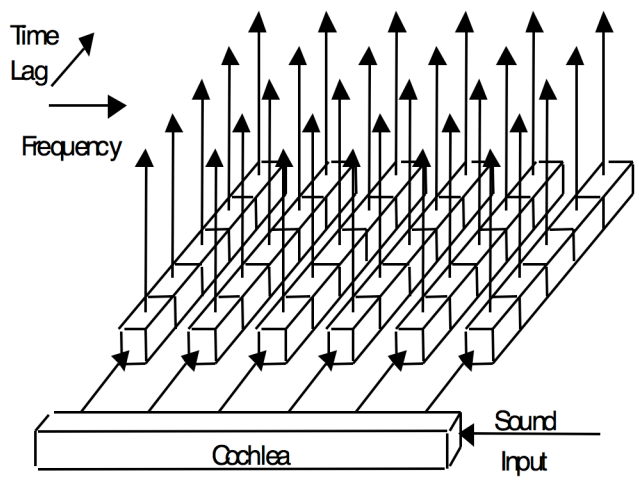




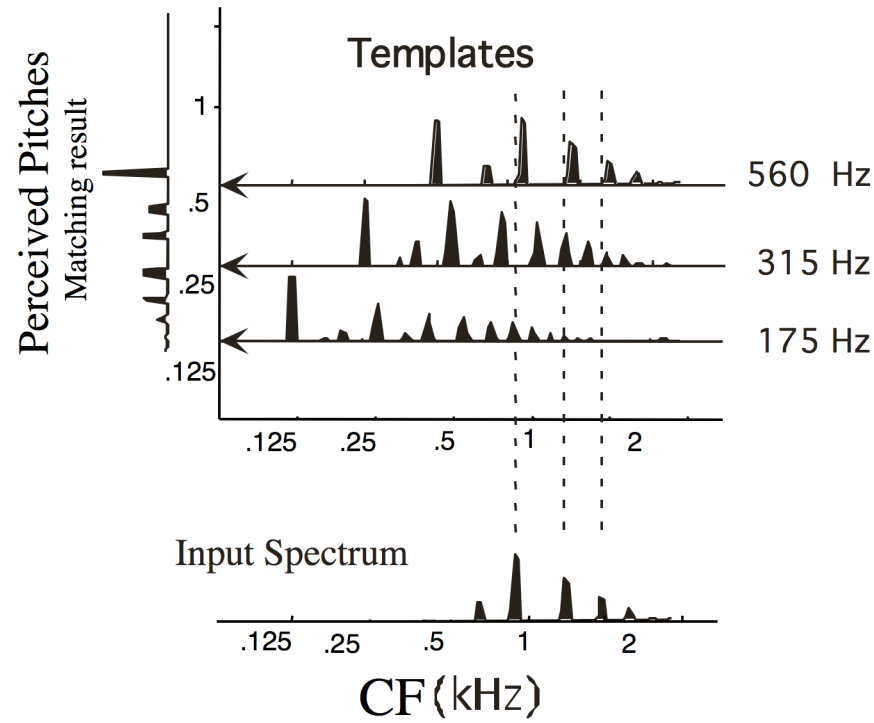


Pitch Theories

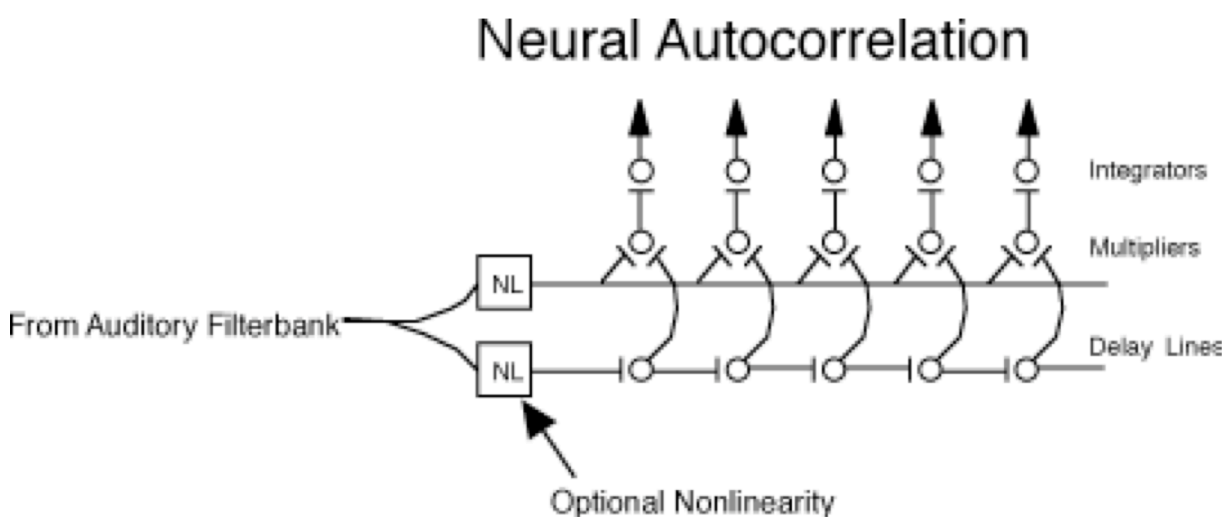
A *"Correlogram"*
Temporal implementations with neural delay lines



B *Central Pattern Match Theory*



Correlogram Implementation



Today
- FFT

Licklider
- Neural Delay

Patterson
- Strobed-temporal integration

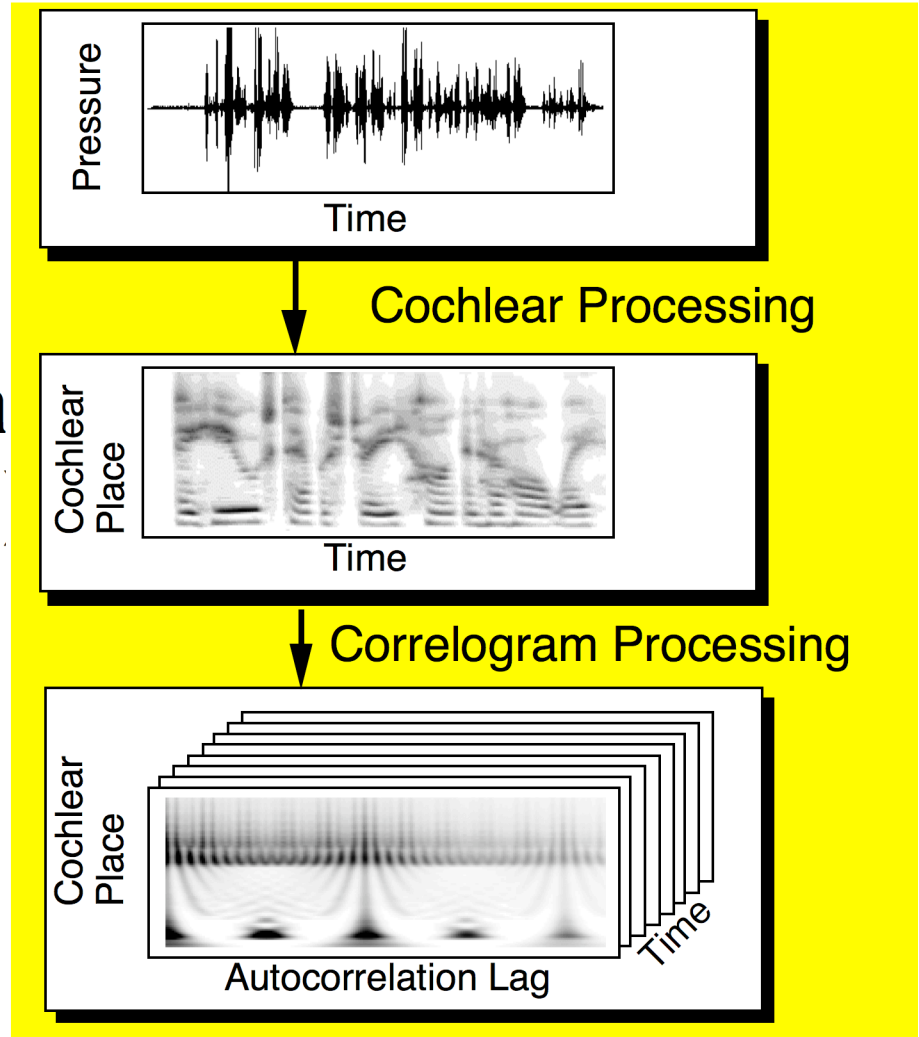
Alternative
- Oscillators
- AM Modulator

Processing Chain

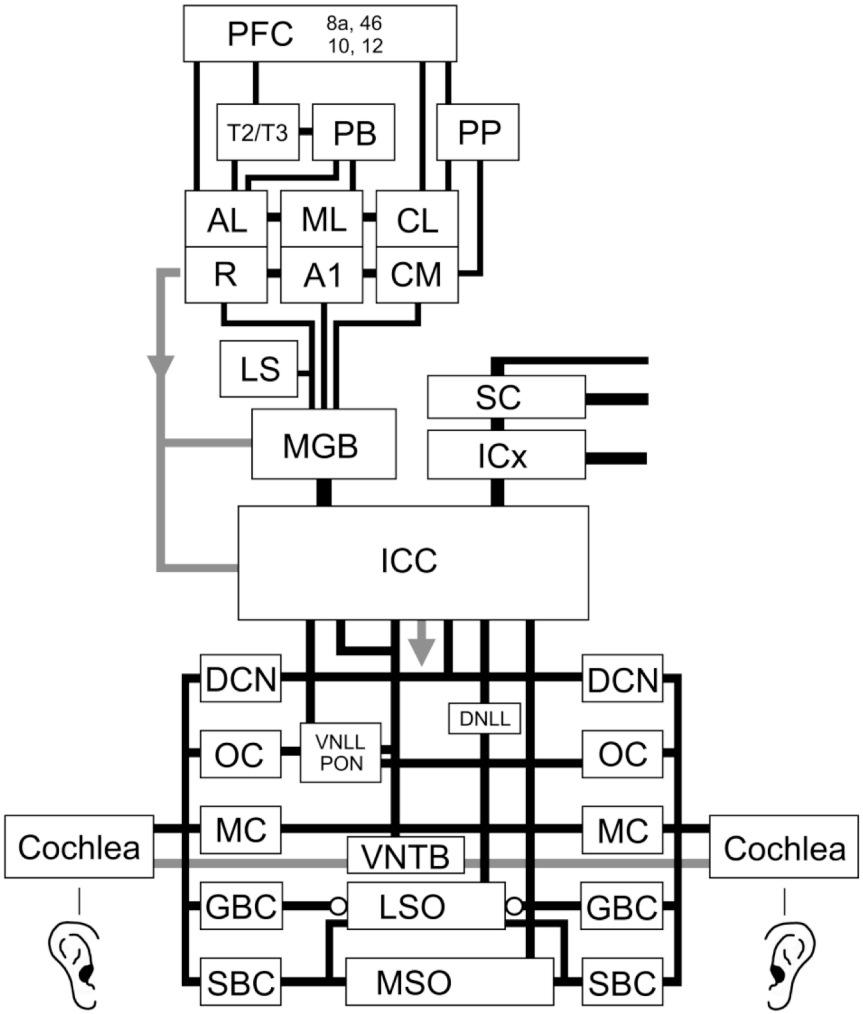
One Dimensional (waveform)

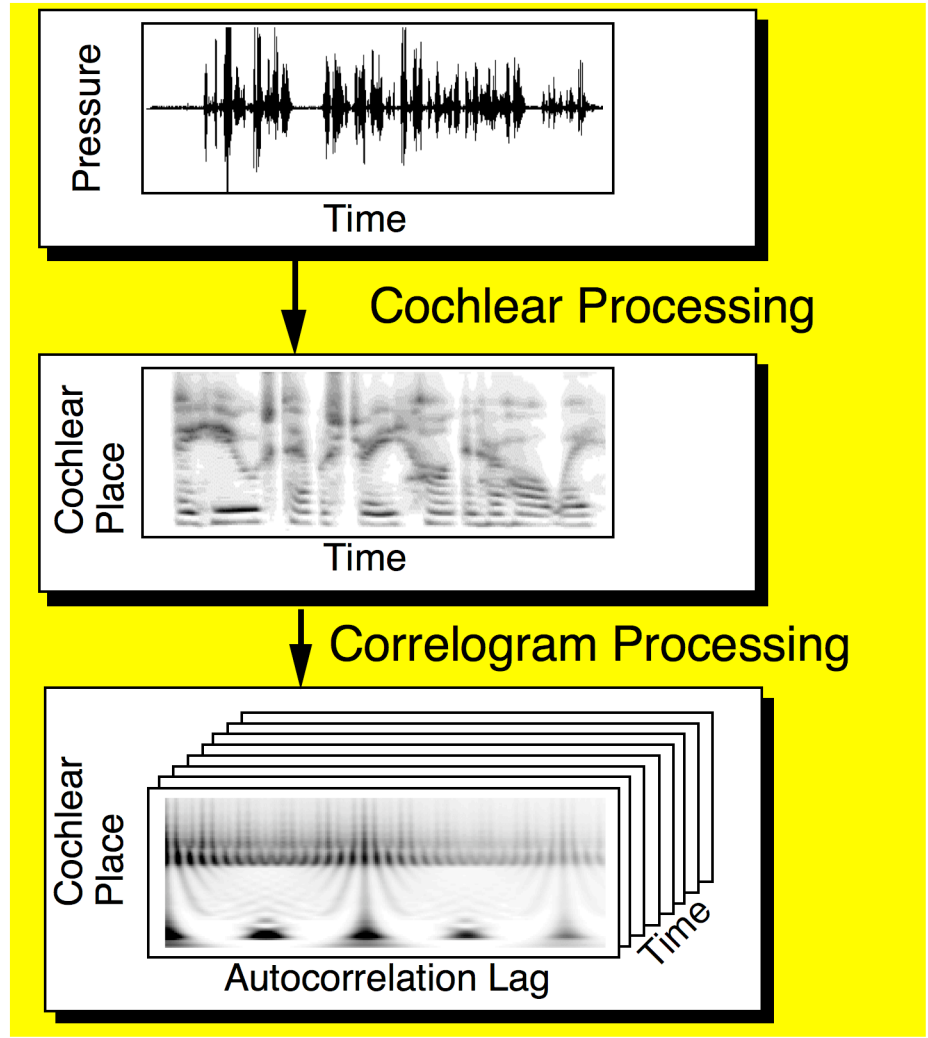
Two Dimensiona (not a spectrogram

Three Dimensional (neural movie)



Block Diagrams





Problems

Problems

- Continuation
  - Tone and Noise
  - Parliament Cough

Problems

- Continuation
  - Tone and Noise
  - Parliament Cough
- What do you hear?
  - Waveforms?
  - Ideas?

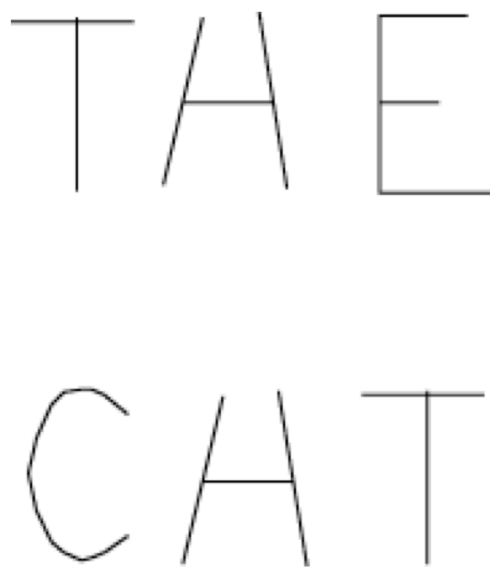
Speech Examples

Wedding

Sine

Natural

Visual Context



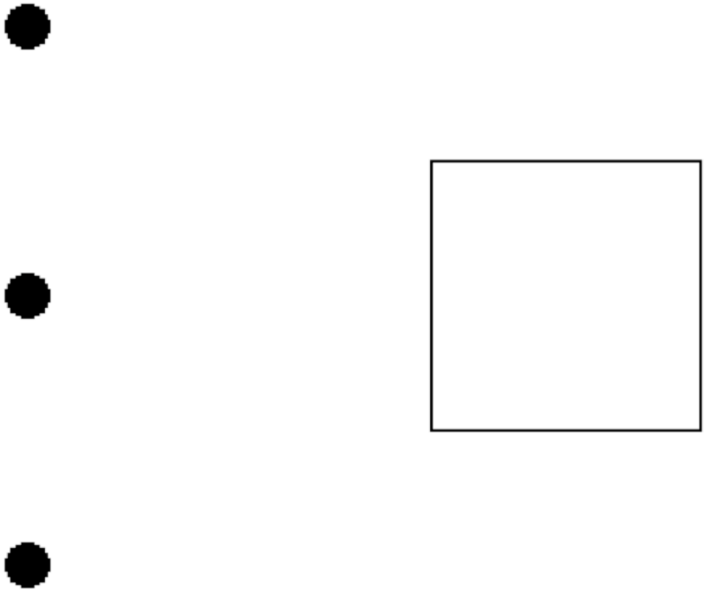
What Vowel is This?

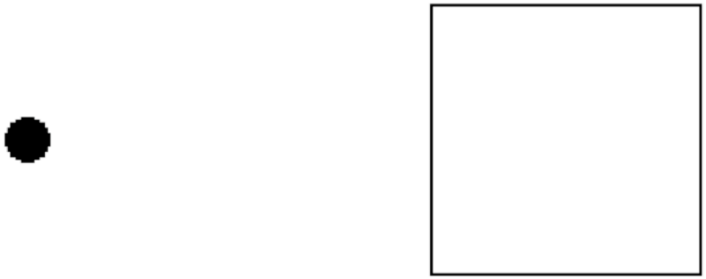
Peter Ladefoged

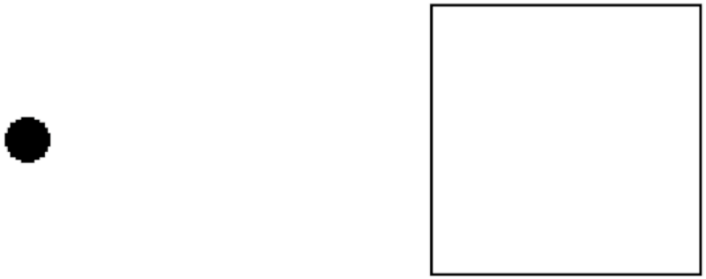
Word 1

Word 2

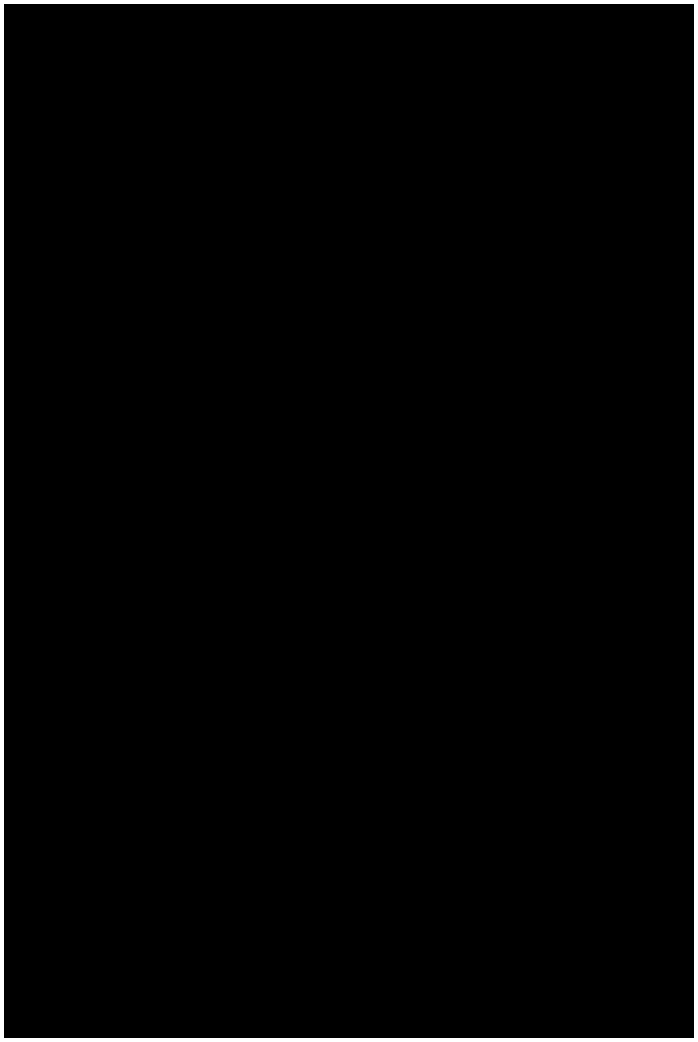
Word 3

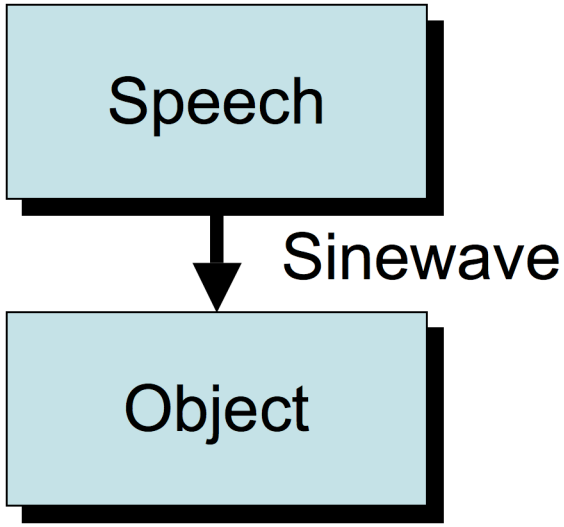


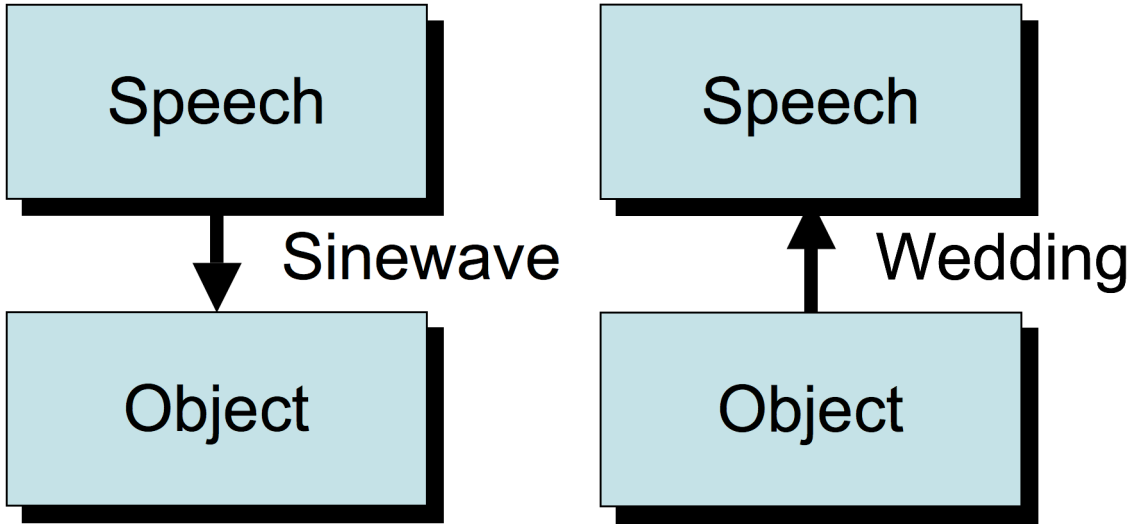


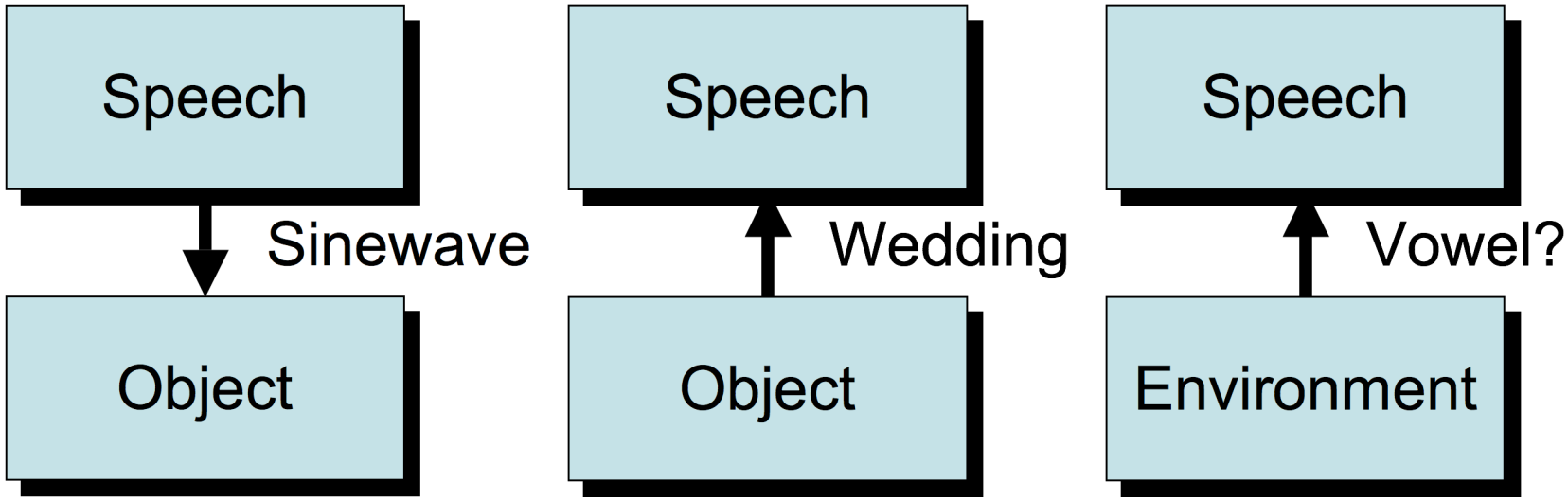


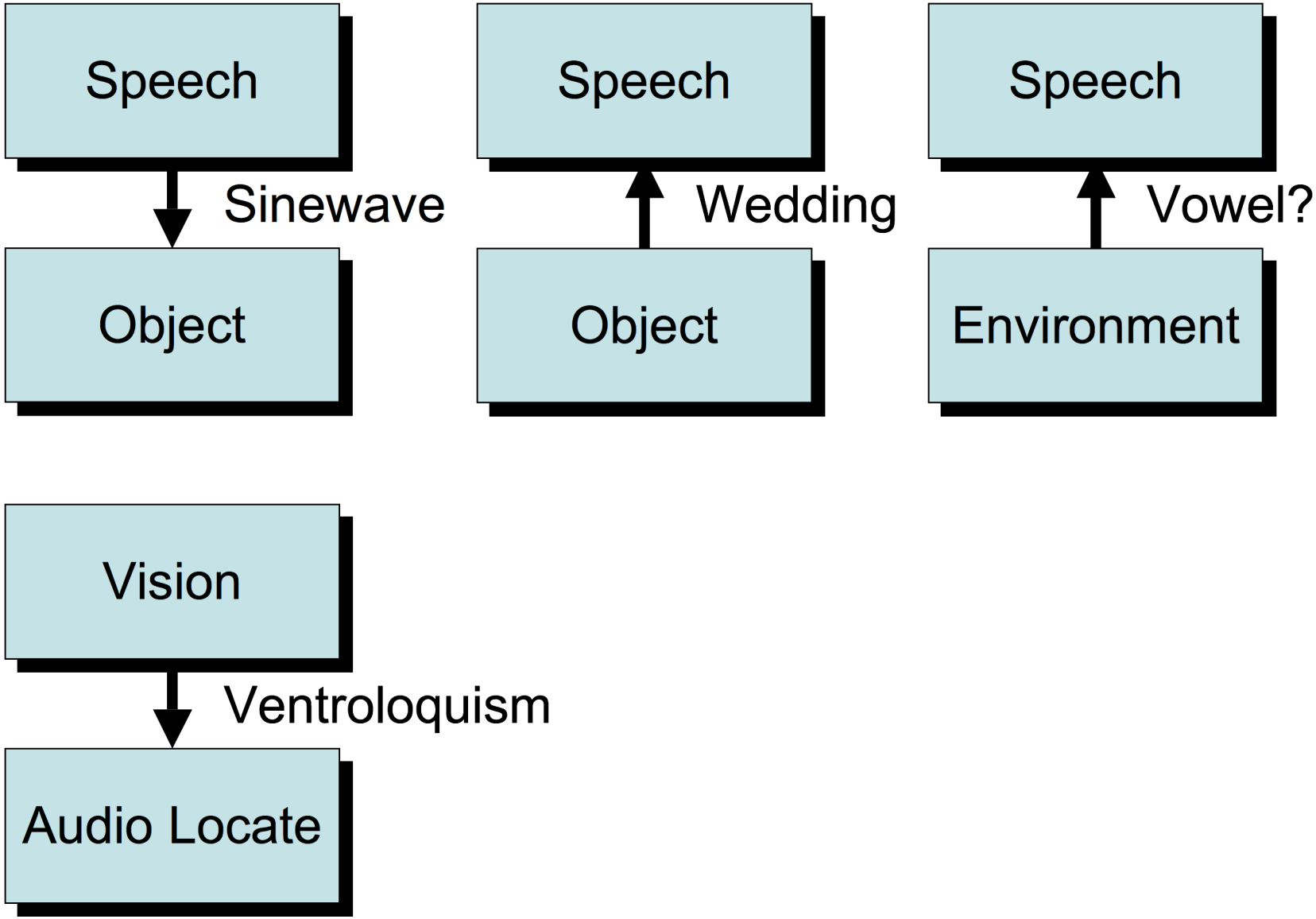
McGurk

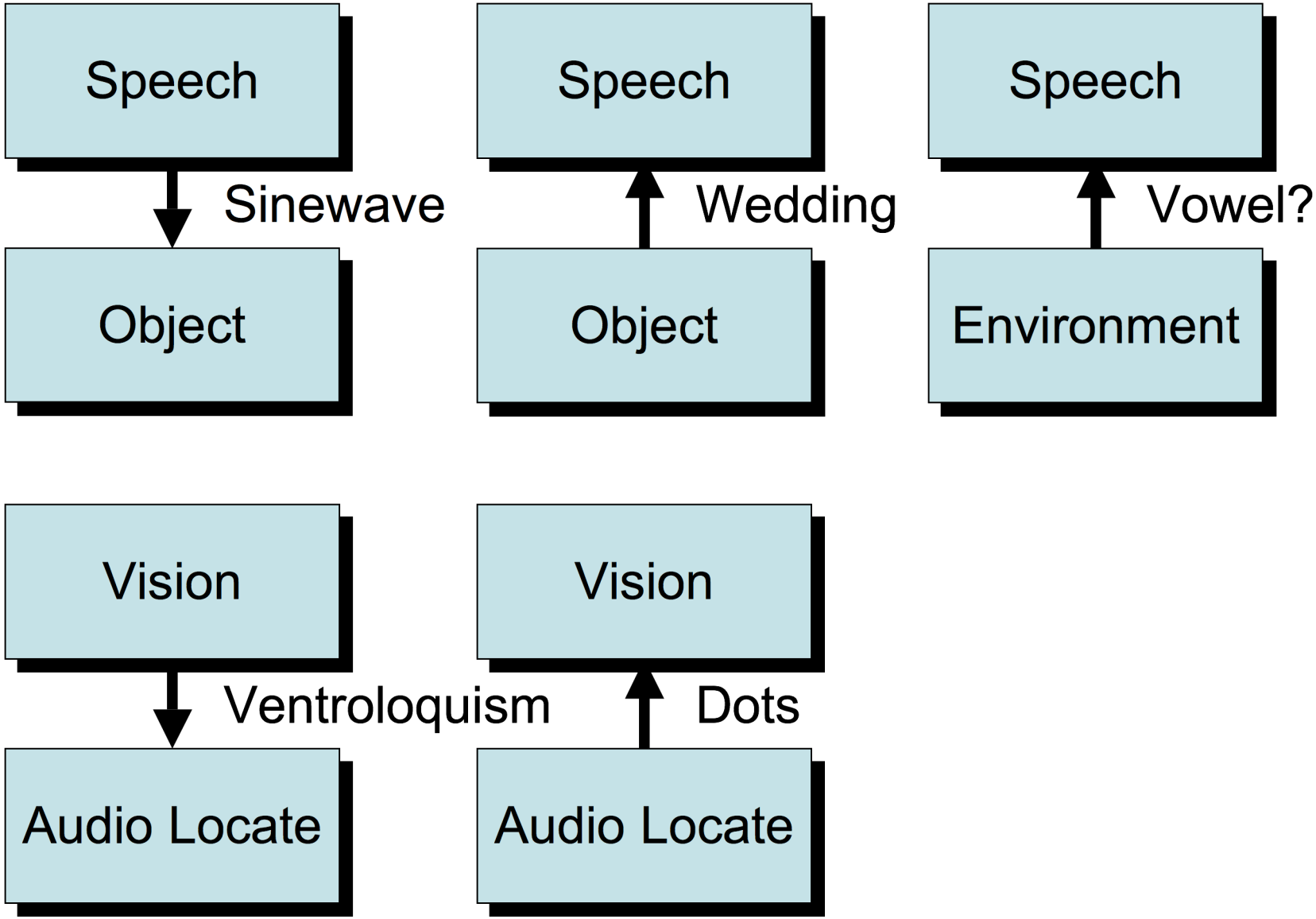


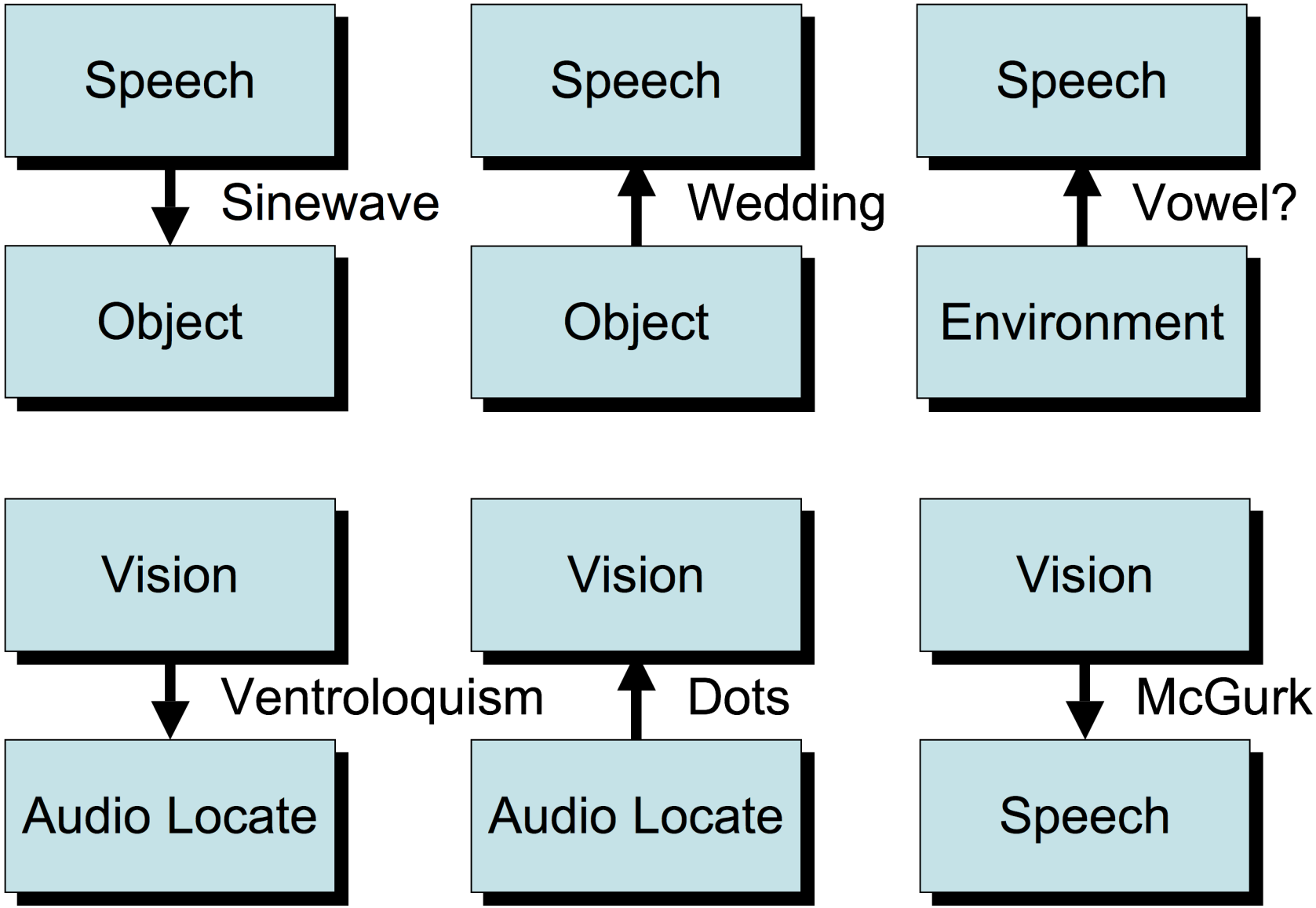

Speech
Sinewave
Object


Speech
Sinewave
Object
Speech
Wedding
Object

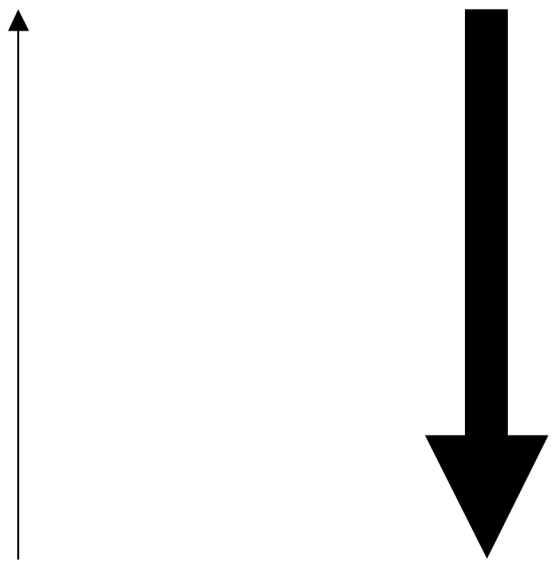

Speech
Sinewave
Object
Speech
Wedding
Object
Speech
Vowel?
Environment


Speech
Sinewave
Object
Speech
Wedding
Object
Speech
Vowel?
Environment
Vision
Ventroloquism
Audio Locate


Speech
Sinewave
Object
Speech
Wedding
Object
Speech
Vowel?
Environment
Vision
Ventroloquism
Audio Locate
Vision
Dots
Audio Locate


Speech
Sinewave
Object
Speech
Wedding
Object
Speech
Vowel?
Environment
Vision
Ventroloquism
Audio Locate
Vision
Dots
Audio Locate
Vision
McGurk
Speech

Information Flow

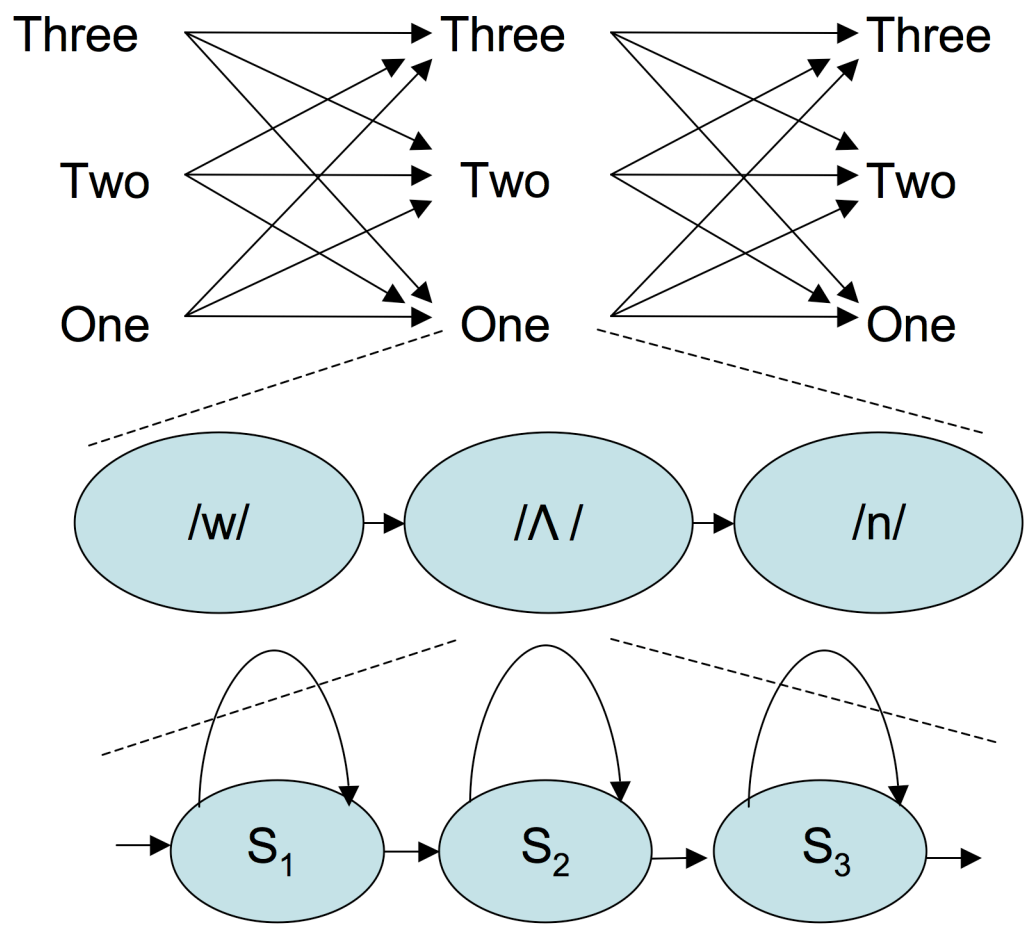


ASR

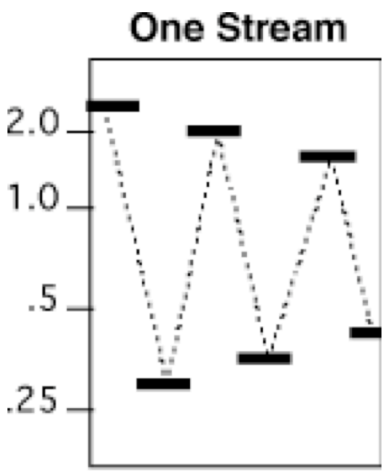
Language model for the words: “one”, “two”, “three”

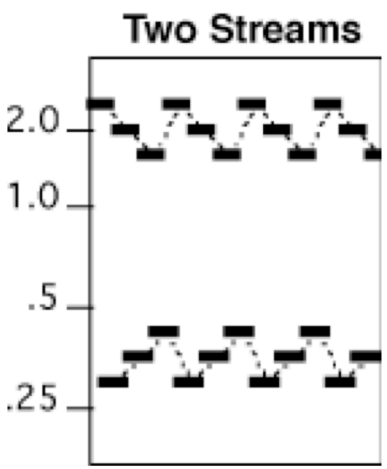
Word model showing phonemes for the word one

Acoustic (phoneme) model for the phoneme /Λ /



Streaming of Tones





Auditory Scene Analysis

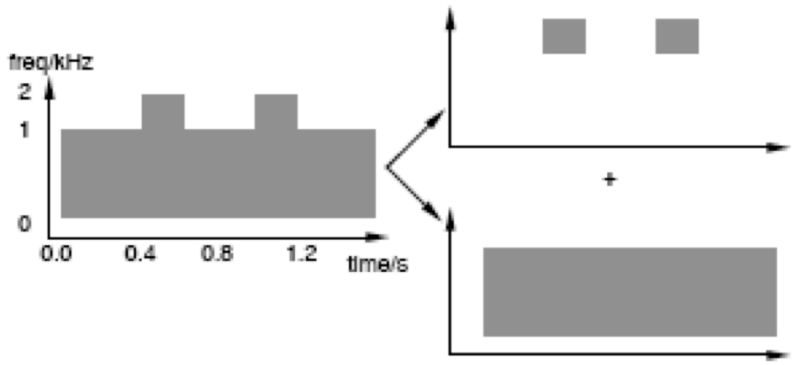
All harmonics - Oboe

Even harmonics - Soprano
Odd harmonics - Clarinet

Old plus New Principle

The effect of context

- **Context can create an 'expectation': i.e. a bias towards a particular interpretation**
- **e.g. Bregman's "old-plus-new" principle:**
  A change in a signal will be interpreted as an *added* source whenever possible



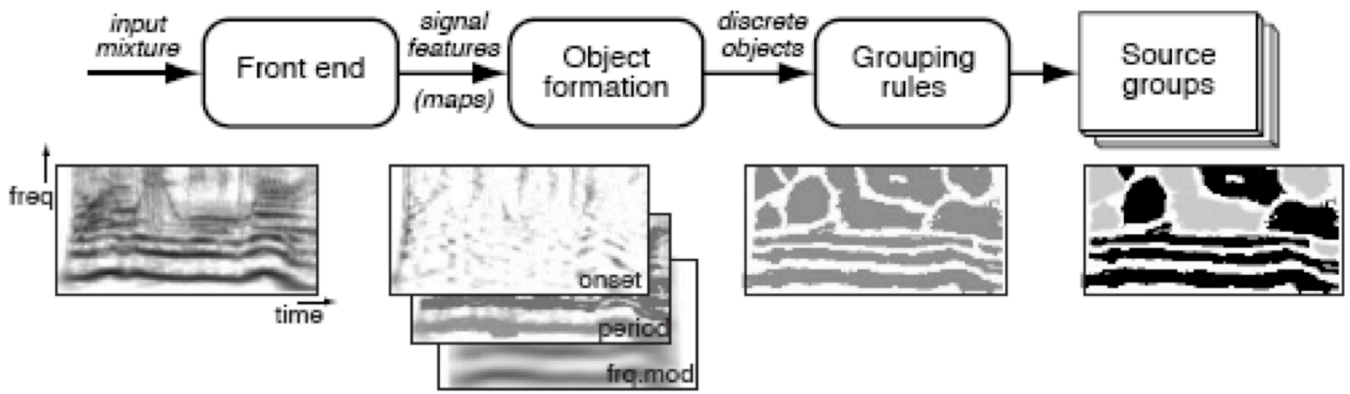
- a different division of the same energy depending on what preceded it

Slide by Dan Ellis (Columbia

Conventional Scene Analysis

Computer models of grouping

- "Bregman at face value" (e.g. Brown 1992):

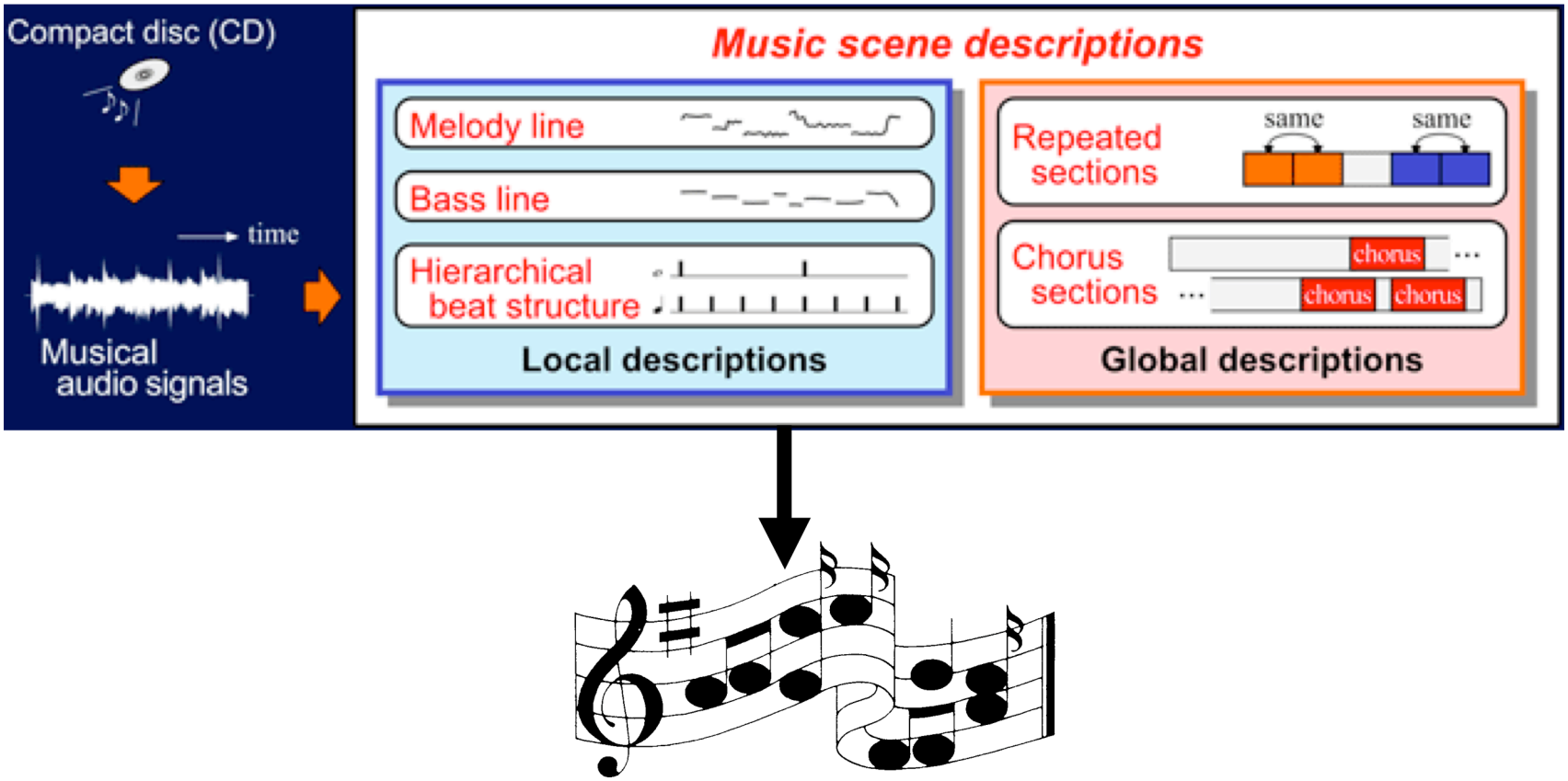


- feature maps
- periodicity cue
- common-onset boost
- resynthesis

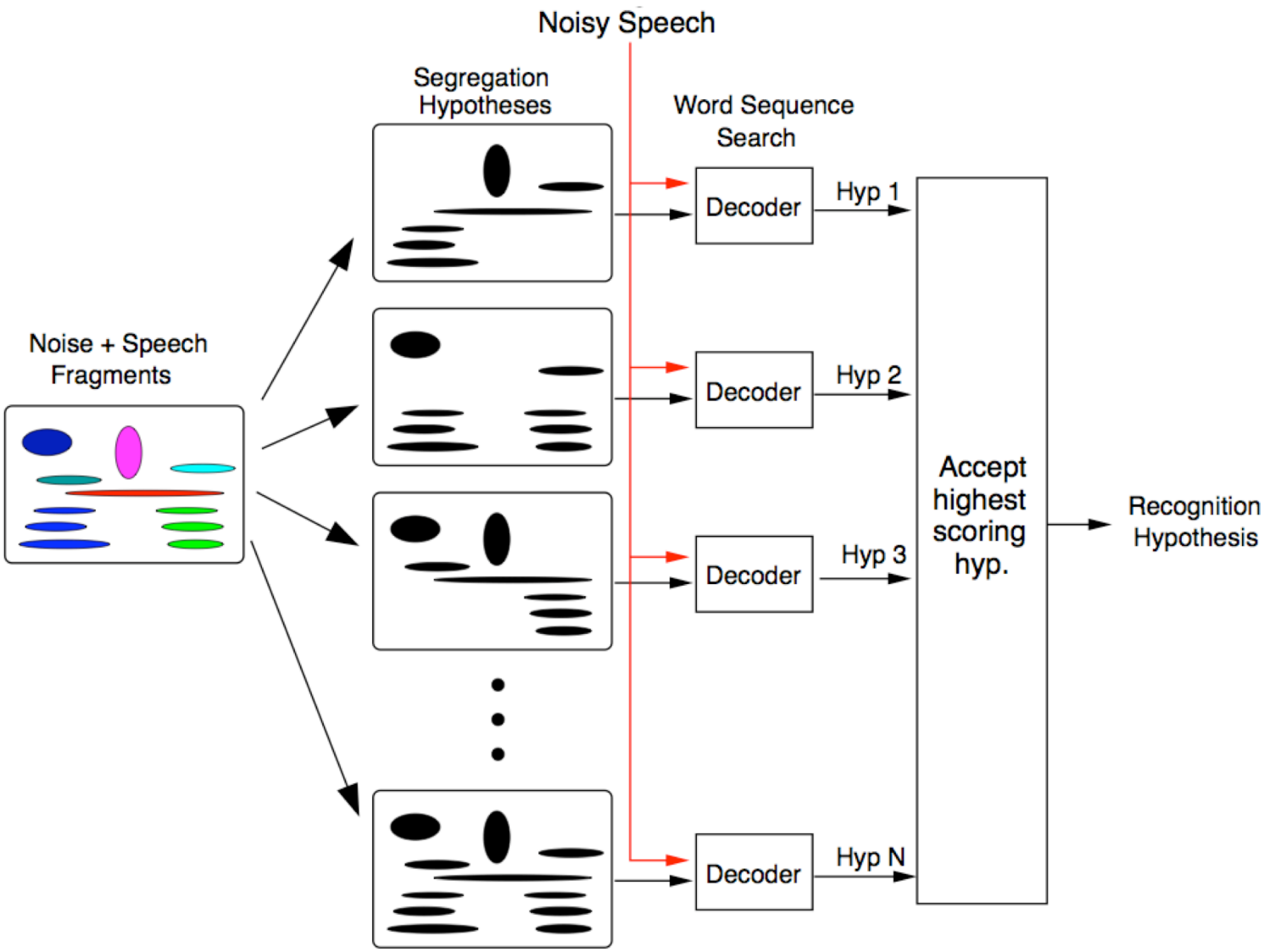
Slide by Dan Ellis (Columbia

Goto—CASA with MIDI

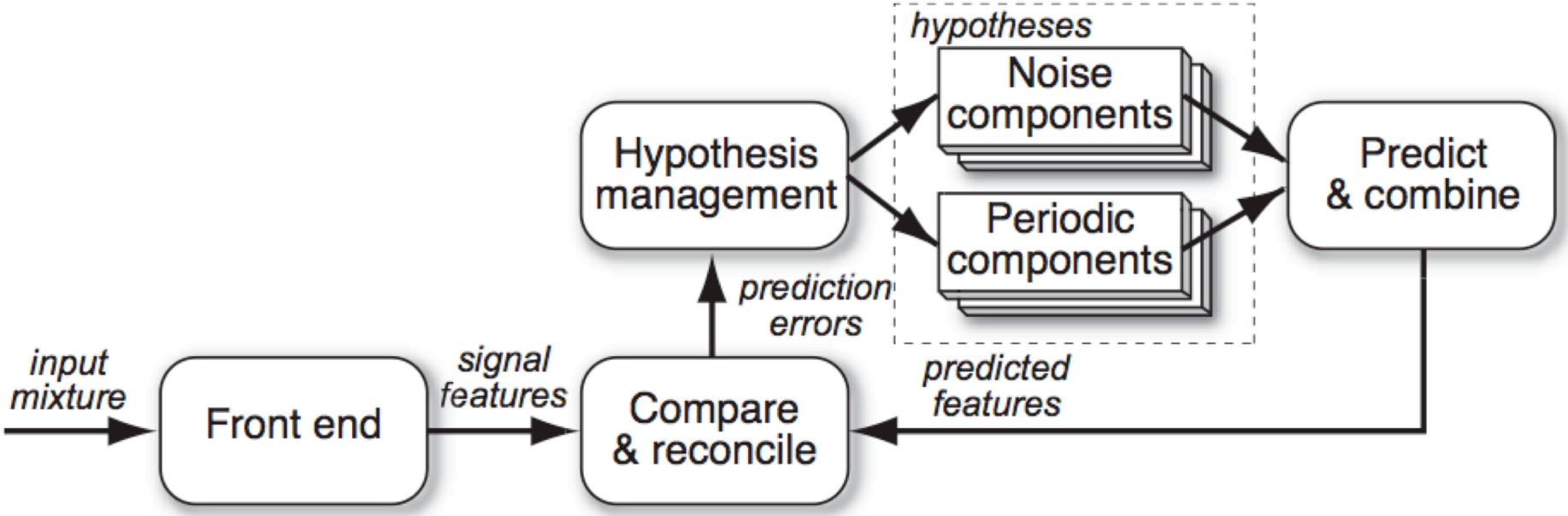
MIDI Sequence



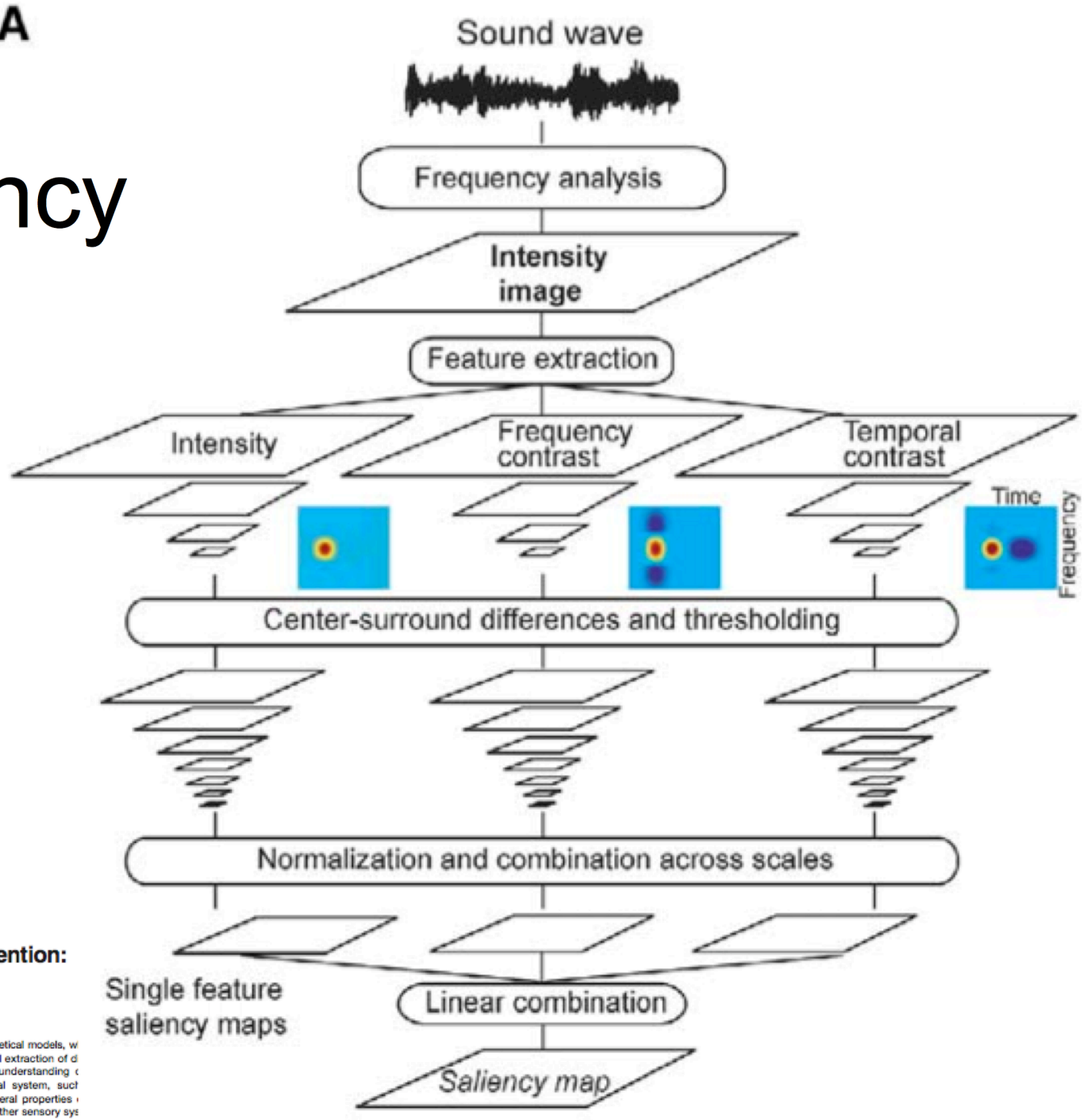
Barker—ASR


Noisy Speech
Segregation Hypotheses
Word Sequence Search
Noise + Speech Fragments
Decoder
Hyp 1
Decoder
Hyp 2
Decoder
Hyp 3
Decoder
Hyp N
Accept highest scoring hyp.
Recognition Hypothesis

Ellis—Prediction Driven



Saliency



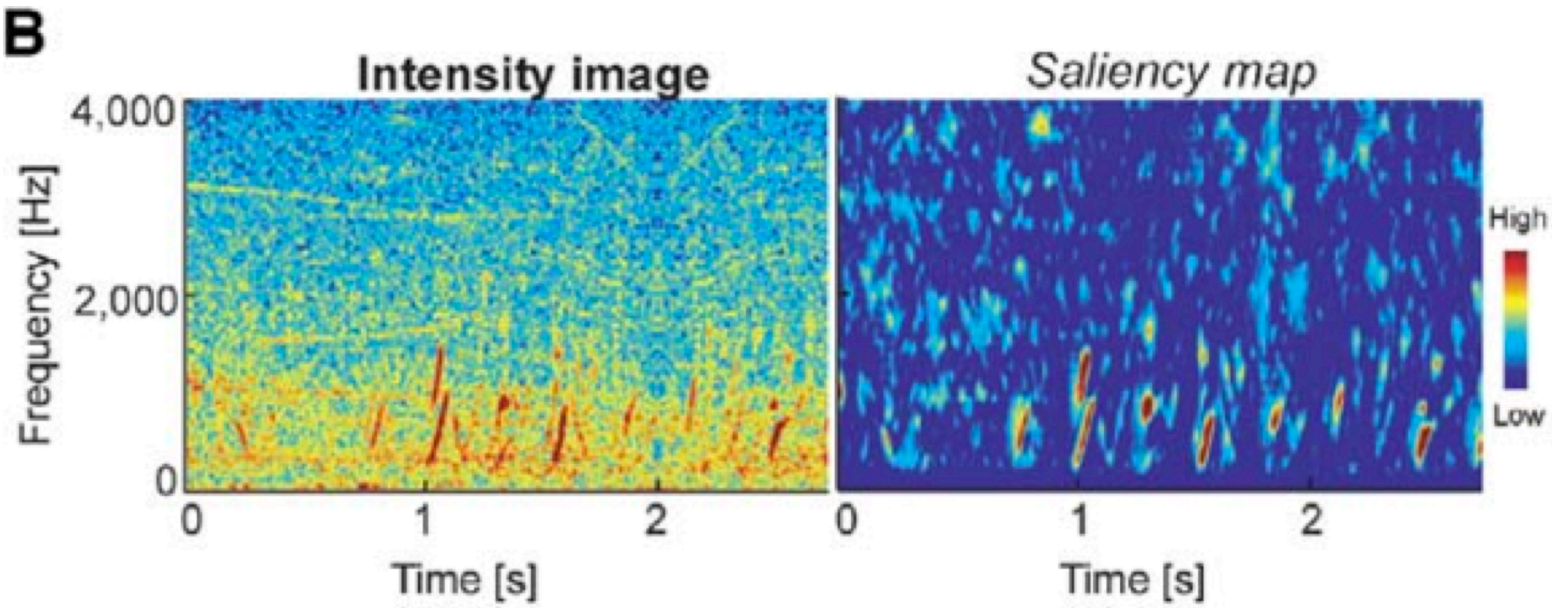
Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map

Christoph Kayser,* Christopher I. Petkov,
Michael Lippert, and Nikos K. Logothetis
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen
Germany

conceptualized as theoretical models, w
hierarchical and parallel extraction of d
and build on existing understanding c
cessing. For the visual system, suc
shown to replicate several properties
attention [2, 8, 9]. For other sensory sys

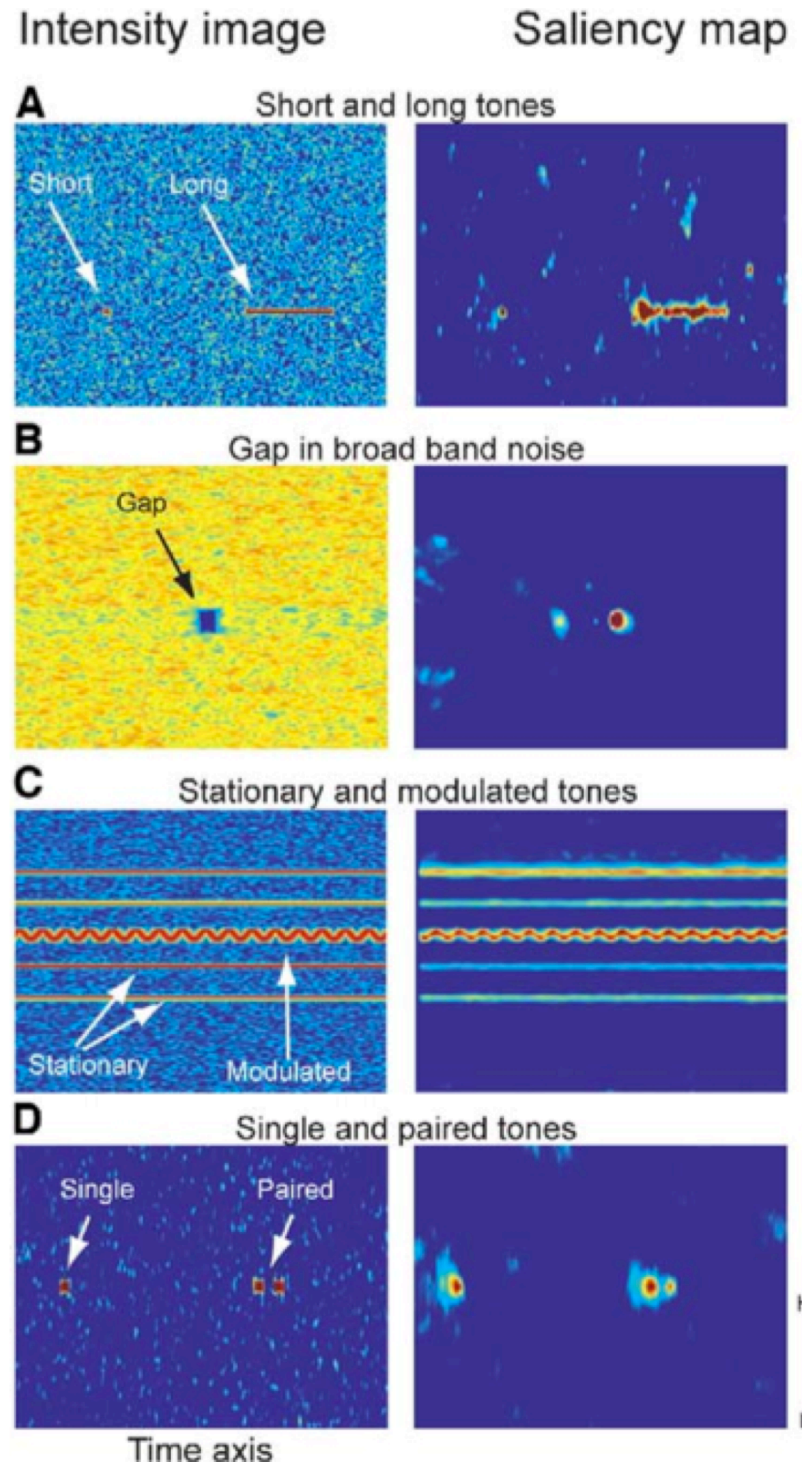
Saliency Example

- Time-frequency display
- Saliency map shows high-interest locations



Saliency Maps

- Longer tones better
- Missing parts salient
- Modulation more salient
- Forward masking works

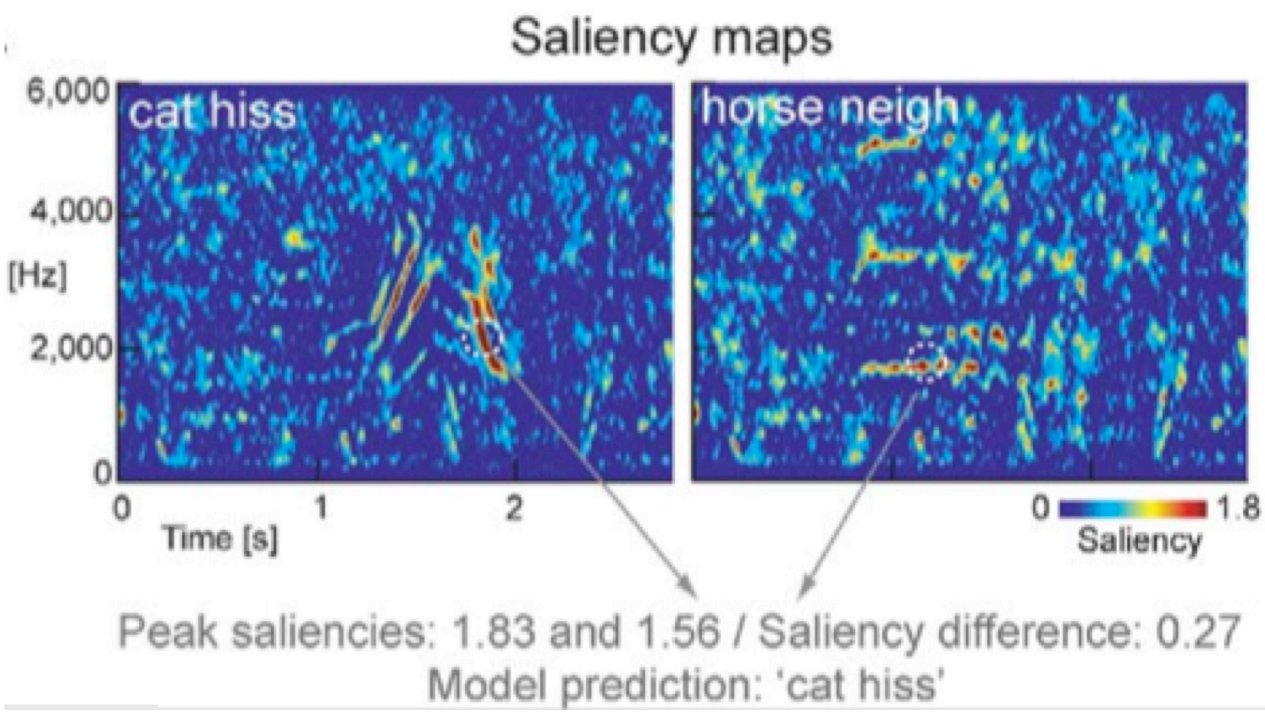


Sound Examples

- Birds
- Calls
- Cows
- Horse
- Waterfall

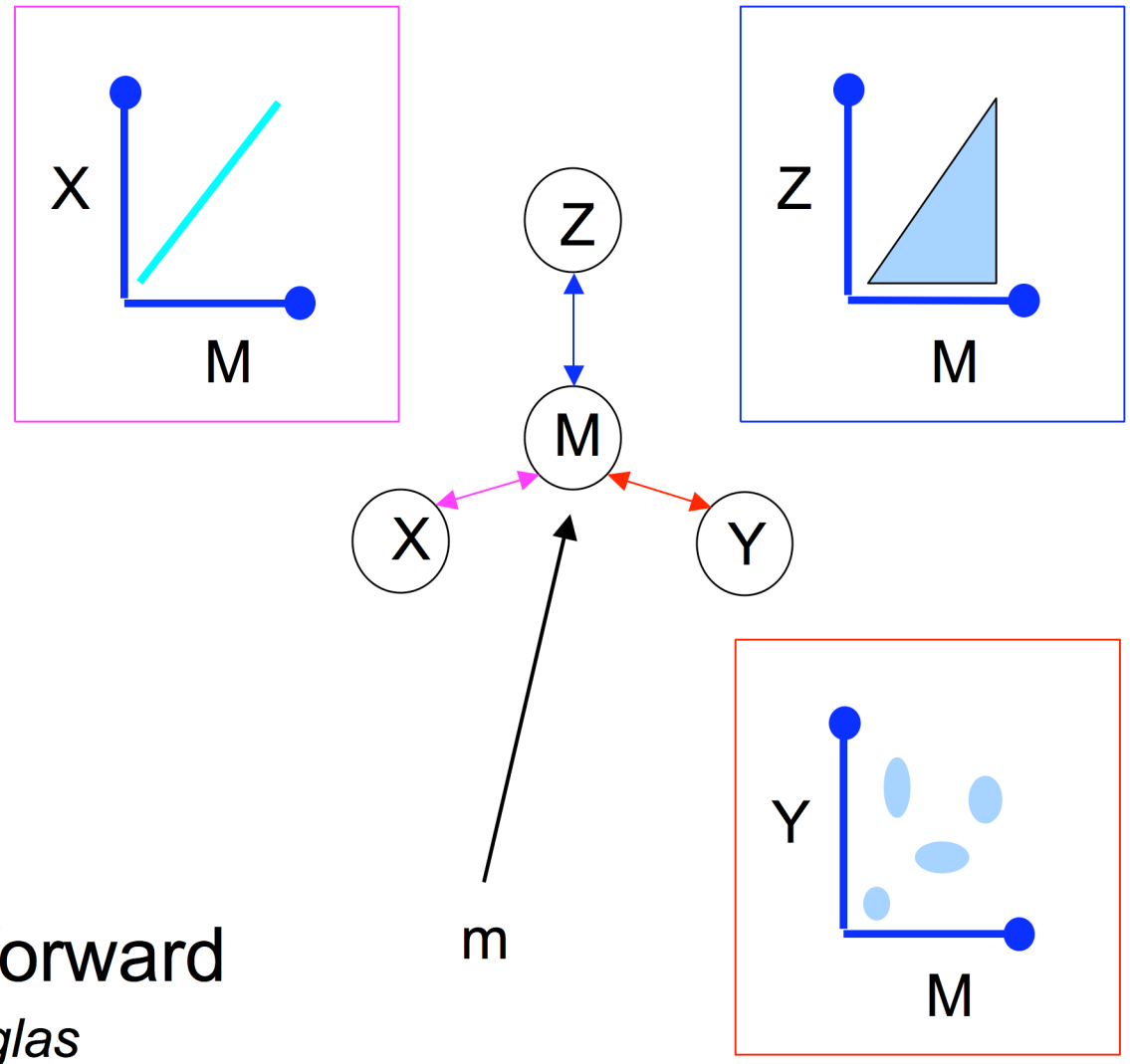
Saliency Comparison

- Details of saliency comparison
- Model predictions



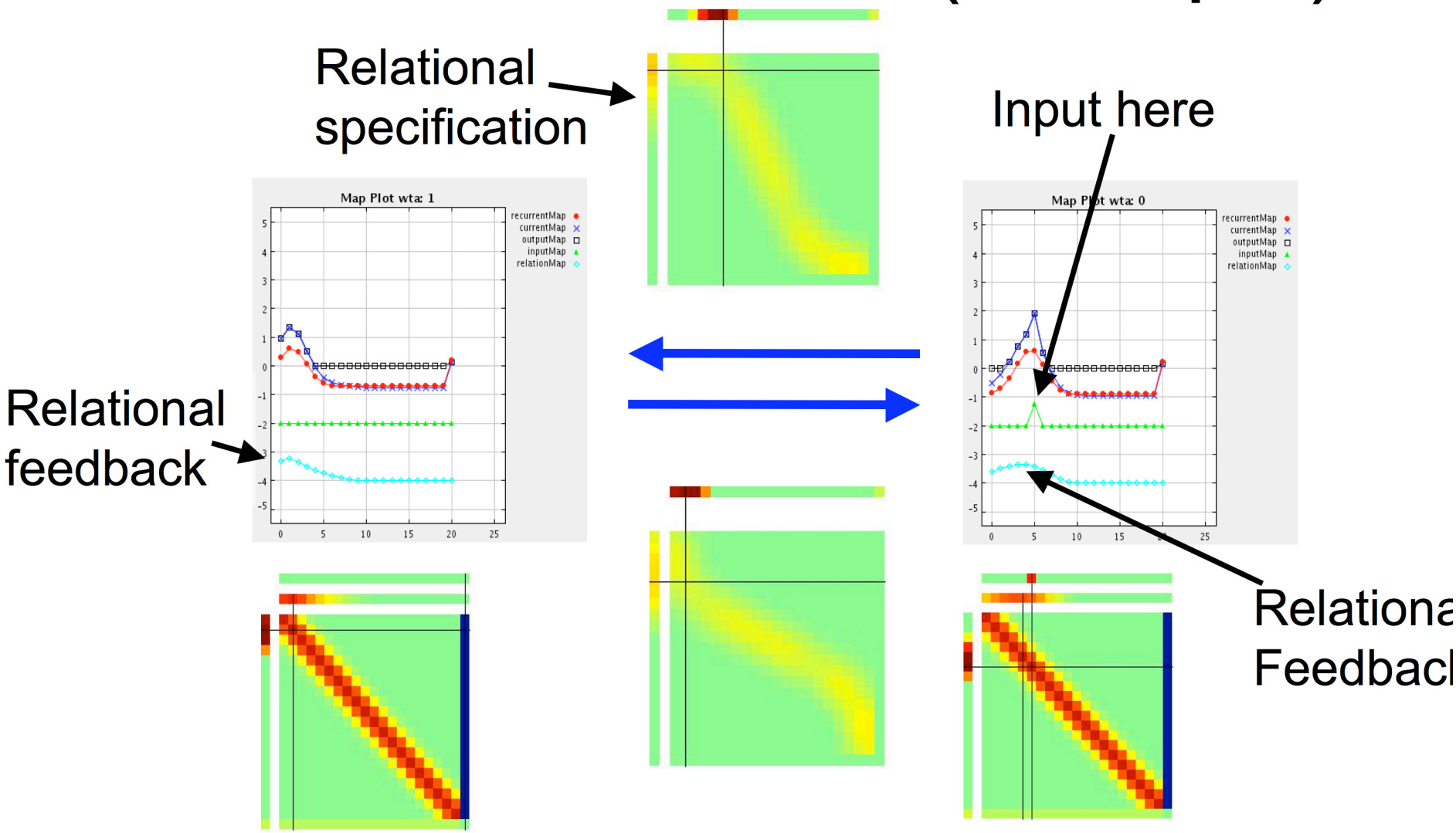
Relational Network (Simple)

- Patches of neurons
- Each measure one quantity
- Bidirectional relations for feedback/feedforward



Thanks to Rodney Douglas

Relational Network (example)



ASR Relational Network

Bidirectional links enforce phoneme/word constraints

Cochlea

Phone Recognizer

Word Recognizer

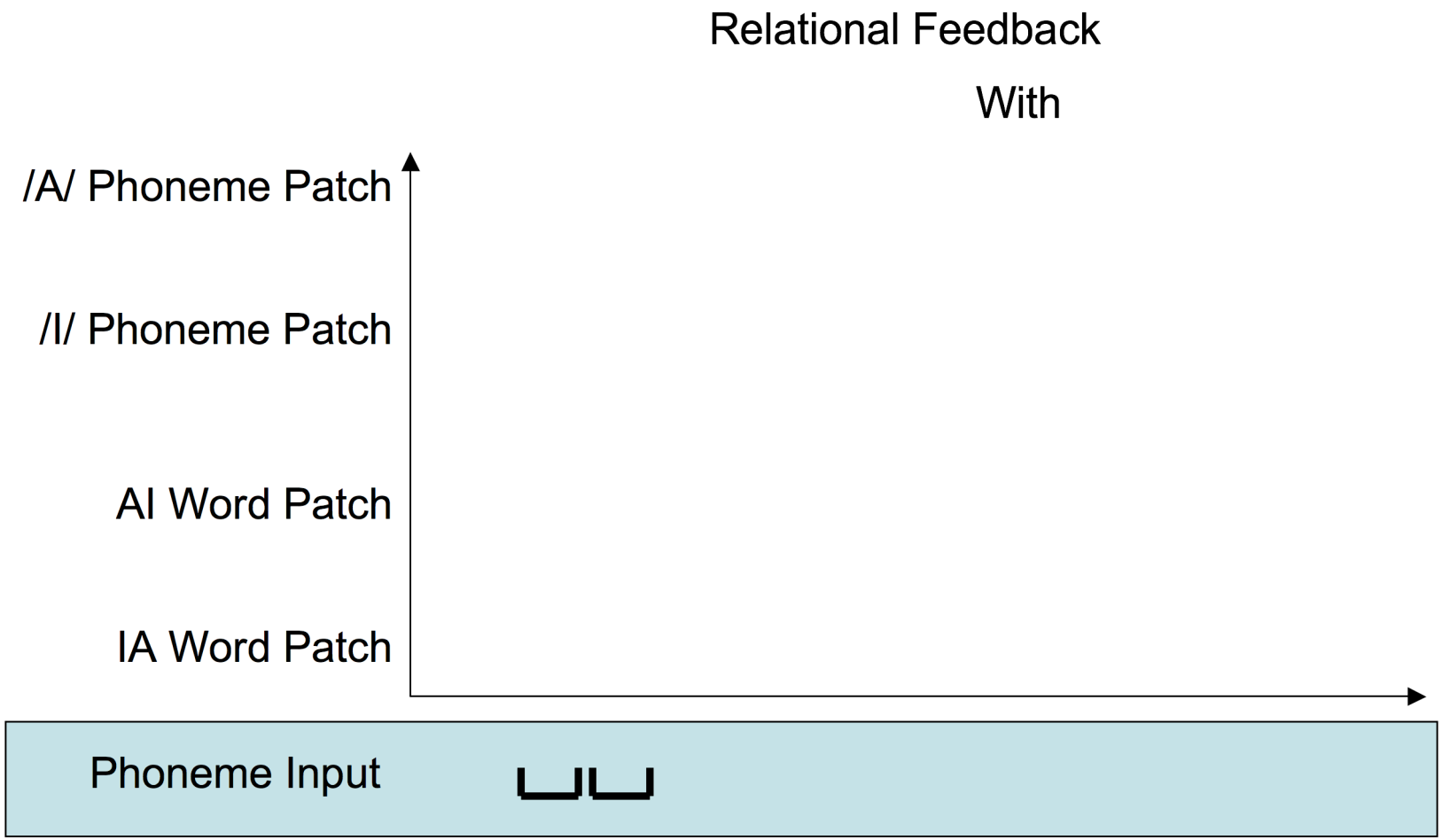
Delay

Phone Recognizer

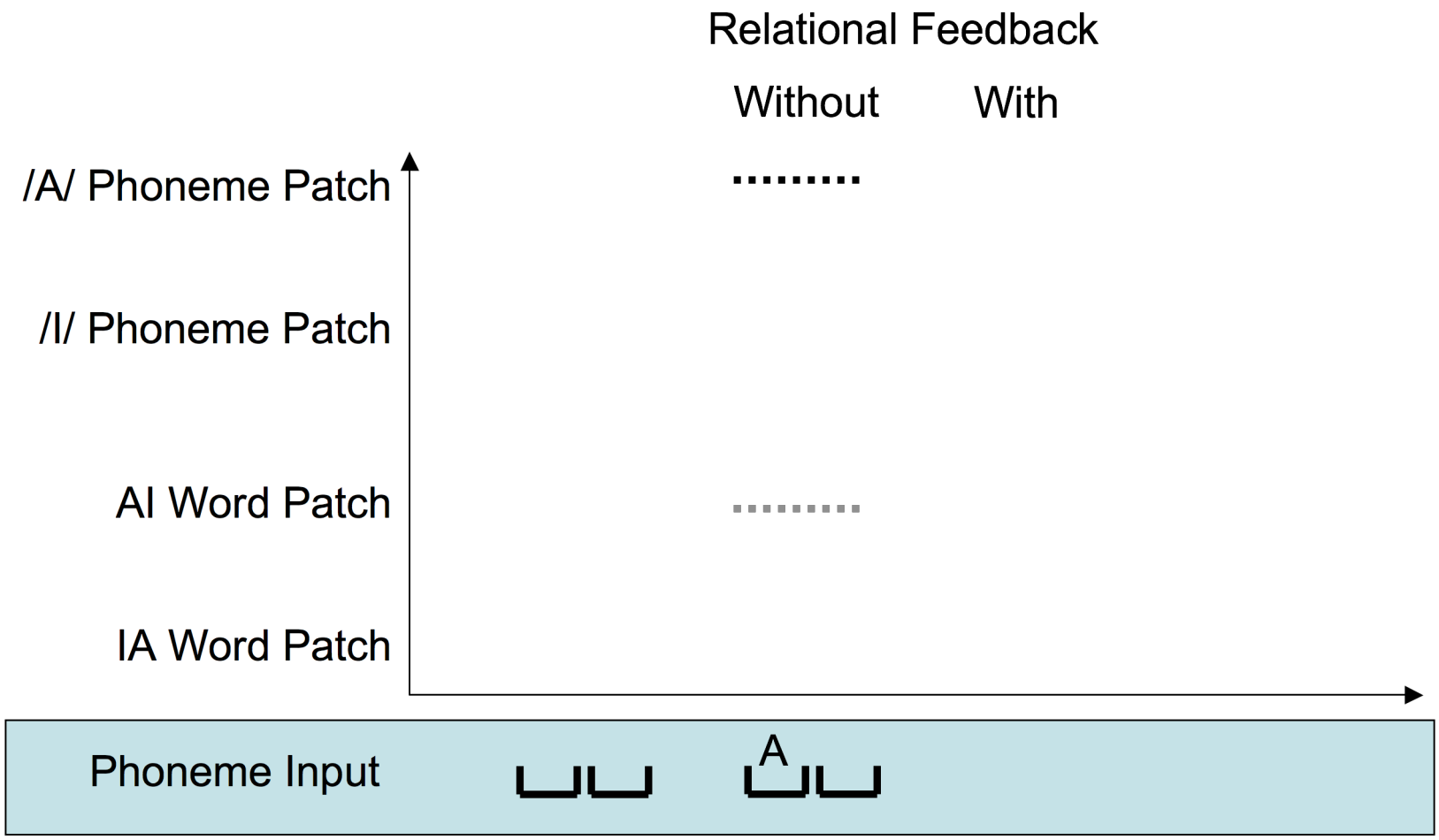
Note: We don't know how to represent delays

A patch of neurons (one of N output)

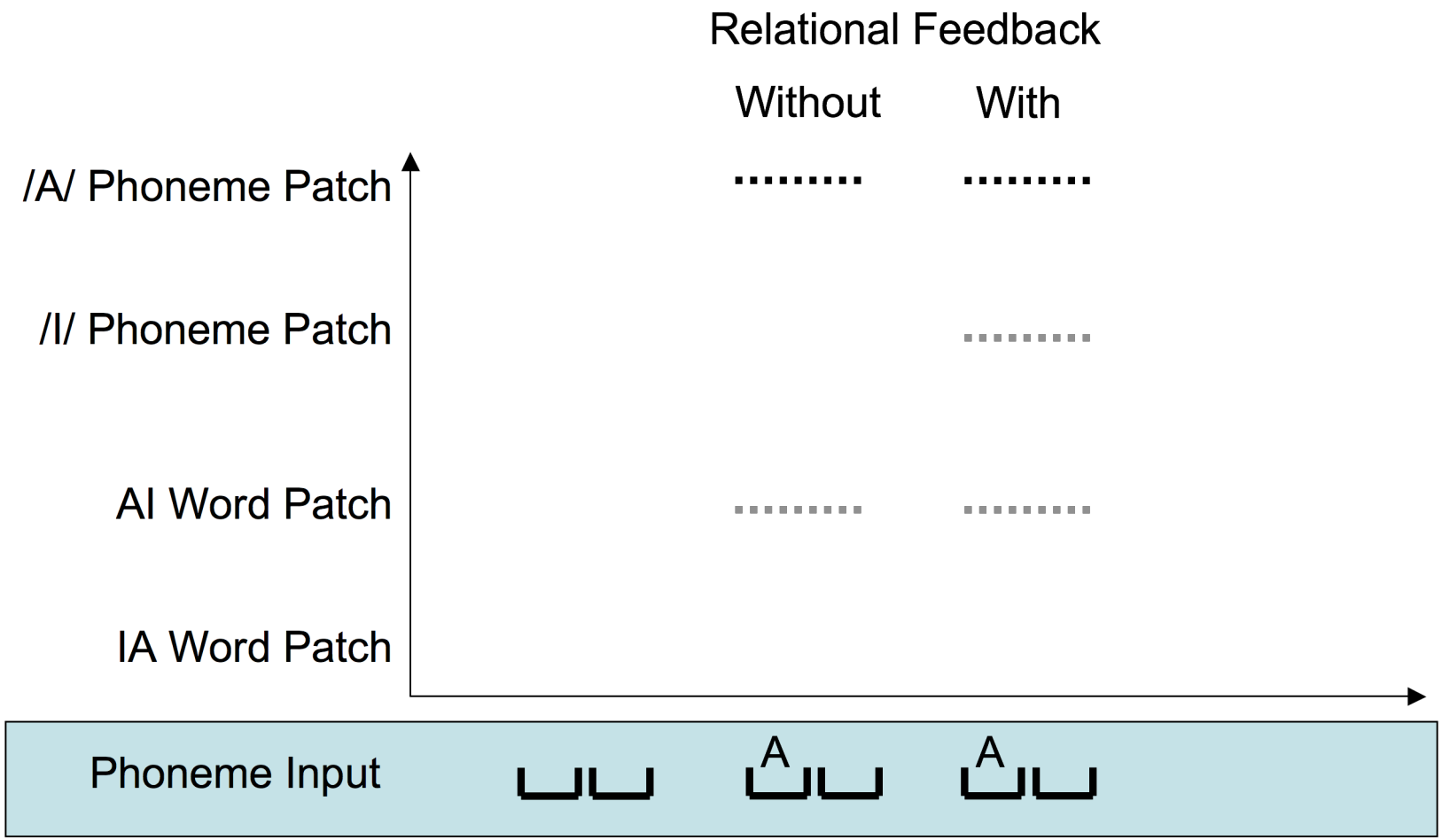
Desired Results



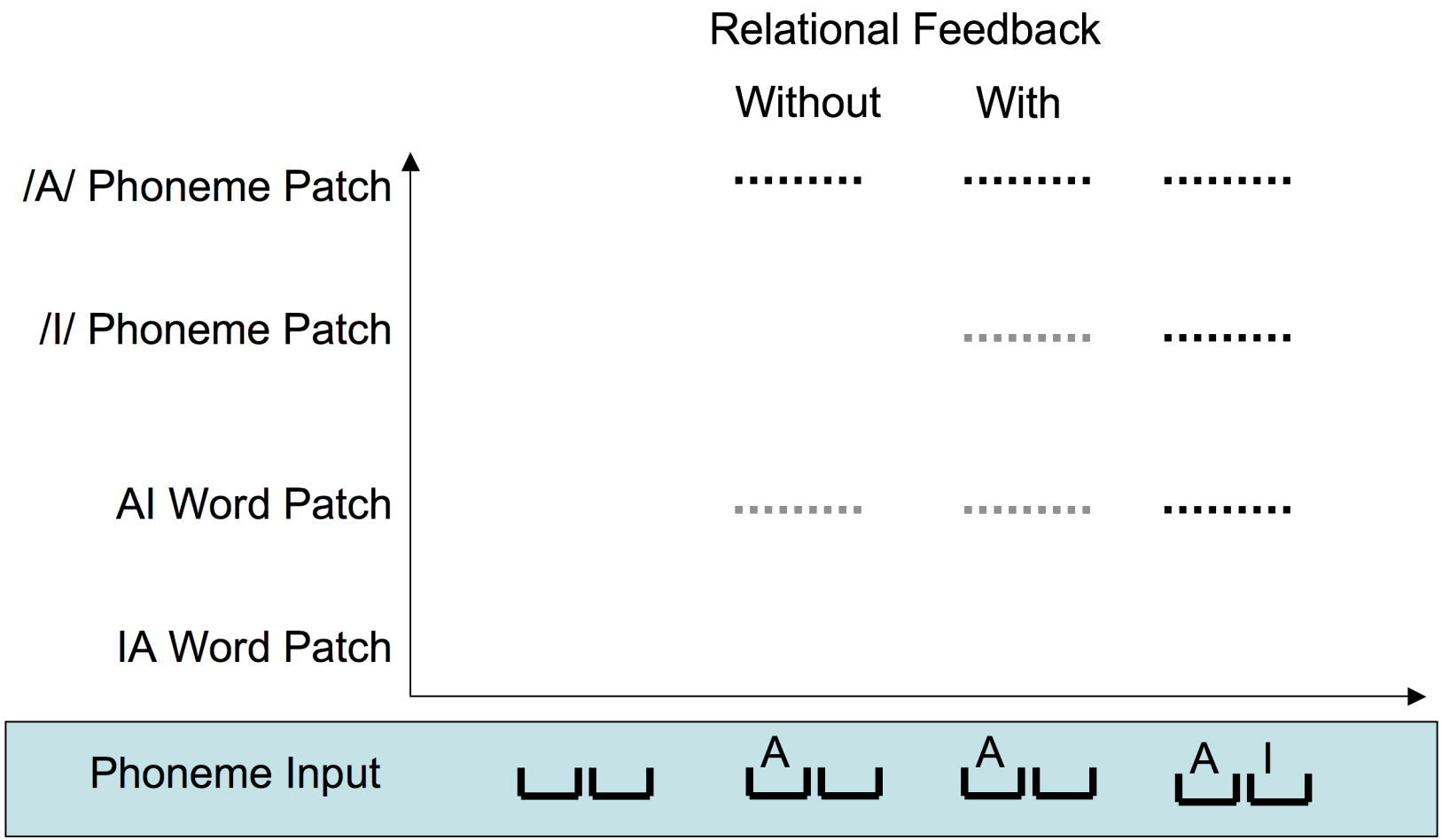
Desired Results



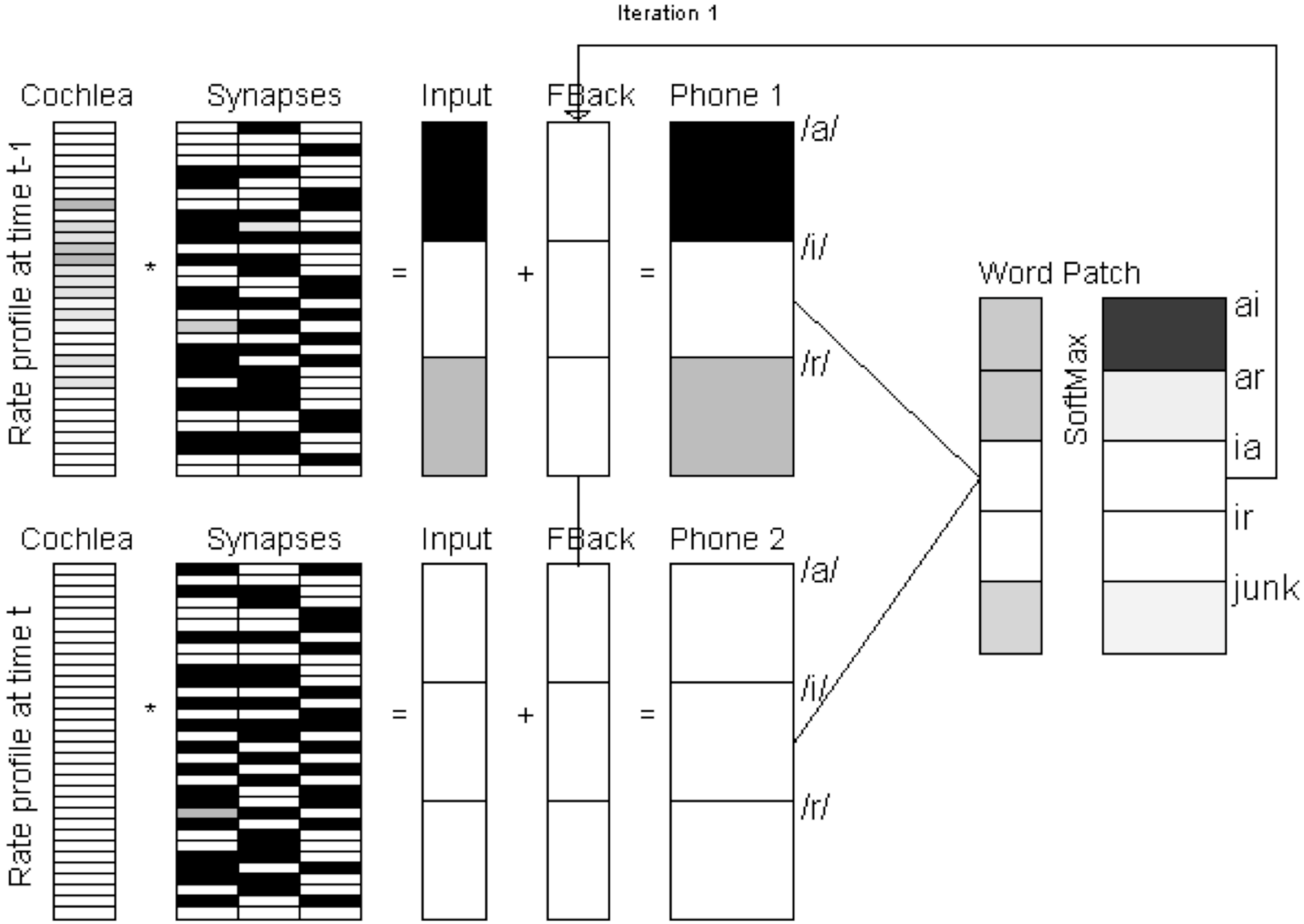
Desired Results



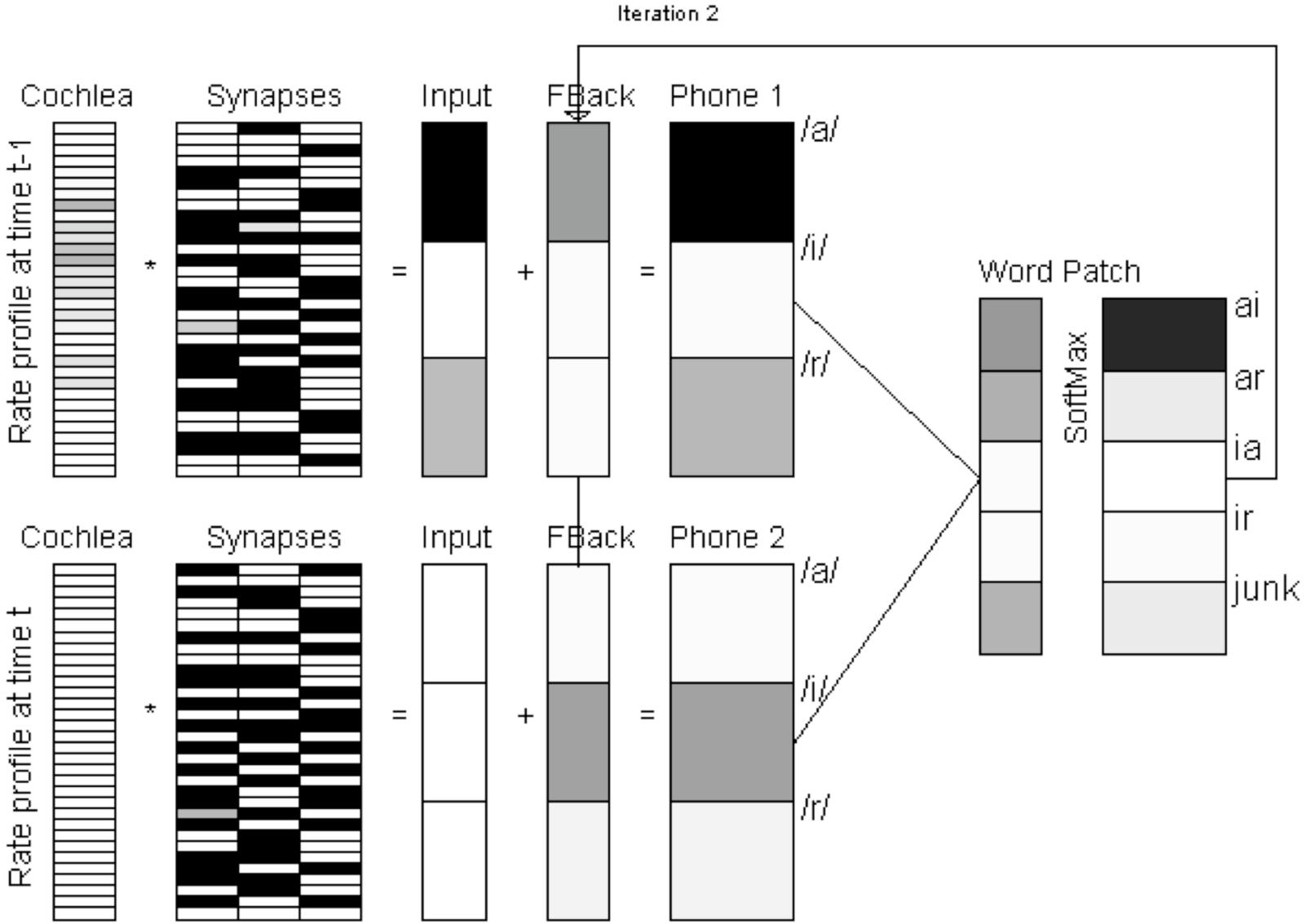
Desired Results



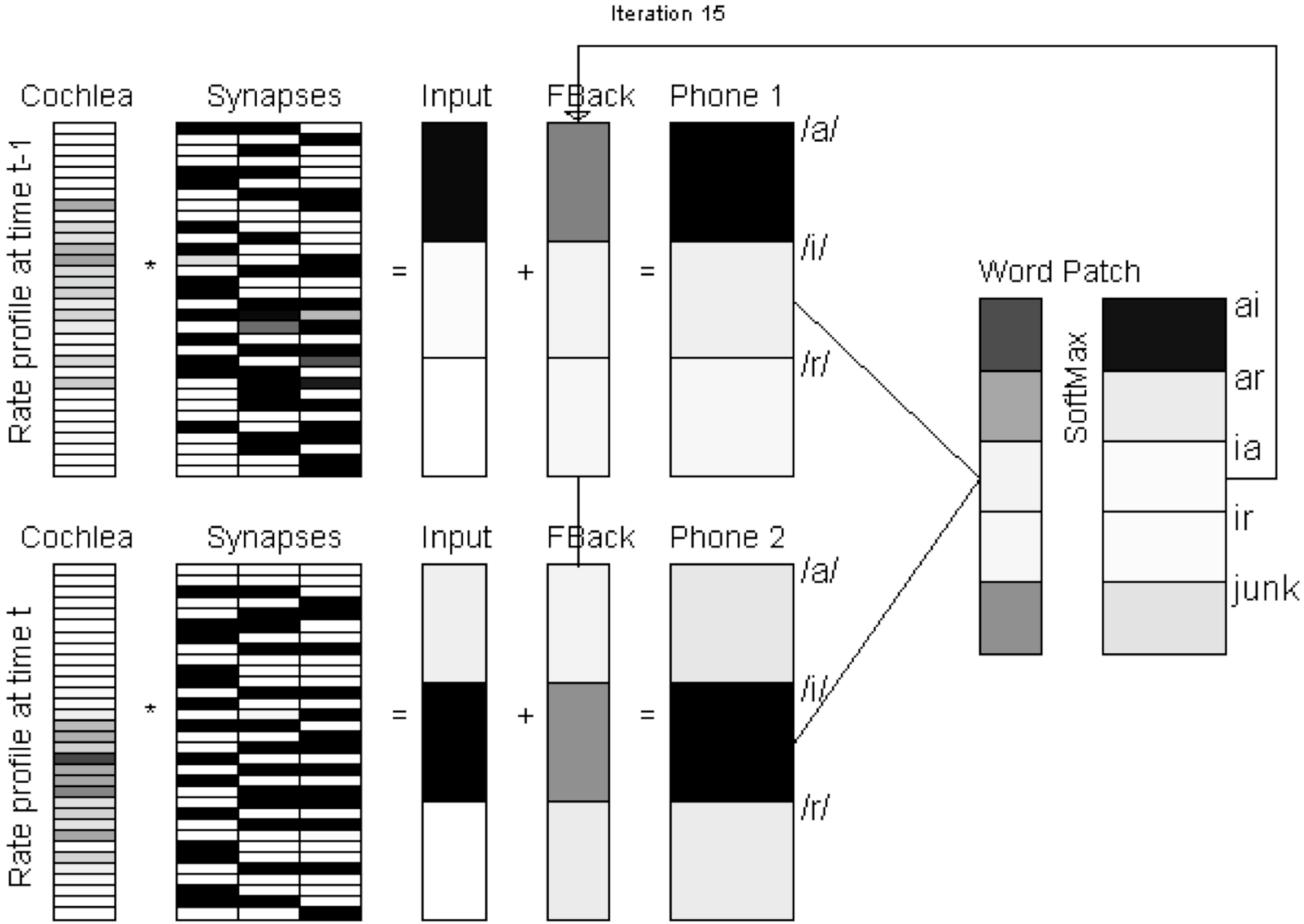
Simulation


Iteration 1
Cochlea
Synapses
Input
FBack
Phone 1
/a/
/i/
/r/
Rate profile at time t-1
*
=
+
=
Word Patch
SoftMax
ai
ar
ia
ir
junk
Cochlea
Synapses
Input
FBack
Phone 2
/a/
/i/
/r/
Rate profile at time t
*
=
+
=

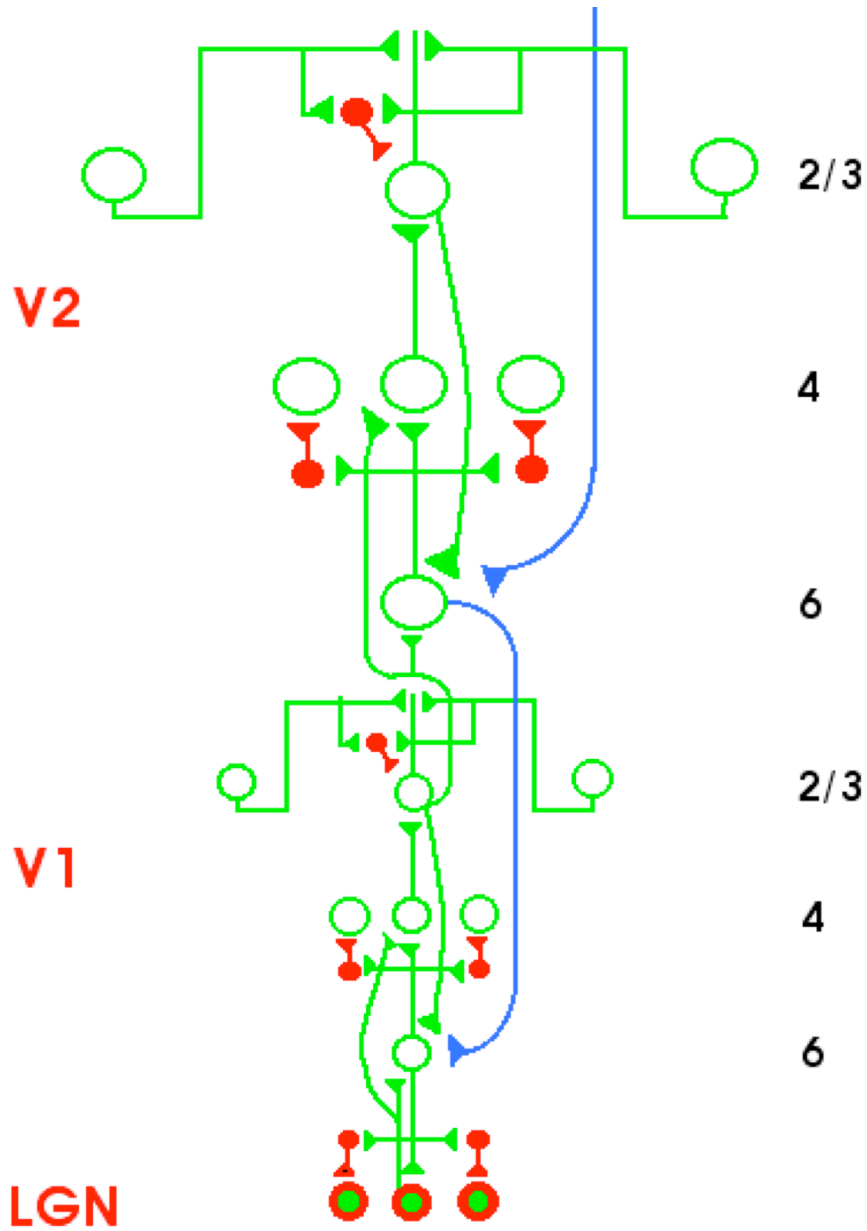
Simulation 2



Simulation 3



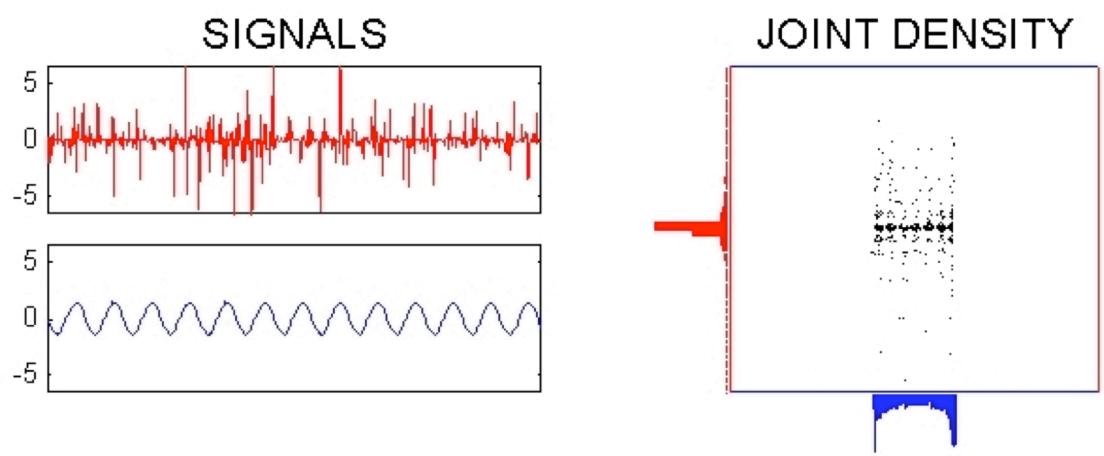
Grossberg—ART



The LAMINART model. From Grossberg, S. (1999). How does the cerebral cortex work? Learning, attention and grouping by the laminar circuits of visual cortex. *Spatial Vision*, **12**, 163-186

Statistical Means

- ICA
  - Different distributions
- One Microphone
  - GMM models of distribution



Separated signals after 4 steps of FastICA

Semantic-Audio Retrieval

Slaney 2003

Semantic-Audio Retrieval

- Predicted words:
  - anim (.108), hors (.105), left (.086), trot (.065), approach (.059), track (.047), walk (.040), depart(.037)

Slaney 2003

Semantic-Audio Retrieval

- Predicted words:
  - anim (.108), hors (.105), left (.086), trot (.065), approach (.059), track (.047), walk (.040), depart(.037)
- True label:
  - animals: horses, two horses trot past on rough track left to right

Slaney 2003

Semantic-Audio Retrieval

- Predicted words:
  - anim (.108), hors (.105), left (.086), trot (.065), approach (.059), track (.047), walk (.040), depart(.037)
- True label:
  - animals: horses, two horses trot past on rough track left to right

- Predicted words:
  - bird (.11), ambienc (.107), jungl (.104), morn (.094), Africa (.093), anim (.054), bark (.029), dog (.020), cricket (.018)

Slaney 2003

Semantic-Audio Retrieval

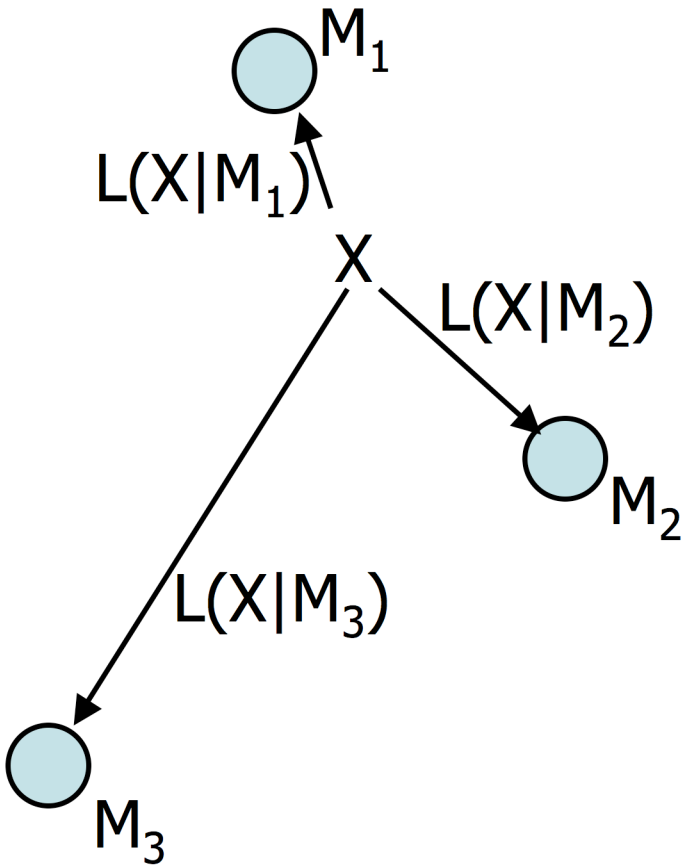
- Predicted words:
  - anim (.108), hors (.105), left (.086), trot (.065), approach (.059), track (.047), walk (.040), depart(.037)
- True label:
  - animals: horses, two horses trot past on rough track left to right

- Predicted words:
  - bird (.11), ambienc (.107), jungl (.104), morn (.094), Africa (.093), anim (.054), bark (.029), dog (.020), cricket (.018)
- True label:
  - jungle, Africa, Africa: morning ambience, birds

Slaney 2003

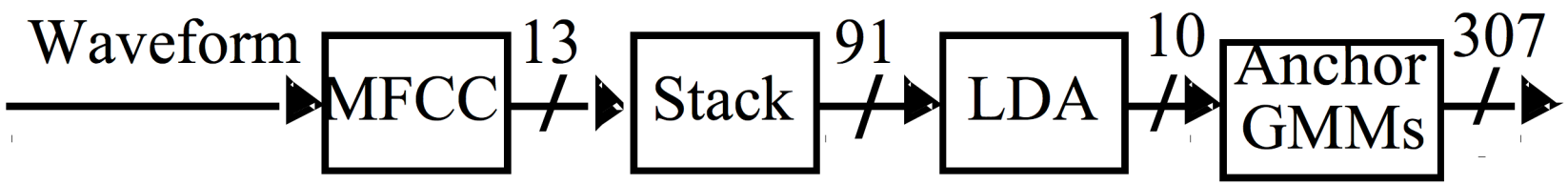
Audio Anchors

- Process
  - Build models for known sounds (GMM/HMM)
  - Measure likelihood of new sound
  - Form vector
- Advantages
  - Summarizes entire sound
  - Agglomerative clustering using distances



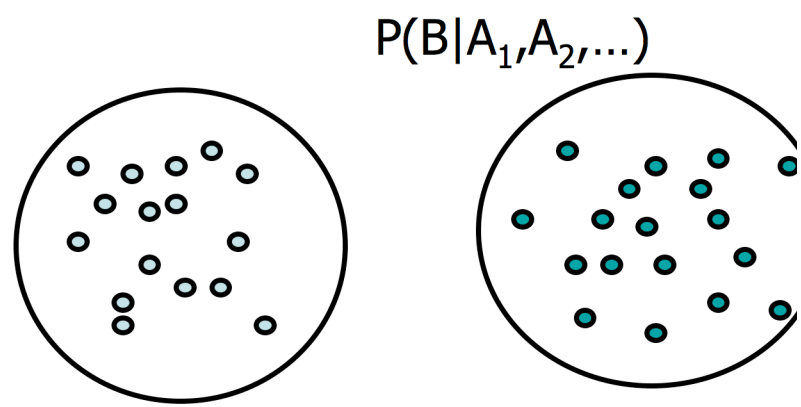
Audio Features

- MFCC/Rasta at 100Hz
  - 13 cepstral coefficients per frame
- Stack frames across time
  - +/- 3 frames
- Linear discriminant analysis (LDA)
  - 10 classes guide best feature decision



Model Clustering Algorithm

- E-Step
  - Build model for all signals in cluster
  - GMM (ignores time dependence)
  - HMM (considers time dependence)
- M-Step (for each signal)
  - Measure likelihood for all models
  - Assign to model with highest likelihood



Multinomial Distribution

- Bag of words model
- Likelihood of document given model

$$L = \prod w_i^{n_i}$$

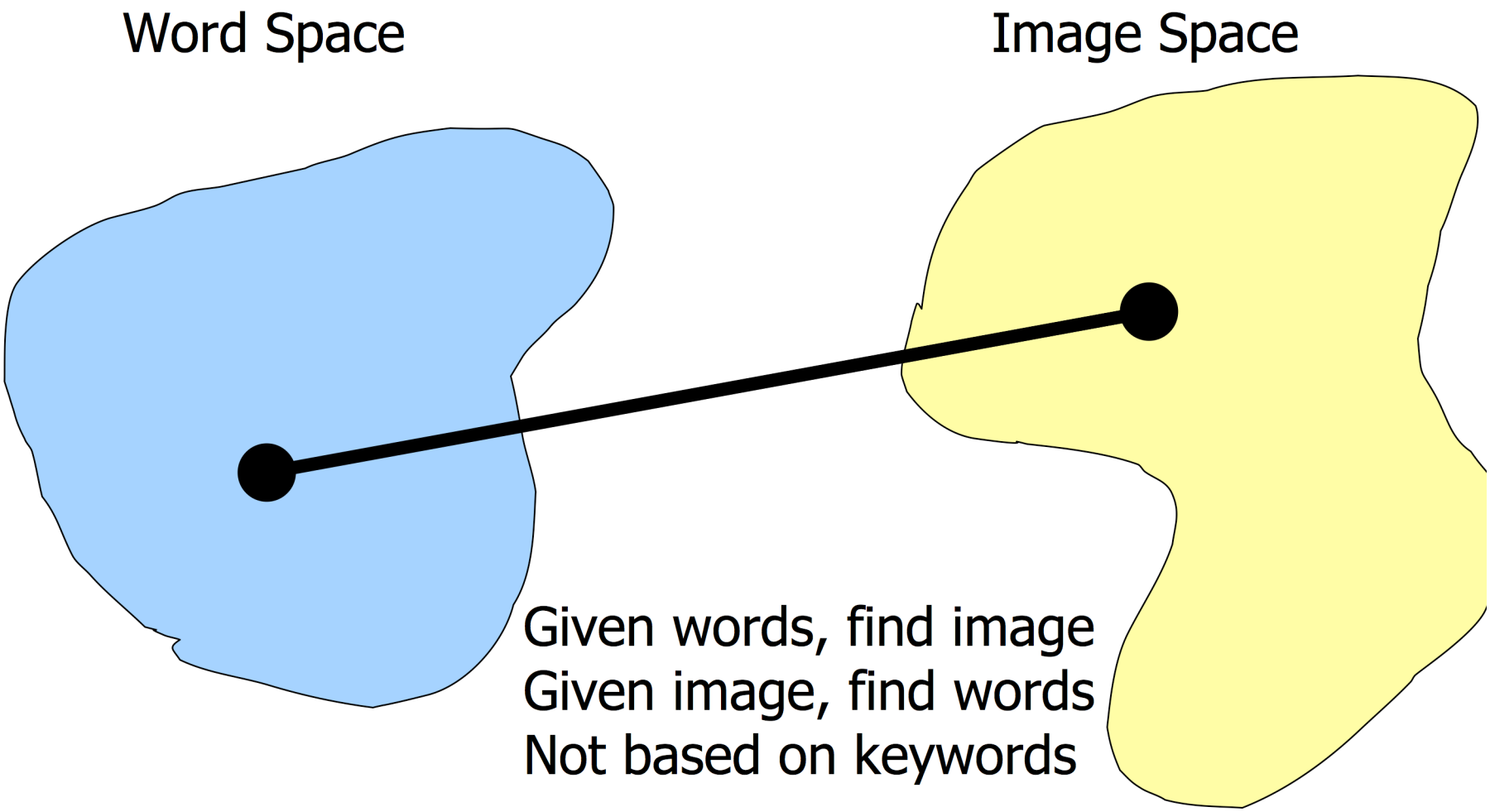
- w_i is probability of word i
- n_i is count of word i
- L is likelihood of document

Semantic Clustering

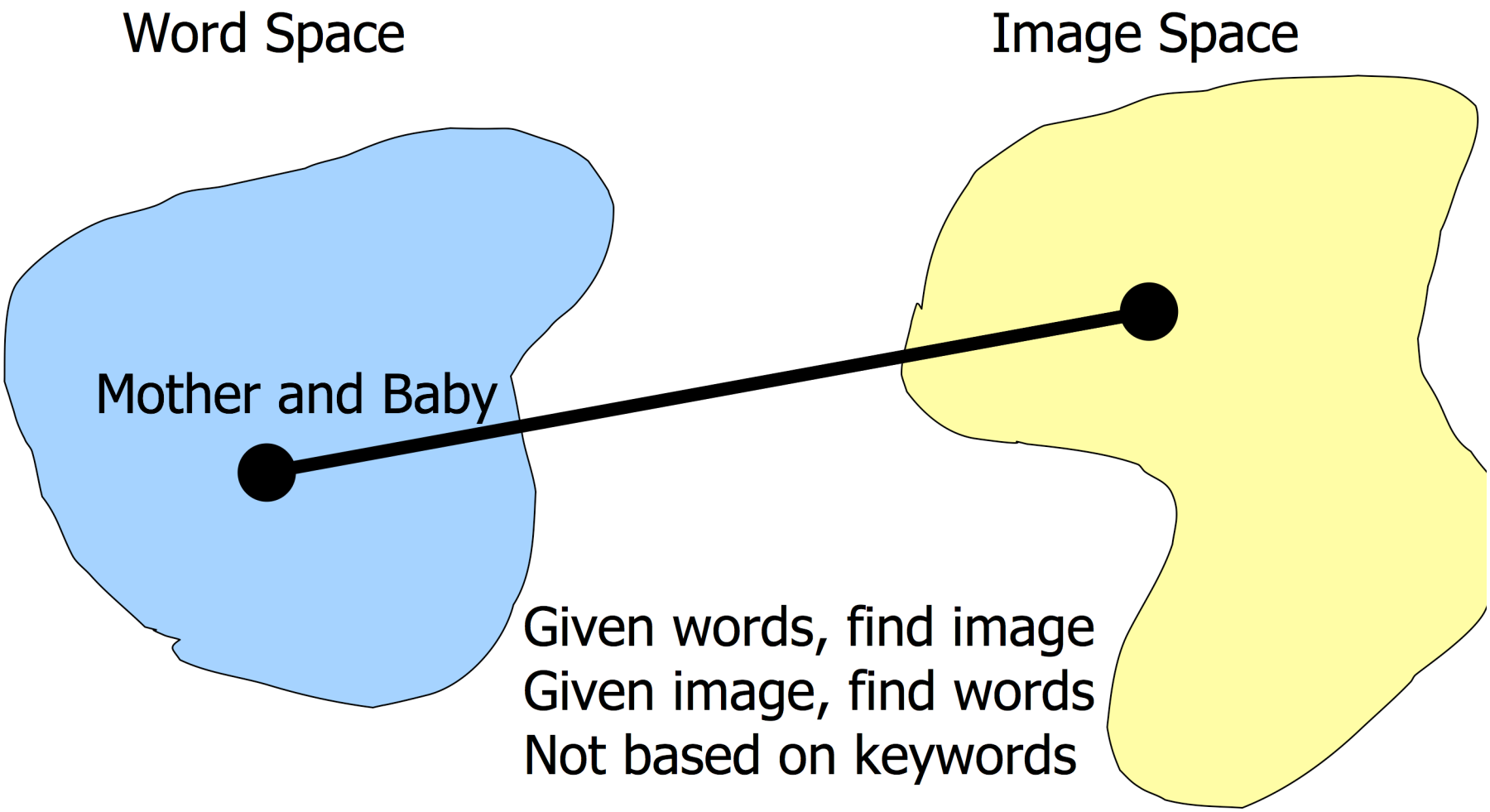
- Find related documents
- Multinomial EM (Nigam)
  - Probabilistic model of word probabilities
- Include smoothing in model

Document 10	**Cluster 383**	**Cluster 388**
anim (0.2)	anim (0.125)	anim (0.137931)
cattl (0.2)	sheep (0.125)	gallop (0.137931)
farm (0.2)	exterior (0.0833333)	hors (0.137931)
restless (0.2)	past (0.0833333)	pass (0.103448)
shed (0.2)	run (0.0833333)	singl (0.103448)
abrupt (0)	tree (0.0833333)	dirt (0.0689655)
abruptli (0)	wind (0.0833333)	gravel (0.0689655)
abuko (0)	concret (0.0416667)	concret (0.0344828)
acanthi (0)	flock (0.0416667)	folei (0.0344828)
acoust (0)	gate (0.0416667)	group (0.0344828)

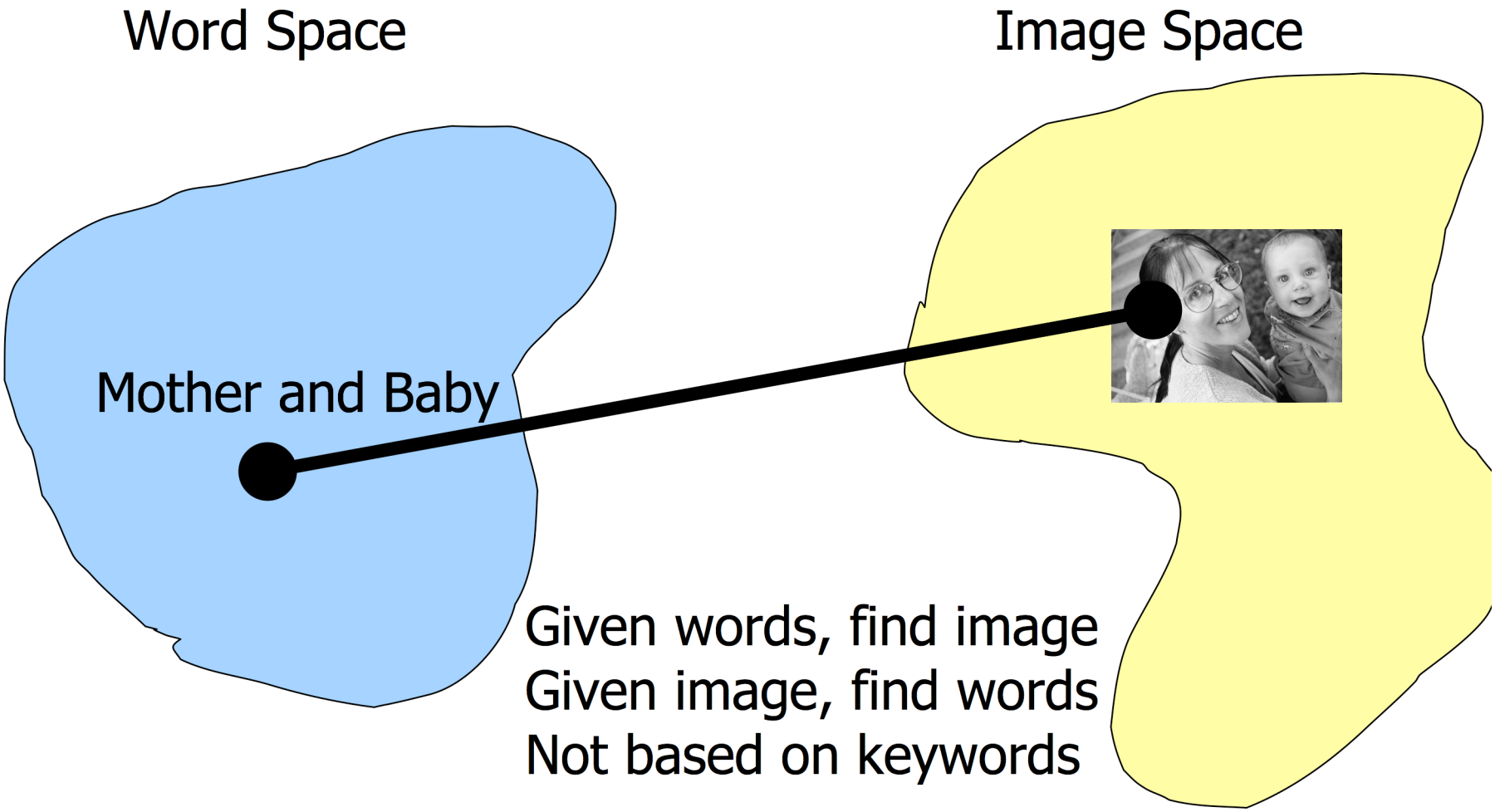
Words and Image Spaces



Words and Image Spaces

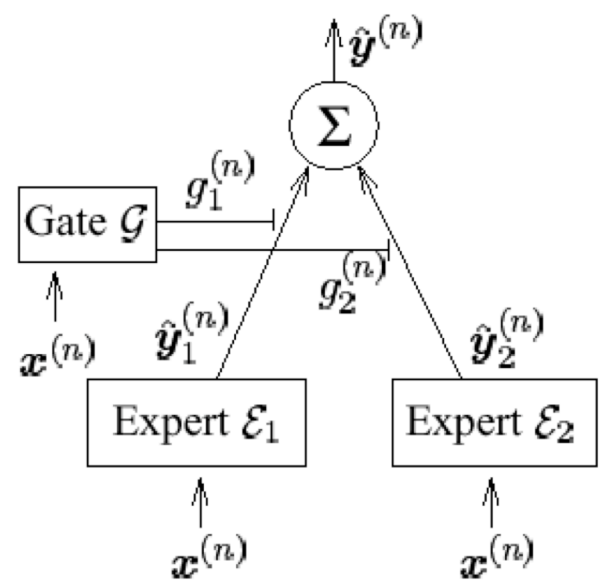


Words and Image Spaces



Mixture of Experts

- $x^{(n)}$ is input query
- Different experts are good with different kinds of queries
- Gate chooses expert
- $y^{(n)}$ is classification
- Train with EM
- Want probability estimates



MPE Mathematics

$$P(a \mid q) = \sum_{c} P(a \mid c) P(c \mid q)$$

Prob(audio|query)

Prob(audio|cluster)

Prob(cluster|query)

(Mixture model, interpolation)

MPE Mathematics

$$P(a \mid q) = \sum_c P(a \mid c) P(c \mid q)$$

Prob(audio|query)

Prob(audio|cluster)

Prob(cluster|query)

$$P(a \mid q) = \sum_c P(a \mid c) P(c) P(q \mid c) / P(q)$$

Ignore

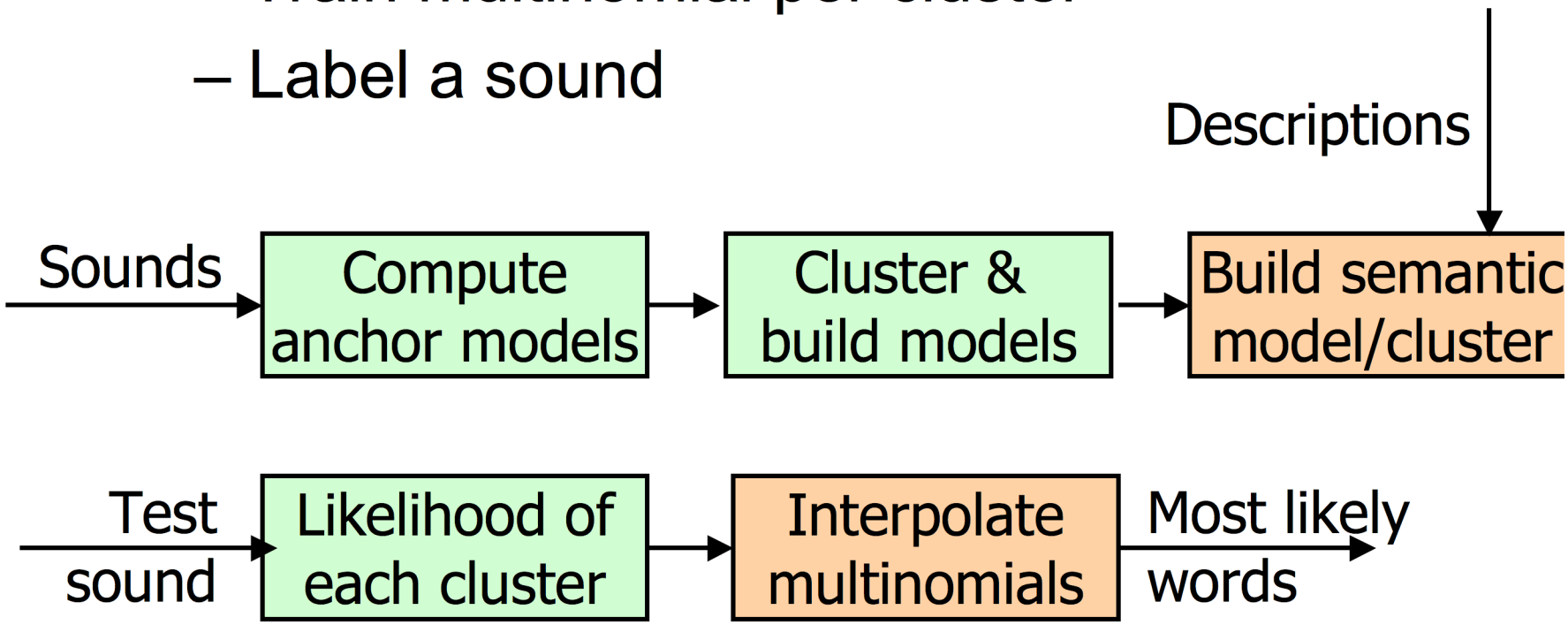
Multinomial model

Gaussian mixture model

(Mixture model, interpolation)

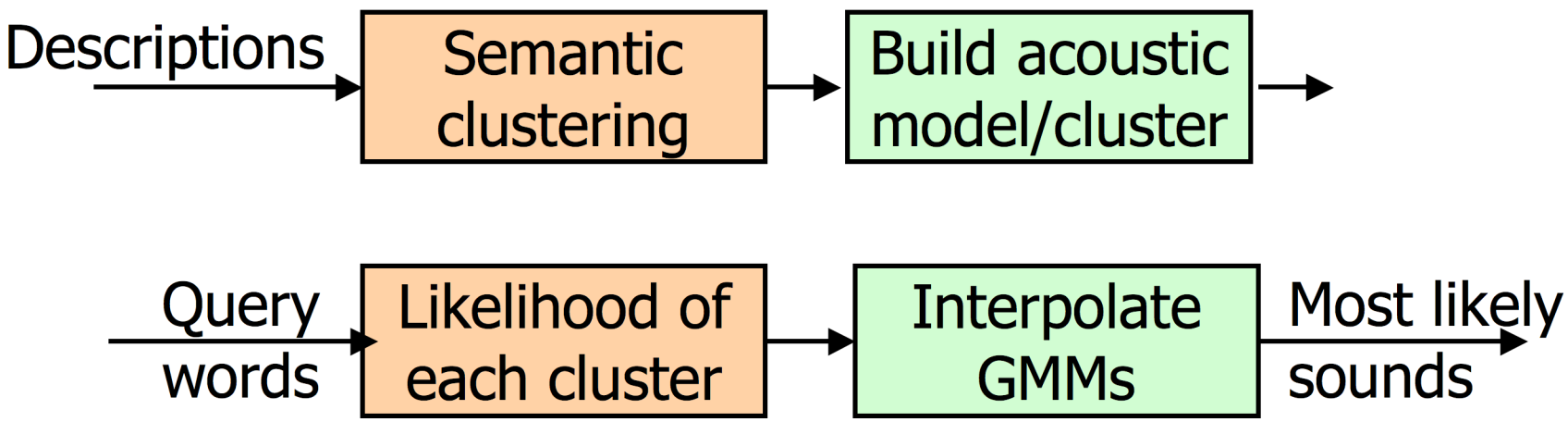
Semantic Retrieval

- Acoustic to Semantic
  - Train multinomial per cluster
  - Label a sound

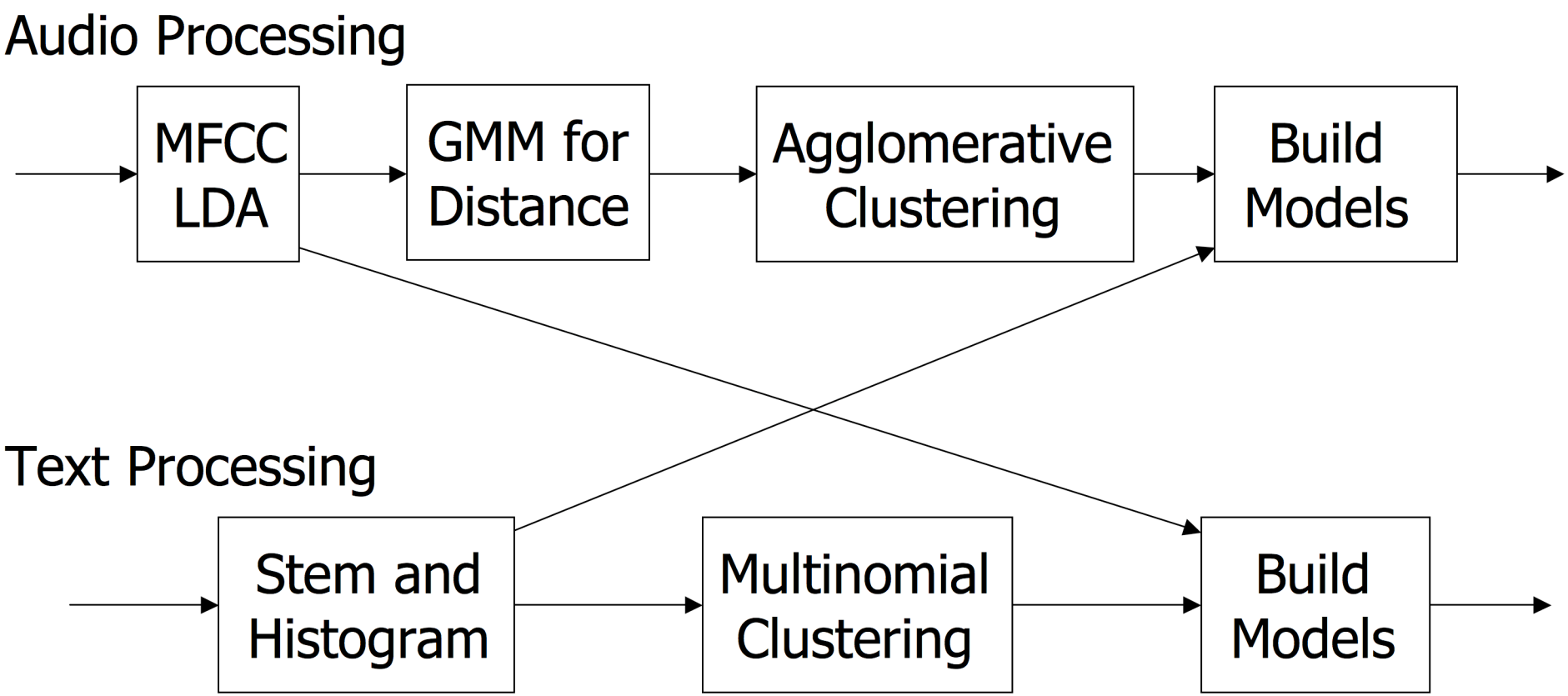


Acoustic Retrieval

- Semantic to Acoustic
  - Train semantic multinomial per cluster
  - Label a sound

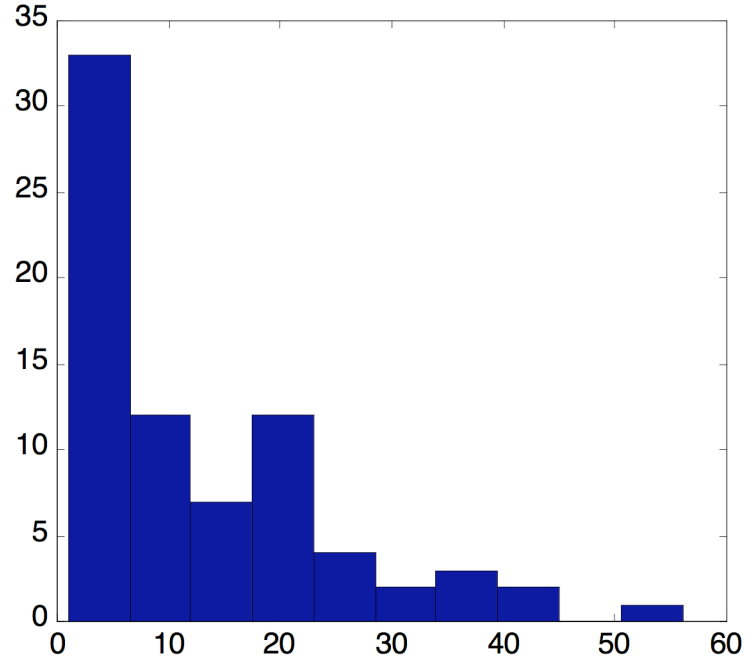


Training Overview

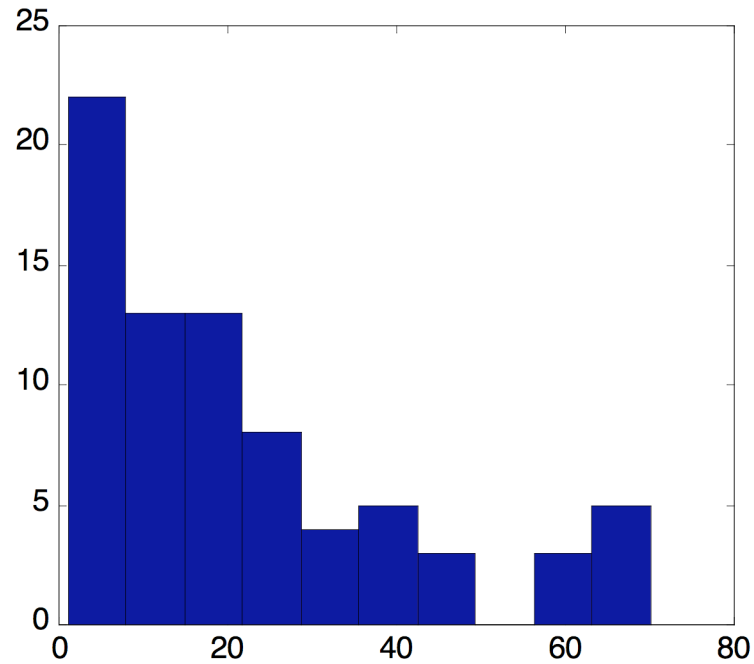


Test Results

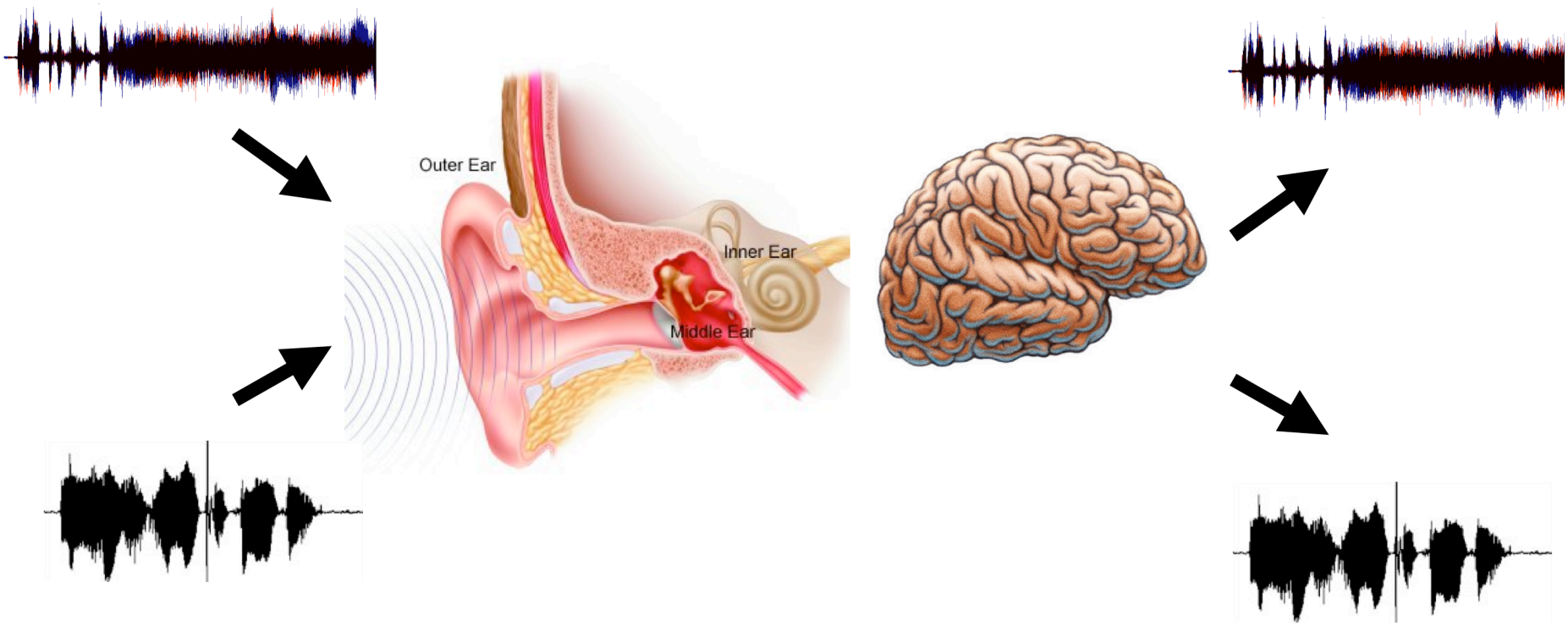
- Semantic → Acoustic
  - Median rank is 9



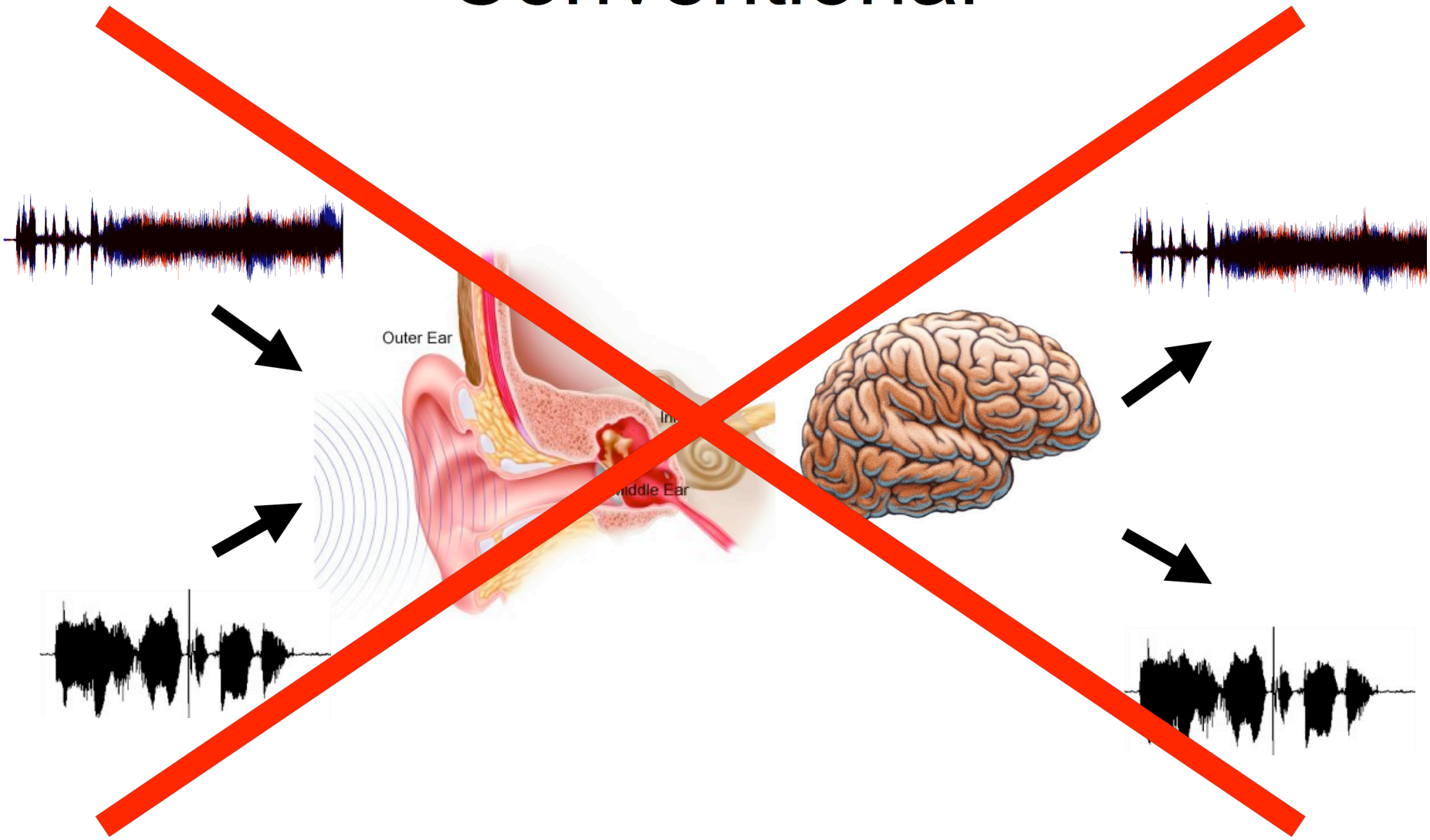
- Acoustic → Semantic
  - Median rank is 17.5.



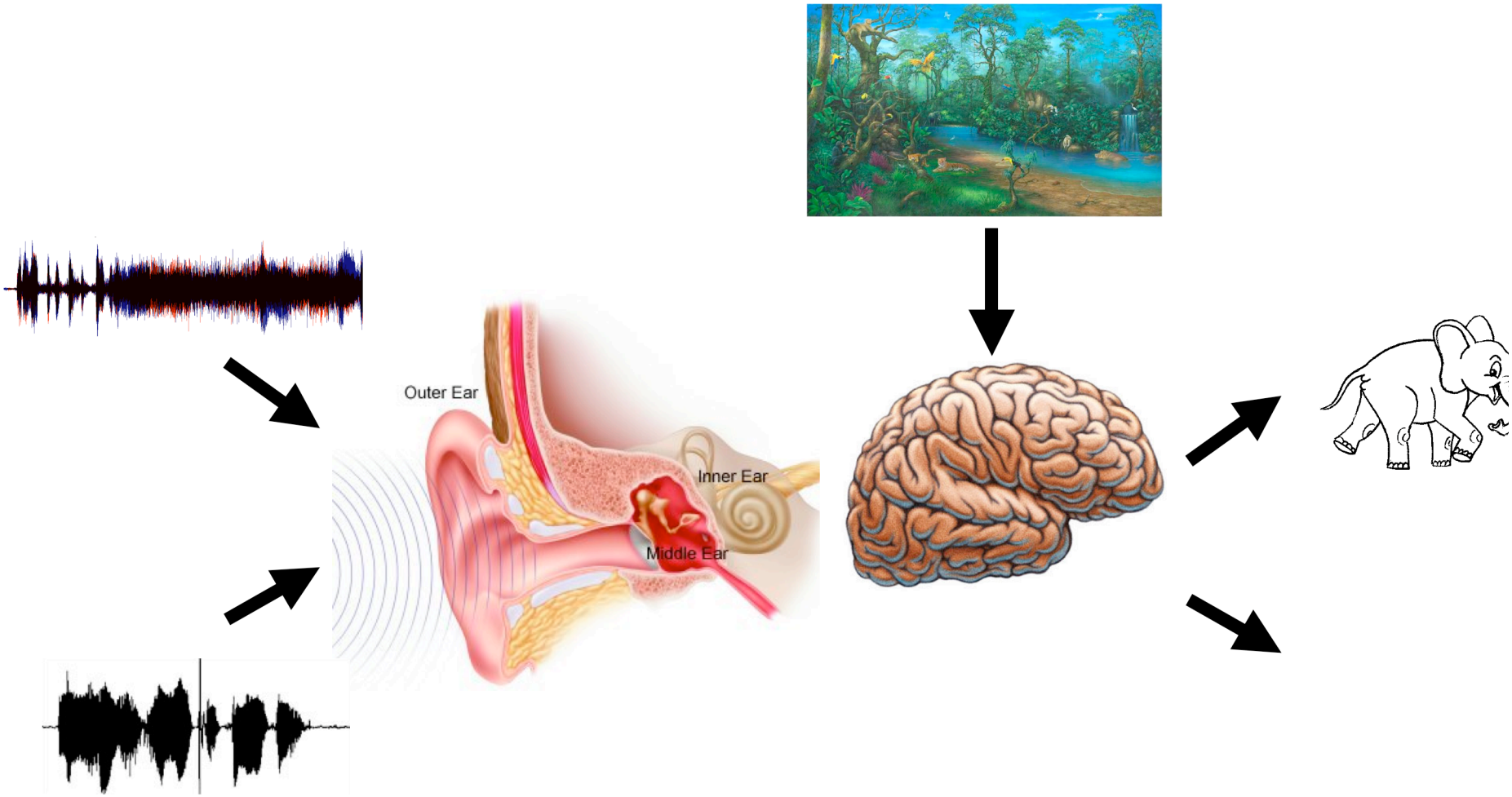
Conventional



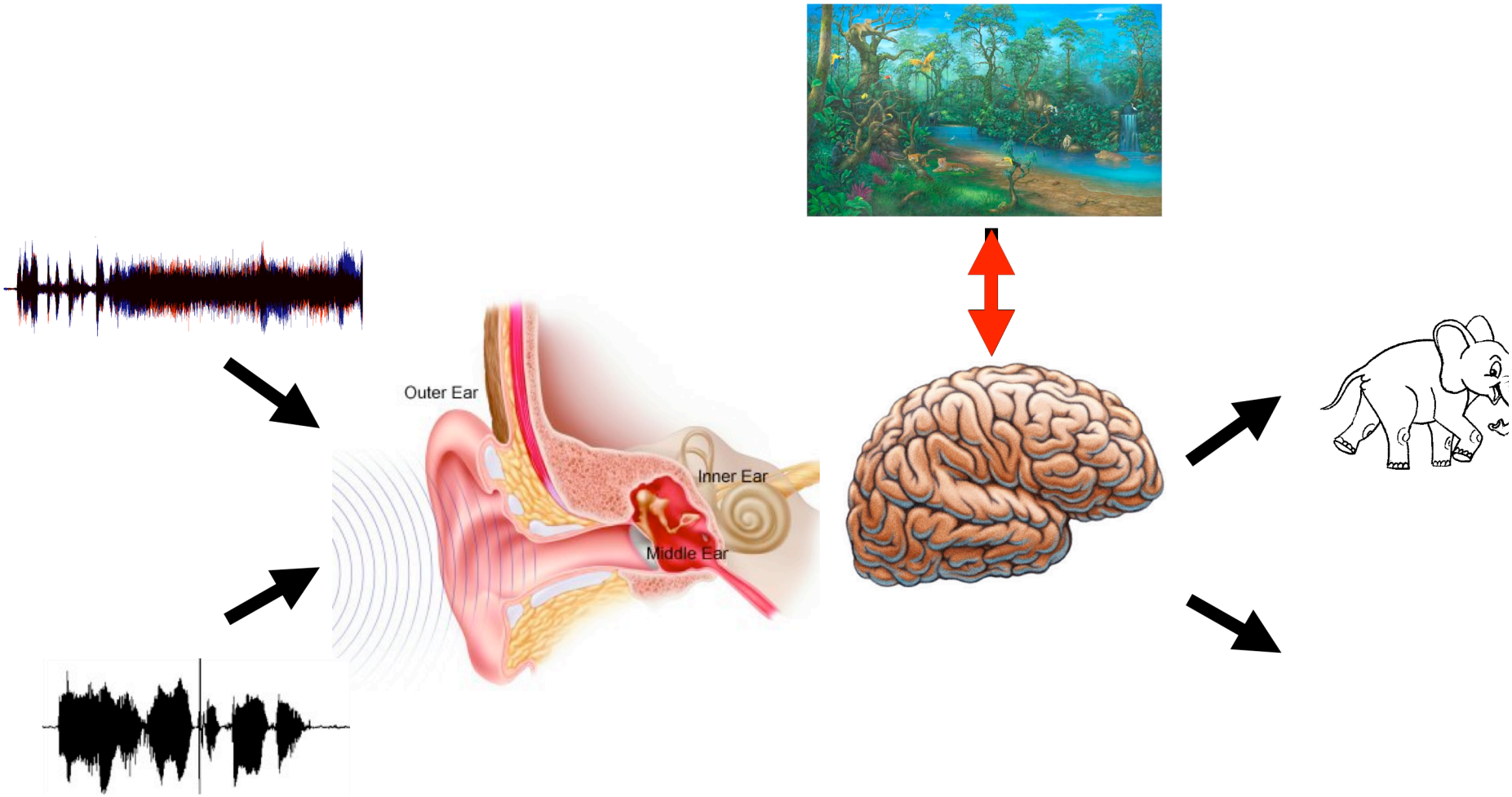
Conventional



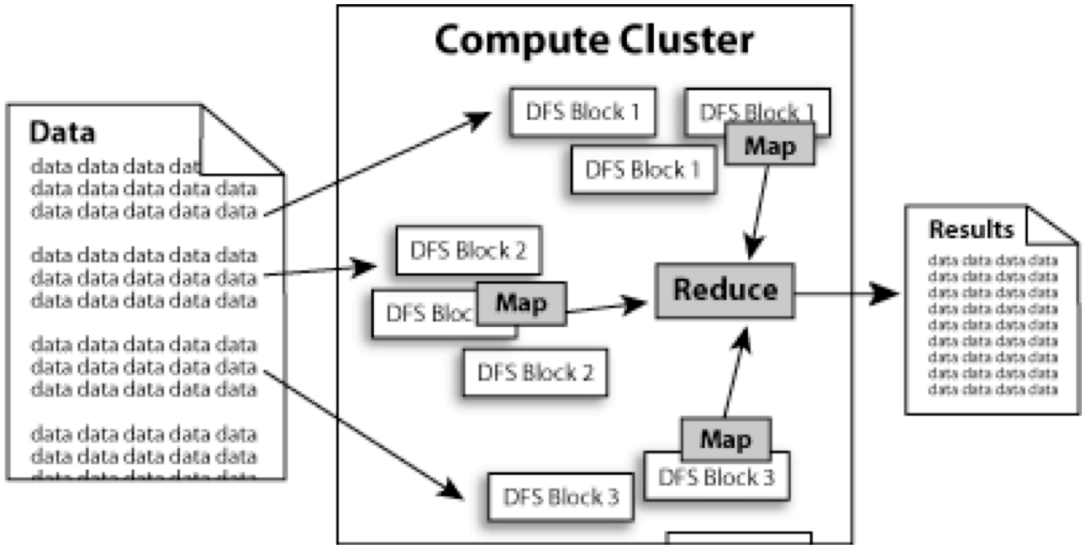
Better?



Better?



Hadoop


Compute Cluster
Data
data data data data
data data data data data
data data data data data
data data data data data
data data data data data
data data data data data
data data data data data
data data data data data
data data data data data
data data data data data
DFS Block 1
DFS Block 1
DFS Block 1
Map
DFS Block 2
DFS Bloc
Map
DFS Block 2
Reduce
Map
DFS Block 3
DFS Block 3
Results
data data data data
data data data data
data data data data
data data data data
data data data data
data data data data
data data data data
data data data data
data data data data

Thanks

malcolm@ieee.org