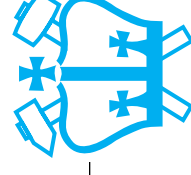


## Lecture 4: Auditory Perception

- 1 **Motivation: Why & how**
- 2 **Auditory physiology**
- 3 **Psychophysics: Detection & discrimination**
- 4 **Pitch perception**
- 5 **Speech perception**
- 6 **Auditory organization & Scene analysis**

Dan Ellis <[dpwe@ee.columbia.edu](mailto:dpwe@ee.columbia.edu)>  
<http://www.ee.columbia.edu/~dpwe/e6820/>

Columbia University Dept. of Electrical Engineering  
Spring 2006



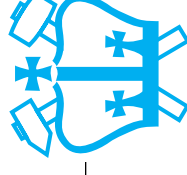
---

---

# 1

## Why study perception?

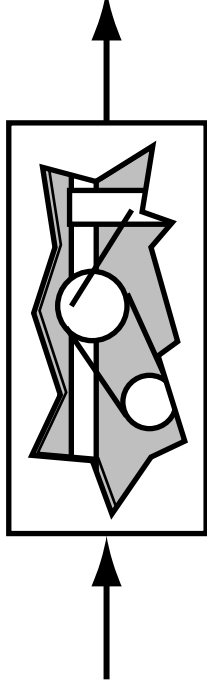
- **Perception is messy: Can we avoid it?**  
**No!**
- **Audition provides the ‘ground truth’ in audio**
  - what is relevant and irrelevant
  - subjective importance of distortion (coding etc.)
  - (there could be other information in sound...)
- **Some sounds are ‘designed’ for audition**
  - **co-evolution** of speech and hearing
- **The auditory system is very successful**
  - we would do extremely well to duplicate it
- **We are now able to model complex systems**
  - faster computers, bigger memories



# How to study perception?

Three different approaches:

- Analyze the **example**: **physiology**



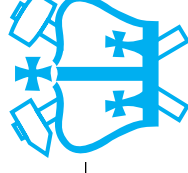
- dissection & nerve recordings

- **Black box input/output**: **psychophysics**



- fit simple models of simple functions

- **Information processing models**
  - investigate and model complex functions
  - e.g. scene analysis, speech perception

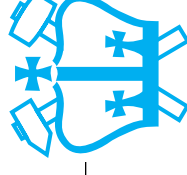


---

---

# Outline

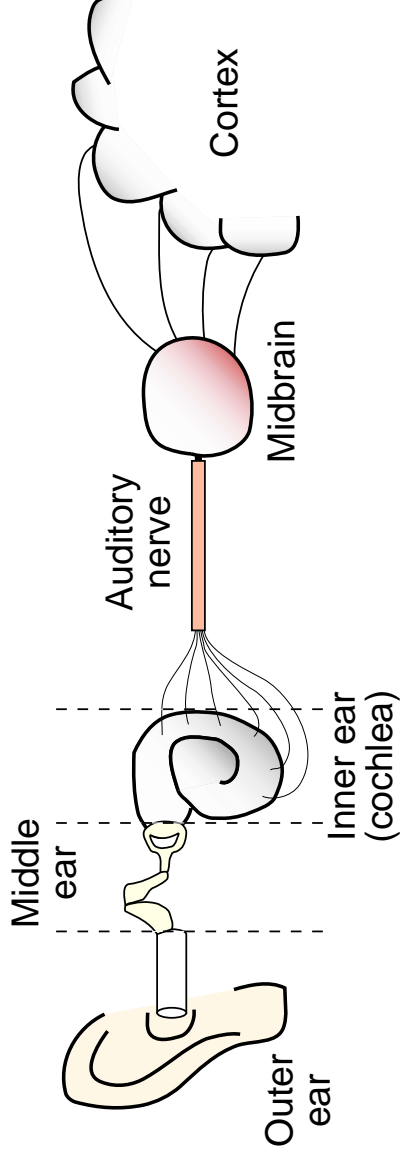
- 1 Motivation**
- 2 Physiology**
  - Outer, middle & inner ear
  - The Auditory Nerve and beyond
  - Models
- 3 Psychophysics**
- 4 Pitch perception**
- 5 Speech perception**
- 6 Scene analysis**



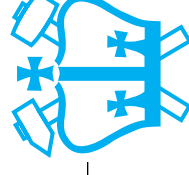
## 2

# Physiology

- Processing chain from air to brain:



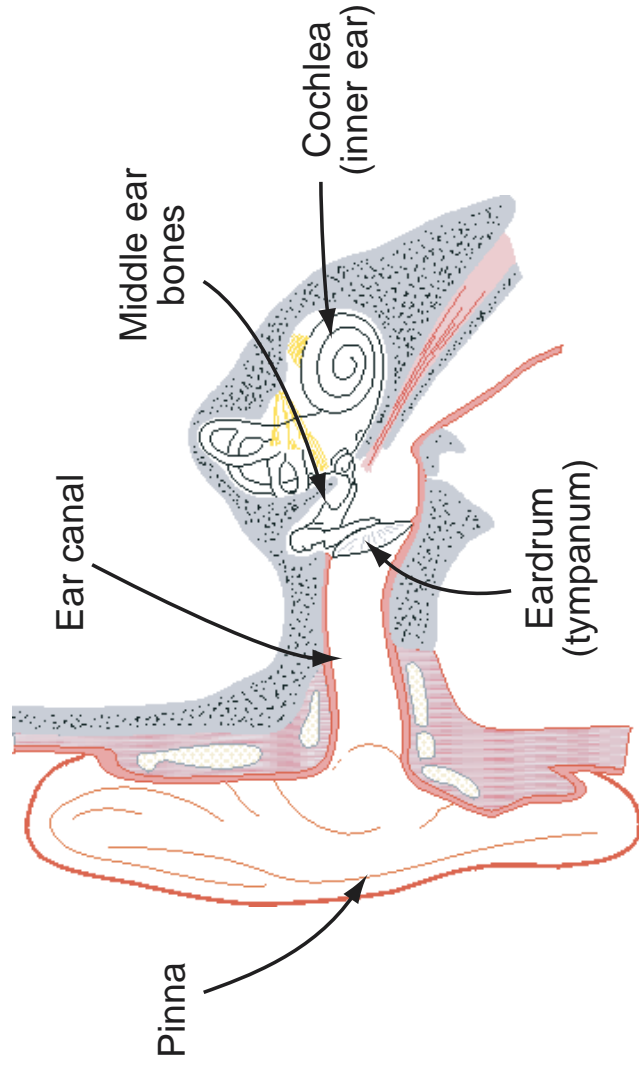
- **Study via:**
  - anatomy
  - nerve recordings
- **Signals flow in both directions**



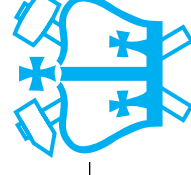
---

---

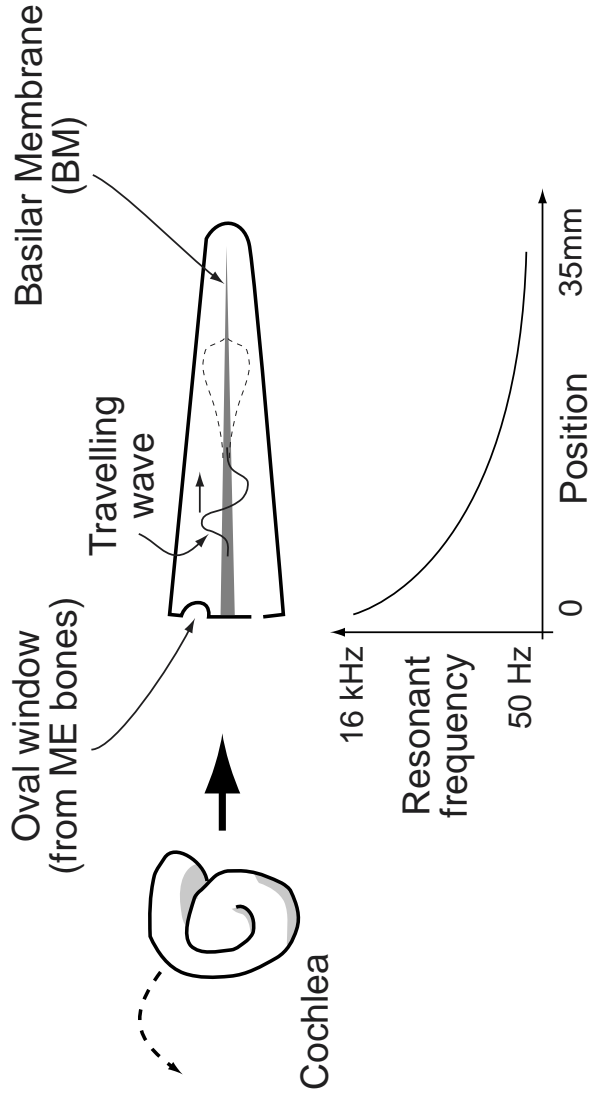
## Outer & middle ear



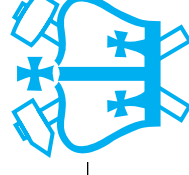
- **Pinna 'horn'**
  - complex reflections give spatial (elevation) cues
- **Ear canal**
  - **acoustic tube**
- **Middle ear**
  - bones provide impedance matching



# Inner ear: Cochlea



- **Mechanical input from middle ear starts traveling wave moving down Basilar Membrane**
- **Varying stiffness and mass of BM gives results in continuous variation of resonant frequency**
- **At resonance, traveling wave energy is dissipated in BM vibration**  
→ **Frequency (Fourier) analysis**

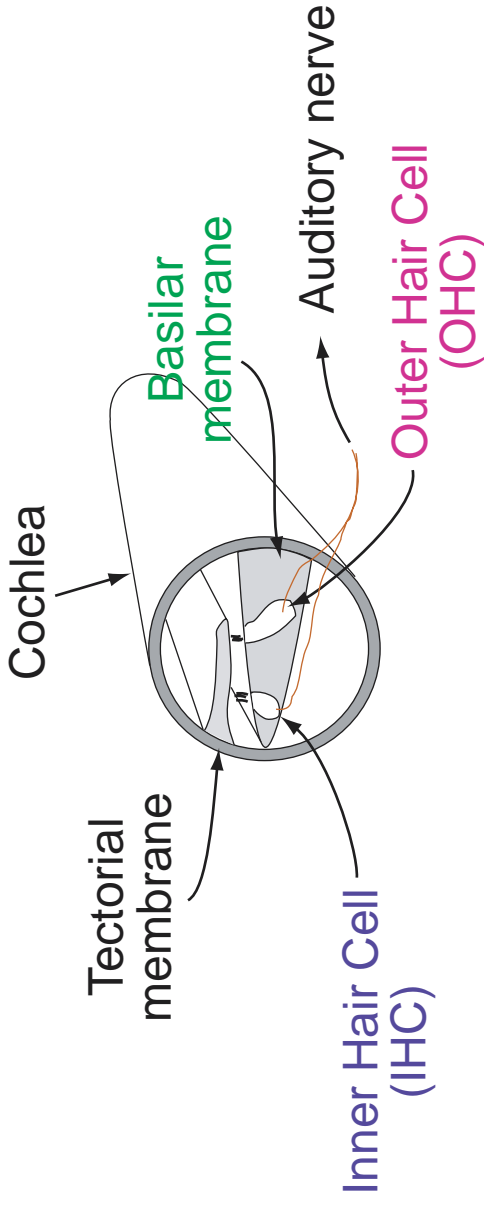


---

---

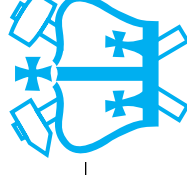
## Cochlea hair cells

- Ear converts sound to **BM** motion;  
Each point on **BM** corresponds to a frequency



- Hair cells on **BM** convert motion into nerve impulses (firings)
- Inner Hair Cells detect motion
- **Outer Hair Cells?** Variable damping?

**[BM animation]**



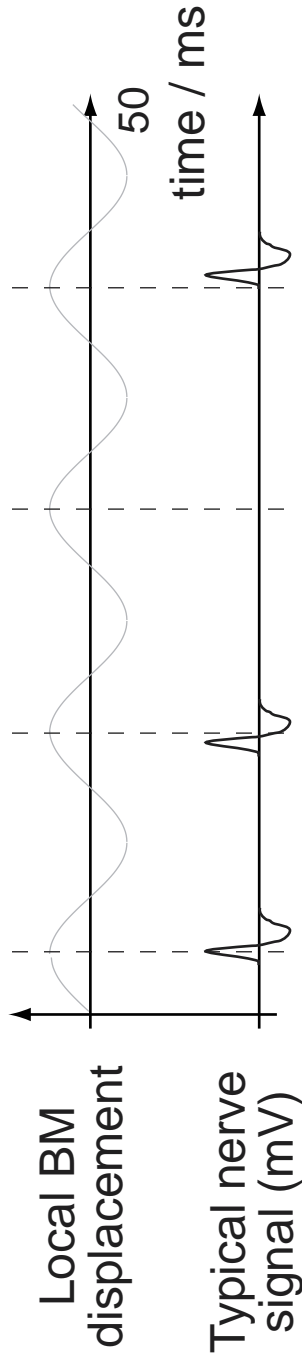


---

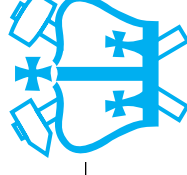
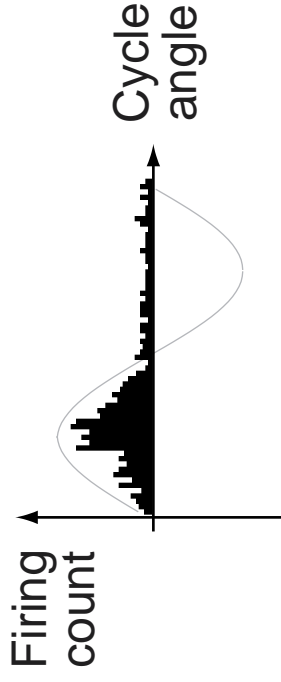
---

## Inner Hair Cells

- IHCs convert BM vibration into **nerve firings**
- Human hear has ~3500 IHCs;  
Each IHC has ~7 connections to Auditory Nerve
- Each nerve **fires** (sometimes) near peak displacement:



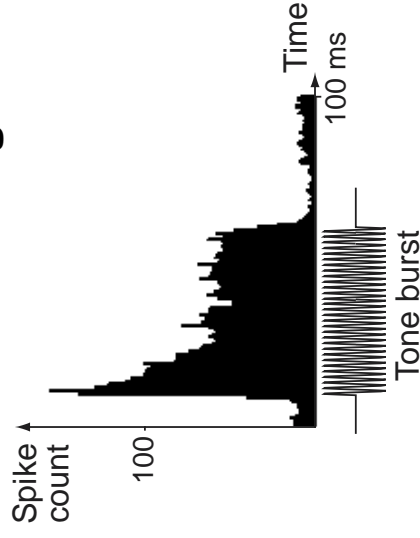
- **Histogram to get firing probability:**



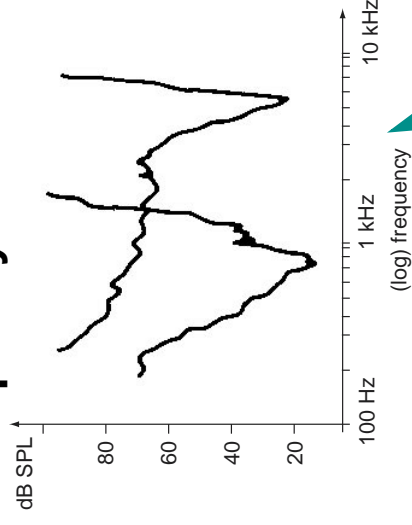
# Auditory nerve (AN) signals

- Single nerve measurements:

## Tone burst histogram

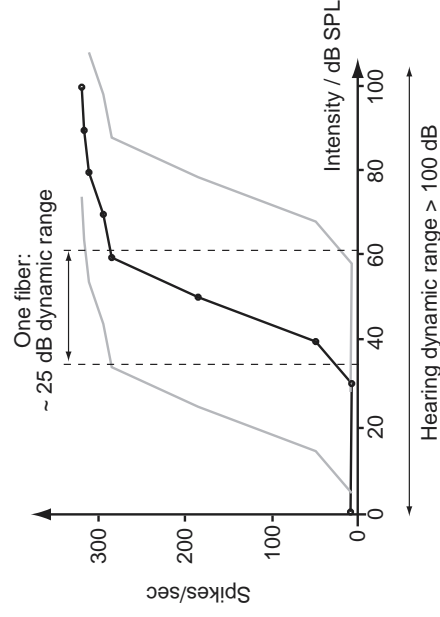


## Frequency threshold

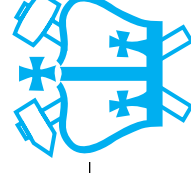


(approx. constant-Q)

## Rate vs. intensity

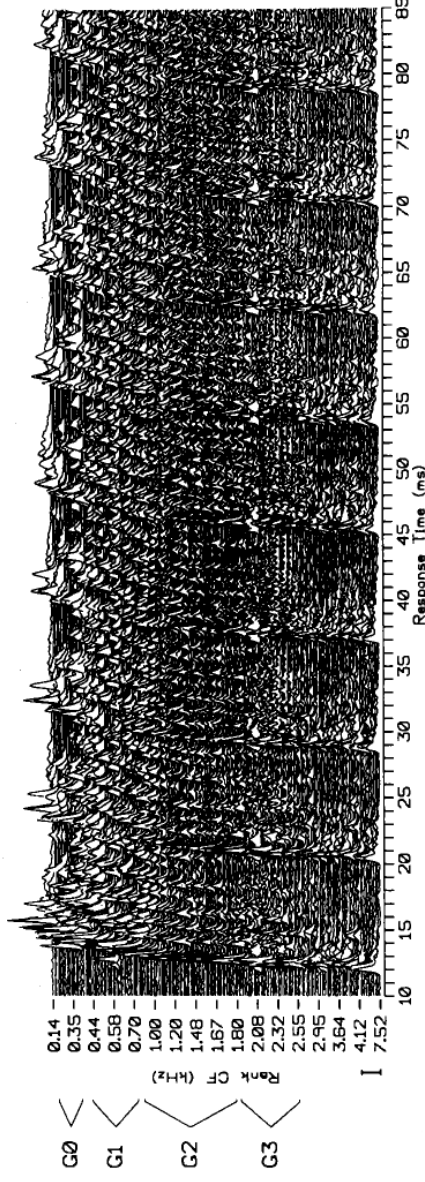


- Hard to measure: probe living ANs?

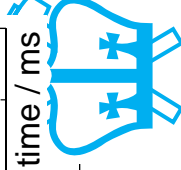
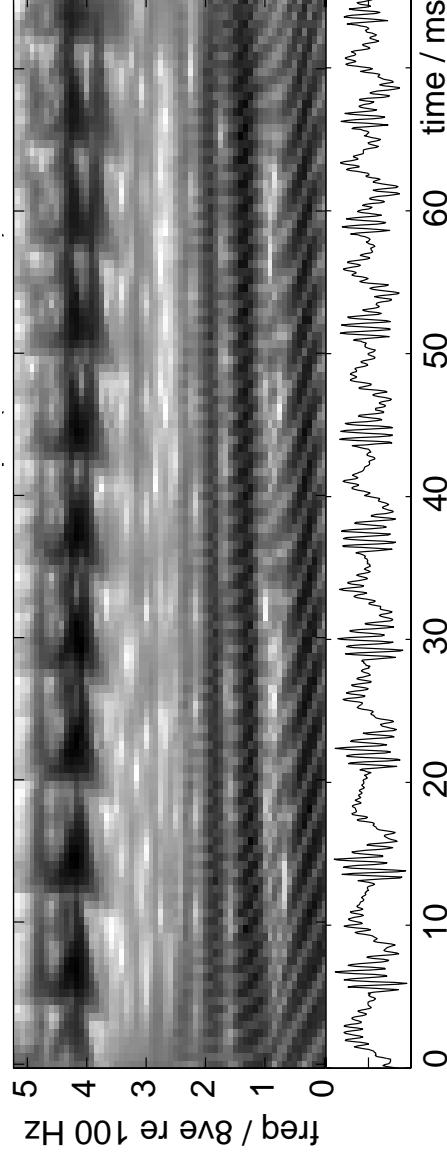


## AN population response

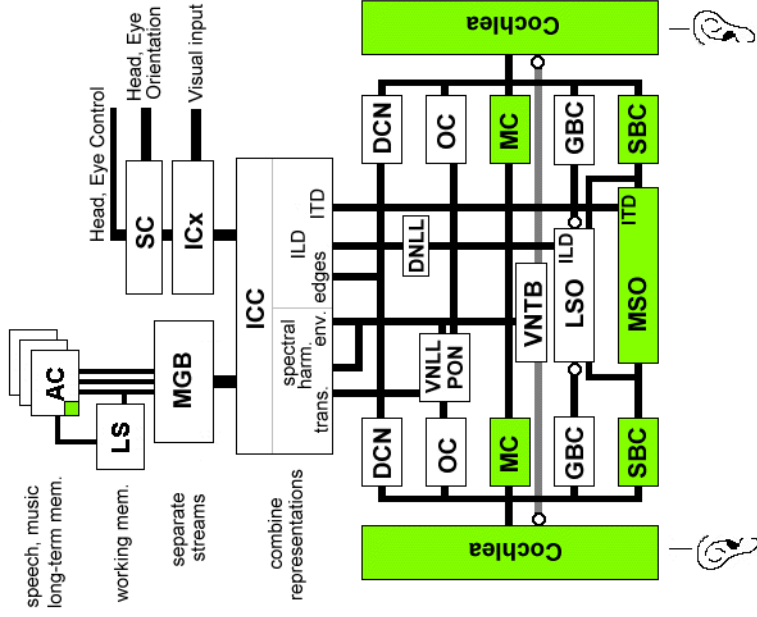
- All the information the brain has about sound:
  - average rate & spike timings on 30,000 fibers



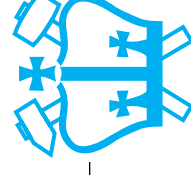
- Not unlike a (constant-Q) spectrogram?



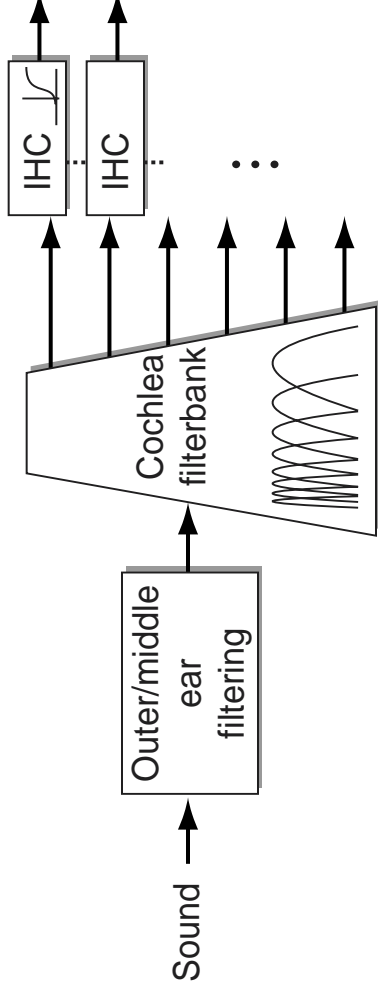
# Beyond the auditory nerve



- **Ascending and descending**
- **Tonotopic × ?**
  - modulation - position - source??

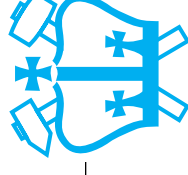
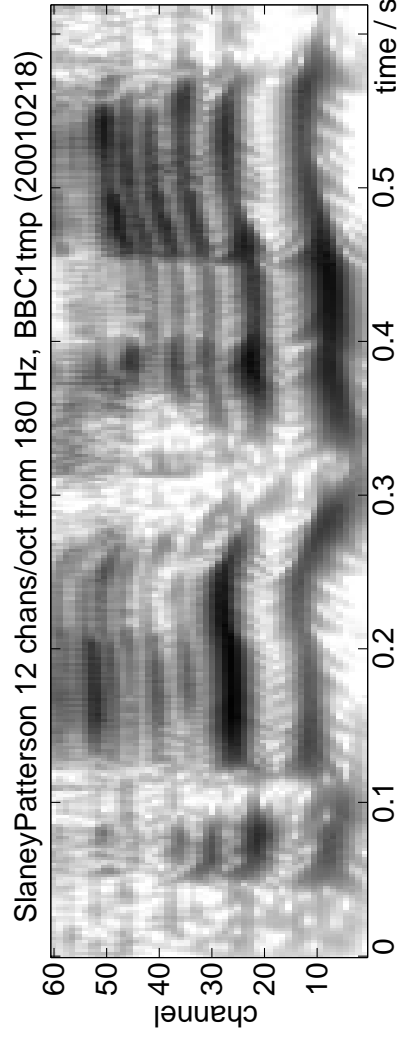


# Periphery models



- **Modeled aspects:**
  - outer/middle ear
  - cochlea filtering
  - hair cell transduction
  - efferent feedback?

- **Result: 'neurogram' / 'cochleagram'**

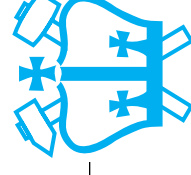


---

---

# Outline

- 1 Motivation
- 2 Physiology
- 3 **Psychophysics**
  - Detection theory modeling
  - Intensity perception
  - Masking
- 4 Pitch perception
- 5 Speech perception
- 6 Scene analysis



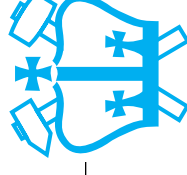
---

---

# 3

## Psychophysics

- **Physiology** looks at the implementation;  
**Psychology** looks at the function/behavior
- **Analyze audition as signal detection:  $p(\omega|O)$** 
  - psychological tests reflect internal decisions
  - assume optimal decision process
  - infer nature of internal representations, noise, ...  
→ lower bounds on more complex functions
- **Different aspects to measure**
  - time, frequency, intensity
  - tones, complexes, noise
  - binaural
  - pitch, detuning



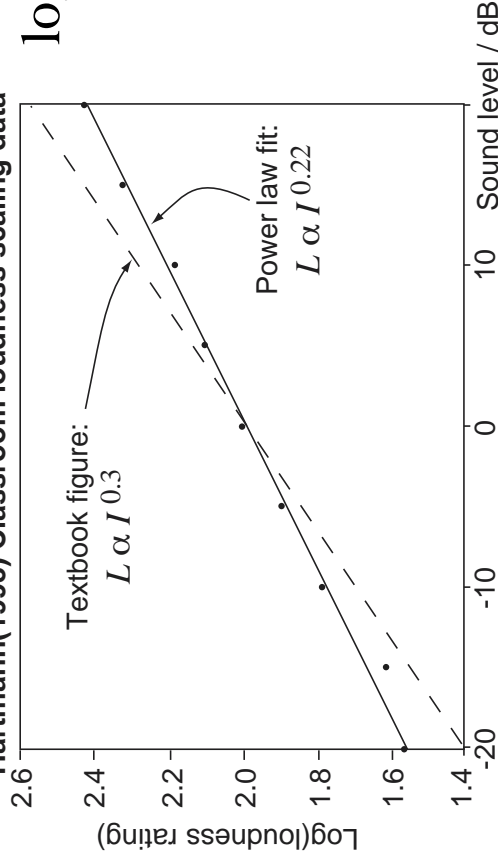
---

---

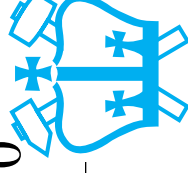
## Basic psychophysics

- Relate **physical** and **perceptual** variables
  - e.g. **intensity** → loudness  
**frequency** → pitch
- **Methodology: subject tests**
  - just noticeable difference (jnd)
  - magnitude scaling e.g. ‘adjust to twice as loud’
- **Results for Intensity vs. Loudness:**  
**Weber’s law**  $\Delta I \propto I \rightarrow \log(L) = k \cdot \log(I)$

Hartmann(1993) Classroom loudness scaling data



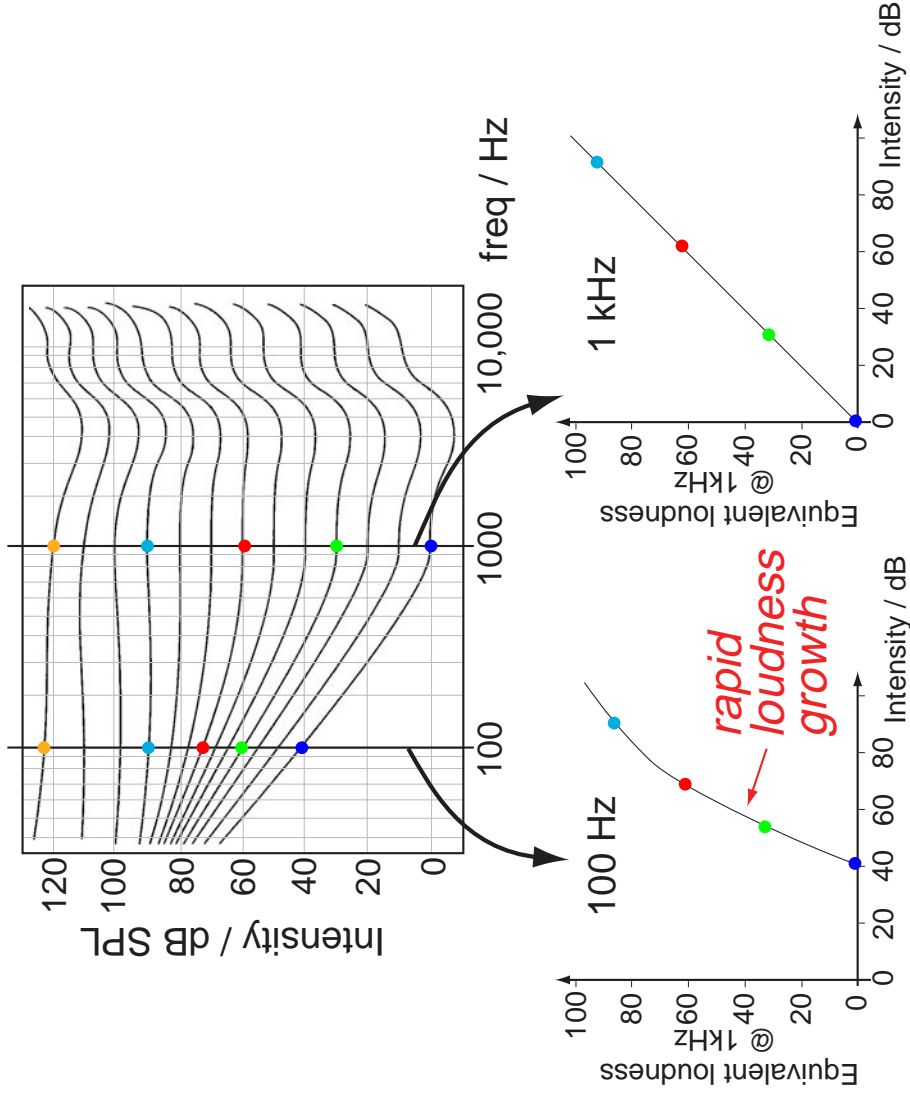
$$\begin{aligned}\log_2(L) &= 0.3 \log_2(I) \\ &= 0.3 \cdot \frac{\log_{10} I}{\log_{10} 2} \\ &= \frac{0.3}{\log_{10} 2} \cdot \frac{dB}{10} \\ &= dB / 10\end{aligned}$$



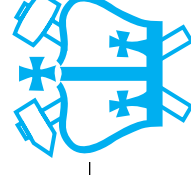


# Loudness as a function of frequency

- Fletcher-Munson equal-loudness curves:

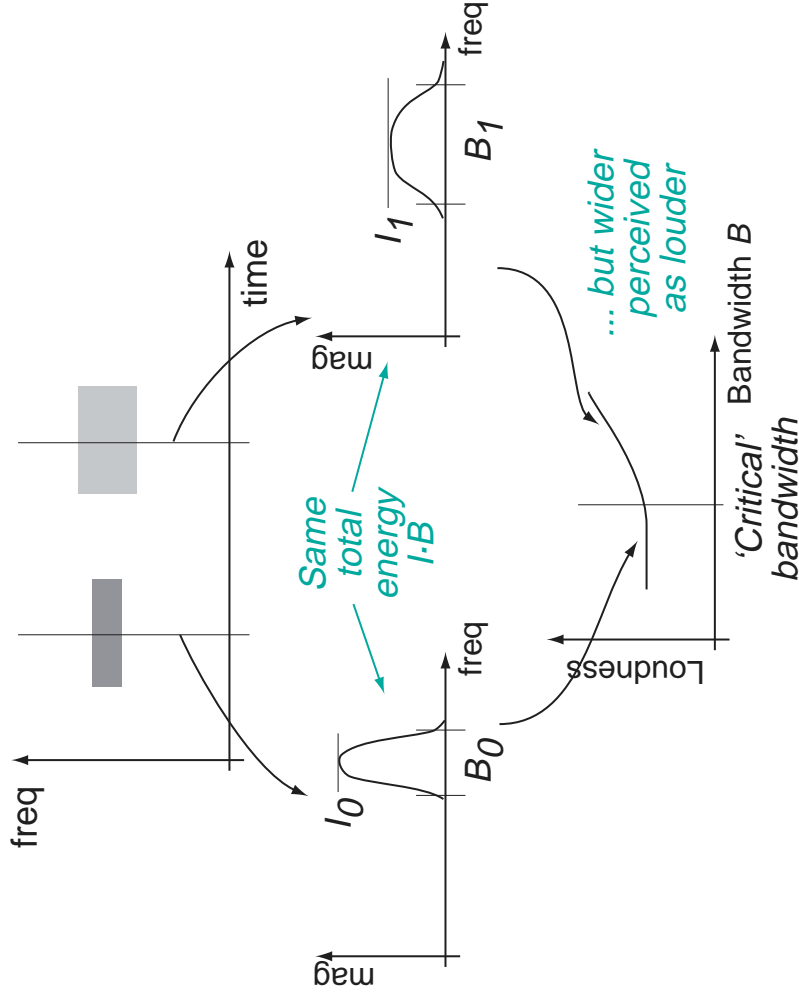


- Hearing impairment: exaggerates



# Loudness as a function of bandwidth

- Same total energy, different distribution:

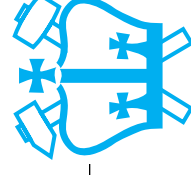


- e.g. 2 chans at -6 dB (not -10 dB)

- **Critical bands: independent freq. channels**

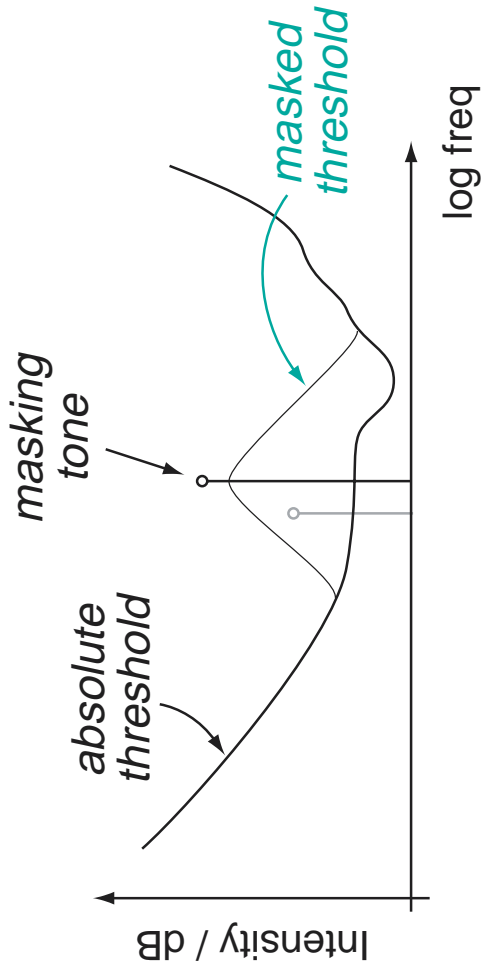
- ~ 25 total (4-6 / octave)

[sindex]

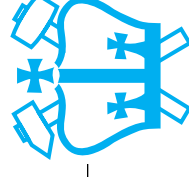
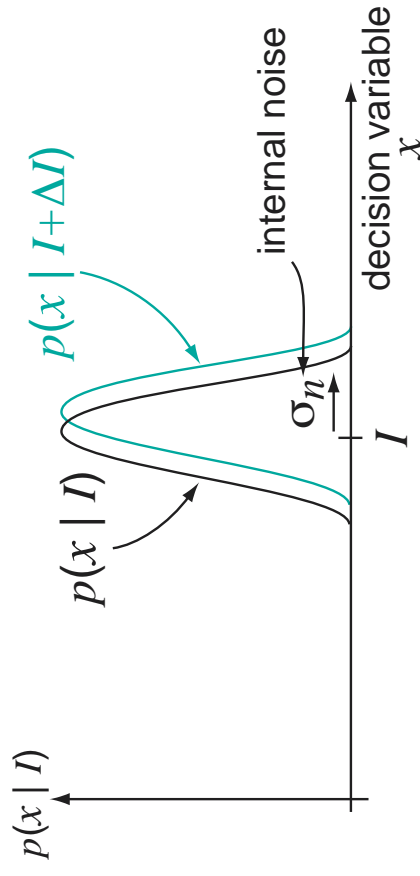


## Simultaneous masking

- A louder tone can **'mask'** the perception of a second tone nearby in frequency:

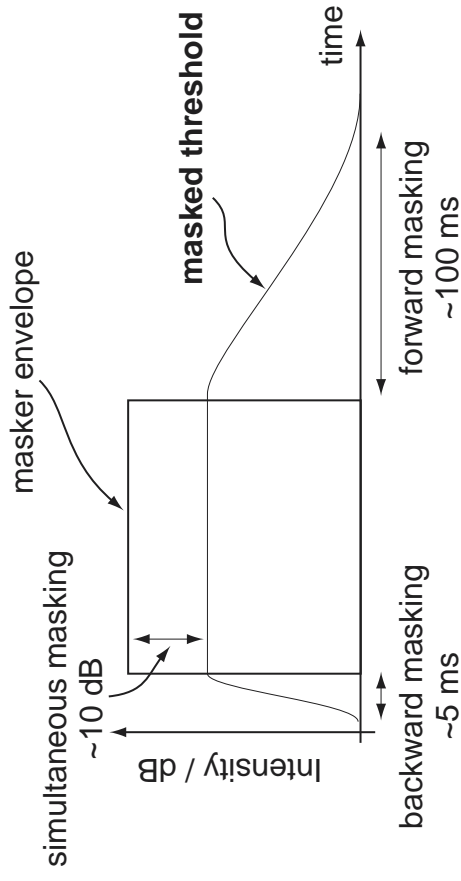


- Suggests an **'internal noise'** model:



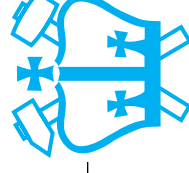
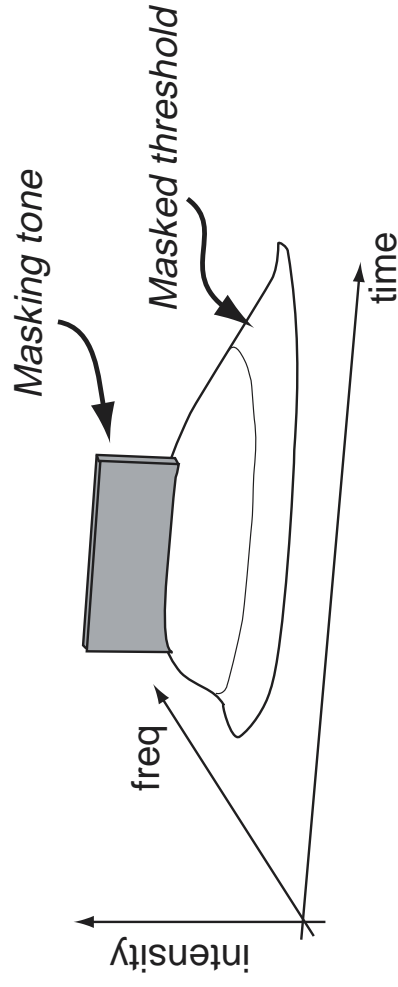
# Sequential masking

- **Backward/forward in time:**



- suggests temporal envelope of decision var.

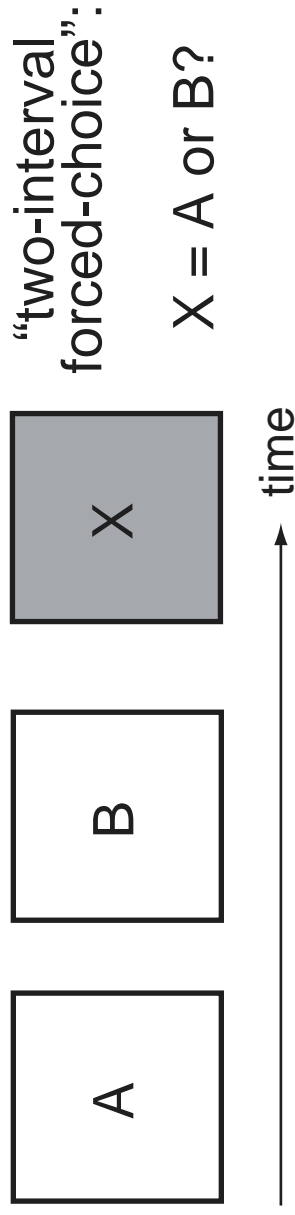
→ **Time-frequency masking 'skirt':**



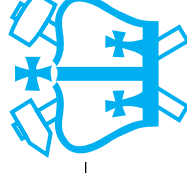
---

---

## What we do and don't hear



- **Timing: 2ms attack resolution, 20ms discrim**
  - but: spectral splatter
- **Tuning: ~ 1% discrimination**
  - but: beats
- **Spectrum: profile changes, formants**
  - variable time-frequency resolution
- **Harmonic phase?**
- **Noisy signals & texture**
- **(Trace vs. categorical memory)**

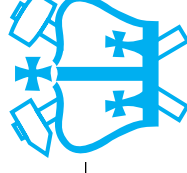


---

---

# Outline

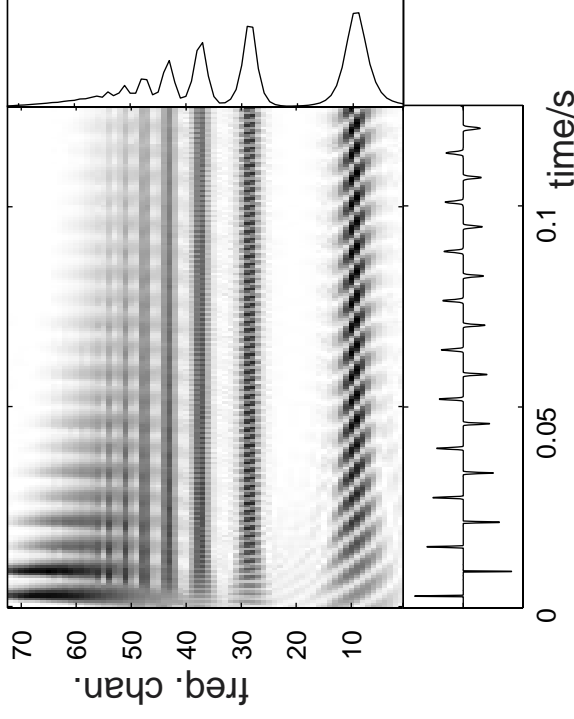
- 1 **Motivation**
- 2 **Physiology**
- 3 **Psychophysics**
- 4 **Pitch perception**
  - 'Place' models
  - 'Time' models
  - Multiple cues & competition
- 5 **Speech perception**
- 6 **Scene analysis**



# 4

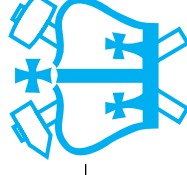
## Pitch perception: A classic argument in psychophysics

- Harmonic complexes are a pattern on AN



- .. but give a *fused* percept (ecological)
- **What determines the pitch percept?**
  - *not* the fundamental
- **How is it computed?**

**Two competing models: *place* and *time***

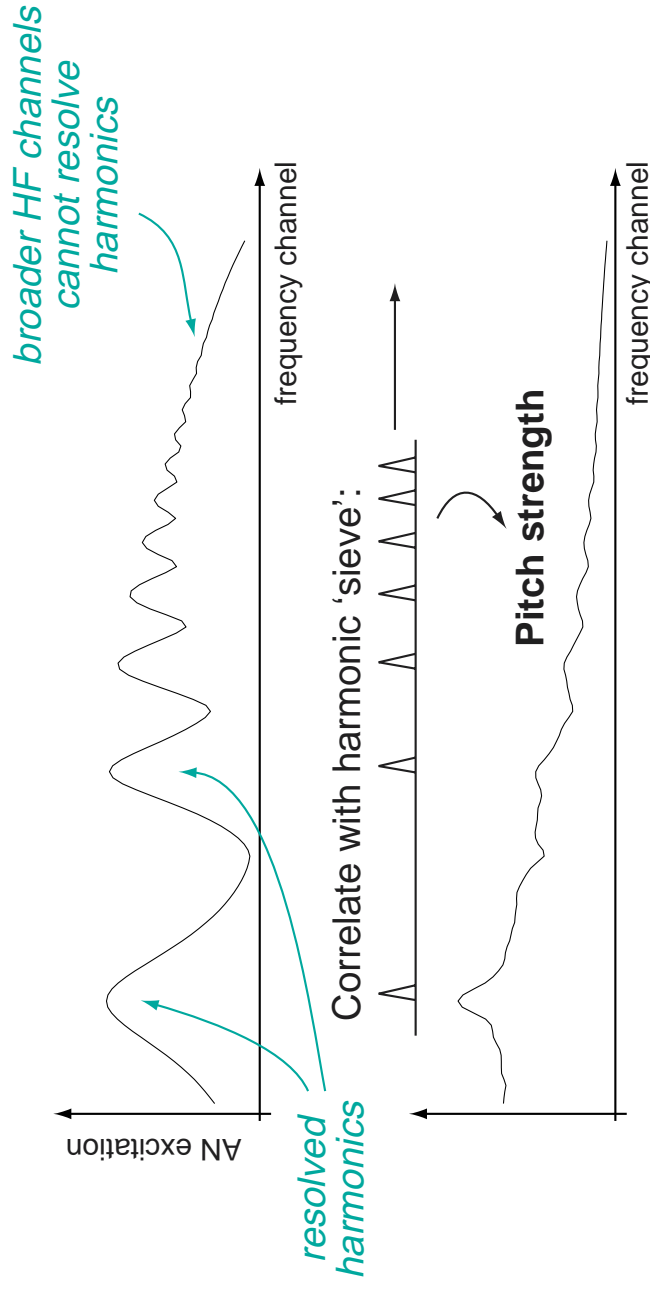


---

---

## Place model of pitch

- AN excitation pattern shows individual peaks
- ‘Pattern matching’ method to find pitch:



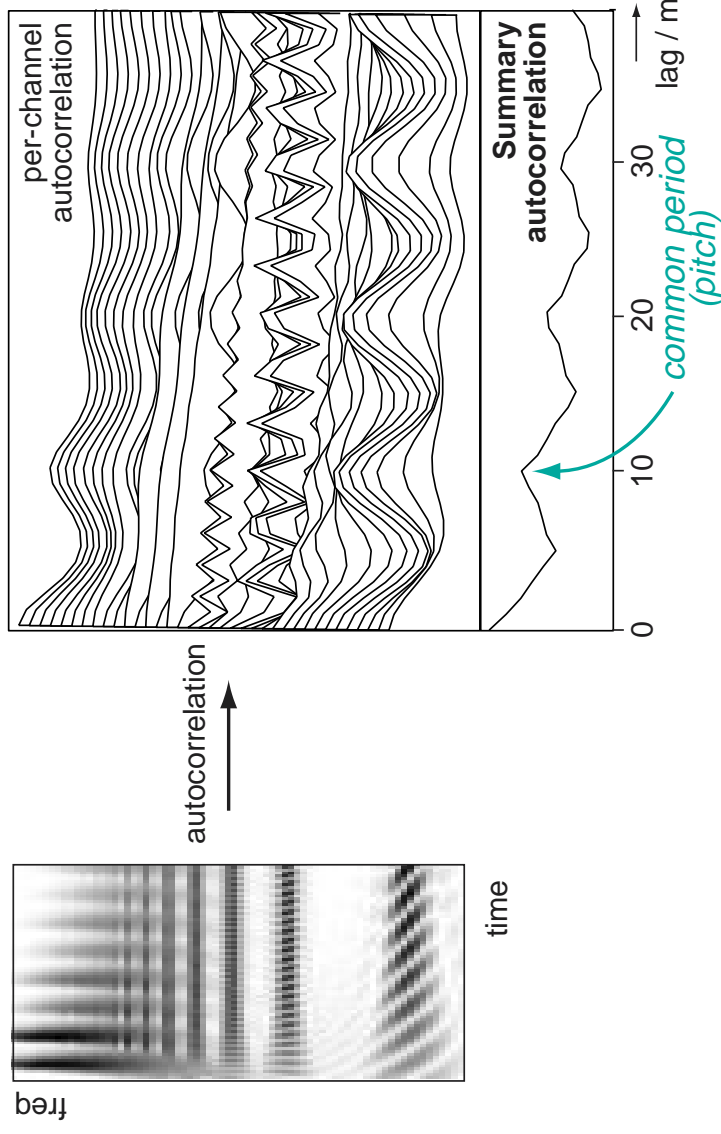
- **Support:**  
**Low harmonics are very important**
- **But: Flat-spectrum noise can carry pitch**



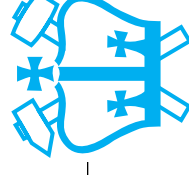


## Time model of pitch

- Timing information is preserved in AN down to ~ 1ms scale
- Extract periodicity by e.g. autocorrelation & combine across frequency chans:



- **But: HF gives weak pitch (in practice)**



---

---

## Alternate & competing cues

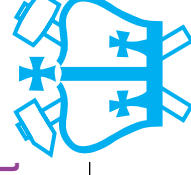
- **Pitch perception could rely on various cues**
  - average excitation pattern
  - summary autocorrelation
  - more complex pattern matching
- **Relying on just one cue is brittle**
  - e.g. missing fundamental

→ **Perceptual system appears to use a flexible, opportunistic combination**

- **Optimal detector justification?**

$$\begin{aligned} & \operatorname{argmax}_{\omega} p(\omega | \mathbf{o}) \\ &= \operatorname{argmax}_{\omega} \frac{p(\mathbf{o} | \omega) \cdot p(\omega)}{p(\mathbf{o})} \\ &= \operatorname{argmax}_{\omega} p(o_1 | \omega) \cdot p(o_2 | \omega) \cdot p(\omega) \end{aligned}$$

if  $o_1$  and  $o_2$  are *conditionally independent*

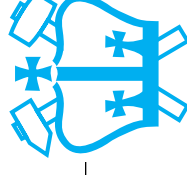


---

---

# Outline

- 1 Motivation**
- 2 Physiology**
- 3 Psychophysics**
- 4 Pitch perception**
- 5 Speech perception**
  - The sounds of speech
  - Phoneme perception
  - Context and top-down influences
- 6 Scene Analysis**



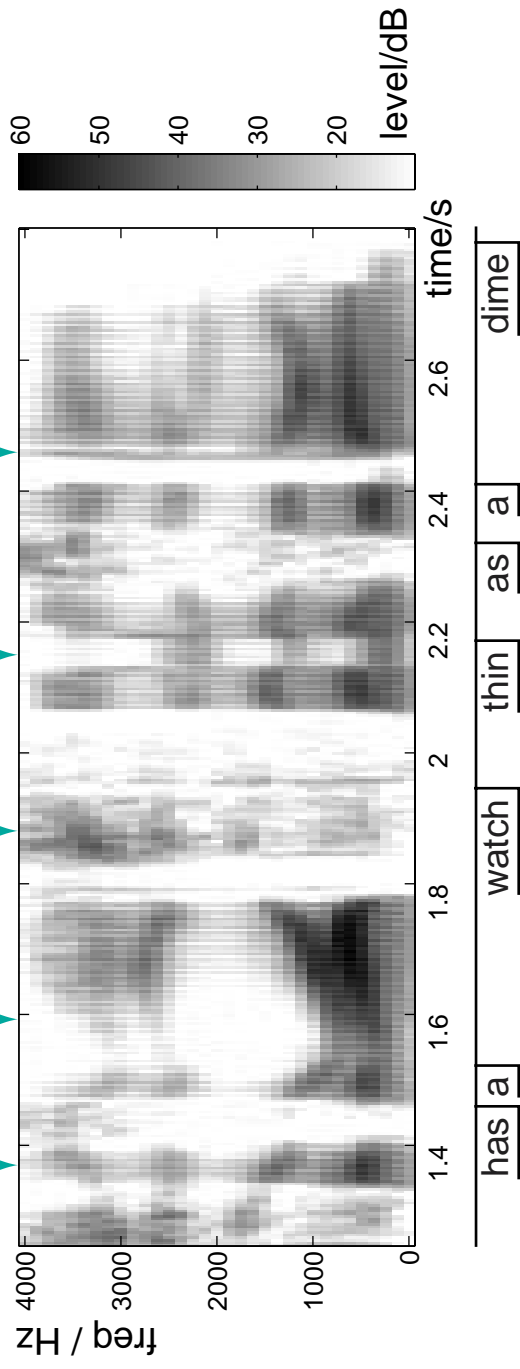
# 5

## Speech perception

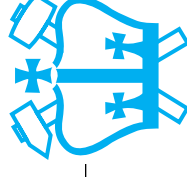
- **Highly specialized function**
  - subsequent to source organization?
  - .. but also can interact

- **Kinds of speech sounds:**

*vowel* ↓ *glide* ↓ *fricative* ↓ *nasal* ↓ *stop burst* ↓

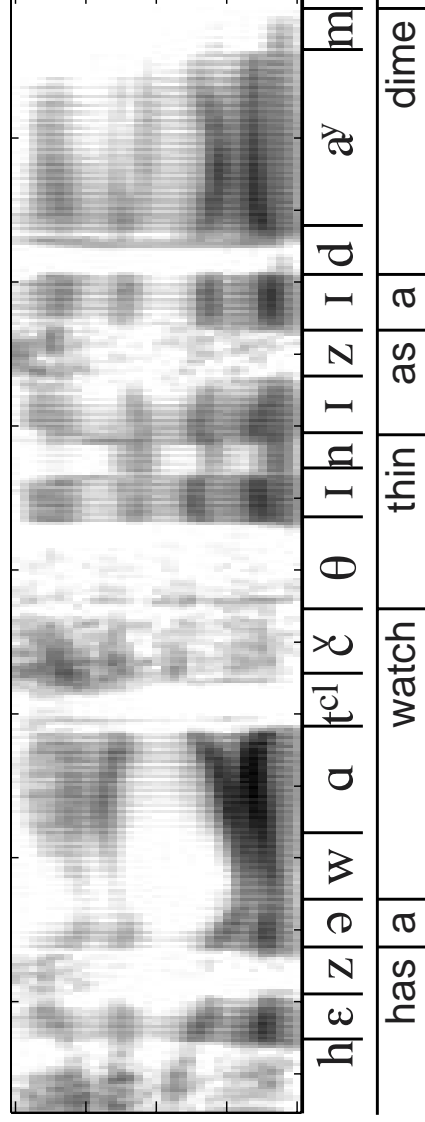


- vowels - glides - nasals
- stops - fricatives ...



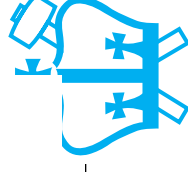
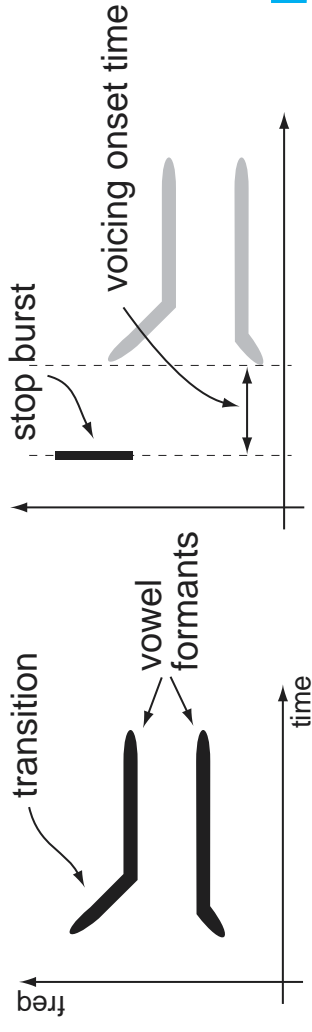
# Cues to phoneme perception

- Linguists describe speech with *phonemes*:



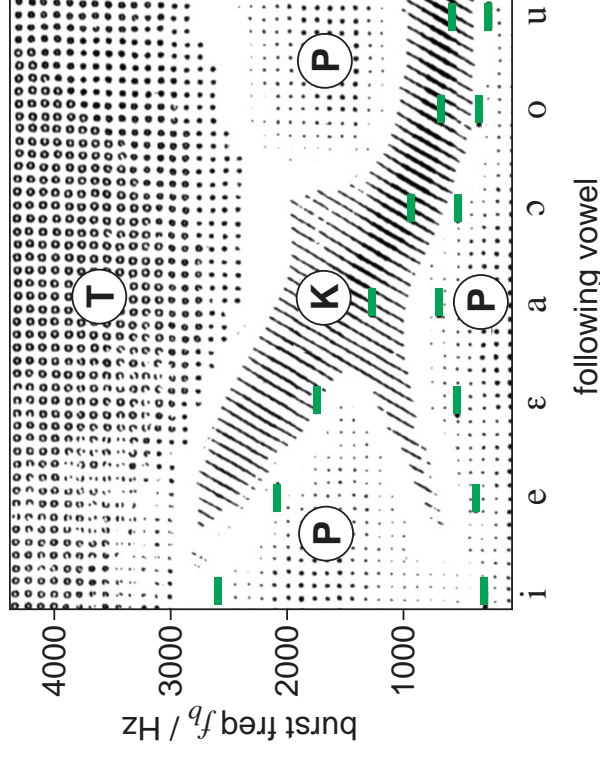
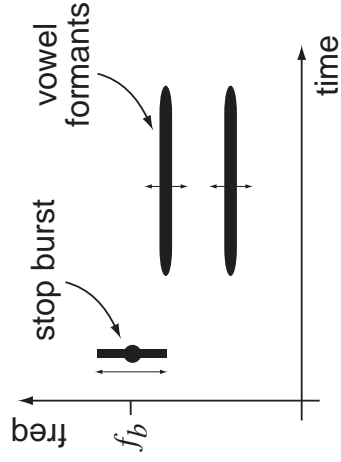
- phonemes define minimal word contrasts

- Acoustic-phoneticians describe phonemes by:
  - formants & transitions
  - bursts & onset times

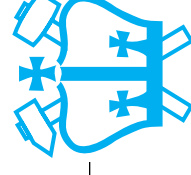


# Categorical perception

- (Some) speech sounds perceived *categorically* rather than *analogically*
  - e.g. stop-burst & timing:



- tokens within category are hard to distinguish
- category boundaries are very sharp
- **Categories are learned for native tongue**
  - “merry” / “mary” / “marry”

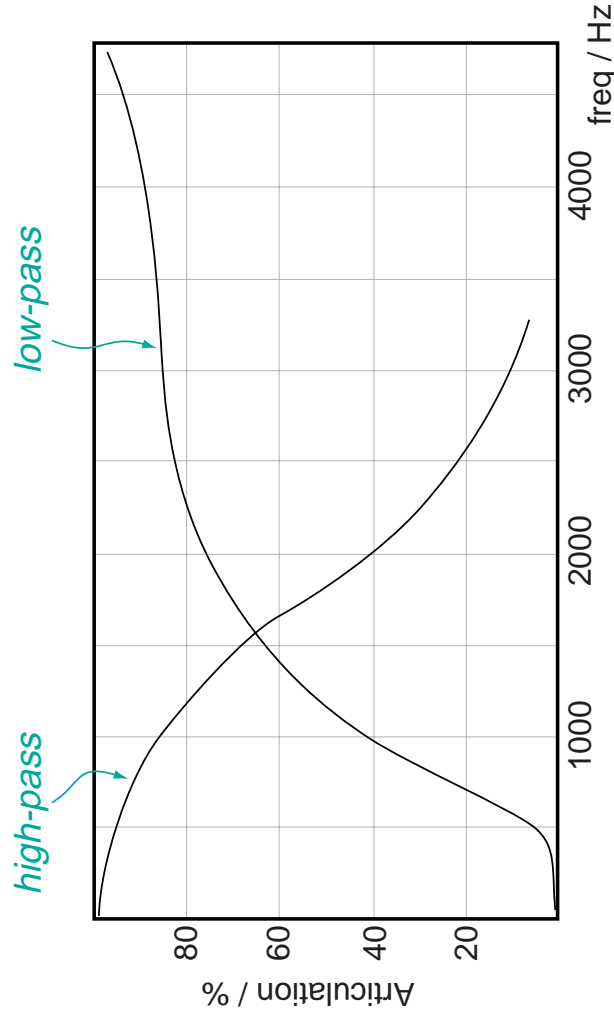


---

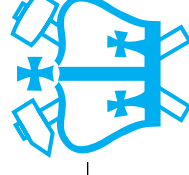
---

## Where is the information in speech?

- ‘Articulation’ of high/low-pass filtered speech:



- sums to more than 1...
- **Speech message is highly redundant**
  - e.g. constraints of language, context
  - listeners can understand with *very few cues*

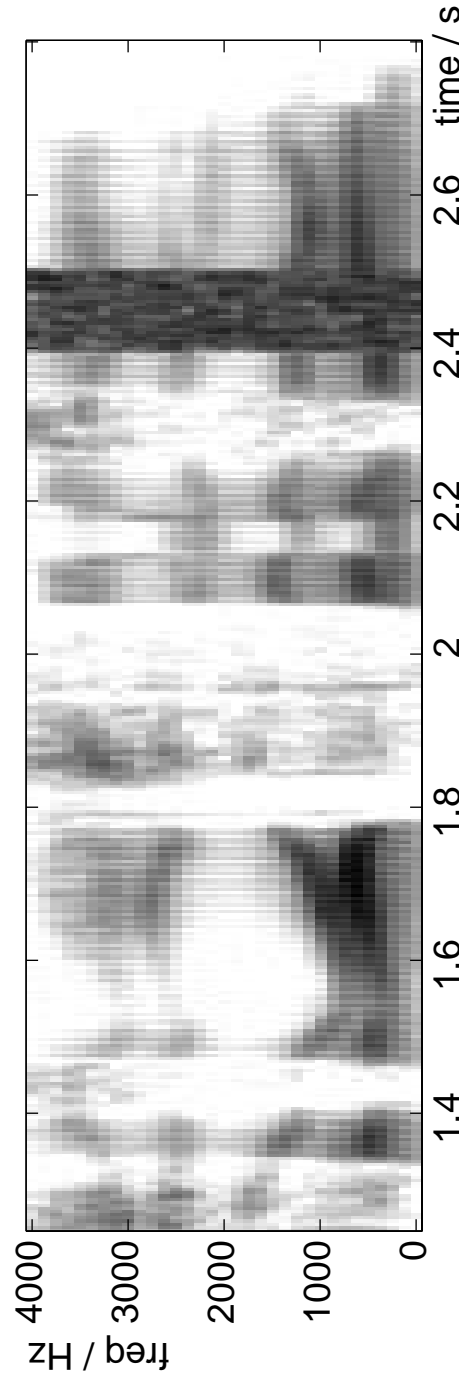


---

---

## Top-down influences: Phonemic restoration (Warren 1970)

- **What if a noise burst obscures speech?**

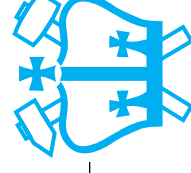


- auditory system 'restores' the missing phoneme

- ... based on **semantic** context

- ... even in **retrospect!**

- **Subjects are typically unaware of which sounds are restored**





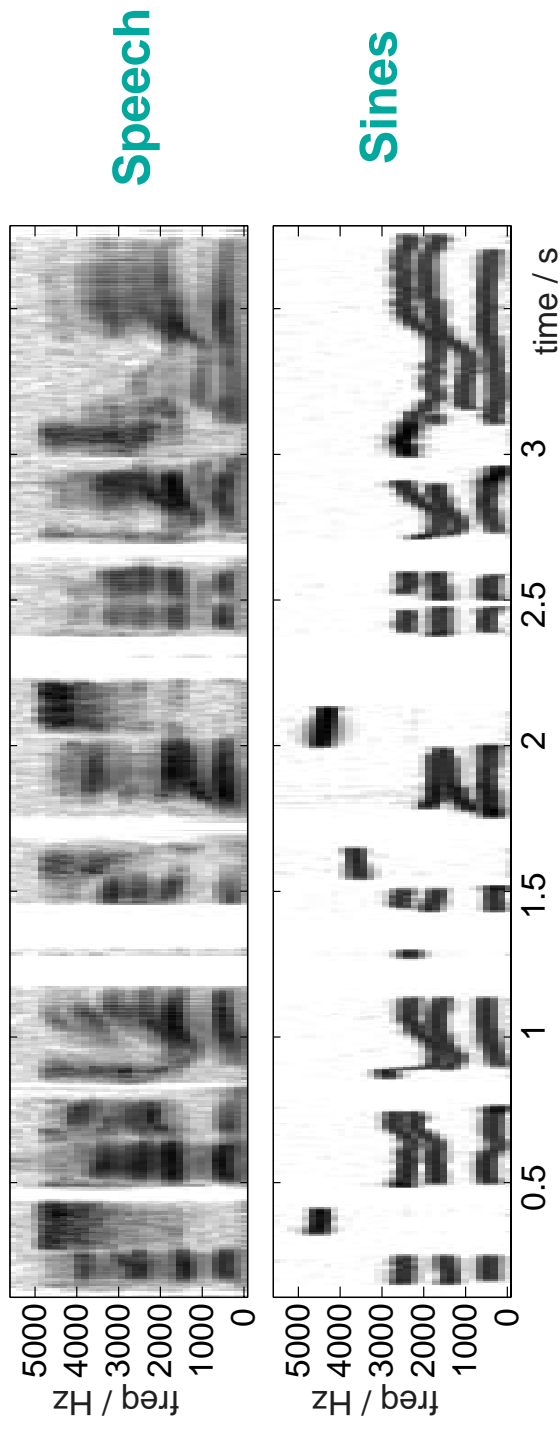
---

---

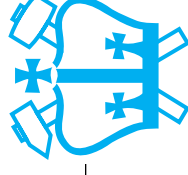
# A predisposition for speech: Sinewave replicas

(Remez et al. 1994)

- Replace each formant with a **single sinusoid**:

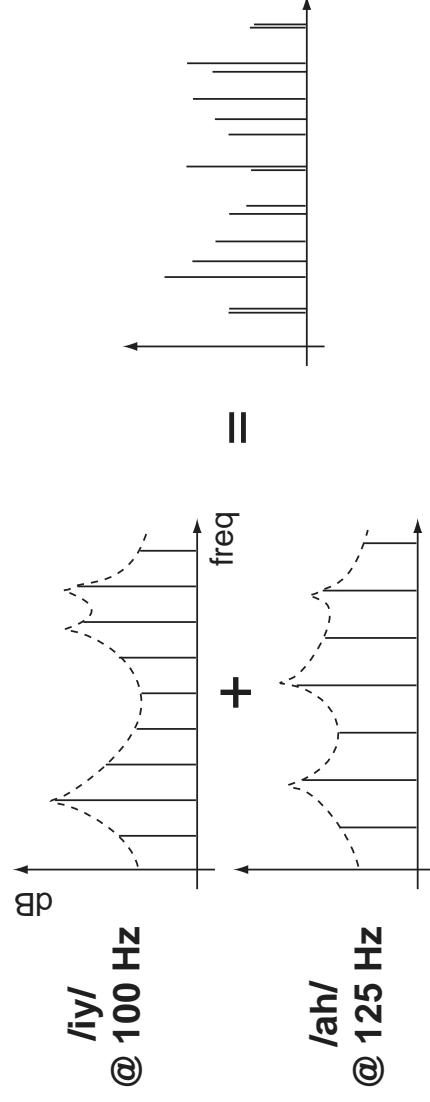


- speech is (somewhat) intelligible
  - people hear both whistles and speech (“duplex”)
  - processed as speech despite un-speech-like
- **What does it take to be speech?**

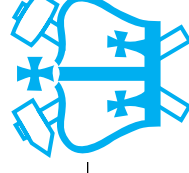
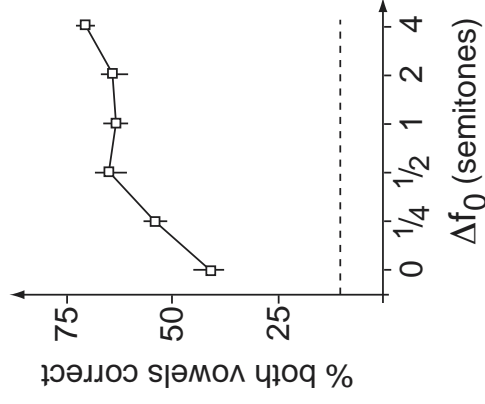


## Simultaneous vowels

- Mix synthetic vowels with different  $f_0$ s:



- Pitch difference helps (though not necessary):

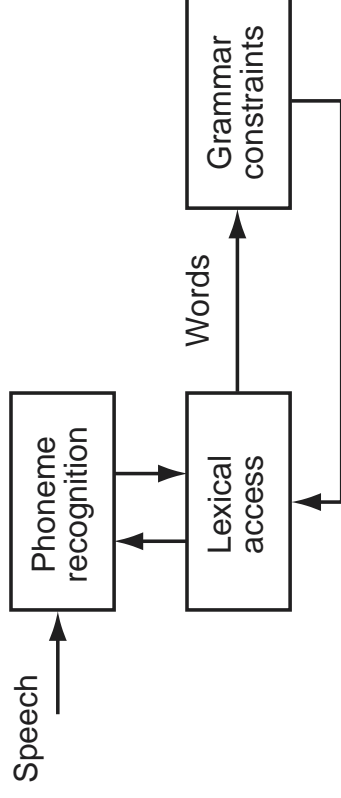


---

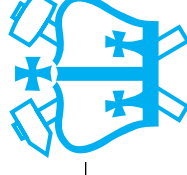
---

# Computational models of speech perception

- Various theoretical - practical models of speech comprehension, e.g.:



- **Open questions:**
  - mechanism of phoneme classification
  - mechanism of lexical recall
  - mechanism of grammar constraints
- **ASR is a practical implementation (?)**

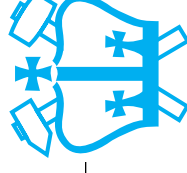


---

---

# Outline

- 1 Motivation
- 2 Physiology
- 3 Psychophysics
- 4 Pitch perception
- 5 Speech perception
- 5 **Scene analysis**
  - Events and sources
  - Fusion and streaming
  - Continuity & restoration
  - Simultaneous vowels

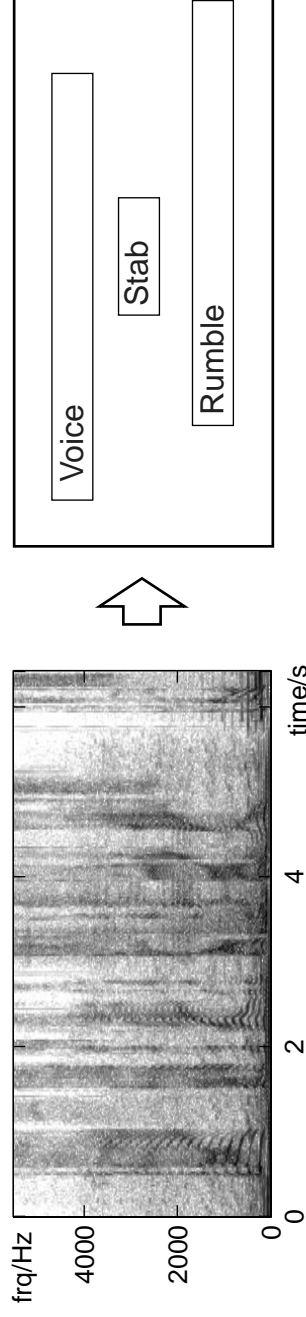


## 6

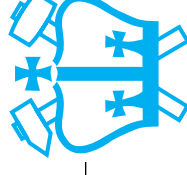
# Auditory Organization

- Detection model is huge simplification
- Real role of hearing is much more general:  
Recover useful information from outside world

→ **Sound organization** into events and sources:



- **Research questions:**
  - what determines perception of sources?
  - how do humans separate mixtures?
  - how much can we tell about a source?

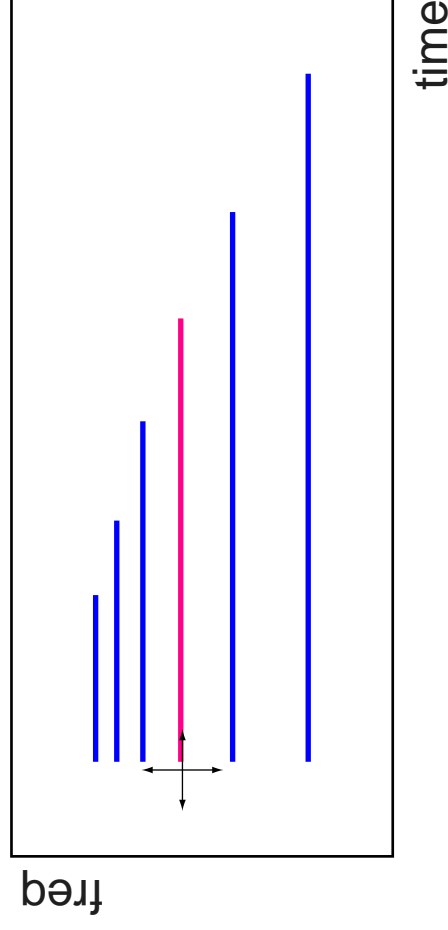


---

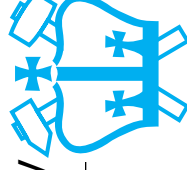
---

## Auditory scene analysis: Simultaneous fusion

- Harmonics are distinct on AN, but perceived as one sound (“fused”):



- depends on *common onset*
- depends on *harmonicity* (common period)
- **Methodologies:**
  - ask subject how many ‘objects’
  - match attributes e.g. object pitch
  - manipulate higher level e.g. vowel identity

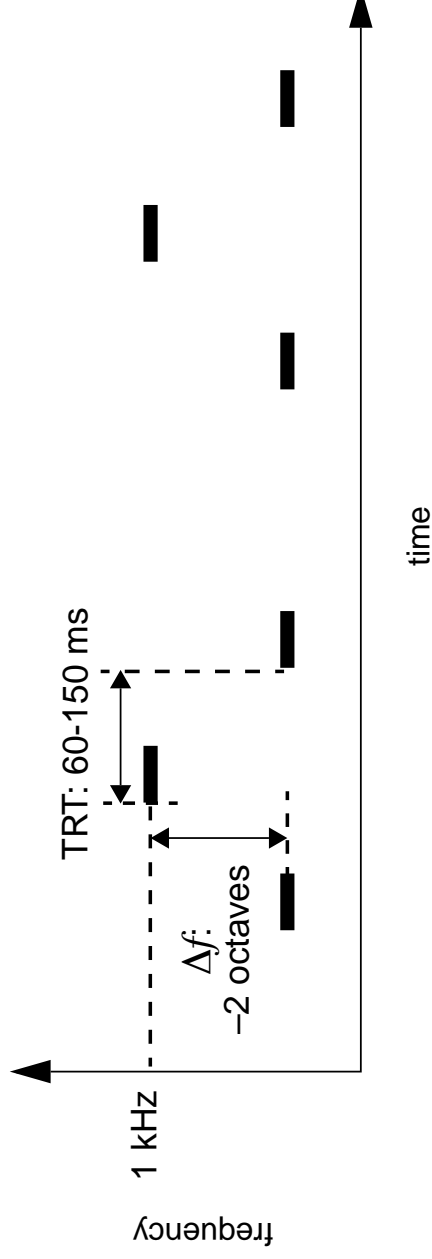


---

---

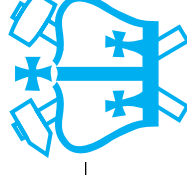
## Sequential grouping: streaming

- **Pattern / rhythm: property of set of objects**
  - subsequent to fusion :. employs fused events?



- **Measure by relative timing judgments**
  - cannot compare between streams
- **Separate 'coherence' and 'fusion' boundaries**
- **Can interact and compete with fusion**

[sindex]

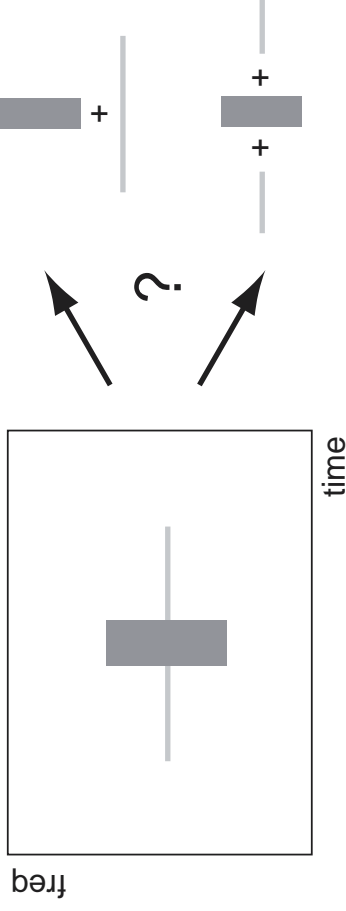


---

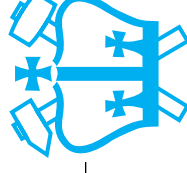
---

## Continuity & restoration

- Tone is interrupted by noise burst:  
What happened?



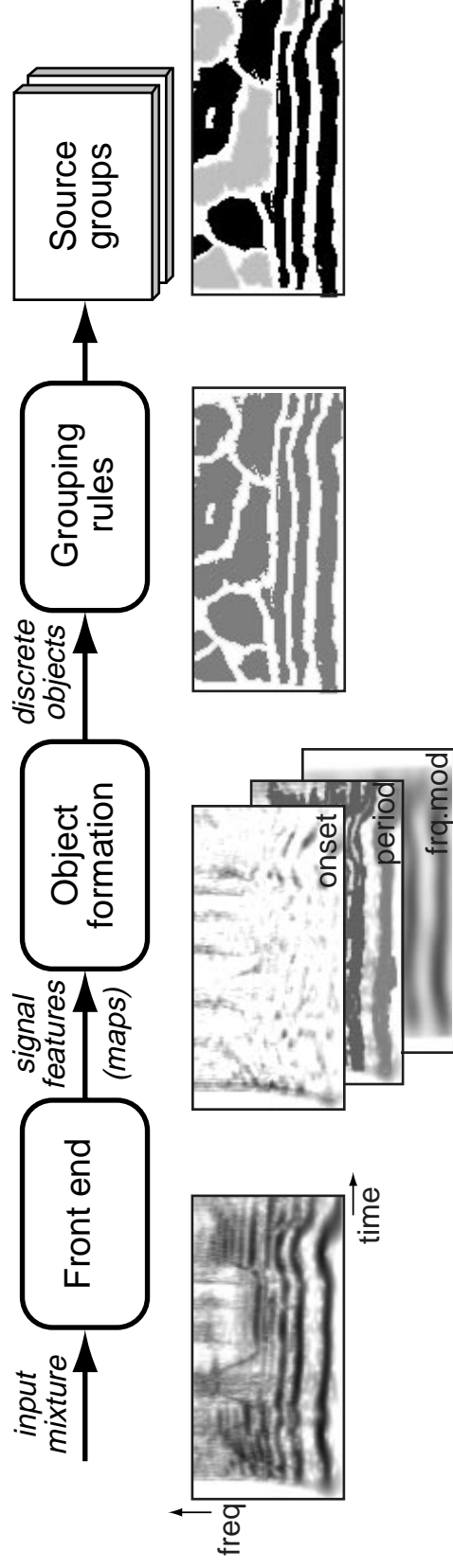
- masking makes tone undetectable during noise
- **Need to infer most probable real-world events**
  - observation equally likely for either explanation
  - *prior* on continuous tone much higher → choose
- **Top-down influence on perceived events...**  
**pulsation threshold [sndex]**





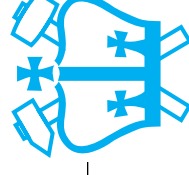
# Models of auditory organization

- Psychological accounts suggest bottom-up:



- (Brown 1991)
- **Complications in practice:**
  - formation of separate elements
  - contradictory cues
  - influence of top-down constraints (context, expectations)

...

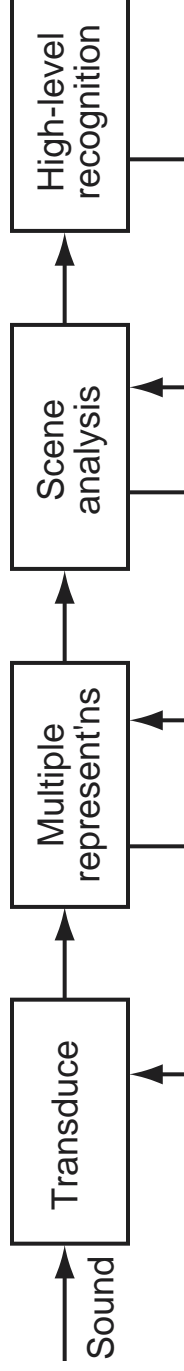


---

---

## Summary

- **Auditory perception** provides the ‘ground truth’ underlying audio processing
- **Physiology** specifies information available
- **Psychophysics** measures basic sensitivities
- **Sound sources** requires further **organization**
- **Strong contextual effects** in **speech** perception



**Parting thought:**

**Is pitch central to communication? Why?**

