PHYS 536

R. J. Wilkes

Session 12

Speech modeling and synthesis

Human hearing

2/9/2023

Course syllabus and schedule – updated

See : http://courses.washington.edu/phys536/syllabus.htm

Session	date	Day	Readings:	K=Kinsler, H=Heller	Topic
8	26-Jan	Thu	K: Ch. 7	H: Ch. 7	Absorption losses; Pulsating spheres and simple sources; pistons and dipoles; Near field, far field; Radiation impedance; Waves in pipes ***UPDATED BELOW HERE:***
9	31-Jan	Tue	K. Ch. 8-10	H: Ch. 13	Rectangular cavities; Helmholtz resonators; Resonant bubbles; Acoustic impedance; physical acoustic filters; Doppler effect; Interference effects
10	2-Feb	Thu	K. Ch 9	H: Chs. 23-25	Musical acoustics: pitch, musical tones and frequency; timbre; beats
11	7-Feb	Tue		H: Chs. 16, 18	Musical instruments: winds and string instruments
12	9-Feb	Thu		H: Chs. 17, 19	Musical instruments: piano, human voice REPORT 1 PAPER DUE by 7 PM; REPORT 2 PROPOSED TOPIC DUE
13	14-Feb	Tue	K. Ch. 11	H: Ch. 21	Human hearing: the inner ear; pitch perception; acoustics of speech
14	16-Feb	Thu	K. Ch. 12	H: Chs. 21-22	Decibels and sound level measurements Environmental acoustics and noise criteria; industrial and community noise regulations; noise mitigation;
15	21-Feb	Tue	K. Chs. 13-14	H: Chs. 27-28; Ch. 6	Room acoustics; Transducers for use in air and water: Microphones and loudspeakers; hydrophones and pingers; Underwater acoustics: sound absorption underwater, the sonar equation
16	23-Feb	Thu	K. Ch 15		Underwater acoustics applications: acoustical positioning, seafloor imaging, sub-bottom profiling; Course wrap-up: review
17	28-Feb	Tue			Student report 2 presentations
18	2-Mar	Thu			Student report 2 presentations
19	7-Mar	Tue			Student report 2 presentations
20	9-Mar	Thu			Student report 2 presentations. TAKE-HOME FINAL EXAM ISSUED
--	17-Mar	Fri			FINAL EXAM ANSWERS DUE by 5 PM

Tonight

Class is over after you turn in your take-home exam. No in-person final exam during finals week.

Announcements

- Term paper 1 is due TONIGHT
  - Submit via Canvas Assignments page
  - Submissions before midnight will be accepted as on-time -- after 11:59 pm you will be docked 5 pts for late papers
    - Submission portal on Canvas closes at 11:59 pm on Sunday 2/12
    - It will take at least a week for us to grade the papers – please be patient
- **Don't forget:** term project 2 presentation proposal is also due today
  - Only about 1/3 of class has sent me proposals, as of 5 pm
  - Please email yours to me within a few days

    Goal is to make sure your topic is appropriate, you have adequate sources, and your scope fits into a 15 min talk

Last time

Vocal folds' vibration mechanisms

Adapted from www.phys.unsw.edu.au/jw/voice.html -- see there for cited work

Mechanism 0 (M0) is also called 'creak' or 'vocal fry'.

- Tension of the folds is so low that the vibration is non-periodic: M0 sounds low but has no clear pitch

Mechanism 1 (M1) is usually associated with what women singers call the 'chest' register and men call their normal voice.

- Virtually all of the mass and length of the vocal folds vibrates (Behnke, 1880) and frequency is regulated by muscular tension (Hirano et al., 1970) but is also affected by air pressure. The glottis opens for a relatively short fraction of a vibration period (Henrich et al., 2005).

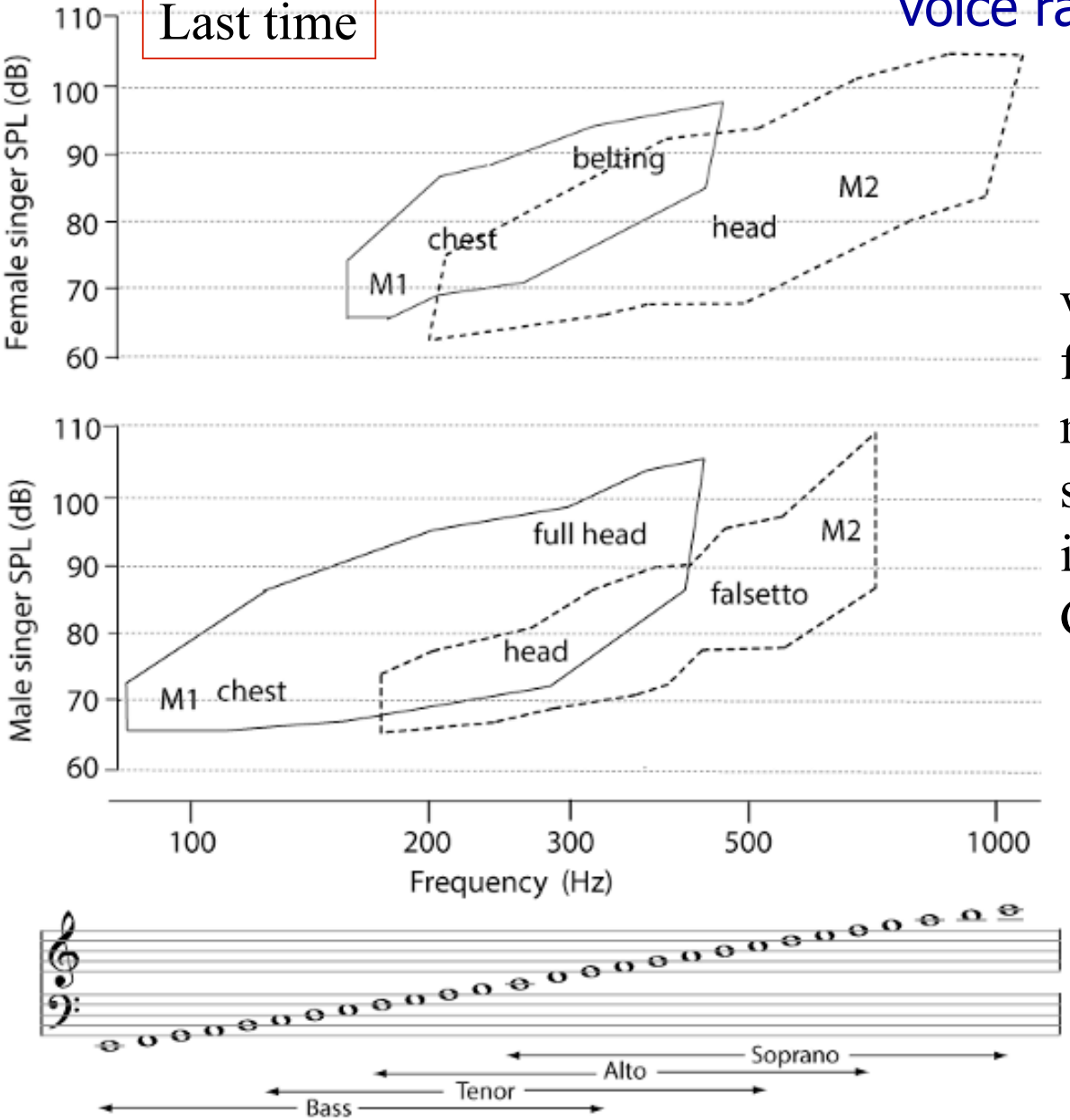
Mechanism 2 (M2) is associated with the 'head' register of women and the 'falsetto' register in men.

- A fraction of the vocal fold mass vibrates. The moving section involves about two thirds of their length, but less of the breadth. The glottis is open for a longer fraction of the vibration period (Henrich et al., 2005).

Mechanism 3 (M3) is the 'whistle' or 'flageolet' register (not to be confused with whistling) (Miller and Shutte, 1993; Garnier et al, 2010; 2012.)

Last time

voice range profiles



voice range profiles for a woman's and man's voice, with some registers indicated (from Garnier et al, 2020)

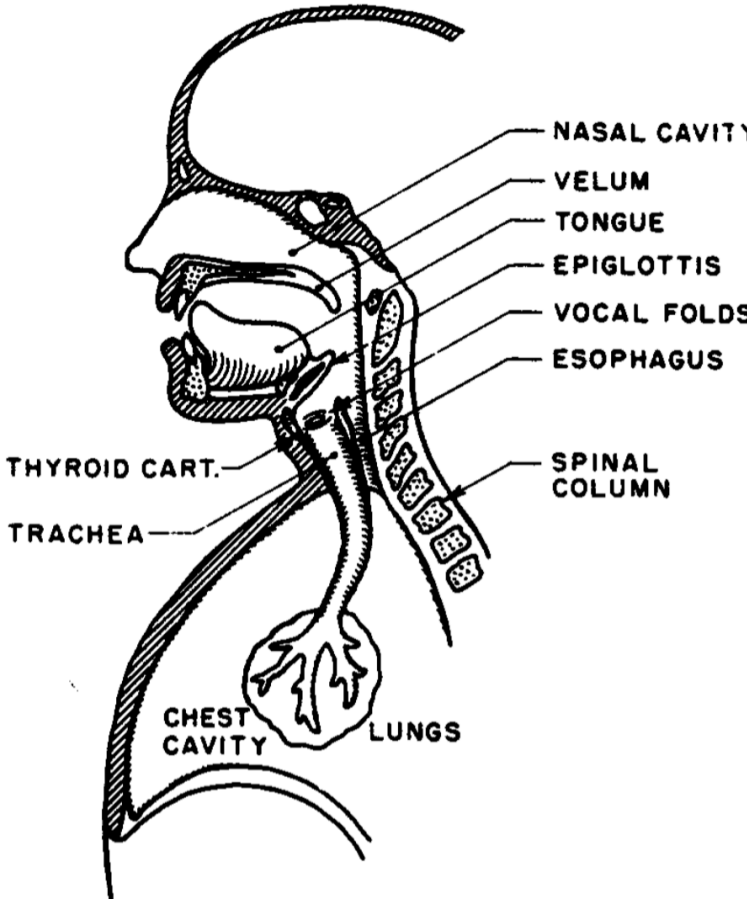
Yodeling

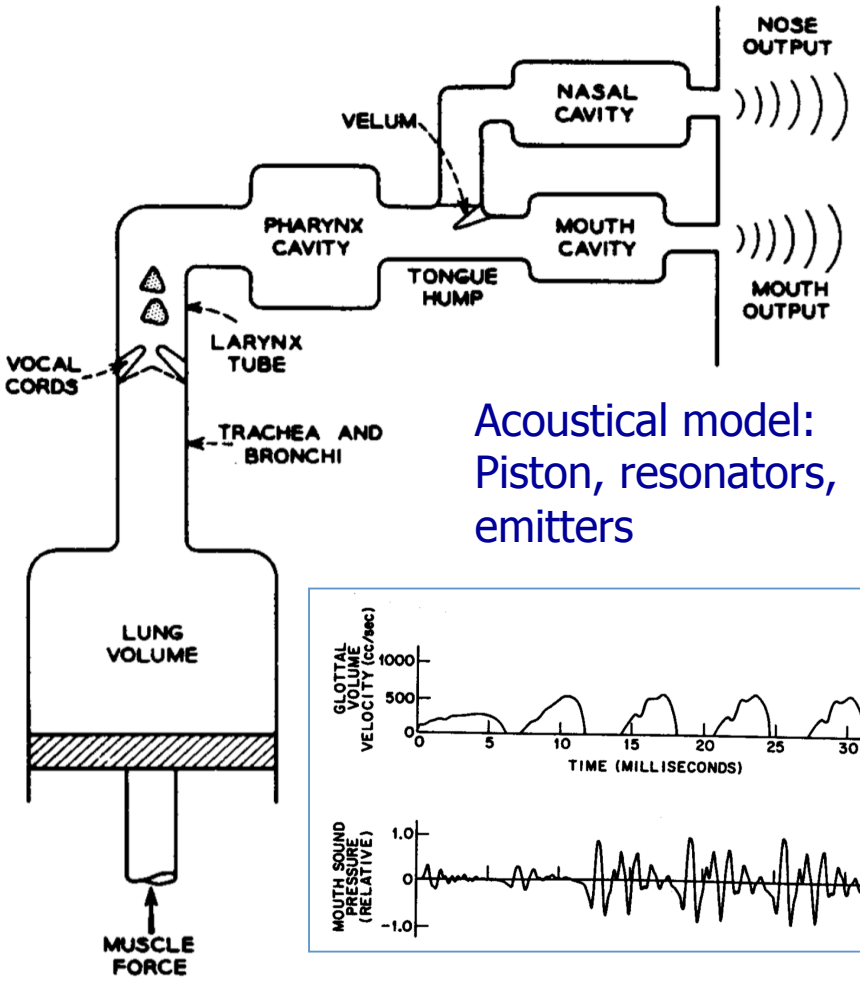
- "Yodeling" → from German *jodeln*, "to utter the syllable *jo*."
- Singing a long note with repeated sharp shifts in pitch
  - Cycle between the low-pitch chest register (or "chest voice") and the high-pitch head register "(falsetto")
  - Opera singers are experts at smoothing over this break, while yodellers accentuate it
- Origins:
  - Marco Polo discovered that Tibetan monks routinely used yodelling to communicate over long distances.
  - Soon after his return German and Swiss mountain dwellers began yodelling to each other across alpine valleys.
  - Immigrants took the style to the US → slower and simpler yodels used by cowboys.

https://archive.org/details/78_swiss-yodel_britt-brothers_gbia0004159b/ (US, 1933)

Acoustics of speech in one slide

Parts diagram



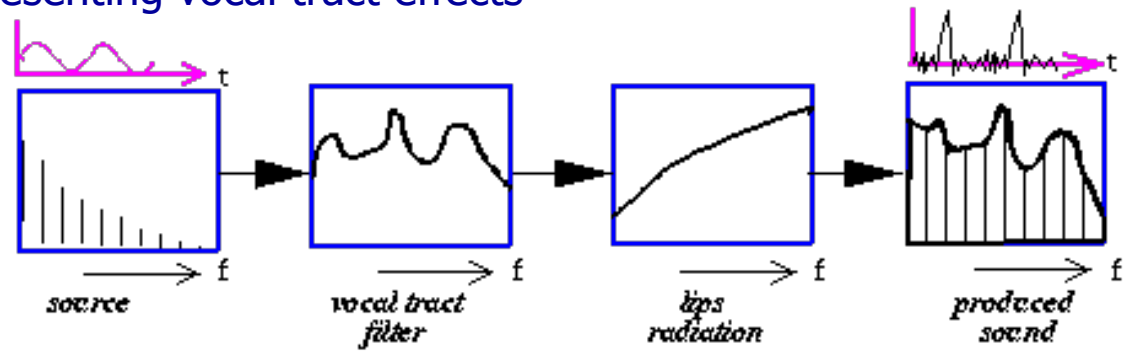


Acoustical model: Piston, resonators, emitters

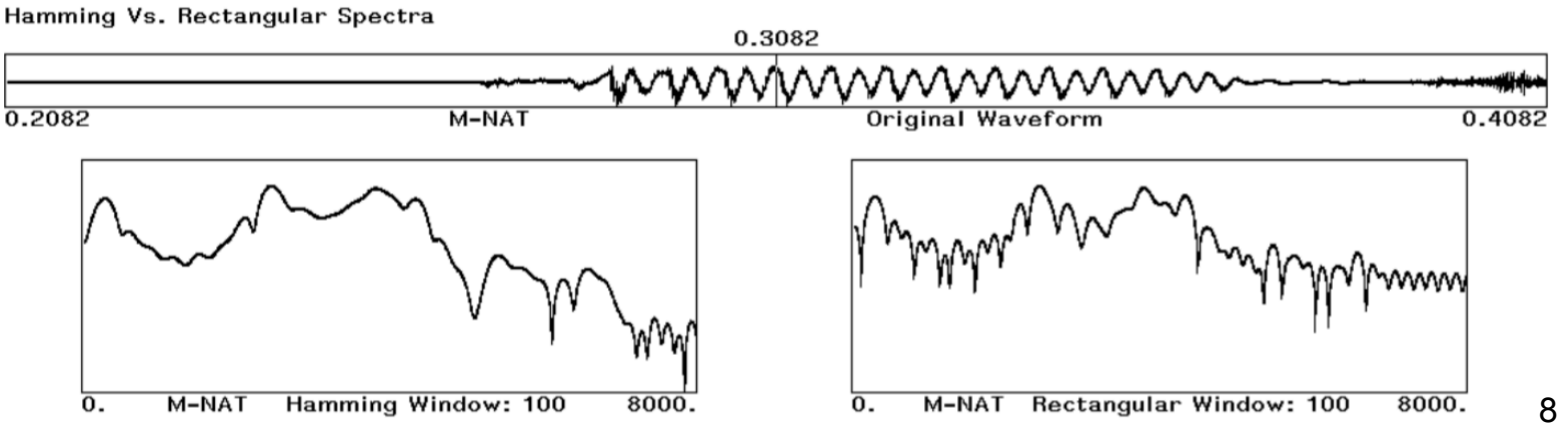
(figures from *Fundamentals of Speech Recognition*, Rabiner & Hwang-Juang, 1993)

Acoustics of speech

- Filter model (G. Fant, 1960)
  - Output sound = source (glottis) spectrum modified by filter function representing vocal tract effects

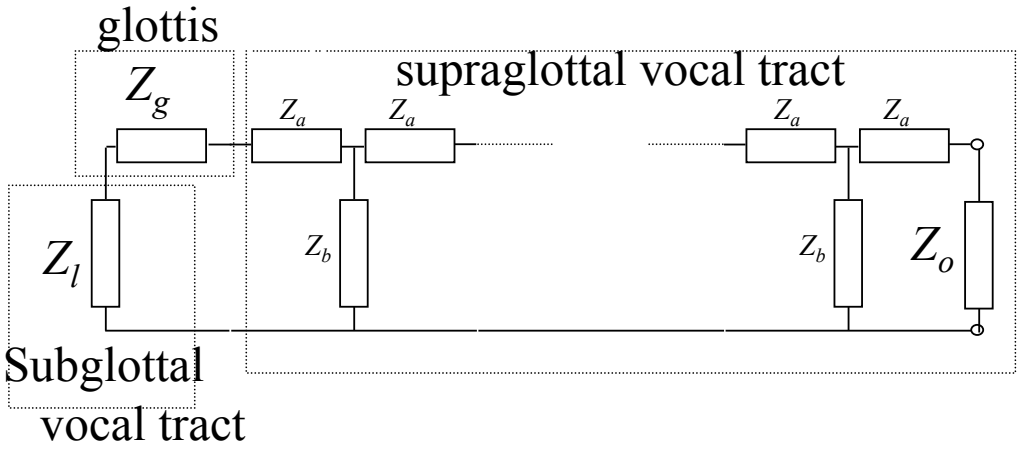


  - Note: must be careful with spectra: what windowing was done (if any)?

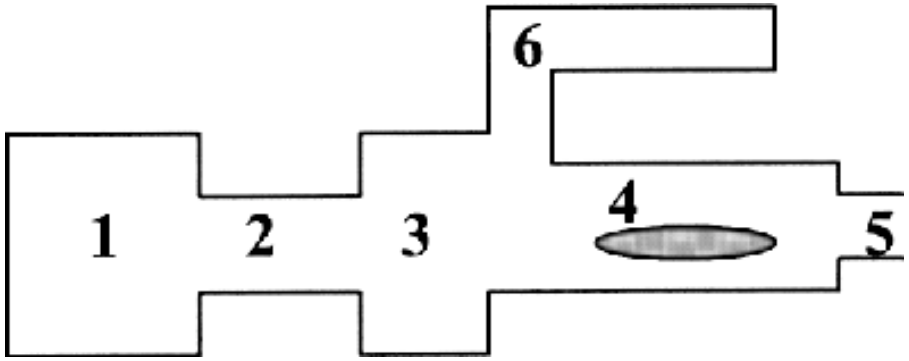


Acoustics of speech

- Lumped parameter model of vocal tract as transmission line



- Vocal tract model example*:
  1. back cavity
  2. pharyngeal constriction
  3. middle cavity
  4. tongue cavity
  5. front cavity
  6. Nasal cavity



* Zhang, JASA 115:1274 (2004)

Acoustics of speech

Labeling convention:
S=silence
V=voiced (vocal cords active)
U=unvoiced (vocal cords inactive)

All vowels are voiced (unless whispered).
Consonants can be either

VOICELESS		VOICED	
/p/	park	/b/	bark
/t/	town	/d/	down
/k/	coat	/g/	goat
/f/	fan	/v/	van
/s/	sip	/z/	zip
"sh"	sure	"zh"	treasure
"ch"	chain	"j"	Jane
"th"	thigh	"th"	thy

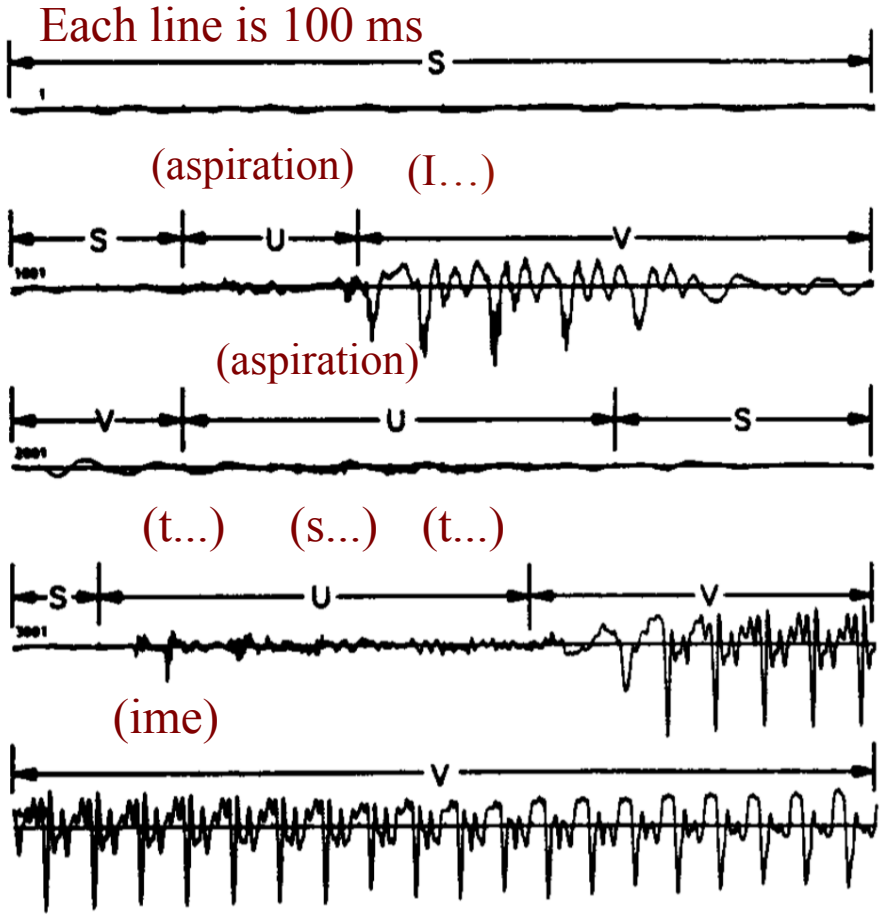


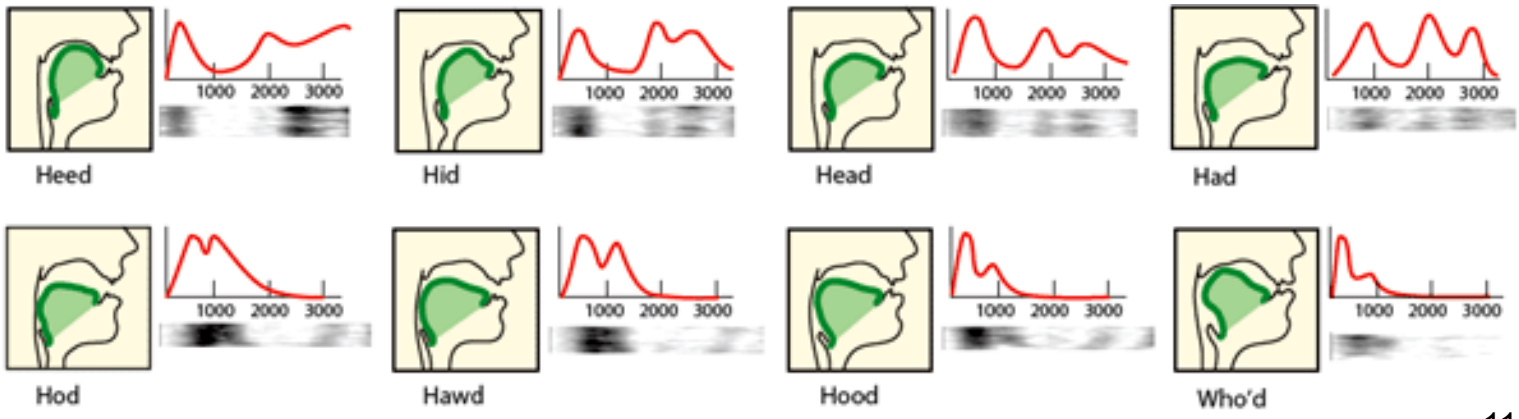
Figure 2.7 Waveform plot of the beginning of the utterance "It's time."

Examples of speech signals (Rabiner & Hwang-Juang, 1993)

Acoustics of speech

- Formants in acoustics = local maximum in the spectrum
  - For harmonic sounds it is the harmonic partial that is augmented by a resonance
  - formants characterize the produced sound
- In phonetics/speech therapy formant = spectral shape due to acoustic resonance of the human vocal tract
  - formants characterize the production mechanisms of a sound

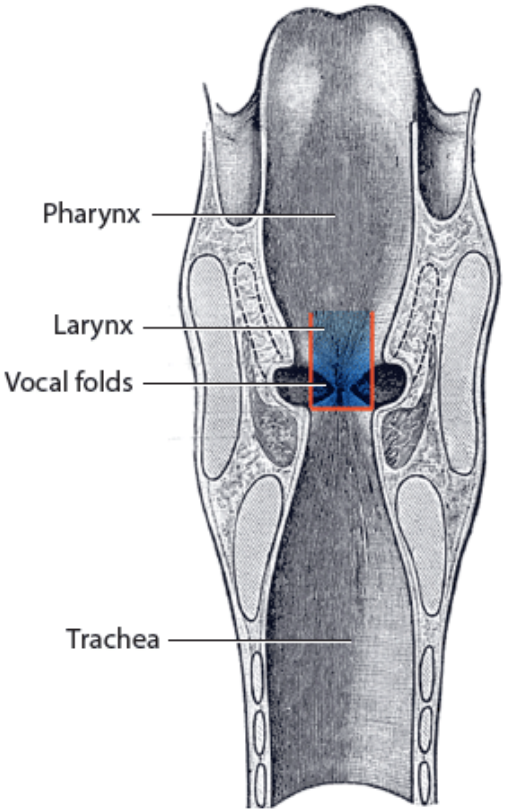
vocal tract shape, predicted formants (red), and measured formants for related vowels. Positions of the lips, opening of the jaw, and the tongue are shown (Heller p. 362)

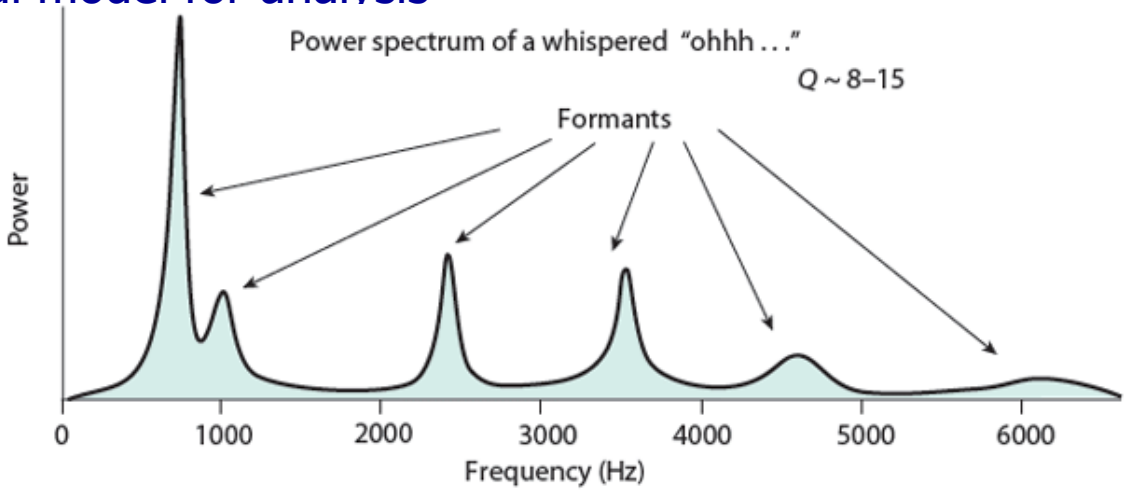


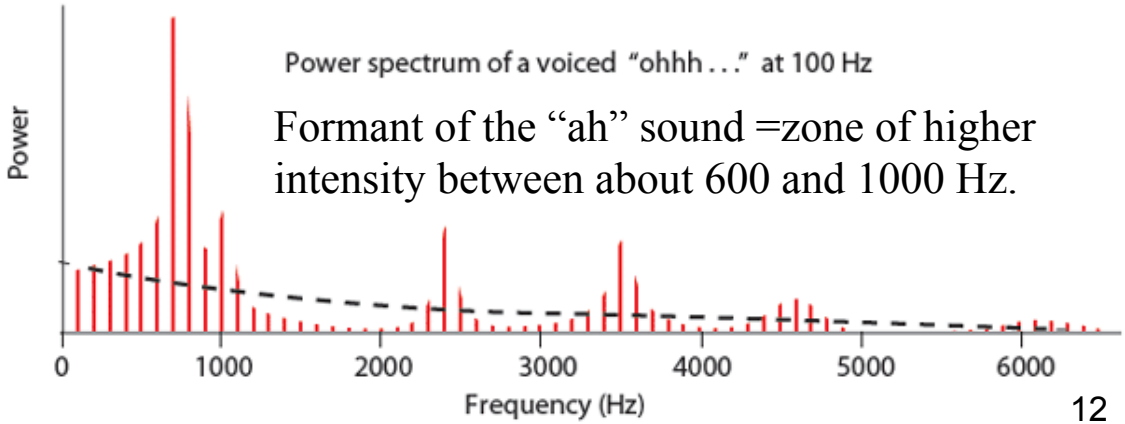
Acoustics of speech

- Filter model does not allow a role for glottal resonances → still used as a physical model for analysis

L=3 cm cylinder has resonance ~2800 Hz





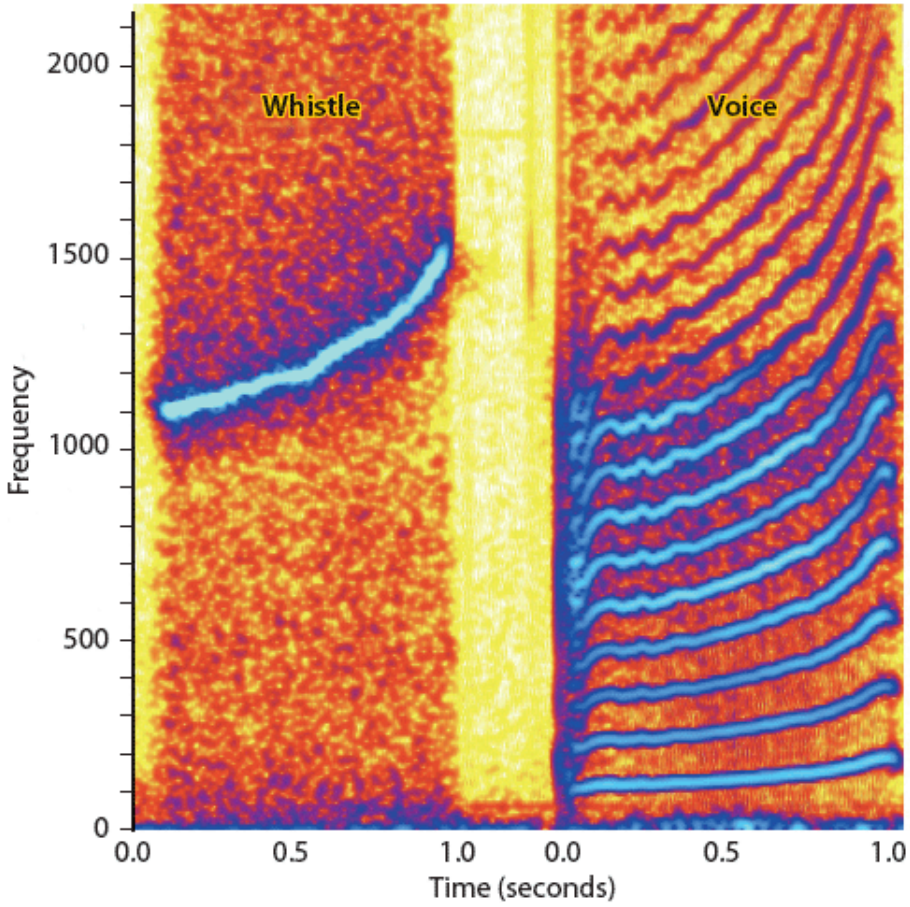


Formant of the "ah" sound =zone of higher intensity between about 600 and 1000 Hz.

Acoustics of speech

Left: Sonogram of a rising whistle (pure tones, starting at about 1100 Hz and ending at 1500 Hz)
Right: rising voice singing "ah" (starting at about 100 Hz and ending at about 200 Hz)

fx-swanee-whistle-up-42036.mp3



Acoustics of speech

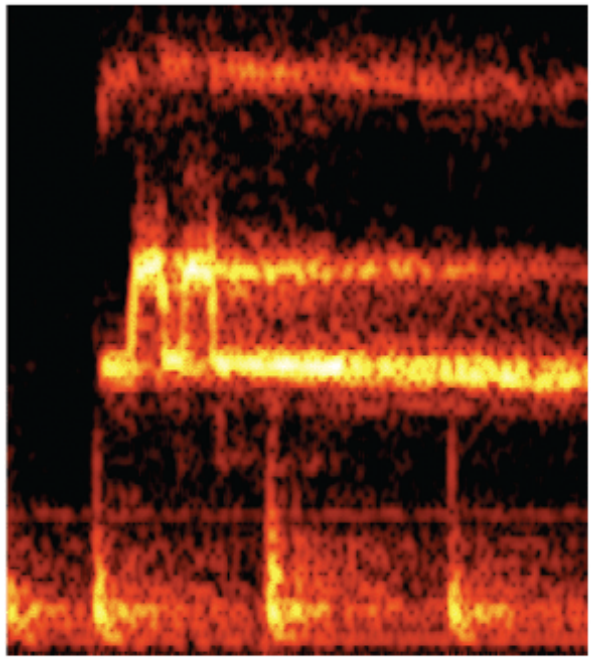
- Another example of misleading windowing effects

Sonograms of the famous whistling in the theme song for the movie The Good, the Bad, and the Ugly with different sample intervals.

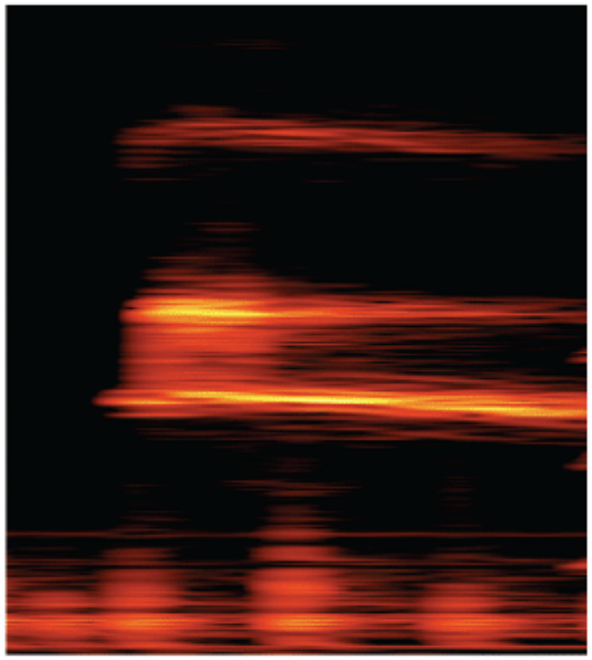
The whistle rapidly switches between two frequencies differing by the interval of a fourth, two notes low and two high.

If the sample window is too long, rapid transitions between notes are wiped out

The "wah-wah sound" is a male voice that has been altered to sound like a harmonica



Window = 256



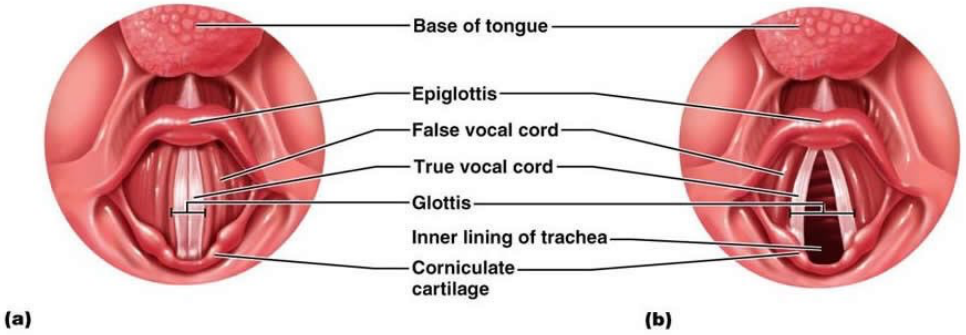
Window = 2048

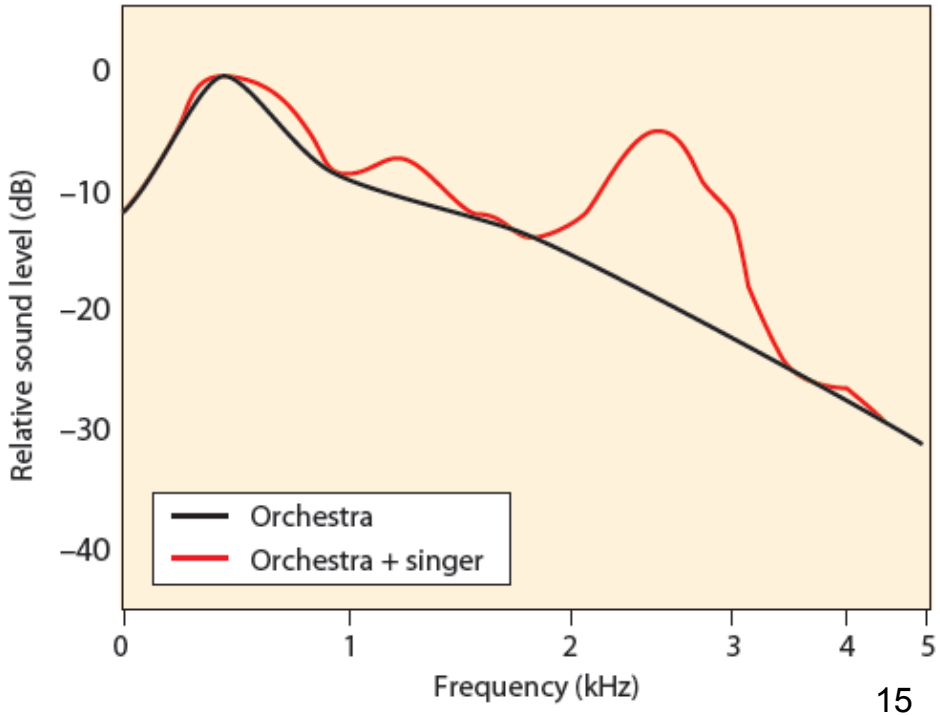
good_bad_ugly.mp3
good_bad_ugly2kfrom11ksampling

Acoustics of speech

The good, the bad, and the beautiful: Singer's formant

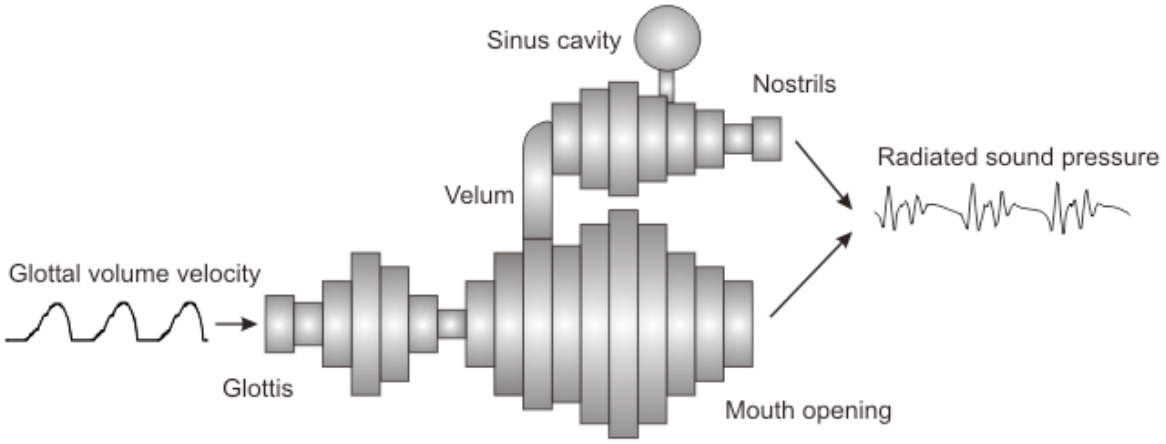
- Trained singers shape larynx resonances using the false vocal fold.
- Average spectrum of a skilled tenor when singing with loud orchestral accompaniment had a pronounced peak around 2800 Hz, with a Q of 3 or 4.
- Singer's formant produces a distinct ringing tone to the voice.
- The orchestra spectrum peaks at much lower frequencies ~500 Hz → audience can easily separate the voice





Speech synthesis

- Articulatory speech synthesis = *direct* simulation of speech production
  - Determine the vocal tract filter by modeling vocal tract geometry
- Synthesizer includes
  - Geometric description of vocal tract → a set of articulatory parameters
    - *piecewise constant* area function = vocal tract composed of cylindrical tube sections
  - Mechanism to adjust the parameters during an utterance
  - Generation of a sound source
- Vocal tract is excited by a glottal volume velocity function (acoustic source) and radiates an acoustic pressure wave at the nostrils and the mouth opening.

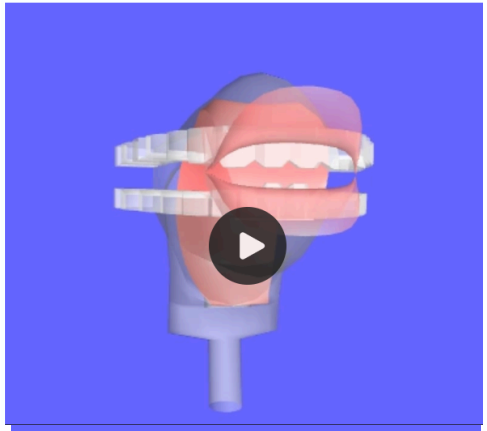


Synthesis examples

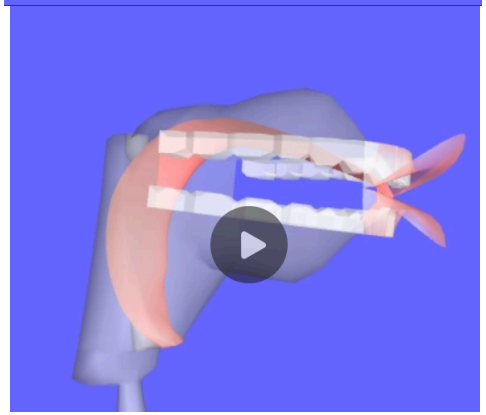
- From VocalTractLab, https://www.vocaltractlab.de/

https://www.vocaltractlab.de/index.php?page=vocaltractlab-examples

"Conny glaubt eigentlich nicht mehr an den Osterhasen."
(Conny actually no longer believes in the Easter Bunny.)
vtl2.3-example03.webm
vtl2.3-example03-synth.wav
vtl2.3-example03-orig.wav

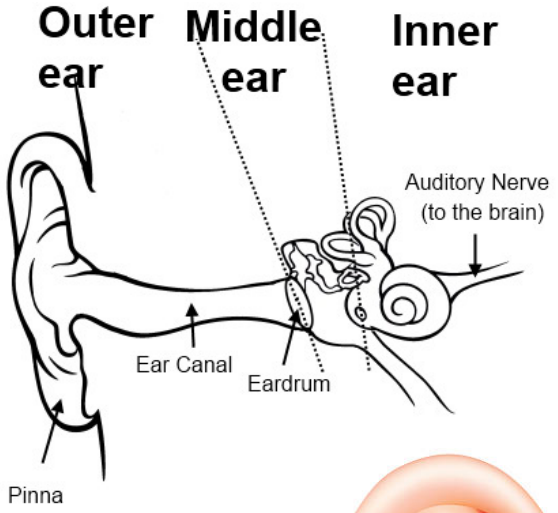


"I think I have a German accent."
vtl2.3-example06.webm
vtl2.3-example06-synth.wav
vtl2.3-example06-orig.wav

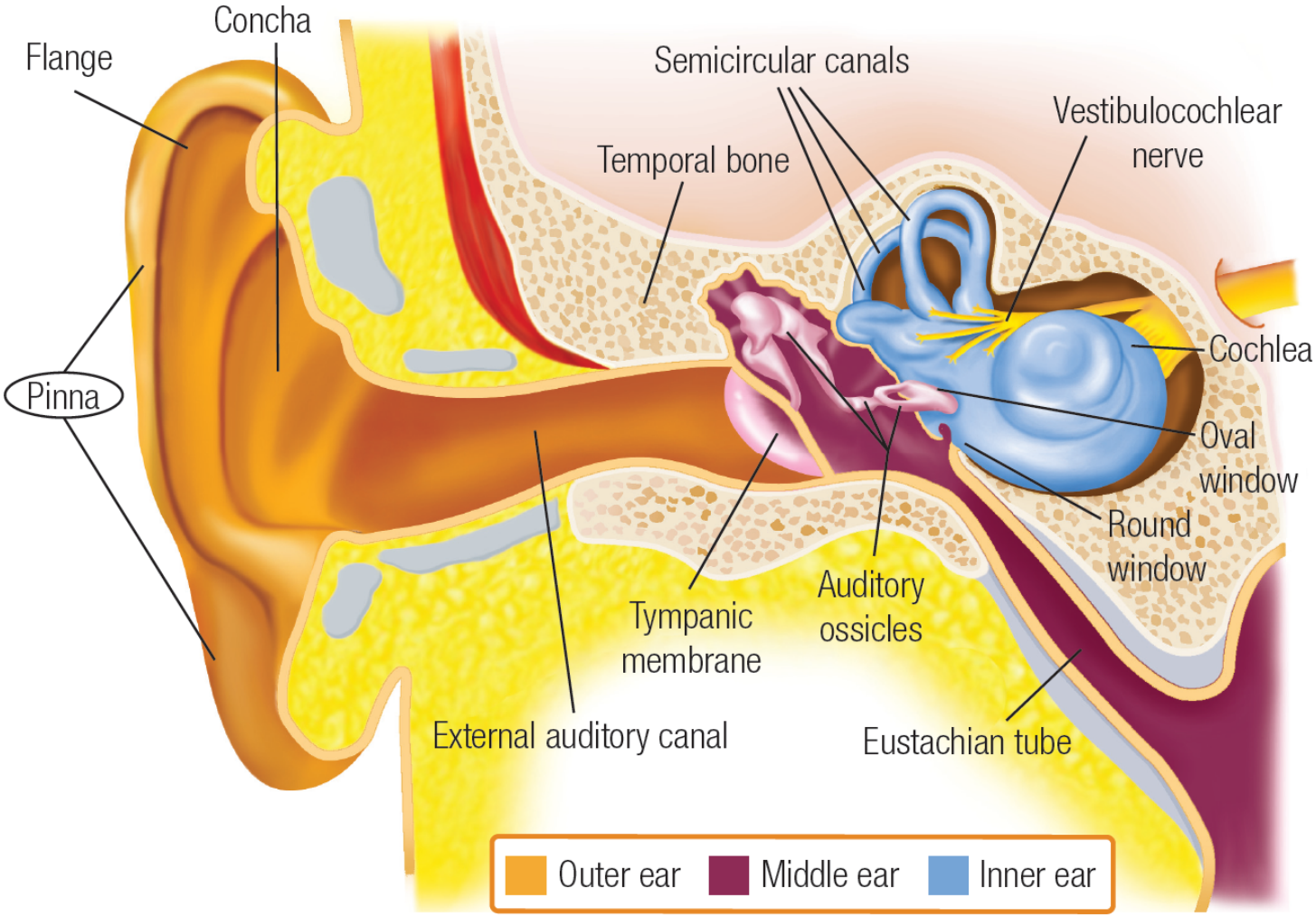


Hearing

- Our personal input transducer = human ear
- Really complex! Parts of the ear:
  (diagrams follow…)
  - Outer Ear
    - Pinna (auricula)
    - Auditory canal (Ear canal)
  - Middle Ear
    - Tympanic membrane (Ear drum)
    - Ossicles: Maleus (Hammer), Incus (Anvil), Stapes (stirrup)
    - Oval window and Round window
    - Two tiny muscles (stapedius and tensor tymphani) connected with the ossicles
    - Eustachian tube (3.5 cm long) connects middle ear with pharynx, balancing air pressures
  - Inner Ear - Cochlea with Organ of Corti

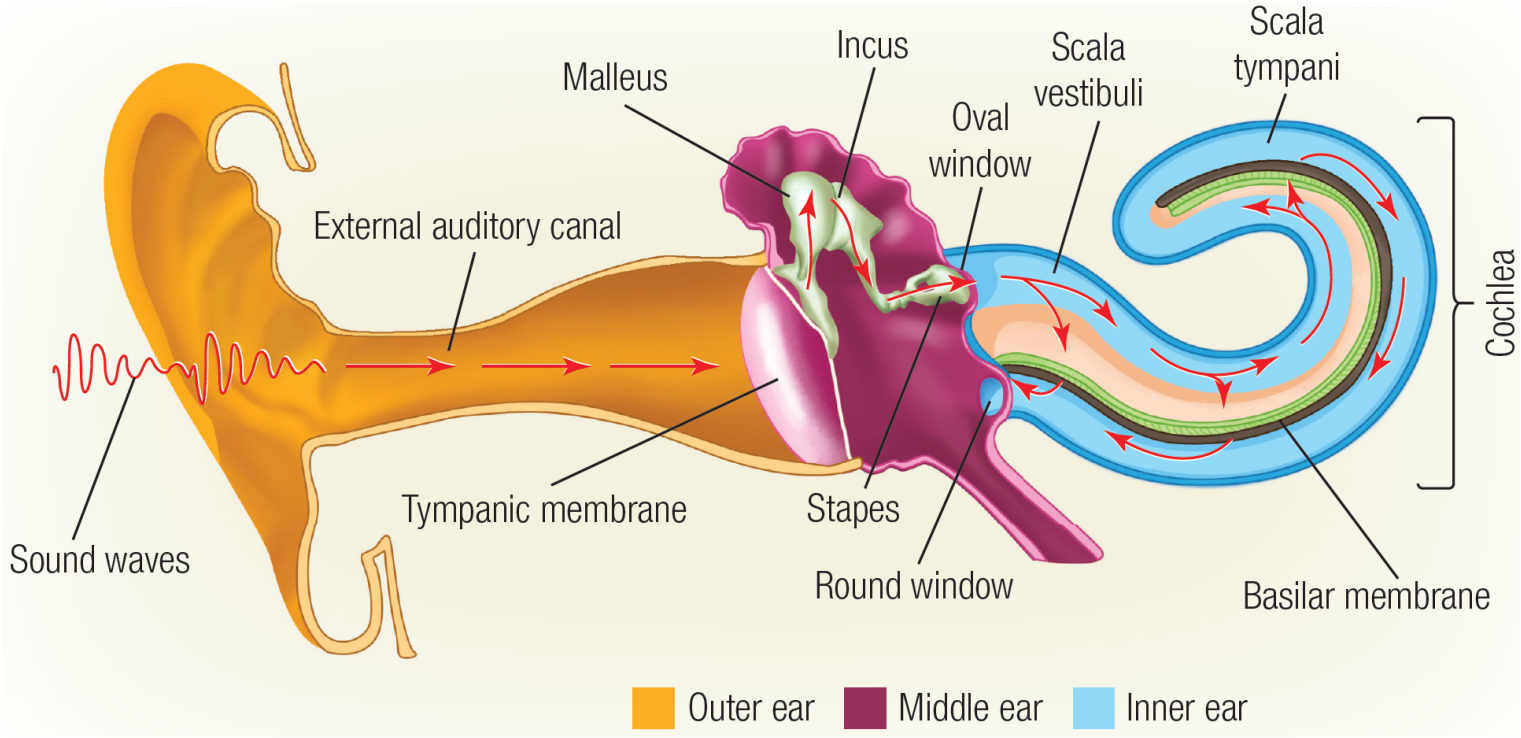


Anatomy of the Ear

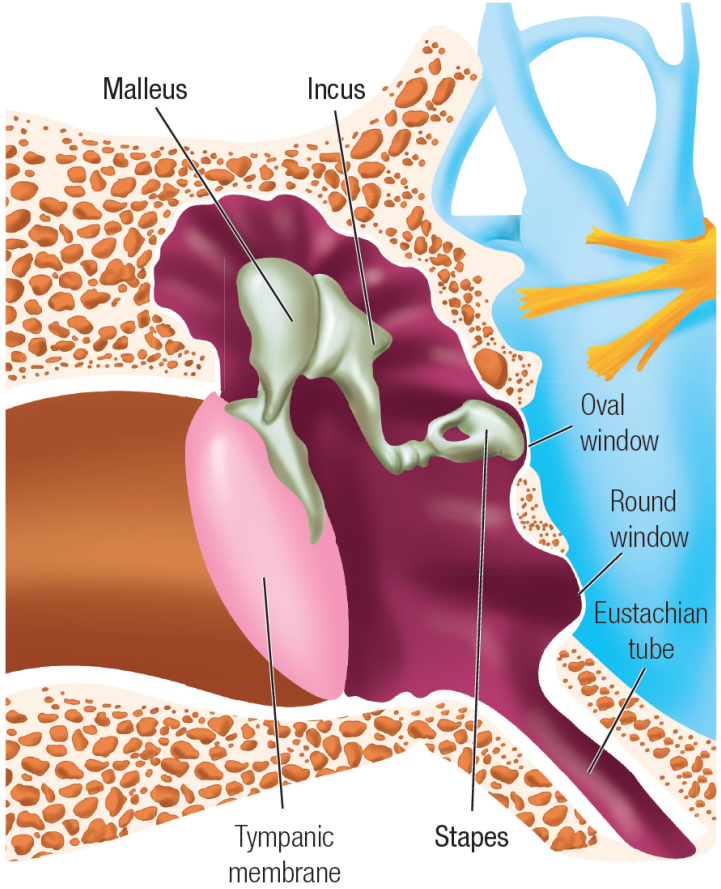


Flow of Acoustic Energy

(The “Impedance Problem”)



The "Impedance Problem"



99.9% of sound energy in the air is reflected at the cochlea's air/water boundary
(10 log(0.1/100)) = **-30 dB loss**) (1/1000x)

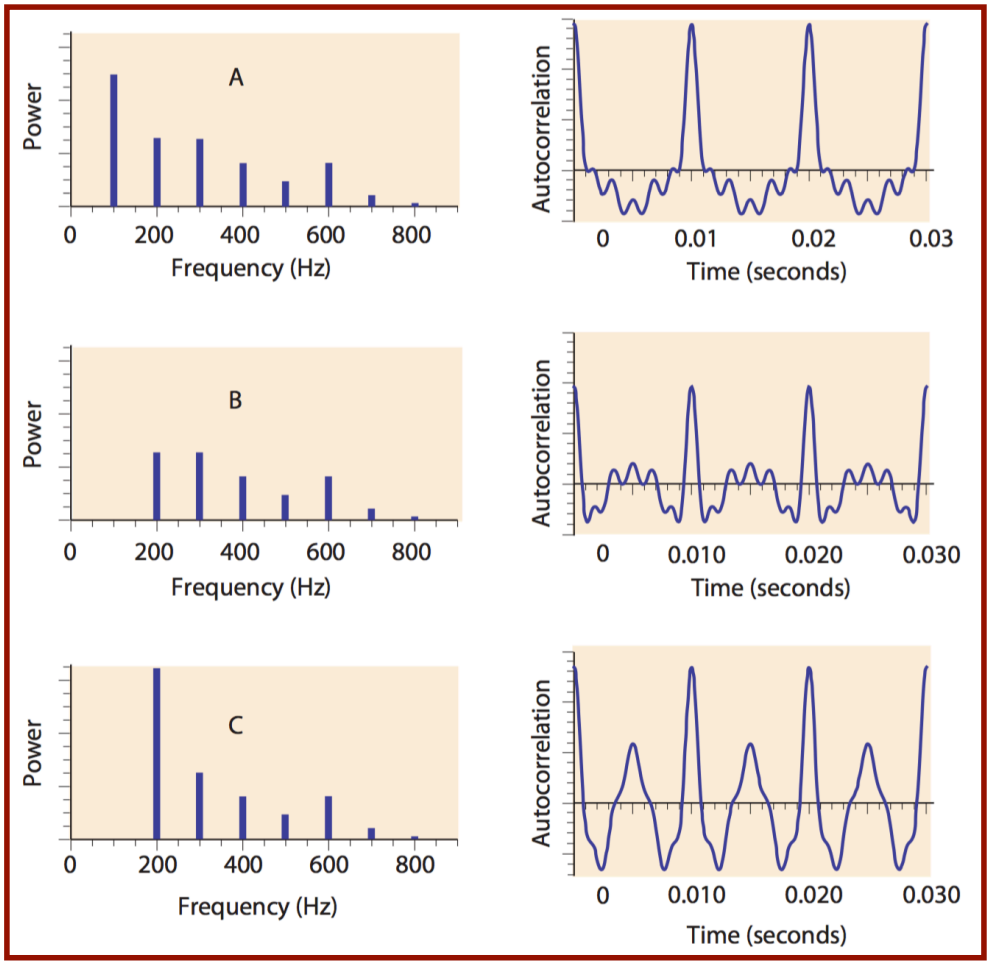
How does the ear compensate for this loss as sound energy is transmitted from the air to the fluid that filled the cochlea?

Last month

Pitch and autocorrelation

- Seebeck's sirens show perceived pitch may be frequency of missing fundamental component
  - "Missing fundamental" effect – Heller calls it "residue pitch"
- Signal with period T has autocorrelation peaks at t=nT, n=0, 1, 2,…
  - Same true for (signal)^2

A. Power spectrum of sound with f_1 =100 Hz and several partials
B. Fundamental removed → same autocorrelation peaks
C. Increase power in f_2 =100 Hz → autocorrelation peaks appear at half-intervals → perceived as sound with 200 Hz fundamental



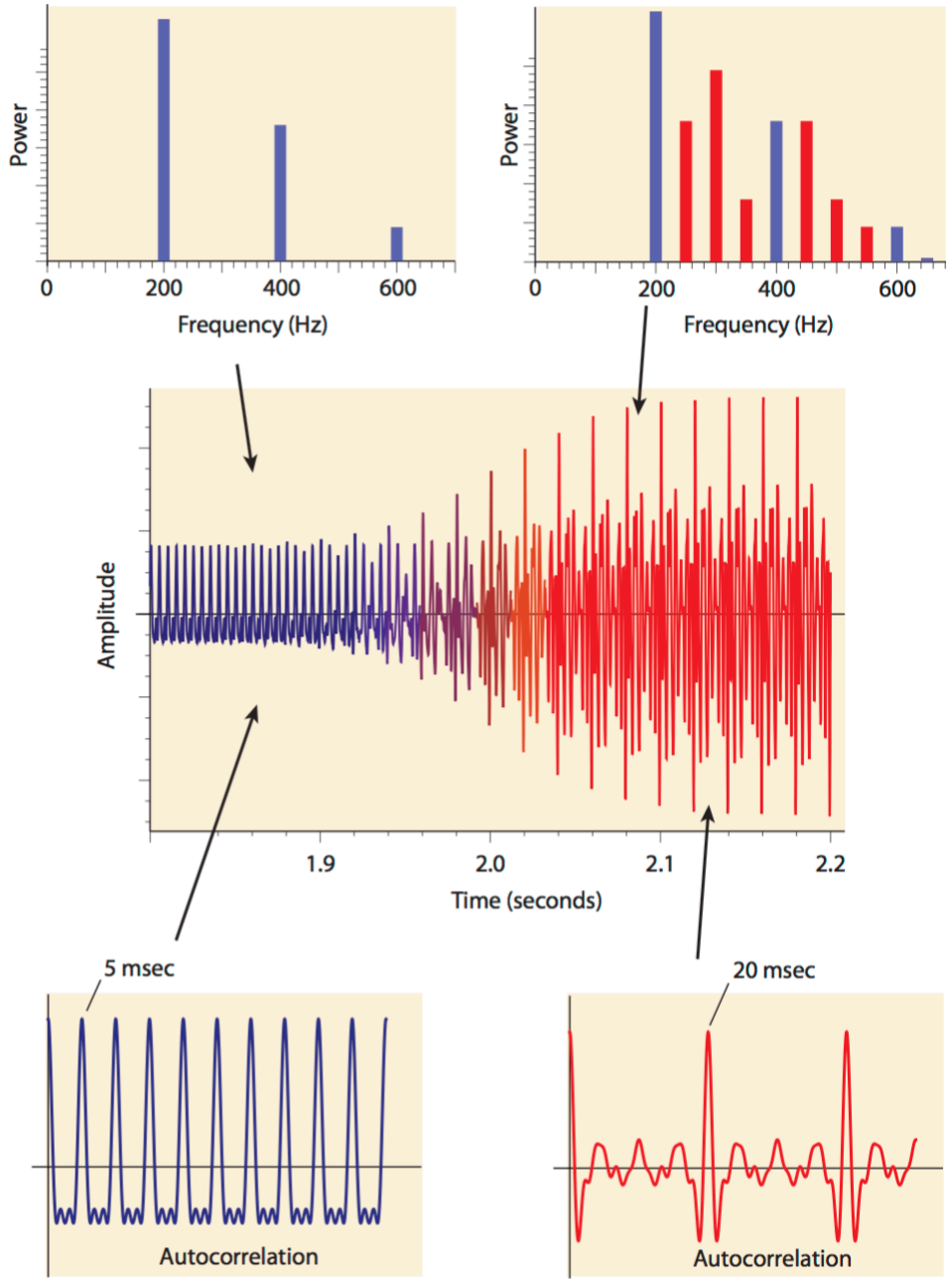
Pitch and autocorrelation

- Example in Heller textbook (p. 448) heller-50HzMissingFund.wav
- Initial signal has 3 partials, 200/400/600 Hz
  - Perceived as 200 Hz sound
- Second signal has additional partials at 50 Hz intervals
  - Perceived as 50 Hz tone

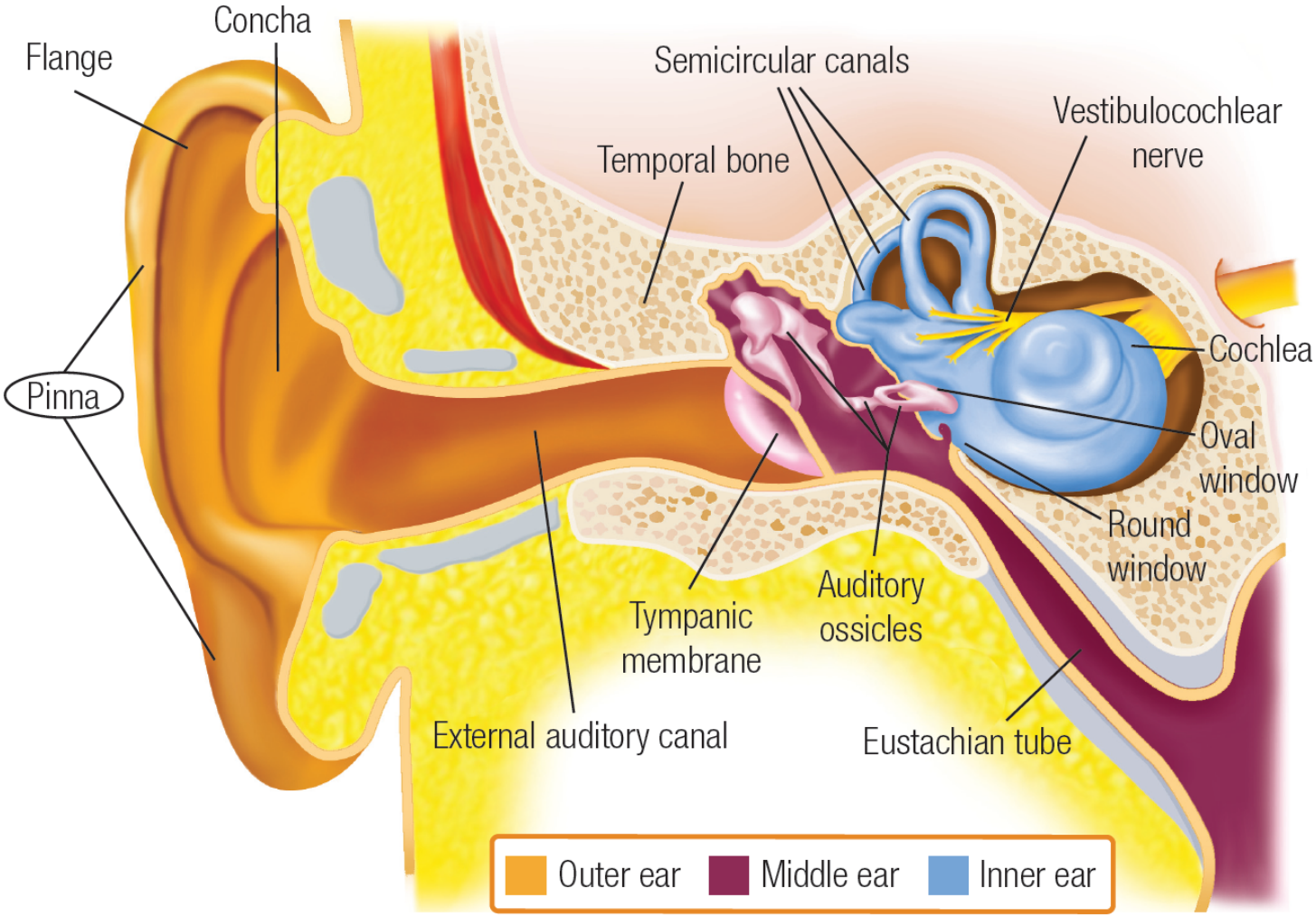
BTW: recall 1 octave = 1200 cents

to get cents from f ratio :

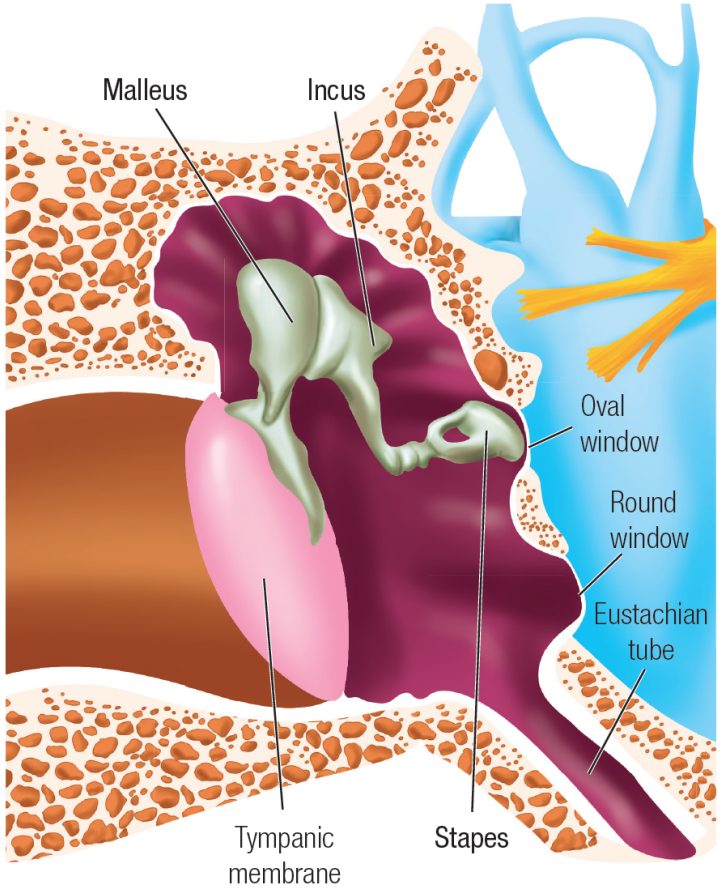
$$I(cents) = 1200 \log\left(\frac{f_2}{f_2}\right) / \log(2)$$



The Ear: Outer, Middle, Inner



The "Impedance Problem": solutions



- Round window vibrates out of phase with oval window (stapes)
  - Pressure relief: Allows vibrations in ~ incompressible cochlear fluid
  - Protects against loud noise

99.9% of sound energy in the air is reflected at the air:water boundary (10 log(0.1/100)) = **-30 dB loss**) (1/1000x)

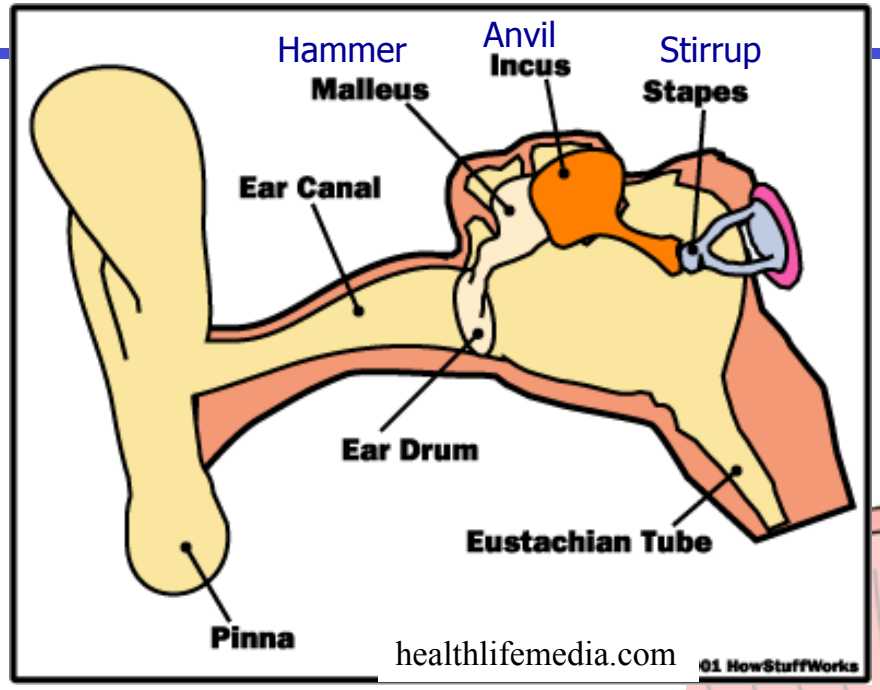
How does the ear compensate for this loss as sound energy is transmitted from the air to the fluid that filled the cochlea?

2 dB gain via ossicular leverage (1.6x)

25 dB gain via surface area condensation (eardrum → stapes) (316x)

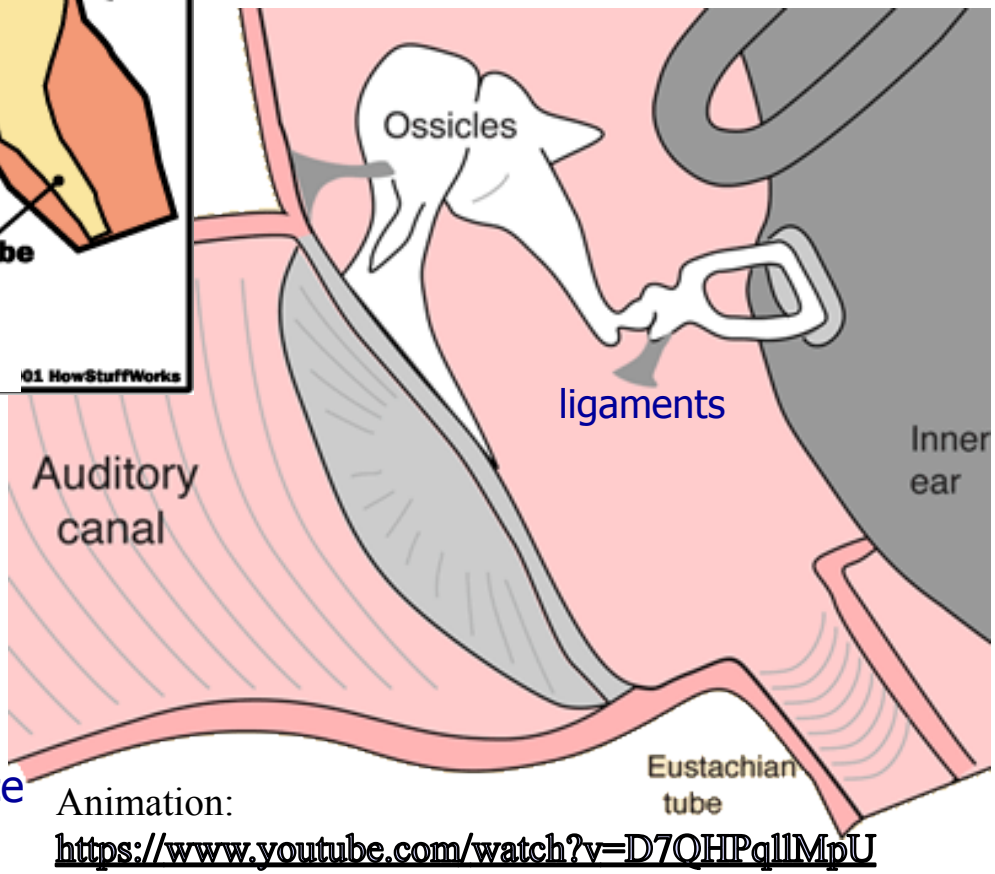
~5 dB gain at mid-frequencies (3x) due to pinna and auditory canal resonance

Middle ear structures



Stapes acts as a piston on inner ear fluid

Middle ear transfers sound p from eardrum to stapes, amplifying with 2 kinds of leverage:



- Hydraulic leverage:

  $S_{EARDRUM} \sim 55\ mm^2$

  $S_{STAPES} \sim 3\ mm^2$

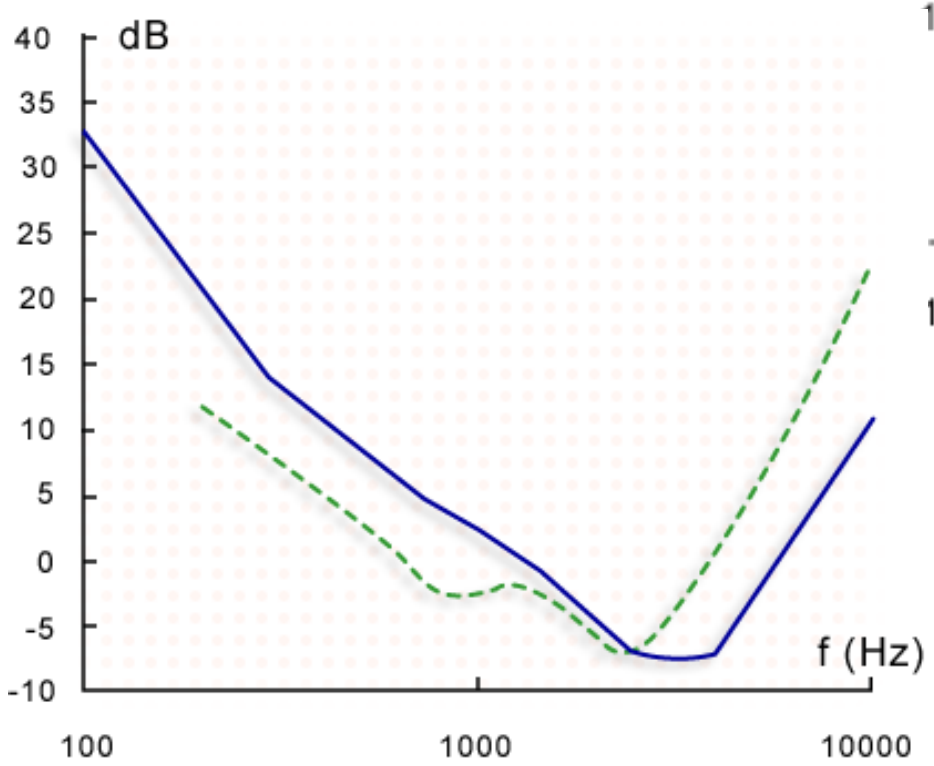
- Mechanical leverage

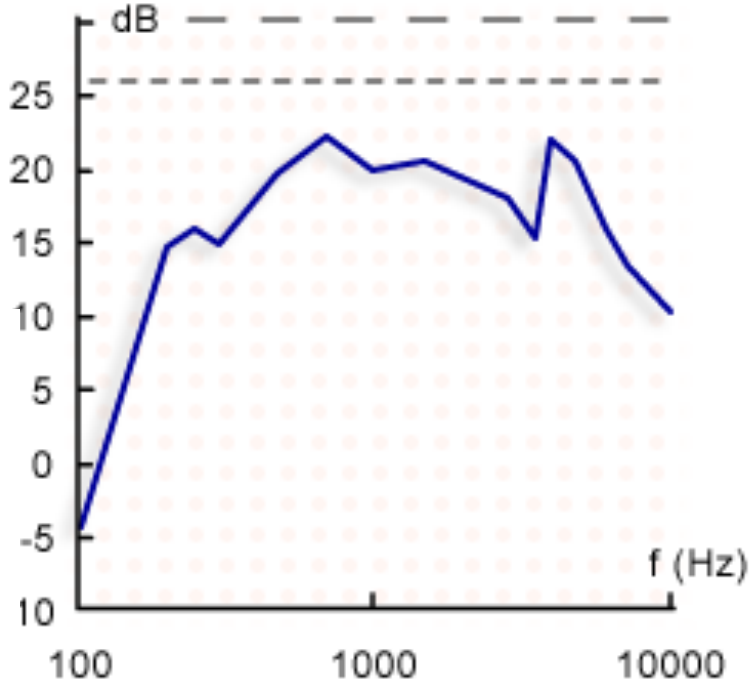
  Malleus is longer than incus, moves a greater distance, so incus moves with greater force (leverage: $E = F \cdot d = const$).

Animation:
https://www.youtube.com/watch?v=D7QHPq1lMpU

Middle ear pressure amplification

- Middle ear transfer function: ratio of sound p at stapes/eardrum: ~20 dB gain at center of hearing range

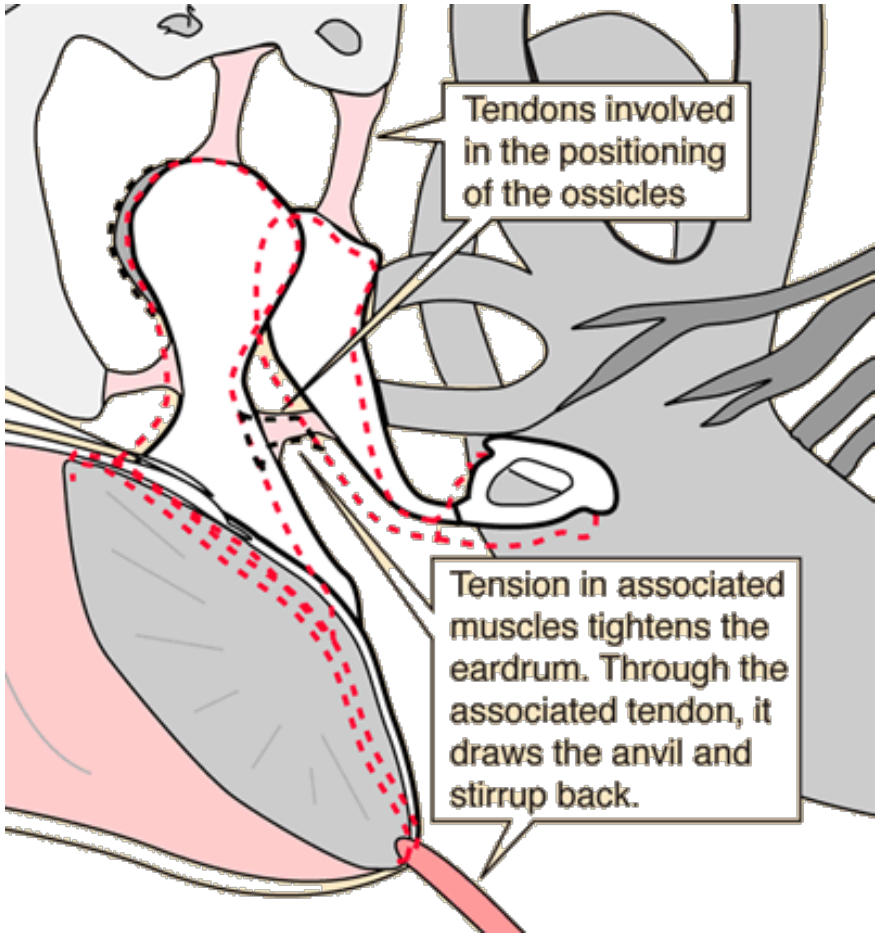




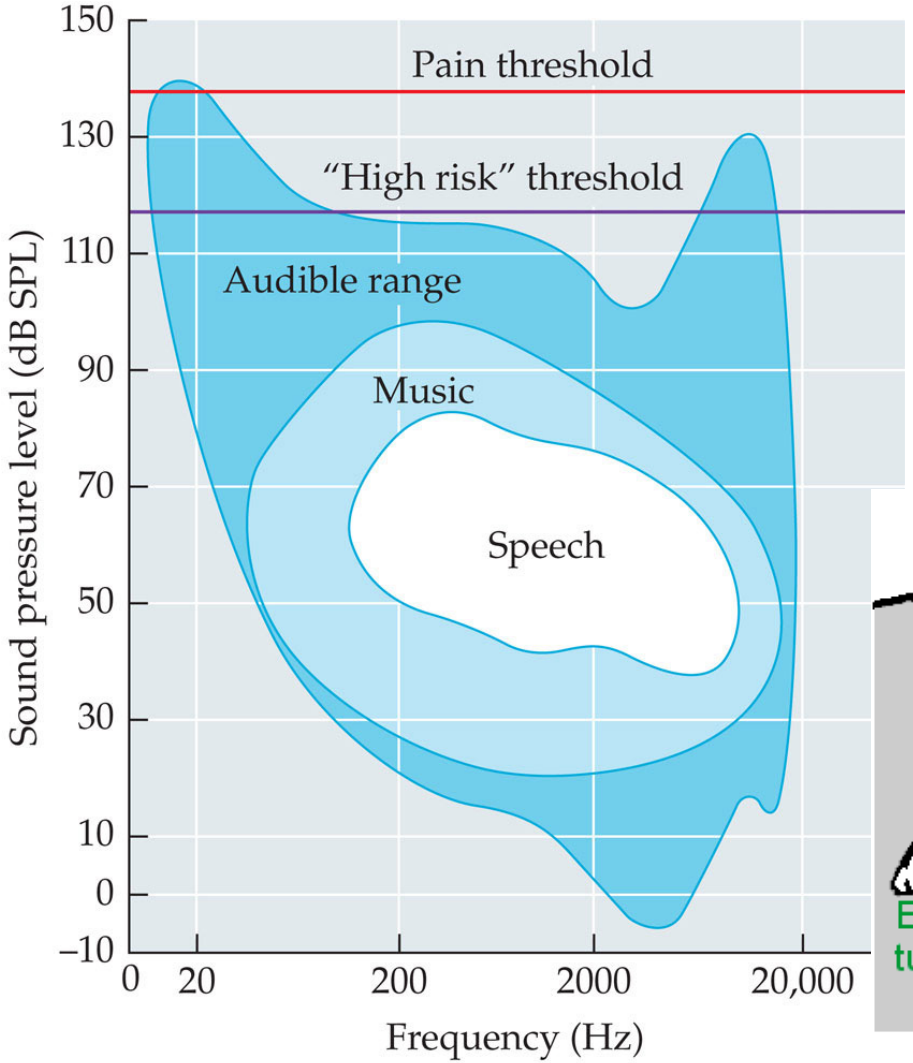
- Human sensitivity curve (solid) vs contribution of external and middle ears (dashed)
  - Similar for all mammals

Middle ear: overload protection

- Response to loud sounds:
  - Tensor tympani muscle tightens the eardrum
  - Via tendons, also shifts the stirrup backward from the oval window of the inner ear.
- Shifting back the ossicles reduces force transmitted to inner ear, protecting it - but
  - Relatively slow response: cannot protect the ear from sudden loud sounds

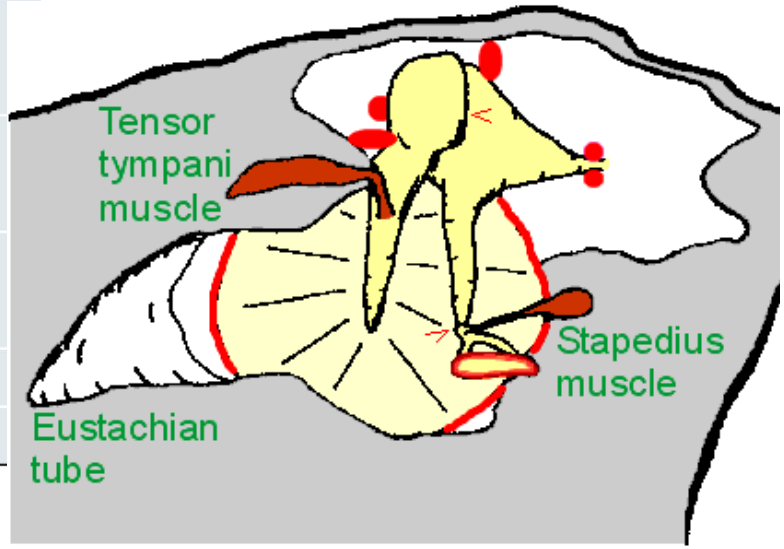


Human "Earscape": SPL vs f range

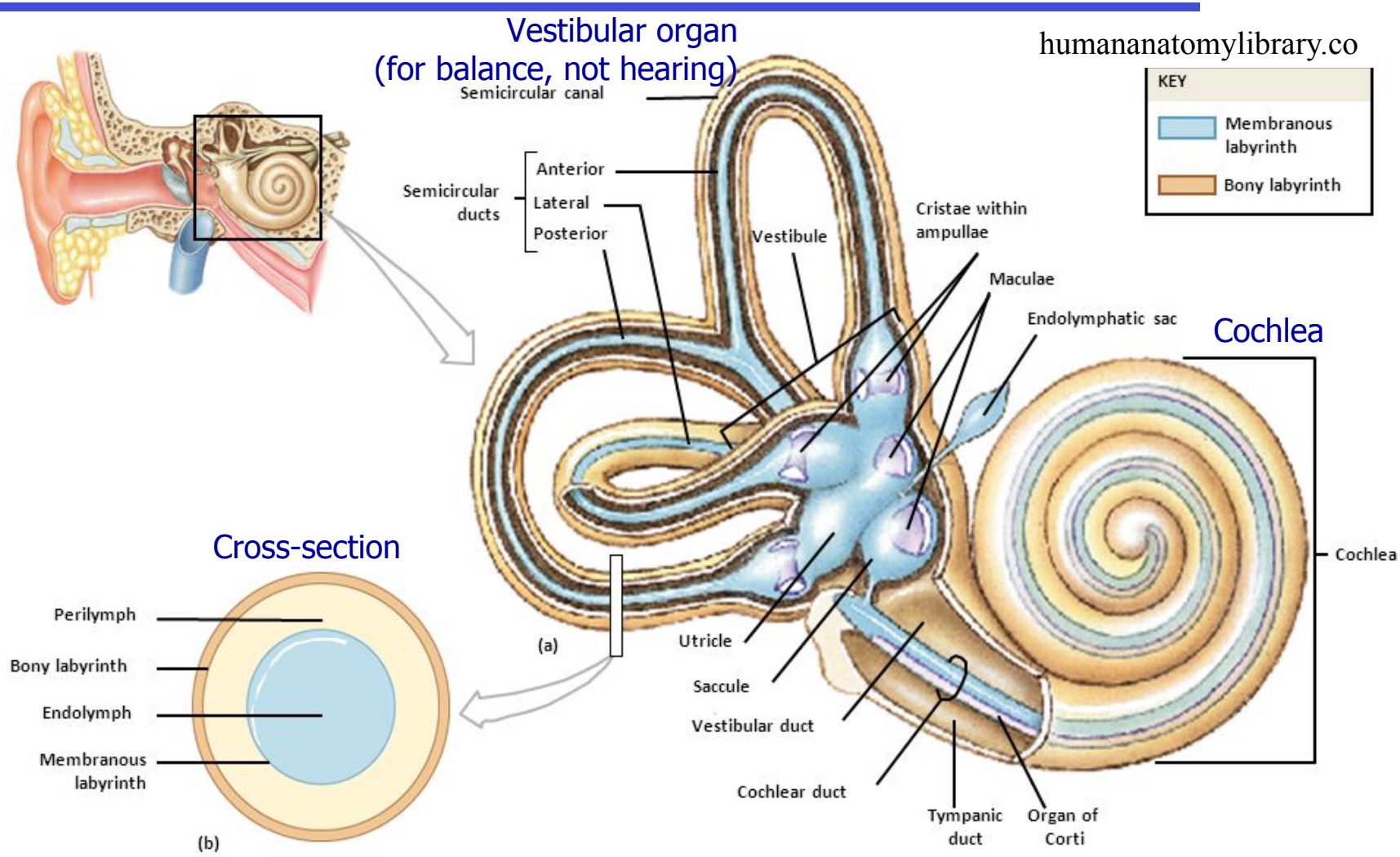


SENSATION & PERCEPTION 4e, Figure 9.3

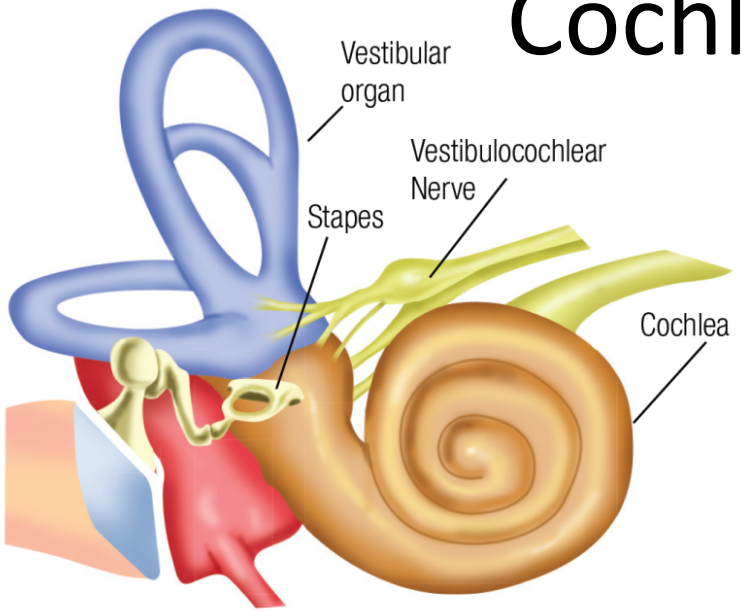
- Middle Ear Ligaments
  - restrict leveraging effects of ossicles
  - limit ossicle motion to reduce the chance of damage to the inner ear

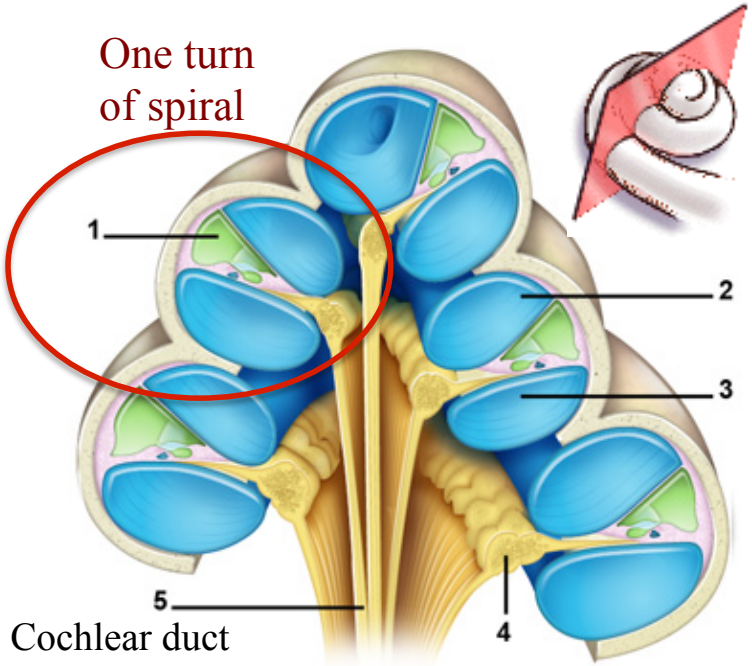


Inner ear: Cochlea and Organ of Corti

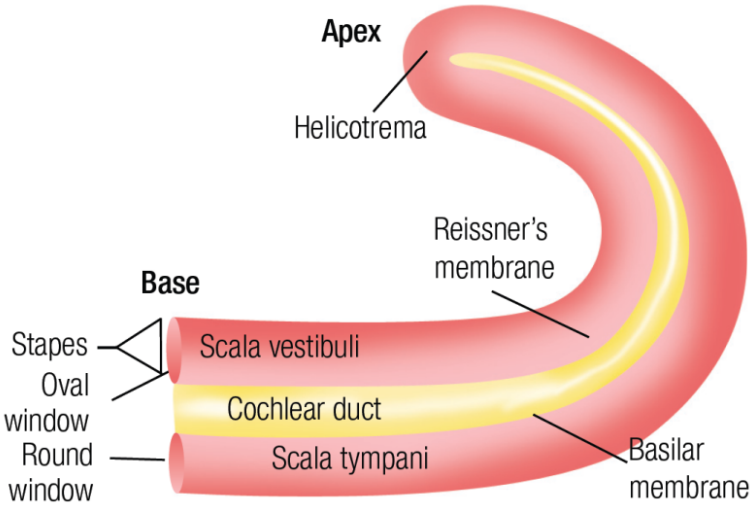


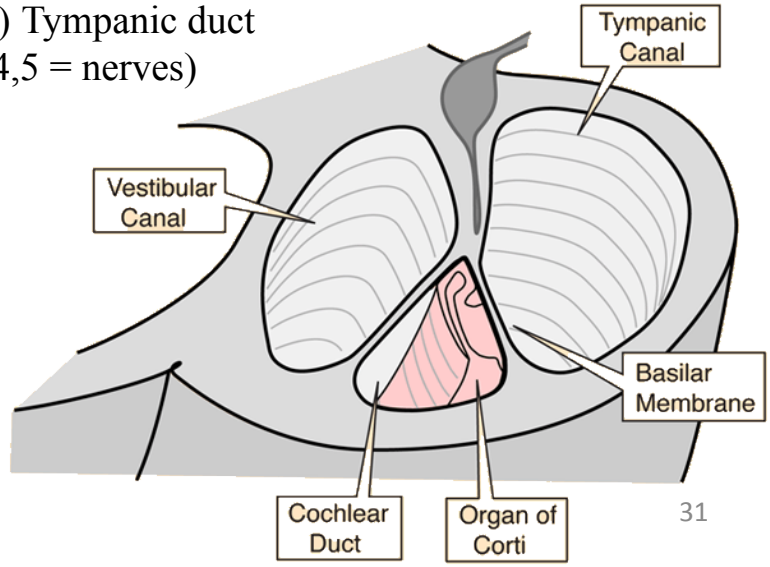
Cochlea



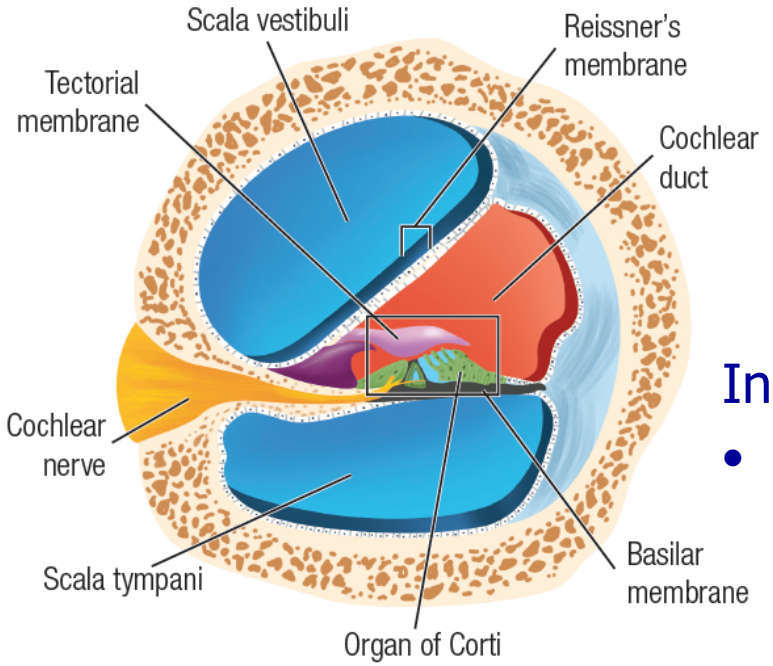


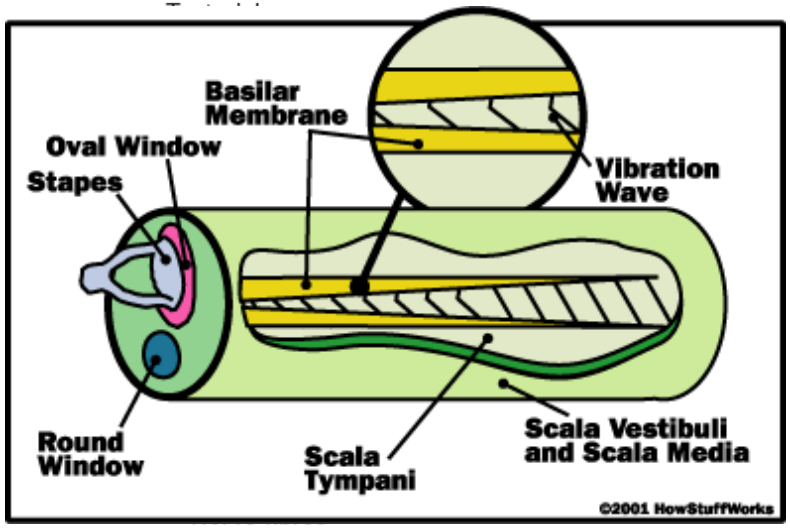
1) Cochlear duct
2) Vestibular duct
3) Tympanic duct
(4,5 = nerves)





Organ of Corti



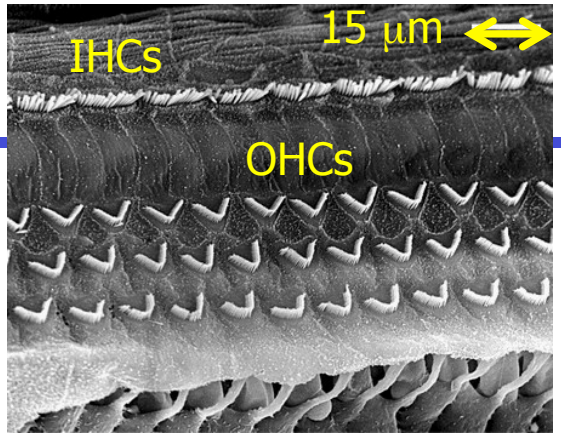


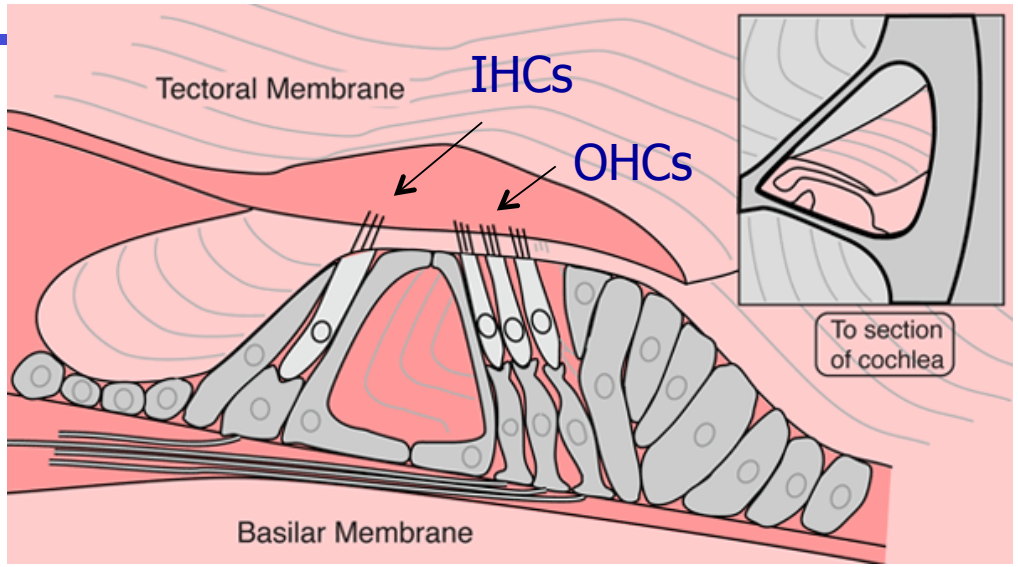
Organ of Corti (in central membrane of cochlea) transforms sound vibrations into nerve signals to brain via its hair cells

Inner and outer hair cells (IHCs/OHCs)

- human cochlea has 3,500 IHCs and ~12,000 OHCs.
  - Tiny numbers: compare to millions of receptors in the retina and the nose!
  - IHCs are primary input devices ("Passive mechanism" of hearing)
  - OHCs work as feedback devices to improve perception ("active mechanism")

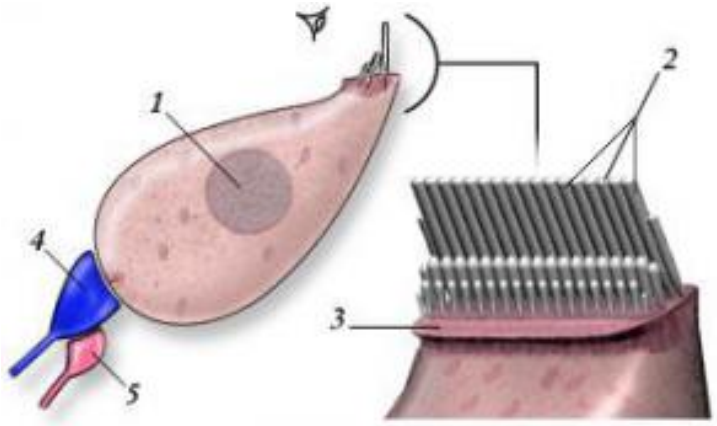
Inner and Outer hair cells (IHCs/OHCs)



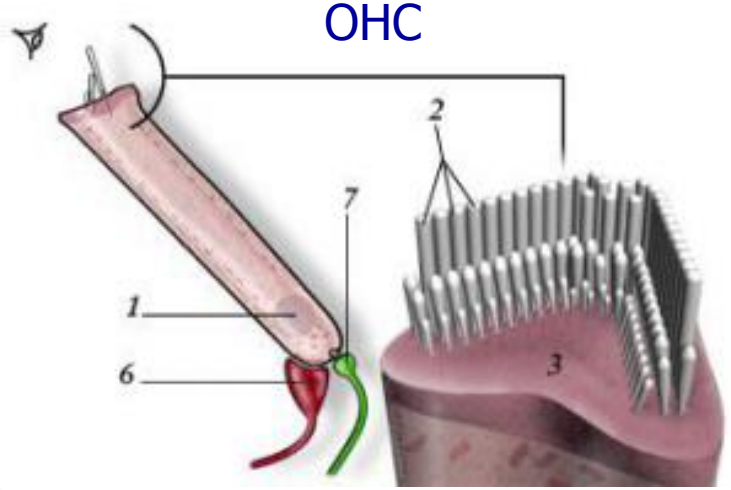


Cell parts
1: Nucleus, 2: Stereocilia,
3: Cuticular plate,
4, 5 and 6, 7: nerve endings
(dendrites of neurons)

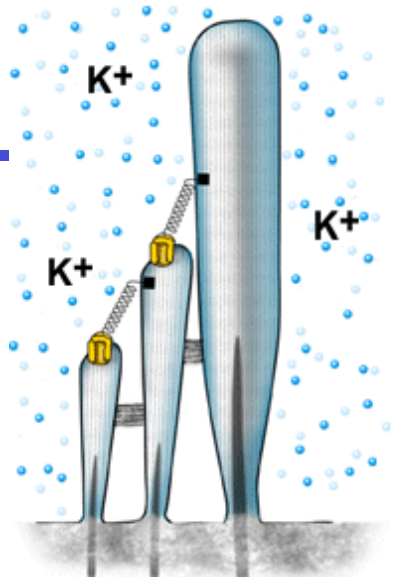
IHC



OHC



Outer hair cells (OHCs)



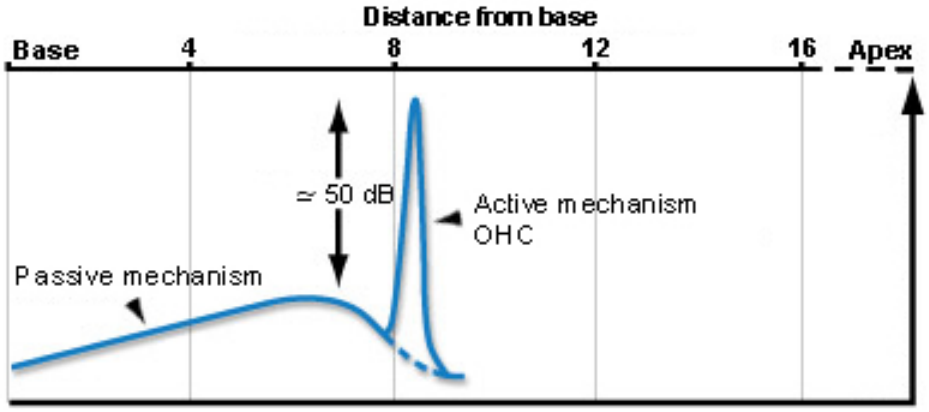
- Like IHCs, acoustic vibrations → motion of of the OHC's stereocilia → modulation in cellular potential
  - Open to admit K+ ions needed for nerve firing
- Unlike IHCs, variations in OHC's potential causes them to change length
  - Due to voltage-sensitive shape of prestin = protein in the cellular membrane

Result: OHCs become oscillators that reinforce the incoming mechanical vibration: positive feedback

- electro-mechanical transducer ("active mechanism")

Increases sensitivity 50~60 dB over a small range of cochlea

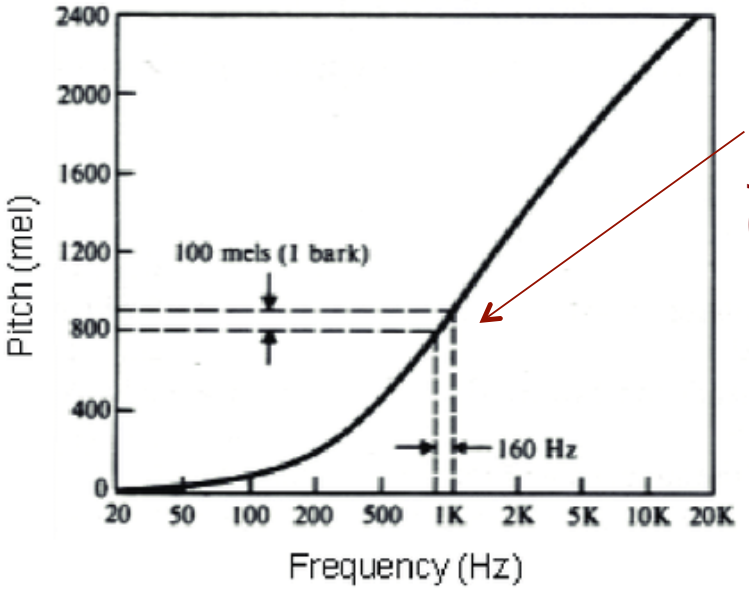
→ Closely spaced f's activate separate cochlear regions: fine-scale frequency selectivity



Pitch perception

Perception of pitch involves fine f discrimination

- "Just Noticeable Difference" (JND): two tones that average listener can just distinguish
- Pitch scale vs frequency: unit of subjective pitch is the *mel* : Perceived pitch is "twice as high" when number of mels is doubled
  - "Critical bandwidth" = Δf of effective bandpass filtering by cochlear hairs

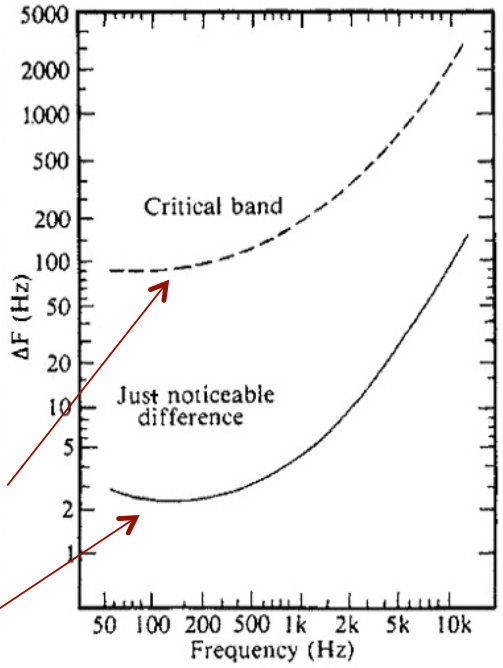


1 bark = critical bandwidth at given f

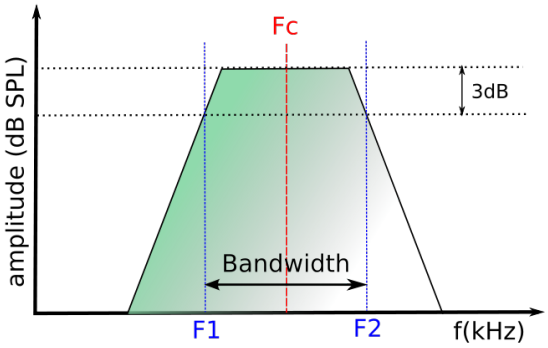
1 bark ~ 100 mels

JND is ~ constant 0.5% of CB

→ CB is connected to pitch perception

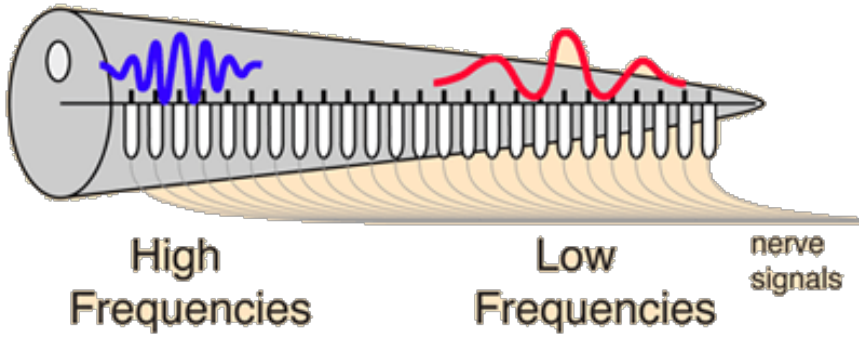


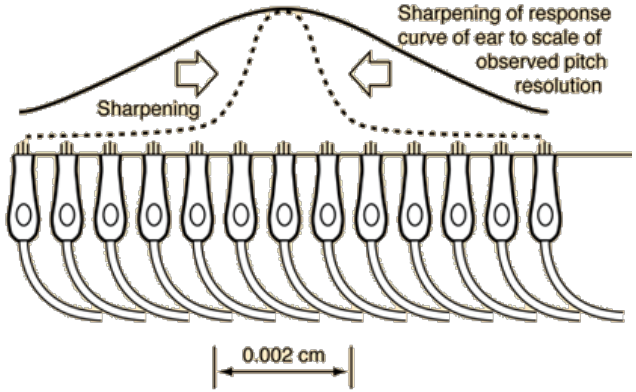
effective bandpass filter for critical bandwidths



Theories of pitch perception

- Place theory: pitch is determined by location on organ of Corti base membrane where hair cells resonate with signal
  - Known that position along cochlea ~ corresponds to frequency: high f near oval window end, low f near spiral tip
  - Works for high frequencies, but have not found locales for lowest f's
  - Narrow CBs (corresponding to JND) → too fine-grained for place theory to explain
- Frequency theory: hairs + neurons act as analog-to-pulse converters -- brain analyzes rate of nerve firings to determine pitch
  - Doesn't work! Neurons cannot fire fast enough for high f's
    - We hear up to 20 kHz, but nerve firing rate ~ 500 Hz max



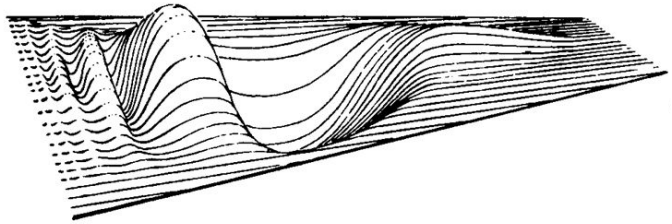


Pitch perception theory evolved…

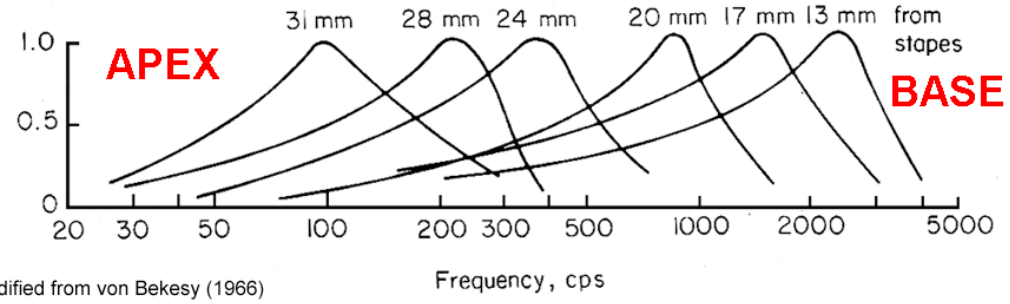
- Place theory (Boltzmann, 1863), Frequency theory (Rutherford, 1886) → Wave theory (Georg von Békésy, 1947)
- Békésy observed surface wave motion of the basilar membrane
  - Cochlea and basilar membrane structures → amplitude maxima of the waves at different locations along the basilar membrane

→ Positive feedback theory (Thomas Gold, 1948)

- Need feedback from OHCs to explain fine discrimination
  - Confirmed by David Kemp (1978): observed otoacoustic emission (OAEs) = sound emitted from inner ear -- now used to check cochlea for damage
  - Bekesy used ears from cadavers: "…dead men don't have OAEs." –J. Hall

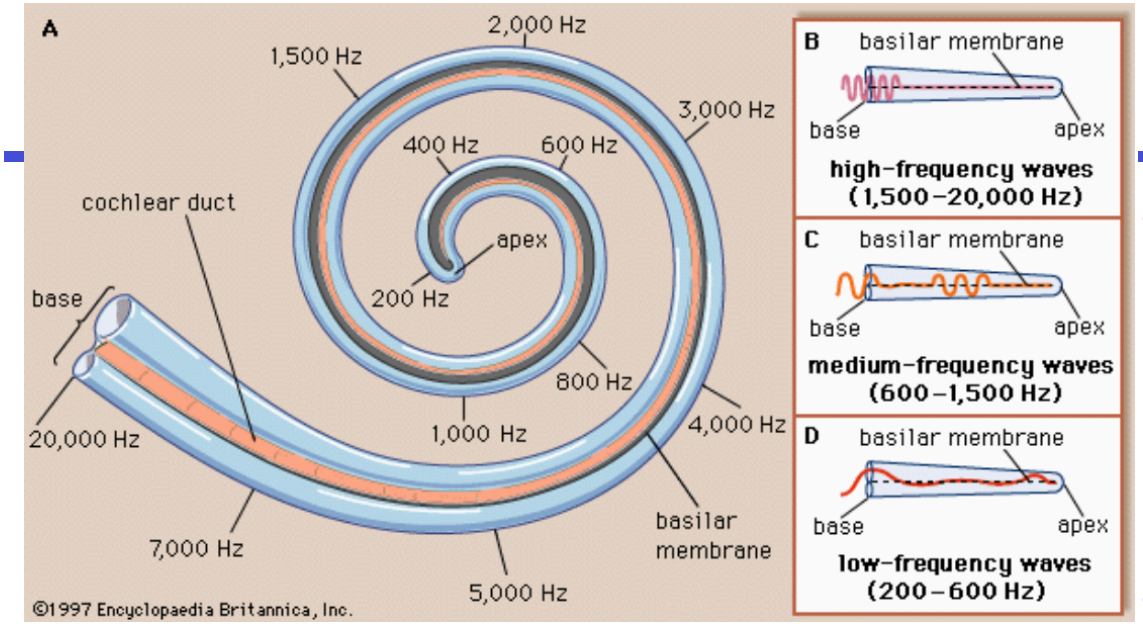


Waves traveling along basilar membrane peak at positions ~ f

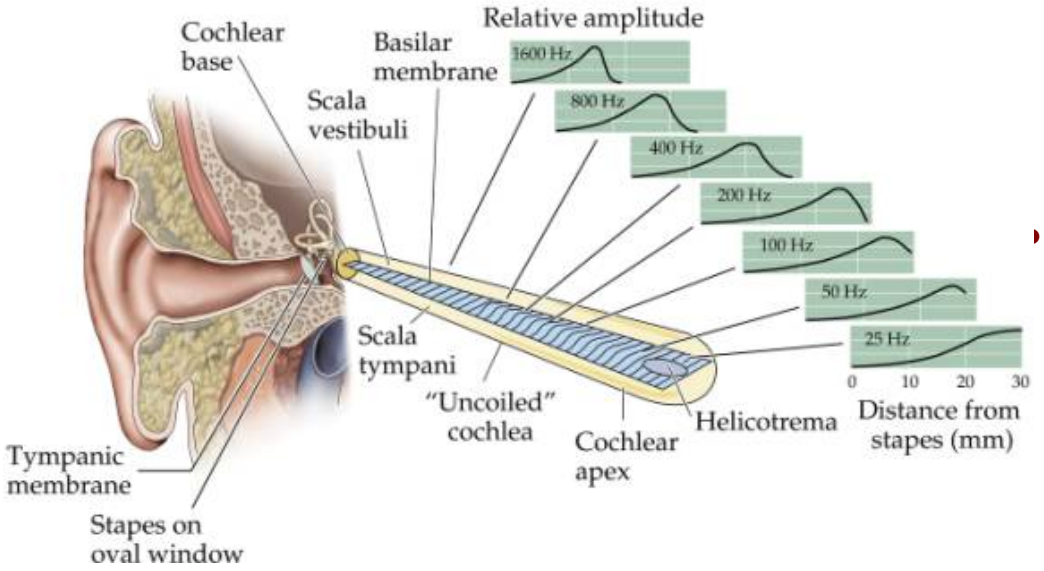


Modified from von Bekesy (1966)

Békésy map



Map with cochlea uncoiled to show locations

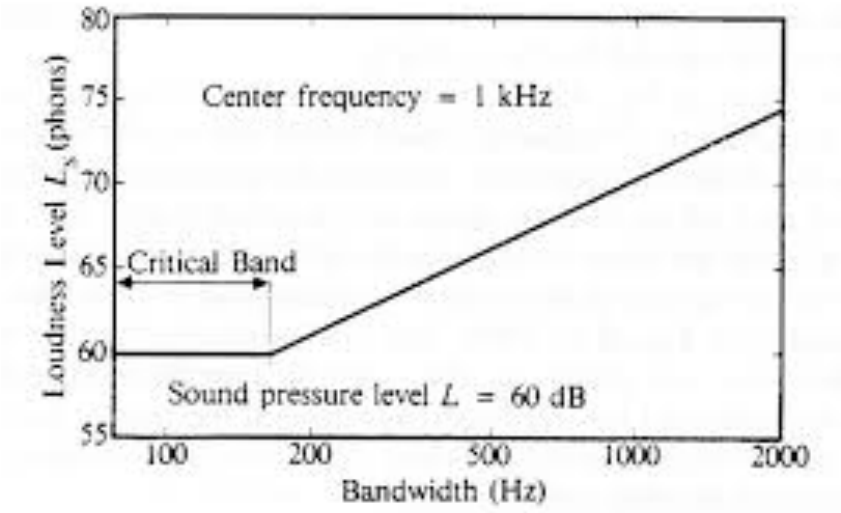


- Georg von Békésy (1899--1972) used strobe photography to observe that the basilar membrane moves like a surface wave when stimulated by sound.
- Different frequencies cause the maximum amplitudes of the waves to occur at different places along the coil of the cochlea.
- Won the 1961 Nobel Prize in Physiology or Medicine.

Critical bandwidth and the cochlea

- Critical band = Δf within which a second tone interferes with perception of the first tone
- One way of observing:
  - Increase bandwidth of a *noise* signal while decreasing amplitude to keep power constant.
  - When Δf is greater than CB, subjective stimulus covers >1 auditory CB

    → perceived loudness increases



Standard audiometric critical bands

CB rate	Center frequency	Frequency	CB bandwidth
Bark	*Hz*	*Hz*	*Hz*
0	50	20	80
1	150	100	100
2	250	200	100
3	350	300	100
4	450	400	110
5	570	510	120
6	700	630	140
7	840	770	150
8	1000	920	160
9	1170	1080	190
10	1370	1270	210
11	1600	1480	240
12	1850	1720	280
13	2150	2000	320
14	2500	2320	380

Critical bands have
Δf ~ constant below 500 Hz :
Compare to Békésy map of cochlea location vs frequency