Chapter 4

Hearing, Auditory Models, and Speech Perception

听觉，听觉模型与语音感知

Topics to be Covered

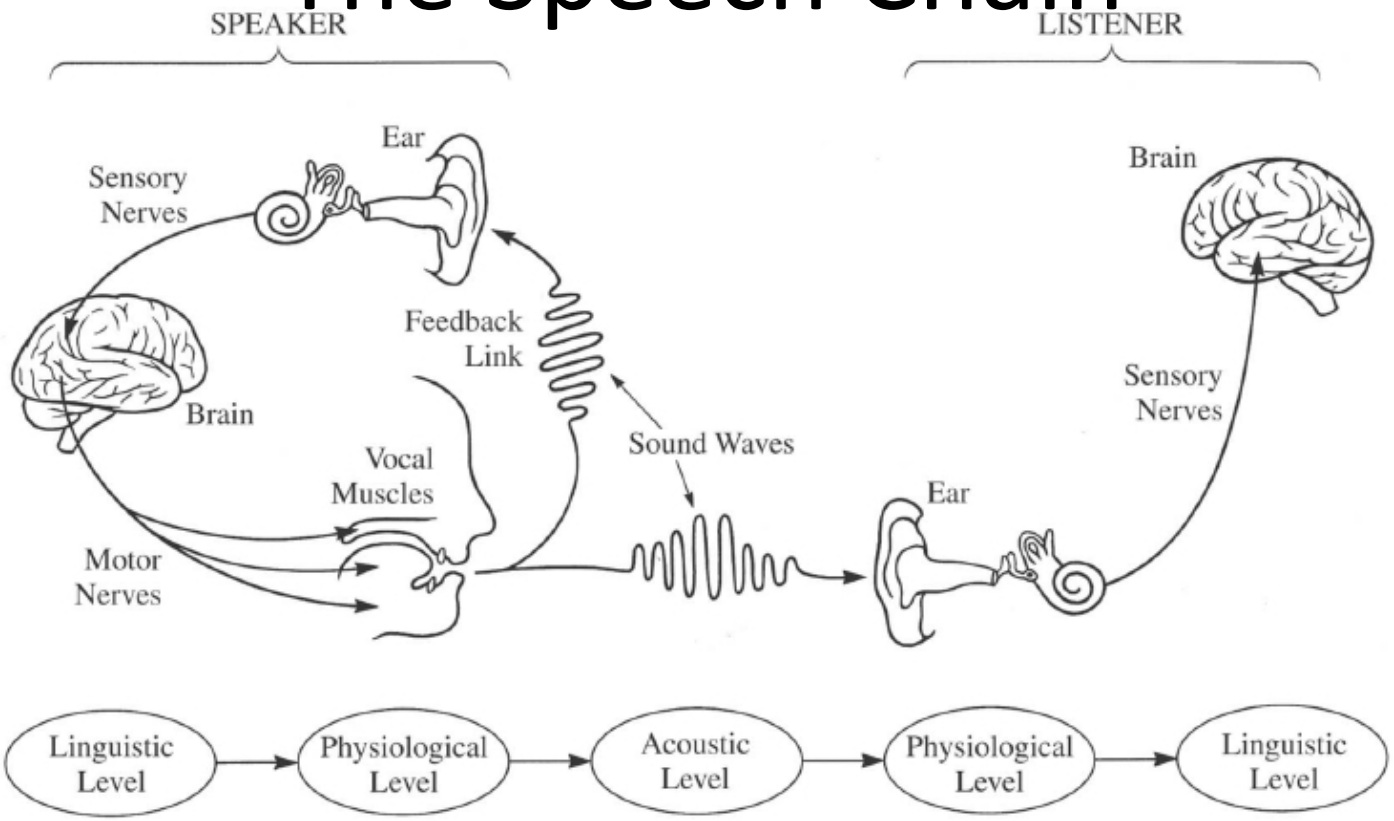
- The Speech Chain (语音链) – Production and Human Perception
- Auditory mechanisms (听觉机理)— the human ear and how it converts sound to auditory representations
- Speech perception (语音感知) and what we know about physical and psychophysical measures of sound
- Auditory masking (听觉掩蔽)
- Sound and word perception in noise

Auditory Mechanisms

Speech Perception

- understanding how we hear sounds and how we perceive speech leads to 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
  - speech and audio coding (MP3 audio, cellphone speech)
  - speech recognition
- try to understand speech perception by looking at the physiological models of hearing

The Speech Chain

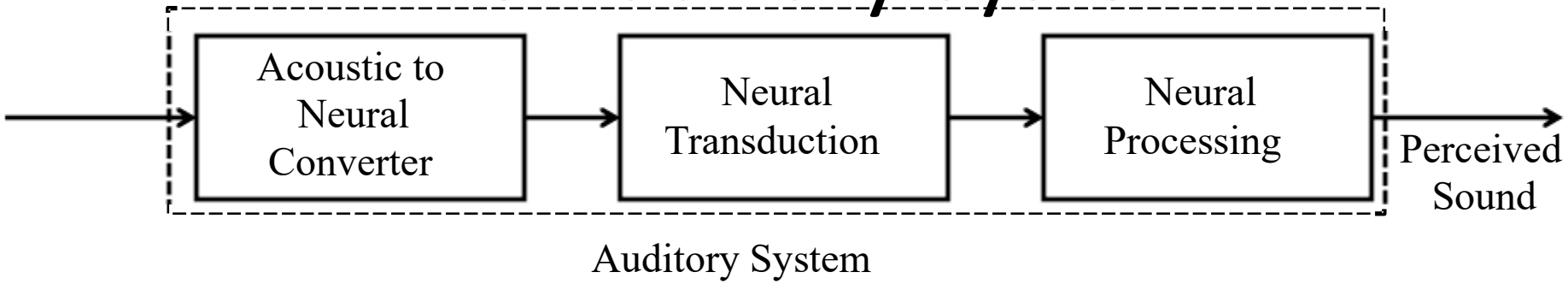


- The Speech Chain comprises the processes of:
  - speech production,
  - auditory feedback to the speaker,
  - speech transmission (through air or over an electronic communication system) to the listener, and
  - speech perception and understanding by the listener.

The Speech Chain

- The message to be conveyed by speech goes through five levels of representation between the speaker and the listener, namely:
  - the linguistic level (where the basic sounds of the communication are chosen to express some thought of idea)
  - the physiological level (where the vocal tract components produce the sounds associated with the linguistic units of the utterance)
  - the acoustic level (where sound is released from the lips and nostrils and transmitted to both the speaker (sound feedback) and to the listener)
  - the physiological level (where the sound is analyzed by the ear and the auditory nerves), and finally
  - 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

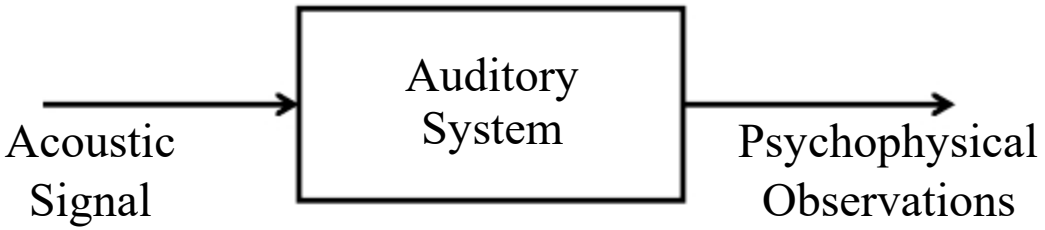


- the acoustic signal first converted to a neural representation by processing in the ear
  - the conversion takes place in stages at the outer, middle and inner ear
  - these processes can be measured and quantified
- the neural transduction step takes place between the output of the inner ear and the 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 McGurk Effect

The Black Box Model of the Auditory System

- researchers have resorted to a “black box” behavioral model of hearing and perception
  - model assumes that an acoustic signal enters the auditory system causing behavior that we record as psychophysical (精神物理学) observations
  - psychophysical methods and sound perception experiments determine how the brain processes signals with different loudness levels, different spectral characteristics, and different temporal properties
  - 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
  - we then determine how various attributes of sound (or speech) are processed by the auditory system

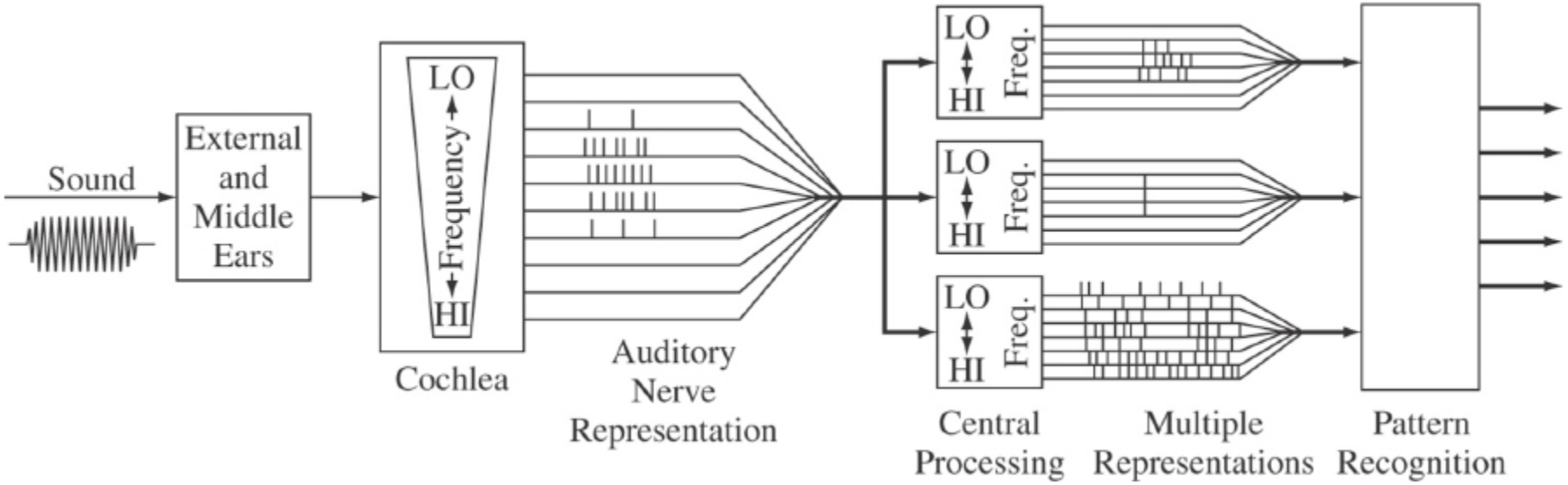


The Black Box Model Examples

Physical Attribute	Psychophysical Observation
Intensity 强度	Loudness 响度
Frequency 频率	Pitch 音高

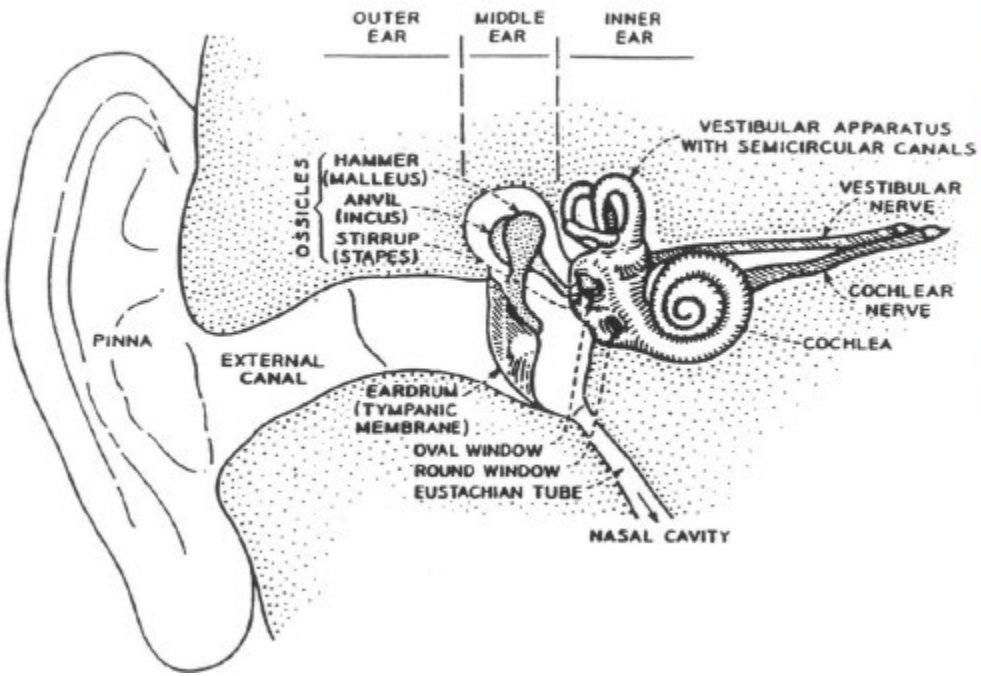
- Experiments with the “black box” model show:
  - correspondences between sound intensity and loudness, and between frequency and pitch are complicated and far from linear
  - attempts to extrapolate from psychophysical measurements to the processes of speech perception and language understanding are, at best, highly susceptible to misunderstanding of exactly what is going on in the brain

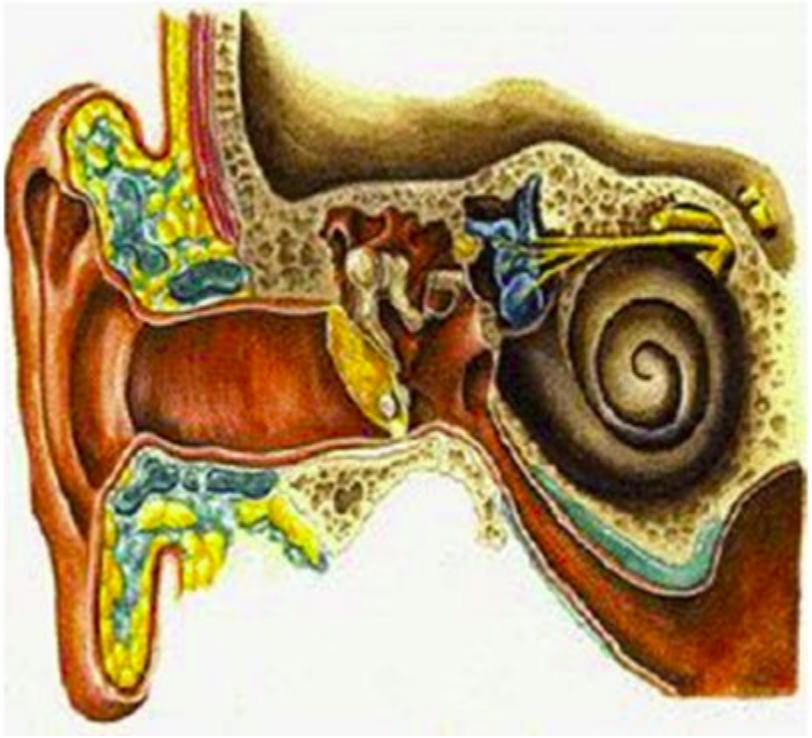
Overview of Auditory Mechanism



- begin by looking at ear models including processing in cochlea (耳蜗)

The Human Ear



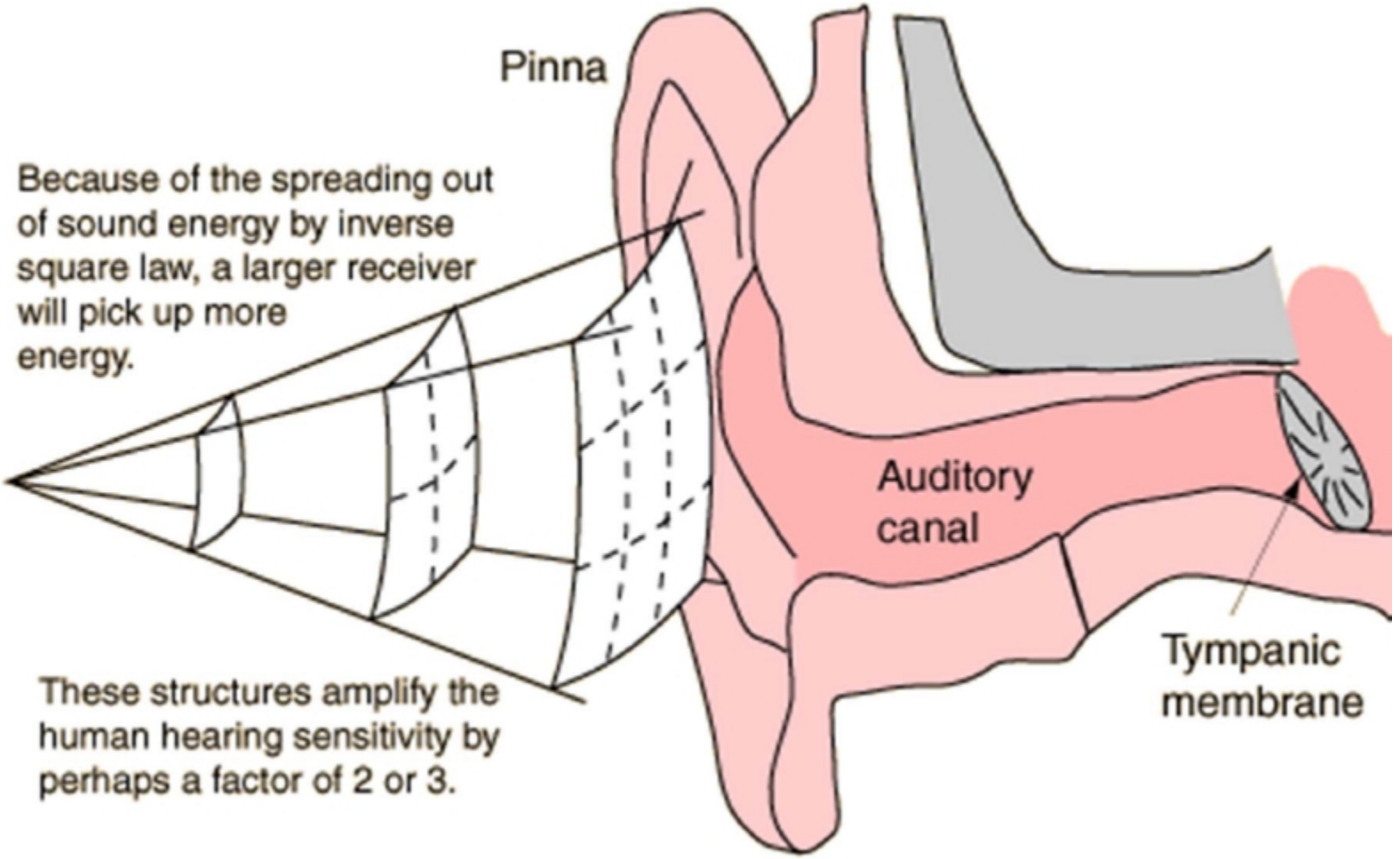


- Outer ear (外耳): pinna (耳廓) and external canal
- Middle ear (中耳): tympanic membrane (鼓膜) or eardrum
- Inner ear (内耳): cochlea(耳蜗), neural connections

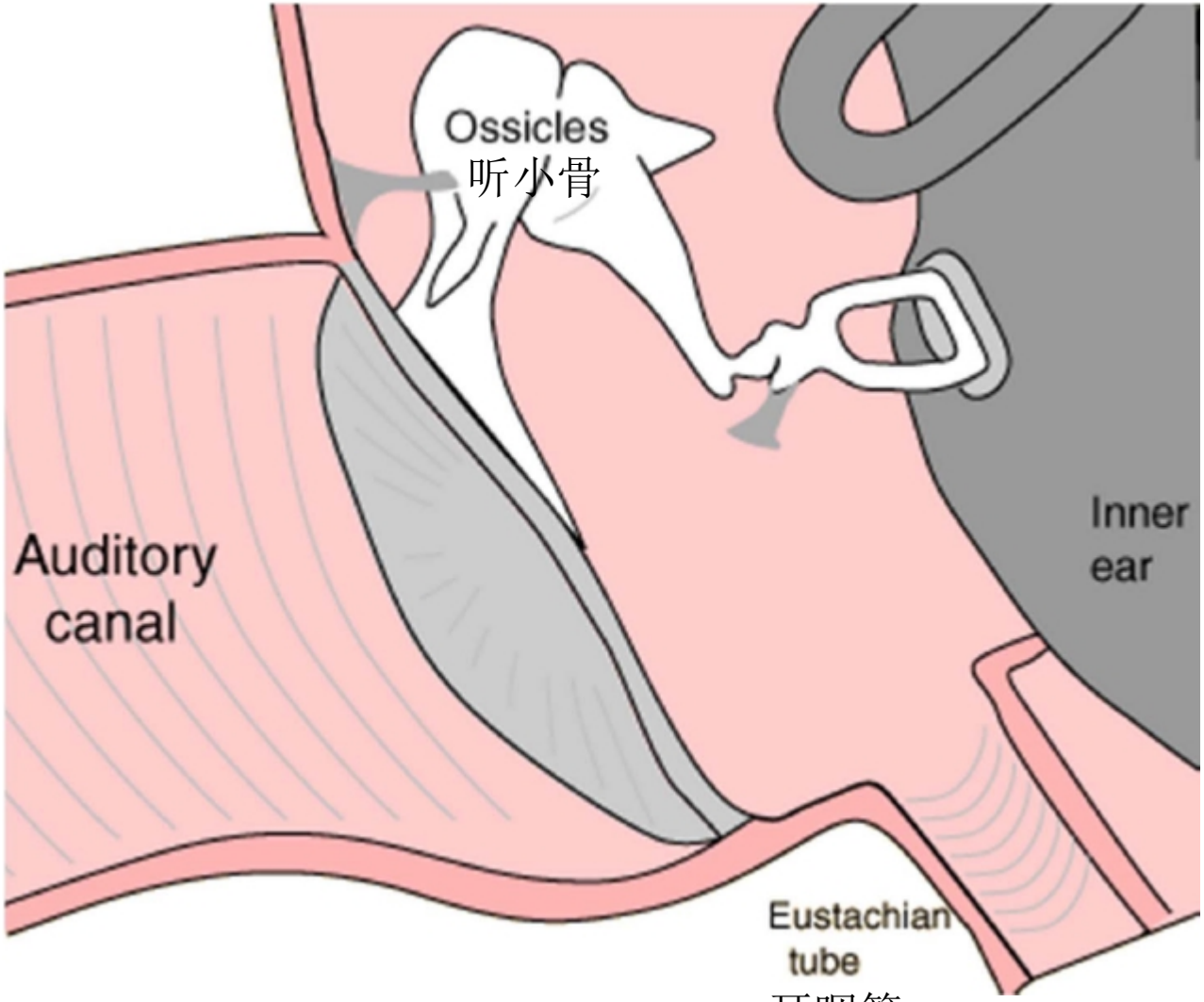
Human Ear

- Outer ear: funnels (使经过漏斗) sound into ear canal
- Middle ear: sound impinges (撞击) on tympanic membrane; this causes motion
  - middle ear is a mechanical transducer, consisting of the hammer (锤骨), anvil (砧骨) and stirrup (镫骨); 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 (基底膜)
  - 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

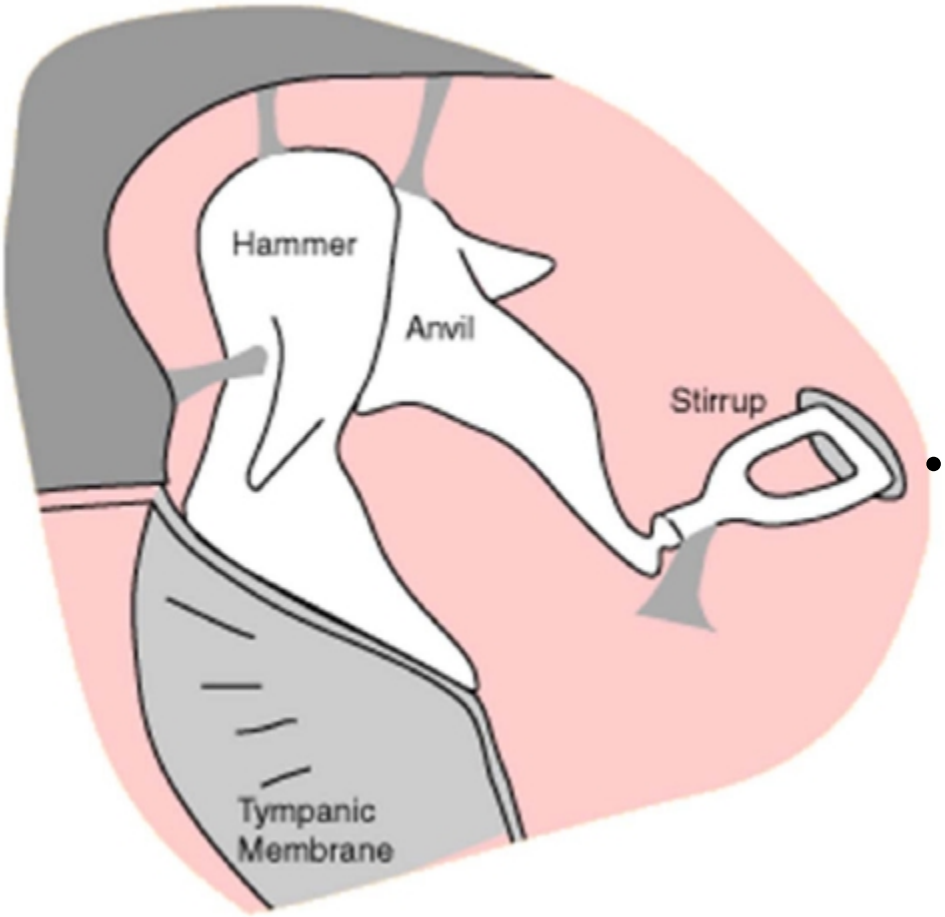
The Outer Ear



The Outer Ear

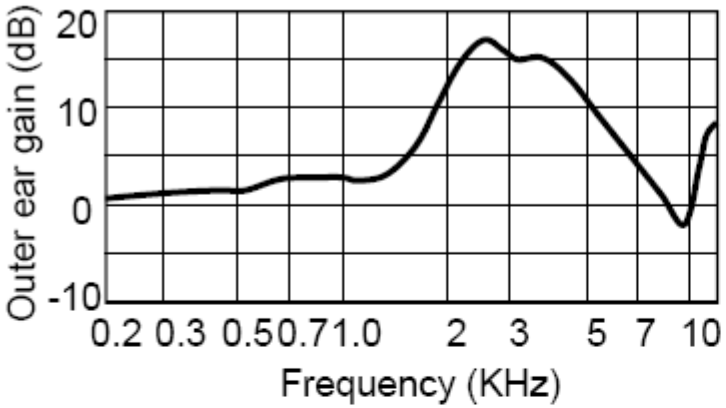


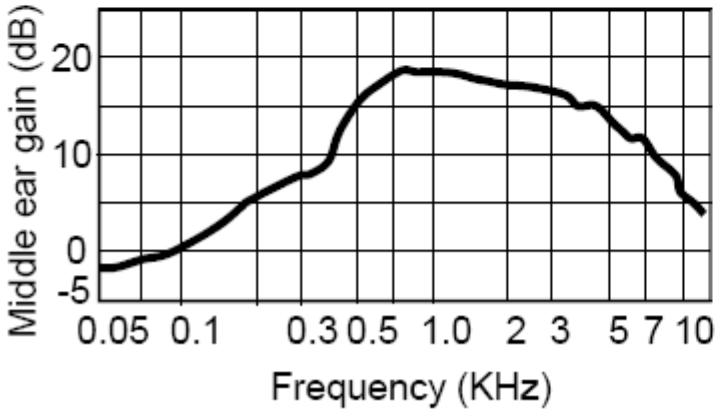
The Middle Ear

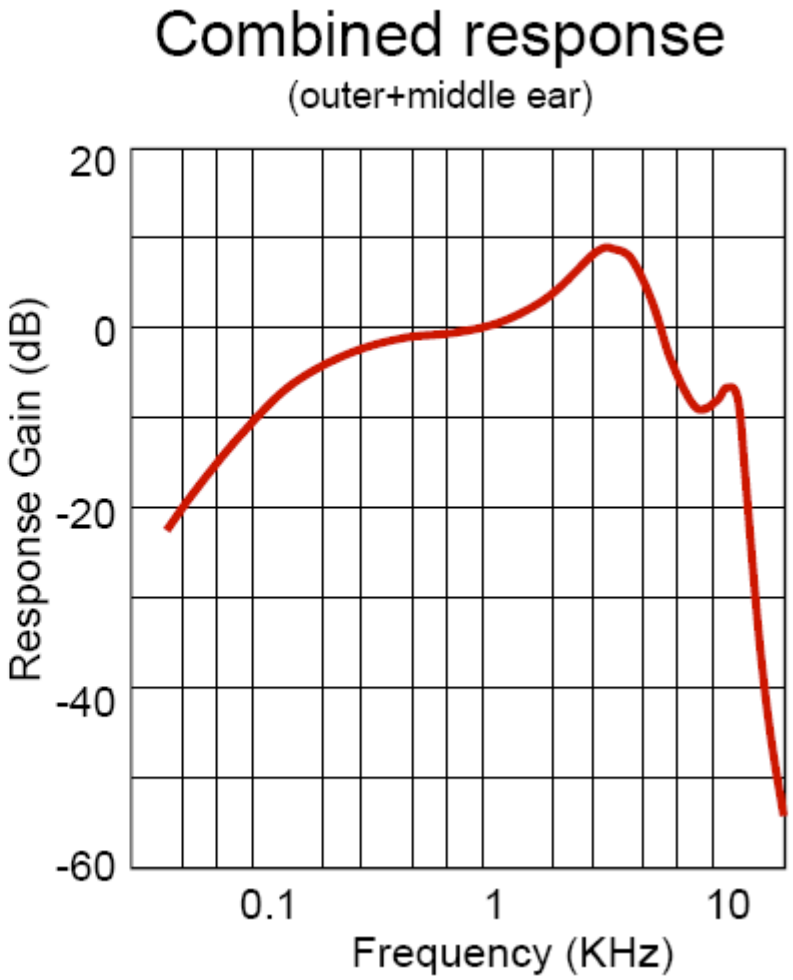


- The Hammer (锤骨), Anvil (砧骨) and Stirrup (镫骨) are the three tiniest bones in the body. Together they form the coupling between the vibration of the eardrum and the forces exerted on the oval window (卵圆窗) of the inner ear.
- These bones can be thought of as a compound lever which achieves a multiplication of force—by a factor of about three under optimum conditions. (They also protect the ear against loud sounds by attenuating the sound.)

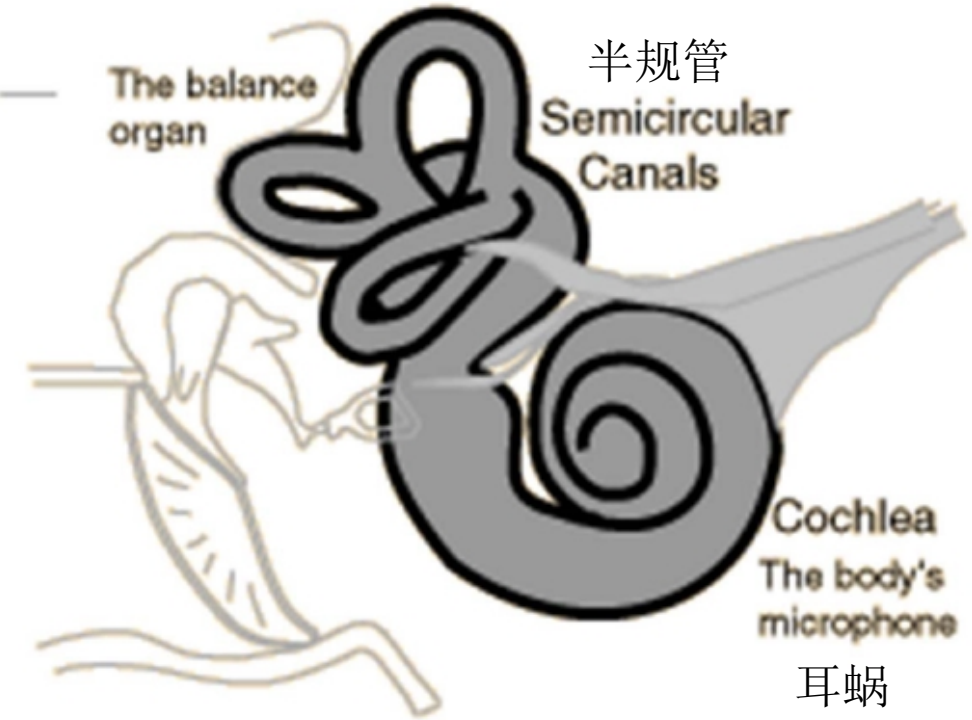
Transfer Functions at the Periphery





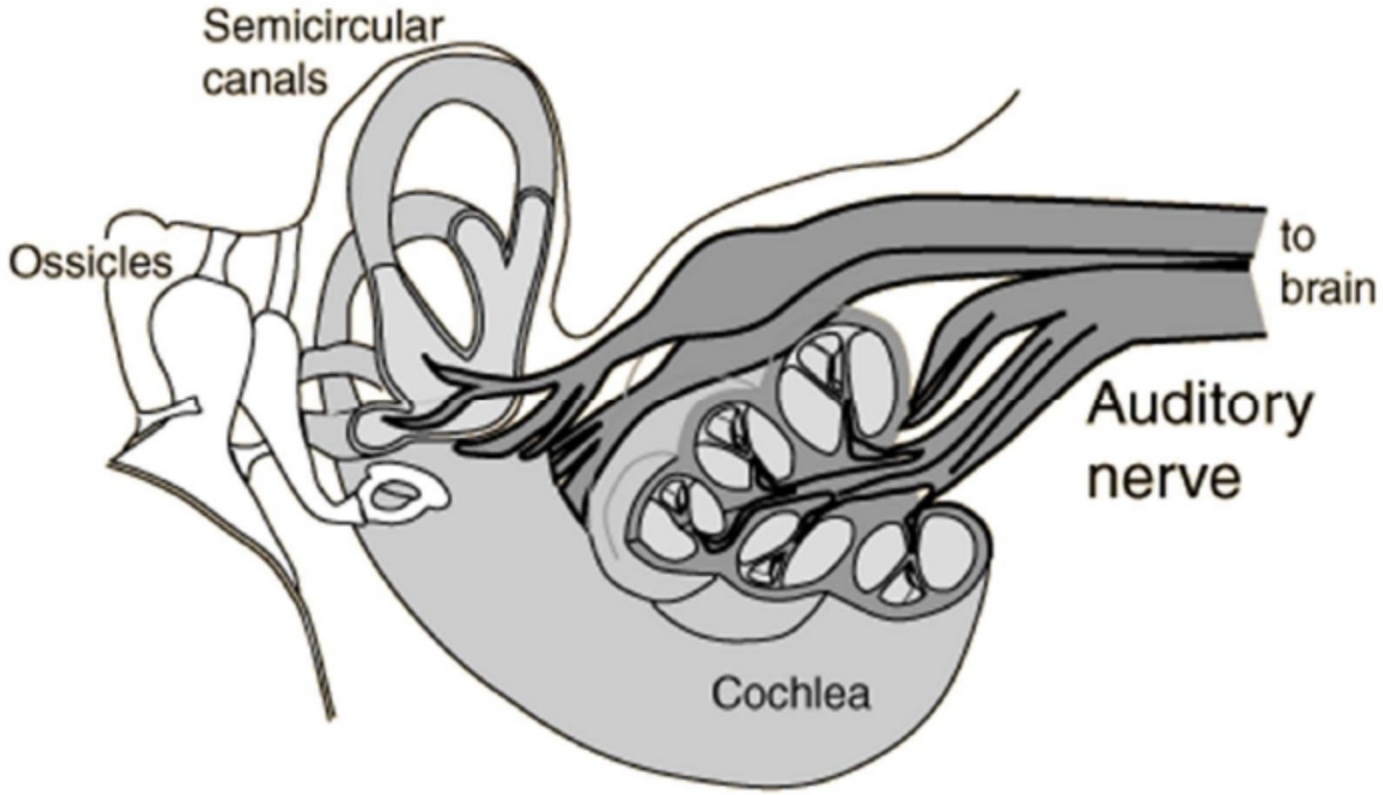


The Inner Ear



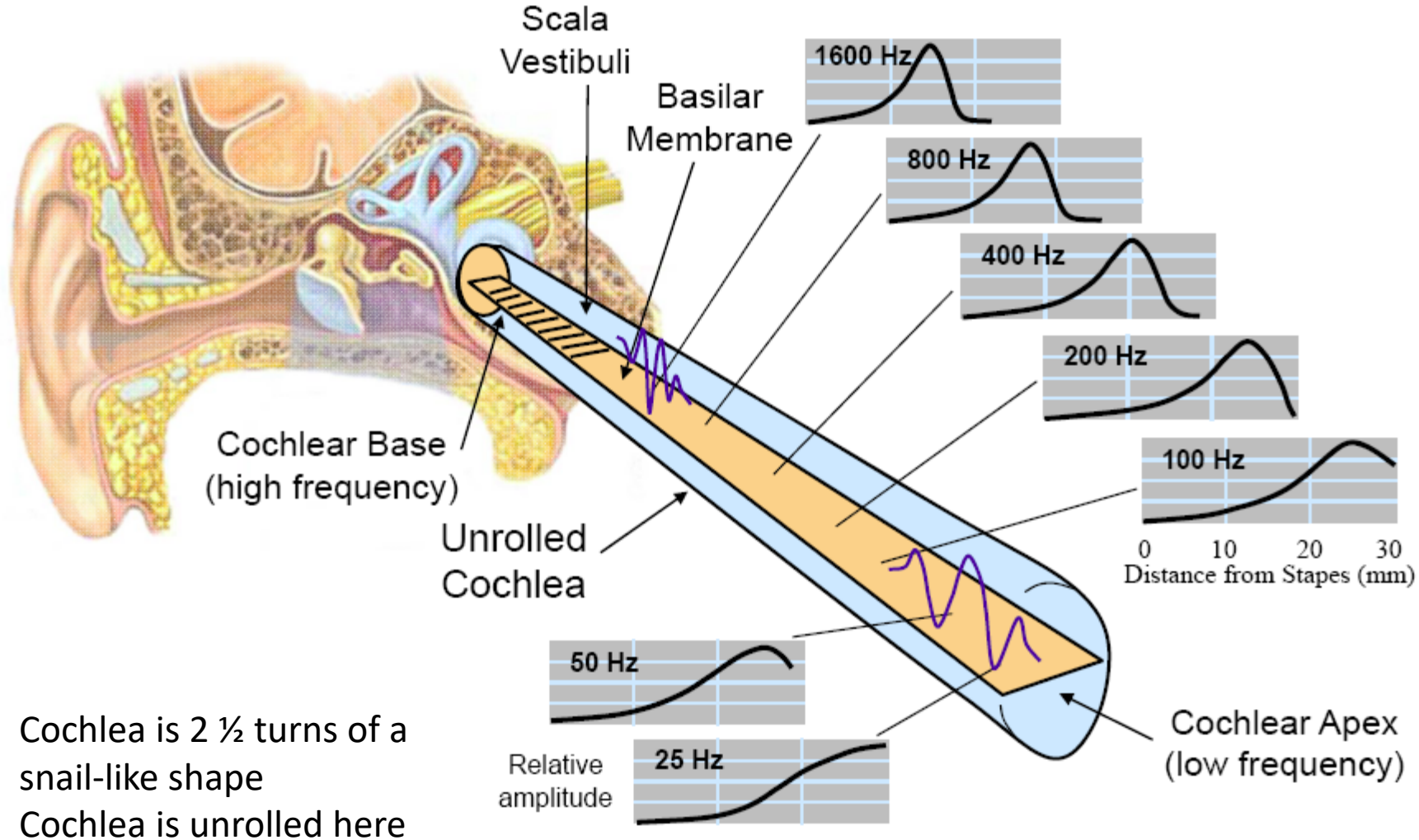
- The inner ear can be thought of as two organs, namely
  - the semicircular canals which serve as the body's balance organ and
  - the cochlea which serves as the body's microphone, converting sound pressure signals from the outer ear into electrical impulses which are passed on to the brain via the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

Stretched Cochlea & Basilar Membrane



- Cochlea is 2 ½ turns of a snail-like shape
- Cochlea is unrolled here

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 about 3000 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 Mechanics

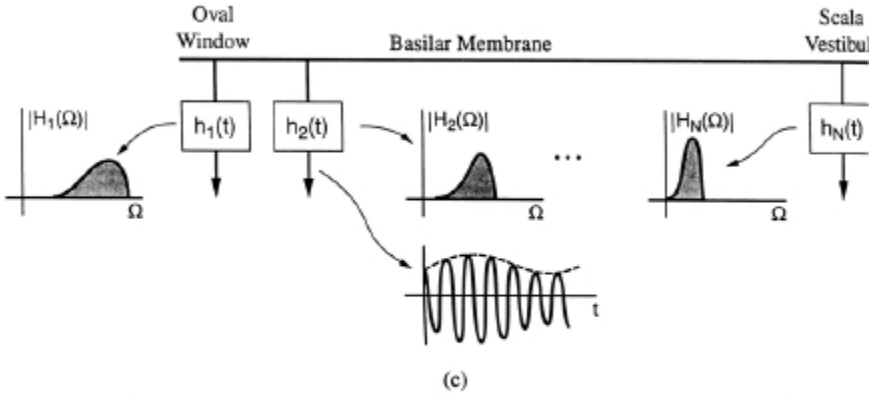
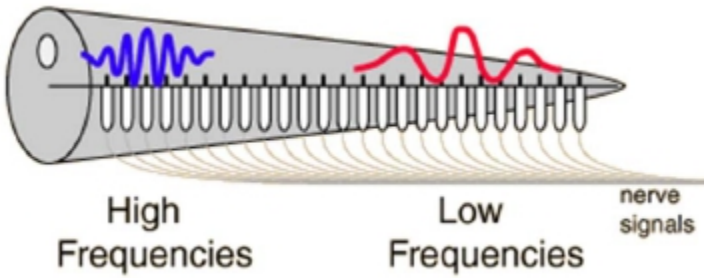


Figure 8.25 Schematic of front-end auditory processing and its model as a wavelet transform: (a) uncoiled cochlea; (b) the transduction to neural firings of the deflection of hairs that protrude from the in hair cells along the basilar membrane; (c) a signal processing abstraction of the cochlear filters 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.



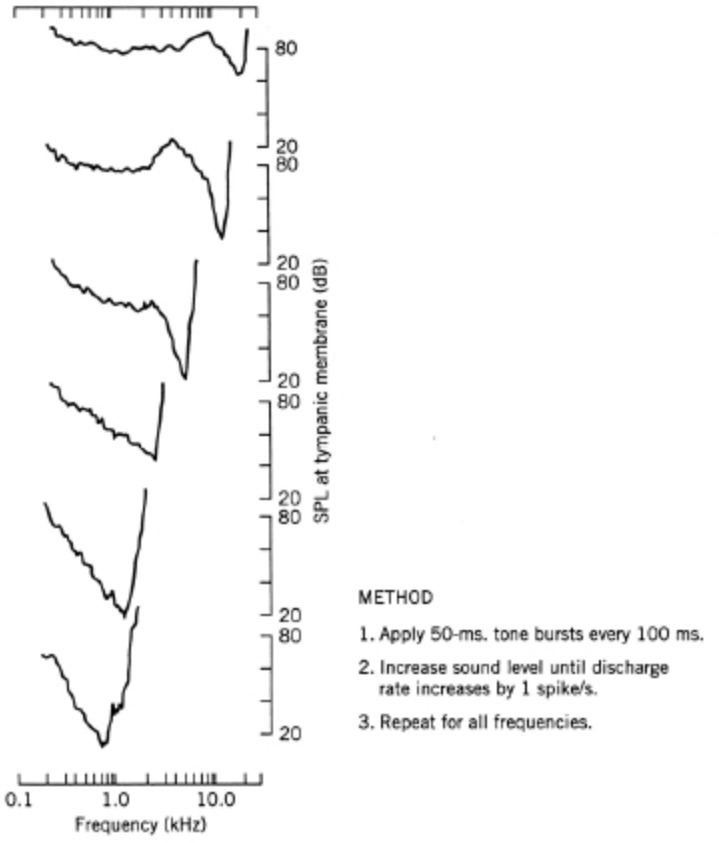


FIGURE 14.10 Tuning curves of six auditory nerve fibers. From [6].

Speech Perception

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

Sound Intensity

- Intensity (音强) 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:

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

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:

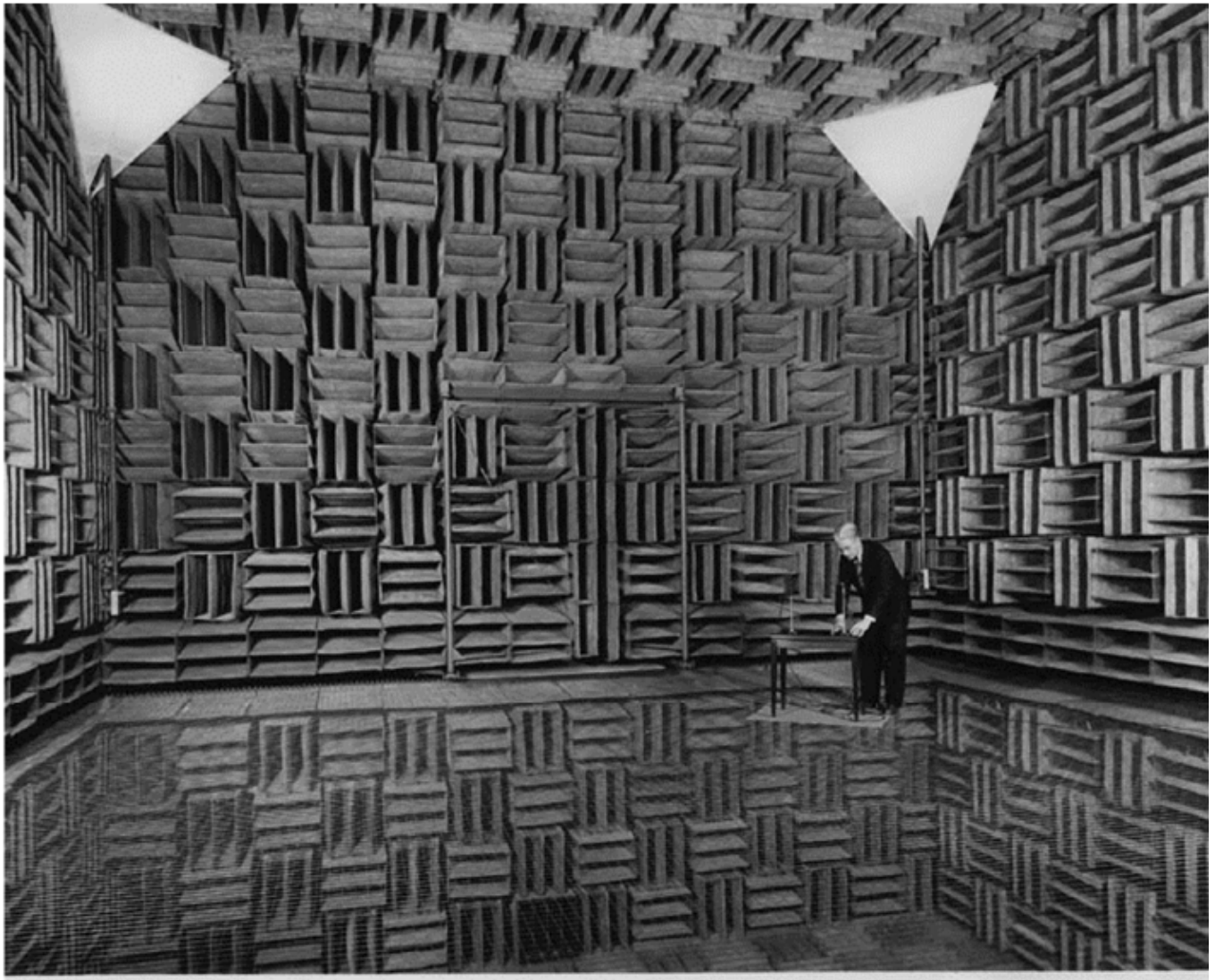
$$SPL = 10 \log_{10}\left(\frac{P^2}{P_0^2}\right) = 20 \log_{10}\left(\frac{P}{P_0}\right) \text{ dB}$$

where $P_0 = 2 \times 10^{-5}$ 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 quantization noise

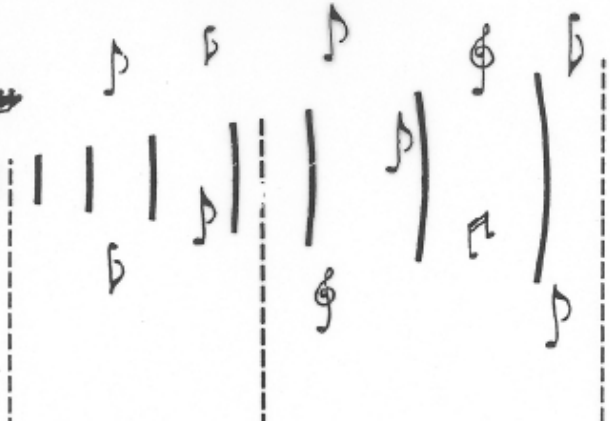
Anechoic Chamber (no Echos)

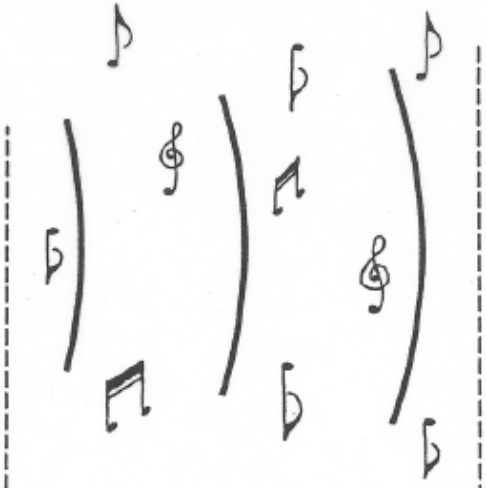


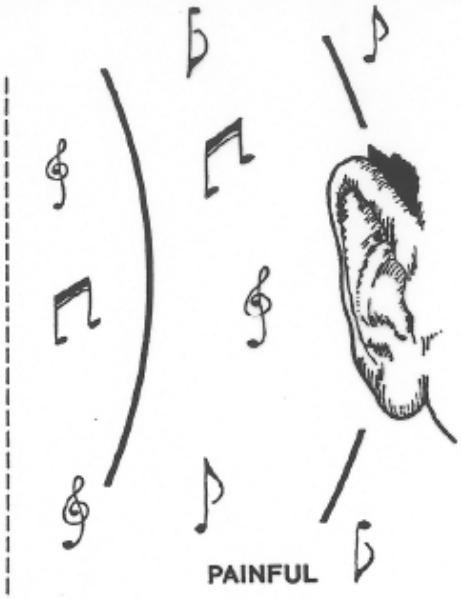
1.1.2-6 . . . . . .ANECHOIC CHAMBER TESTS 13745-11 (8/60)

Anechoic Chamber (no Echos)

Decibel Levels



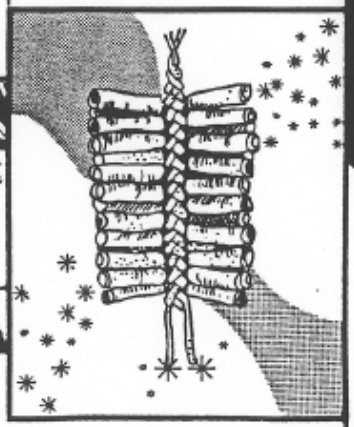




FAINT	MODERATE	VERY LOUD	EXTREMELY LOUD	PAINFUL
30–40 dB	50–70 dB	80–100 dB	110–130 dB	140–170 dB

WHISPER

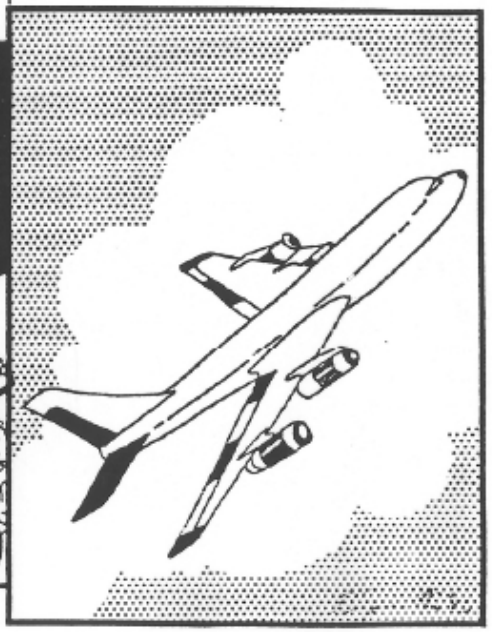
CONVERSATION



FIRE CRACKERS

(at 10 feet)

ROCK CONCERT



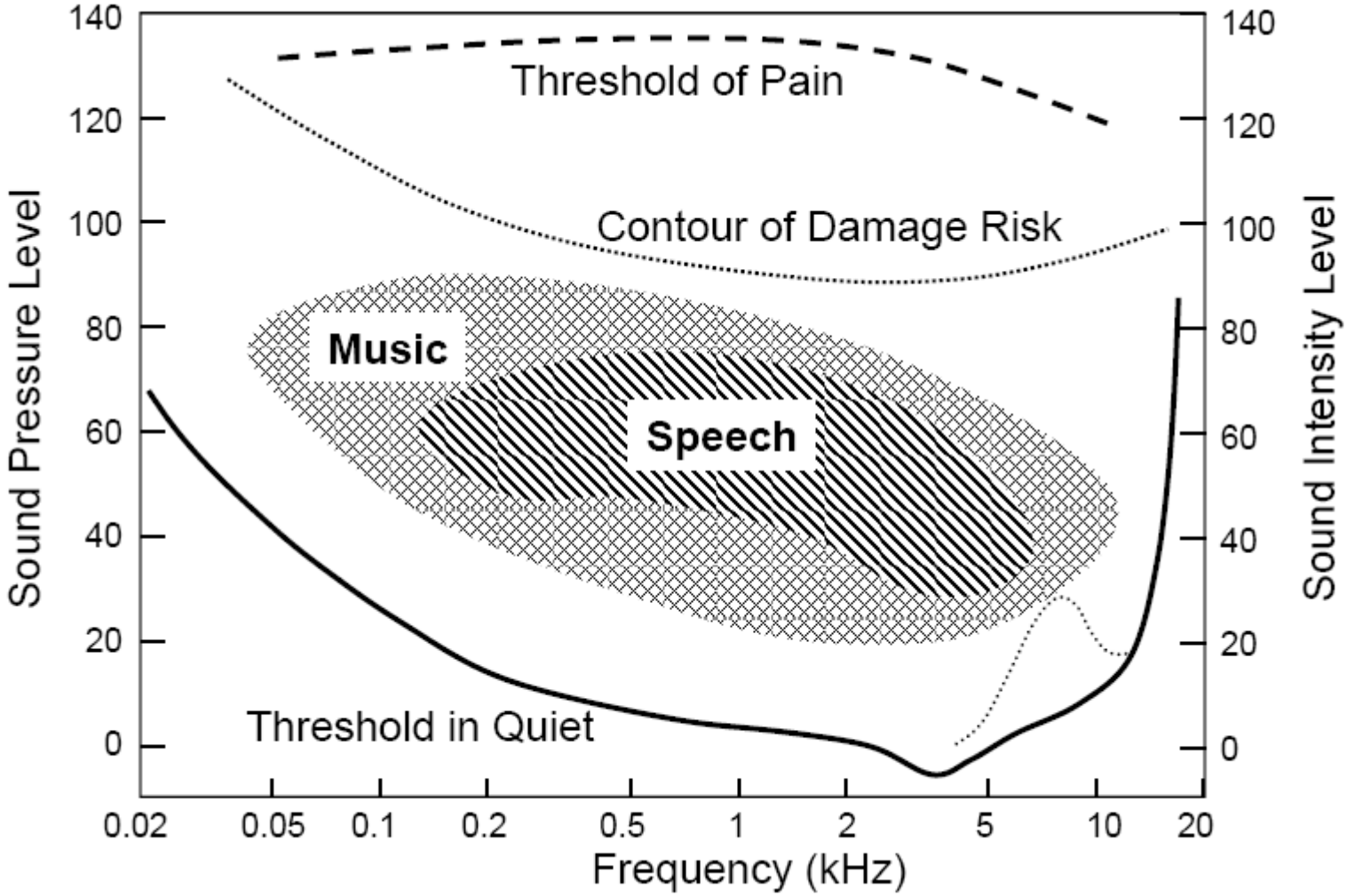
AIRPLANE

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***

Range of Human Hearing

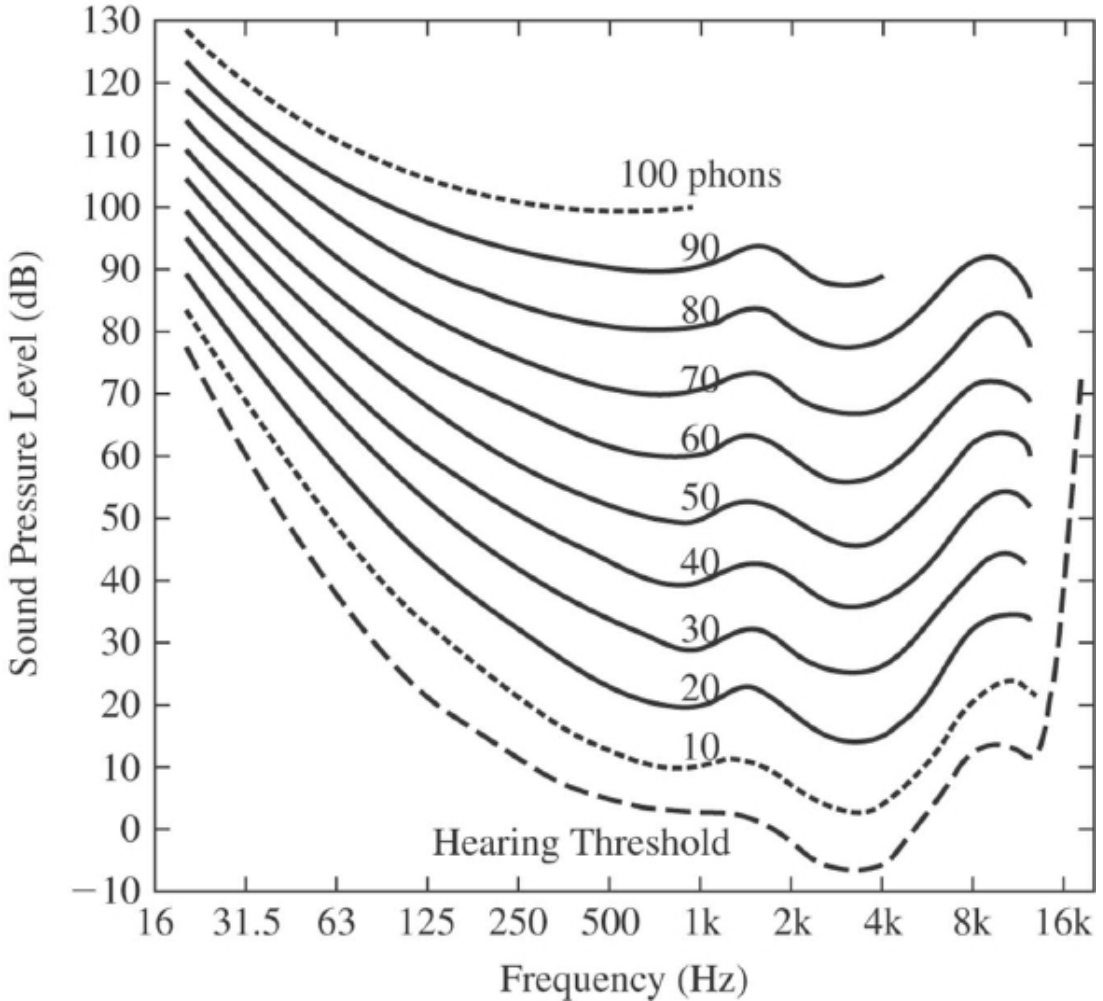


Hearing Thresholds 听阈

- 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
- Thresholds vary with frequency and from person-to-person
- Maximum sensitivity is at about 3000 Hz

Loudness Level

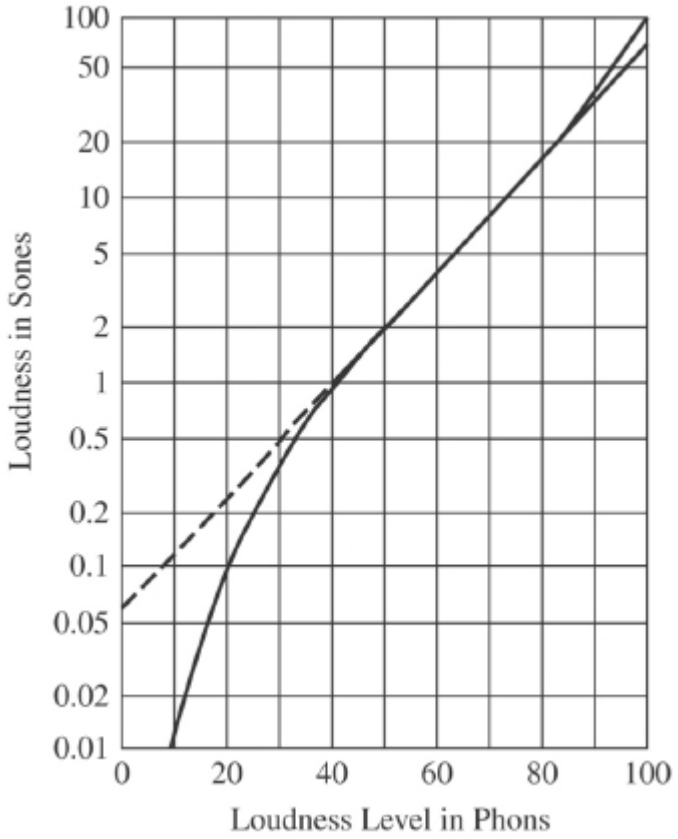
- Loudness Level (响度级 LL) is equal to the IL of a 1000 Hz tone that is judged by the average observer to be equally loud as the tone



phon 方

Loudness

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



$$L = 2^{(LL-40)/10}$$

$$\text{Log } L = 0.03(LL-40)$$

$$= 0.03LL-1.2$$

- for a frequency of 1000Hz, 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

$$\log L = 0.03(10\log I+120)-1.2$$

$$= 0.3 \log I+2.4$$

Which reduces to

$$L = 251I^{0.3}$$

Pitch

- Pitch(音高) and fundamental frequency(基频) are not the same thing
- we are quite sensitive to changes in pitch
  - *F < 500 Hz, ΔF ≈ 3 Hz*
  - *F > 500 Hz, ΔF/F ≈ 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

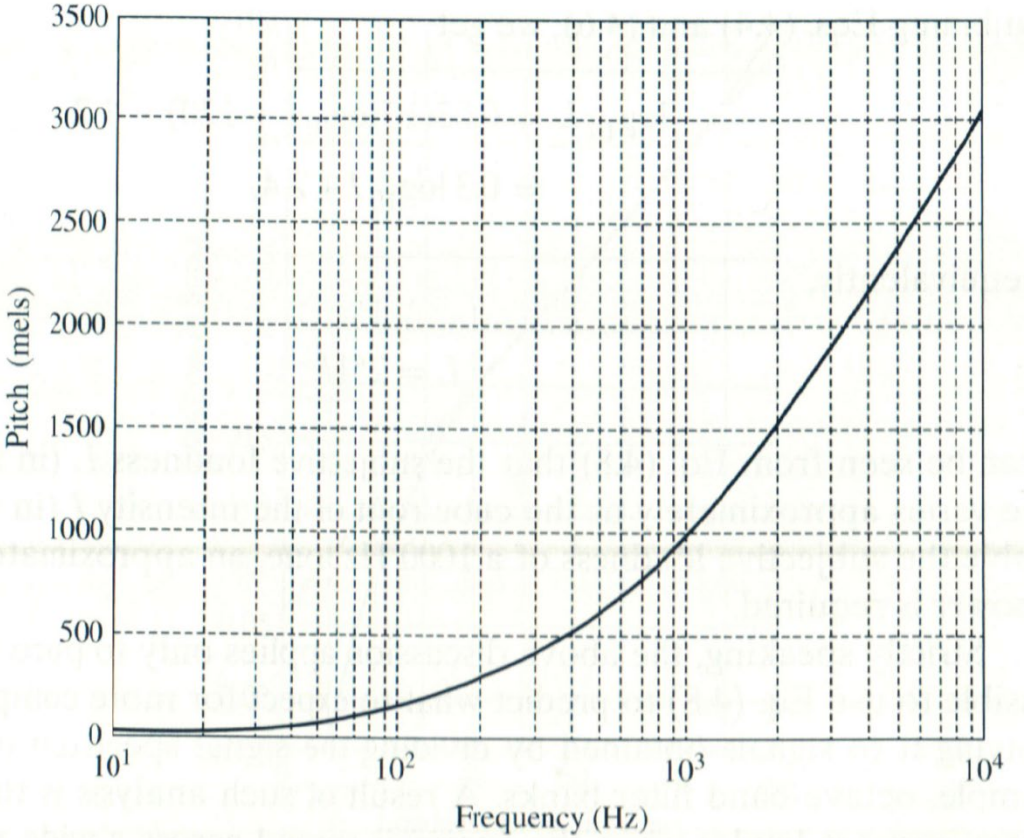


FIGURE 4.16
Chart of the subjective pitch (in mels) versus the actual frequency (in Hz) of a pure tone.

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

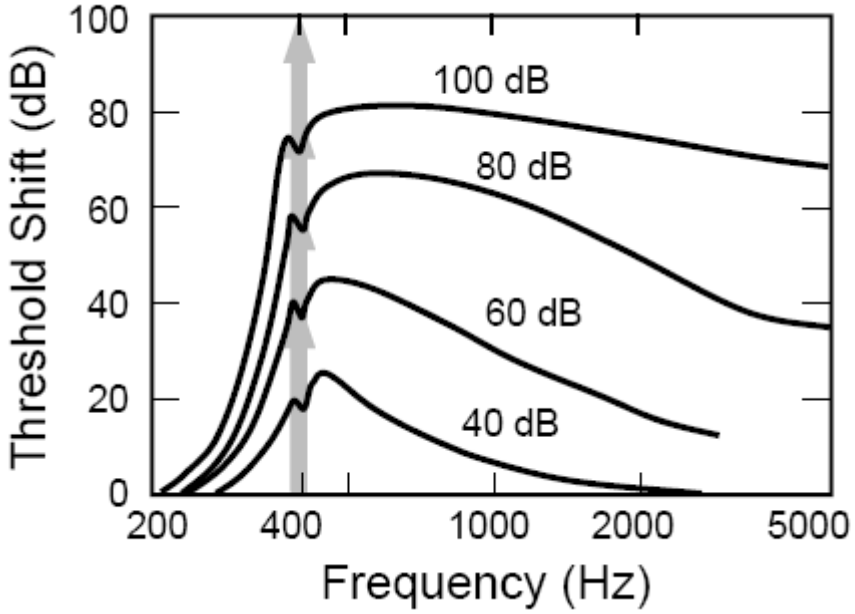
Perception of Frequency

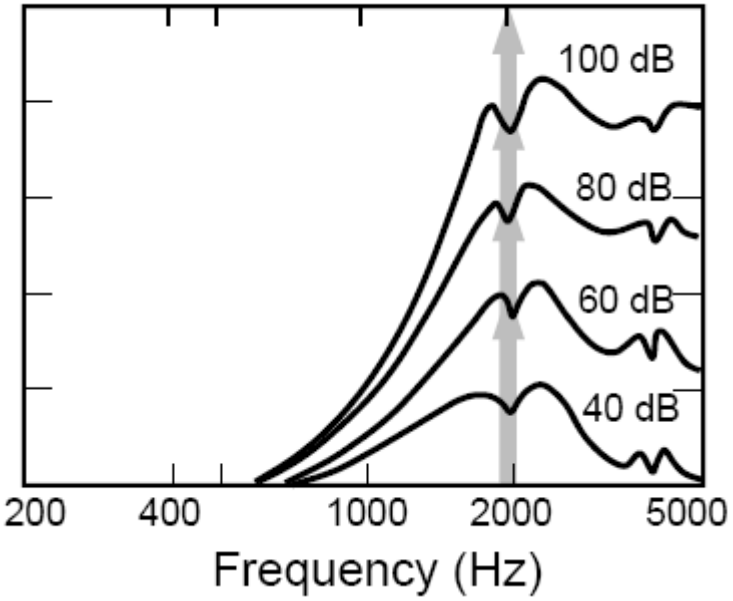
- Pure tone
  - Pitch is a perceived quantity while frequency is a physical one (cycle per second or Hertz)
  - Mel is a scale that doubles whenever the perceived pitch doubles; start with 1000 Hz = 1000 mels, increase frequency of tone until listener perceives twice the pitch (or decrease until half 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

Auditory Masking

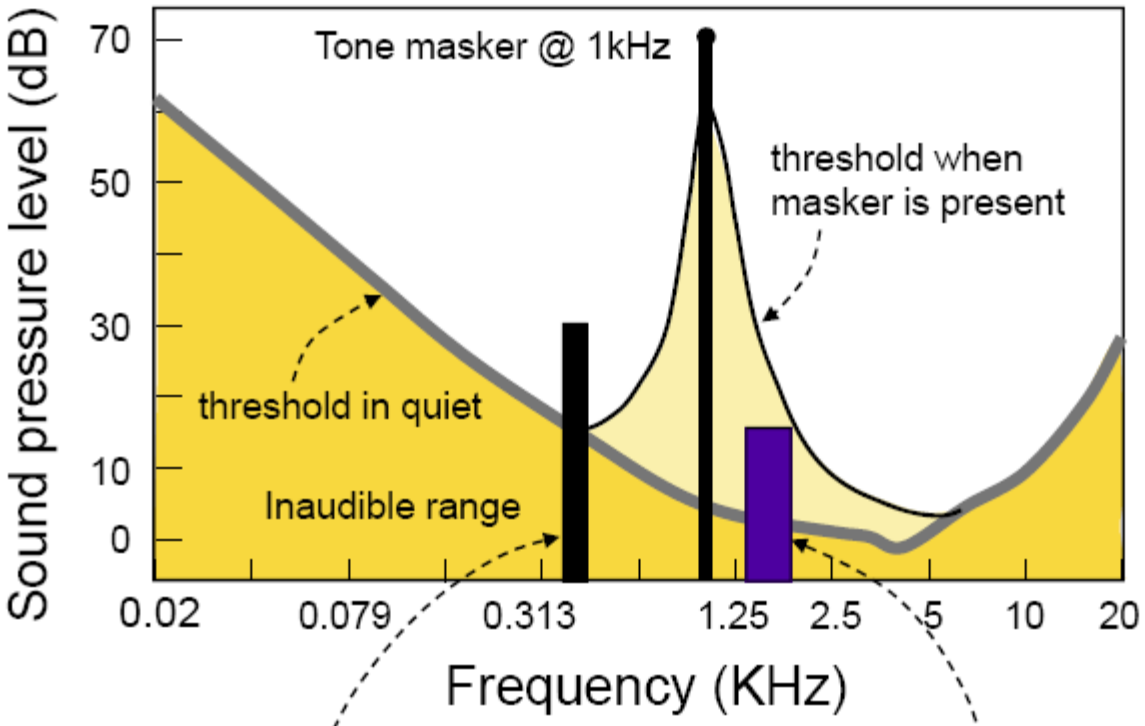
Pure Tone Masking

- Masking is the effect whereby some sounds are made less distinct or even inaudible by the presence of other sounds
- Make threshold measurements in presence of masking tone; plots below show shift of threshold over non-masking thresholds as a function of the level of the tone masker





Auditory Masking

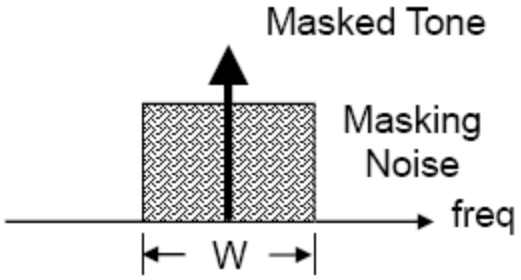


Signal perceptible even in the presence of the tone masker

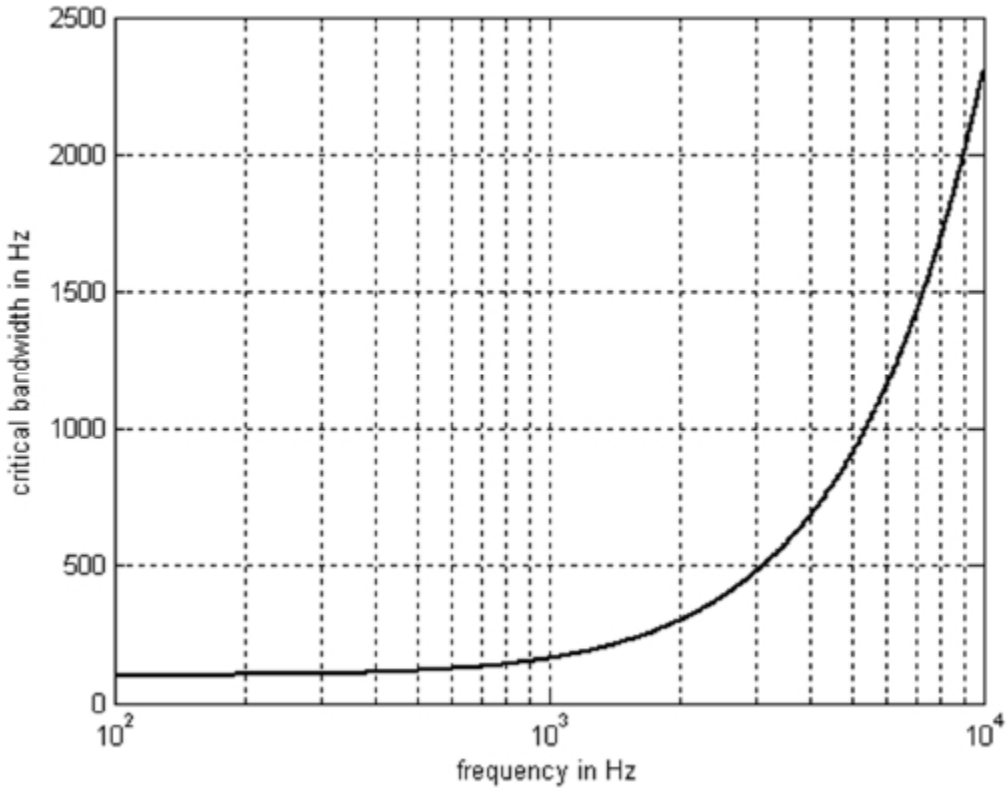
Signal not perceptible due to the presence of the tone masker

Masking and Critical Bandwidth

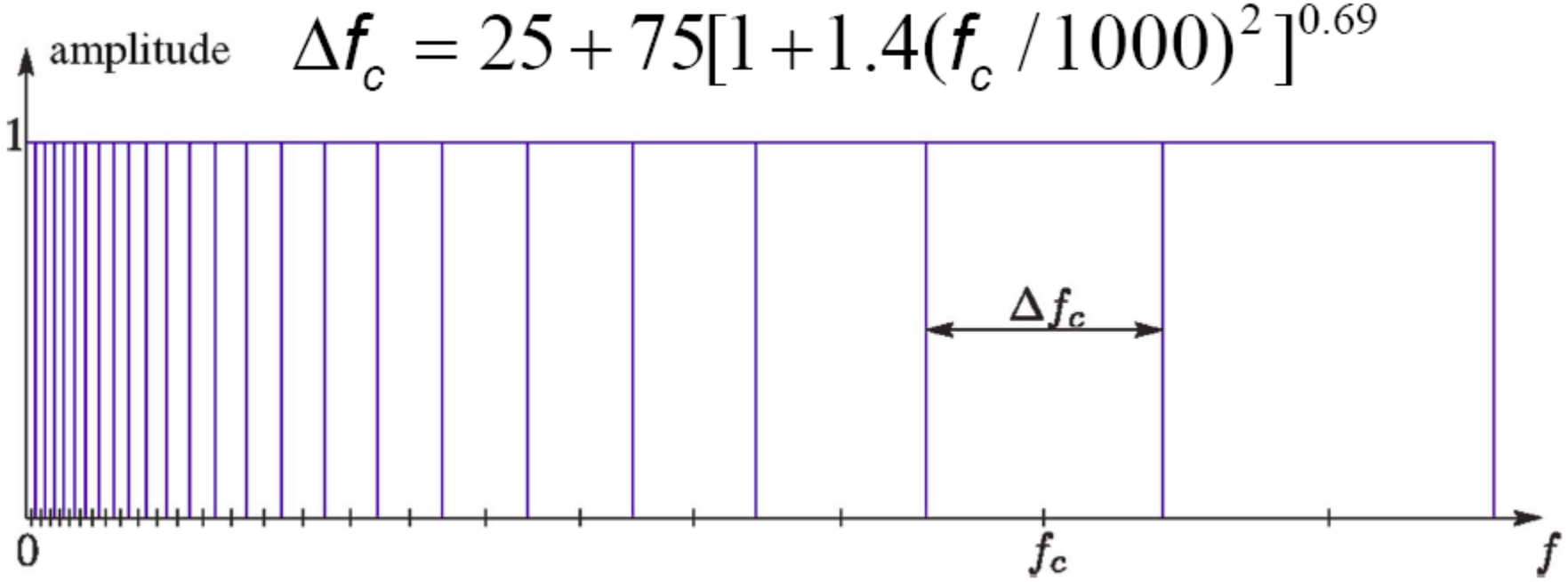
- Critical Bandwidth (临界带宽) is the bandwidth of masking noise beyond which further increase in bandwidth has little or no effect on the amount of masking of a pure tone at the center of the band



The noise spectrum used is essentially rectangular, thus the notion of equivalent rectangular bandwidth (ERB)



Critical Bands

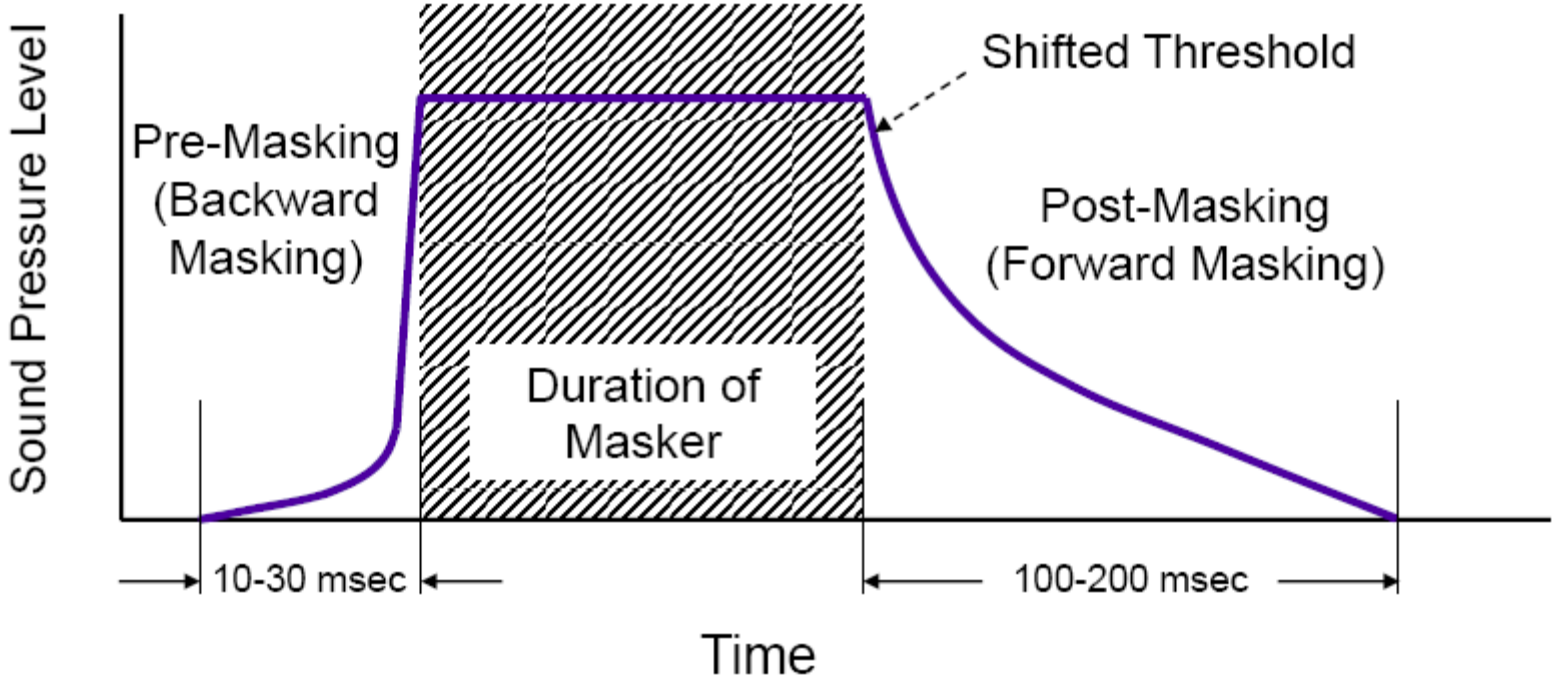


- Idealized basilar membrane filter bank
  - Center Frequency of Each Bandpass Filter: f_c
  - Bandwidth of Each Bandpass Filter: Δf_c
  - Real BM filters overlap significantly

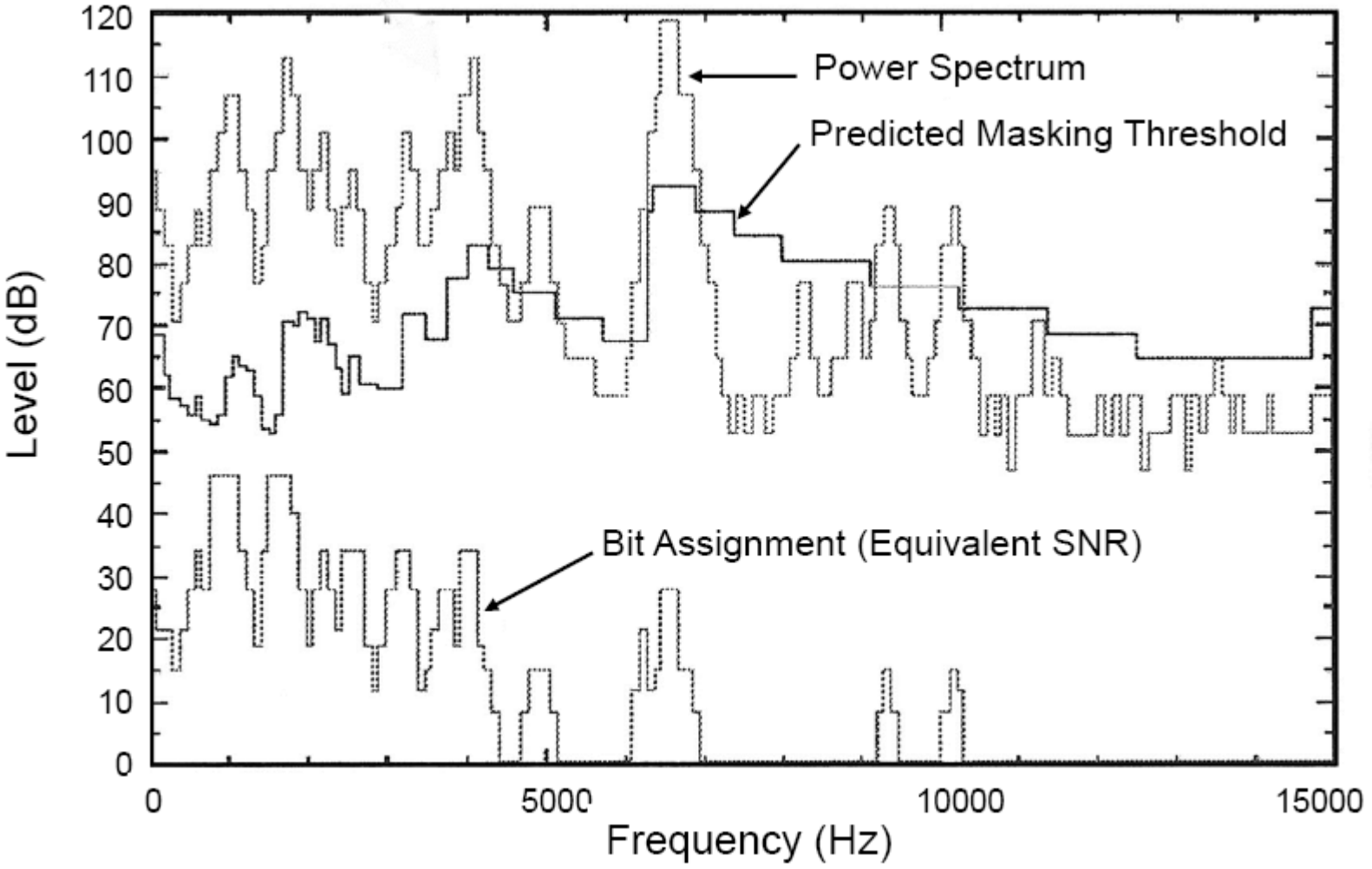
TABLE 4.2 Critical band rate z and lower (F_l) and upper (F_u) frequencies of critical bandwidths ΔF_G centered at F_c. (After Zwicker and Fastl [428].)

z	F_l,F_u	F_c	z	ΔF_G	z	F_l,F_u	F_c	z	ΔF_G
Bark	Hz	Hz	Bark	Hz	Bark	Hz	Hz	Bark	Hz
0	0				12	1720			
		50	0.5	100			1850	12.5	280
1	100				13	2000			
		150	1.5	100			2150	13.5	320
2	200				14	2320			
		250	2.5	100			2500	14.5	380
3	300				15	2700			
		350	3.5	100			2900	15.5	450
4	400				16	3150			
		450	4.5	110			3400	16.5	550
5	510				17	3700			
		570	5.5	120			4000	17.5	700
6	630				18	4400			
		700	6.5	140			4800	18.5	900
7	770				19	5300			
		840	7.5	150			5800	19.5	1100
8	920				20	6400			
		1000	8.5	160			7000	20.5	1300
9	1080				21	7700			
		1170	9.5	190			8500	21.5	1800
10	1270				22	9500			
		1370	10.5	210			10,500	22.5	2500
11	1480				23	12,000			
		1600	11.5	240			13,500	23.5	3500
12	1720				24	15,500			
		1850	12.5	280					

Temporal Masking



Exploiting Masking in Coding



Parameter Discrimination

JND – Just Noticeable Difference (最小可觉差)
Similar names: differential limen (DL), …

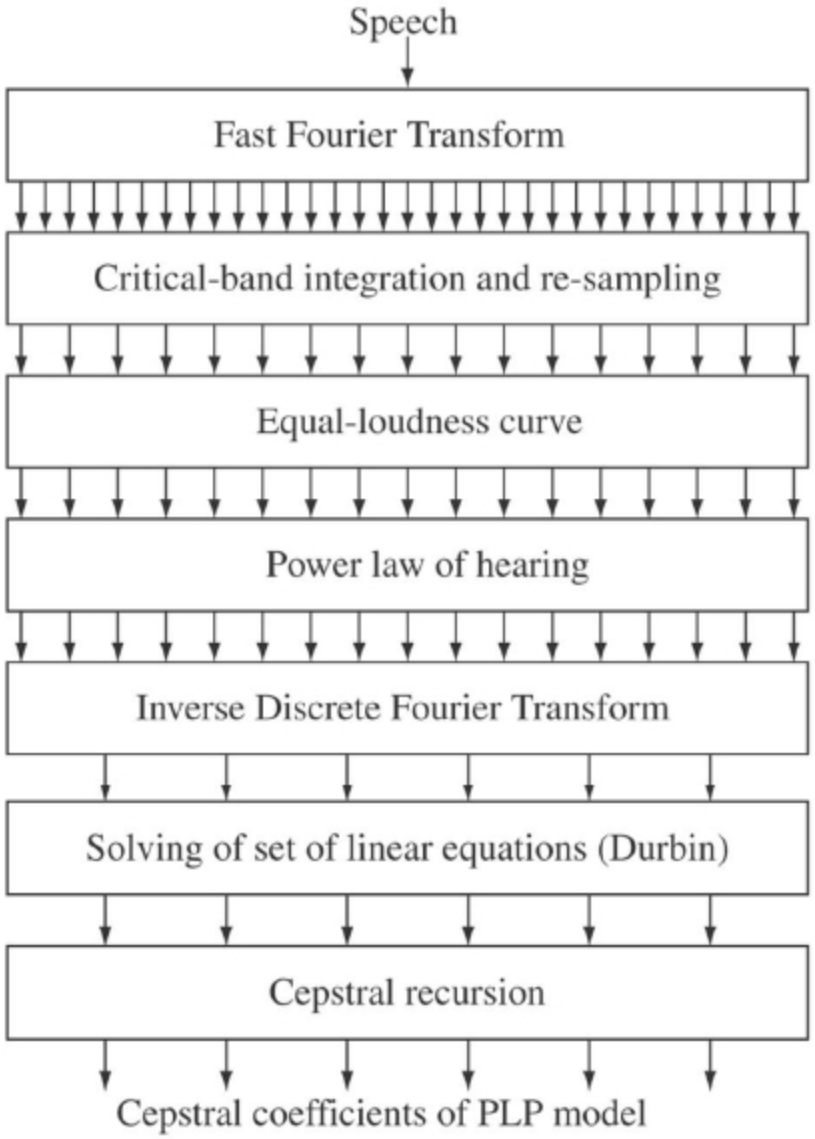
Parameter	JND/DL
Fundamental Frequency	0.3-0.5%
Formant Frequency	3-5%
Formant bandwidth	20-40%
Overall Intensity	1.5 dB

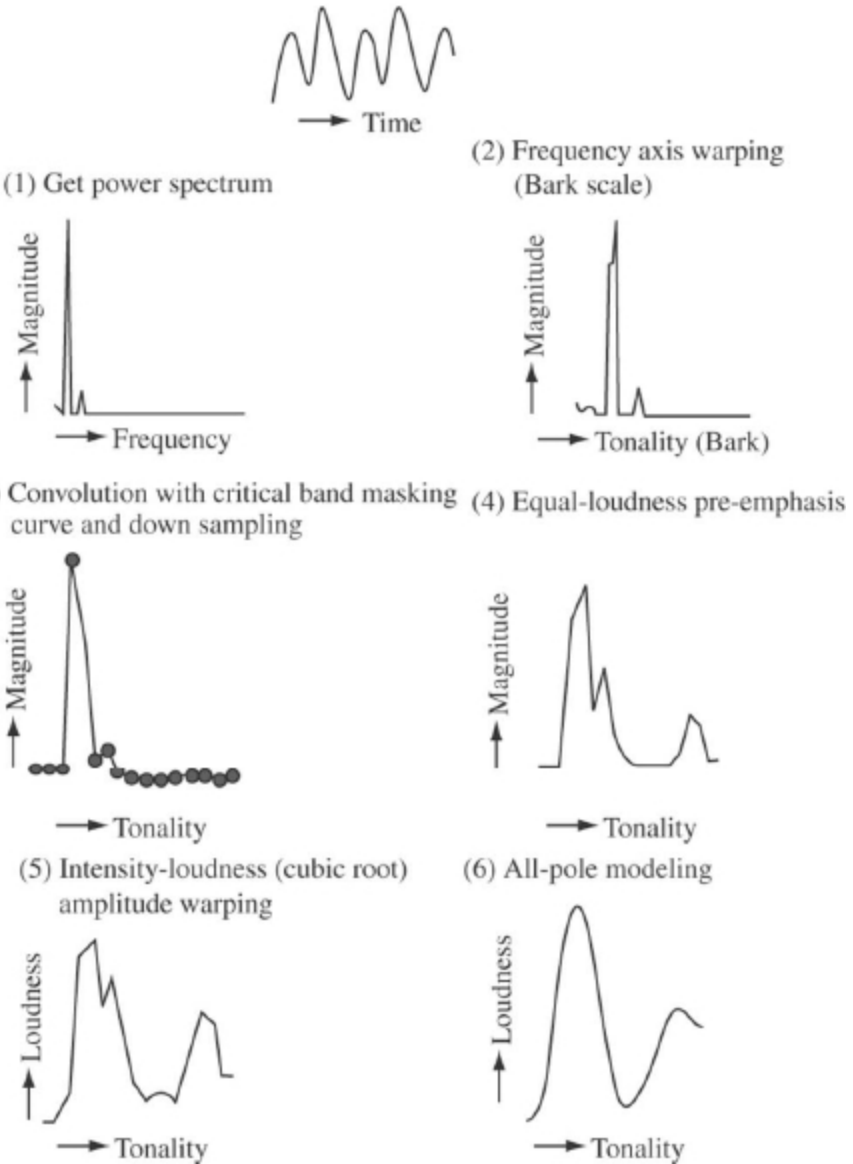
Auditory Models

Auditory Models

- Auditory models
  - To predict auditory phenomena for speech applications
- 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)

Perceptual Linear Prediction





Perceptual Linear Prediction

- Included perceptual effects in PLP
  - critical band spectral analysis using a Bark frequency scale with variable bandwidth trapezoidal shaped filters
  - asymmetric auditory filters with a 25 dB/Bark slope at the high frequency cutoff and a 10 dB/Bark slope at the low frequency cutoff
  - use of the equal loudness contour to approximate unequal sensitivity of human hearing to different frequency components of the signal
  - use of the non-linear relationship between sound intensity and perceived loudness using a cubic root compression method on the spectral levels
  - a method of broader than critical band integration of frequency bands based on an autoregressive, all-pole model utilizing a fifth order analysis

Human Speech Perception Experiments

Sound Perception in Noise

	p	*t*	*k*	*f*	θ	*s*	ʃ	*b*	*d*	*g*	ν	δ	*z*	ʒ	*m*	*n*
p	240		41	2	1											
t	1	252	1	1						1						
k	18	3	219													
f				225	24			5			2					
θ	9		1	69	185			3				1				
s						232										
ʃ							236									
b					1			242			24	12	1			
d									213	22			1			
g					1				33	203		3				
ν								6			171	30			1	
δ					1			1		3	22	208	4			1
z									2	4	1	7	238			
ʒ														244		
m												1			274	1
n																252

FIGURE 17.4 Confusion matrix for S/N = +12 dB and a frequency response of 200–6500 Hz. From [13].

Confusions as to sound PLACE, not MANNER

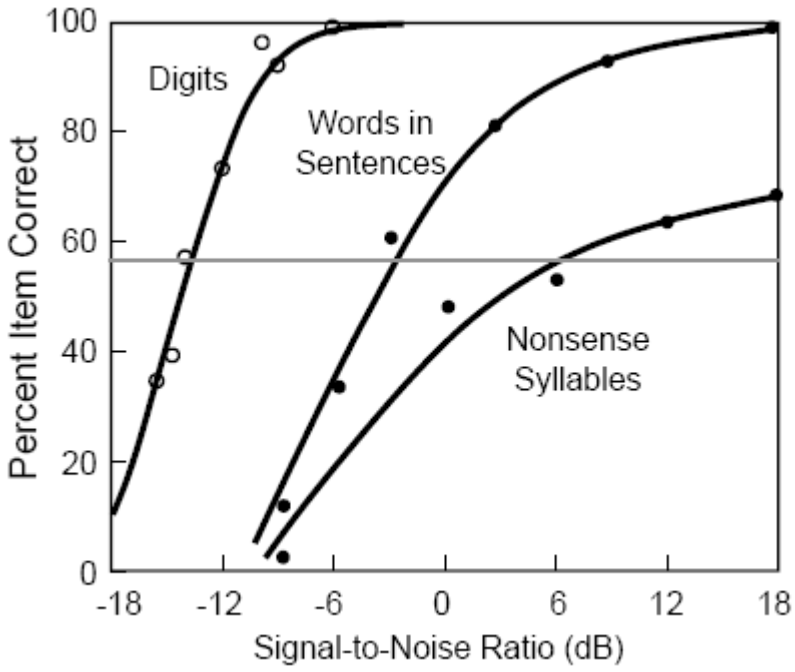
Sound Perception in Noise

	p	*t*	*k*	*f*	θ	*s*	ʃ	*b*	*d*	*g*	τ	δ	*z*	3	*m*	*n*
p	80	43	64	17	14	6	2	1	1			1				
t	71	84	55	5	9	3	8	1				1	2			
k	66	76	107	12	8	9	4					1				
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	65	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		
ʃ	1	6	3	4	6	29	195		3							
b	1			5	4	4		136	10	9	47	16	6	1		
d							8	5	80	45	11	20	20	26		
g					2			3	63	66	3	19	37	56		
τ				2		2		48	5	5	145	45	12			
δ					6			31	6	17	86	58	21	5		
z					1	1	1	7	20	27	16	28	94	44		
3								1	26	18	3	8	45	129		
m	1							4			4	1	3		17	
n					4			1	5	2		7	1	6	4	

FIGURE 17.5 Confusion matrix for S/N = −6 dB and a frequency response of 200–6500 Hz. From [13].

Confusions in both sound PLACE and MANNER

Speech Perception



50% S/N level for correct responses:

- -14 db for digits
- -4 db for major words
- +3 db for nonsense syllables

- Speech Perception depends on multiple factors including the perception of individual sounds (based on distinctive features) and the predictability of the message (think of the message that comes to mind when you hear the preamble 'To be or not to be …', or 'Four score and seven years ago …')
- the importance of linguistic and contextual structure cannot be overestimated (e.g., the Shannon Game where you try to predict the next word in a sentence i.e., 'he went to the refrigerator and took out a …' where words like plum, potato etc are far more likely than words like book, painting etc.)

Word Intelligibility

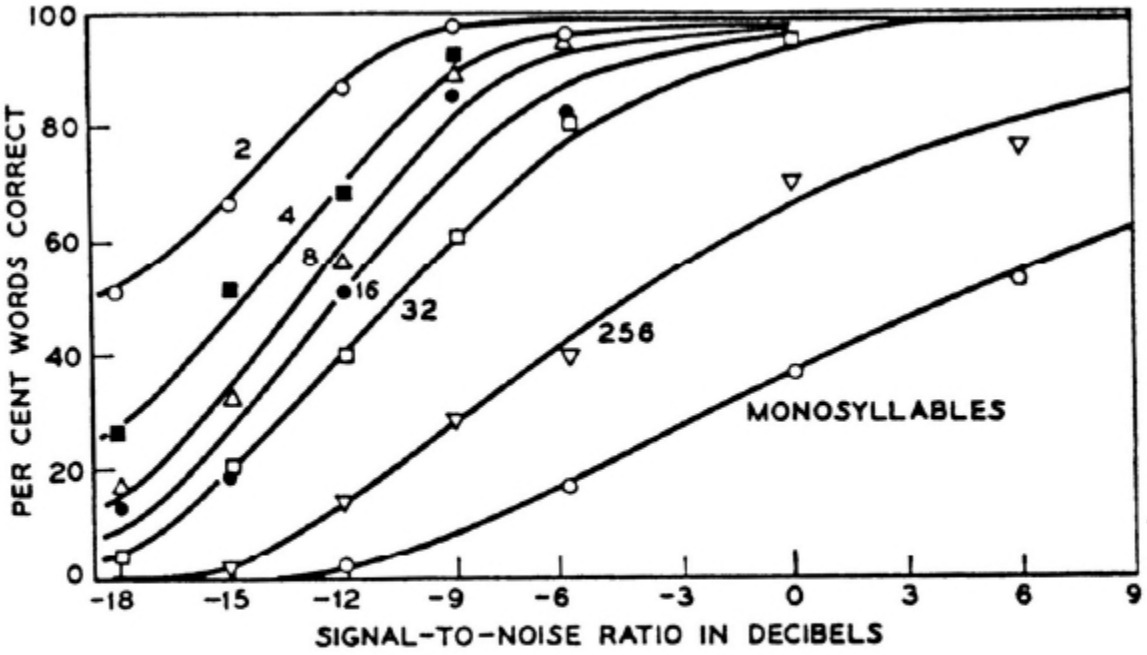


Fig. 7.22. Effects of vocabulary size upon the intelligibility of monosyllabic words. (After MILLER, HEISE and LICHTEN)

Intelligibility - Diagnostic Rhyme Test (诊断押韵测试)

Voicing		Nasality		Sustenation		Sibilation		Graveness		Compactness	
veal	feel	meat	beat	vee	bee	zee	thee	weed	reed	yield	wield
bean	peen	need	deed	sheet	cheat	cheep	keep	peak	teak	key	tea
gin	chin	mitt	bit	vill	bill	jilt	gilt	bid	did	hit	fit
dint	tint	nip	dip	thick	tick	sing	thing	fin	thin	gill	dill
zoo	sue	moot	boot	foo	pooh	juice	goose	moon	noon	coop	poop
dune	tune	news	dues	shoes	choose	chew	coo	pool	tool	you	rue
vole	foal	moan	bone	those	doze	joe	go	bowl	dole	ghost	boast
goat	coat	note	dote	though	dough	sole	thole	fore	thor	show	so
zed	said	mend	bend	then	den	jest	guest	met	net	keg	peg
dense	tense	neck	deck	fence	pence	chair	care	pent	tent	yen	wren
vast	fast	mad	bad	than	dan	jab	gab	bank	dank	gat	bat
gaff	calf	nab	dab	shad	chad	sank	thank	fad	thad	shag	sag
vault	fault	moss	boss	thong	tong	jaws	gauze	fought	thought	yawl	wall
daunt	taunt	gnaw	daw	shaw	chaw	saw	thaw	bong	dong	caught	thought
jock	chock	mom	bomb	von	bon	jot	got	wad	rod	hop	fop
bond	pond	knock	dock	vox	box	chop	cop	pot	tot	got	dot

$$DRT = 100 \times \frac{R_d - W_d}{T_d}$$

R = right
W = wrong
T = total
d = one of the six speech dimensions.

Coder	Rate (kb/s)	Male	Female	All	MOS
FS1016	4.8	94.4	89.0	91.7	3.3
IS54	7.95	95.2	91.4	93.3	3.6
GSM	13	94.7	90.7	92.7	3.6
G.728	16	95.1	90.9	93.0	3.9

Quantification of Subjective Quality

Absolute category rating (ACR) – MOS, mean opinion score

Quality description	Rating
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Degradation category rating (DCR) – D(egradation)MOS; need to play reference

Quality description	Rating
Degradation not perceived	5
.. perceived but not annoying	4
.. slightly annoying	3
.. annoying	2
.. very annoying	1

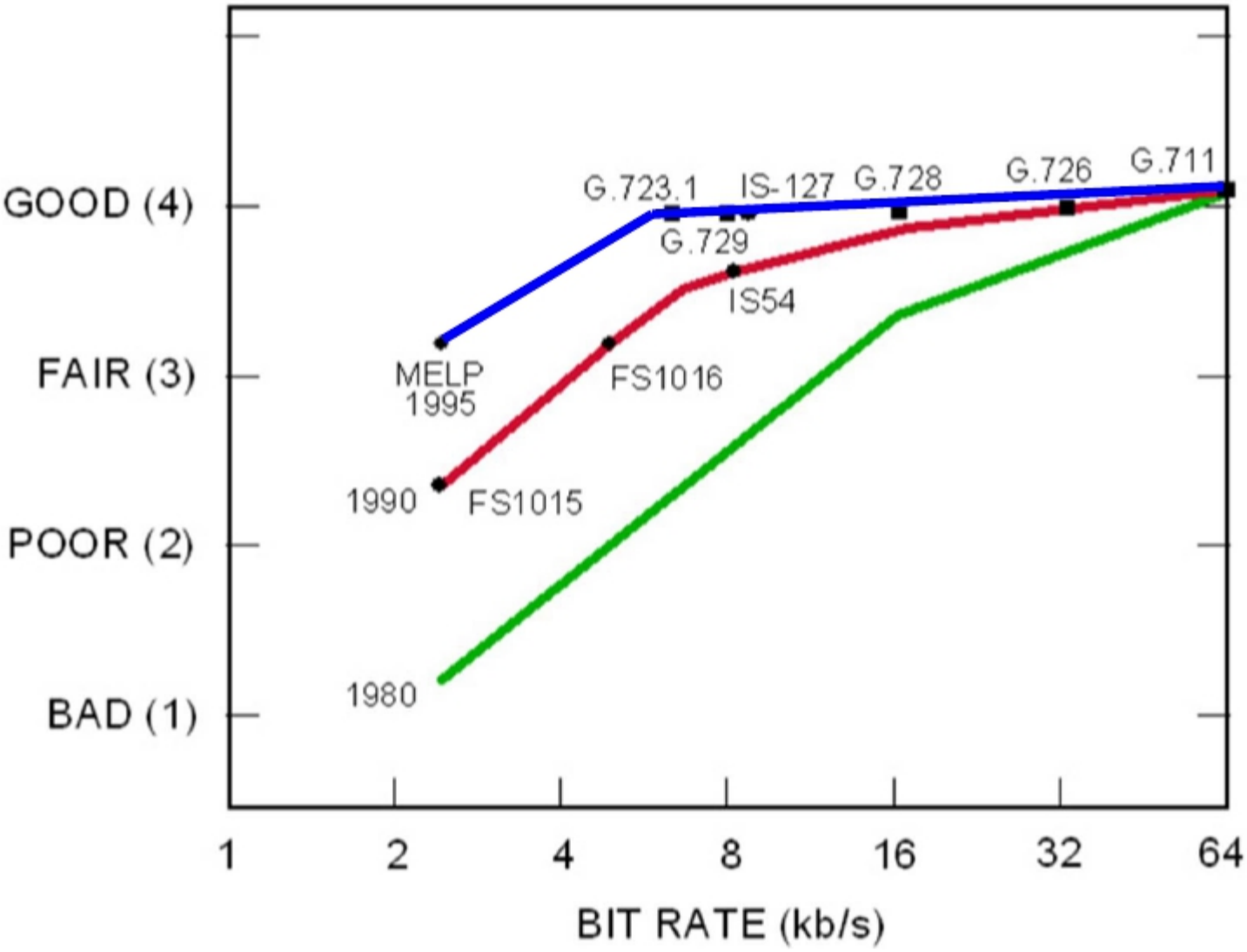
Comparison category rating (CCR) – randomized (A,B) test

Description	Rating
Much better	3
Better	2
Slightly better	1
About the same	0
Slightly worse	-1
Worse	-2
Much worse	-3

MOS (Mean Opinion Scores 平均意见分)

- Why MOS:
  - SNR is just not good enough as a subjective measure for most coders (especially model-based coders where waveform is not preserved inherently)
  - noise is not simple white (uncorrelated) noise
  - error is signal correlated
    - clicks/transients
    - frequency dependent spectrum—not white
    - includes components due to reverberation and echo
    - noise comes from at least two sources, namely quantization and background noise
    - delay due to transmission, block coding, processing
    - transmission bit errors—can use Unequal Protection Methods
    - tandem encodings

MOS for Range of Speech Coders



Lecture Summary

- the ear acts as a sound canal, transducer, spectrum analyzer
- the cochlea acts like a multi-channel, logarithmically spaced, constant Q filter bank
- frequency and place along the basilar membrane are represented by inner hair cell transduction to events (ensemble intervals) that are processed by the brain
  - this makes sound highly robust to noise and echo
- hearing has an enormous range from threshold of audibility to threshold of pain
  - perceptual attributes scale differently from physical attributes—e.g., loudness, pitch
- masking enables tones or noise to hide tones or noise => this is the basis for perceptual coding (MP3)
- perception and intelligibility are tough concepts to quantify—but they are key to understanding performance of speech processing systems