Computational Perception

15-485/785

February 5, 2008

Auditory Coding 1

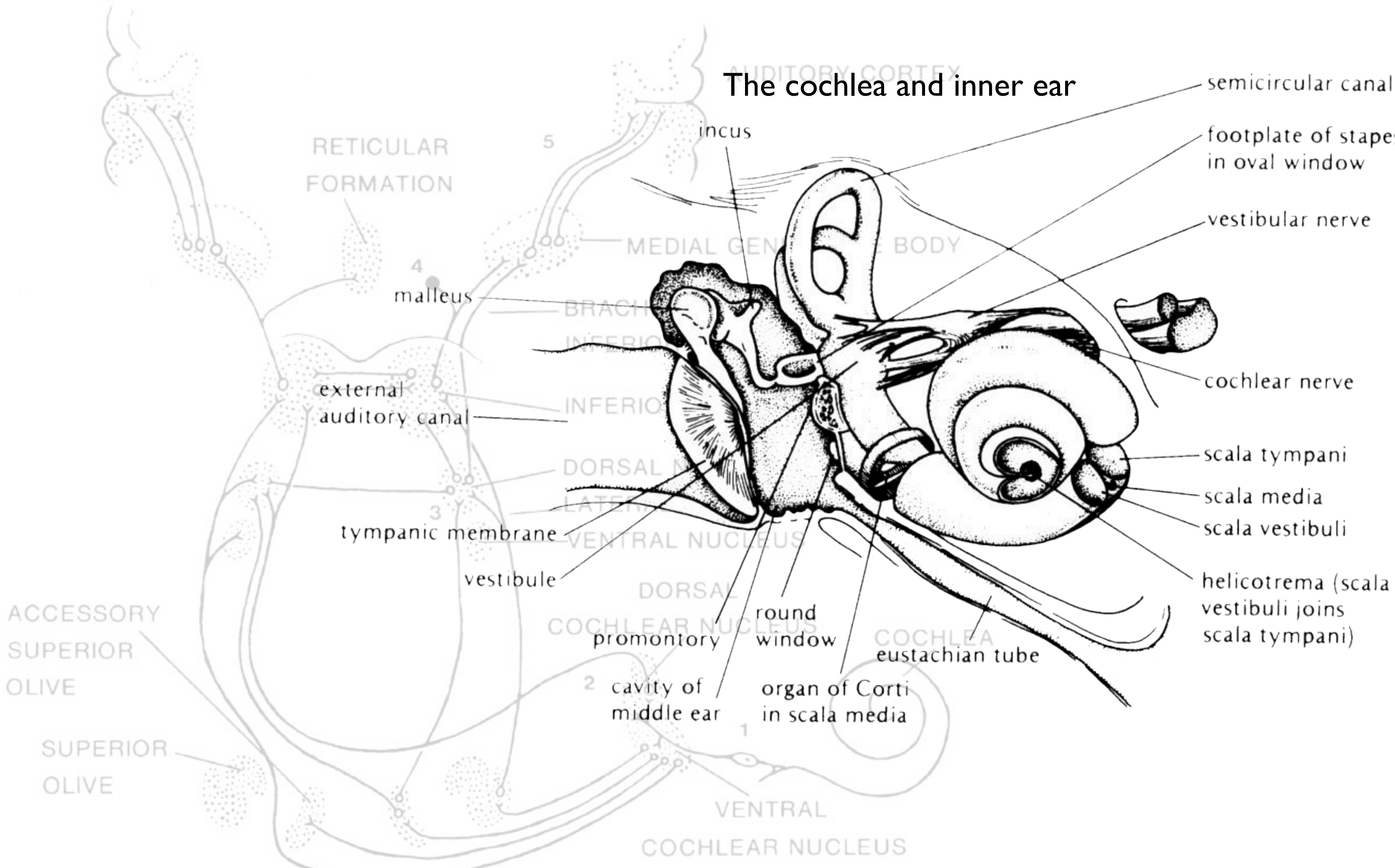
What are the problems of sensory coding?

- What should the sensor sense?
- How is energy transduced?
- How to deal with noise?
- How to compress dynamic range?
- How to prevent the sensor from being damaged?

Two approaches to the study of systems

1. Experimental/behavioral approach:
   - describe and characterize behavior
   - understand range and limitations
   - investigate system properties and organization
   - develop theories to better understand functional roles
2. Theoretical/computational approach:
   - define problem
   - develop models and algorithms
   - understand range and limitations
   - develop more algorithms: more general/specialized; faster/less resources

The complexity of the auditory system



from Warren, 1999

What principles should guide the choice of representation?

Unsupervised approaches:

- find useful "features"
- adapted to the patterns of interest
- useful in a wide range of tasks

Supervised approaches:

- Maximize performance on given task

At low-levels, we have to use unsupervised approaches.

Linear superposition

Goal is to describe the data to desired precision.
Code signal by linear superposition of basis functions:

$$\begin{aligned} \mathbf{x} &= \vec{a}_1 s_1 + \vec{a}_2 s_2 + \cdots + \vec{a}_L s_L + \vec{\epsilon} \\ &= \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon} \end{aligned}$$

- x(t) is represented by a vector $\mathbf{x}$
- $\vec{a}_i$ are the *basis vectors*
- $\mathbf{A}$ is the *basis* (could be Fourier, wavelet, etc.)
- s_i are the *coefficients*

Can solve for $\hat{\mathbf{s}}$ in the no noise case

$$\hat{\mathbf{s}} = \mathbf{A}^{-1}\mathbf{x}$$

An information theoretic approach

Want algorithm to choose optimal $\mathbf{A}$ (basis matrix).

Generative model for data is:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}$$

Probability of pattern $\mathbf{x}$ given representation $\mathbf{s}$

$$P(\mathbf{x}|\mathbf{A}, \mathbf{s}) \quad \sim \quad f(\mathbf{x} - \mathbf{A}\mathbf{s}, \boldsymbol{\Sigma}, I)$$

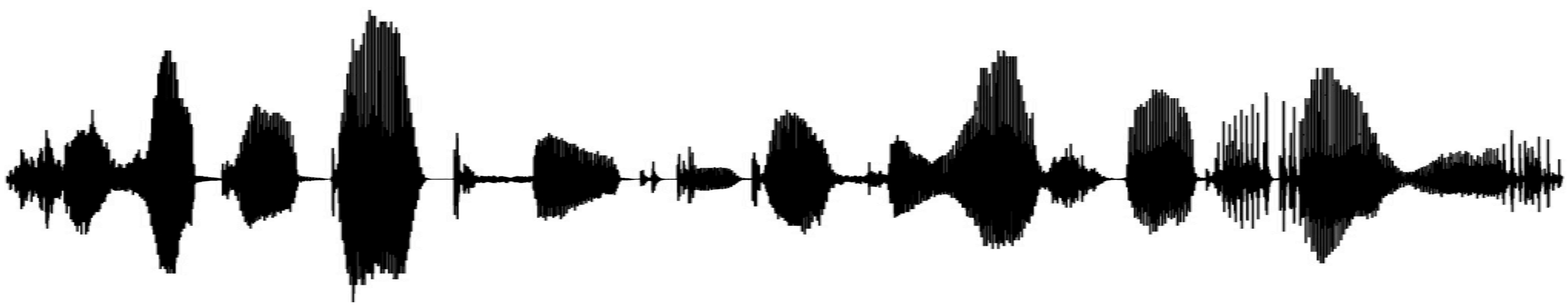
Learning objective

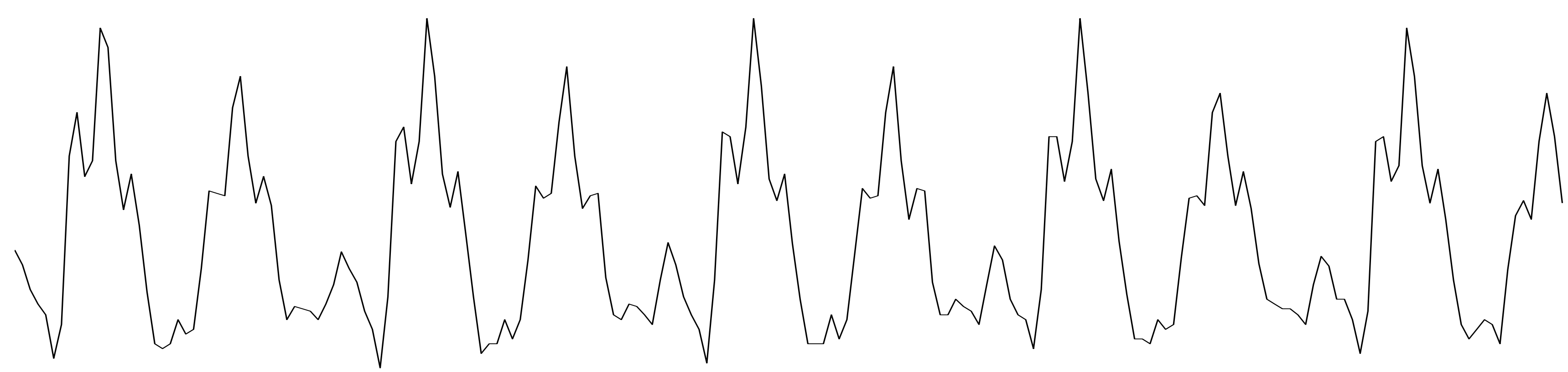
Objective: maximize coding efficiency
$\Rightarrow$ maximize probability of data ensemble

Probability of pattern ensemble is:

$$P(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N | \mathbf{A}) = \prod_k P(\mathbf{x}_k | \mathbf{A})$$

Optimal coding of an acoustic waveform





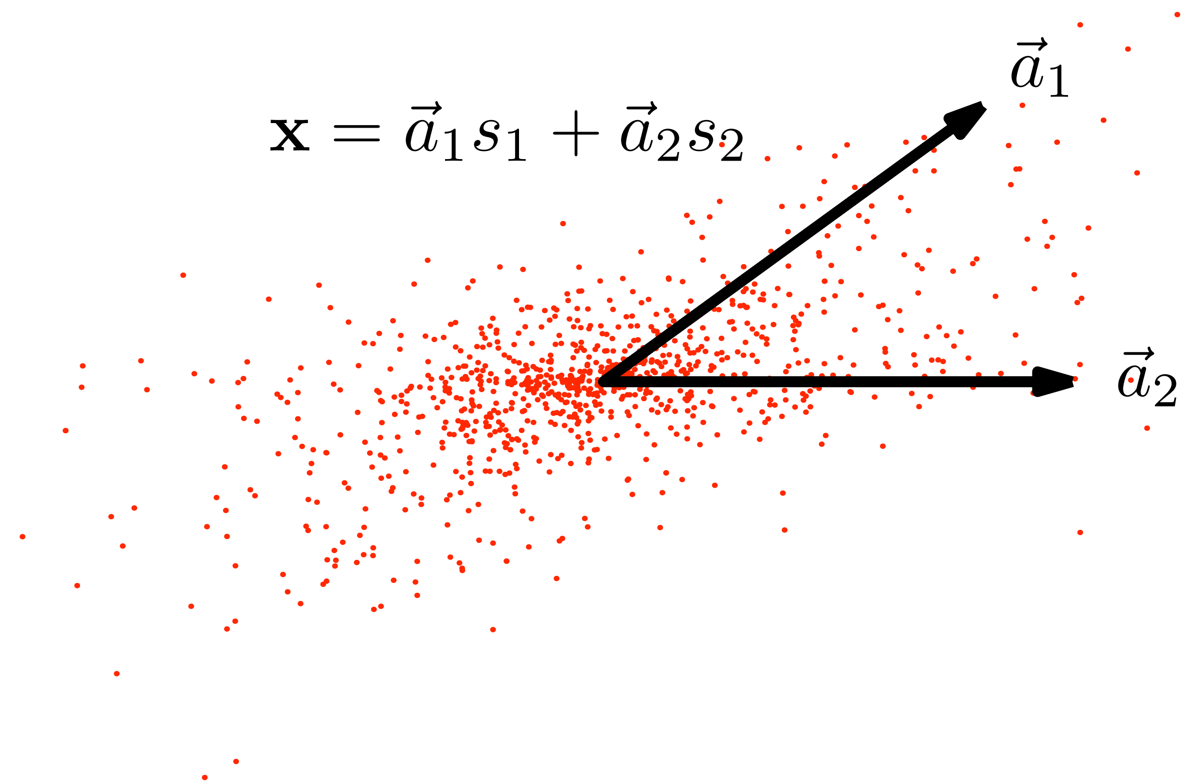
- We do *not* assume a Fourier or spectral representation.
- Goal:

  Predict optimal transformation of acoutsic waveform from statistics of the acoustic environment.
- Use a simple model: bank of linear filters

Coding patterns with a statistical model

Goal: Encode the patters to desired precision:

$$\begin{aligned} \mathbf{x} &= \vec{a}_1 s_1 + \cdots + \vec{a}_L s_L + \vec{\epsilon} \\ &= \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon} \end{aligned}$$



Posterior:

$$P(\mathbf{s}|\mathbf{x}, \mathbf{A}) = \frac{P(\mathbf{s})P(\mathbf{x}|\mathbf{s}, \mathbf{A})}{P(\mathbf{x}|\mathbf{A})}$$

Prior: s_i's are *independent* and *sparse*:

$$P(\mathbf{s}) = \prod_i P(s_i)$$

$$P(s_i) \propto \exp\left[-\left|\frac{s_i}{\lambda_i}\right|^{q_i}\right]$$

Coding patterns with a statistical model

Goal: Encode the patters to desired precision:

$$\begin{aligned}\mathbf{x} &= \vec{a}_1 s_1 + \cdots + \vec{a}_L s_L + \vec{\epsilon} \\ &= \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon}\end{aligned}$$

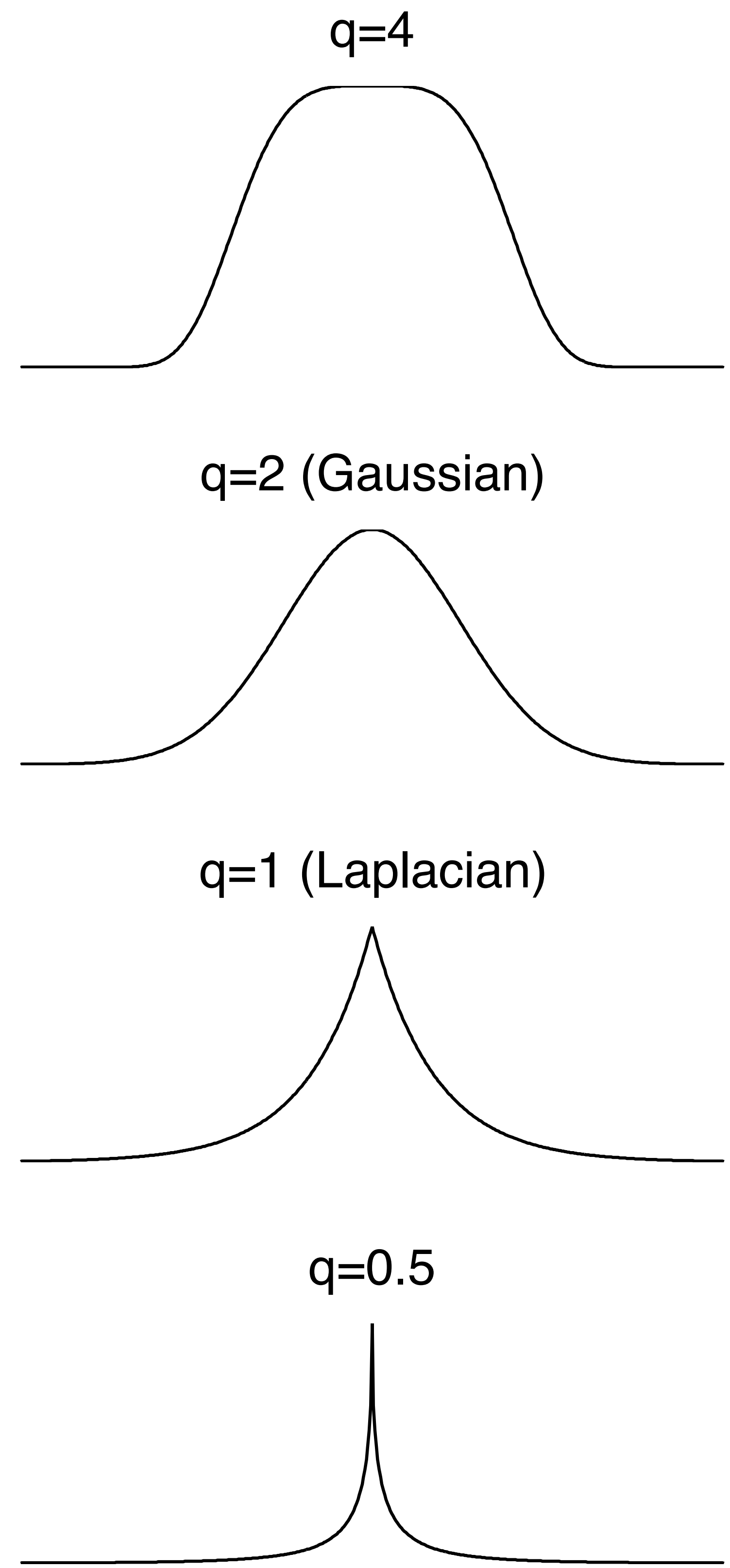
Posterior:

$$P(\mathbf{s}|\mathbf{x},\mathbf{A}) = \frac{P(\mathbf{s})P(\mathbf{x}|\mathbf{s},\mathbf{A})}{P(\mathbf{x}|\mathbf{A})}$$

Prior: s_i's are *independent* and *sparse*:

$$P(\mathbf{s}) = \prod_i P(s_i)$$

$$P(s_i) \propto \exp\left[-\left|\frac{s_i}{\lambda_i}\right|^{q_i}\right]$$



Coding patterns with a statistical model

Goal: Encode the patters to desired precision:

$$\begin{aligned} \mathbf{x} &= \vec{a}_1 s_1 + \cdots + \vec{a}_L s_L + \vec{\epsilon} \\ &= \mathbf{A}\mathbf{s} + \boldsymbol{\epsilon} \end{aligned}$$

Posterior:

$$P(\mathbf{s}|\mathbf{x}, \mathbf{A}) = \frac{P(\mathbf{s})P(\mathbf{x}|\mathbf{s}, \mathbf{A})}{P(\mathbf{x}|\mathbf{A})}$$

Prior: s_i's are *independent* and *sparse*:

$$P(\mathbf{s}) = \prod_i P(s_i)$$

$$P(s_i) \propto \exp\left[-\left|\frac{s_i}{\lambda_i}\right|^{q_i}\right]$$

Likelihood: Assume $\boldsymbol{\epsilon} \sim$ Gaussian,

$$P(\mathbf{x}|\mathbf{s}, \Sigma) \propto \exp\left[-\frac{1}{2}\boldsymbol{\epsilon}^T \Sigma^{-1} \boldsymbol{\epsilon}\right]$$

Inference: use the MAP value:

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}, \mathbf{A})$$

Simple special case: no noise (ICA)

$$\hat{\mathbf{s}} = \mathbf{A}^{-1}\mathbf{x}$$

Inference (or recognition or coding):

finds most efficient representation of pattern $\mathbf{x}$ in a given basis $\mathbf{A}$

Learning: Optimizing the model parameters

Learning objective:

maximize coding efficiency
$\Rightarrow$ maximize $P(\mathbf{x}|\mathbf{A})$ over $\mathbf{A}$.

Probability of pattern ensemble is:

$$P(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N|\mathbf{A}) = \prod_k P(\mathbf{x}_k|\mathbf{A})$$

$P(\mathbf{x}|\mathbf{A})$ is obtained by marginalization:

$$\begin{aligned} P(\mathbf{x}|\mathbf{A}) &= \int d\mathbf{s}\, P(\mathbf{x}|\mathbf{A}, \mathbf{s}) P(\mathbf{s}) \\ &= \frac{P(\mathbf{s})}{|\det \mathbf{A}|} \end{aligned}$$

Use *independent component analysis* (ICA) to learn $\mathbf{A}$:

$$\begin{aligned} \Delta\mathbf{A} &\propto \mathbf{A}\mathbf{A}^T \frac{\partial}{\partial\mathbf{A}} \log P(\mathbf{x}|\mathbf{A}) \\ &= -\mathbf{A}(\mathbf{z}\mathbf{s}^T - \mathbf{I}), \end{aligned}$$

where $\mathbf{z} = (\log P(\mathbf{s}))'$. Assume generalized Gaussians:

$$P(s_i) \sim \mathcal{N}^{q_i}(s_i|\mu, \sigma).$$

This learning rule:

- *learns the feature set that captures the most structure*
- *optimizes basis to maximize the efficiency of the code*

Learning the optimal codes

Goal:

Predict optimal transformation of sound waveform from statistics of the acoustic environment

Learning procedure:

- random sound segments (8 msec)
- optimize features using ICA

What sounds to use?

What tasks are auditory systems adapted to do?

- localization ⇒ environmental sounds
- communication ⇒ vocalizations
- general sound recognition

Use a variety of sound ensembles:

- non-harmonic *environmental sounds* (e.g. footsteps, stream sounds, etc.)
- *animal vocalizations* (rainforest mammals, e.g. chirps, screeches, cries, etc.)
- *speech* (samples from 100 male & female speakers from the TIMIT corpus)

Natural sounds

vocalizations	environmental sounds: transient	environmental sounds: ambient
fox	walking on leaves	rustling leaves
squirrel	cracking branches	stream by waterfall

Natural sounds

	environmental sounds	
vocalizations	transient	ambient
fox	walking on leaves	rustling leaves
squirrel	cracking branches	stream by waterfall

Natural sounds

	environmental sounds	
vocalizations	transient	ambient
fox	walking on leaves	rustling leaves
squirrel	cracking branches	stream by waterfall

Natural sounds

vocalizations	environmental sounds: transient	environmental sounds: ambient
fox	**walking on leaves**	rustling leaves
squirrel	cracking branches	stream by waterfall

Natural sounds

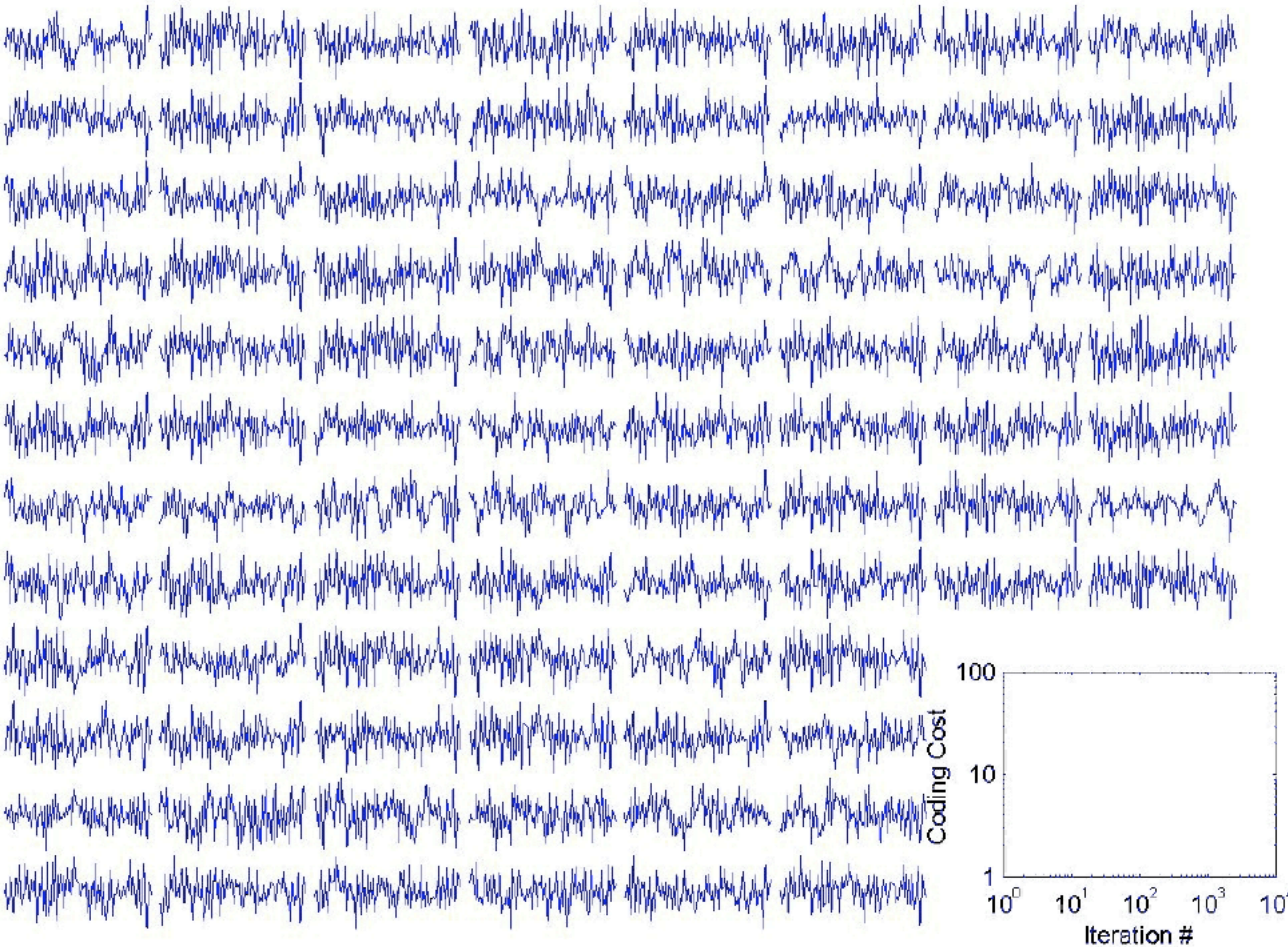
	environmental sounds	
vocalizations	transient	ambient
fox	walking on leaves	rustling leaves
squirrel	**cracking branches**	stream by waterfall

Natural sounds

	environmental sounds	
vocalizations	transient	ambient
fox	walking on leaves	**rustling leaves**
squirrel	cracking branches	stream by waterfall

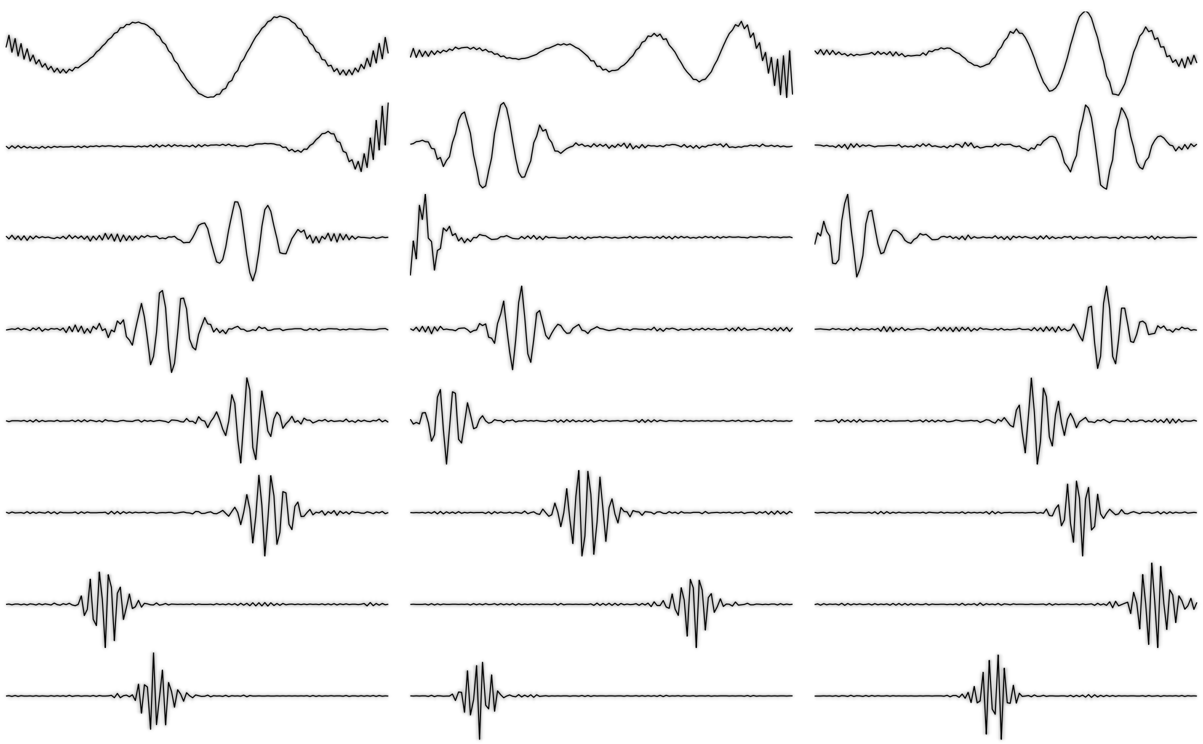
Natural sounds

	environmental sounds	
vocalizations	transient	ambient
fox	walking on leaves	rustling leaves
squirrel	cracking branches	**stream by waterfall**

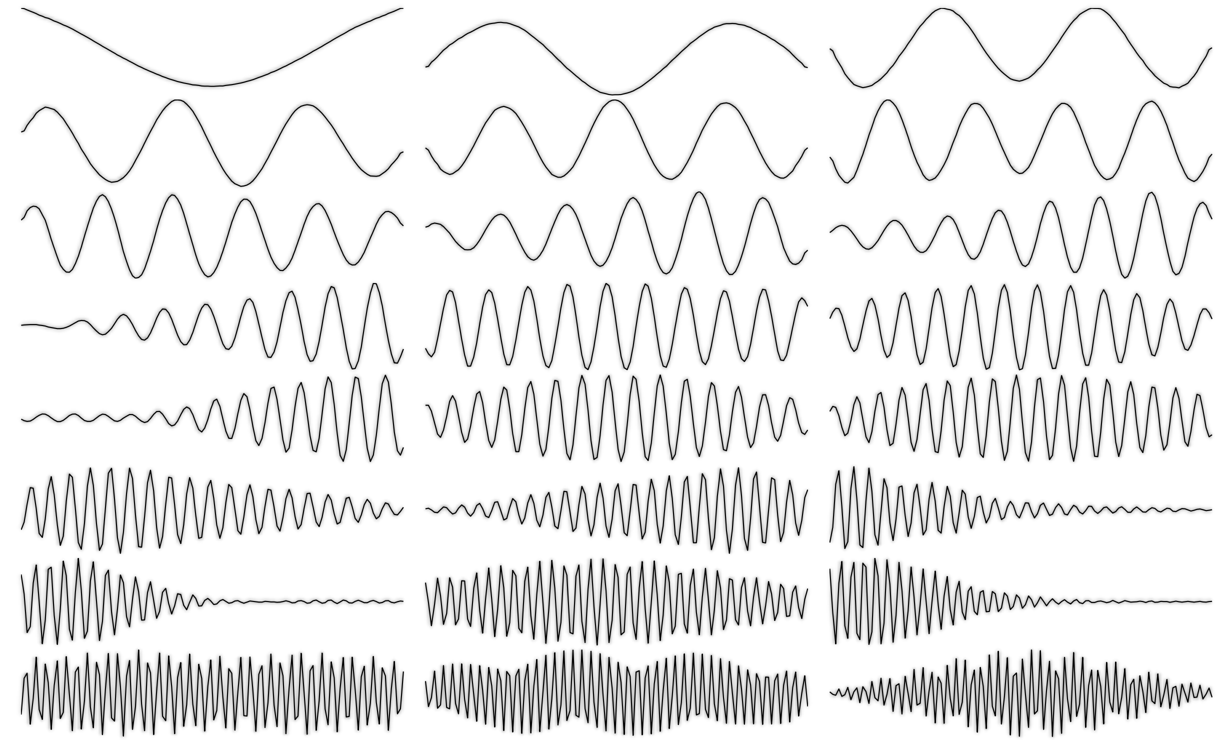


100
10
1
Coding Cost
10^0
10^1
10^2
10^3
10^4
Iteration #

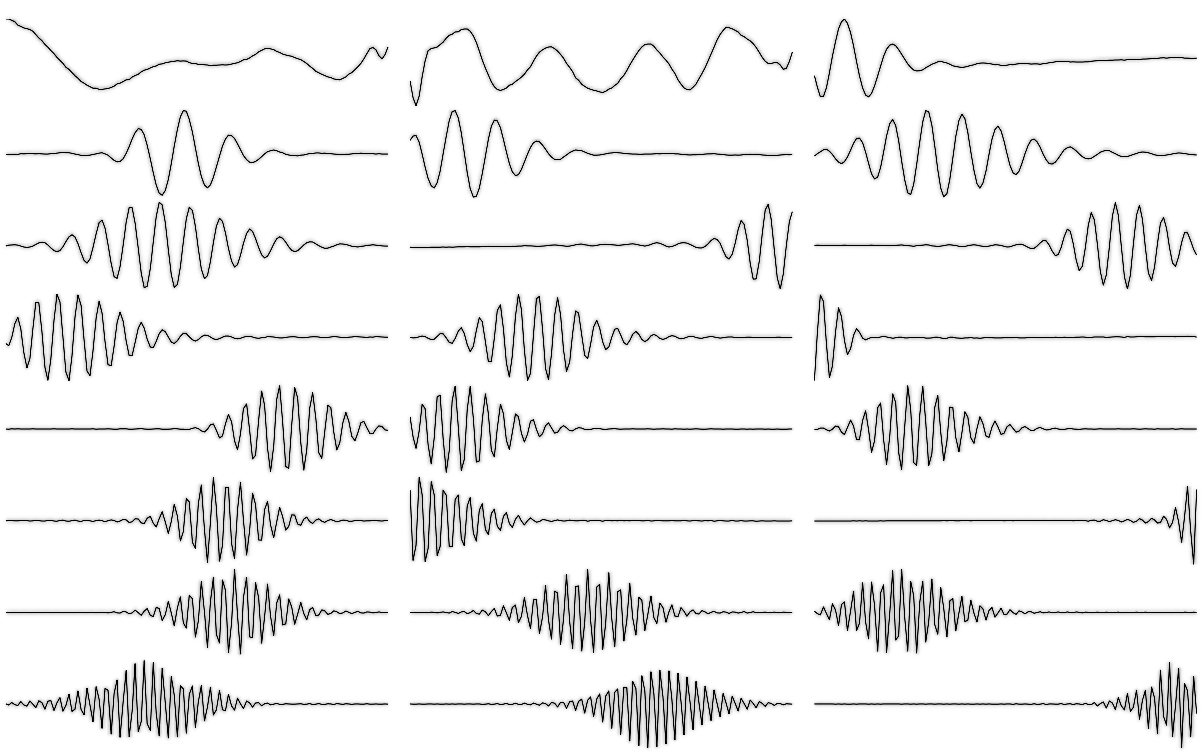
Optimal linear filters for natural sounds



environmental sounds



vocalizations



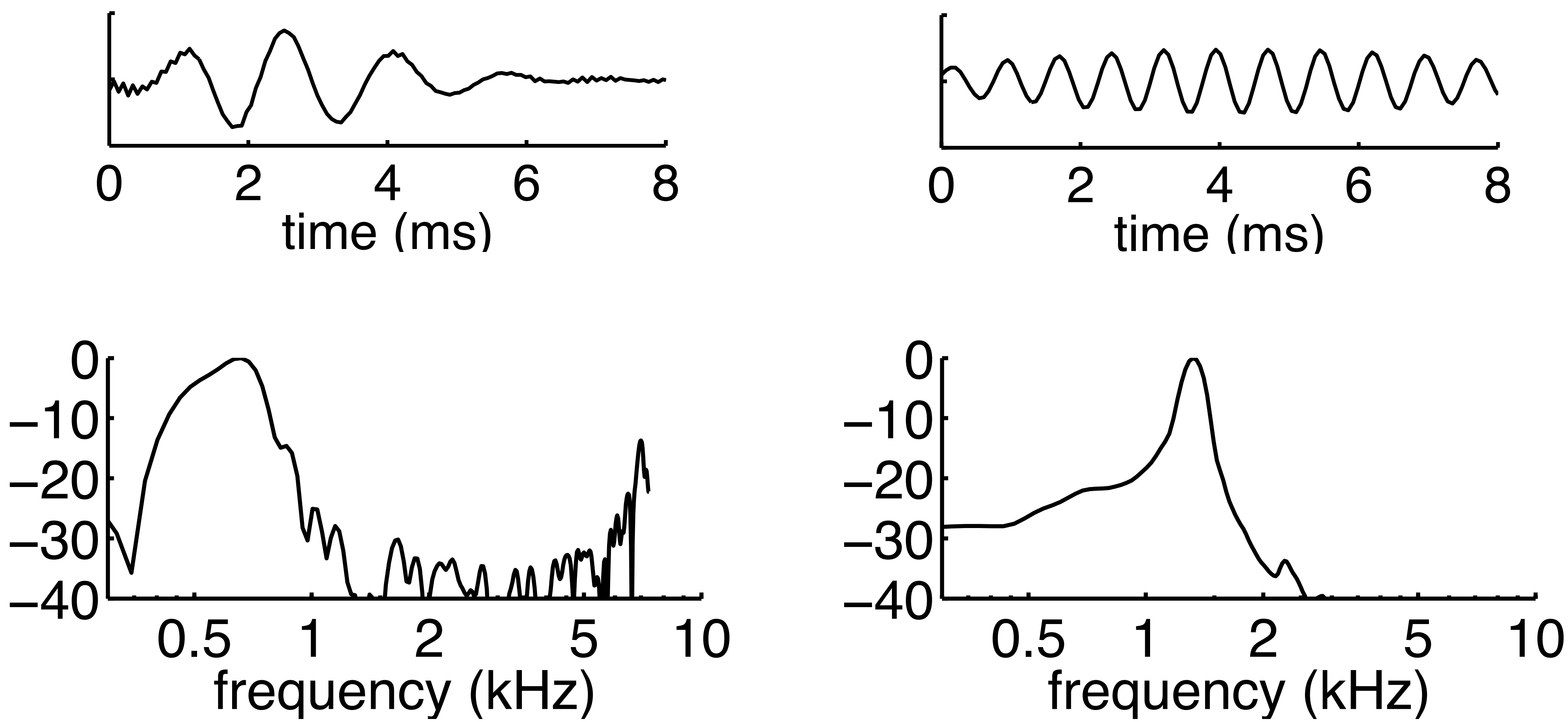
speech

The optimal code depends on the class of sounds being encoded:

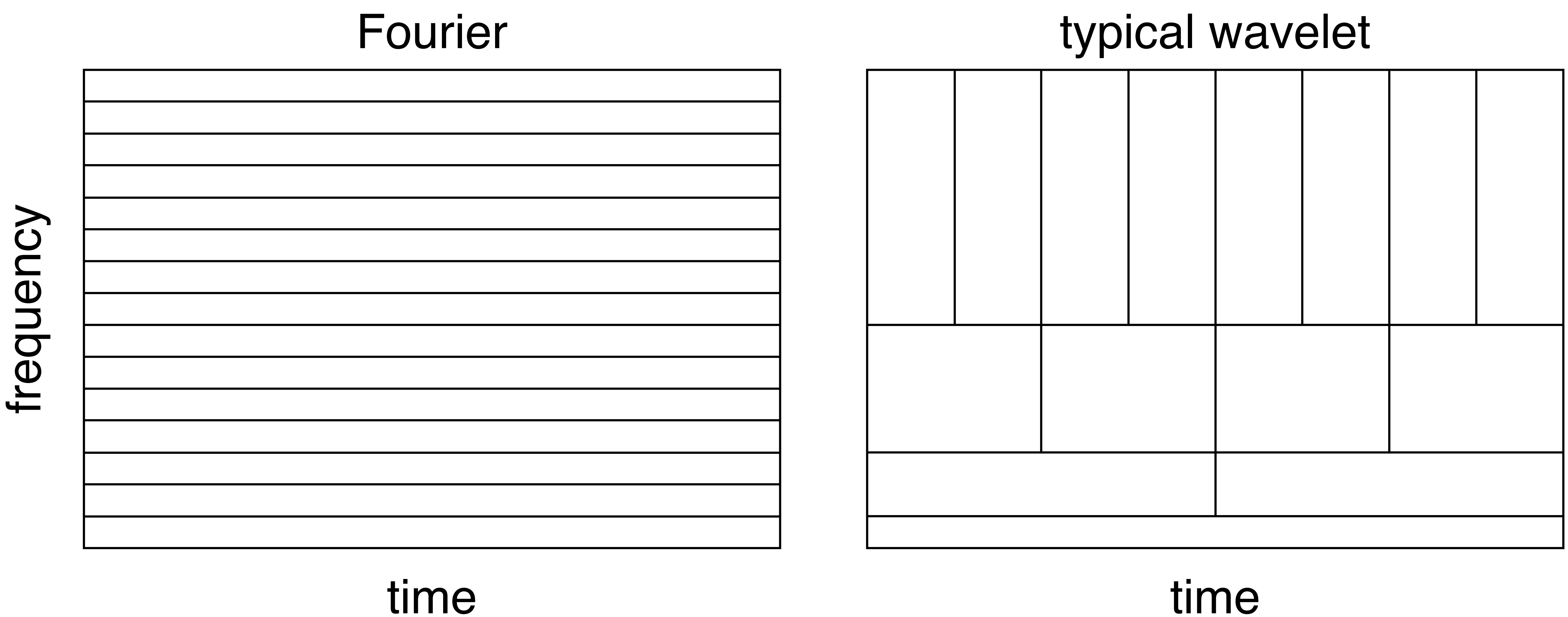
- a wavelet-like transform is best for environmental sounds
- a Fourier-like transform is best for vocalizations
- an intermediate transform is best for speech *or general natural sounds*

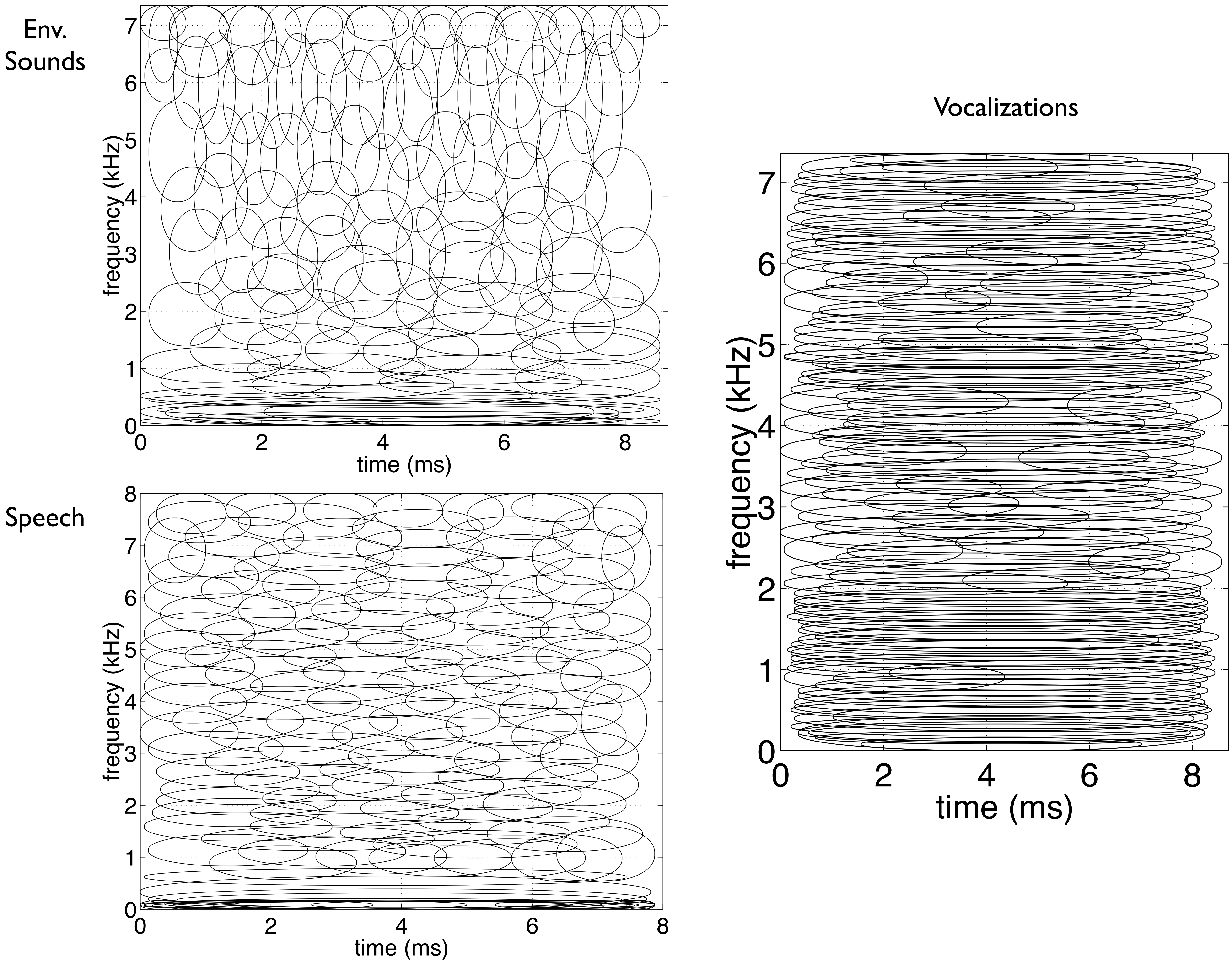
Characterizing the filter population

time-frequency distributions

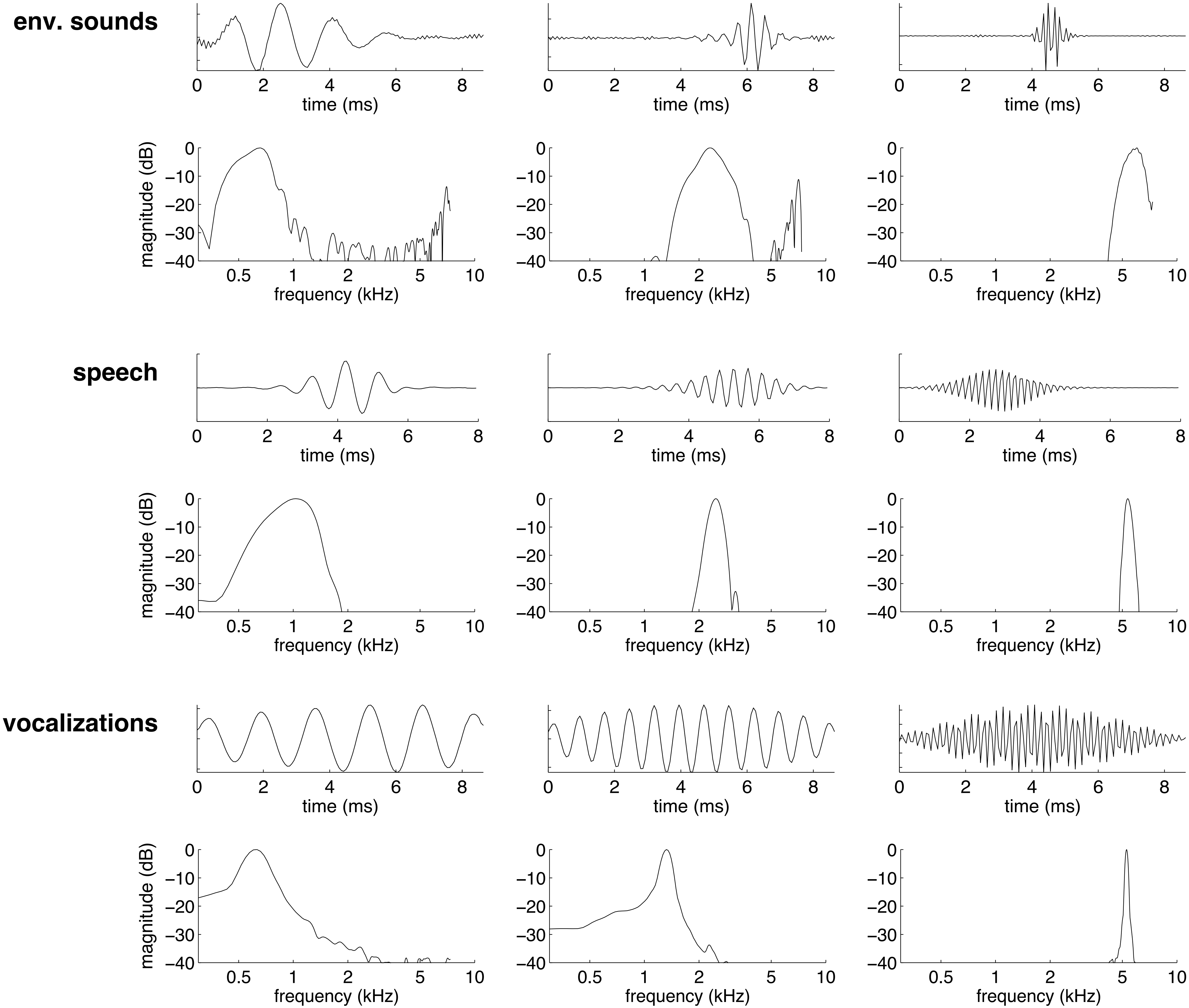


Schematic time-frequency distributions



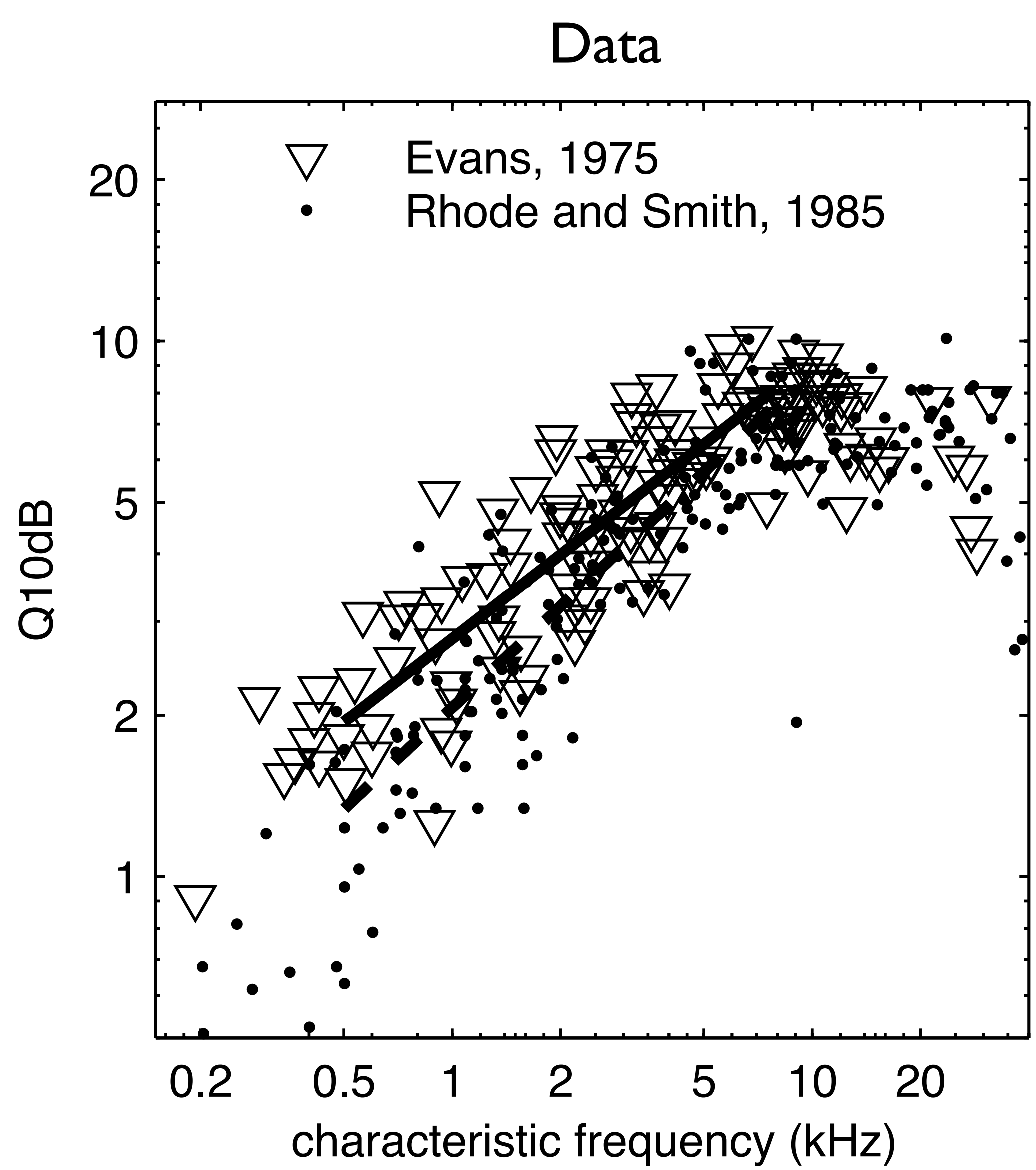


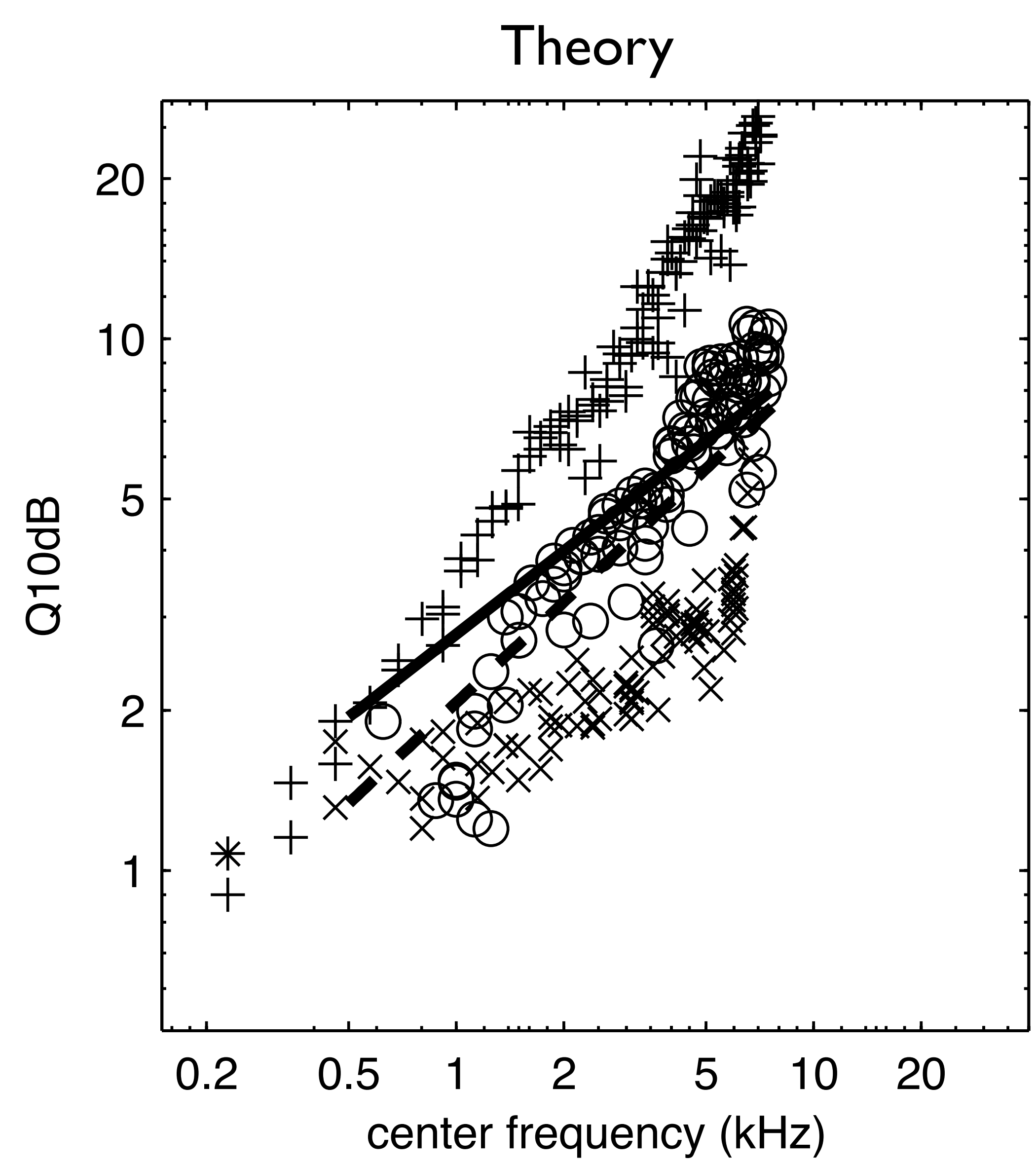
Env.
Sounds
frequency (kHz)
time (ms)
Speech
frequency (kHz)
time (ms)
Vocalizations
frequency (kHz)
time (ms)



env. sounds
speech
vocalizations
time (ms)
magnitude (dB)
frequency (kHz)

Comparison to cat auditory nerve data





Filter sharpness:

$$Q_{10\mathrm{dB}} = f_c / w_{10\mathrm{dB}}$$

+ vocalizations
o speech/combined
x environmental sounds

Next time:

non-linear coding