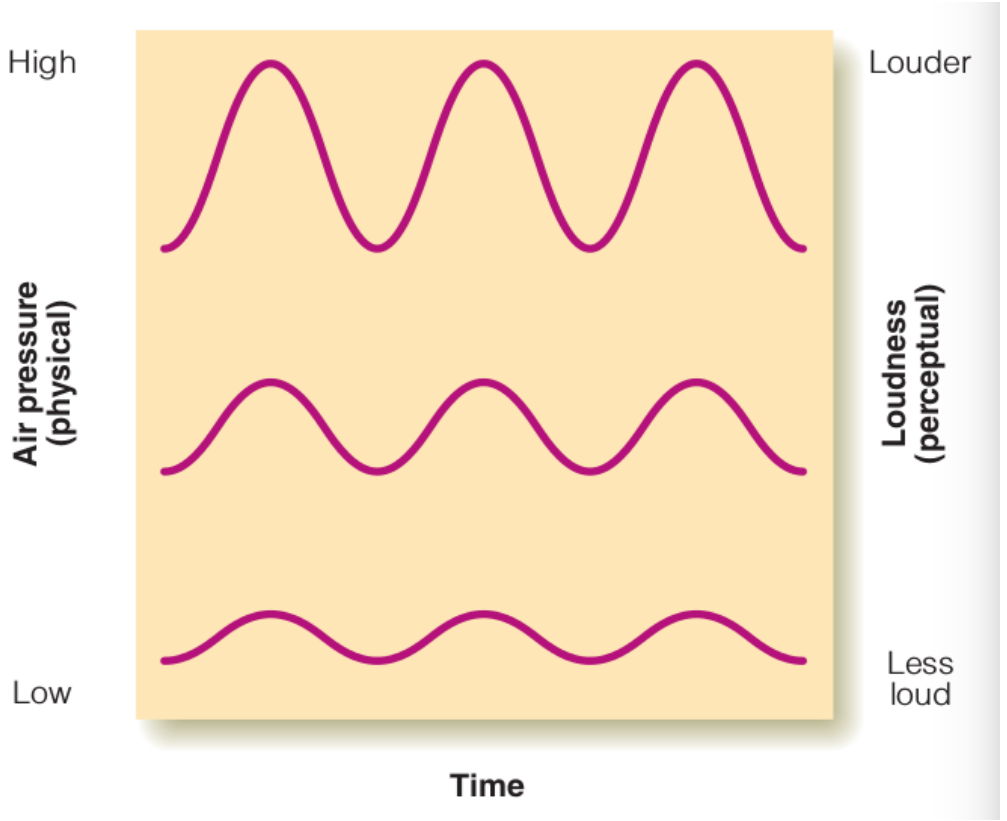
ELL 788
Computational Perception & Cognition

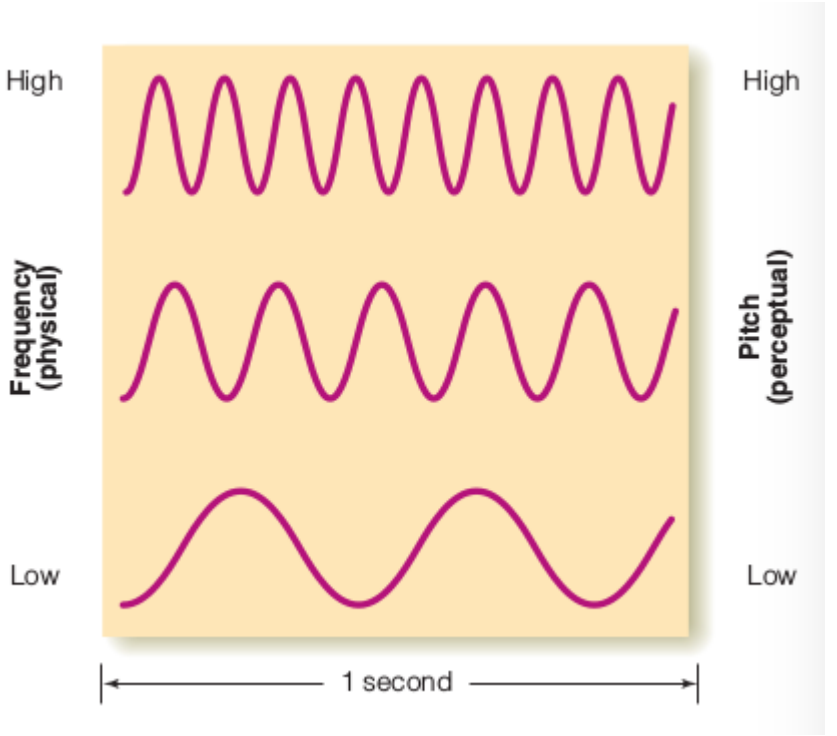
July – November 2015

Module 2

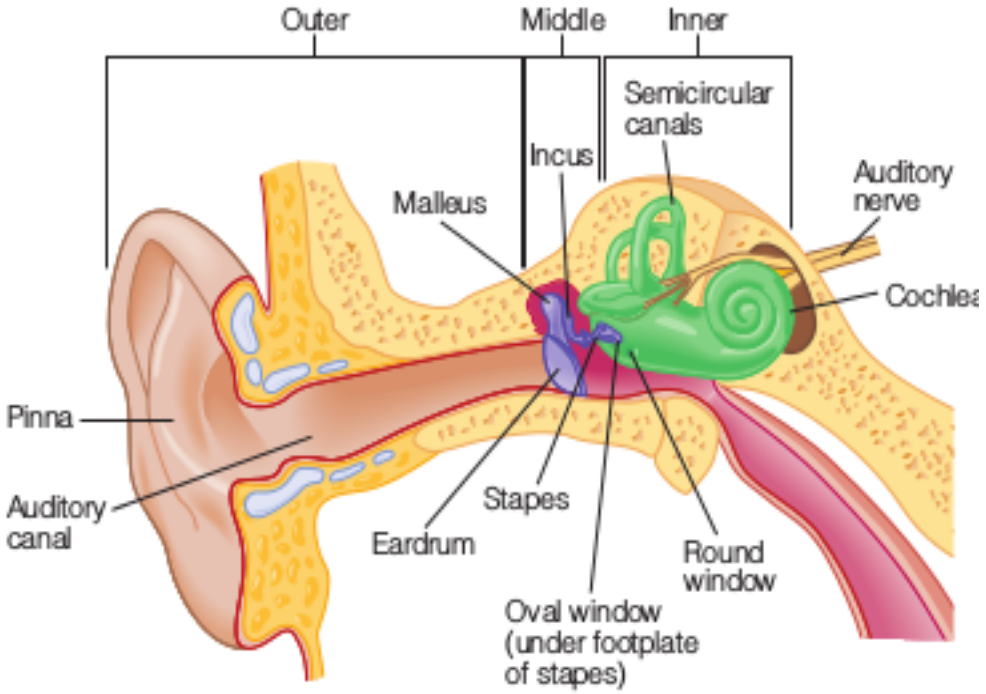
Auditory perception

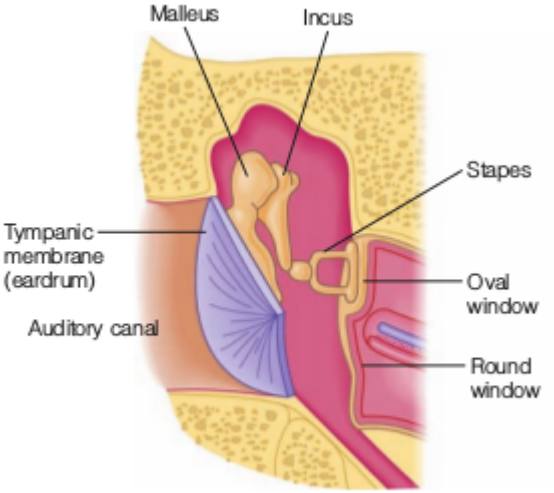
Loudness and pitch





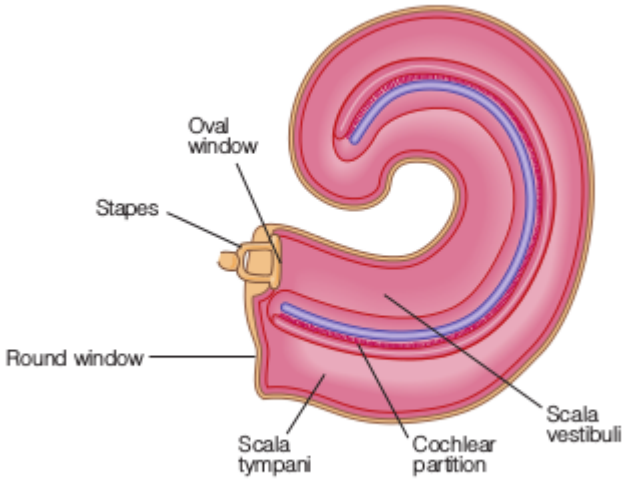
Anatomy of human ear

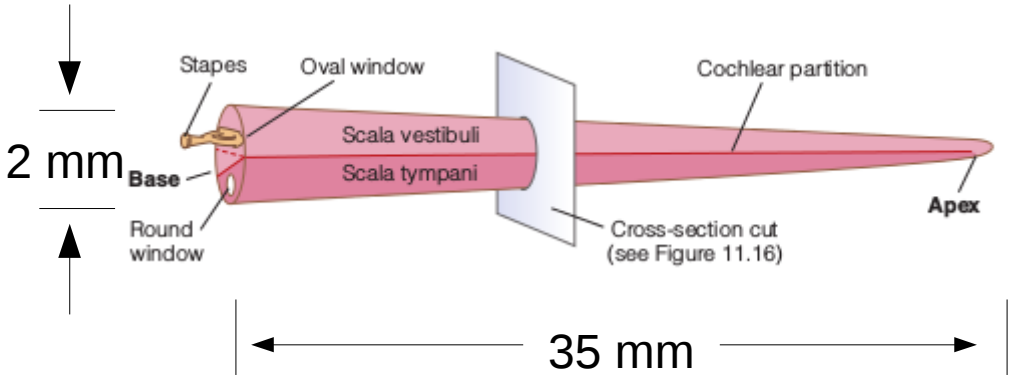


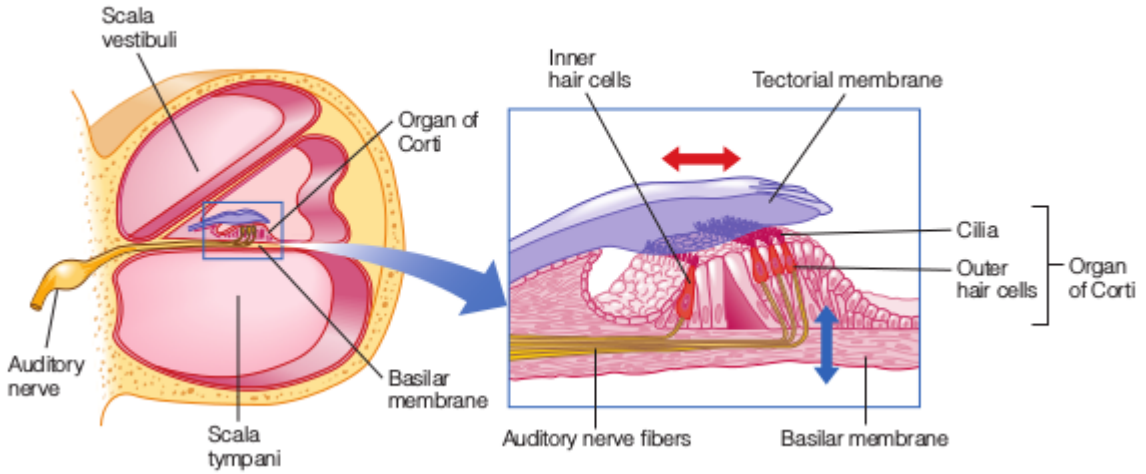


Source: Goldstein (e-book)

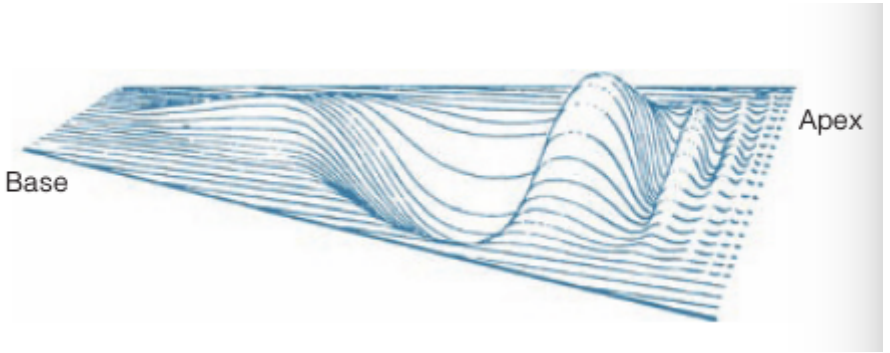
The inner ear (cochlear)

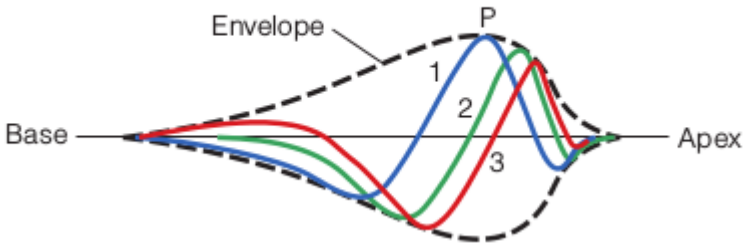


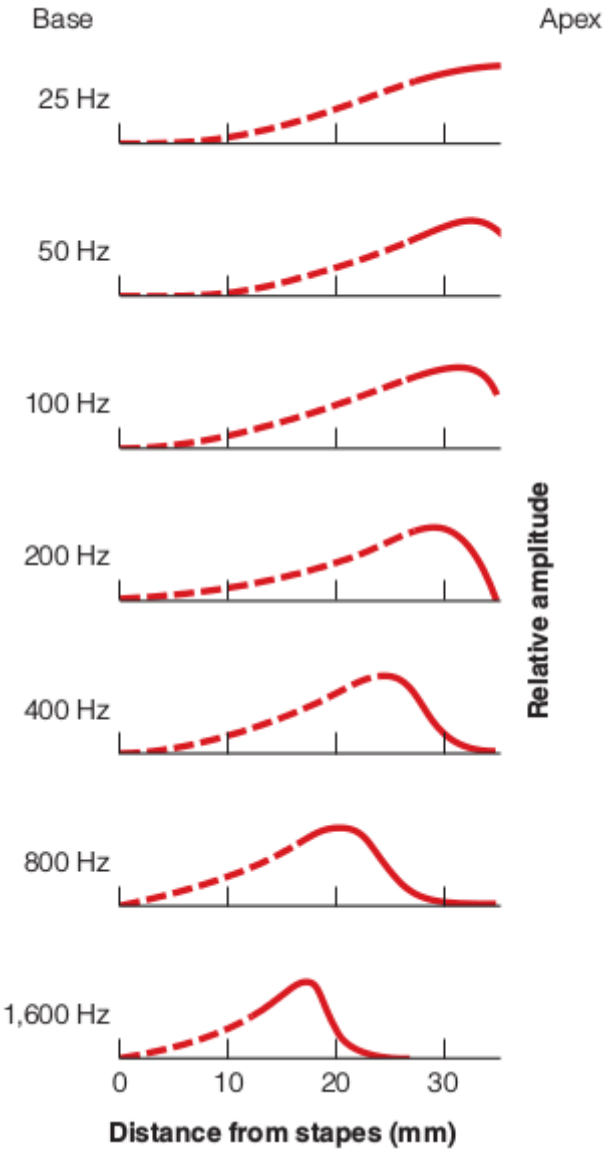




Vibration of basilar membrane

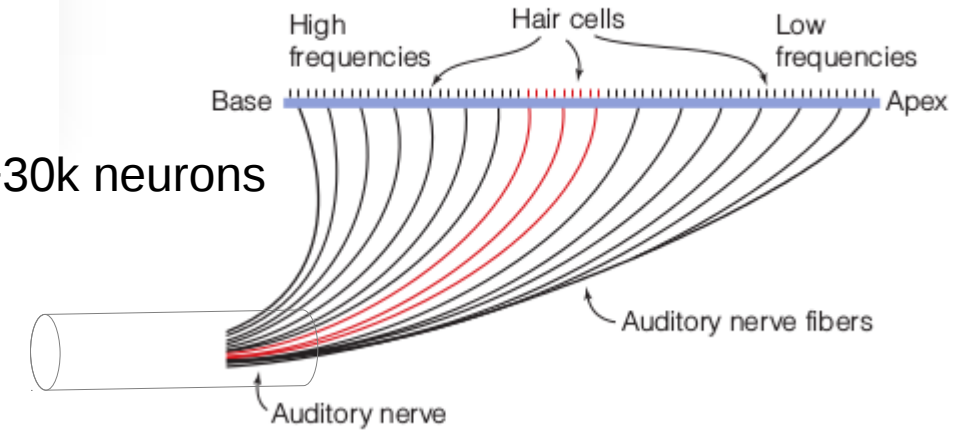


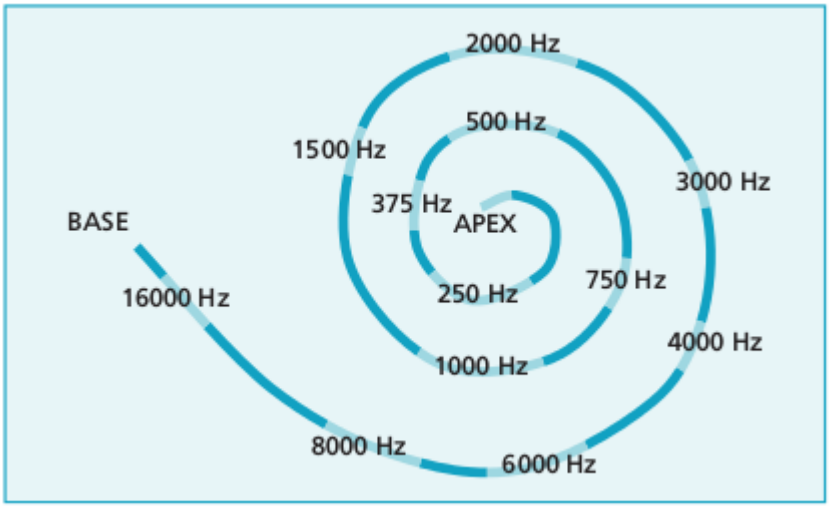




Frequency sensitivity in cochlea

~30k neurons



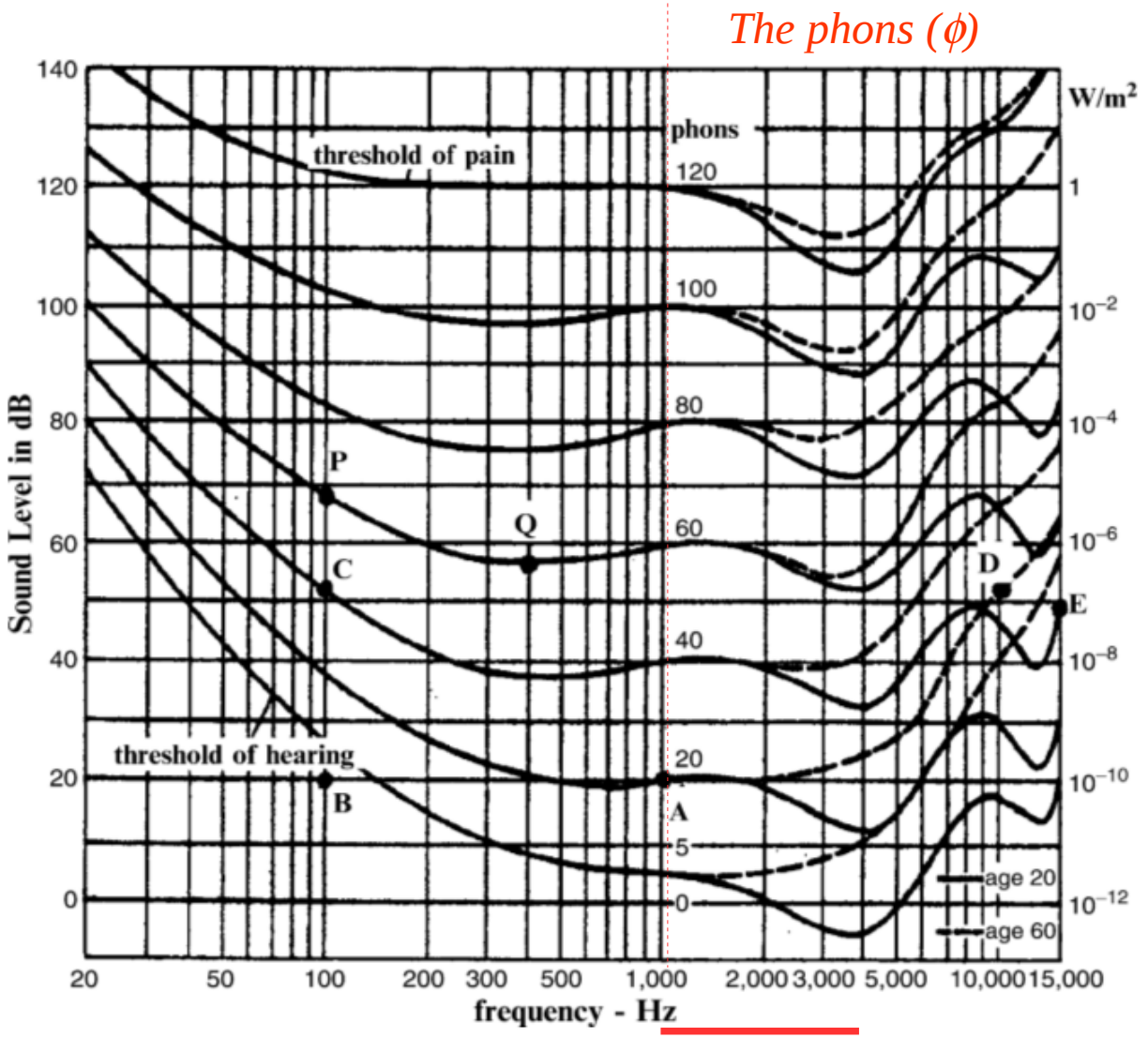


Central nerves originate from apex: cary low frequency
Peripheral nerves originate from base: carry high frequency

Increase in sound intensity produces a greater rate of firing in the neurons in the auditory nerve

- Temporal resolution: ~200 impulses per sec upto 60 dB
- Less sensitive nerves can cope up upto 100 dB

Perception of Loudness



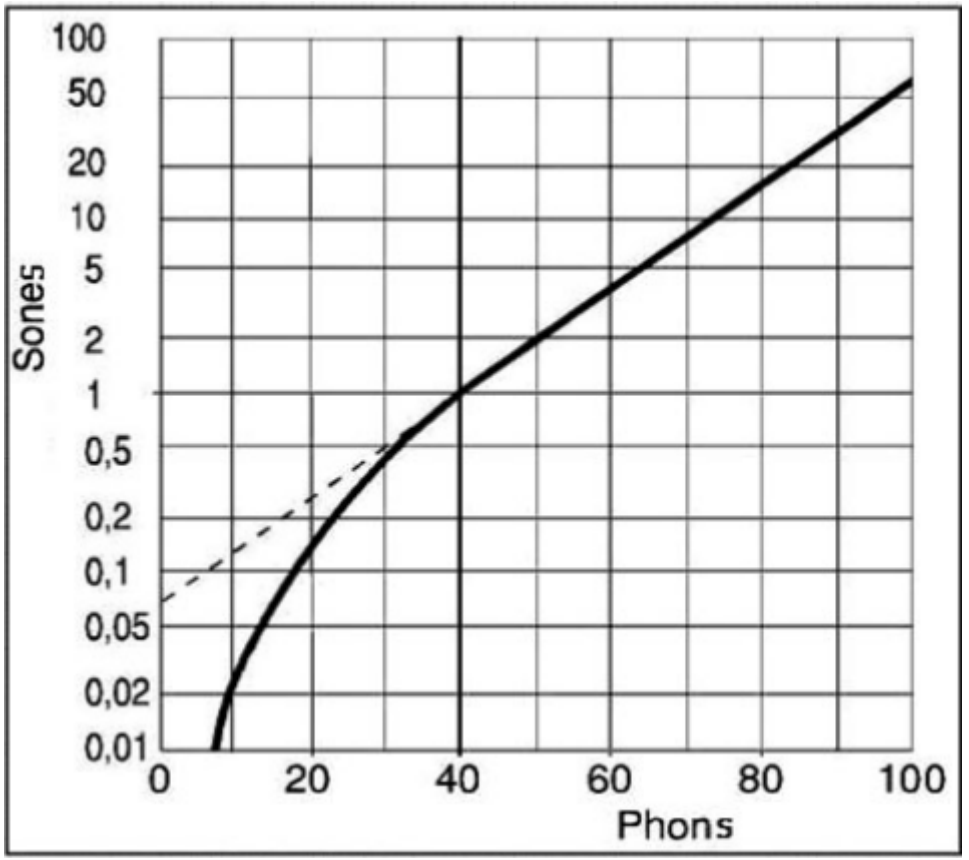
Age: 60 yrs

Age: 20 yrs

Equal Loudness Curves

Max sensitivity
1-4 kHz

Perception of loudness ... more



$$s = 2^{\frac{\phi - 40}{10}} \qquad \text{for } \phi > 40 \text{ phons}$$

$$s = (\phi/40)^{2.86} - 0.005 \qquad \text{for } \phi < 40$$

[Recall Weber law]

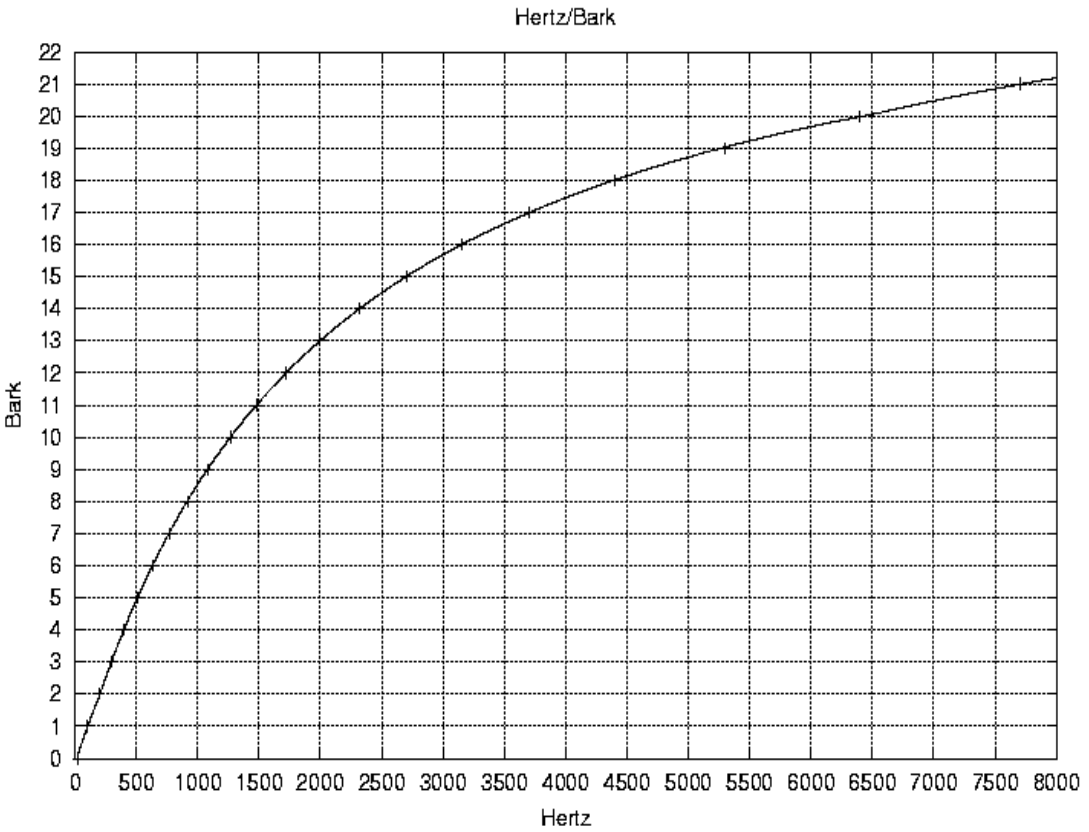
Phon level	40	50	60	70
Loudness in sones	1	2	4	8

Perception of pitch

- Rate-place code
  - Frequency is coded by the place on the basilar membrane that is activated
- Temporal code
  - Frequency is coded by the pattern of phase-locked firing in the auditory nerve
- Temporal code does not work at freq > 5 kHz
  - Upper limit of perception of musical melodies
- Perception of difference of frequency is also guided by temporal code
- Frequency resolution is higher at lower frequency [Weber law]

The frequency scale for human perception is not linear

Bark Scale



Source: Wikipedia

$$\text{Bark} = 13 \arctan(0.00076 f) + 3.5 \arctan((f/7500)^2)$$

$$\text{Critical band rate (bark)} = [(26.81 f)/(1960 + f)] - 0.53$$

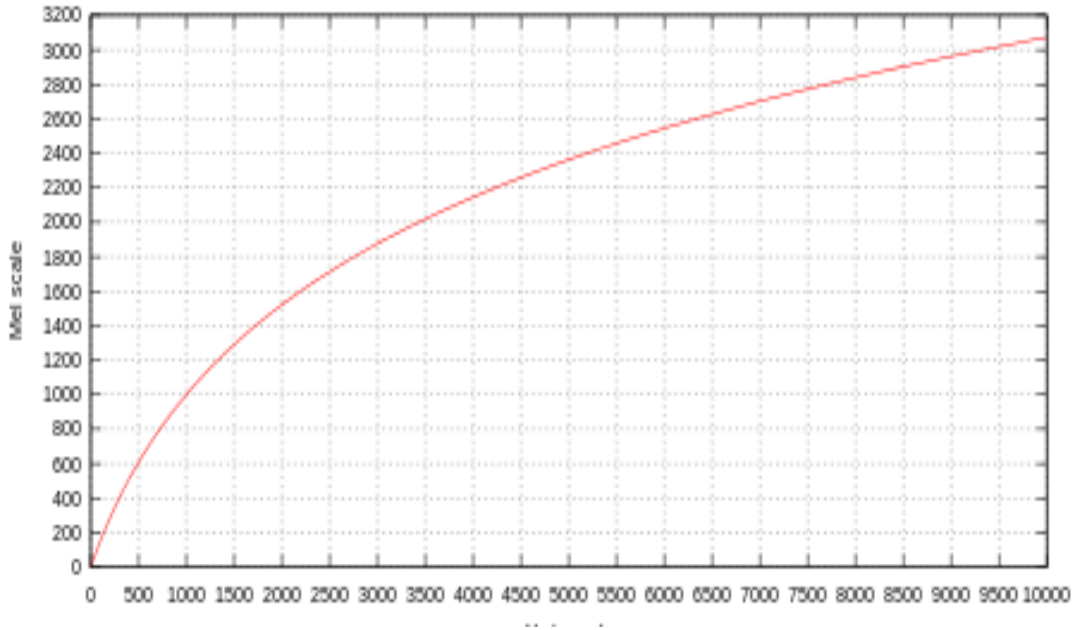
if result < 2 add 0.15*(2-result)

if result > 20.1 add 0.22*(result-20.1)

Other perceptual representations

- Equivalent Rectangular Bandwidth (ERB) Scale
  - Simplified rectangular band-pass filters

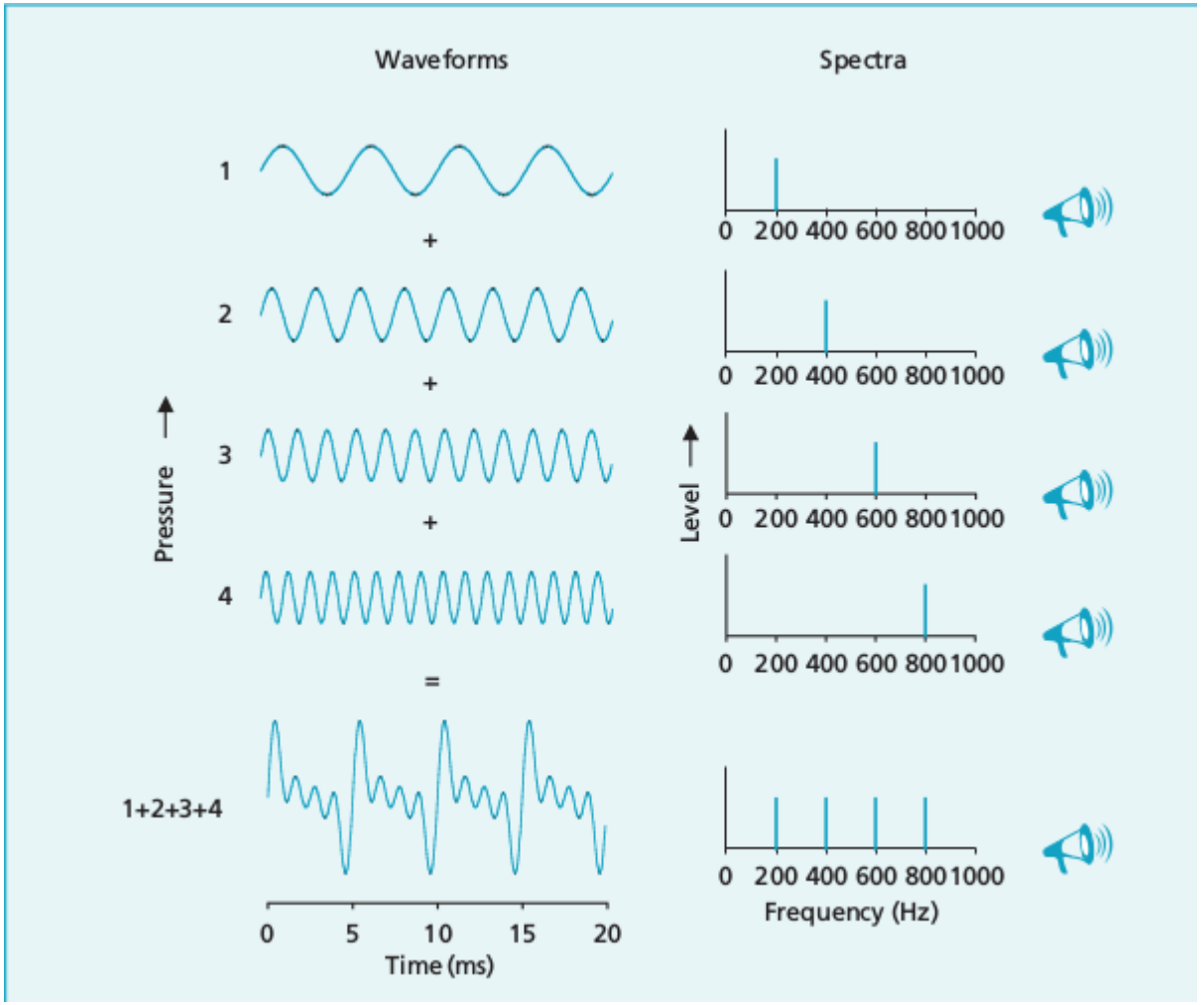
- Mel scale



$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Source: Wikipedia

Audio with multiple tones



Source: Plack

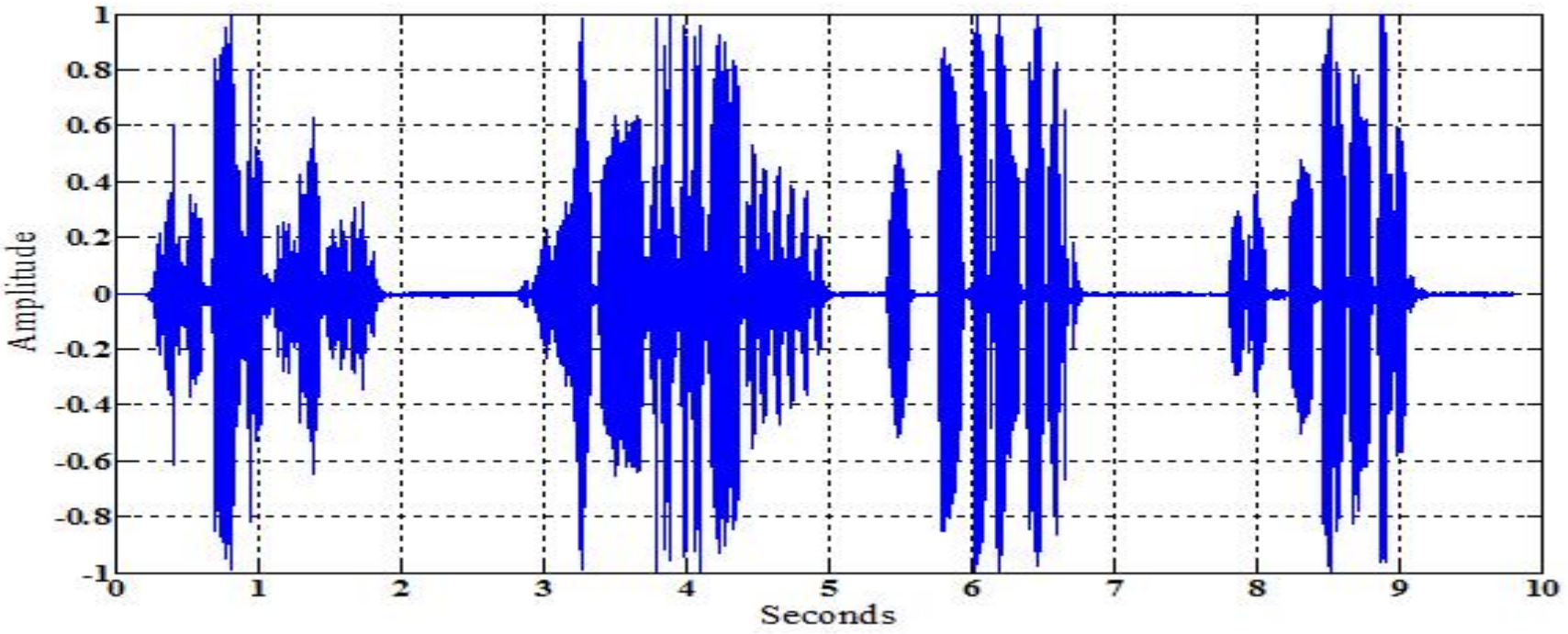
Fourier analysis

Transforms a signal from time domain to frequency domain and vice-versa

Source: Wikipedia

Time-frequency analysis

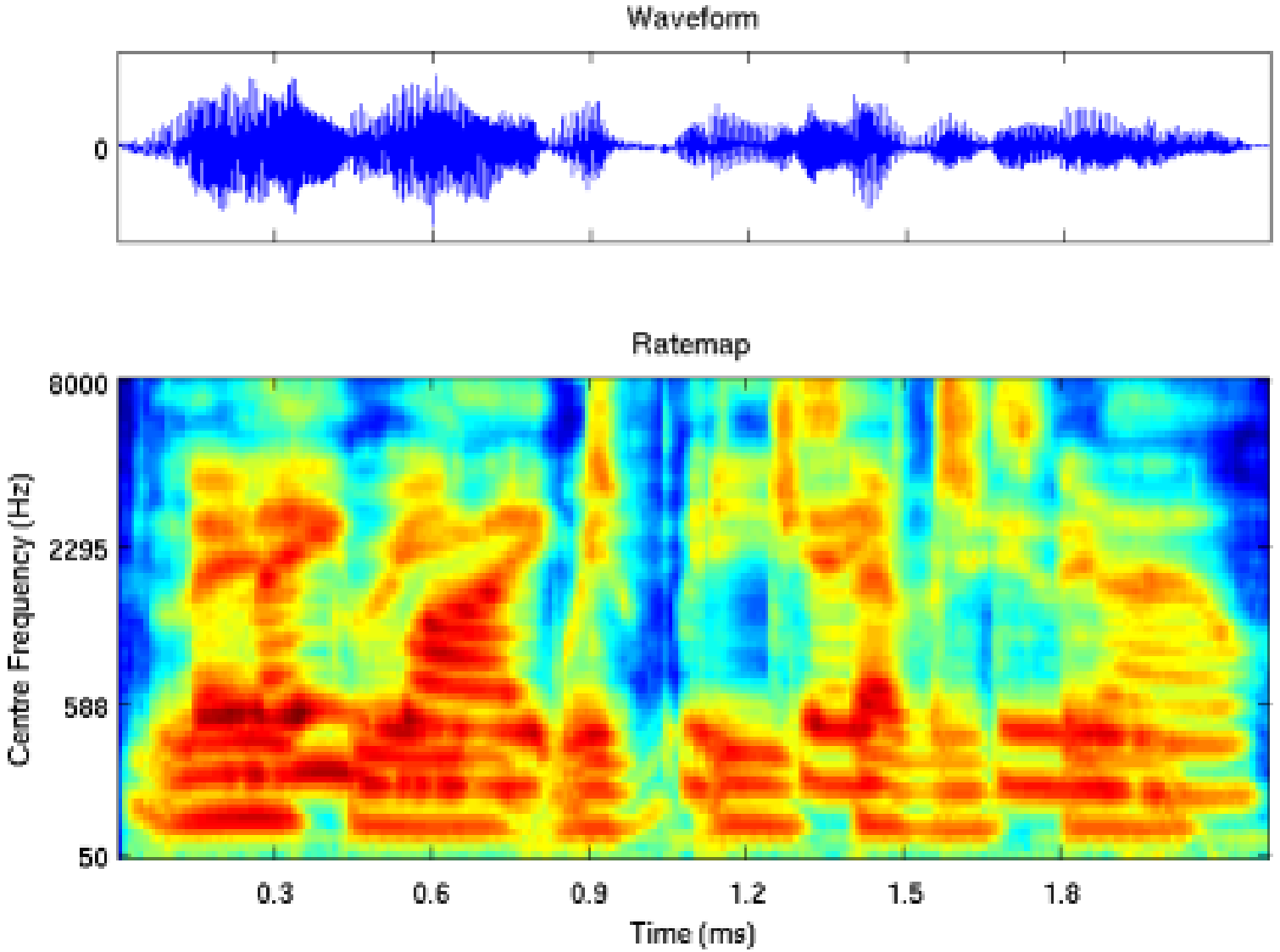
- Non-stationary random process
- Different stationary process of short durations on the same timeline
- DFT in time window ~10 ms



Source: Internet

Cochleagram

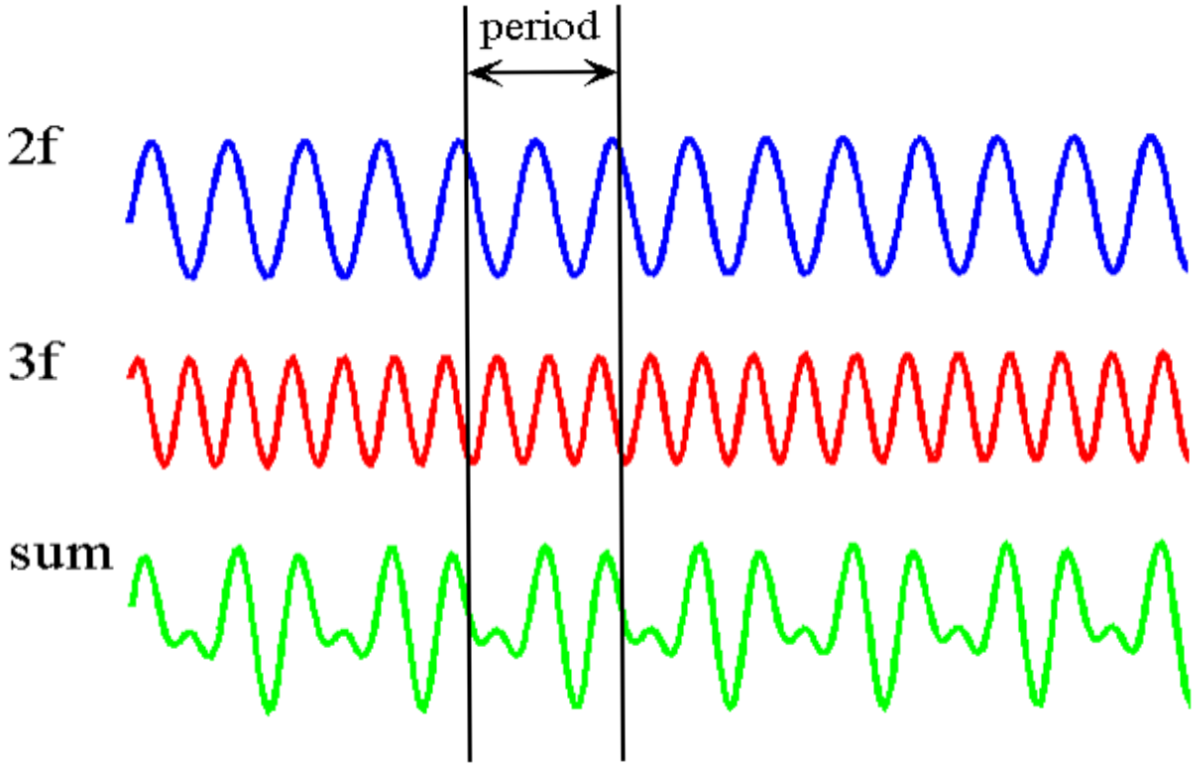
A time-frequency representation of audio signal



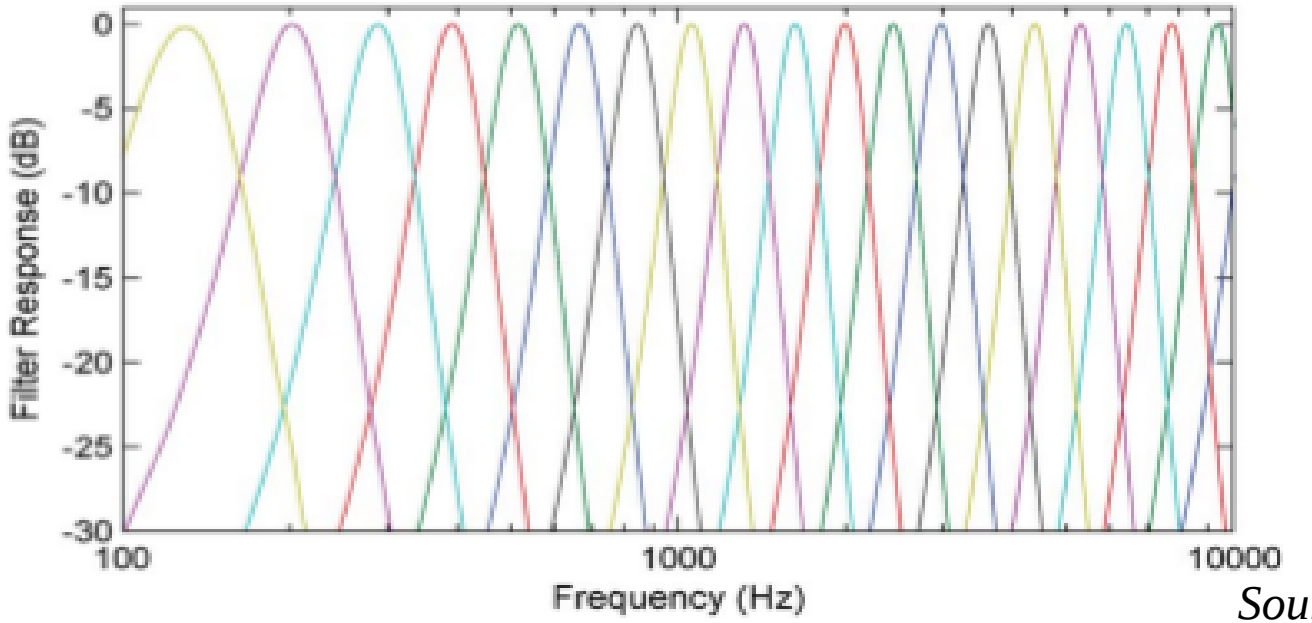
Source: Internet

Timbre

- “Quality” of sound – distinguishes one musical instrument from other
  - Determined by proportion of higher harmonics
- Pitch of a complex sound = fundamental frequency
- Pitch is perceived as the fundamental frequency even when it is missing.



Functional model of cochlea

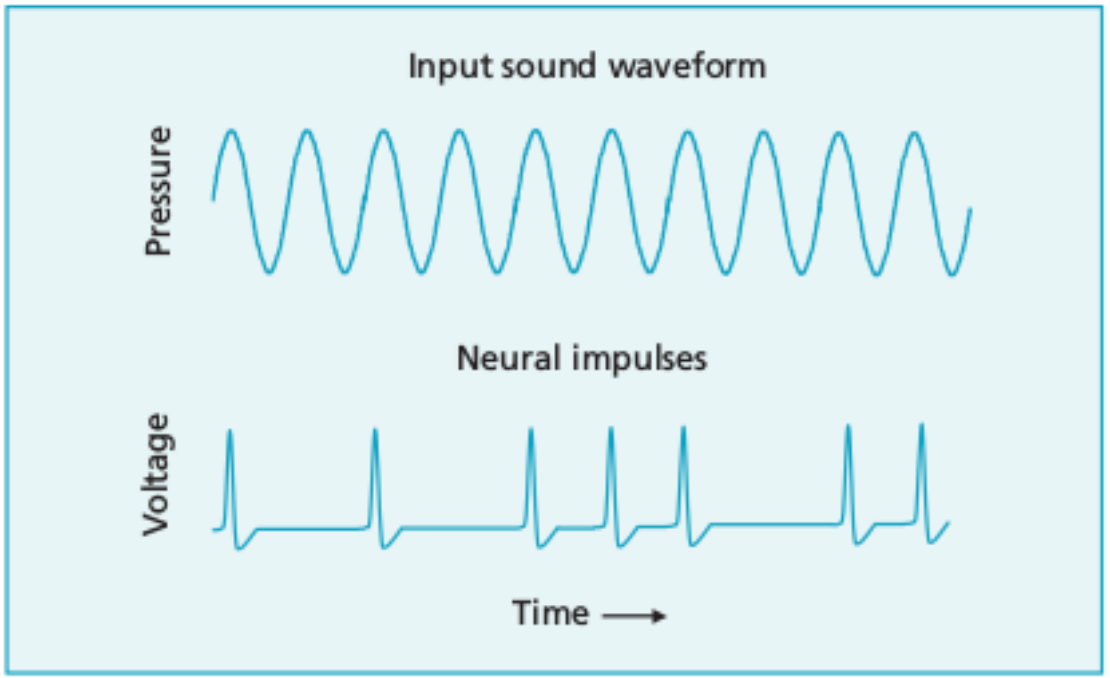


Note:
The Frequency axis is in logarithmic scale

Source: Brandenburg, et al., 2013

- A set of 24 filter banks located at different parts of the cochlea
- Each with a different frequency response
  - The center frequency follows a log scale
  - Bandwidth are narrow for low frequencies and wider at higher frequencies
- High-frequency resolution for low-frequency components
- Higher time resolution for high-frequency components

When do the neurons fire: Phase locking

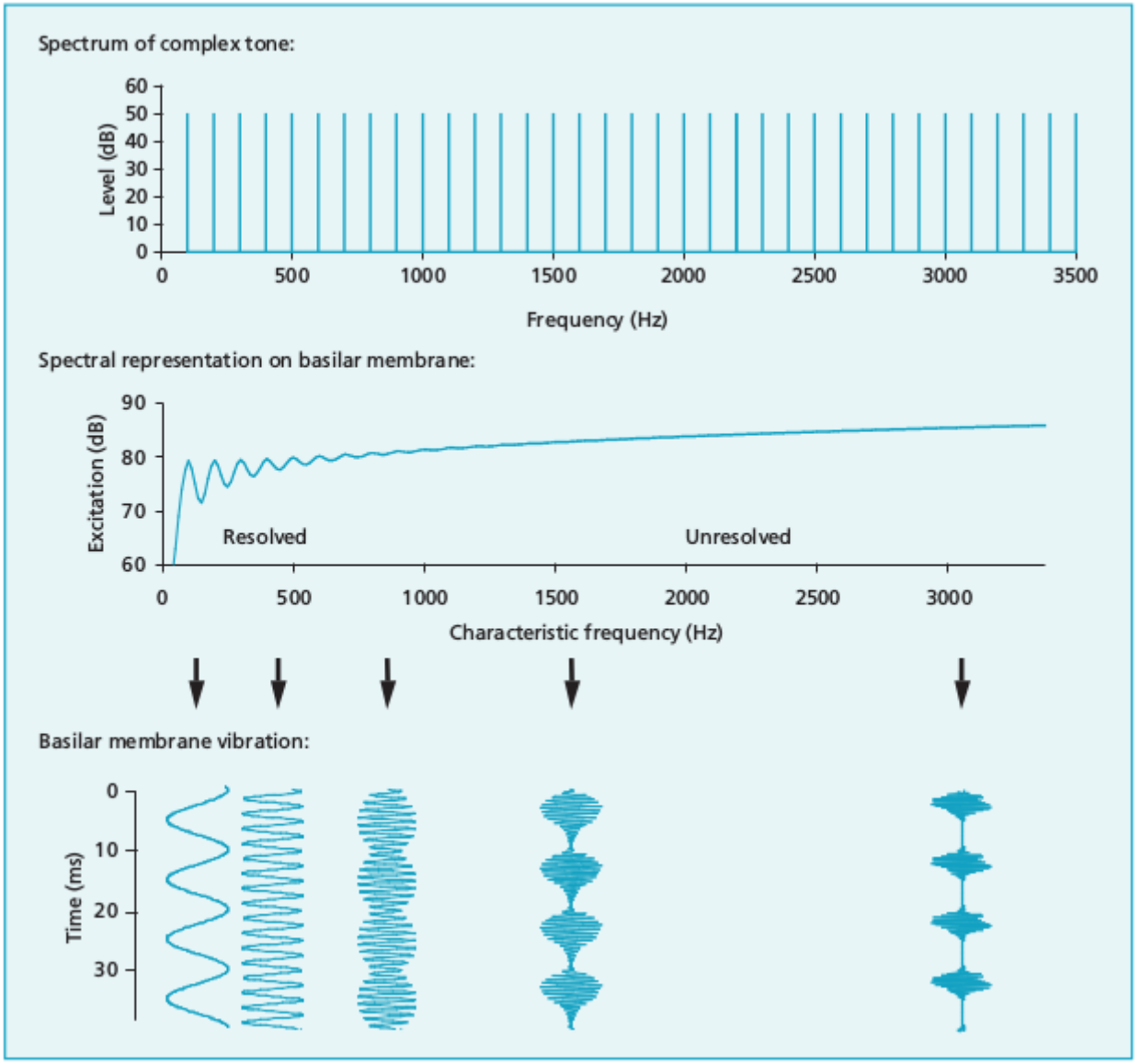


- Neurons fire when there is maximum movement of basilar membrane in one direction
- Pattern of firing is phase locked with the sound wave
- At freq > 200 Hz, neurons may not fire on every cycle
  - remain phase locked to the waveform

- At freq > 5000 Hz, neurons cannot remain phase locked to the waveform
  - phase locked to overall amplitude

- Conveys more information than the overall frequency components

Pitch perception for complex tones



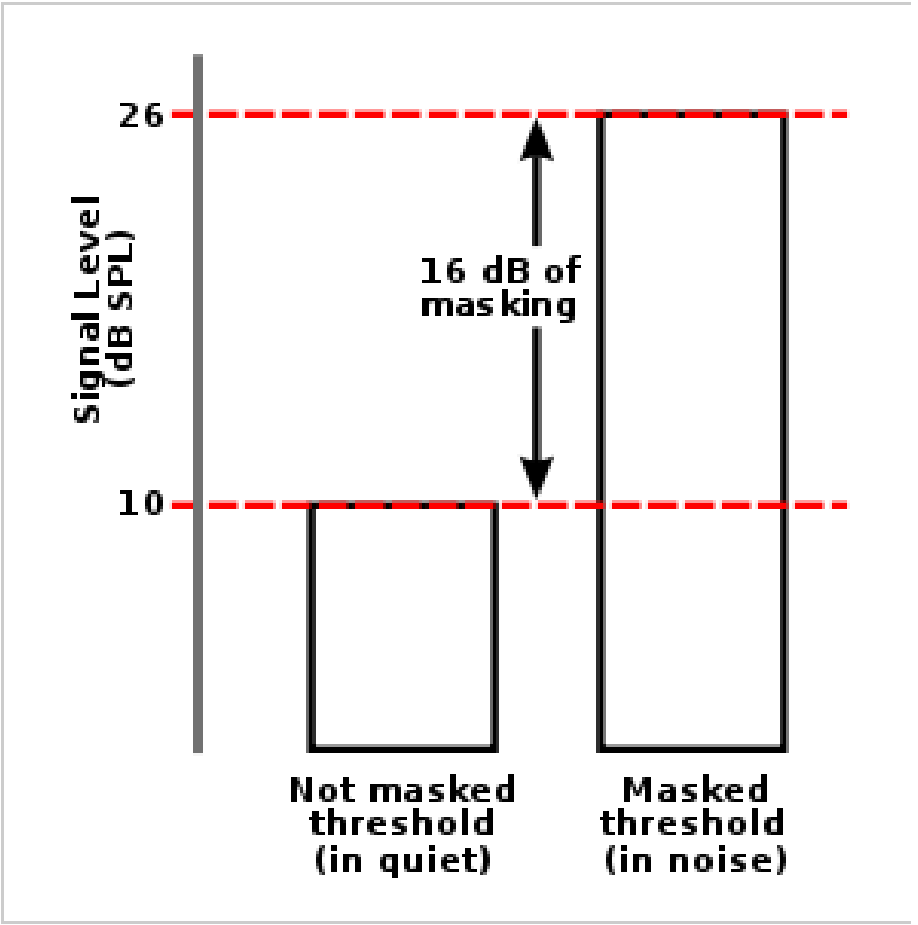
Pitch perception for complex tones ... more

- Pattern recognition theory (Goldstein, 1973; Terhardt, 1974)
  - The resolved harmonics form a pattern that is characteristic of any fundamental frequency.
  - If harmonics of 300, 400, and 500 Hz are present, the auditory system can deduce that the fundamental frequency is 100 Hz.
  - This mechanism requires that the harmonics are resolved, so that their frequencies can be independently determined.
- Temporal theory (Schouten, 1940, 1970)
  - Pitch may be derived directly from the repetition rate of the waveform produced by the interacting unresolved harmonics.
- *Does a combination work?*

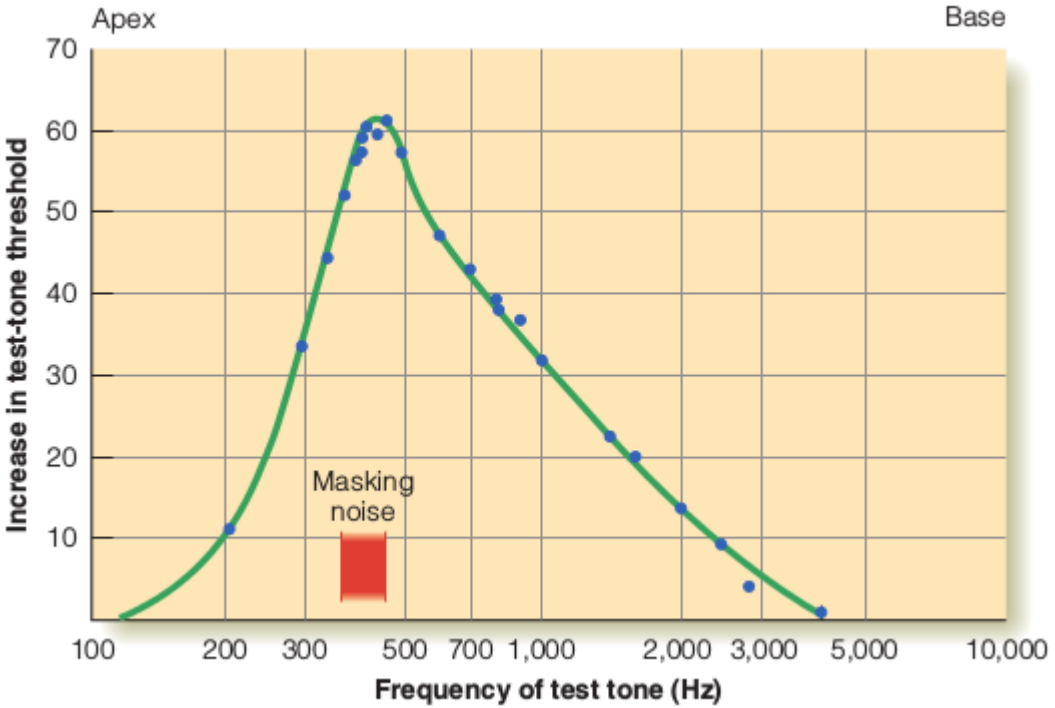
Auditory masking

*Inability to hear a sound in presence of other sounds_*

- Unmasked threshold
  - quietest level of the signal which can be perceived without a masking signal present
- Masked threshold
  - quietest level of the signal perceived when combined with a specific masking signal
- Amount of masking
  - the difference between the masked and unmasked thresholds
- Depends on
  - Frequency of signal being masked
  - Frequency of masking signal
  - Individual listener



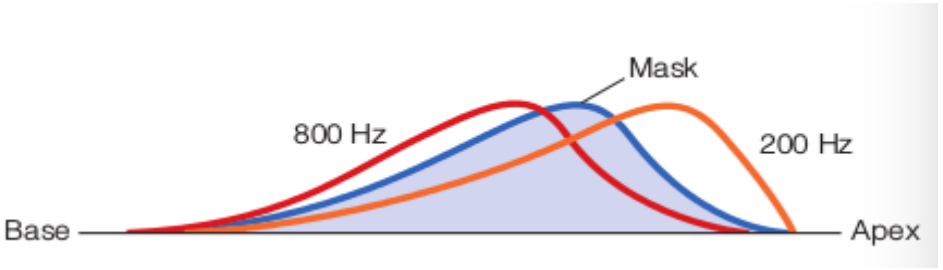
Source: Wikipedia



Sound of the same fequency is masked the most

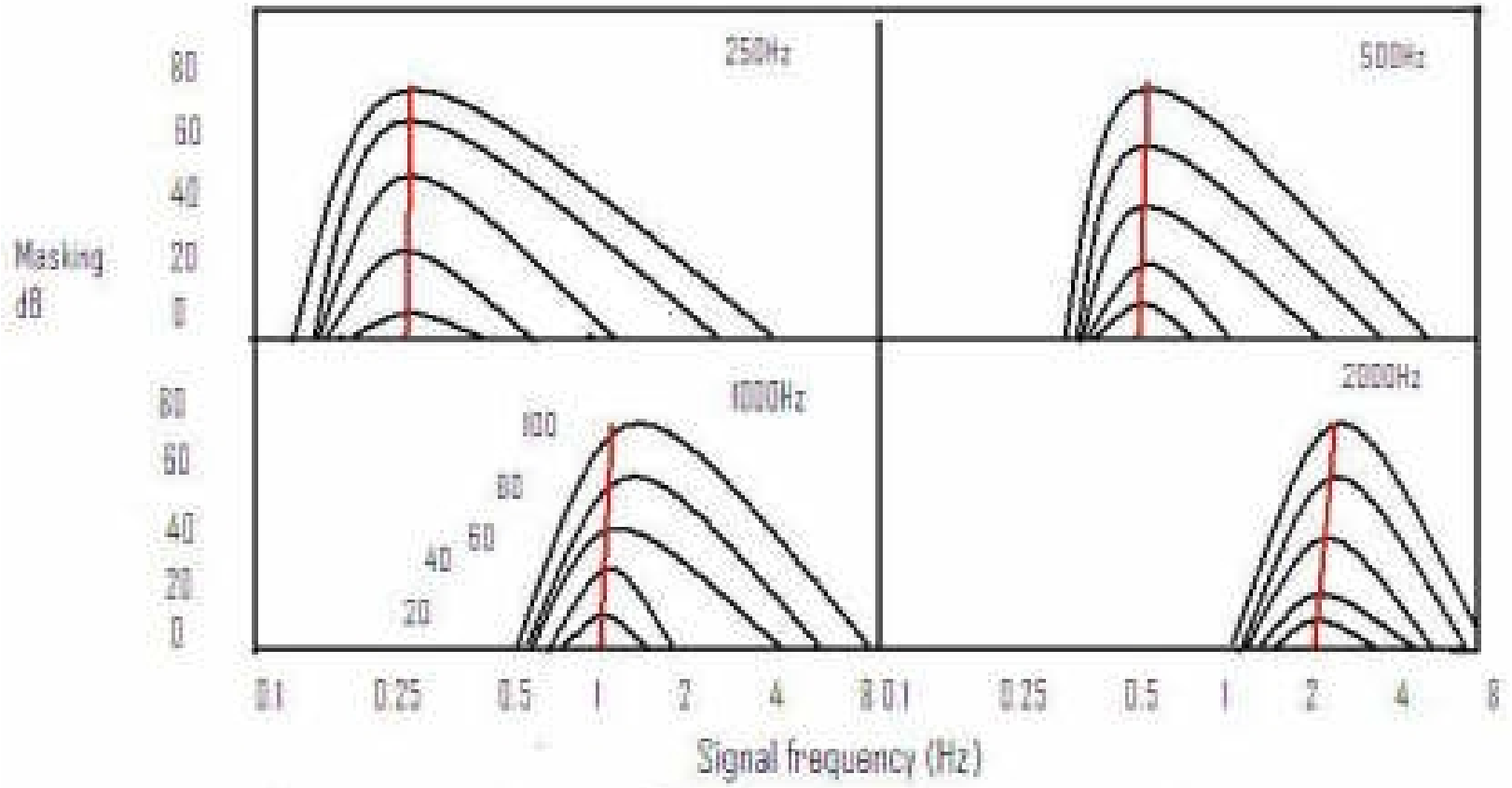
Higher frequencies are masked more than the lower ones

Mask 400 Hz



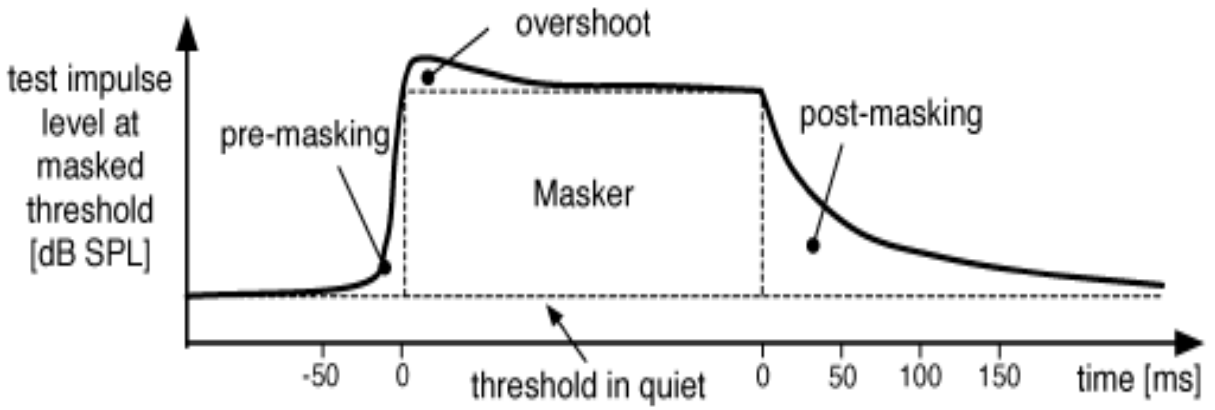
Basilar membrane vibration patterns

Masking audiograms



Source: Wikipedia

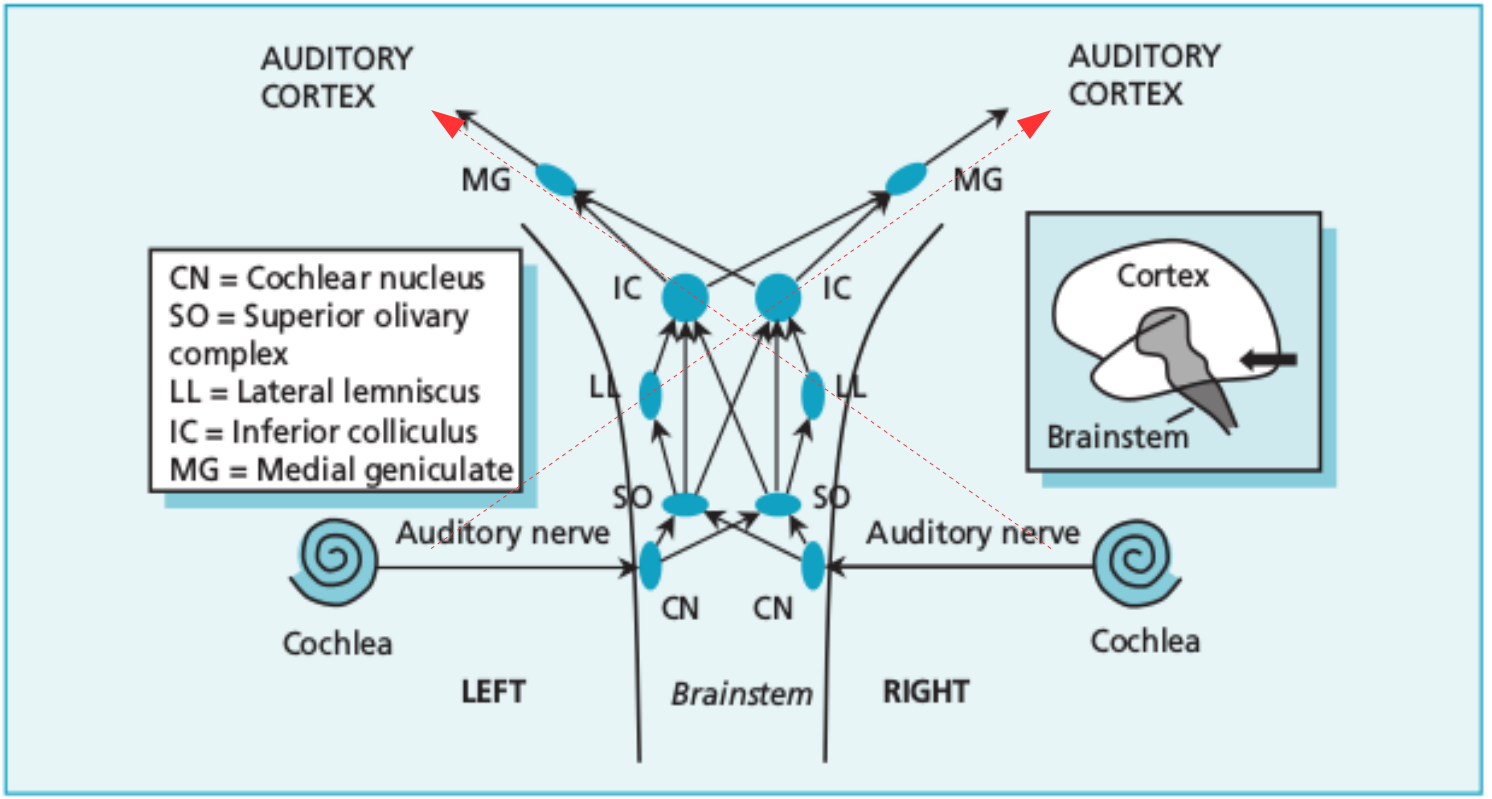
Temporal masking



Source: Brandenburg, et al., 2013

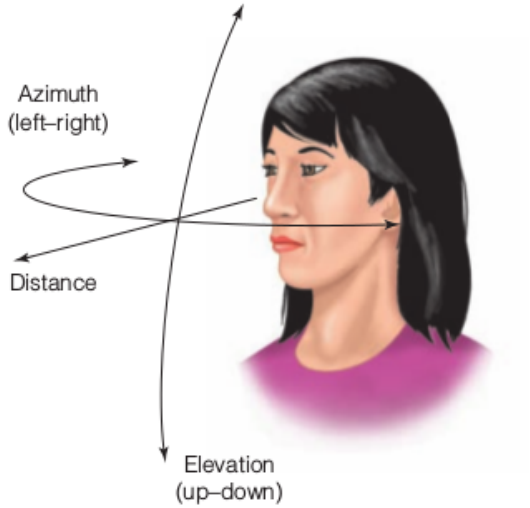
- **Temporal masking or non-simultaneous masking:** a sudden stimulus sound makes inaudible other sounds which are present immediately preceding or following the stimulus.
- **Backward masking / Pre-masking:** Masking which obscures a sound immediately preceding the masker
- **Forward masking / Post-masking:** masking which obscures a sound immediately following the masker

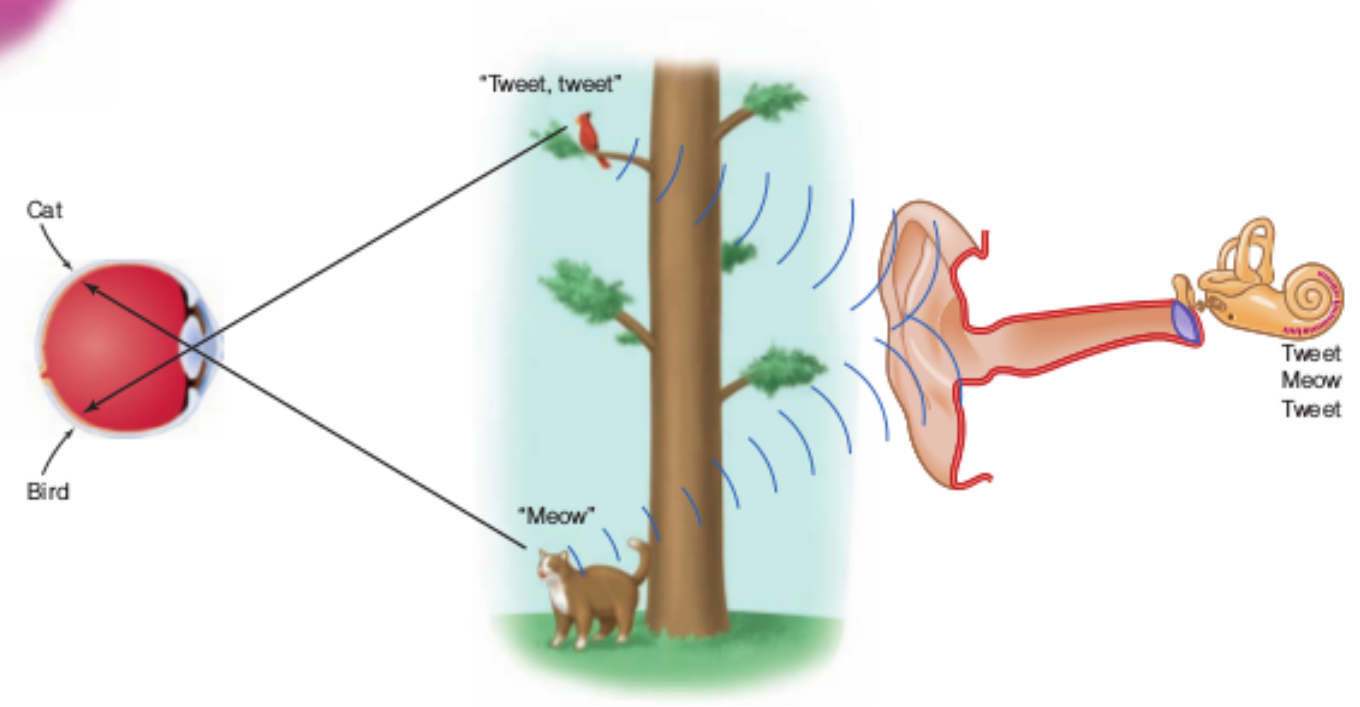
From cochlea to auditory cortex



Sound localization – auditory scene

Visual localisation *vs.* Sound localization

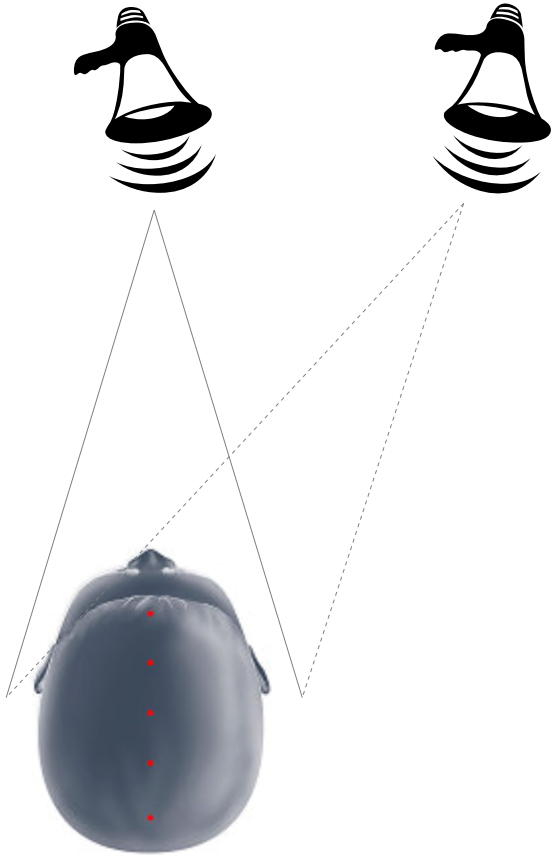




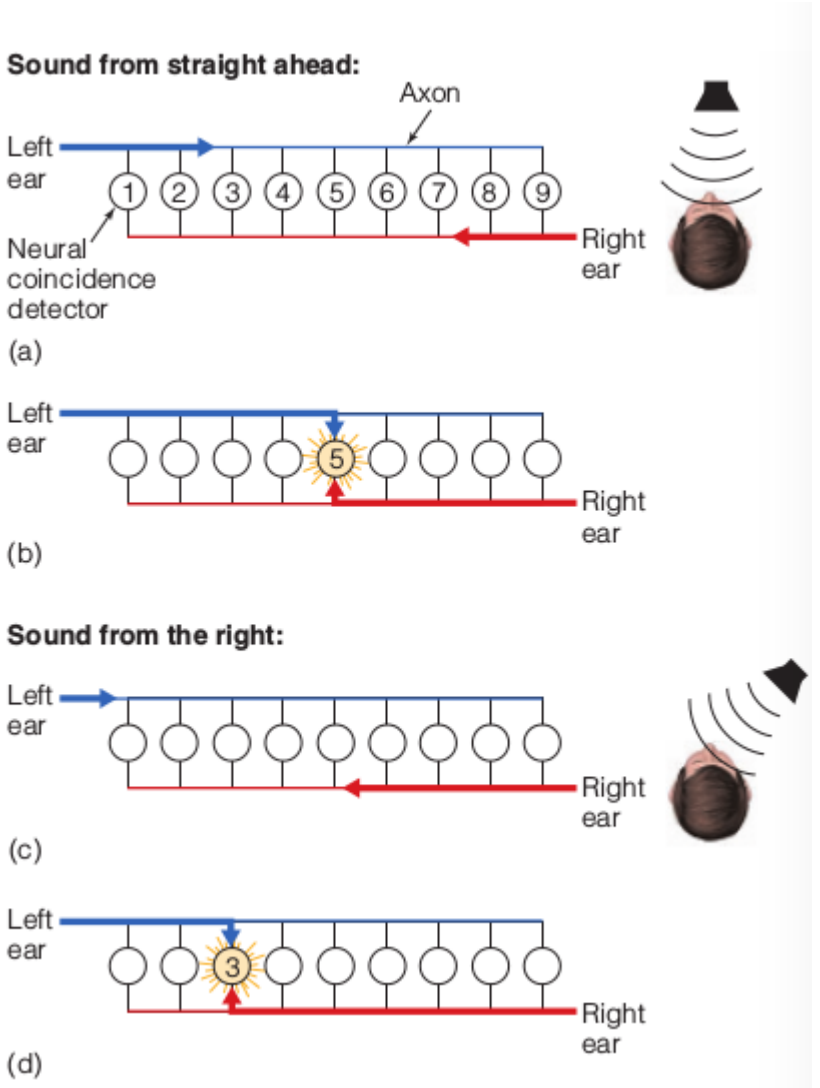
Cues for sound localization

- Binaural cues
  - Interaural Time Difference (ITD)
  - Interaural Level Difference (ILD)
  - Assymetric spectral reflection from body parts
  - Ratio of direct signal and reverberations (echoes)
- Monaural cues
  - Spectral cue
- Visual cues

Inter-aural Time Difference (ITD)

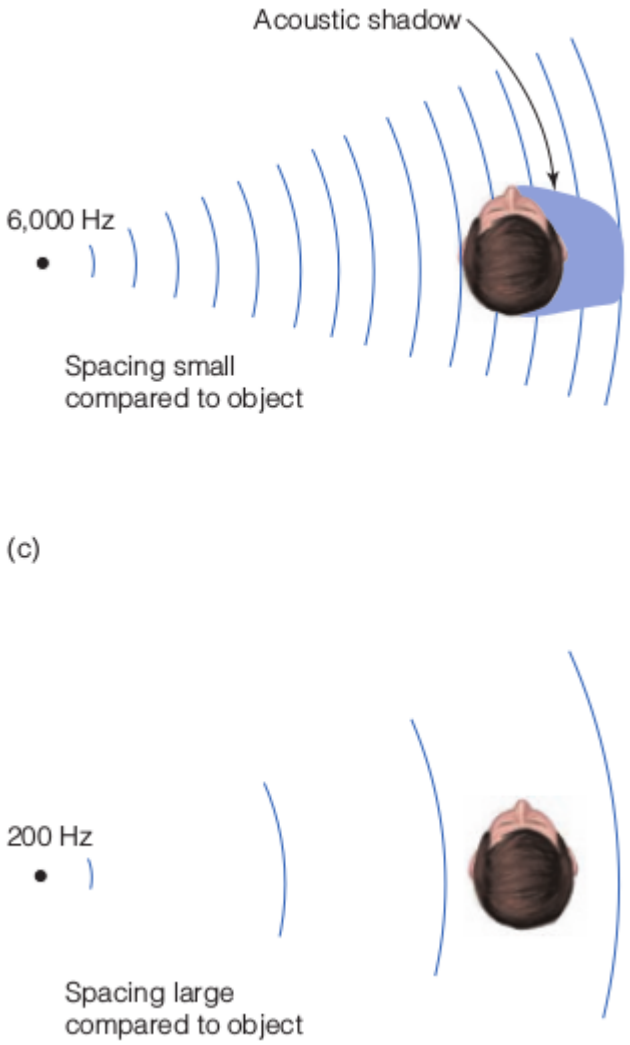


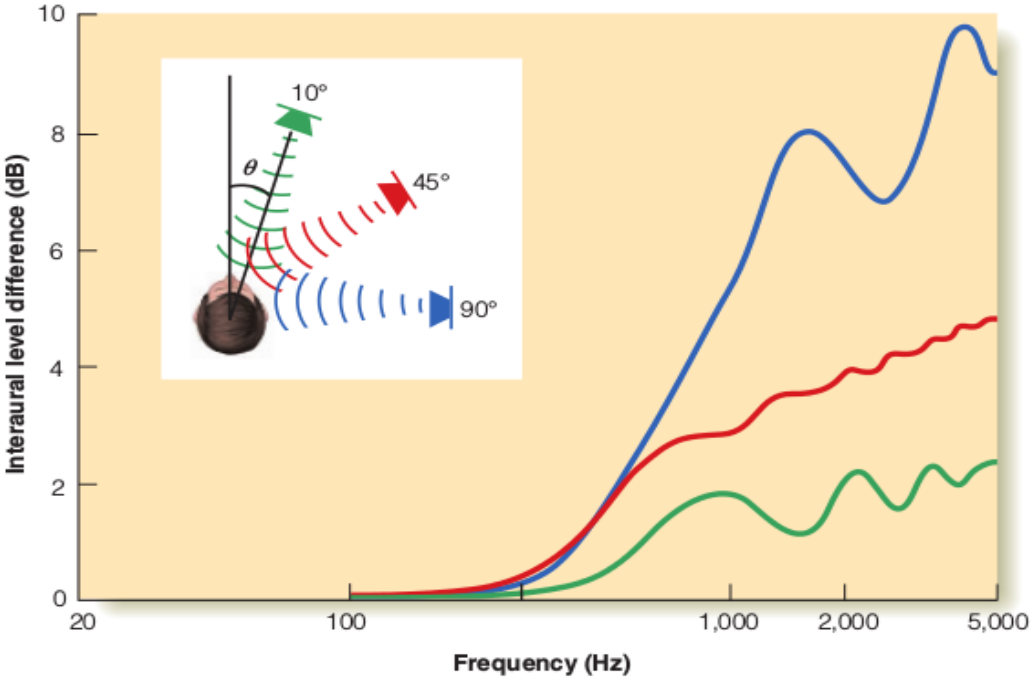
Effective cue for location of low-frequency sounds



Narrowly tuned ITD neurons
- Jeffres (1948)

Inter-aural Level Difference (ILD)

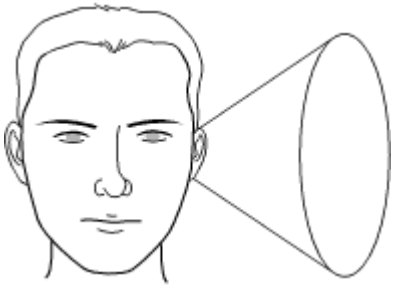




Effective cue for location of high-frequency sounds

Binaural cues - summary

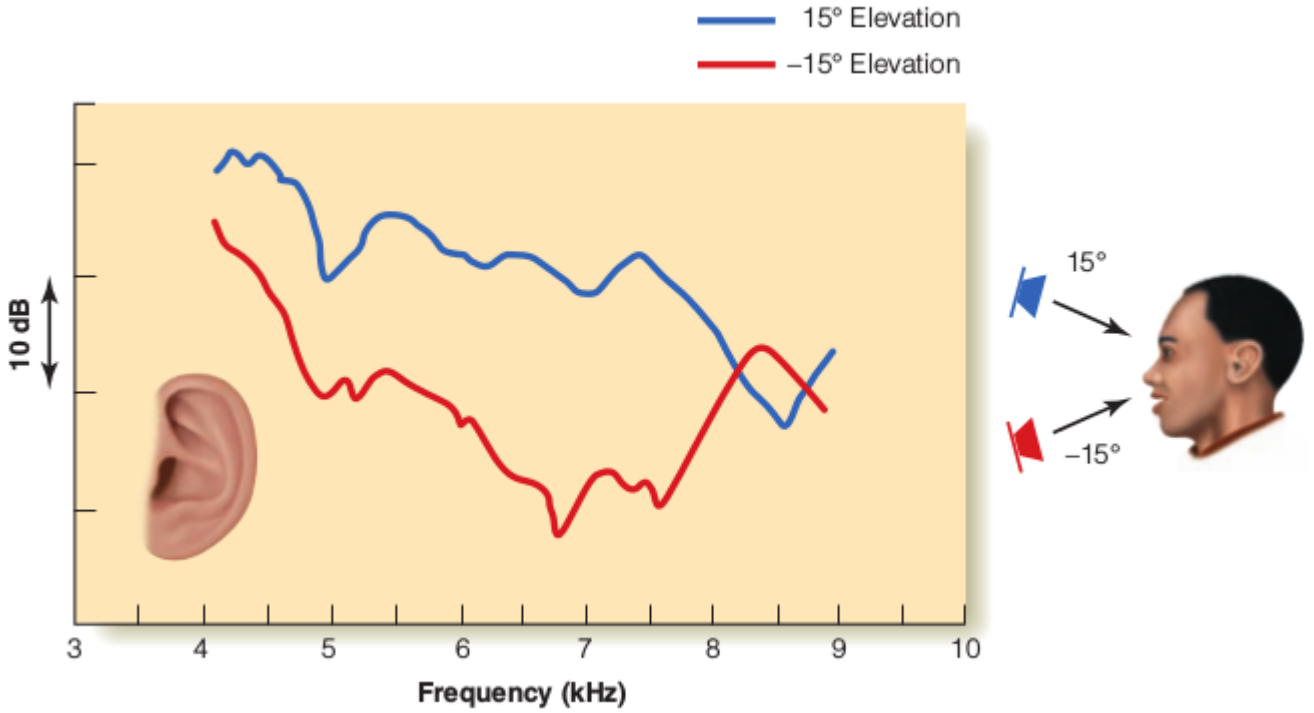
- ITD and ILD together is good for azimuthal discrimination
- Does not provide elevation information
- Cone of confusion
  - All point on periphery of the cone have same ITD and ILD
  - Cannot be distinguished



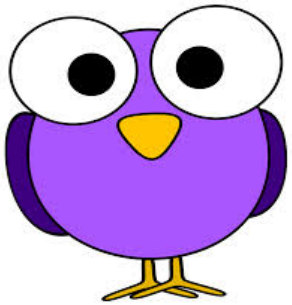
Cone of confusion

Monaural cues:
Spectral cue

- Reflections on the pinnae
- Differences in the spectrum of frequencies that reach the ear from different locations
- Provides elevation information



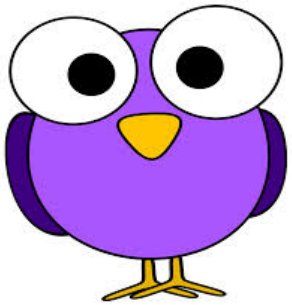
Visual cue



Tweet

Meow

What if there are contradictory cues ?



Meow

Tweet

Ventriloquism

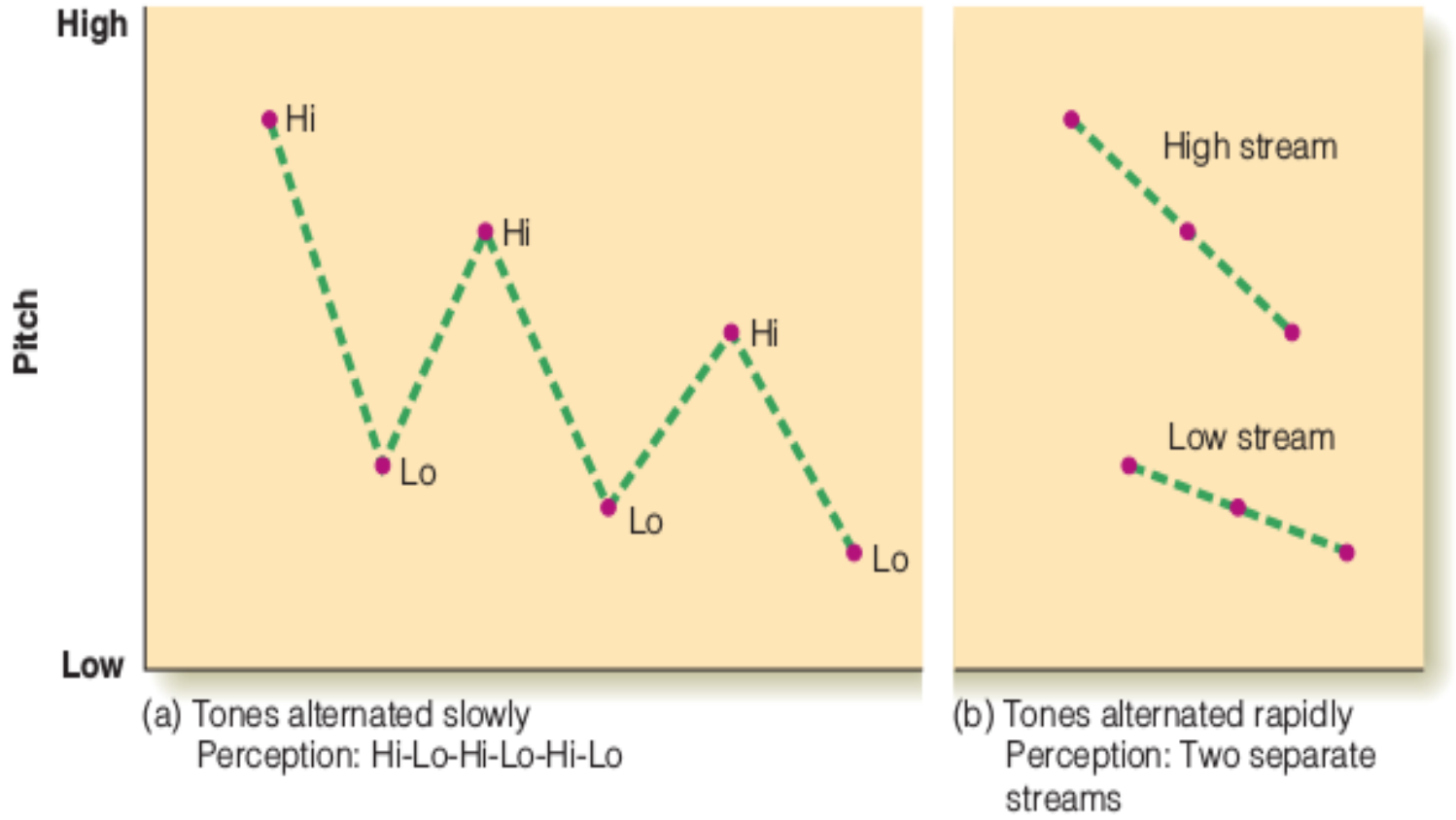
Perceptual grouping:
Auditory scene analysis

- *How do we converse in a noisy room, e.g. in a party?*
- *How do we distinguish musical instruments (and vocal) in a mono player?*

Perceptual grouping:
Auditory scene analysis

- Location
- Similarity of timre and pitch
- Proximity in time
- Auditory continuity
  - Similar to Gestalt principle of visual continuity

Auditory stream seggregation



(a) Tones alternated slowly
Perception: Hi-Lo-Hi-Lo-Hi-Lo

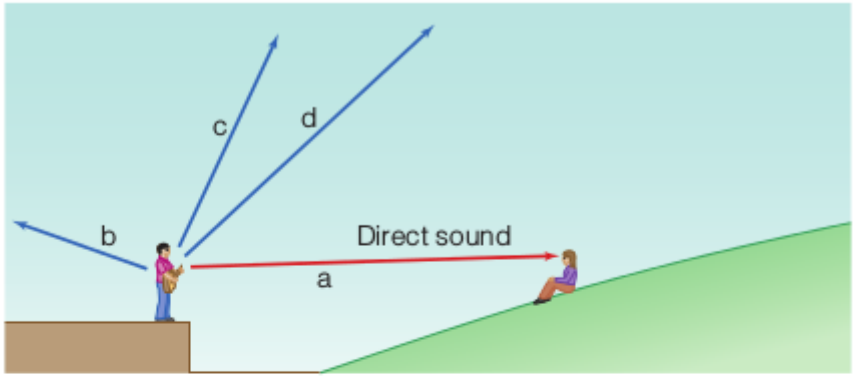
(b) Tones alternated rapidly
Perception: Two separate streams

Some perceptually motivated features

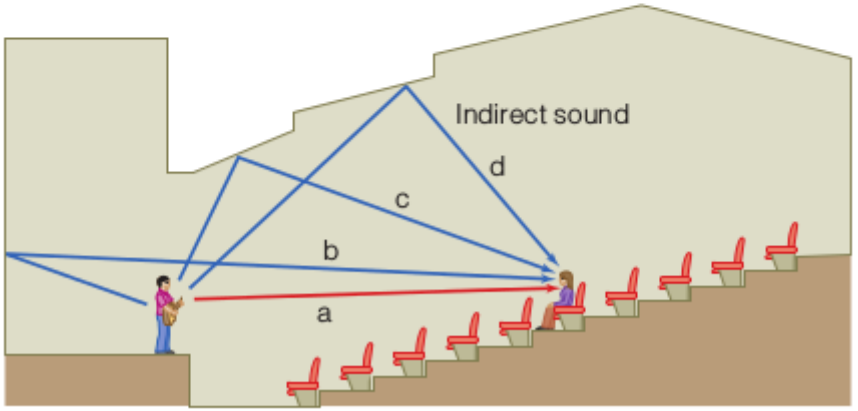
Features	Description
Loudness	Is the subjective impression of the intensity of a sound.
Spectral Centroid	Spectral centroid is the weighted mean of the magnitude frequency spectrum; it is commonly described as one of the main dimensions of timbre perception.
Sharpness	Can be interpreted as a perceptual spectral centroid.
Perceptual spread	Is a measure of the timbral width of a given sound.
Signal to Mask ratio	Is the difference between the signal intensity and the intensity of the signal perceptual mask.
Local energy in Bark scale	Represents the relative importance of local energy distribution in Bark bands.
Spectral Flux	Is the spectral magnitude Euclidean distance between neighboring audio frames.
Sub-band flux	Represents the fluctuation of frequency content in ten octave-scaled bands.
High energy / low Energy	Represents the ratio of energy above and below a given frequency.
Roughness	Is a basic psychoacoustical sensation for rapid amplitude variations.
Relative entropy	Provides an estimate of the whiteness of a signal.
MFCC	Mel-Frequency Cepstral coefficients; Estimate the spectral envelope using (limited) perception principles.
Cortical Representations	Multiscale or Multi linear representations that model various spectro-temporal properties in the central auditory system.

Source: Richard, et al. 2013

Indoor hearing



(a)



(b)

<u>If delay > 50 ms</u>

Two distinct sounds are perceived (from different locations) -- echo

<u>If delay < 50 ms</u>

No localization effect
The first received sound takes precedence

References

- Leon Gunther. The Physics of music and color (e-book)
- Goldstein. Sensation and perception (e-book)
- Plack. Auditory perception http://socialscientist.us/nphs/psychIB/psychpdfs/PIP_Auditory_Perception.pdf