Lecture-10

Speech Perception

Speech Perception

- Understanding ***how we hear sounds and how we perceive speech*** □ ***better design and*** **implementation** of robust and efficient systems for analyzing and representing speech

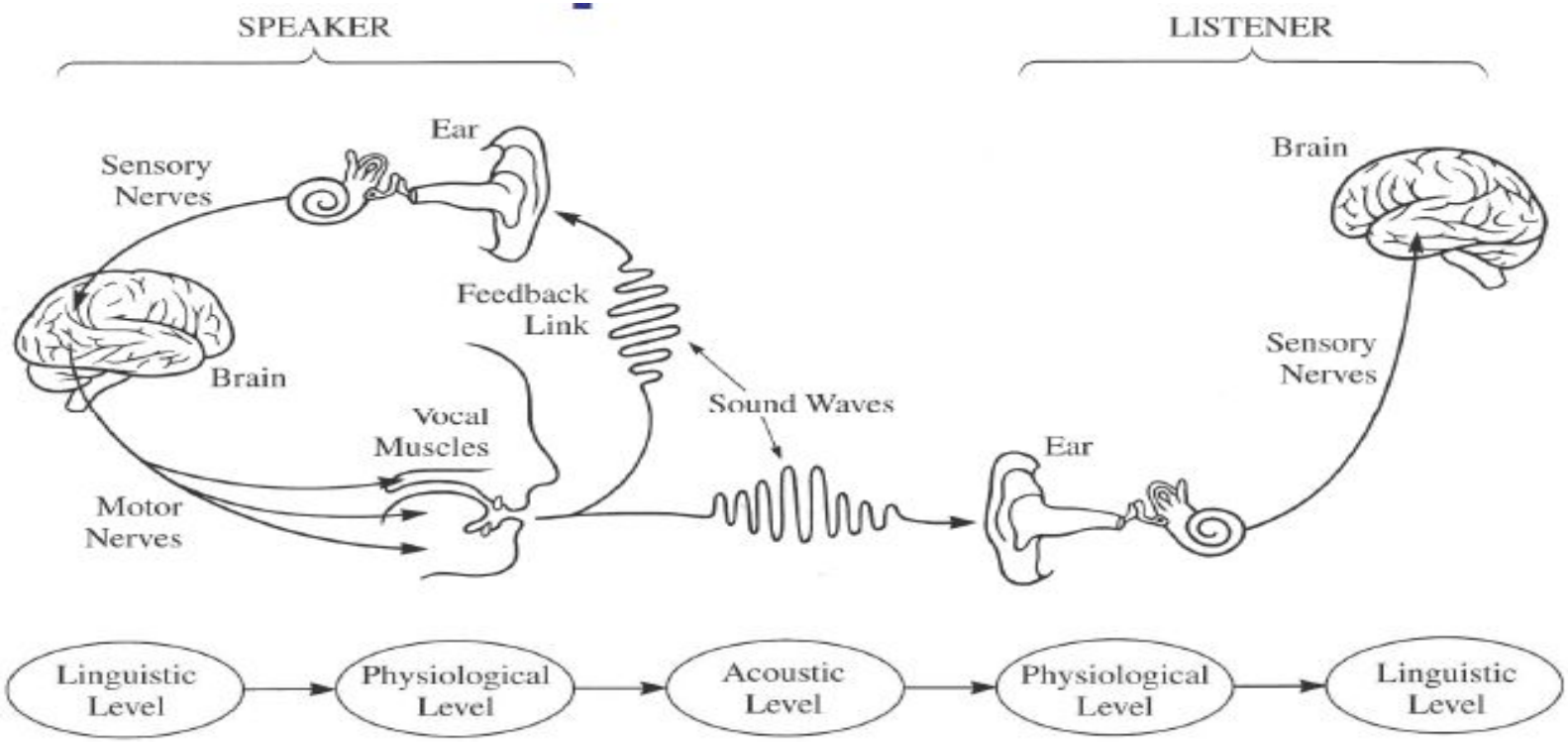
  the better we understand signal processing in the human auditory system, the better we can (at least in theory) design practical speech processing systems like:

  – speech coding

  – speech recognition

- Try to understand speech perception by looking at the ***physiological models of hearing***

The Speech Chain



The Speech Chain comprises the processes of:

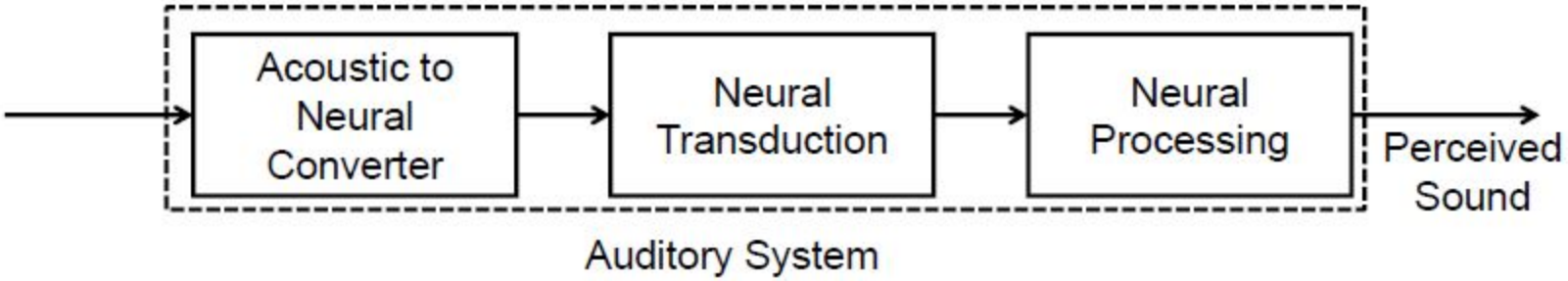
- ❑ Speech production,
- ❑ Auditory feedback to the speaker,
- ❑ Speech transmission (through air or over an electronic
- ❑ Communication system (to the listener), and
- ❑ Speech perception and understanding by the listener.

The Speech Chain

The message to be conveyed by speech goes through five levels of representation between the speaker and the listener

I. the linguistic level (where the basic sounds of the communication are chosen to express some thought of idea)
II. the physiological level (where the vocal tract components produce the sounds associated with the linguistic units of the utterance)
III. the acoustic level (where sound is released from the lips and nostrils and transmitted to both the speaker (sound feedback) and to the listener
IV. the physiological level (where the sound is analyzed by the ear and the auditory nerves), and finally
V. the linguistic level (where the speech is perceived as a sequence of linguistic units and understood in terms of the ideas being communicated)

The Auditory System

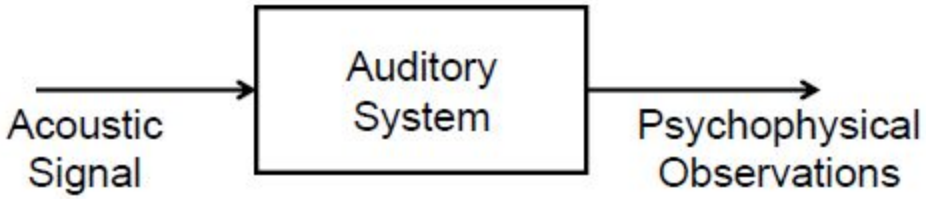


- The acoustic signal first converted to a neural representation by processing in the ear
  - The conversion takes place in stages at the outer, middle and inner ear
  - These processes can be measured and quantified
- The neural transduction step takes place between the output of the inner ear and the neural pathways to the brain
  - Consists of a statistical process of nerve firings at the hair cells of the inner ear, which are transmitted along the auditory nerve to the brain
  - Much remains to be learned about this process
- The nerve firing signals along the auditory nerve are processed by the brain to create the perceived sound corresponding to the spoken utterance
  - **These processes not yet understood**

The Black Box Model of the Auditory System

Researchers have resorted to a "black box" behavioral model of hearing and perception

I. Model assumes that an acoustic signal enters the auditory system causing behavior that we record as psychophysical observations
II. Psychophysical methods and sound perception experiments determine how the brain processes signals with different loudness levels, different spectral characteristics, and different temporal properties
III. Characteristics of the physical sound are varied in a systematic manner and the psychophysical observations of the human listener are recorded and correlated with the physical attributes of the incoming sound
IV. Then determine how various attributes of sound (or speech) are processed by the auditory system



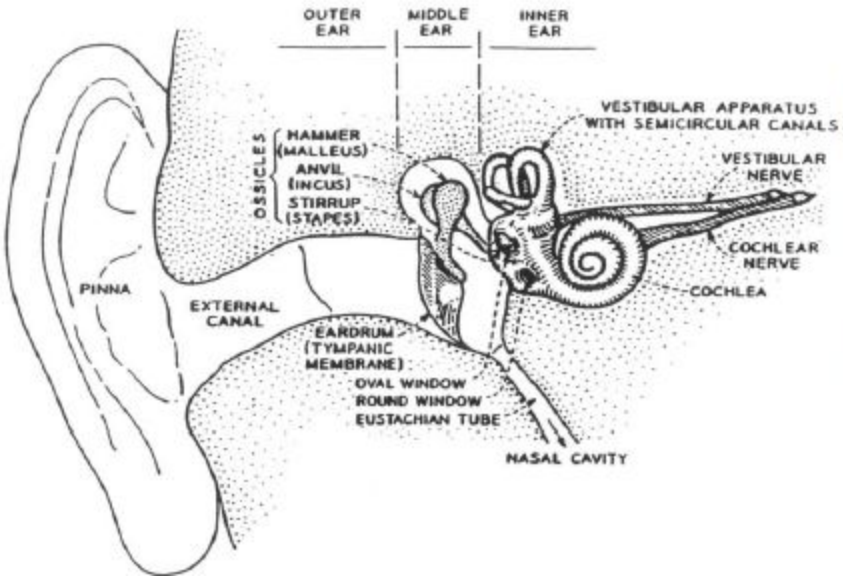
Why Do We Have Two Ears

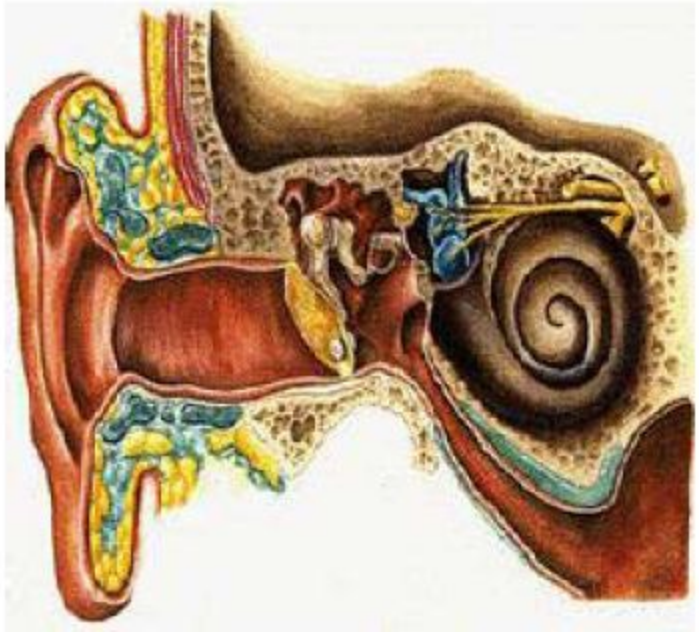
- ***Sound localization – spatially locate*** sound sources in 3-dimensional sound fields

- ***Sound cancellation – focus attention on*** a 'selected' sound source in an array of sound sources – 'cocktail party effect'

- Effect of ***listening over headphones*** **=>** localize sounds inside the head (rather than spatially outside the head)

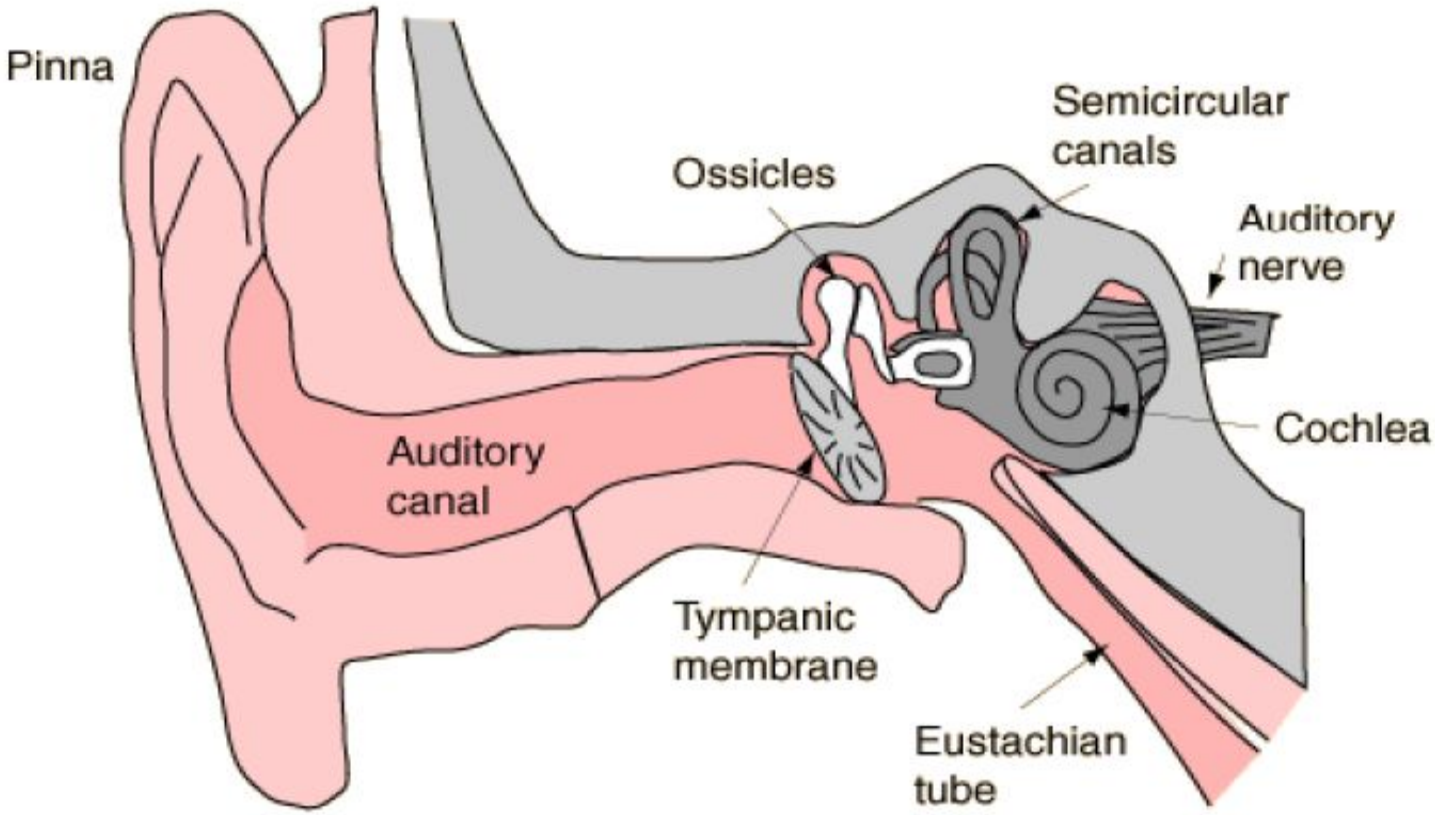
The Human Ear





Outer ear: pinna and external canal
Middle ear: tympanic membrane or eardrum
Inner ear: cochlea, neural connections

Ear and Hearing



Human Ear

Outer ear: funnels sound into ear canal

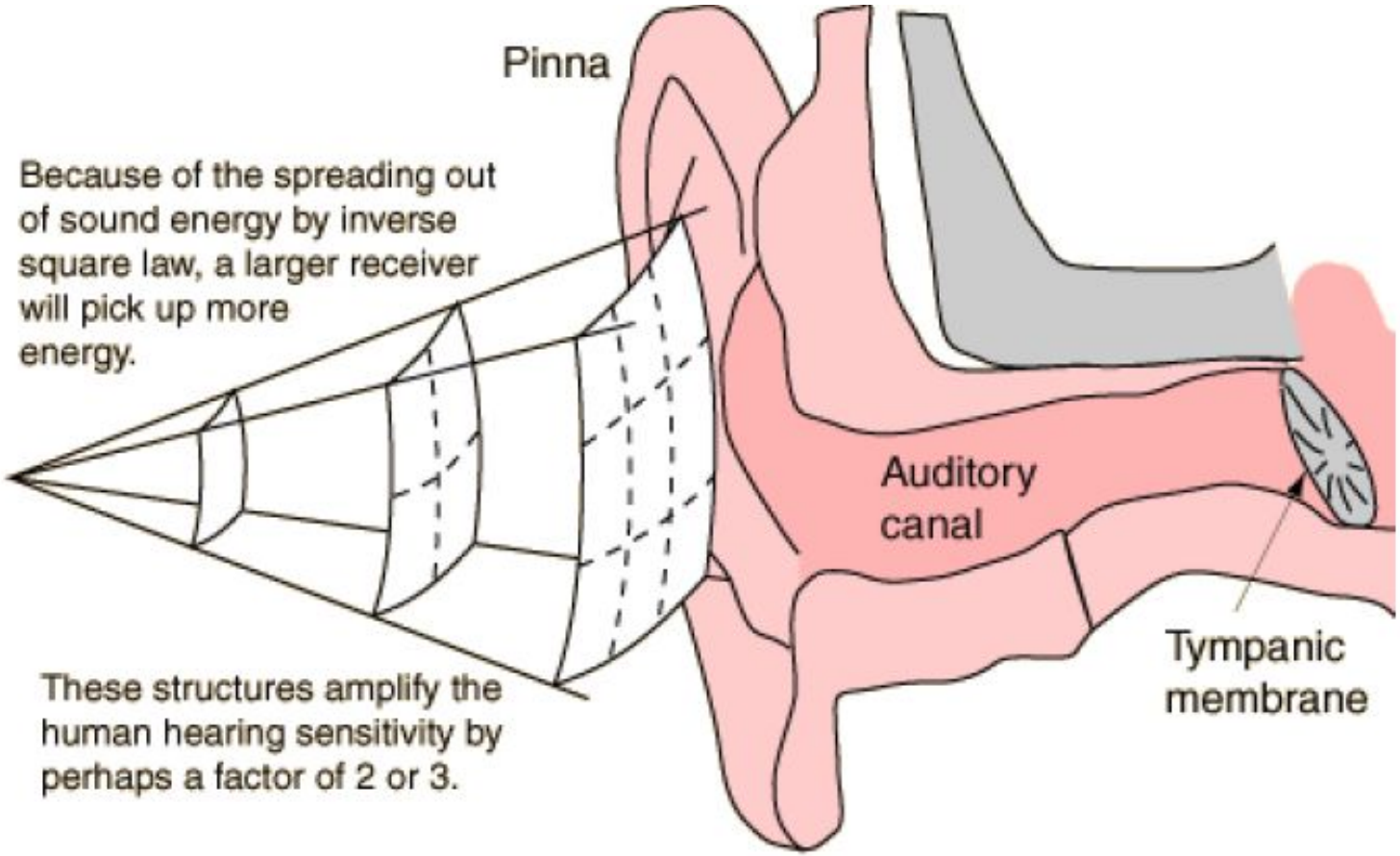
Middle ear: sound impinges on tympanic membrane; this causes motion

- Middle ear is a mechanical transducer, consisting of the hammer, anvil and stirrup; it converts acoustical sound wave to mechanical vibrations along the inner ear

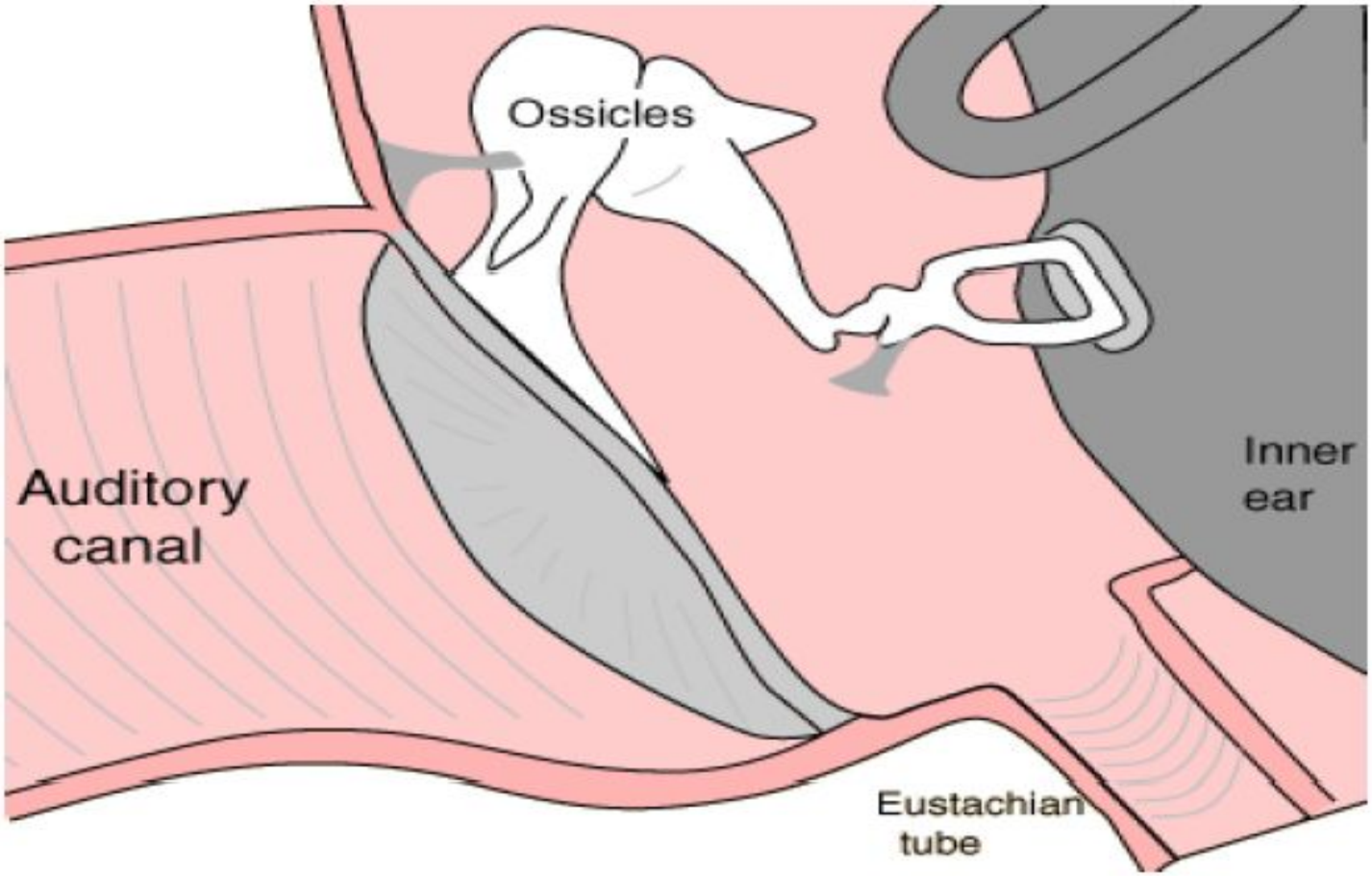
Inner ear: the cochlea is a fluid-filled chamber partitioned by the basilar membrane

- The auditory nerve is connected to the basilar membrane via inner hair cells
- Mechanical vibrations at the entrance to the cochlea create standing waves (of fluid inside the cochlea) causing basilar membrane to vibrate at frequencies commensurate with the input acoustic wave frequencies (formants) and at a place along the basilar membrane that is associated with these frequencies

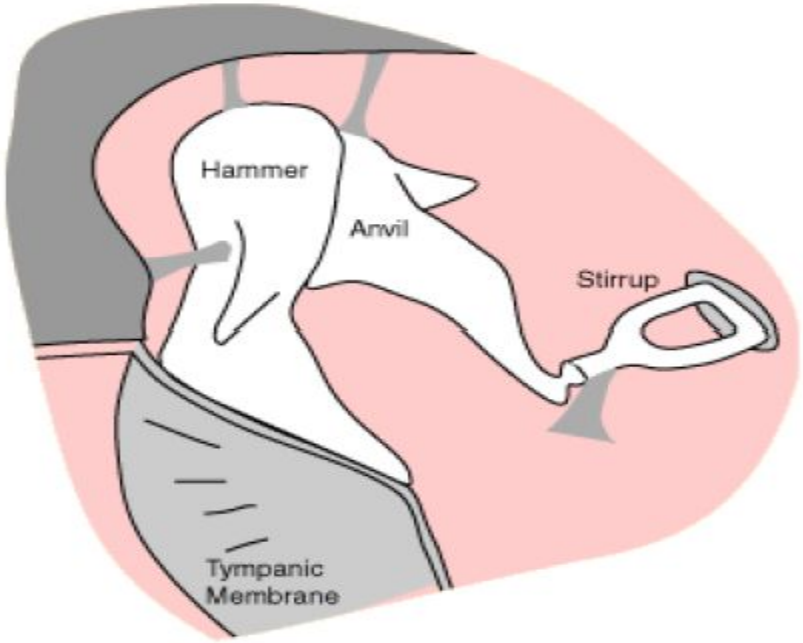
The Outer Ear



The Outer Ear



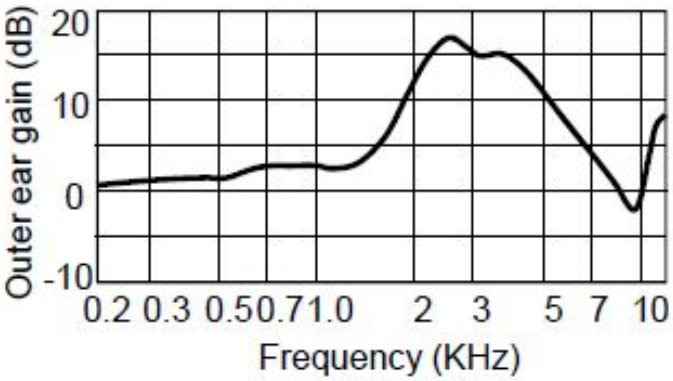
The Middle Ear

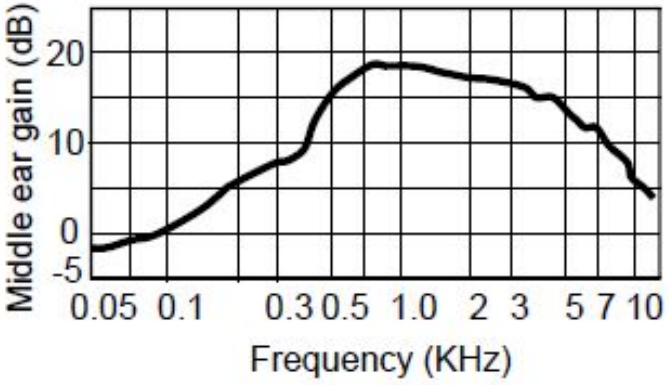


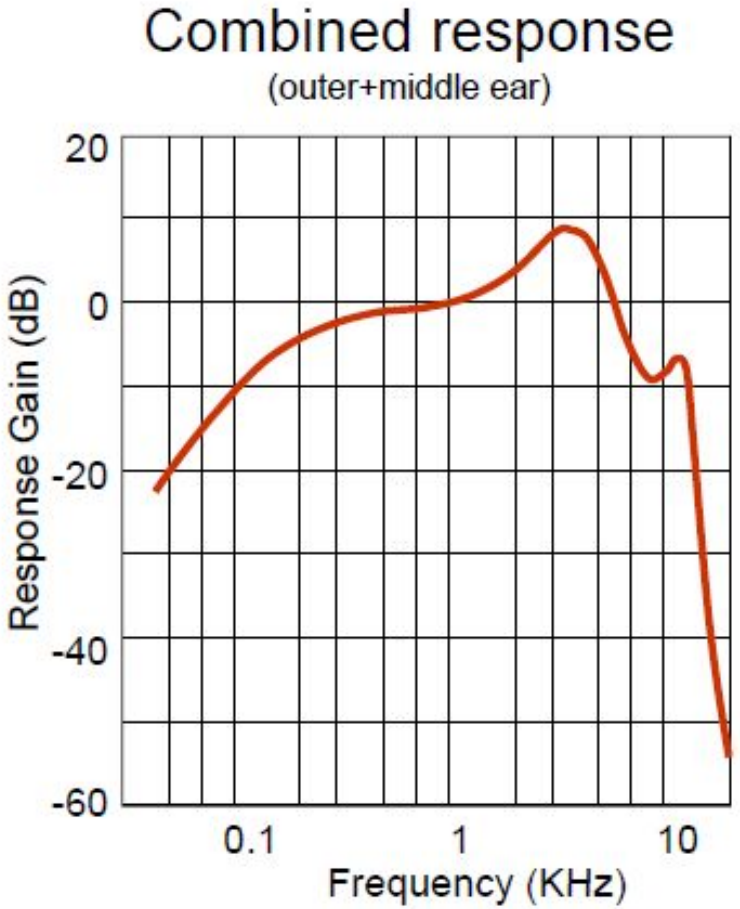
The Hammer (Malleus), Anvil (Incus) and Stirrup (Stapes) are the three tiniest bones in the body. Together they form the coupling between the vibration of the eardrum and the forces exerted on the oval window of the inner ear.

These bones can be thought of as a compound lever which achieves a multiplication of force—by a factor of about three under optimum conditions. (They also protect the ear against loud sounds by attenuating the sound.)

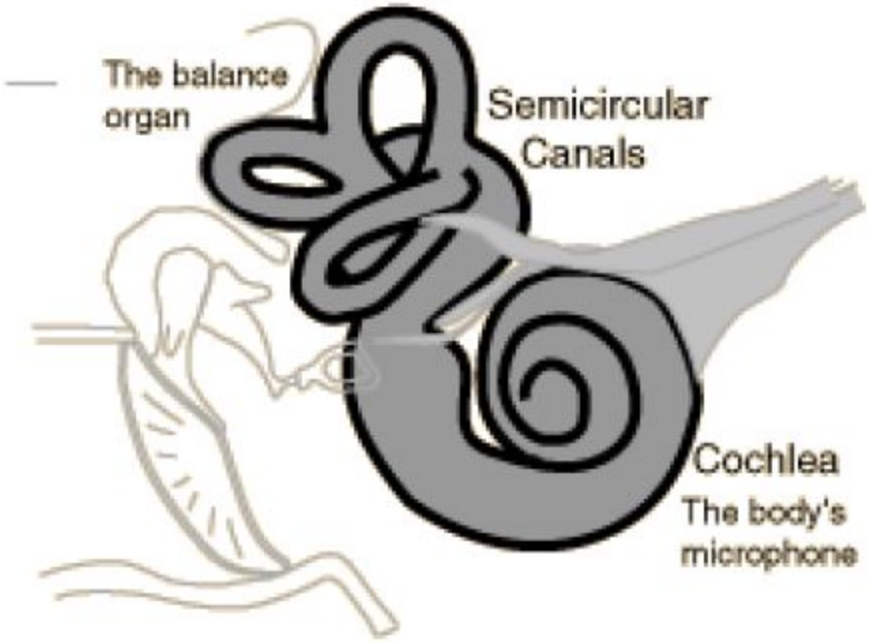
Transfer Functions at the Periphery







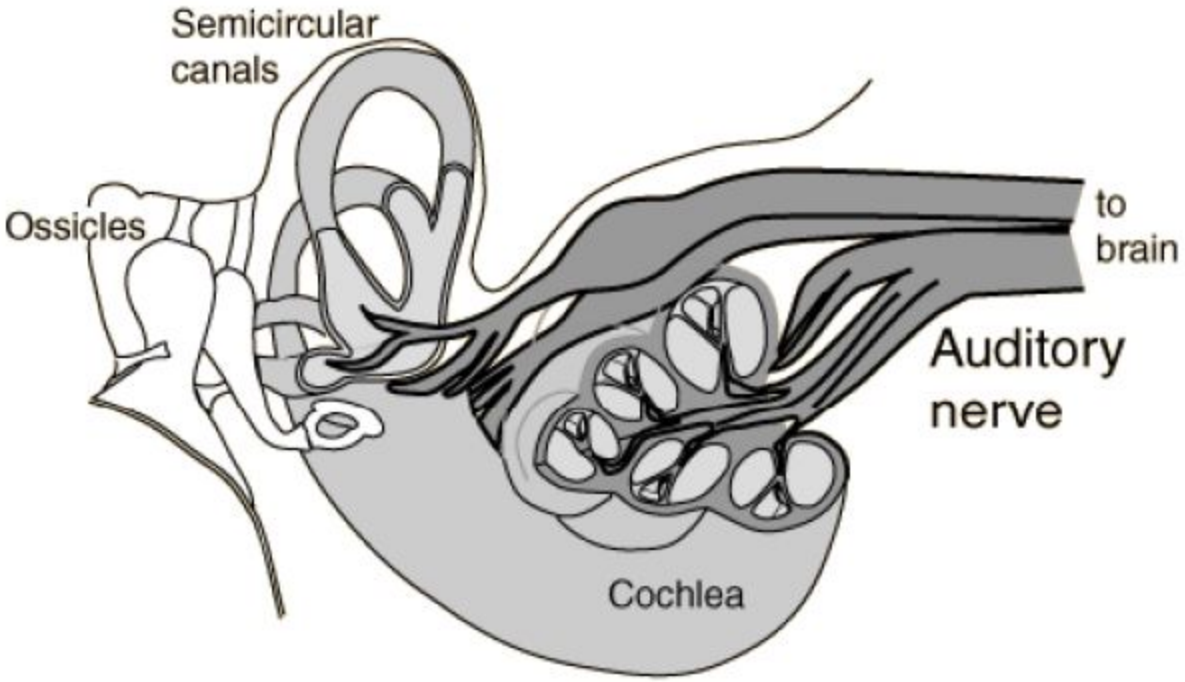
The Inner Ear



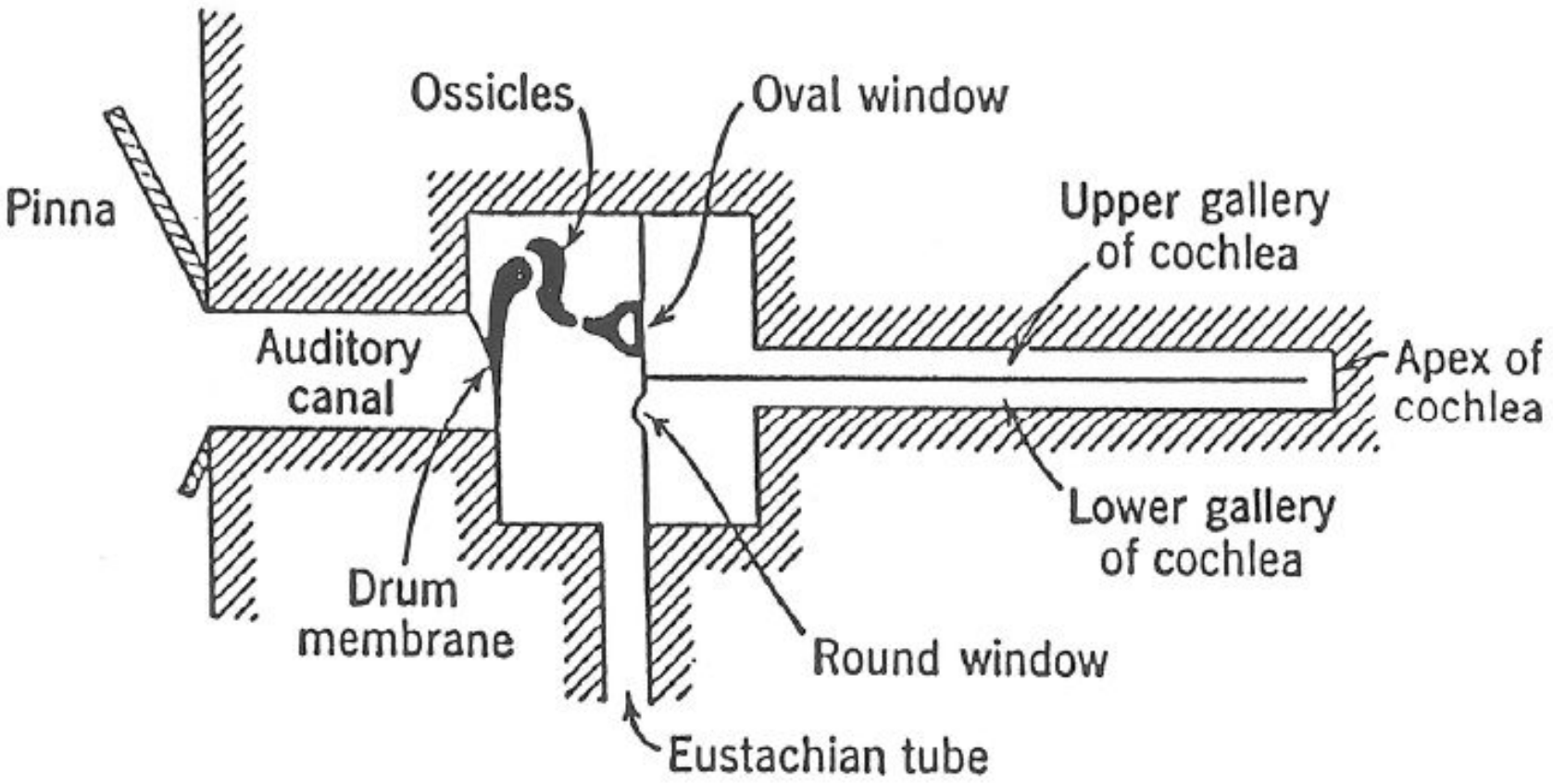
The inner ear can be thought of as two organs, namely the **semicircular canals** which serve as the body's balance organ and the **cochlea** which serves as the body's microphone, converting sound pressure signals from the outer ear into electrical impulses which are passed on to the brain via the auditory nerve.

The Auditory Nerve

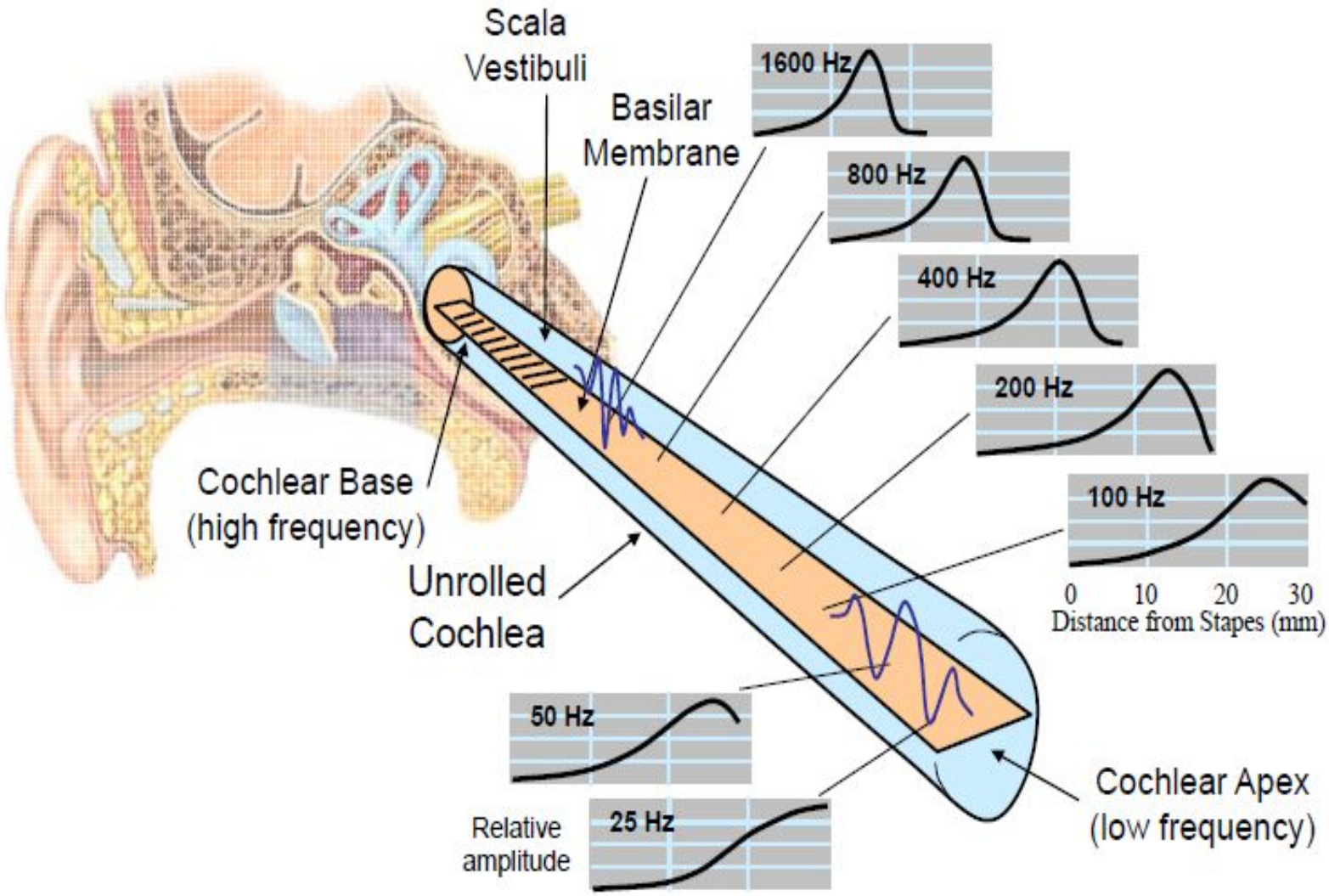
Taking electrical impulses from the cochlea and the semicircular canals, the auditory nerve makes connections with both auditory areas of the brain.



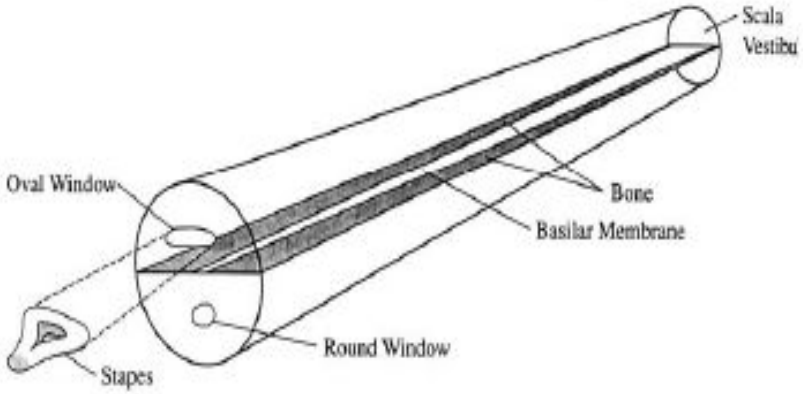
Schematic Representation of the Ear



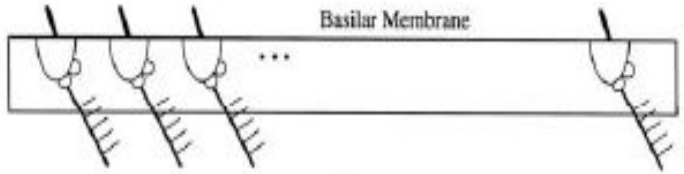
Stretched Cochlea & Basilar Membrane



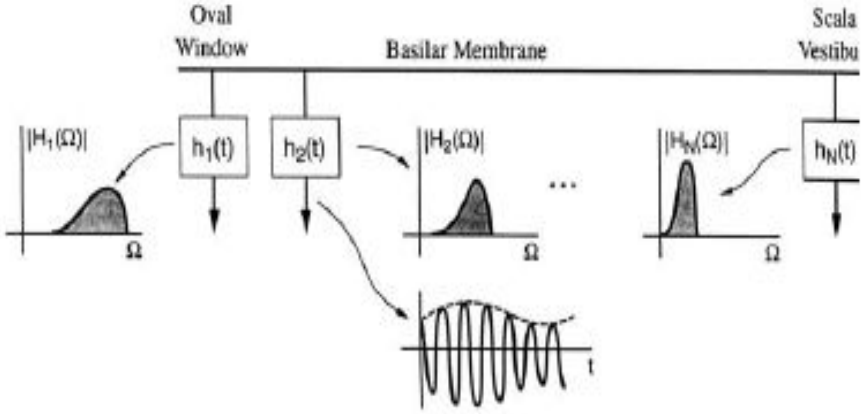
Basilar Membrane Mechanics



(a)

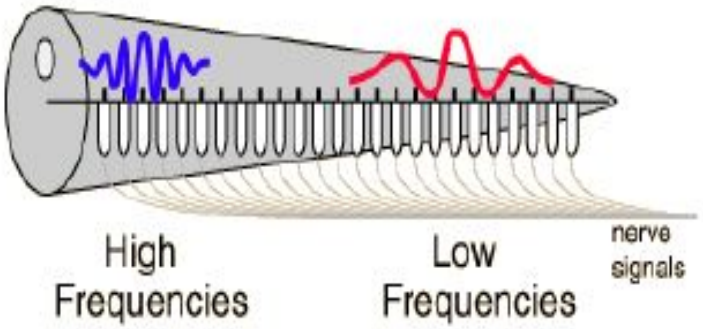


(b)



(c)

Figure 8.25 Schematic of front-end auditory processing and its model as a wavelet transform: (a) uncoiled cochlea; (b) the transduction to neural firings of the deflection of hairs that protrude from the in hair cells along the basilar membrane; (c) a signal processing abstraction of the cochlear filters along basilar membrane. The filter tuning curves, i.e., frequency responses, are roughly constant-Q with bandwi decreasing logarithmically from the oval window to the scala vestibuli.



Basilar Membrane Mechanics

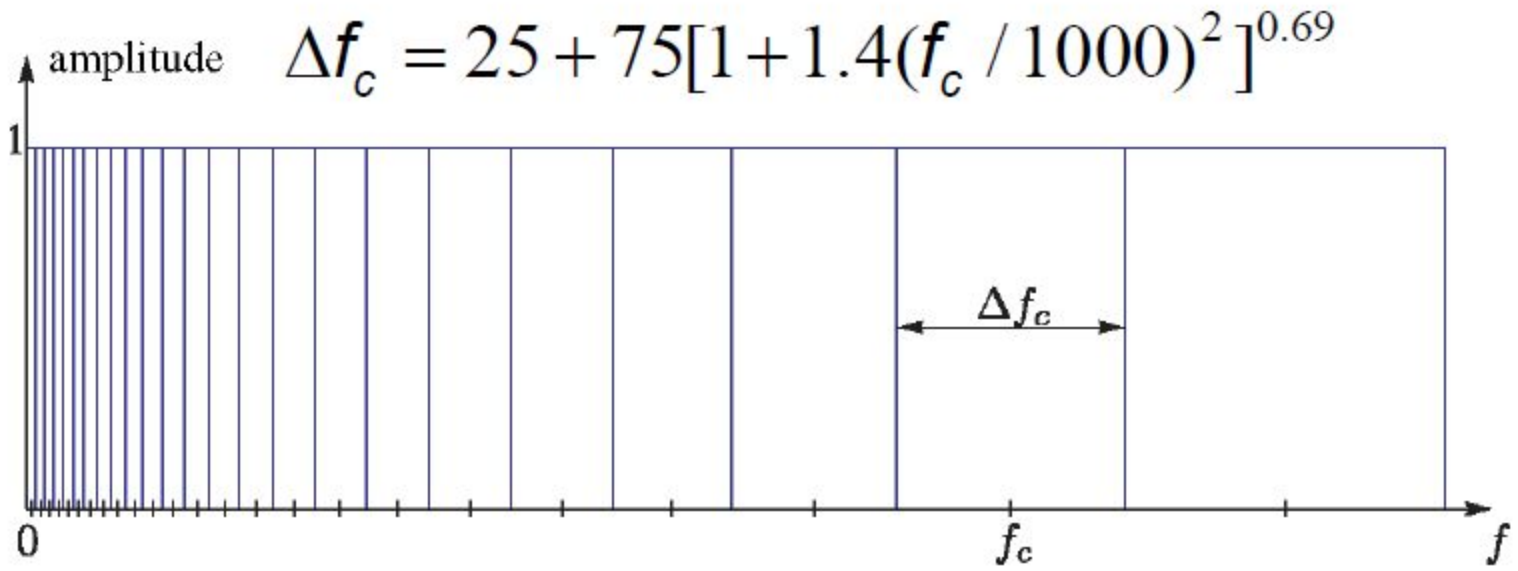
- Characterized by a set of ***frequency responses at different points*** along the membrane
- Mechanical realization of a ***bank of filters***
- Filters are roughly ***constant Q (center frequency/bandwidth) with*** logarithmically decreasing bandwidth
- Distributed along the Basilar Membrane is a set of sensors called ***Inner Hair Cells (IHC) which act as mechanical motion-to-neural*** activity converters
- Mechanical motion along the BM is sensed by local IHC causing ***firing activity at nerve fibers that innervate bottom of each IHC***
- Each IHC connected to about 10 ***nerve fibers, each of different*** diameter => thin fibers fire at high motion levels, thick fibers fire at lower motion levels
- 30,000 nerve fibers link IHC to ***auditory nerve***
- Electrical pulses run along auditory nerve, ultimately reach higher levels of auditory processing in brain, perceived as ***sound***

Basilar Membrane Motion

The ear is excited by the input acoustic wave which has the spectral properties of the speech being produced

- Different regions of the BM respond maximally to different input frequencies => frequency tuning occurs along BM
- The BM acts like a bank of non-uniform cochlear filters
- Roughly logarithmic increase in BW of filters (<800 Hz has equal BW) => constant Q filters with BW decreasing as we move away from cochlear opening
- Peak frequency at which maximum response occurs along the BM is called the characteristic frequency

Critical Bands



Real BM filters overlap significantly

The Perception of Sound

Key questions about sound perception:

--what is the `resolving power' of the hearing mechanism

– how good an estimate of the fundamental frequency of a sound do we need so that the perception mechanism basically `can't tell the difference'

– how good an estimate of the resonances or formants (both center frequency and bandwidth) of a sound do we need so that when we synthesize the sound, the listener can't tell the difference

– how good an estimate of the intensity of a sound do we need so that when we synthesize it, the level appears to be correct

Parameter Discrimination

JND – Just Noticeable Difference
Similar names: differential limen (DL), …

Parameter	JND/DL
Fundamental Frequency	0.3-0.5%
Formant Frequency	3-5%
Formant bandwidth	20-40%
Overall Intensity	1.5 dB

Sound Intensity

- Intensity of a sound is a physical quantity that can be measured and quantified
- Acoustic Intensity (I) defined as the average flow of energy (power) through a unit area, measured in watts/square meter
- Range of intensities between 10^{-12} watts/square meter to 10 watts/square meter; this corresponds to the range from the threshold of hearing to the threshold of pain

Threshold of hearing defined to be:

$I_0 = 10^{-12}$ watts/m^2

The intensity level of a sound, IL is defined relative to I_0 as:

$$IL = 10 \log_{10}\left(\frac{I}{I_0}\right) \text{ in dB}$$

For a pure sinusoidal sound wave of amplitude P, the intensity is proportional to P^2 and the sound pressure level (SPL) is defined as:

$$SPL = 10 \log_{10}\left(\frac{P^2}{P_0^2}\right) = 20 \log_{10}\left(\frac{P}{P_0}\right) \text{ dB}$$

where $P_0 = 2 x 10^{-5}$ Newtons/m^2

Some Facts About Human Hearing

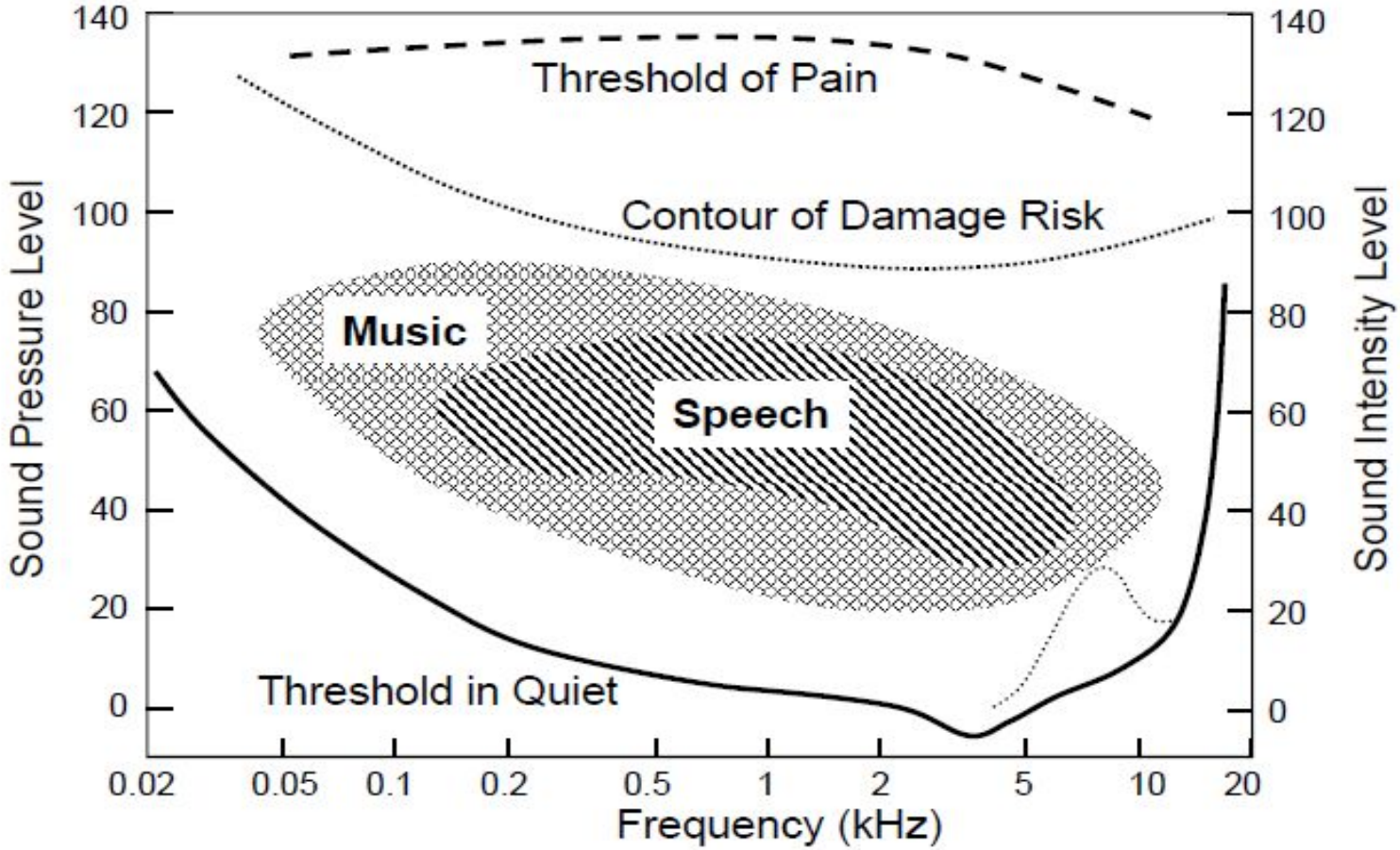
- the ***range of human hearing*** is incredible
  - ***threshold of hearing*** — thermal limit of Brownian motion of air particles in the inner ear
  - ***threshold of pain*** — intensities of from 10**12 to 10**16 greater than the threshold of hearing
- human hearing perceives both ***sound frequency*** 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)

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***

Range of Human Hearing



Hearing Thresholds

Threshold of Audibility is the acoustic intensity

level of a pure tone that can barely be heard at a particular frequency

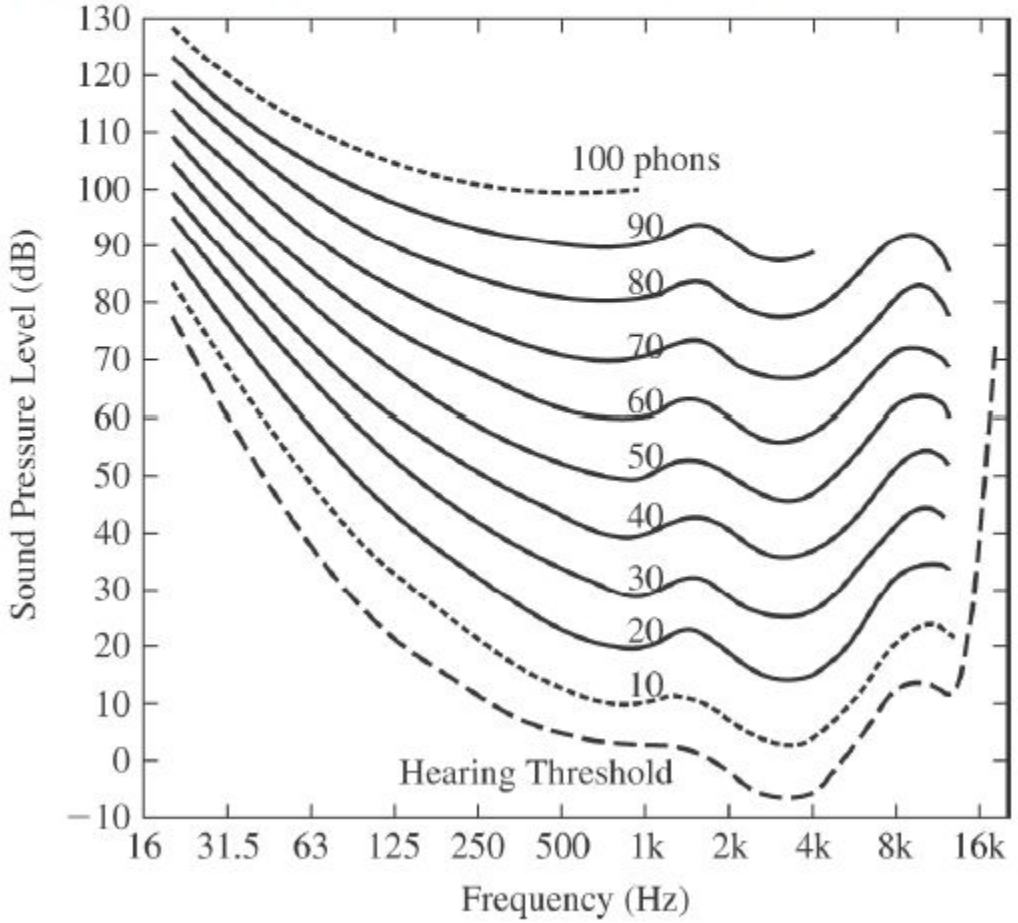
– *threshold of audibility ≈ 0 dB at 1000 Hz*
– *threshold of feeling ≈ 120 dB*
– *threshold of pain ≈ 140 dB*
– *immediate damage ≈ 160 dB*

• *Thresholds vary with frequency and from person-to-person*

• *Maximum sensitivity is at about 3000 Hz*

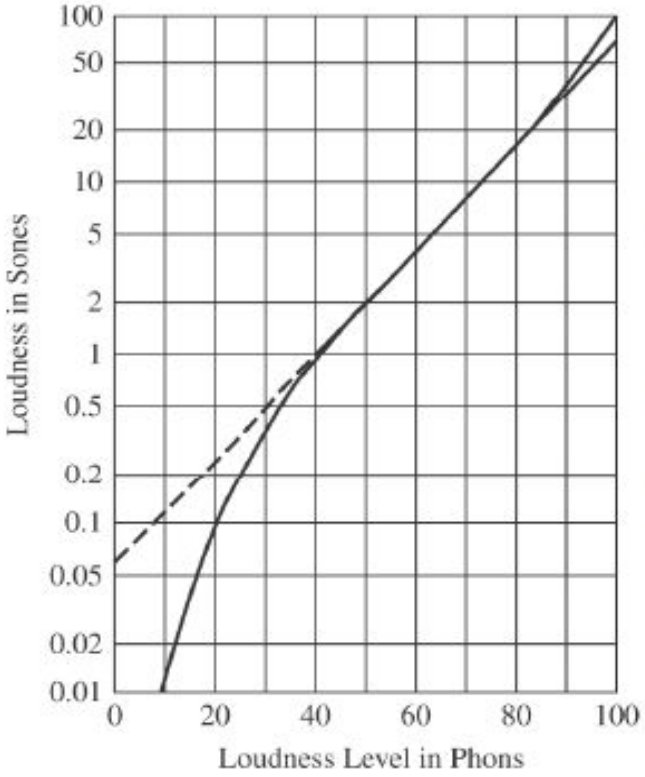
Loudness Level

- ***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

- ***Loudness (L)*** (in sones) is a scale that doubles whenever the *perceived* loudness doubles



$$\log L = 0.033\,(LL - 40)$$
$$= 0.033LL - 1.32$$

• for a frequency of 1000 Hz, the loudness level, LL, in phons is, by definition, numerically equal to the intensity level IL in decibels, so that the equation may be rewritten as

$$LL = 10\log(I / I_0)$$

or since $I_0 = 10^{-12}$ watts/m^2

$$LL = 10\log I + 120$$

Substitution of this value of *LL* in the equation gives

$$\log L = 0.033(10\log I + 120) - 1.32$$
$$= 0.33\log I + 2.64$$

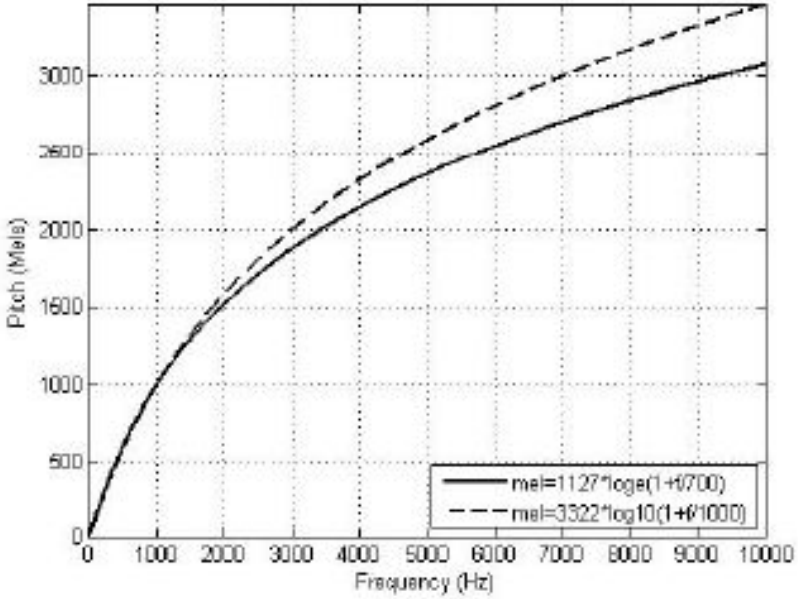
which reduces to

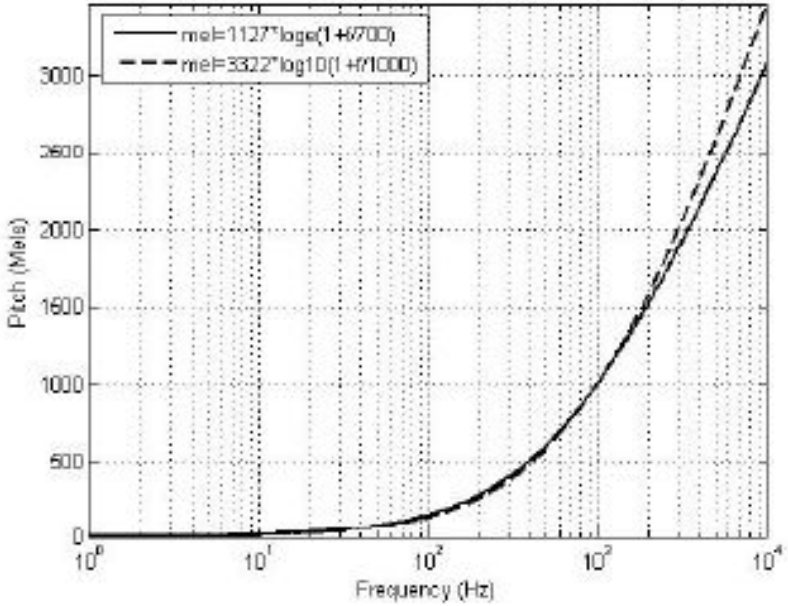
$$L = 445I^{0.33}$$

Pitch

- ***pitch*** and ***fundamental frequency*** are not the same thing
- we are quite sensitive to changes in pitch
  - $F < 500$ Hz, $\Delta F \approx 3$ Hz
  - $F > 500$ Hz, $\Delta F/F \approx 0.003$
- relationship between pitch and fundamental frequency is not simple, even for pure tones
  - the tone that has a pitch half as great as the pitch of a 200 Hz tone has a frequency of about 100 Hz
  - the tone that has a pitch half as great as the pitch of a 5000 Hz tone has a frequency of less than 2000 Hz
- the pitch of complex sounds is an even more complex and interesting phenomenon

Pitch-The Mel Scale





$$\text{Pitch } (mels) = 3322 \log_{10}(1 + f / 1000)$$

Alternatively, we can approximate curve as:

$$\text{Pitch } (mels) = 1127 \log_{e}(1 + f / 700)$$

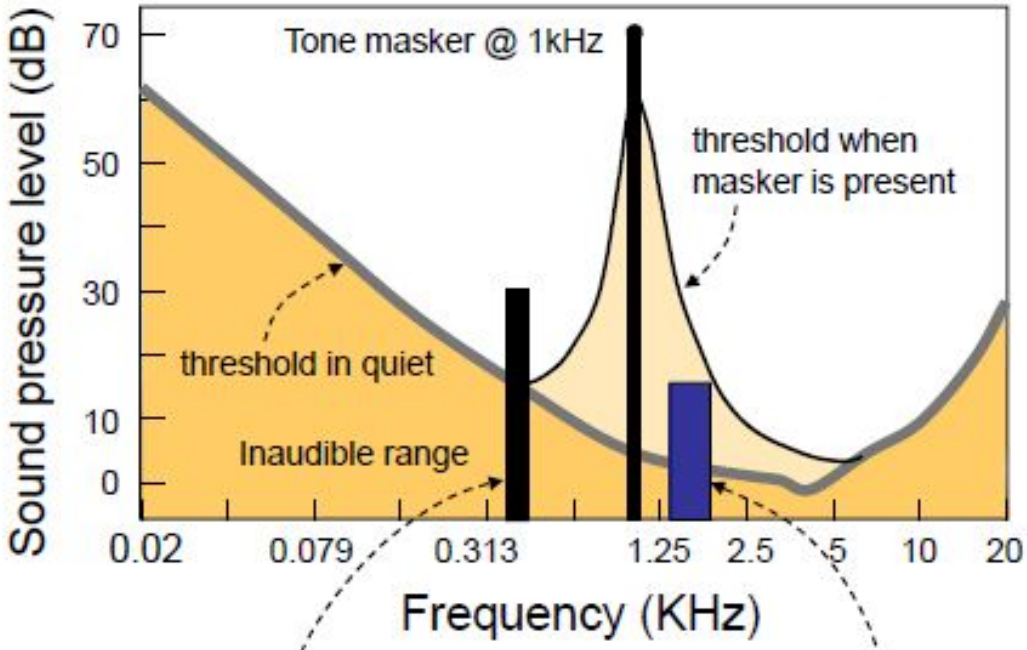
Perception of Frequency

- Pure tone
  - **Pitch** is a perceived quantity while **frequency** is a physical one (cycle per second or Hertz)
  - **Mel** is a scale that doubles whenever the perceived pitch doubles; start with 1000 Hz = 1000 mel, increase frequency of tone until listener perceives twice the pitch (or decrease until half the pitch) and so on to find mel-Hz relationship
  - The relationship between pitch and frequency is non-linear
- Complex sound such as speech
  - **Pitch** is related to **fundamental frequency** but not the same as fundamental frequency; the relationship is more complex than pure tones
- **Pitch period** is related to time.

END

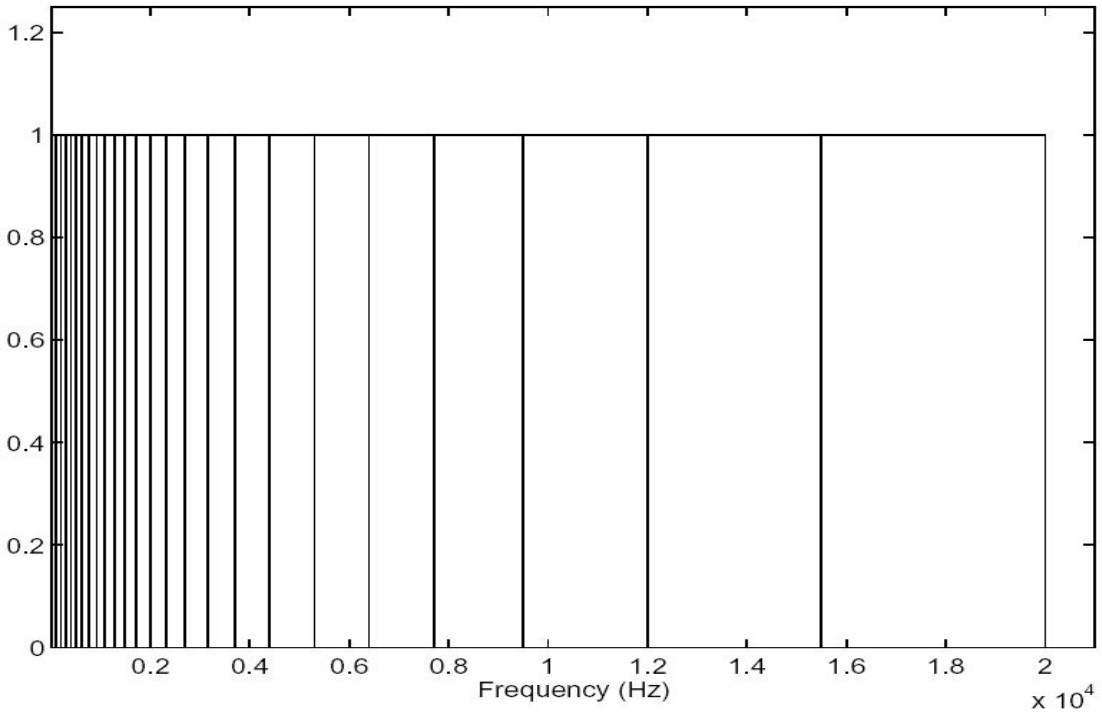
Masking as defined by the American Standards Association (ASA) is the amount (or the process) by which the threshold of audibility for one sound is raised by the presence of another (masking) sound (B.C.J. Moore 1982, p. 74)

Auditory Masking



Signal perceptible even in the presence of the tone masker

Signal not perceptible due to the presence of the tone masker



- About 75% of the critical bands are below 5 kHz
- The hearing system receives more information from these low frequencies.
- Approximately, the critical bandwidth is 100 Hz up to 500 Hz
- Above 500 Hz, the critical bandwidth is approximately 20% of the center frequency .

Nonuniform AUDITORY FILTERBANK

A critical band is the bandwidth around a center frequency beyond which subjective responses of the hearing system abruptly change

The hearing system contains a bank of overlapping band pass linear filters.

Therefore, the peripheral section of the hearing system can be modeled as a non uniform filter bank consisting of band pass filter with bandwidths equal to critical bandwidths.

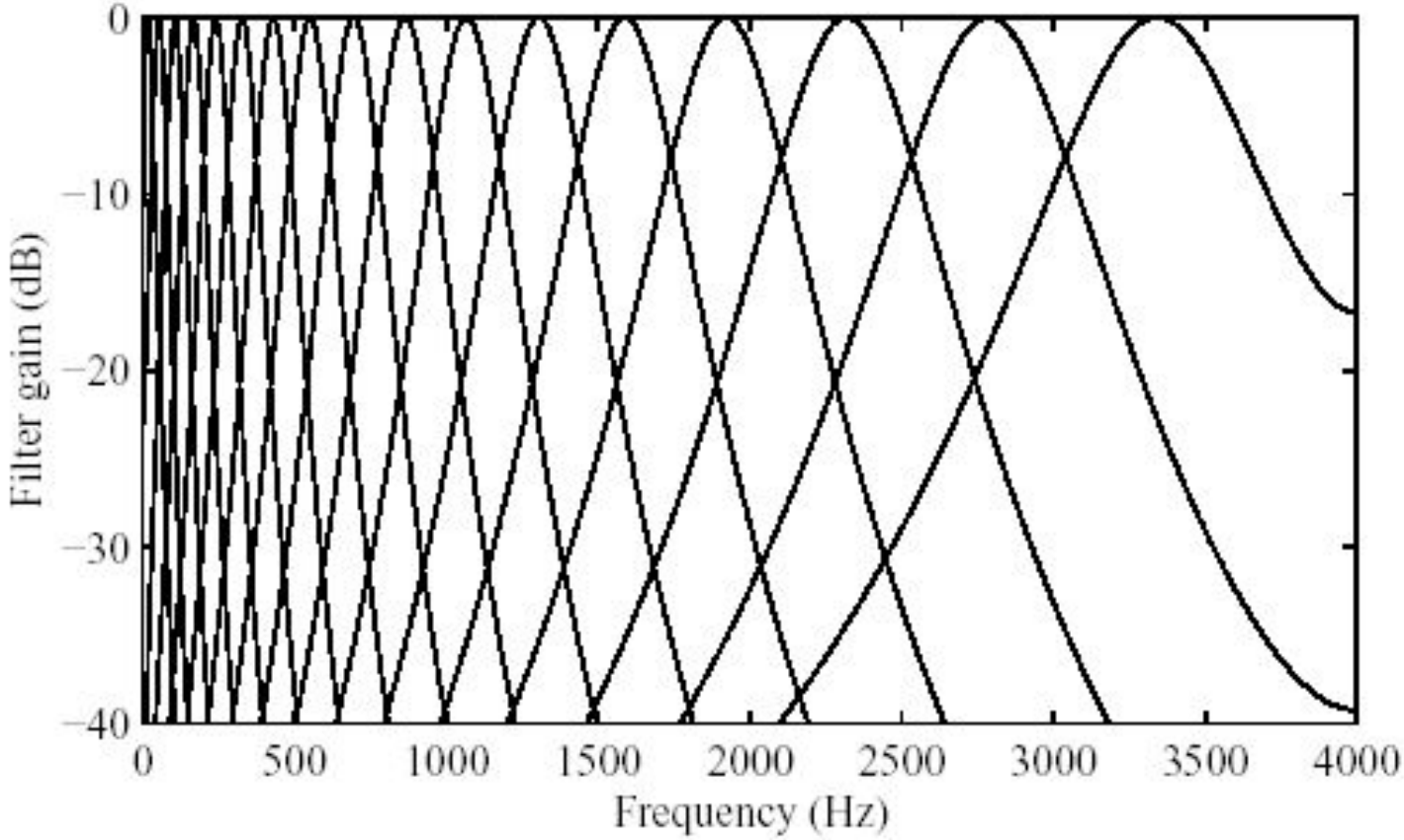
Critical Bands

- The ear cannot distinguish sounds within the same band that occur simultaneously.
- Each band is called a critical band
- The auditory system can be roughly modeled as a filterbank, consisting of 25 overlapping bandpass filters, from 0 to 20 KHz
- The bandwidth of each critical band is about 100 Hz for signals below 500 Hz, and increases non-linearly after 500 Hz up to 5000 Hz
- 1 bark = width of 1 critical band

$$\text{Bark} = \begin{cases} f/100, & f \le 500\text{Hz} \\ 9 + 4\log_2(f/1000), & f > 500\text{Hz} \end{cases}$$

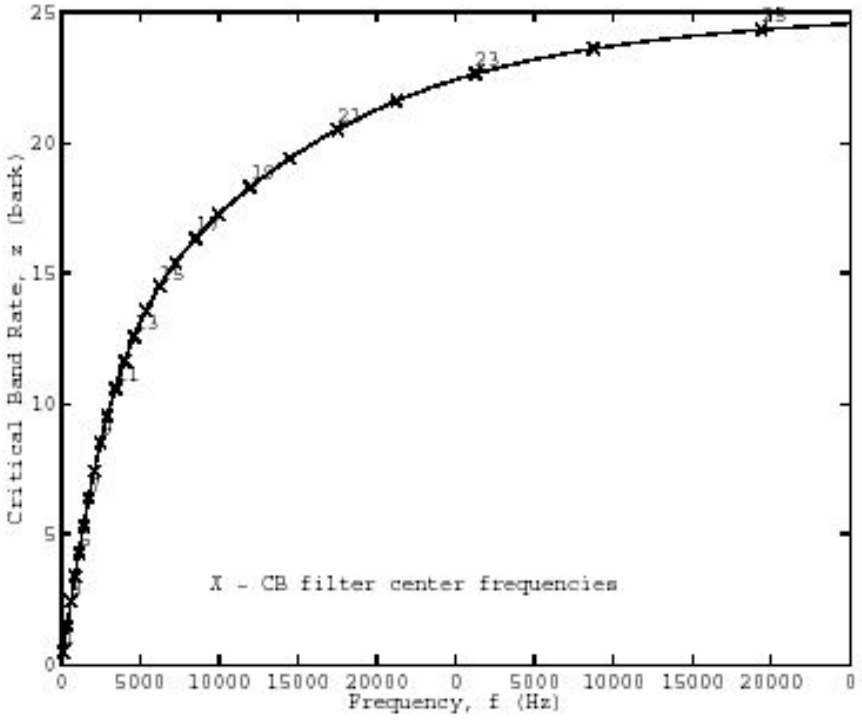
Critical Bands

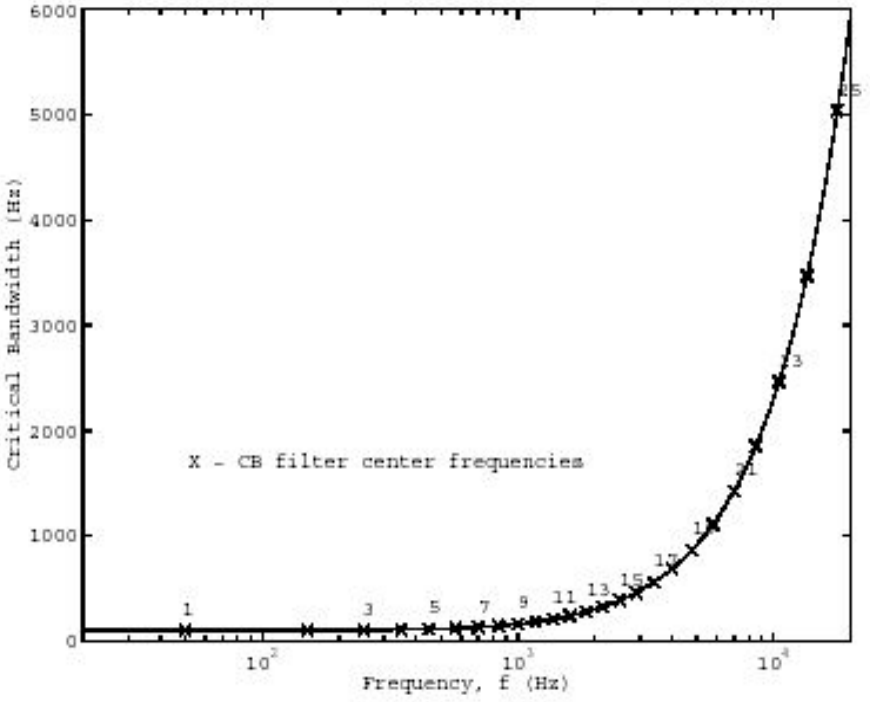
Critical bands: The widths of the masking bands for different masking tones are different, increasing with the frequency of the masking tone.



Although the function *BWc* is continuous, it is useful when building practical systems to treat the ear as a discrete set of bandpass filters that conforms to Eq.

$$BW_c(f) = 25 + 75\left[1 + 1.4(f/1000)^2\right]^{0.69}$$





Cochlea-Overlapping Bandpass Filter

1. A frequency-to-place transformation takes place in the cochlea (inner ear), along the basilar membrane.
2. Distinct regions in the cochlea, each with a set of neural receptors, are “tuned” to different frequency bands.
3. In fact, the cochlea can be viewed as a bank of highly overlapping band-pass filters.
4. The magnitude responses are asymmetric and non-linear (level-dependent).
5. Moreover, the cochlear filter pass-bands are of non-uniform bandwidth, and the bandwidths increase with increasing frequency.
6. The “critical bandwidth” is a function of frequency that quantifies the cochlear filter pass-bands.

Cochlea-Overlapping Bandpass Filter

A distance of 1 critical band is commonly referred to as “one bark” in the literature.

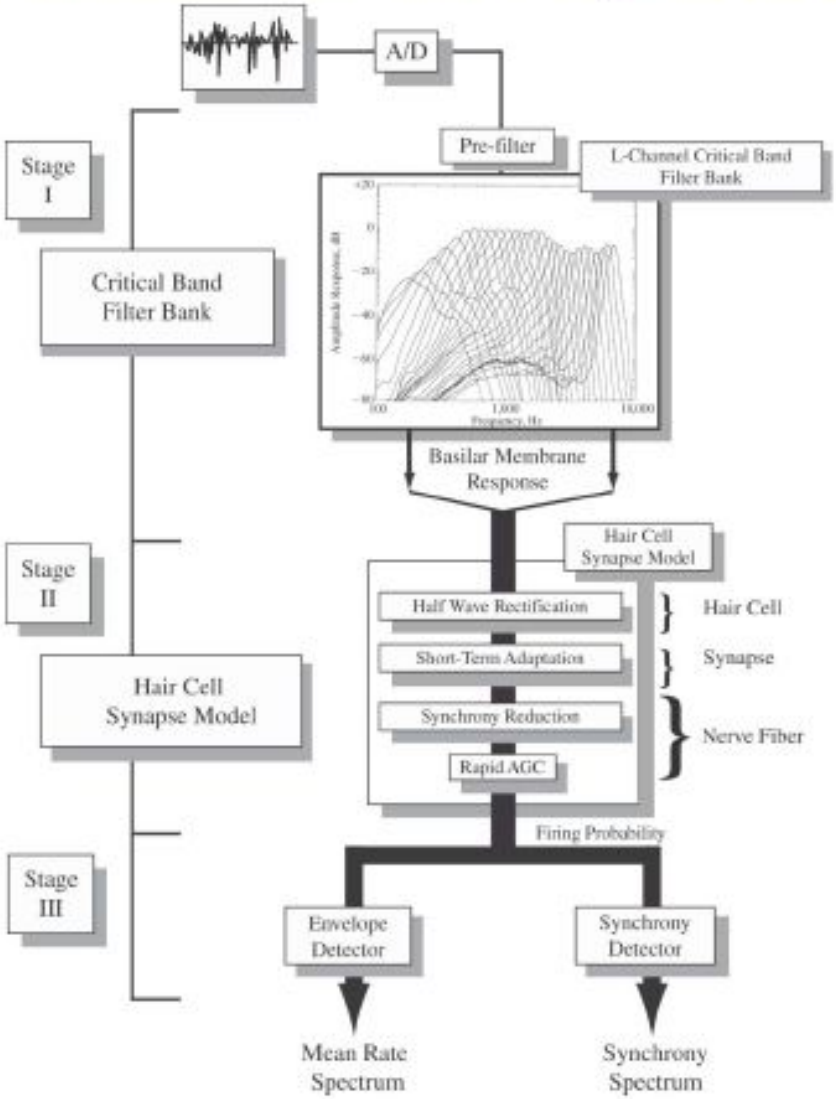
$$z(f) = 13\arctan(.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \text{ (Bark)}$$

Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)	Band No.	Center Freq. (Hz)	Bandwidth (Hz)
1	50	-100	10	1175	1080-1270	19	4800	4400-5300
2	150	100-200	11	1370	1270-1480	20	5800	5300-6400
3	250	200-300	12	1600	1480-1720	21	7000	6400-7700
4	350	300-400	13	1850	1720-2000	22	8500	7700-9500
5	450	400-510	14	2150	2000-2320	23	10,500	9500-12000
6	570	510-630	15	2500	2320-2700	24	13,500	12000-15500
7	700	630-770	16	2900	2700-3150	25	19,500	15500-
8	840	770-920	17	3400	3150-3700			
9	1000	920-1080	18	4000	3700-4400			

Auditory Models

- Perceptual effects included in most auditory models:
  - spectral analysis on a non-linear frequency scale (usually mel or Bark scale)
  - spectral amplitude compression (dynamic range compression)
  - loudness compression via some logarithmic process
  - decreased sensitivity at lower (and higher) frequencies based on results from equal loudness contours
  - utilization of temporal features based on long spectral integration intervals (syllabic rate processing)
  - auditory masking by tones or noise within a critical frequency band of the tone (or noise)

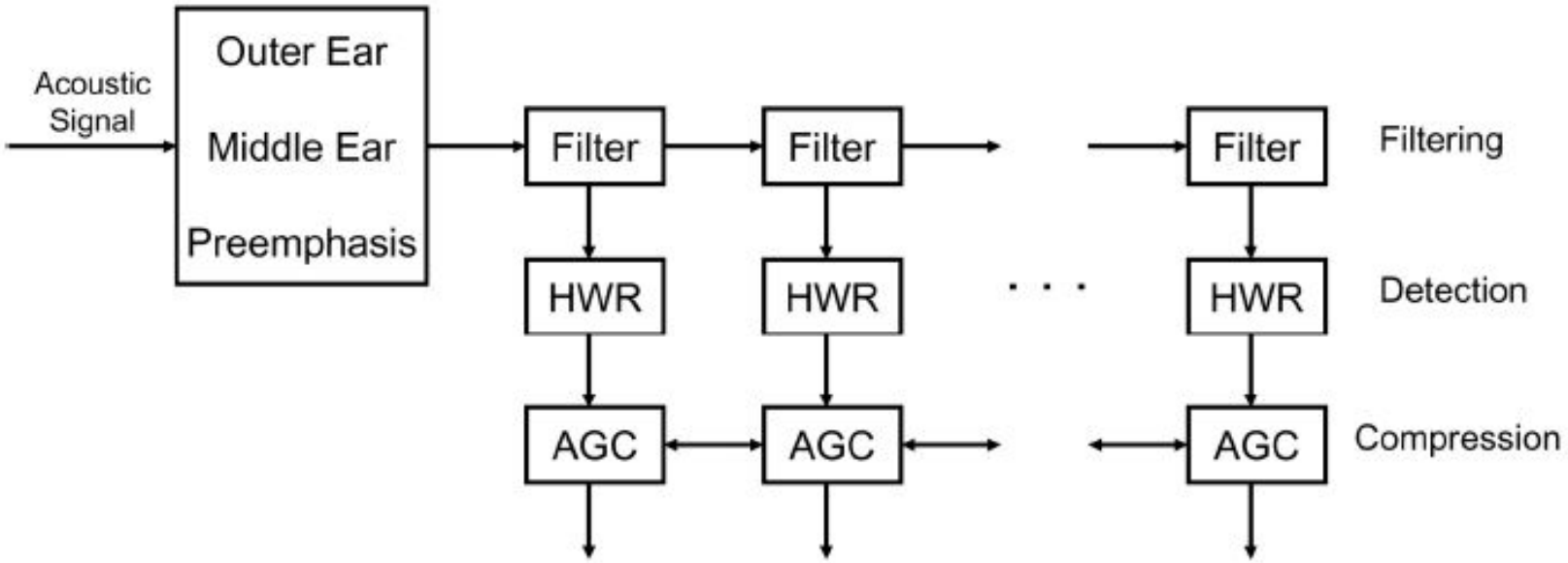
Seneff Auditory Model



Seneff Auditory Model

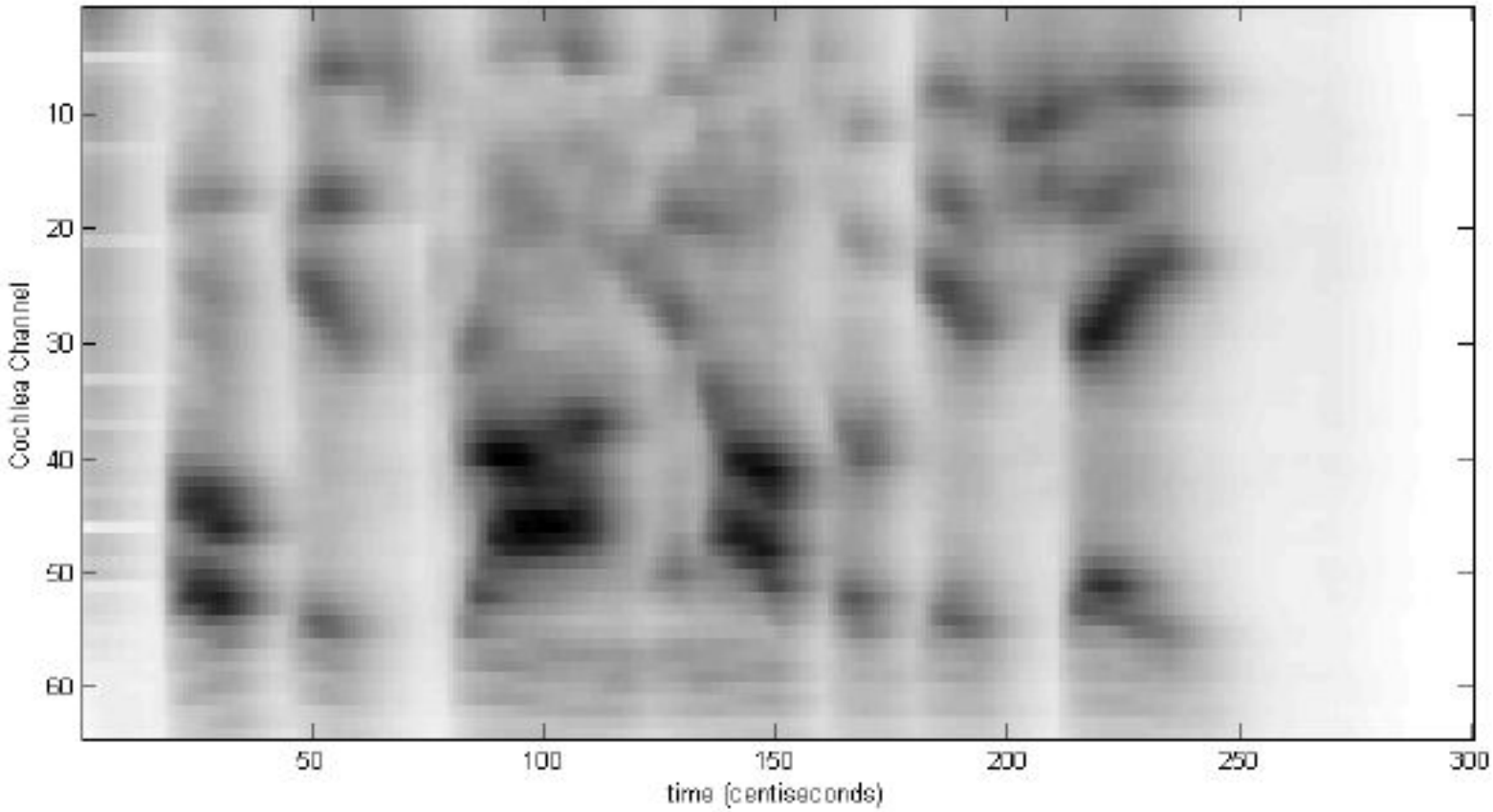
- This model tried to capture essential features of the response of the cochlea and the attached hair cells in response to speech sound pressure waves
- Three stages of processing:
  - stage 1 pre-filters the speech to eliminate very low and very high frequency components, and then uses a 40-channel critical band filter bank distributed on a Bark scale
  - stage 2 is a hair cell synapse models which models the (probabilistic) behavior of the combination of inner hair cells, synapses, and nerve fibers via the processes of half wave rectification, short-term adaptation, and synchrony reduction and rapid automatic gain control at the nerve fiber; outputs are the probabilities of firing, over time, for a set of similar fibers acting as a group
  - stage 3 utilizes the firing probability signals to extract information relevant to perception; i.e., formant frequencies and enhanced sharpness of onset and offset of speech segments; an Envelope Detector estimates the Mean Rate Spectrum (transitions from one phonetic segment to the next) and a Synchrony Detector implements a phase-locking property of nerve fibers, thereby enhancing spectral peaks at formants and enabling tracking of dynamic spectral changes

Lyon's Cochlear Model



- Pre-processing stage (simulating effects of outer and middle ears as a simple pre-emphasis network)
  - three full stages of processing for modeling the cochlea as a non-linear filter bank
  - first stage is a bank of 86 cochlea filters, space non0uniformly according to mel or Bark scale, and highly overlapped in frequency
  - second stage uses a half wave rectifier non-linearity to convert basilar membrane signals to Inner Hair Cell receptor potentials or Auditory Nerve firing rates
  - third stage consists of inter-connected AGC circuits which continuously adapt in response to activity levels at the outputs of the HWRs of the second stage to compress the wide range of sound levels into a limited dynamic range of basilar membrand motion, IHC receptor potential and AN firing rates

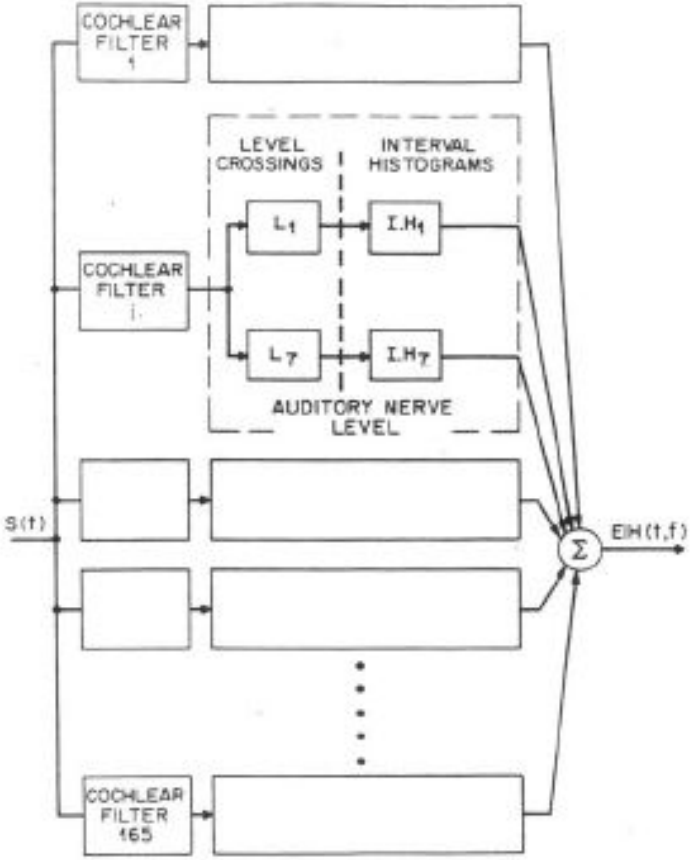
Lyon's Cochleargram



Cochleagram is a plot of model intensity as a function of place (warped frequency) and time; i.e., a type of auditory model spectrogram.

Ensemble Interval Histogram (EIH)

• model of cochlear and hair cell transduction => filter bank that models frequency selectivity at points along the BM, and nonlinear processor for converting filter bank output to neural firing patterns along the auditory nerve



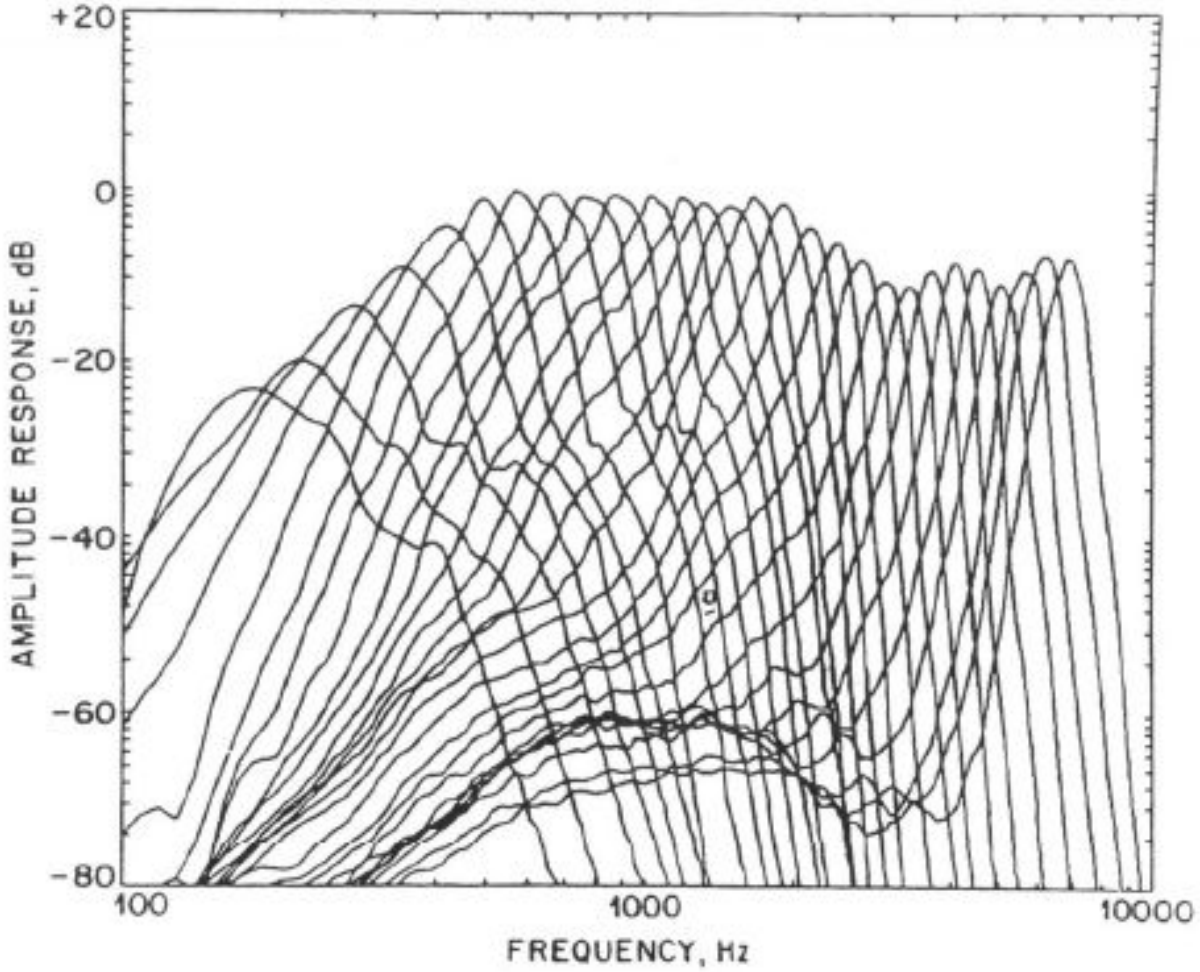
• 165 channels, equally spaced on a log frequency scale between 150 and 7000 Hz

• cochlear filter designs match neural tuning curves for cats => minimum phase filters

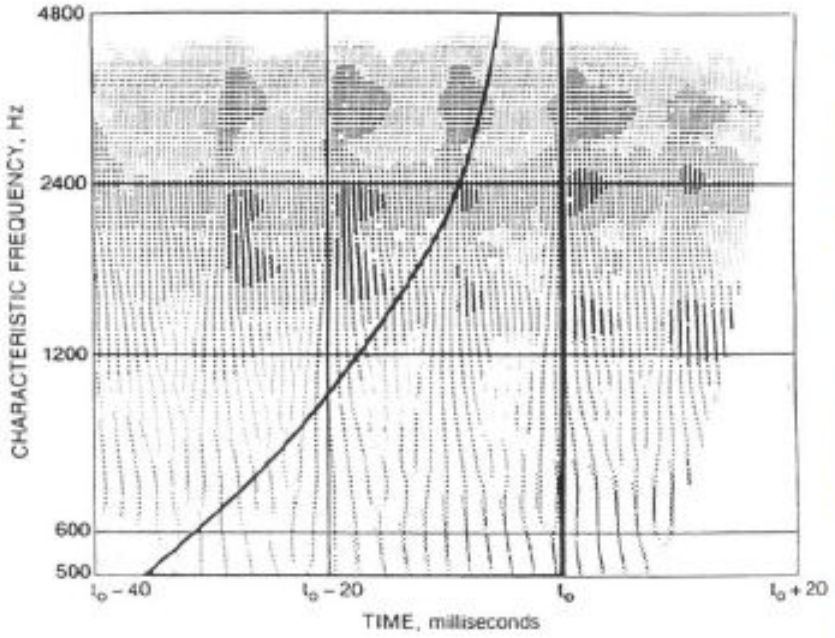
• array of level crossing detectors that model motion-to-neural activity transduction of the IHCs

  • detection levels are pseudo-randomly distributed to match variability of fiber diameters

Cochlear Filter Designs



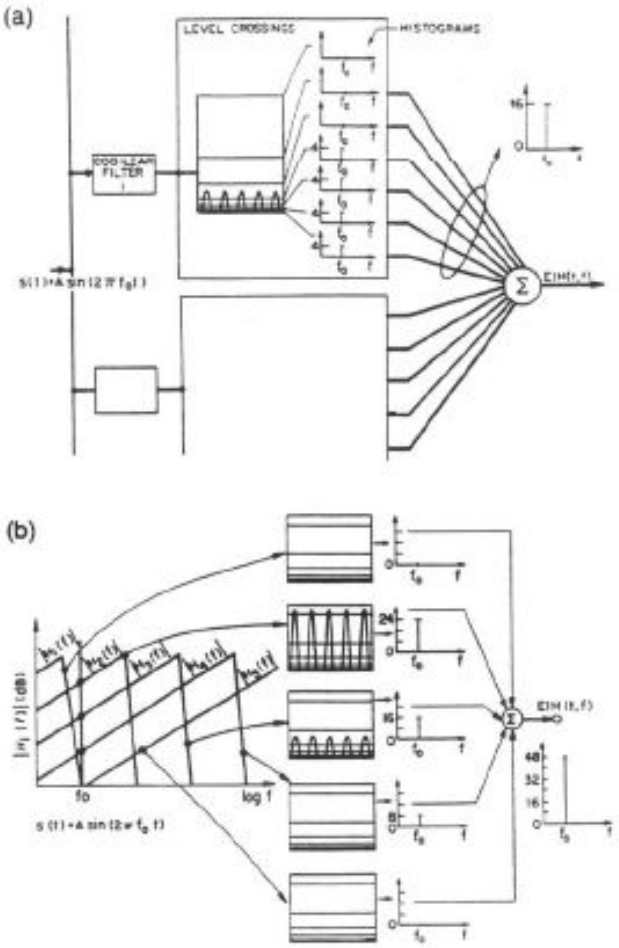
EIH Responses



- plot shows simulated auditory nerve activity for first 60 msec of /o/ in both time and frequency of IHC channels
- log frequency scale
- level crossing occurrence marked by single dot; each level crossing detector is a separate trace
- for filter output low level—1 or fewer levels will be crossed
- for filter output high level—many levels crossed => darker region

Overall EIH

- EIH is a measure of spatial extent of coherent neural activity across auditory nerve
- it provides estimate of short term PDF of reciprocal of intervals between successive firings in a characteristic frequency-time zone
- EIH preserves signal energy since threshold crossings are functions of amplitude
  - as A increases, more levels are activated



response to pure sinusoid

Why Auditory Models

- Match human speech perception
  - Non-linear frequency scale – mel, Bark scale
  - Spectral amplitude (dynamic range) compression – loudness (log compression)
  - Equal loudness curve – decreased sensitivity at lower frequencies
  - Long spectral integration – “temporal” features

What Do We Learn From Auditory Models

- Need both short (20 msec for phonemes) and long (200 msec for syllables) segments of speech
- Temporal structure of speech is important
- Spectral structure of sounds (formants) is important
- Dynamic (delta) features are important