

Primer on Auditory Processing

Mounya Elhilali
Department of Electrical & Computer Engineering
Johns Hopkins University
mounya@jhu.edu

601.467/667 Introduction to Human Language Technology

1

Topics

1. Speech signals as waves
   - Physical/Perceptual dimensions of speech waves
2. How do we perceive sounds (the auditory system)
3. How is speech produced (the production system)

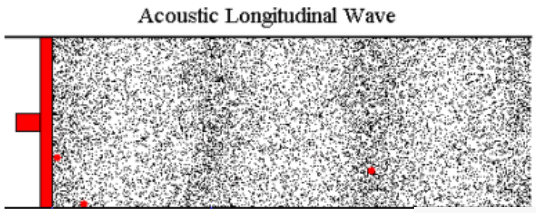
2

Speech as waves

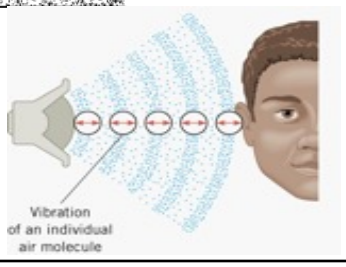
3

Sound is a wave

- Sound is a mechanical wave caused by a vibrating source
- The vibrating source that causes the matter around it to move



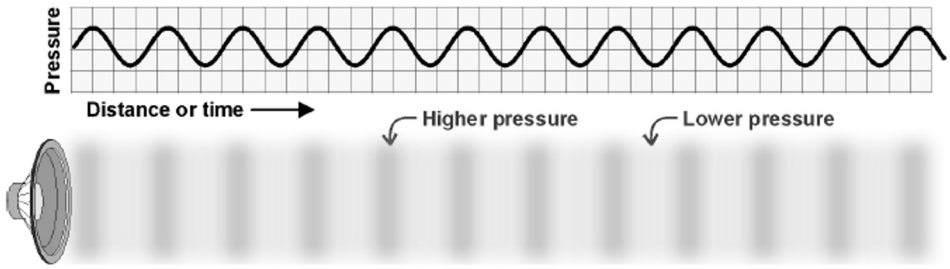
- No sound is produced in a vacuum
  - Matter (air, water, earth) must be present
- Individual air molecules do not move the wave. A given molecule vibrates forth about a fixed location.



4

Sound waves

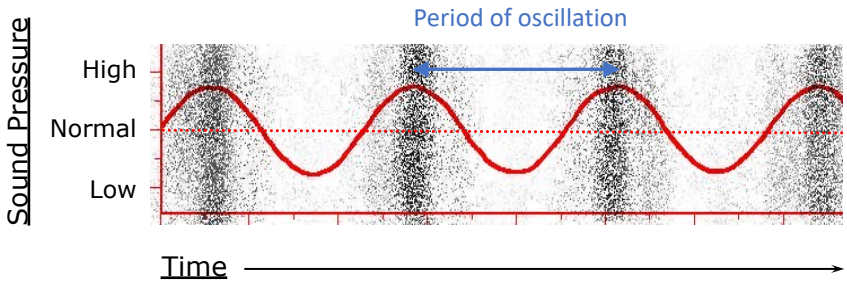
A sound wave is a wave of alternating high-pressure and low-pressure regions of air.



- Vibrating object compresses the air around it (high pressure)
- Pushes air away leaving an area of low pressure (low pressure)
- then compresses again creating a periodic pattern

5

Sound waves

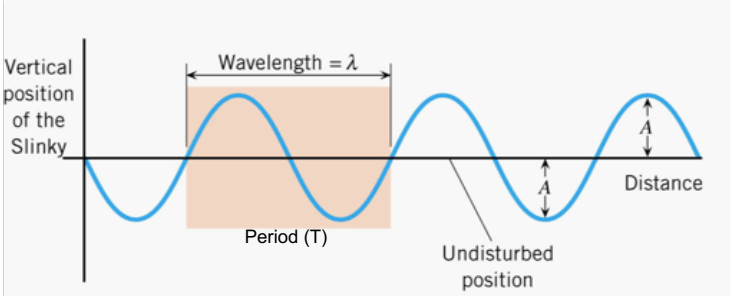


- Motion air particles do not travel, they oscillate around a point in space
- The rate of oscillation is called frequency (f)
  - ✓ denoted in cycles per second (cps) or hertz (Hz).

6

6

Physical Dimensions of Sound



Wavelength (λ)

Distance traveled by one cycle

Affected by medium (how sound travels)

Amplitude

Height of a cycle

Related to measure/perception of loudness

Frequency (F)

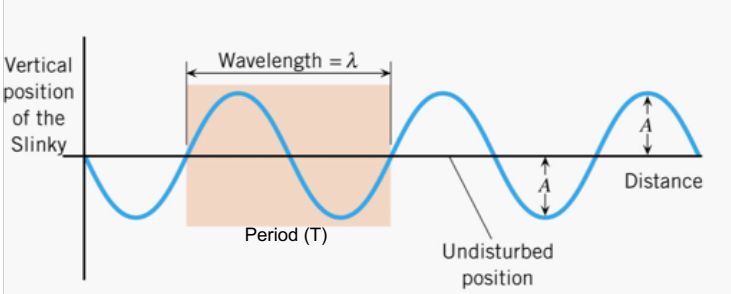
Cycles per second

Related to measure/perception of pitch/spectrum

7

7

Physical Dimensions of Sound



Wavelength (λ)

Distance traveled by one cycle

Affected by medium (how sound travels)

Amplitude

Height of a cycle

Related to measure/perception of loudness

Frequency (F)

Cycles per second

Related to measure/perception of pitch/spectrum

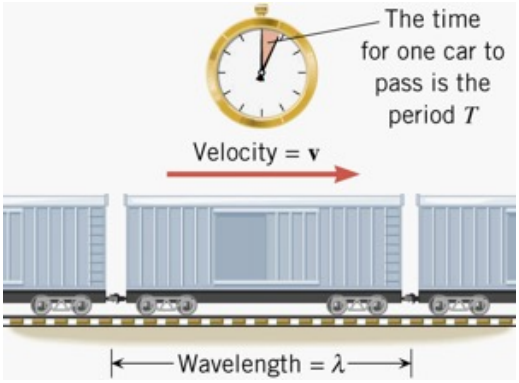
8

8

Physical Dimensions: Wavelength

- **Wavelength** is the distance over which one wave cycle repeats
- It is measured in meters (m)
- Relationship between period (T), velocity (v) and wavelength (λ):

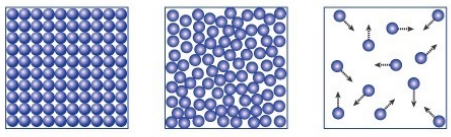
$$\text{velocity} = \lambda F$$



9

Effect of medium

- Sound velocity changes very much in different media.
  - It travels faster in non-porous solids than in air (molecules in solids are closer together)
  - It is the "compression" that travels not the medium, *per se*.



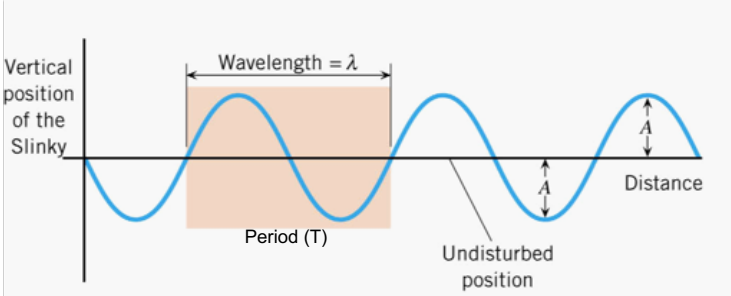
Medium	Sound Speed (metres/second)
Gas	
Carbon dioxide	268
Air	331
Helium	972
Liquid	
Ethanol	1,130
Fresh water	1,402
Sea water	1,522
Solid	
Brass	4,700
Steel	5,790
Aluminum	6,420

* Slightly colder than ambient temperature (20◦)

10

10

Physical Dimensions of Sound



Wavelength (λ)

Distance traveled by one cycle

Affected by medium (how sound travels)

Amplitude

Height of a cycle

Related to measure/perception of loudness

Frequency (F)

Cycles per second

Related to measure/perception of pitch/spectrum

11

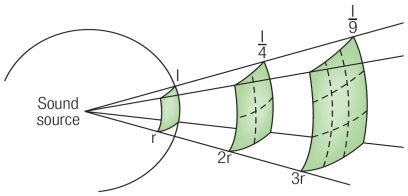
11

Sound Amplitude

- Sound amplitude is measured as intensity
  - *Intensity = rate of energy flow over an area*

$$I = \frac{\text{Power}}{\text{Area}} = \frac{\text{Energy}/\text{Time}}{\text{Area}} = \frac{\text{Power}}{4\pi r^2} \left(\text{Watts}/\text{Meter}^2\right)$$

Inverse square law



  - Sounds get fainter away from a source [same power, bigger area]

12

Amplitude ↔ Loudness

- Sound Pressure Level (SPL) is a relative measure of sound intensity

$$L = 10log_{10}\frac{I}{I_0}$$

Intensity of target sound (*rate of energy flow over an area in* W/m^2)

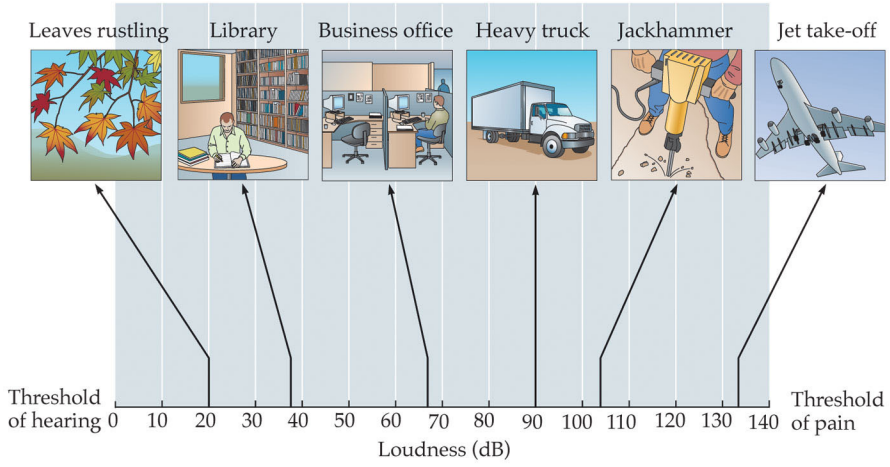
Reference Intensity:
Human absolute hearing threshold
$I_0 = 1 \times 10^{-12}\ (W/m^2)$
(related to $P_0 = 20\ \mu Pa$)

- SPL is technically a unitless measure; but uses unit of Decibels or dB or dB-SPL
- Decibels provide a **relative** measure of sound intensity.

13

13

Sound pressure level of everyday sounds

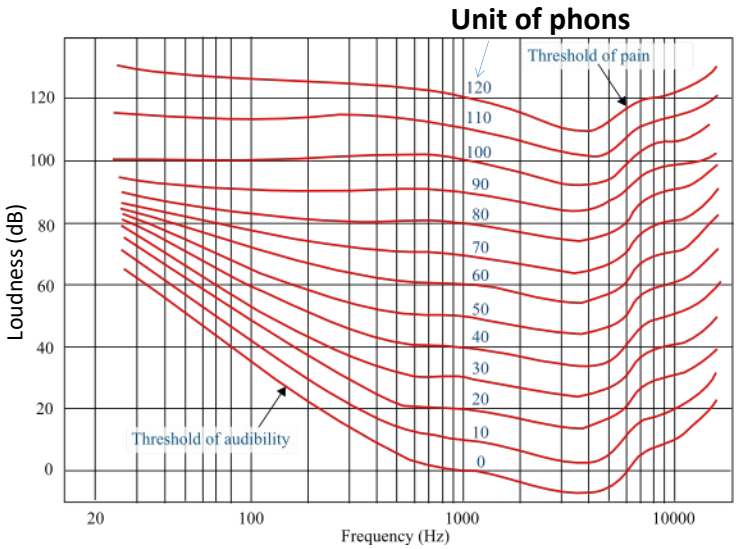


Note: Listening to loud music will gradually damage your hearing!

14

14

Physical sound pressure level vs. **perceived** loudness

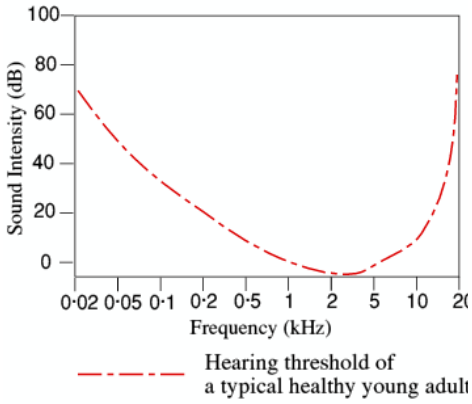


Each contour represents equally-perceived tones

15

15

Threshold of hearing

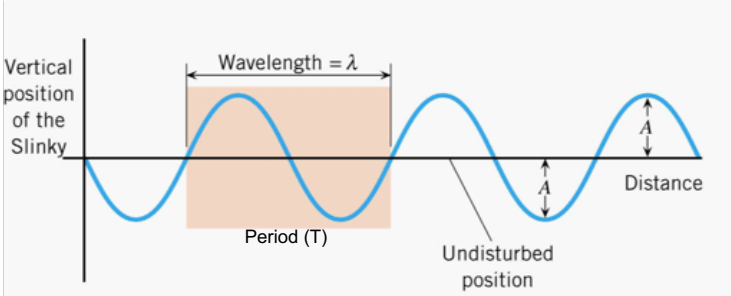


Threshold of hearing is a principle used for perceptually-lossless compression (MP3)

16

16

Physical Dimensions of Sound



Wavelength (λ)

Distance traveled by one cycle

Affected by medium (how sound travels)

Amplitude

Height of a cycle

Related to measure/perception of loudness

Frequency (F)

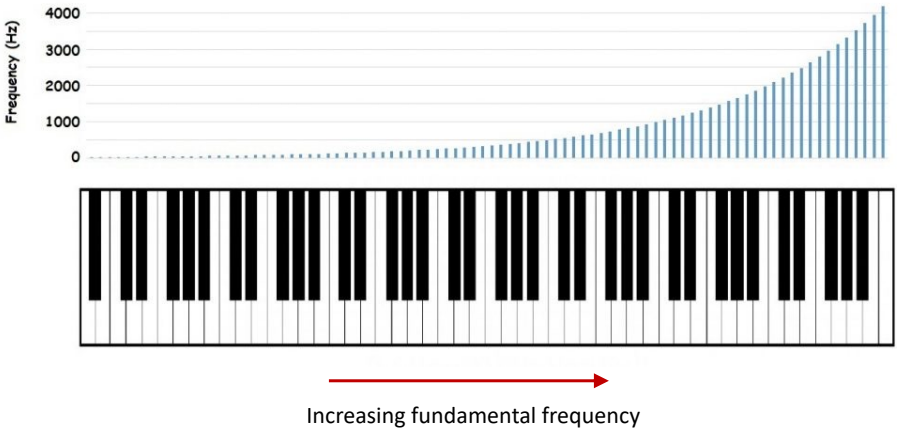
Cycles per second

Related to measure/perception of pitch/spectrum

17

17

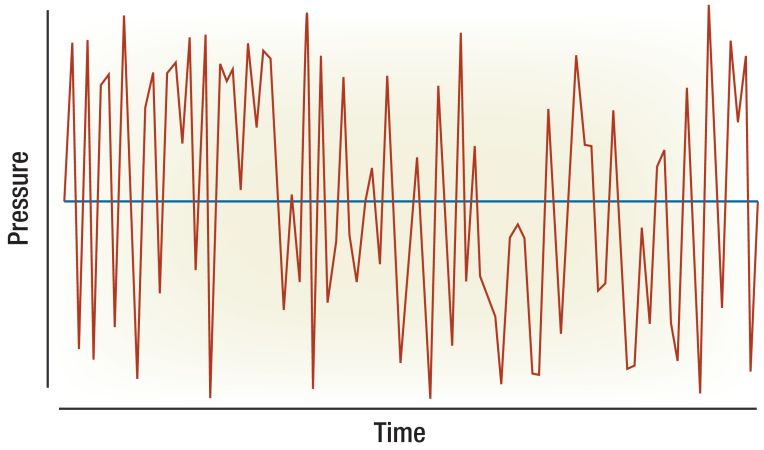
Frequency



18

18

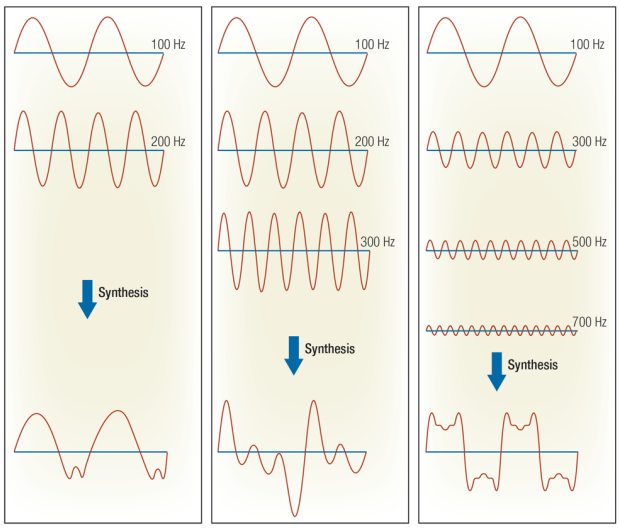
Most sounds (like speech) are not simple tones -> they consist of multiple frequencies



19

19

Complex Sound = Sum of Sinusoids (Fourier Theorem)

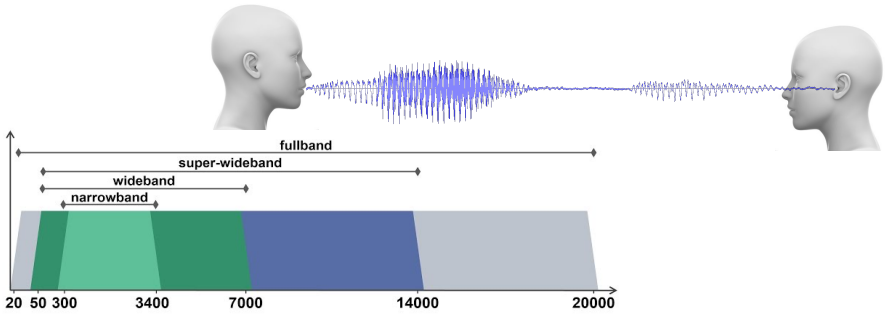


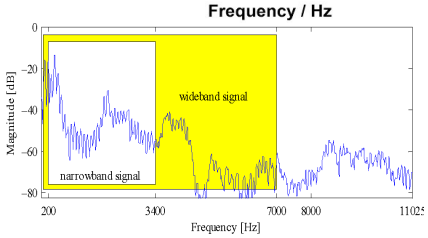


J.B.J. Fourier
(1768-1830)

20

Frequency <-> Speech



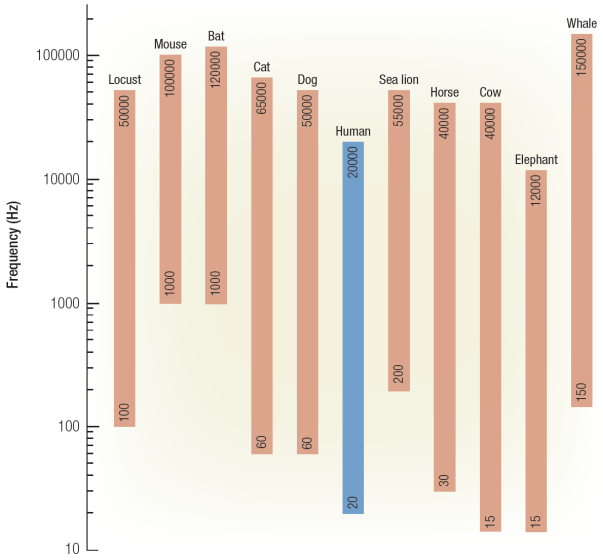


The human brain can hear between 20Hz and 20KHz (reduced with age)

Energy of speech signals is mostly below 5-6KHz

21

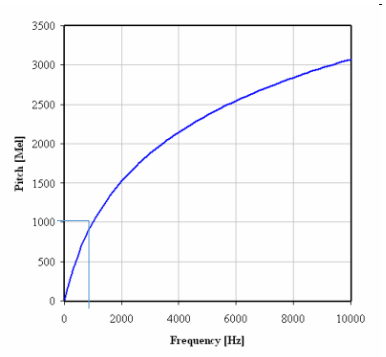
Species-Specific Frequency Range



22

How do we hear the frequency axis?

- Frequencies are represented in nonlinear fashion in the brain: The mel scale (~melody)
  - *How far in frequency do we have to be in order to feel a tone as doubled in pitch?*



➡ It's a relative scale, based on pitch comparisons

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

✓ Mel-scaling is used in signal processing to build filters that approximate human pitch perception (MFCC)

23

23

Masking

- Hearing phenomenon
  - When the perception of one sound is affected by presence of *another* sound
  - one sound being *masked* by another
- Term masking is used to describe effects of noise and interference in sound perception
- We experience masking everyday

24

24

Masking

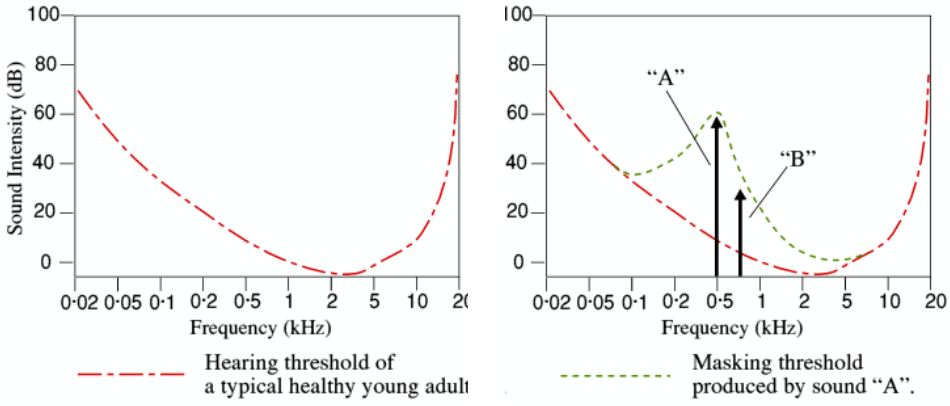


25

25

Frequency Masking

(simultaneous masking)



Masking is a principle used for perceptually-lossless compression (MP3)

26

26

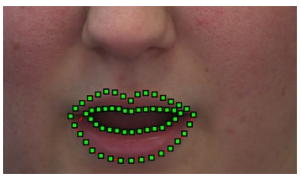
Speech perception

- Speech perception is affected by other factors beyond the characteristics of the signal itself
  - E.g. multimodal cues (Audiovisual integration)

27

Audio-visual integration

- Perception of sounds is heavily affected by other sensory modalities, importantly vision
  - We rely on it heavily especially to complement audio signals and navigate noisy environments .. Particularly true for people with hearing impairment -> **Lip reading**



  - Many cases where the 2 modalities clash

28

28

AV integration – McGurk effect

- "Compromise" between conflicting sound and visual cues in speech understanding
  - Compare auditory perception with eyes open vs. eyes closed



29

29

How do we perceive sounds?

30

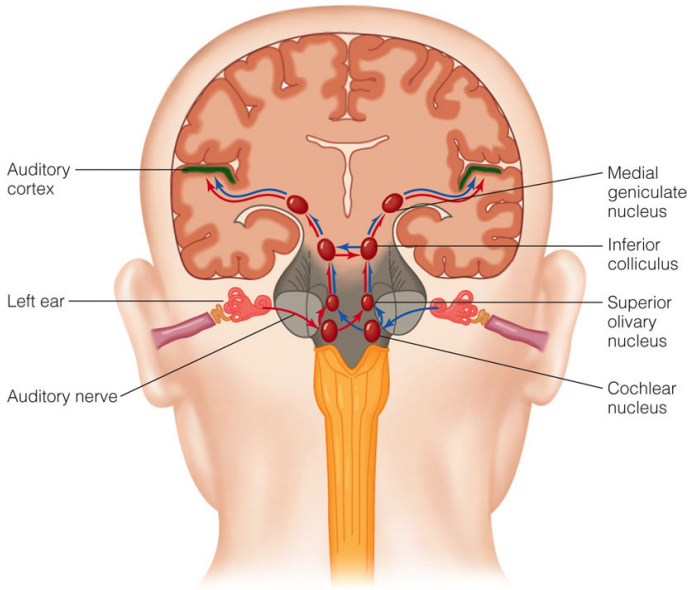
The auditory system

- Two major components in the auditory system
  - The peripheral auditory organs (the ear)
    - Converts sounds pressure into mechanical vibration patterns, which then are transformed into neural firings
  - The auditory nervous system (the brain)
    - Extracts perceptual information in various stages

31

31

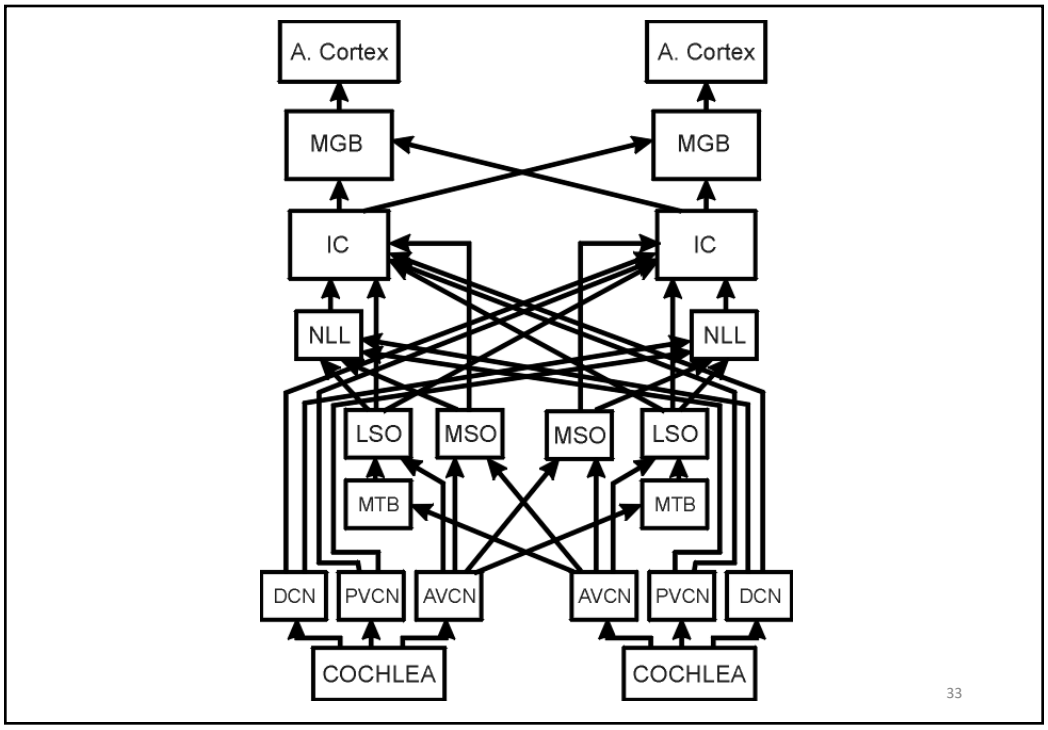
Auditory Pathway



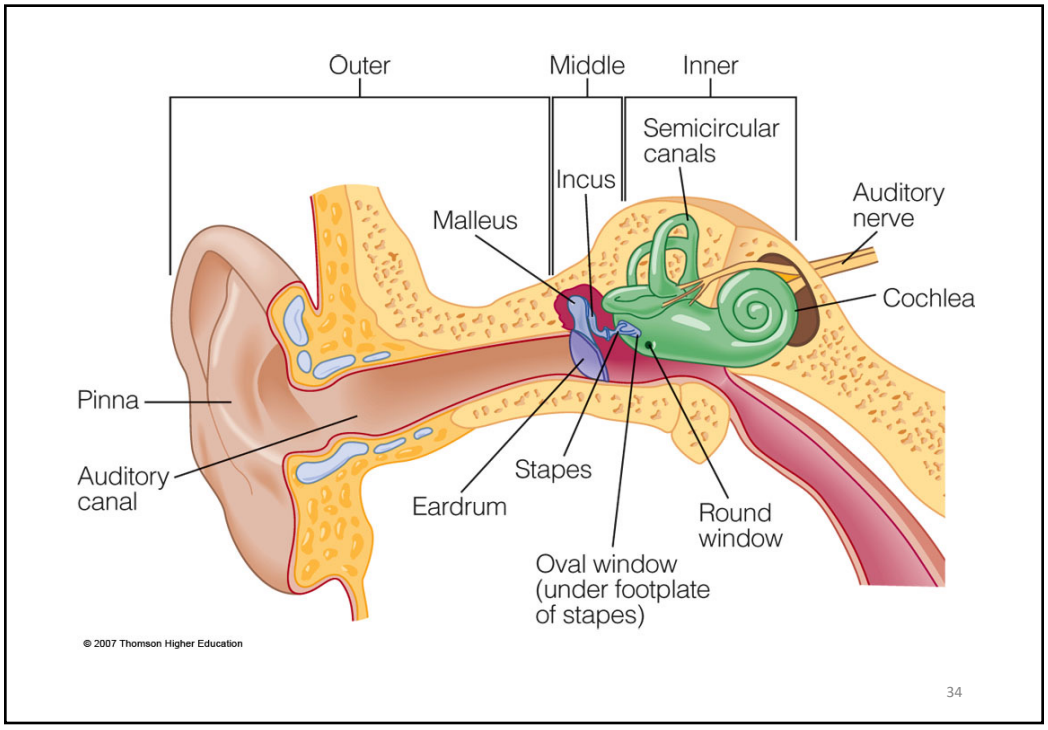
32

© 2007 Thomson Higher Education

32


A. Cortex
A. Cortex
MGB
MGB
IC
IC
NLL
NLL
LSO
MSO
MSO
LSO
MTB
MTB
DCN
PVCN
AVCN
AVCN
PVCN
DCN
COCHLEA
COCHLEA
33

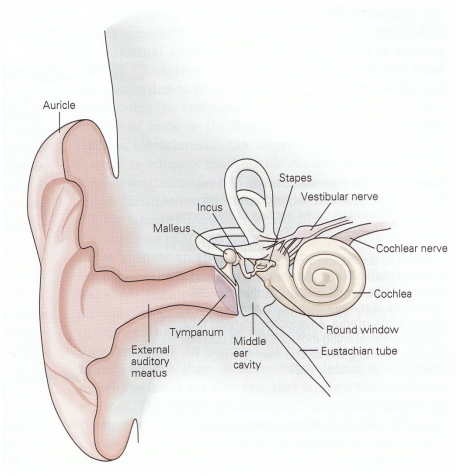
33


Outer
Middle
Inner
Semicircular canals
Incus
Malleus
Auditory nerve
Cochlea
Pinna
Auditory canal
Stapes
Eardrum
Round window
Oval window (under footplate of stapes)
© 2007 Thomson Higher Education
34

34

The ear

- The ear is the organ of hearing
- It changes sound pressure waves from the outside world into a signal of nerve impulses sent to the brain.
- It consists of 3 components:
  - Outer ear
  - Middle ear
  - Inner ear

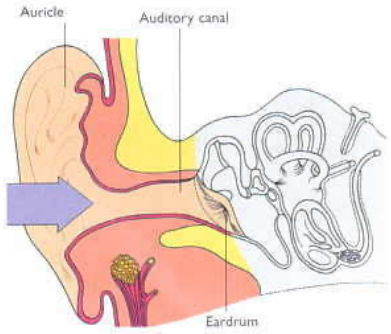


35

35

Organ of hearing outer ear

- The external ear plays the role of an acoustic antenna,
- It diffracts and focuses sound waves (pinna), while the ear canal acts as a resonator => amplifies sounds in 2-5 kHz range
- The end of the canal has an eardrum which vibrates with sound

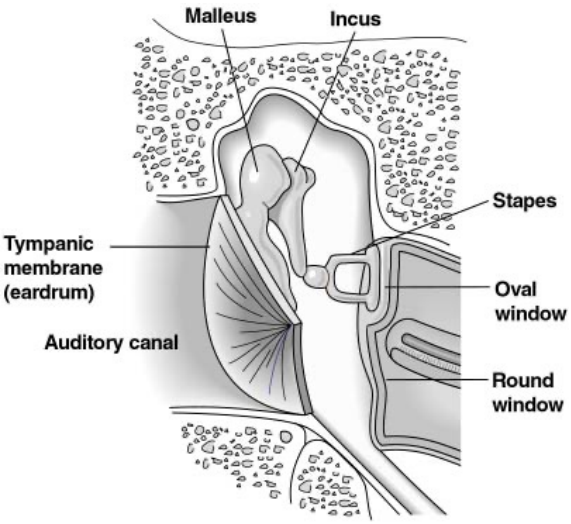


36

36

Organ of hearing middle ear

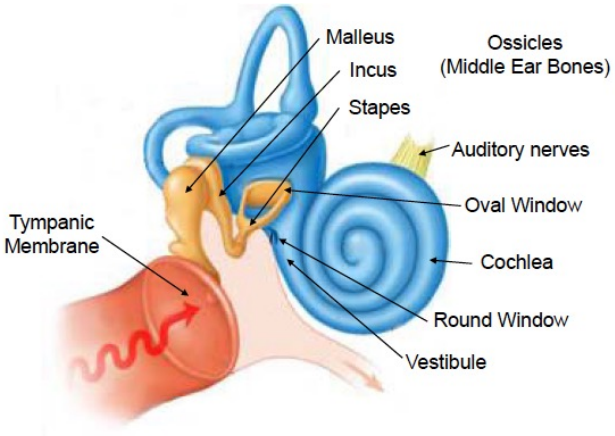
- Eardrum (or tympanic membrane) vibrations cause mechanical motion of the small bones of the middle ear (malleus, incus & stapes) [3 smallest bones in the human body]
- The middle ear acts as an impedance adapter to adjust energy difference between air environment and fluid environment



37

Organ of hearing inner ear

- Cochlea translates physical vibrations into electrical signals for the brain to process
- Cochlea acts a frequency analyzer of sound signals

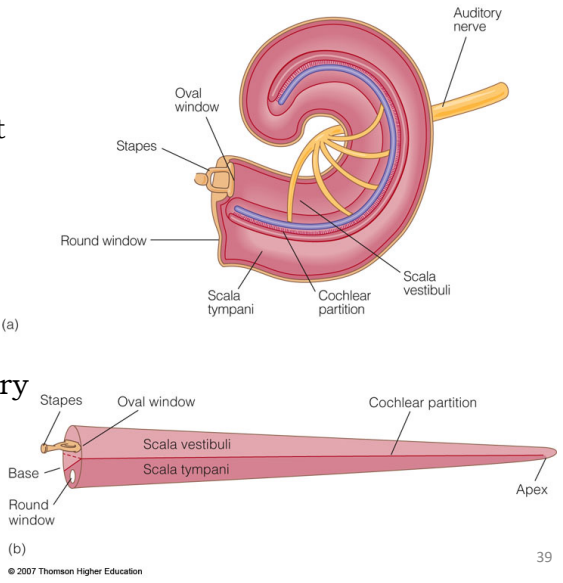


38

38

The Cochlea

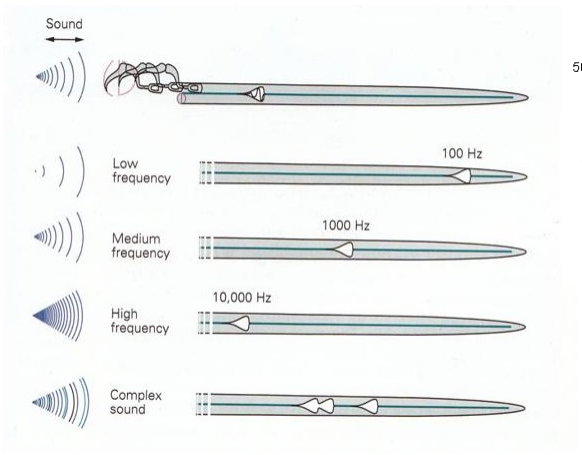
- The cochlea is the inner ear organ that converts sound waves into neural signals.
- The neural signals are passed to the brain via the auditory nerve.

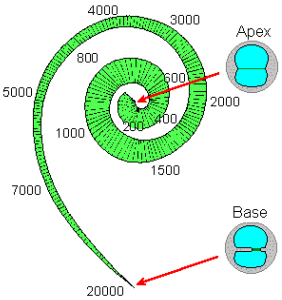


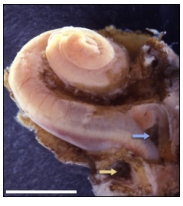
39

39

Cochlea as frequency analyzer







40

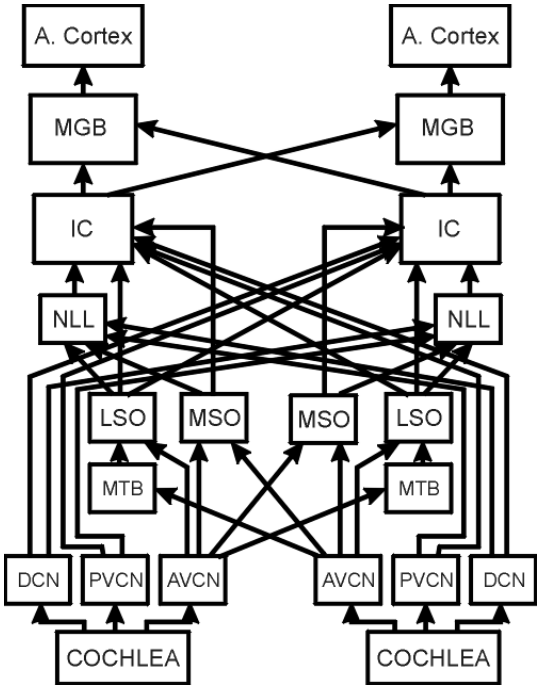
40

How the ear works (review)



41

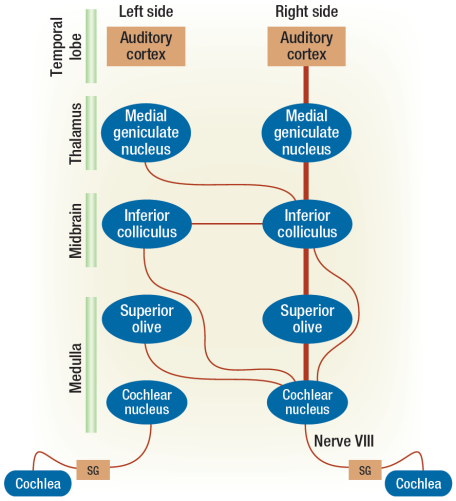
41



42

42

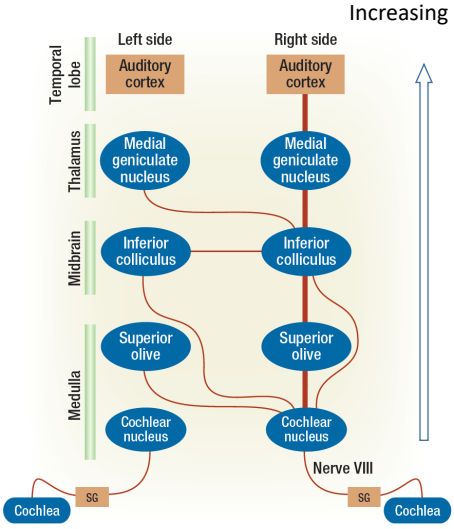
Ascending pathway

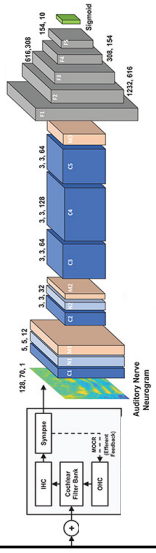


- Very complex. Just some major pathways shown.
- Extensive binaural interactions
- General principle:
  - ✓ Increasing complexity of responses (like vision, touch)

43

Ascending pathway





- Parallels with deep neural networks:
  - ✓ increasing complexity
  - ✓ Highly nonlinear
  - ✓ Recurrent connections
- Differences:
  - ✓ Multiple goals and objectives
  - ✓ Continual learning
  - ✓ Highly adaptive
  - ✓ Spike-based computation

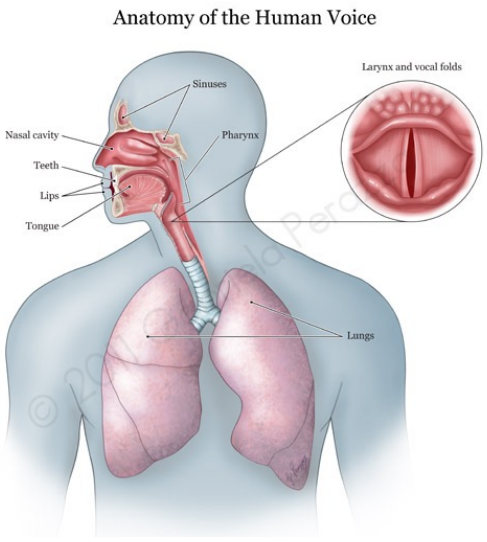
44

How is speech produced?

47

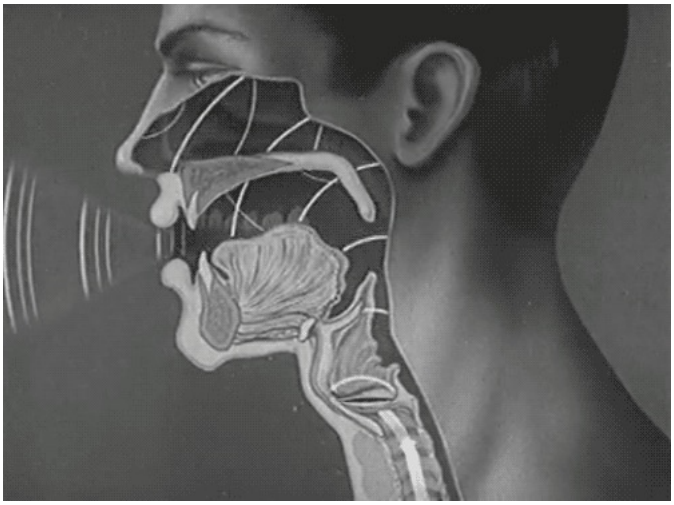
Anatomy & Physiology of speech organs

- Speech production apparatus starts from the lungs to the lips & nose.



48

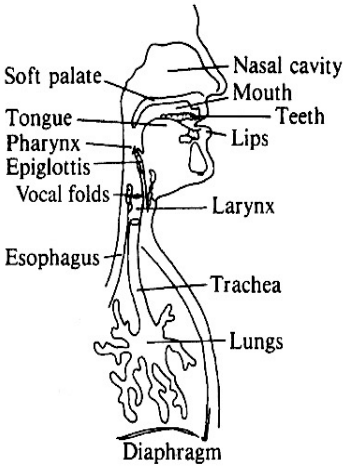
Anatomy & Physiology of speech organs

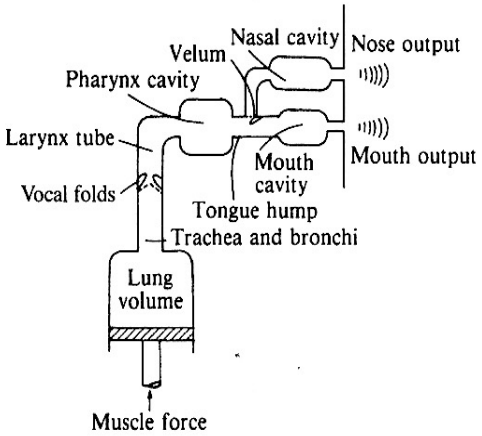


49

49

Anatomy & Physiology of speech organs

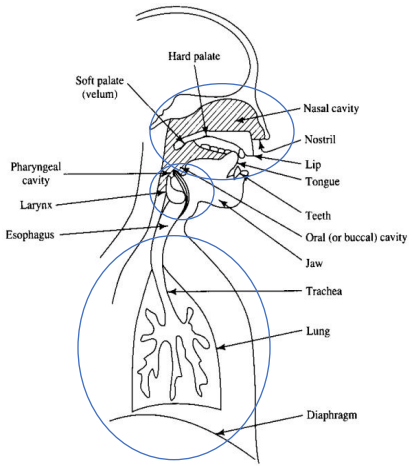




50

Speech Organs

- Speech organs:
  1. Lungs
  2. Larynx (vocal cords)
  3. Vocal tract

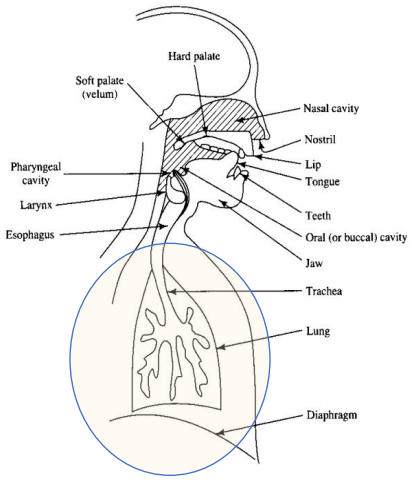


51

51

Speech Organs: Lungs

- Lungs play respiratory role: inhaling and exhaling air
- Inhalation results in expanding chest cavity to fill it with oxygen (lowering diaphragm which separates chest cavity from abdomen).
  - Lowers air pressure in the lungs
  - Causes air to rush through the vocal tract into lungs

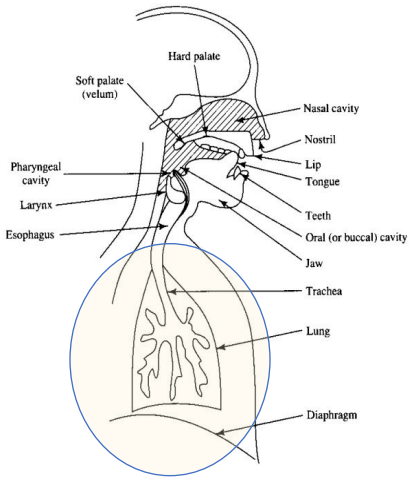


52

52

Speech Organs: Lungs

- Exhalation results in reducing volume of chest (contracting muscles of rib cage)
  - increases air pressure in the lungs
  - Causes air (carbon dioxide) to rush outside the lungs
  - Exhaling normally takes 60% of breathing cycle.

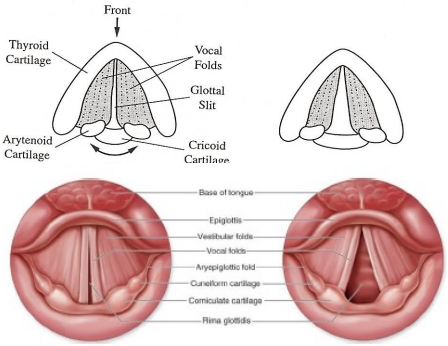


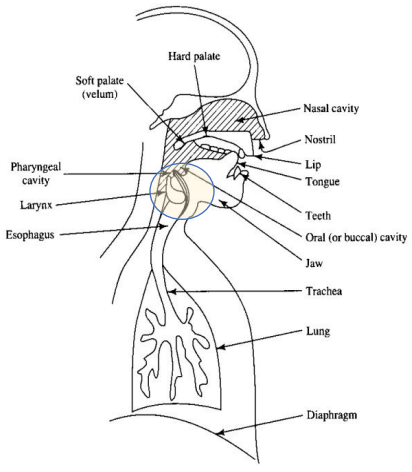
53

53

Speech Organs: Larynx

- Larynx: system of cartilages, muscles and ligaments
- Primary role is to control vocal cords (or vocal folds)



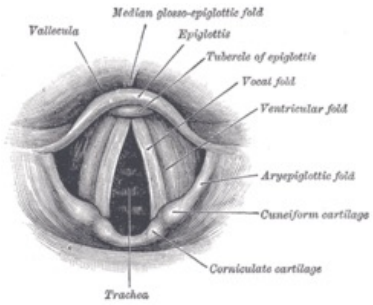


54

54

Speech Organs: Larynx

- Vocal folds have a main function of vibrating
- 3 primary/normal states of vocal folds:
  - breathing
  - Voiced
  - Unvoiced
- There are also abnormal states



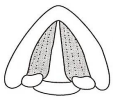
55

55

3 vocal states

Breathing

- The vocal fold muscles are relaxed
- Glottis is wide open
- Air flows from and to lungs with no hindrance

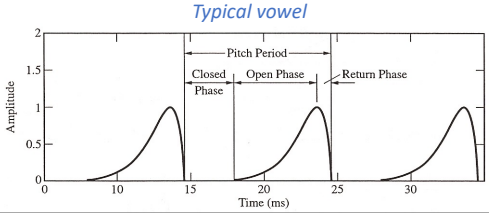


Voicing

cycles between 3 phases

- *Closed* phase: no flow occurs
- *Open* phase: nonzero flow of air up to a maximum velocity
- *Return* phase: interval when glottis start closing after maximal air velocity

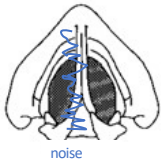
Typical vowel



Unvoicing

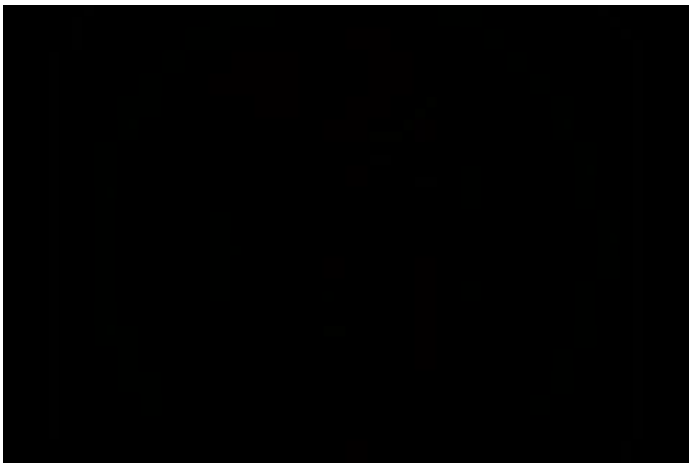
- similar to breathing state
- There is no vocal fold vibration
- But folds are closer together and more tense than in breathing state

/h/ (e.g. he)



56

Vocal fold animation



57

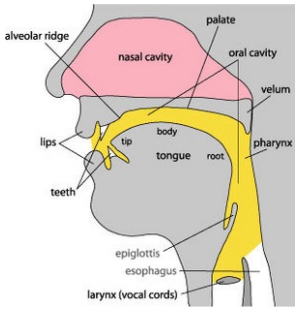
Vocal fold video

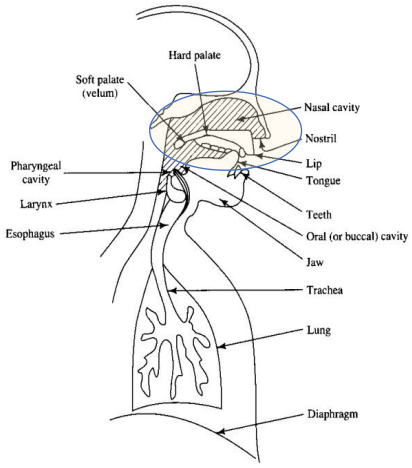


58

Speech Organs: Vocal tract

- Vocal tract is comprised of:
  - Oral cavity (from larynx to the lips)
  - Nasal cavity (coupled with oral tract through velum)



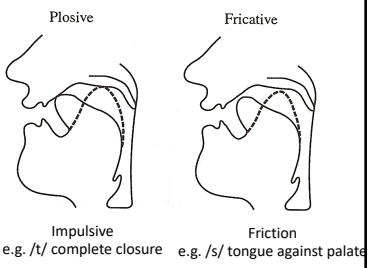


59

59

Speech Organs: Vocal tract

- Vocal tract plays 2 major roles:
  1. Generating sound



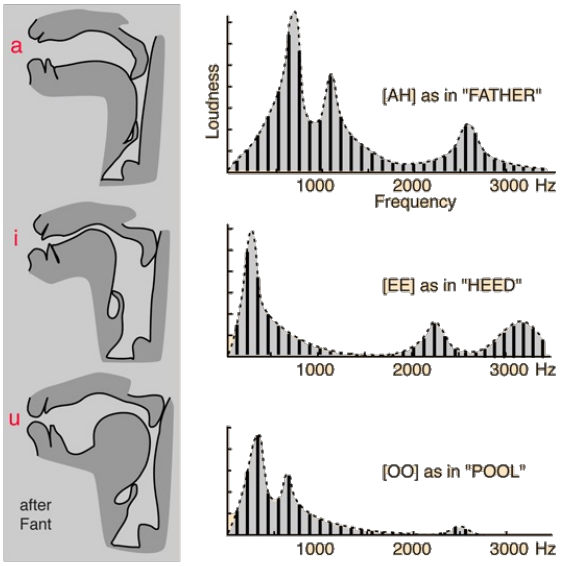
  2. Shaping sound generated at vocal folds
     - Also called: to color sound
     - If we listen to sound at output of vocal folds during voicing, we would hear a buzz-like sound which is not interesting
     - The vocal tract spectrally colors / shapes the sound to make perceptually distinct speech sounds

60

60

Formants

Part of the identity of sounds is defined by position of formants



61

Speech coming out of lips

Output sound

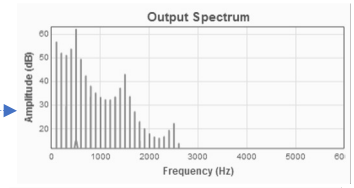
Vocal tract *(resonator)*

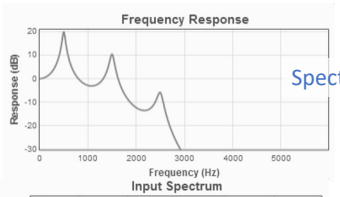
Voice source

Vibrating vocal folds *(oscillator)*

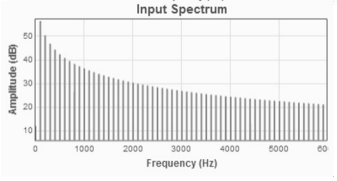
Airstream

Lungs *(power supply)*



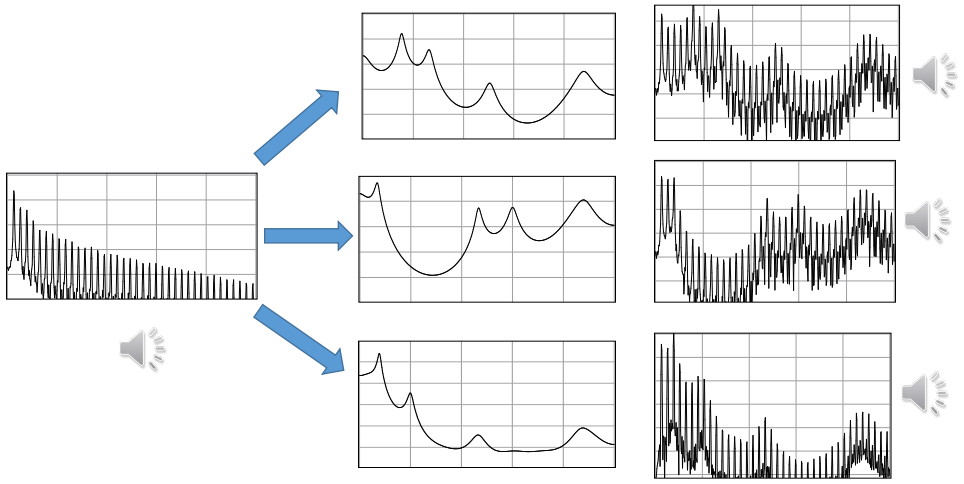


Spectral shaping



62

Same Source, Different Filters



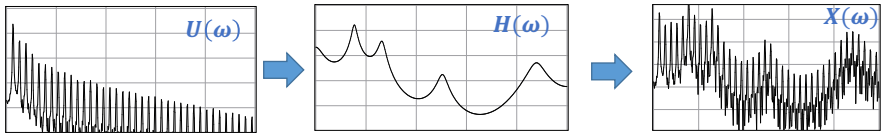
63

Source/Filter model -> Speech Analysis

Source (Phonation)

Filter (Resonator)

Speech (What We Hear)



Input Spectrum

Frequency Response Curve (Transfer Function)

Output Spectrum

Time domain: $x[n] = u[n] * h[n]$

Transform domain: $X(\omega) = U(\omega)H(\omega)$

64

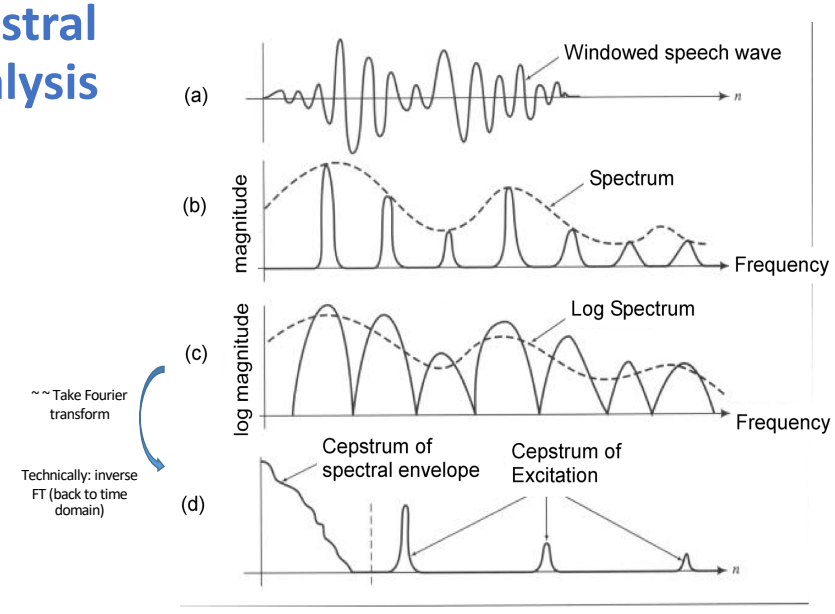
Source/Filter model -> Speech Analysis

- Deconvolution of speech is feasible because
  - $U(\omega)$ and $H(\omega)$ have different spectra
    - $U(\omega)$: Excitation harmonics or noise varies much faster
    - $H(\omega)$: vocal tract response (formant structure) varies very slowly in frequency
  - The contribution of $U(\omega)$ and $H(\omega)$ could be separated if they are in a linear domain
  - Use trick

$$X(\omega) = U(\omega)H(\omega)$$
$$logX(\omega) = logU(\omega) + logH(\omega)$$

65

Cepstral Analysis

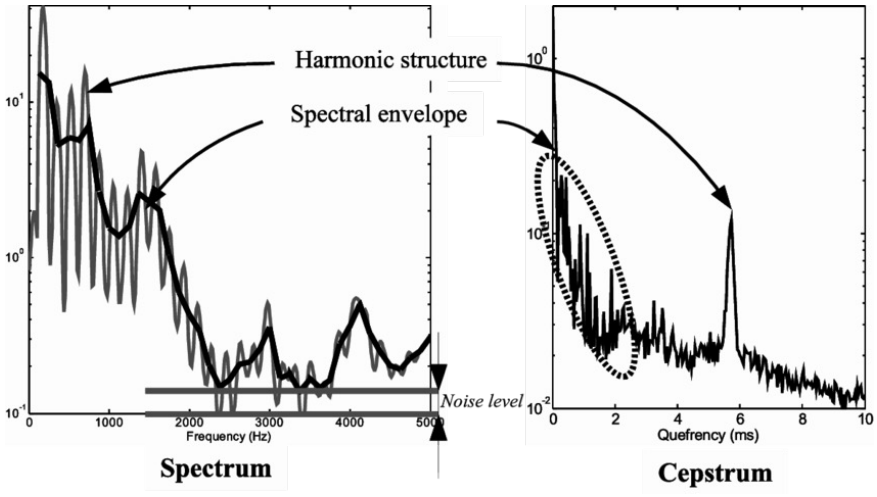


66

66

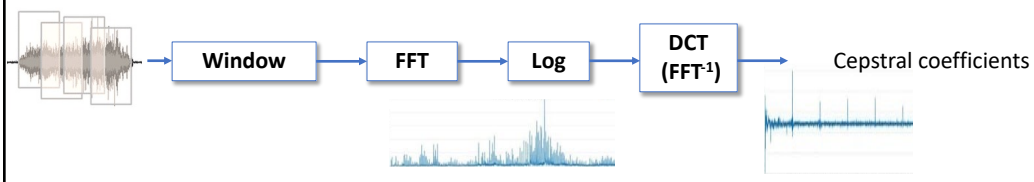
$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} logX(\omega)e^{j\omega n}d\omega$$

Cepstral Analysis

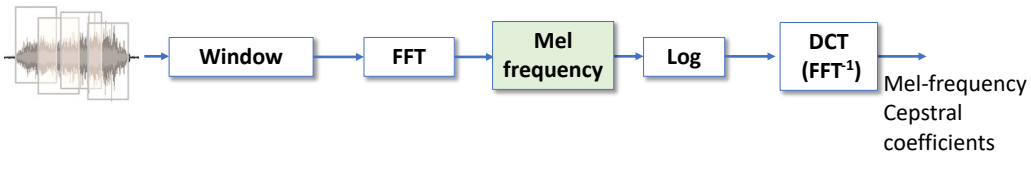


67

Cepstral Analysis



Mel-Frequency Cepstral Analysis



68