Lecture 08:
[Rabiner] Hearing, auditory models, and speech perception

DEEE725 음성신호처리실습

Instructor: 장길진

Original slides from Lawrence Rabiner

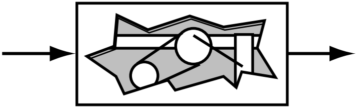
Why study perception?

- 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

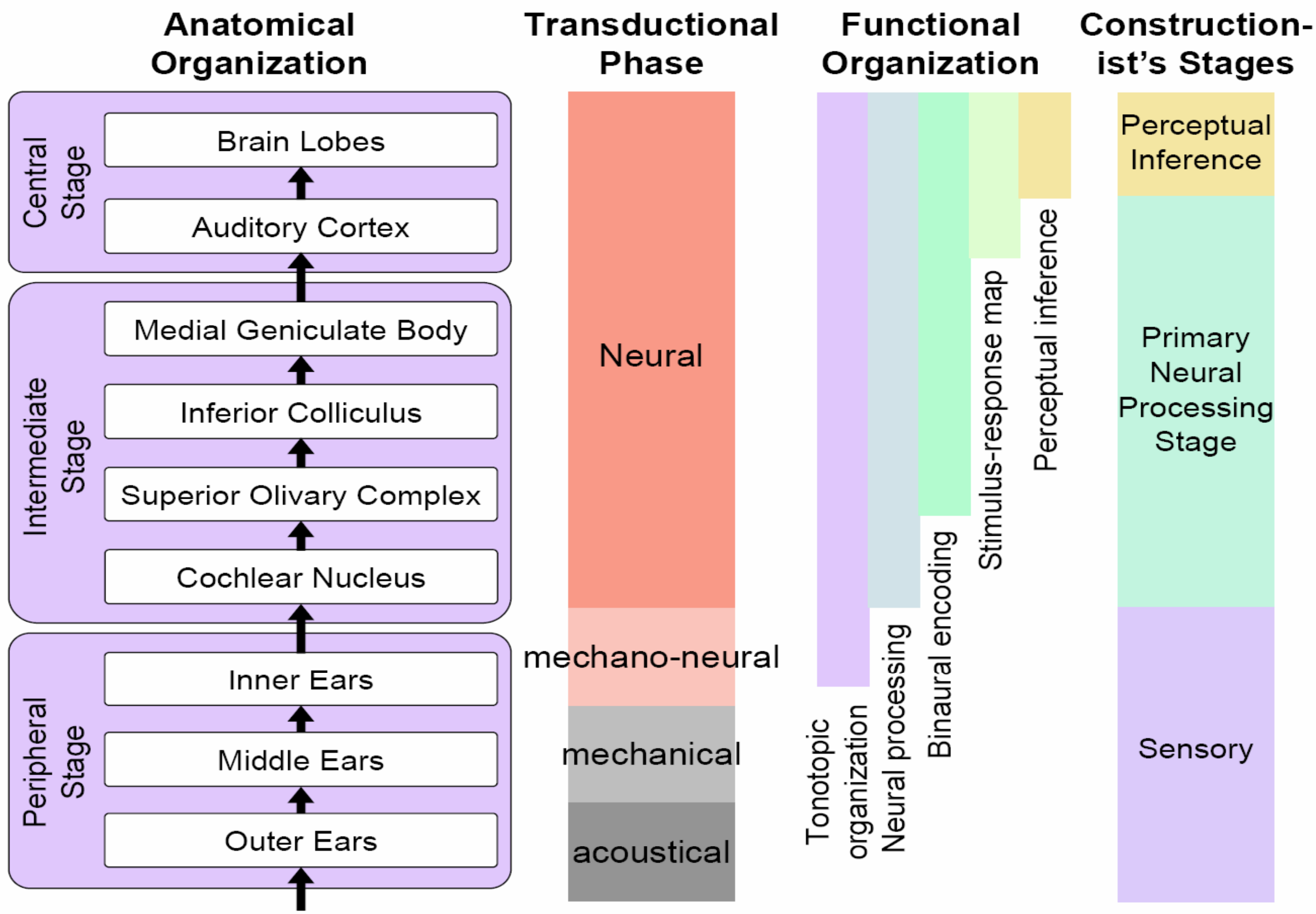
- Black box input/output: psychophysics

  - fit simple models of simple functions

- Information processing models
  - investigate and model complex functions

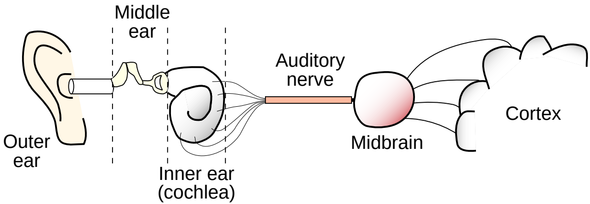
  e.g. scene analysis, speech perception

Anatomical & Functional Organizations



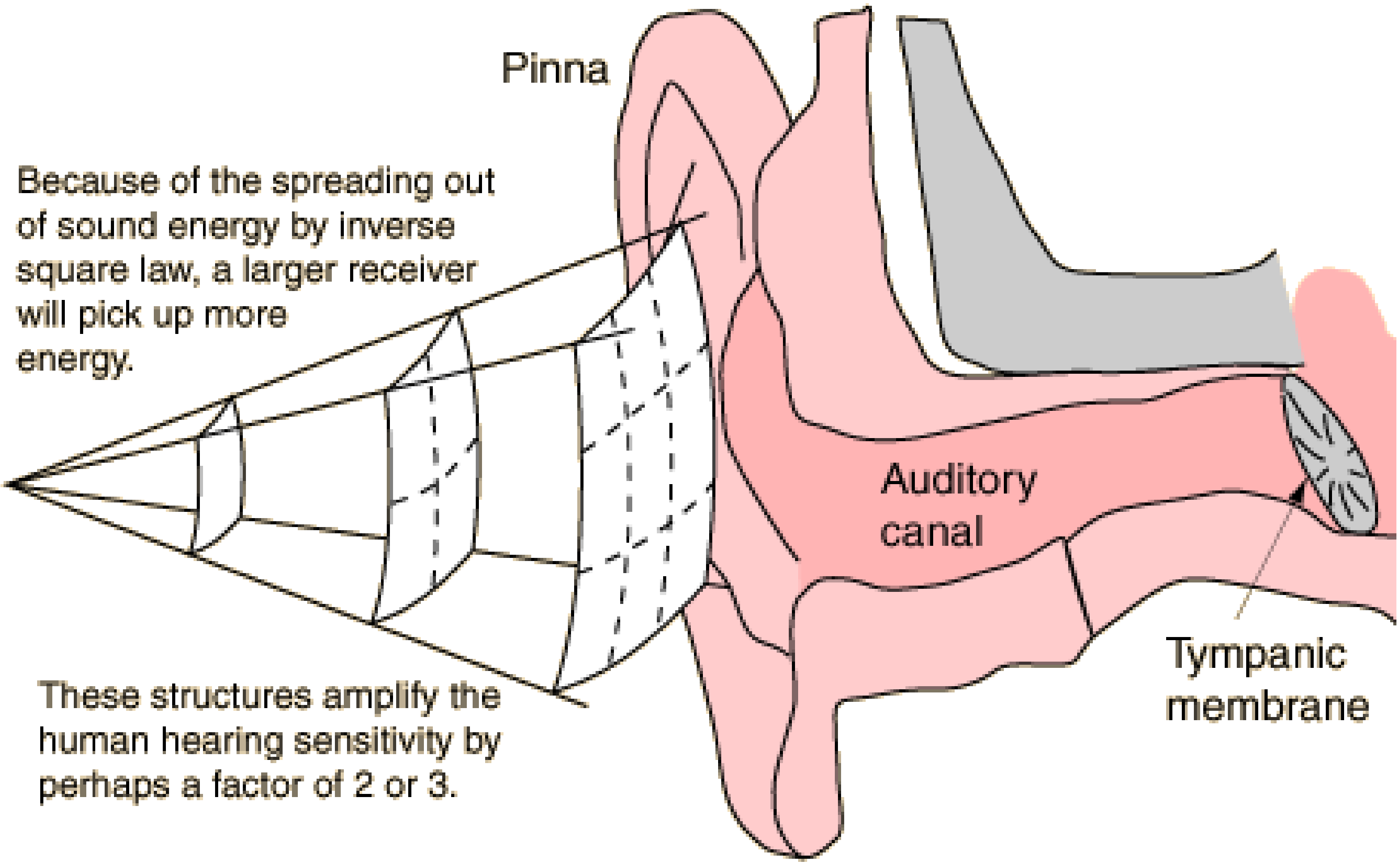
Physiology

- Processing chain from air to brain:

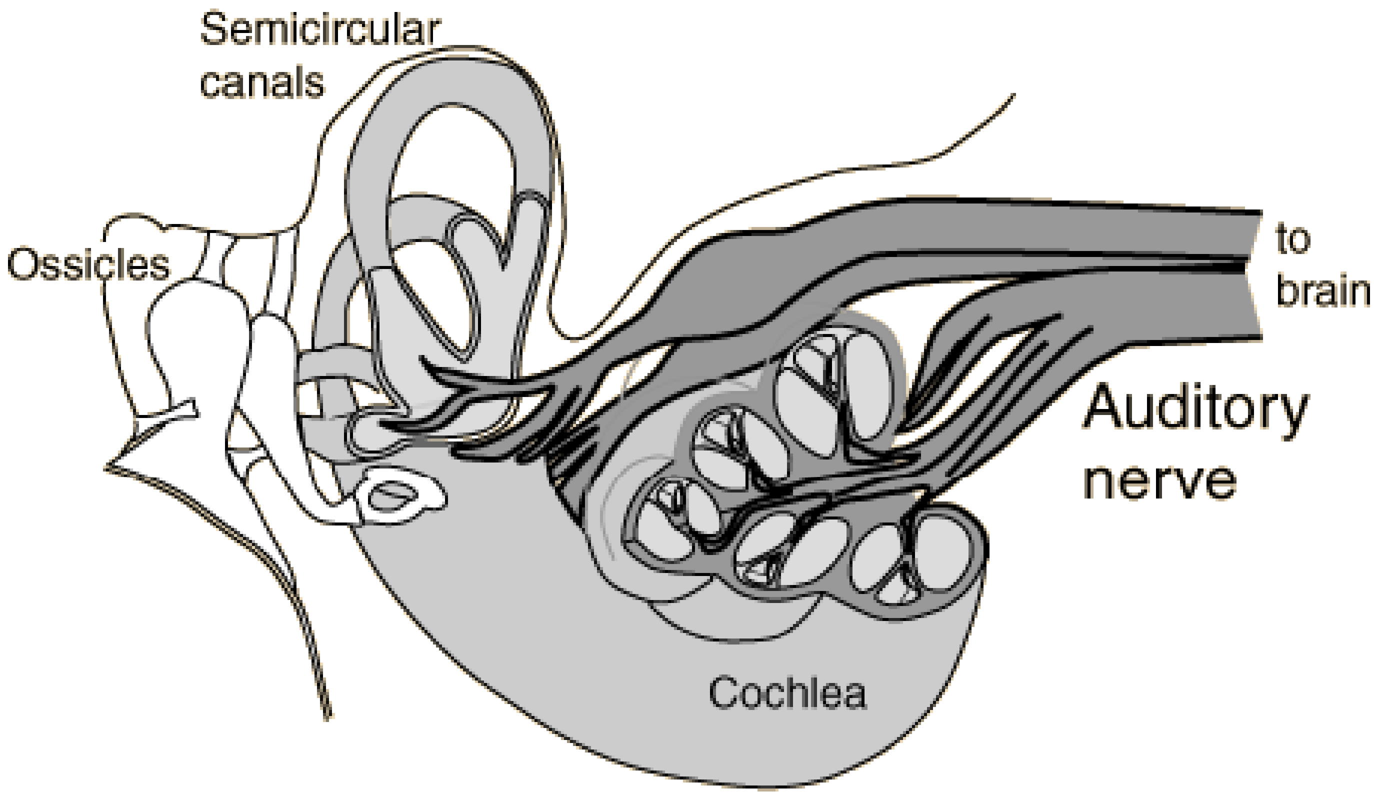


- Study via:
  - anatomy
  - nerve recordings
- Signals flow in both directions

The Outer Ear

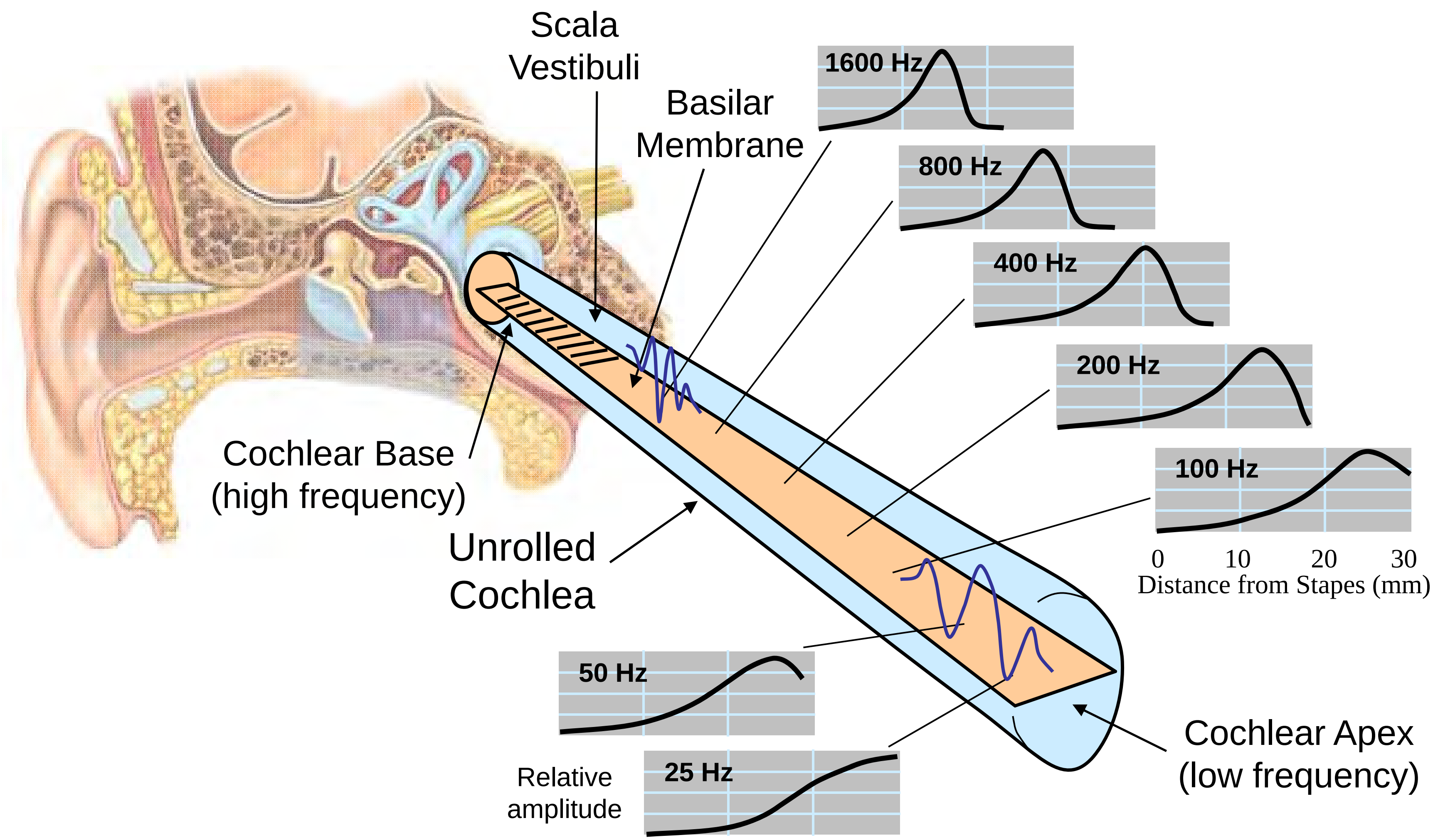


The Auditory Nerve

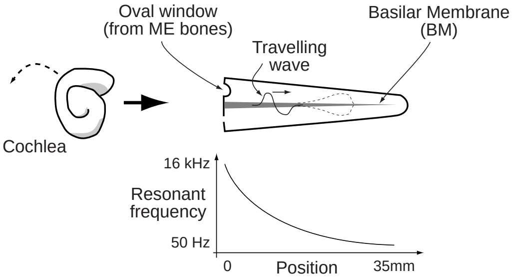


Taking electrical impulses from the cochlea and the semicircular canals, the auditory nerve makes connections with both auditory areas of the brain.

Stretched Cochlea & Basilar Membrane



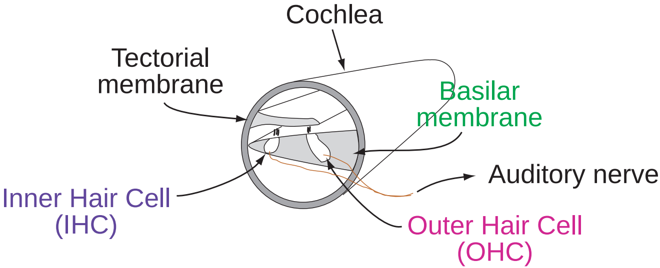
Inner ear: Cochlea



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

Cochlea hair cells

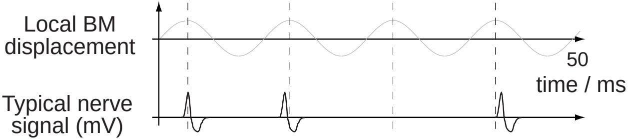
- Ear converts sound to BM motion
  - each point on BM corresponds to a frequency



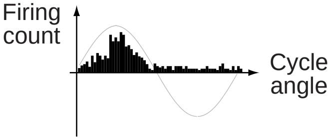
- Hair cells on BM convert motion into nerve impulses (firings)
- Inner Hair Cells detect motion
- Outer Hair Cells? Variable damping?

Inner Hair Cells

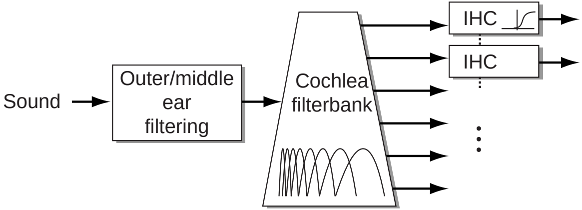
- IHCs convert BM vibration into nerve firings
- Human ear has ∼3500 IHCs
  - each IHC has ∼7 connections to Auditory Nerve
- Each nerve fires (sometimes) near peak displacement



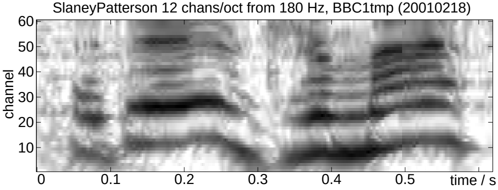
- Histogram to get firing probability



Periphery models



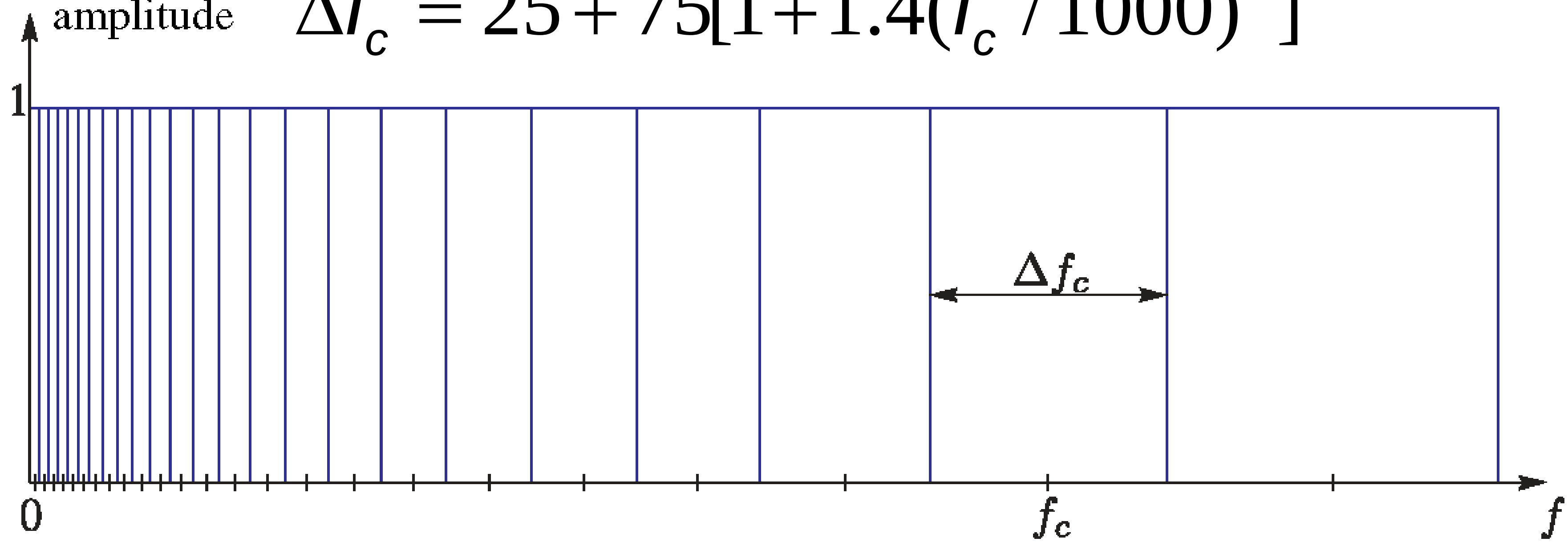
- Modeled aspects
  - outer / middle ear
  - hair cell transduction
  - cochlea filtering
  - efferent feedback?



Results: ‘neurogram’ / ‘cochleagram’

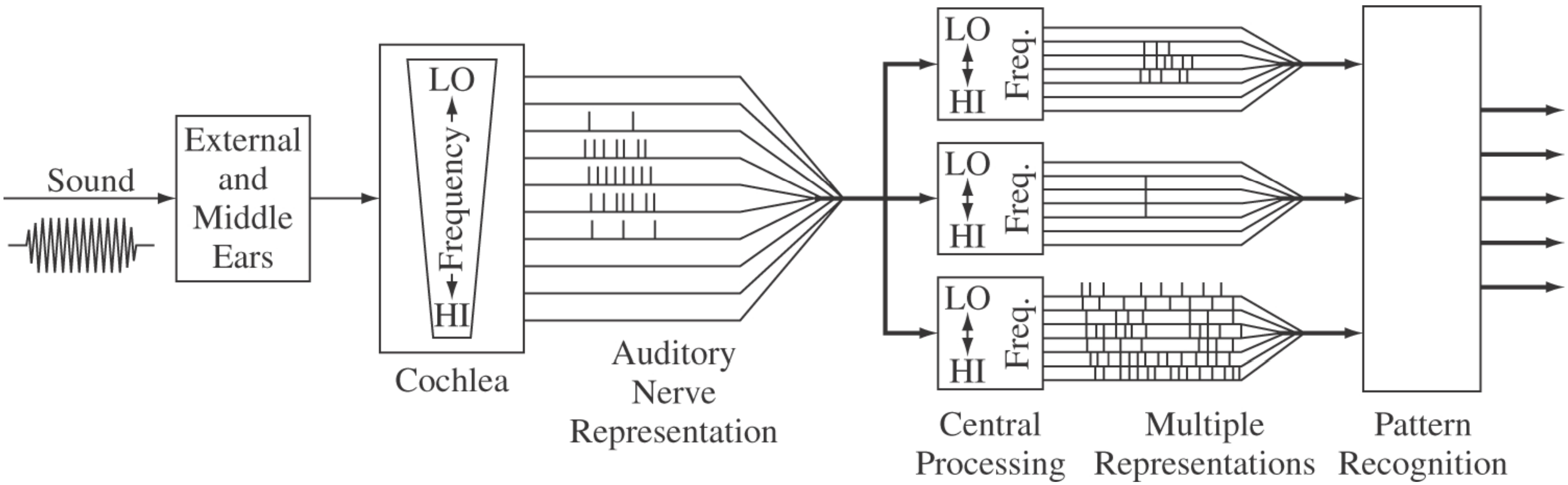
Critical Bands

$$\Delta f_c = 25 + 75[1 + 1.4(f_c / 1000)^2]^{0.69}$$



- Idealized basilar membrane filter bank
  - Center Frequency of Each Bandpass Filter: f_c
  - Bandwidth of Each Bandpass Filter: Δf_c
  - Real BM filters overlap significantly

Overview of Auditory Mechanism

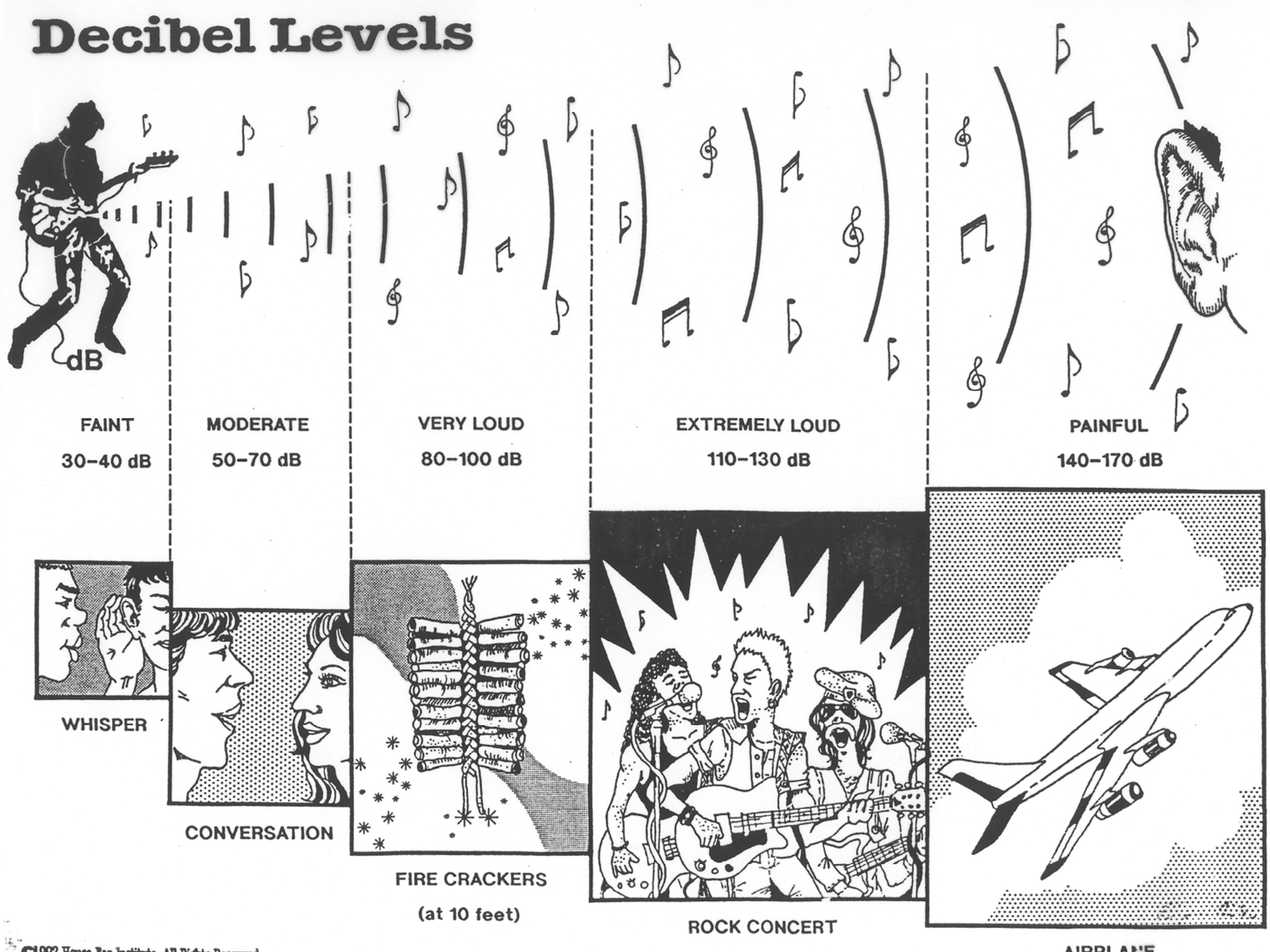


- begin by looking at ear models including processing in cochlea
- give some results on speech perception based on human studies in noise

The Range of Human Hearing

Some Facts About Human Hearing

- the ***range of human hearing*** is incredible
  - ***threshold of hearing*** — thermal limit of Brownian motion of air particles in the inner ear
  - ***threshold of pain*** — intensities of from 10**12 to 10**16 greater than the threshold of hearing
- human hearing perceives both ***sound frequency*** and ***sound direction***
  - can detect weak spectral components in strong broadband noise
- ***masking*** is the phenomenon whereby one loud sound makes another softer sound inaudible
  - masking is most effective for frequencies around the masker frequency
  - masking is used to hide quantizer noise by methods of spectral shaping (similar grossly to Dolby noise reduction methods)


Decibel Levels
dB
FAINT
30–40 dB
MODERATE
50–70 dB
VERY LOUD
80–100 dB
EXTREMELY LOUD
110–130 dB
PAINFUL
140–170 dB
WHISPER
CONVERSATION
FIRE CRACKERS
(at 10 feet)
ROCK CONCERT
AIRPLANE

Sound Pressure Levels (dB)

SPL (dB)—Sound Source

160	Jet Engine — close up
150	Firecracker; Artillery Fire
140	Rock Singer Screaming into Microphone: Jet Takeoff
130	***Threshold of Pain***; .22 Caliber Rifle
120	Planes on Airport Runway; Rock Concert; Thunder
110	Power Tools; Shouting in Ear
100	Subway Trains; Garbage Truck
90	Heavy Truck Traffic; Lawn Mower
80	Home Stereo — 1 foot; Blow Dryer

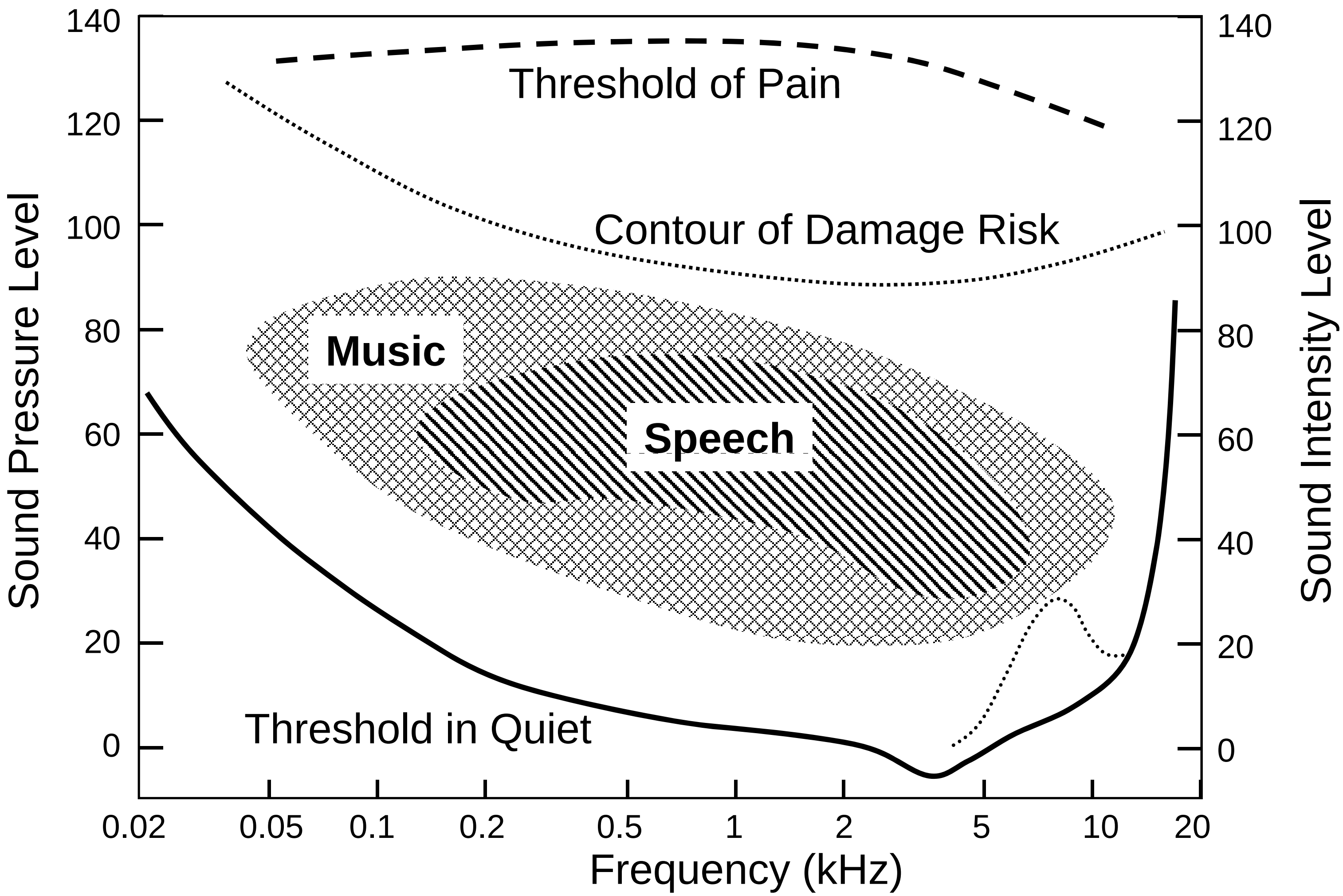
SPL (dB)—Sound Source

70	Busy Street; Noisy Restaurant
60	Conversational Speech — 1 foot
50	Average Office Noise; Light Traffic; Rainfall
40	Quiet Conversation; Refrigerator; Library
30	Quiet Office; Whisper
20	Quiet Living Room; Rustling Leaves
10	Quiet Recording Studio; Breathing
0	***Threshold of Hearing***

Hearing Thresholds

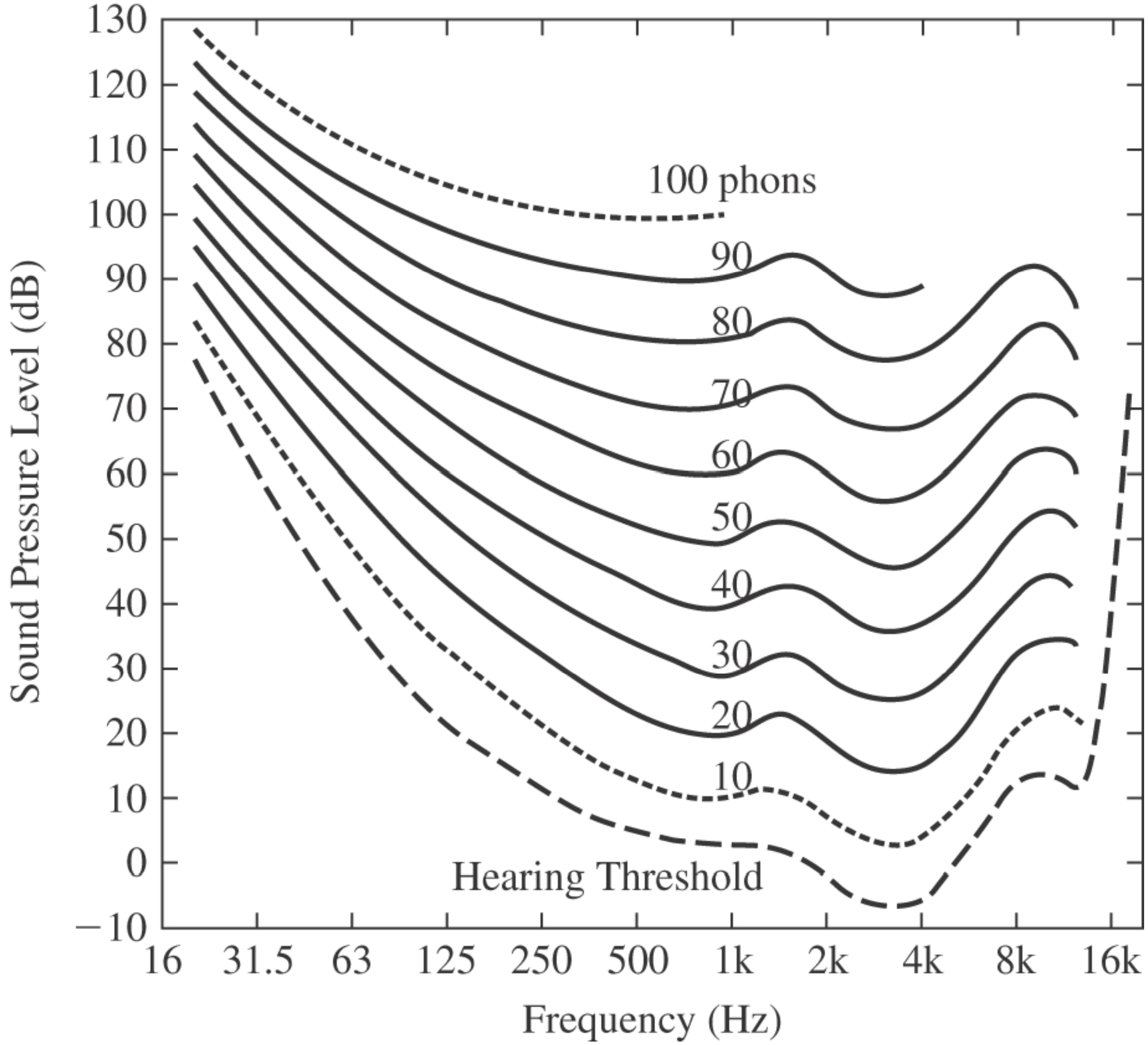
- ***Threshold of Audibility*** is the acoustic intensity level of a pure tone that can barely be heard at a particular frequency
  - *threshold of audibility ≈ 0 dB at 1000 Hz*
  - *threshold of feeling ≈ 120 dB*
  - *threshold of pain ≈ 140 dB*
  - *immediate damage ≈ 160 dB*
- *Thresholds vary with frequency and from person-to-person*
- *Maximum sensitivity is at about 3000 Hz*

Range of Human Hearing



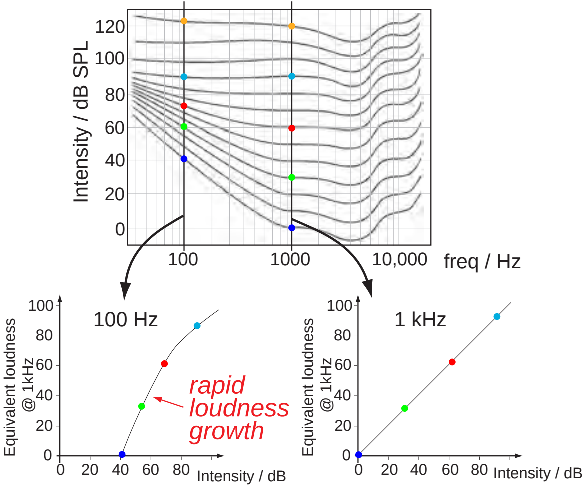
Loudness Level

- ***Loudness Level (LL)*** is equal to the *IL* of a 1000 Hz tone that is judged by the average observer to be equally loud as the tone



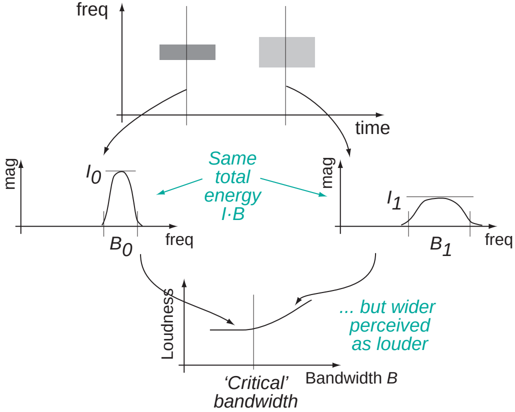
Loudness as a function of frequency

Fletcher-Munsen equal-loudness curves



Loudness as a function of bandwidth

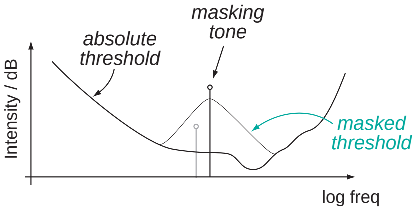
- Same total energy, different distribution
  e.g. 2 channels at −6 dB (not −10 dB)



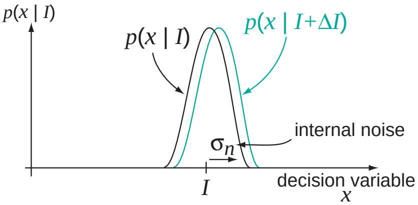
- Critical bands: independent frequency channels
  - ~25 total (4-6 / octave)

Simultaneous masking

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

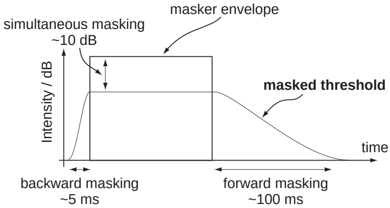


Suggests an 'internal noise' model:

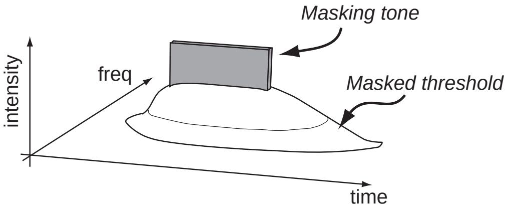


Sequential masking

Backward/forward in time:



→ Time-frequency masking 'skirt':

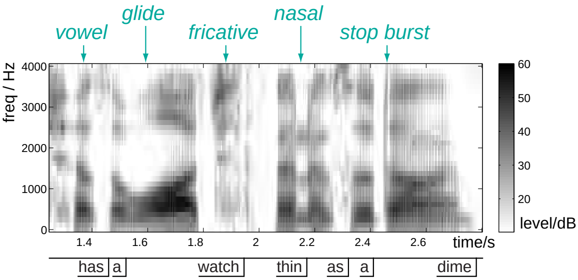


Outline

Speech perception

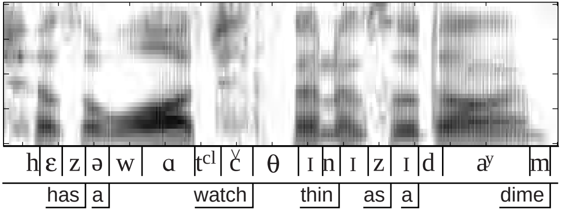
- Highly specialized function
  - subsequent to source organization?

  . . . but also can interact
- Kinds of speech sounds

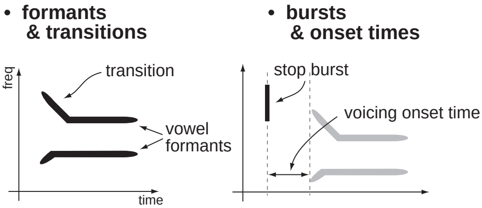


Cues to phoneme perception

Linguists describe speech with phonemes

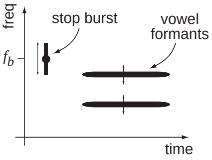


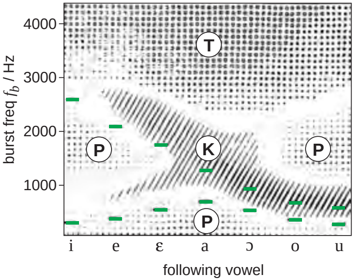
Acoustic-phoneticians describe phonemes by



Categorical perception

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

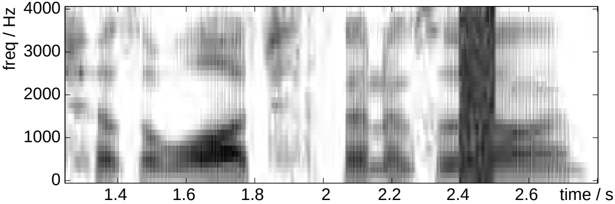




  - tokens within category are hard to distinguish
  - category boundaries are very sharp
- Categories are learned for native tongue
  - "merry" / "Mary" / "marry"

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

What if a noise burst obscures speech?

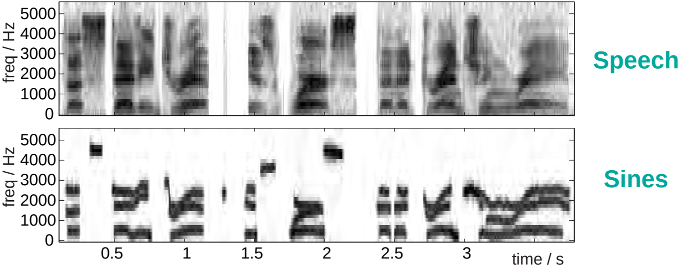


- auditory system 'restores' the missing phoneme
  ... based on semantic content
  ... even in retrospect

Subjects are typically unaware of which sounds are restored

A predisposition for speech: Sinewave replicas

Replace each formant with a single sinusoid (Remez et al., 1981)

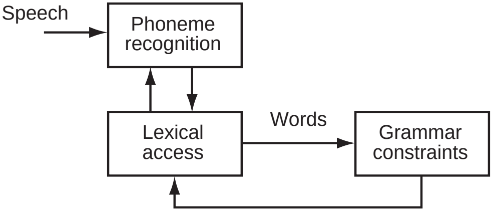


- speech is (somewhat) intelligible
- people hear both whistles and speech ("duplex")
- processed as speech despite un-speech-like

What does it take to be speech?

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 (?)

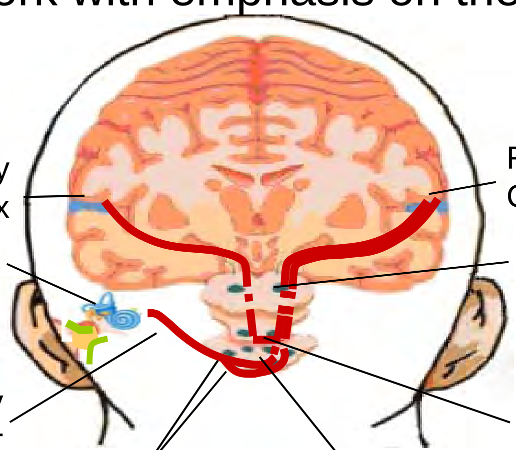
Different Views of Auditory Perception

- Functional: based on studies of psychophysics – relates stimulus (*physics*) to perception (*psychology*): e.g. frequency in Hz. vs. Mel/Bark scale.

Stimulus → **Auditory System** → Sensation, Perception

Black Box

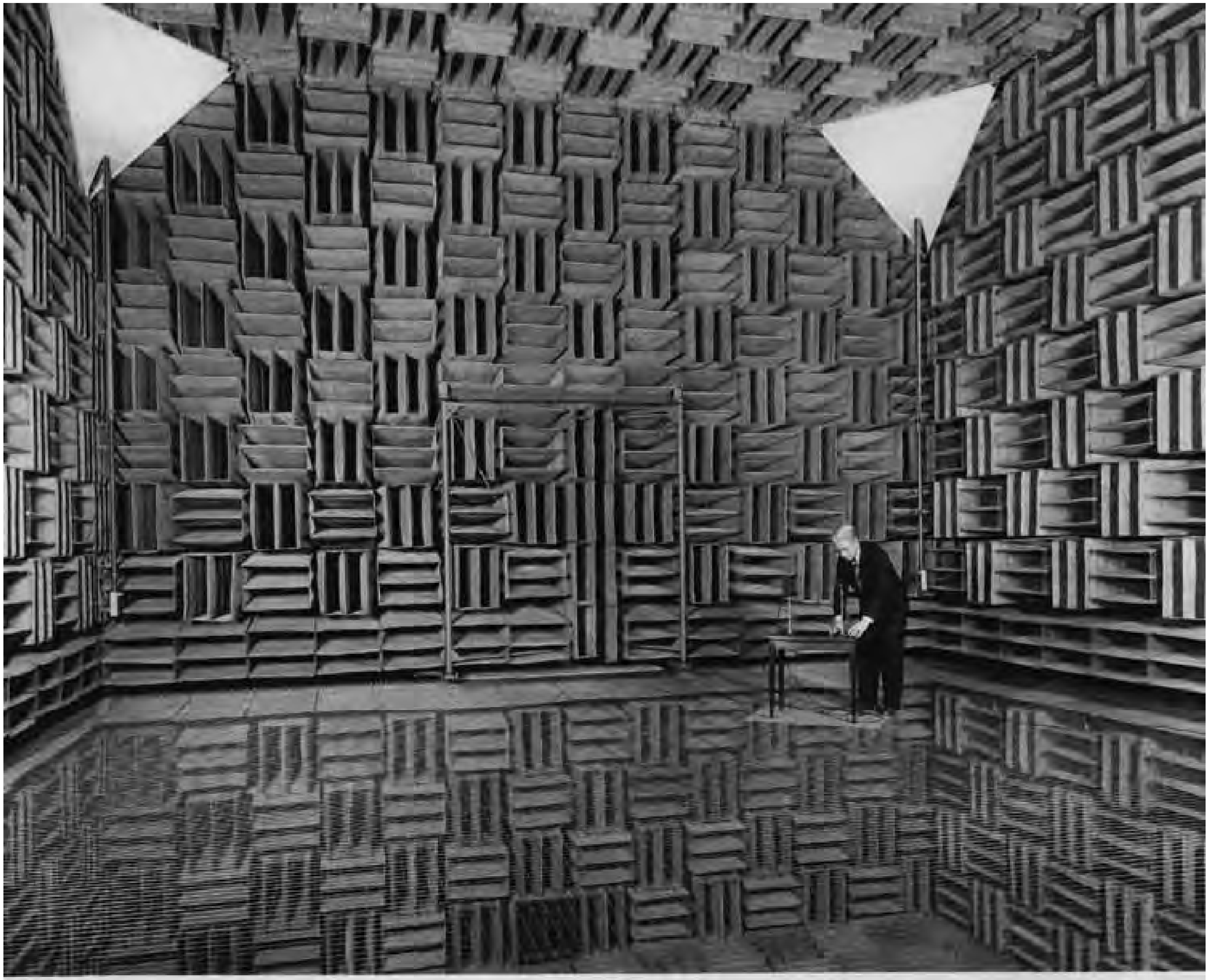
- Structural: based on studies of physiology/anatomy – how various body parts work with emphasis on the process; e.g. neural processing of a sound



Auditory System:

- Periphery: outer, middle, and inner ear
- Intermediate: CN, SON, IC, and MGN
- Central: auditory cortex, higher processing units

Anechoic Chamber (no Echos)



1.1.2-6 . . . . .ANECHOIC CHAMBER TESTS 13745-11 (8/60)

DEEE725 Speech Signal Processing Lab

Gil-Jin Jang

END OF LECTURE 08 HEARING, AUDITORY MODELS, AND SPEECH PERCEPTION