Outline of Today’s Lecture

- 0: Announcements
- I: Basic anatomy and physiology of the Ear
- II: Brief overview of Speech Perception theories
Good books (for today)

• Goldstein, “Sensation and Perception”
• Moore, “An Intro to the psychology of hearing”
• Pickles, “An Intro to the physiology of hearing”
• Clark & Yallop, “An Intro to Phonetics and Phonology”

Outer Ear

• Pinna (sound localization)
• Auditory canal – 3cm long
  – Wax
• “meatus = pinna + auditory canal
• Tympanic membrane = eardrum
  – Vibration (coupling between air pressure wave and bodily movement)
• Resonance in canal
  – Between 2kHz and 5kHz
Middle Ear

- 2 cm^2 in volume
- ossicles – 3 bones, couples together
  - end of outer ear, eardrum
  - beginning of inner ear via “oval window”
- Ossicles – 3 bones
  - malleus (hammer)
  - incus (anvil)
  - stapes (stirrup)
- Why middle ear?
  - acoustic coupler to correct for impedance mismatch between air and relatively incompressible fluid in inner ear
    - concentration larger area eardrum onto smaller area stapes & oval window
    - board balanced on fulcrum not in middle (see-saw)
  - if not, only ~ 3% vibration would be transmitted
- “reflex muscles”, middle ear muscles – will contract and dampen sound after impulsive or intense sound (reduces transduction)
  - prevents loud sound damage (mostly low frequencies)
  - reduction of audibility of self sounds, speech
  - reduce low frequency masking of high freq. sounds
- most efficient at 500-4k Hz
Inner Ear

• “Cochlea” = most important part of inner ear
• filled with lymphatic fluid
• tube coiled like snail shell
  – 35mm long, 2.5 turns
• cochlear partition - 2 membranes divide the cochlea lengthwise
• has “base” and “apex”
• stapes attaches to “scala vestibuli” via oval window, but liquid is not (very) compressible
• round window relieves pressure (like passive radiator, enclosed loudspeakers, but for liquid)
Inner Ear

- **Organ of Corti**
  - contained in cochlear partition
  - contains receptors, called “hair cells”
  - sits atop “basilar membrane”
  - covered by “tectorial membrane”
Inner Ear

• Organ of Corti
  – about 30k sensory hair cells
  – auditory nerve terminates at these hair cells (each nerve ending has about 40-140 hair cells)
• hair cell vibration causes nerve firings
• There are both inner and outer hair cells
  – outer cells have more nerve endings and have different response patterns
  – no final consensus as to the different function of these cells (difficult to measure in a live human without disturbing the entire system)
Inner Ear

• How do hair cells transmit? They bend.
  – oval window vibration causes a vibration wave to be sent along basilar membrane
  – entire organ of Corti vibrates
  – hair cells rub against tectoral membrane at position where vibration occurs

• “Q” of a filter
  – Q = center frequency / bandwidth
  • history: “Q” stands for “quality factor”, high-quality filters are very narrow-band relative to their center frequency.
  – constant Q filterbanks
  • as center frequency increases, so does BW.

Basilar membrane

• Each position has a characteristic frequency (CF)
  – maximum vibration for a given input sound

• Bandwidth is almost constant Q
  – bandwidth increases for increasing freq, better spectral resolution at lower frequencies

• Basilar membrane vibration

The wave displayed here, generated on the basilar membrane by a 400 Hz input tone, is the solution of an integro-differential equation which describes the membrane motion in the linear approximation. Waves peak at a frequency-dependent location and form because: i) each membrane segment interacts instantly with each other through the fluid filling the cochlear duct and ii) membrane stiffness is graded from base to apex.
Traveling Waves

- traveling wave (von Bekesy)
- oval window vibrates – waves travel up basilar membrane
- max amplitude of wave = point on BM where CF matches input frequency
- high freq = at base
- low freq = at apex
- stiffness of BM is a factor

Two views of frequency encoding

- Place Theory, von Bekesy’s traveling way (as described above)
- Timing theory (an alternative theory)
  - neurons fire at a rate that indicates frequency
  - possible for low frequencies
  - not possible (for single neuron) at high frequencies because of refractory period (time after which neuron can not fire after it has fired)
  - Population coding (multiple neurons can coordinate to represent information at rates faster than individual neurons can fire)
  - Still much research on this – current consensus, it is both place and timing
- Much controversy here
Timing – phase locking of nerves

Phase Locking

• Real neural phase locking
Two views of frequency encoding

- **Place**
  - weak at < 1kHz since traveling wave is wide
  - probably not utilized for information encoding in this range

- **Timing**
  - good at < 1kHz since timing can encode information (fire) at rates fast enough for these frequencies

- **Both operate at frequencies between 1k – 5kHz**
  - but still open question

- **Mel-scale frequency warping**
  - linear below 1kHz, log thereafter
  - Used ubiquitously for speech recognition
  - similar to the place-timing view above
  - The “M” in MFCCs

Sound Perception

- **Thresholds**
  - auditory system has a dynamic range
**Sound Perception**

- **Thresholds**
  - thresholds of hearing
  - threshold of “annoyance”
  - thresholds of feeling (pain)
- **Speech frequencies relative const. threshold of hearing**
- **F0’s of male and females (typically)**
  - Males: F0 = 50-250 Hz
  - Females: F0 = 120-500 Hz
  - So easier to hear females at lower volume

**Fletcher & Munson Curves**

- **Loudness ≠ Intensity**
Units of Loudness - Phons

• 80 phons = different SPLs at diff. frequencies
Masking

• perception of one sound is obscured by presence of another sound
  – raises threshold of hearing for other sound
  – “A masks B” means A makes you unable to hear B.

• perceptual auditory coding (e.g., mpeg standards)
  – use fewer bits (more additive noise) by quantizing at a coarser scale where perceptual system might mask a given sound by some louder sound

• simultaneous masking
  – two sounds at once, one masks the other

• temporal masking
  – one sound after the other, one masks the other
  – forward masking/backward masking

Masking

• How to use masking to measure shape of critical band

• Neural Tuning Curves
  – probe a single neuron and measure response (in terms of spiking rate) for a given stimulus
  – typically constant Q

• Tuning curves of anesthetized cats, for each neuron
  – curves show threshold where neuron starts to respond above chance level (spontaneous firing rate)
  – different neurons have different tuning curves
  – different neurons have different characteristic frequencies, or CFs (point of lowest threshold)
Psychophysical tuning curve - Human

- fixed level sinusoidal tone signal in narrow-band noise masker with some center (masker) frequency
- signal is fixed at a low level (~ 10db SPL) - so it will produce signal output mostly in one auditory filter (no lateral spreading at low levels)
- If masker is loud enough, it will prevent listener from hearing signal
- Procedure:
  - fix signal, change masker frequency and intensity
  - find intensity at each freq. to mask tone
  - intensity gives upside-down CF shape
- masking threshold as a function of masker frequency and masking intensity with fixed signal -> CF (linearity assumption)
- there are other ways to do this as well

Speech Perception

- Much higher level in auditory cortex or brain – not well understood
- Goal: find invariant acoustic cues for different speech sounds
  - phonemes, phones, syllables, words, phrases, sentences, etc.
  - Difficult since acoustic cues for an object change depending on context
Cues depend on context

• cues for different ‘d’ sounds are different depending on context
  – F2 totally different
  – This is one of the reasons formants are unreliable to do phone classification
• So must be more than just formants that are used to determine words.
  – temporal patterns, long context, adaptation
Spectral Regions of Speech Perception

- 200-5.5kHz most important
  - by filtering out spectral regions and measuring intelligibility
- ISDN: 4kHz BW
- formal perceptual theories to determine intelligibility
  - ex: Articulation Index by Steeneken and Houtgast

Spectral Regions of Speech Perception

- ex: filter out < 1kHz, then voicing and manner of articulation discrimination decreases (/p/ vs. /b/ vs. /v/)
- ex: filter out > 1.2kHz, place of articulation discrimination drops (/p/ vs. /t/)
- Telephone bandwidth 200-4kHz (good enough for most intelligibility)
- Particularly bad are the infamous “E-set”
  - /p/, /d/, /e/, /g/, /c/, etc.
  - vowel energy
Human Speech Perception

- we are remarkably good at perceiving speech
- SWS Example (next slide)
- face recognition
- evolution
- therefore, hard to identify most important cues, since they all could potentially be used depending on context

SineWave Synthesis
(haskins lab, Yale)
Human Speech Perception

• Speech is redundant
  – checkerboard speech

• Spectral transitions, derivatives

• Gaussian Scaled Speech (example)

• Apparently, no particular location in
time/frequency that contains the crucial
information (we can infer what is missing from the
other bits)

• Redundancy allows us to perceive speech in many
different situations
  – background noise, cocktail party effect