Announcements, Assignments, and Reminders

- Visit the URL links that were covered in previous lectures.
Read chapters 1 and 2 in our book (Huang, Acero, Hon, “Spoken Language Processing”).
We’ve covered speech production, including the mechanisms of speech production, and methods to describe the production of speech.
We’ve covered speech production, including the mechanisms of speech production, and methods to describe the production of speech. Acoustic tube model leads to a simple all-pole filter (which will come up again when we talk about feature extraction)!
We’ve covered speech production, including the mechanisms of speech production, and methods to describe the production of speech.

Acoustic tube model leads to a simple all-pole filter (which will come up again when we talk about feature extraction)!

Speech production analyses for the purposes of speech recognition can be seen as a form of analysis by synthesis: inverse modeling problem, come up with parameters that given signal describe the generative process for that signal.
Outline of today

- Today we begin our discussion of speech perception
- Basic anatomy and physiology of the Ear
- Brief overview of Speech Perception theories
Good books (for today)

- our book (Huang, Acero, Hon, “Spoken Language Processing”)
- 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”
Some good videos to watch

- Overview animation: http://www.youtube.com/watch?v=qgdqp-oPb1Q
- Ep 37: How Sound is Transferred to the Inner Ear http://youtu.be/L4F4zaRqQdk
- Ep 38: How We Hear Different Pitches http://youtu.be/Id-L0_7e9BI
- Ep 40: The Role of Hair Cells in Hearing: http://youtu.be/lDXVZ0U_f_E
The ear is of course how we hear.
The ear is of course how we hear.

All speech perception starts by the sound pressure waves first entering the head through the ear.
The ear is of course how we hear.

All speech perception starts by the sound pressure waves first entering the head through the ear.

The ear is a complicated organ that itself transforms the sounds in quite significant ways, and the way it does so helps to determine the perception of speech.
The ear is of course how we hear.

All speech perception starts by the sound pressure waves first entering the head through the ear.

The ear is a complicated organ that itself transforms the sounds in quite significant ways, and the way it does so helps to determine the perception of speech.

Amazingly, as we will see, the ear can be seen to be doing a form of Fourier analysis of the signal.
The ear is of course how we hear.

All speech perception starts by the sound pressure waves first entering the head through the ear.

The ear is a complicated organ that itself transforms the sounds in quite significant ways, and the way it does so helps to determine the perception of speech.

Amazingly, as we will see, the ear can be seen to be doing a form of Fourier analysis of the signal.

Function of the ear is (at least in part) like a bank of band-pass filters, with increasing bandwidths as the center frequency rises.
The ear is of course how we hear.

All speech perception starts by the sound pressure waves first entering the head through the ear.

The ear is a complicated organ that itself transforms the sounds in quite significant ways, and the way it does so helps to determine the perception of speech.

Amazingly, as we will see, the ear can be seen to be doing a form of Fourier analysis of the signal function of the ear is (at least in part) like a bank of band-pass filters, with increasing bandwidths as the center frequency rises.

The ear itself can be described as having three parts, the outer, middle, and inner ear so we talk about each in turn.
Outer Ear

- Pinna (sound localization)
- Auditory canal - 3cm long, contains Wax
- meatus = pinna + auditory canal
- Tympanic membrane = eardrum: vibration (coupling between air pressure wave and bodily movement)
- Resonance in canal itself, between 2kHz and 5kHz, good for speech.
The Outer and Middle Ears

- Concha
- Pinna
- External auditory meatus
- Tympanic membrane
- Temporal bone
- Tensor tympani tendon
- Incus
- Stapes tendon
- Lateral posterior superior semicircular canals
- Oval window
- Vestibule
- Round window
- Cochlea
- Facial nerve
- Vestibular nerve
- Cochlear nerve
- Eustachian tube
- Parotid gland
- Stapes
The Ear

Sound Perception

Speech Perception

Summary

Scratch

Outer & Middle Ears
Middle Ear

- about 2 cm$^2$ in volume
Middle Ear

- about 2 cm² in volume
- contains ossicles - 3 bones, couples together: 1) end of outer ear, eardrum, and 2) beginning of inner ear via “oval window”
Middle Ear

- about 2 cm$^2$ in volume
- contains ossicles - 3 bones, couples together: 1) end of outer ear, eardrum, and 2) beginning of inner ear via “oval window”
- Ossicles - 3 bones
Middle Ear

- about 2 cm² in volume
- contains ossicles - 3 bones, couples together: 1) end of outer ear, eardrum, and 2) beginning of inner ear via “oval window”
- Ossicles - 3 bones
  - malleus (hammer)
Middle Ear

- about 2 cm² in volume
- contains ossicles - 3 bones, couples together: 1) end of outer ear, eardrum, and 2) beginning of inner ear via “oval window”
- Ossicles - 3 bones
  - malleus (hammer)
  - inncus (anvil)
Middle Ear

- about 2 cm$^2$ in volume
- contains ossicles - 3 bones, couples together: 1) end of outer ear, eardrum, and 2) beginning of inner ear via “oval window”
- Ossicles - 3 bones
  - malleus (hammer)
  - incus (anvil)
  - stapes (stirrup)
Why Middle Ear

Why do we have middle ear?

- acoustic coupler to correct for impedance mismatch between air and relatively incompressible fluid in inner ear
Why do we have 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

[See figure next page]

"reflex muscle", 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 is most efficient at 500-4k Hz
Why Middle Ear

Why do we have 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)

"Reflex muscle", 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

Is most efficient at 500-4k Hz
Why Middle Ear

Why do we have 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)
  - without, only $\approx 3\%$ SPW vibration would be transmitted

- reflex muscle "s", 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
  - low frequency masking of high freq. sounds

- is most efficient at 500-4k Hz
Why Middle Ear

Why do we have 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)
  - without, only $\approx 3\%$ SPW vibration would be transmitted
  - see figure next page
Why Middle Ear

Why do we have 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)
  - without, only \( \approx 3\% \) SPW vibration would be transmitted
  - see figure next page

- “reflex muscles”, middle ear muscles - will contract and dampen sound after impulsive or intense sound (reduces transduction)
Why do we have 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)
  - without, only $\approx 3\%$ SPW vibration would be transmitted
  - see figure next page
- “reflex muscles”, middle ear muscles - will contract and dampen sound after impulsive or intense sound (reduces transduction)
  - prevents loud sound damage (mostly low frequencies)
Why Middle Ear

Why do we have 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)
  - without, only \( \approx 3\% \) SPW vibration would be transmitted
  - see figure next page
- “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
Why do we have 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)
  - without, only ≈ 3% SPW vibration would be transmitted
  - see figure next page

- “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
Why do we have 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)
  - without, only \( \approx 3\% \) SPW vibration would be transmitted
- see figure next page

- “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

- is most efficient at 500-4k Hz
Middle Ear - acoustic coupler

- Malleus
- Incus
- Stapes
- Tympanic membrane (eardrum)
- Oval window
- Round window
- Auditory canal
- Auditory nerve
**Middle Ear - acoustic coupler**

\[
\text{Force} = P_1 A_1 = P_2 A_2
\]

Therefore \(\frac{P_2}{P_1} = \frac{A_1}{A_2}\)

\[
\frac{l_1}{l_2} = \frac{F_1}{F_2} = \frac{v_1}{v_2}
\]
“Cochlea” = arguably, most important part of inner ear
Inner Ear

- “Cochlea” = arguably, most important part of inner ear
- filled with lymphatic fluid
**Inner Ear**

- “Cochlea” = arguably, most important part of inner ear
- filled with lymphatic fluid
- tube coiled like snail shell - 35mm long, 2.5 turns
“Cochlea” = arguably, 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
Inner Ear

- “Cochlea” = arguably, 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
- cochlea has a “base” and an “apex”
“Cochlea” = arguably, 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
- cochlea has a “base” and an “apex”
- stapes attaches to “scala vestibuli” via oval window, but liquid is not (very) compressible
Inner Ear

- “Cochlea” = arguably, 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
- cochlea has a “base” and an “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
Inner Ear - Organ of Corti

- contained in cochlear partition
- contains receptors, called “hair cells”
Inner Ear - Organ of Corti

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

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

- about 30k sensory hair cells
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)
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
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
The Ear

Sound Perception

Speech Perception

Summary

Scratch

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
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)
Cilia, small hairs attached to the hair cells.
Cilia

- Cilia, small hairs attached to the hair cells.
- Hair cells embedded in organ of Corti.
Cilia

- Cilia, small hairs attached to the hair cells.
- Hair cells embedded in organ of Corti.
- Called “hair cells” because protruding from them are bundles of tiny protrusions called “stereocilia”.

Cilia
Cilia, small hairs attached to the hair cells.

Hair cells embedded in organ of Corti.

called “hair cells” because protruding from them are bundles of tiny protrusions called “stereocilia”.

Initially, people thought they were normal hairs but that are actually are microvilli (microscopic cellular membrane protrusions that increase the surface area of cells).
Cilia

- Cilia, small hairs attached to the hair cells.
- Hair cells embedded in organ of Corti.
- Called “hair cells” because protruding from them are bundles of tiny protrusions called “stereocilia”.
- Initially, people thought they were normal hairs but that are actually are microvilli (microscopic cellular membrane protrusions that increase the surface area of cells).
- When the basilar membrane is deflected by fluid waves in the cochlea, deflection of the stereociliary bundles produces electrical signals in their associated hair cell body.
Hair Cells - SEM picture of a mouse hair cell.

How do hair cells transmit? They bend.
Inner Ear

- How do hair cells transmit? They bend.
- oval window vibration causes a vibration wave to be sent along basilar membrane
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.
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
- roughly, hair cells “rub” against tectoral membrane at position where vibration occurs
What is the \( Q \) of a filter?

- \( Q = \frac{\text{center frequency}}{\text{bandwidth}} \)

![Graph showing different Q values](image)

- \( Q = 0.1 \)
- \( Q = 1 \)
- \( Q = 10 \)
What is the “Q” of a filter?

- $Q = \frac{\text{center frequency}}{\text{bandwidth}}$

- History: “Q” stands for “quality factor”, high-quality filters are very narrow-band relative to their center frequency.
What is the “$Q$” of a filter?

- $Q = \text{center frequency} / \text{bandwidth}$

history: “$Q$” stands for “quality factor”, high-quality filters are very narrow-band relative to their center frequency.

- A filterbank is a set of band-pass filters.
What is the “Q” of a filter?

- $Q = \text{center frequency} / \text{bandwidth}$

- History: “Q” stands for “quality factor”, high-quality filters are very narrow-band relative to their center frequency.

- A filterbank is a set of band-pass filters.

- A constant $Q$ filterbank is one where the bandwidth of the filter increases with center frequency. I.e. the $Q$ of each filter is constant.
Basilar membrane

- Each position along membrane has a characteristic frequency (CF)
Basilar membrane

- Each position along membrane has a characteristic frequency (CF)
- The CF corresponds to maximum vibration for a given input sound.
Basilar membrane

- Each position along membrane has a characteristic frequency (CF)
- The CF corresponds to maximum vibration for a given input sound.
- Bandwidth is approximately constant $Q$ - bandwidth increases for increasing freq, better spectral resolution at lower frequencies
Basilar membrane vibration

- Wave generated on the basilar membrane by a 400 Hz input tone
Basilar membrane vibration

- Wave generated on the basilar membrane by a 400 Hz input tone

- Approximation, based on solution of integro-differential equation which describes the membrane motion in the linear approximation.

See [http://www.youtube.com/watch?v=1JE8WduJKV4](http://www.youtube.com/watch?v=1JE8WduJKV4), [http://www.youtube.com/watch?v=qgdqp-oPb1Q](http://www.youtube.com/watch?v=qgdqp-oPb1Q), and for fun see [http://www.youtube.com/watch?v=dyenMluFaUw](http://www.youtube.com/watch?v=dyenMluFaUw)
Basilar membrane vibration

- Wave generated on the basilar membrane by a 400 Hz input tone

- Approximation, based on solution of integro-differential equation which describes the membrane motion in the linear approximation.

- Waves peak at a frequency-dependent location since: 1) each membrane segment interacts with each other through the fluid the cochlear duct and 2) membrane stiffness is graded from base to apex.

Basilar membrane vibration

- Wave generated on the basilar membrane by a 400 Hz input tone

- Approximation, based on solution of integro-differential equation which describes the membrane motion in the linear approximation.
- Waves peak at a frequency-dependent location since: 1) each membrane segment interacts with each other through the fluid the cochlear duct and 2) membrane stiffness is graded from base to apex.
- See http://www.youtube.com/watch?v=1JE8WduJKV4, http://www.youtube.com/watch?v=qgdqp-oPb1Q, and for fun see http://www.youtube.com/watch?v=dyenMluFaUw
Traveling Waves

- traveling wave (von Bekesy)
Traveling Waves

- traveling wave (von Bekesy)
- oval window vibrates - waves travel up basilar membrane
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
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
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
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)
Two views of frequency encoding

- Place Theory, von Bekesy’s traveling way (as described above)
- Timing theory (an alternative theory)
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
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
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)
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)
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 have part
Two views of frequency encoding

- **Place Theory**, von Bekesy’s traveling wave (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 have part

- Some disagreement here still
Timing - phase locking of nerves

Stimulus

1 2 3 4 5 6 7 8 9 10 11 12

a

b

c

d

e

Total response
Timing - phase locking of nerves, more realistic
Two views of frequency encoding

- Place
Two views of frequency encoding

- **Place**
  - weak at < 1kHz since traveling wave is wide
Two views of frequency encoding

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

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

- **Timing**
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
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)
Two views of frequency encoding

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

- **Timing**
  - good at $< 1$kHz 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
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
Two views of frequency encoding

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

- **Timing**
  - good at $< 1$kHz 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
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
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
  - Mel-scale is the “M” in MFCCs, which we will learn about.
Sound Perception

- Thresholds - auditory system has a dynamic range
Sound Perception

- Thresholds

- Thresholds of hearing
- Thresholds of annoyance
- Thresholds of feeling (pain)

- Speech frequencies relative const. threshold of hearing
  - Males: F₀ = 50-250 Hz
  - Females: F₀ = 120-500 Hz

So easier to hear females at lower volume
Sound Perception

- Thresholds
  - thresholds of hearing
Sound Perception

- Thresholds
  - thresholds of hearing
  - threshold of “annoyance”
Sound Perception

- Thresholds
  - thresholds of hearing
  - threshold of "annoyance"
  - thresholds of feeling (pain)
Sound Perception

- Thresholds
  - thresholds of hearing
  - threshold of “annoyance”
  - thresholds of feeling (pain)

- Speech frequencies relative const. threshold of hearing
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)
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
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
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**

![Graph showing Fletcher & Munson Curves](image-url)
Units of Loudness - Phons

- 80 phons = different SPLs at different frequencies
Units of Loudness - Phons

- 80 phons = different SPLs at different 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.
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
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
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
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
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
How to use masking to measure shape of critical band
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
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
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)
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
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)
Neural tuning curve - Cat
“Neural” (or really psychophysical) tuning curve - Human

![Graph showing tuning curve](image-url)
Two signals mixed together, the **signal** and the **masker**.
Psychophysical tuning curve - Human

- Two signals mixed together, the **signal** and the **masker**.
- **fixed level sinusoidal tone** signal occurs in narrow-band noise masker having a center (masker frequency).
Psychophysical tuning curve - Human

- Two signals mixed together, the signal and the masker.
- Fixed level sinusoidal tone signal occurs in narrow-band noise masker having a center (masker frequency).
- Signal is fixed at a low level ($\approx 10\text{db SPL}$) - so it will produce signal output mostly in one auditory filter (no lateral spreading at low levels).
Psychophysical tuning curve - Human

- Two signals mixed together, the signal and the masker.
- Fixed level sinusoidal tone signal occurs in narrow-band noise masker having a center (masker frequency).
- Signal is fixed at a low level ($\approx 10\text{db SPL}$) - 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 (masker will mask).
Psychophysical tuning curve - Human

- Two signals mixed together, the **signal** and the **masker**.
- **fixed level sinusoidal tone signal** occurs in narrow-band noise **masker** having a 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 (masker will mask).
- **Procedure:**
Psychophysical tuning curve - Human

- Two signals mixed together, the **signal** and the **masker**.
- fixed level sinusoidal tone **signal** occurs in narrow-band noise **masker** having a 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 (masker will mask).
- Procedure:
  - fix signal, change masker frequency and intensity
Psychophysical tuning curve - Human

- Two signals mixed together, the signal and the masker.
- fixed level sinusoidal tone signal occurs in narrow-band noise masker having a 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 (masker will mask).
- Procedure:
  - fix signal, change masker frequency and intensity
  - find intensity at each masker freq that masks signal
Psychophysical tuning curve - Human

- Two signals mixed together, the signal and the masker.
- fixed level sinusoidal tone signal occurs in narrow-band noise masker having a center (masker frequency).
- signal is fixed at a low level ($\approx 10$db 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 (masker will mask).
- Procedure:
  - fix signal, change masker frequency and intensity
  - find intensity at each masker freq that masks signal
  - intensity profile gives upside-down CF shape
Psychophysical tuning curve - Human

- Two signals mixed together, the signal and the masker.
- fixed level sinusoidal tone signal occurs in narrow-band noise masker having a center (masker frequency).
- signal is fixed at a low level ($\approx 10$db 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 (masker will mask).

Procedure:
- fix signal, change masker frequency and intensity
- find intensity at each masker freq that masks signal
- intensity profile gives upside-down CF shape

- masking threshold as a function of masker frequency and masking intensity with fixed signal $\Rightarrow$ CF (linearity assumption)
Psychophysical tuning curve - Human

- Two signals mixed together, the signal and the masker.
- fixed level sinusoidal tone signal occurs in narrow-band noise masker having a 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 (masker will mask).
- Procedure:
  - fix signal, change masker frequency and intensity
  - find intensity at each masker freq that masks signal
  - intensity profile 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
Speech Perception

- Much higher level in auditory cortex or brain - not well understood
- Goal: find invariant acoustic cues for different speech sounds
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. Is there an ideal realization?
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. Is there an ideal realization?
- Difficult since acoustic cues for an object change depending on context, other perceptual modalities (McGurk effect), prior beliefs, etc.
Cues depend on context

Graph showing formant transitions for the sounds di and du.
Cues depend on context

- cues for different “d” sounds are different depending on context
Cues depend on context

- cues for different “d” sounds are different depending on context
  - F2 totally different
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 even considering acoustics, percept is dependent on more than just acoustic cues at the current time. Hence, must be more than just formants that are used to determine words. E.g., temporal patterns, long context, adaptation (time dependence, dependence on what you've previously been exposed to), etc.
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 even considering acoustics, percept is dependent on more than just acoustic cues at the current time.
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 even considering acoustics, percept is dependent on more than just acoustic cues at the current time.

Hence, must be more than just formants that are used to determine words. E.g., temporal patterns, long context, adaptation (time dependence, dependence on what you’ve previously been exposed to), etc.
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
- Acoustic/Phonetic level: group convergence to common rate-of-speech, pitch range, style of turn timing, loudness quality, prosodic pattern (J. Local, 2003).
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
- Acoustic/Phonetic level: group convergence to common rate-of-speech, pitch range, style of turn timing, loudness quality, prosodic pattern (J. Local, 2003).
- Linguistic level: conversations are “easier” to ascertain than monologues (cell phones) because of alignment processes such as entrainment (Brennan & Clark, Pickering & Garrod).
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
- Acoustic/Phonetic level: group convergence to common rate-of-speech, pitch range, style of turn timing, loudness quality, prosodic pattern (J. Local, 2003).
- Linguistic level: conversations are “easier” to ascertain than monologues (cell phones) because of alignment processes such as entrainment (Brennan & Clark, Pickering & Garrod).
- Behavioral level: gestures, mannerisms (e.g., gait), other social features also eventually converge (Dijksterhuis & Bargh, Wyatt et. al.).
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
- Acoustic/Phonetic level: group convergence to common rate-of-speech, pitch range, style of turn timing, loudness quality, prosodic pattern (J. Local, 2003).
- Linguistic level: conversations are “easier” to ascertain than monologues (cell phones) because of alignment processes such as entrainment (Brennan & Clark, Pickering & Garrod).
- Behavioral level: gestures, mannerisms (e.g., gait), other social features also eventually converge (Dijksterhuis & Bargh, Wyatt et. al.).
- Could percept change as a function of recent social context (i.e., at signal level, two identical cues might be interpreted entirely different depending on recent social context)?
Social Structure of Speech

- Turns “influence” each other (i.e., knowing the constituent parts of one turn allows either us, or in theory a statistical model, to make more informed predictions about aspects of another turn).
- Humans interact and influence each other in many different ways.
- Acoustic/Phonetic level: group convergence to common rate-of-speech, pitch range, style of turn timing, loudness quality, prosodic pattern (J. Local, 2003).
- Linguistic level: conversations are “easier” to ascertain than monologues (cell phones) because of alignment processes such as entrainment (Brennan & Clark, Pickering & Garrod).
- Behavioral level: gestures, mannerisms (e.g., gait), other social features also eventually converge (Dijksterhuis & Bargh, Wyatt et. al.).
- Could percept change as a function of recent social context (i.e., at signal level, two identical cues might be interpreted entirely different depending on recent social context)?
- Current ASR systems do not fully exploit these phenomena.
200-5.5 kHz most important — by filtering out spectral regions and measuring intelligibility
200-5.5kHz most important — by filtering out spectral regions and measuring intelligibility

Hence, ISDN: 4kHz BW, early digital speech channel, prior to coding and compression being ubiquitous
200-5.5kHz most important — by filtering out spectral regions and measuring intelligibility

Hence, ISDN: 4kHz BW, early digital speech channel, prior to coding and compression being ubiquitous

formal perceptual theories to determine intelligibility. E.g., the “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/)
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/)
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)
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
we are remarkably good at perceiving speech

Recall the sine-wave speech, and the 1-bit speech from a few lectures ago.
we are remarkably good at perceiving speech

Recall the sine-wave speech, and the 1-bit speech from a few lectures ago.

face recognition, when do we see a face?
we are remarkably good at perceiving speech

Recall the sine-wave speech, and the 1-bit speech from a few lectures ago.

face recognition, when do we see a face?

therefore, hard to identify most important cues, since they all could potentially be used depending on context
Human Speech Perception

- Speech contains much redundant information, much can be removed w/o impacting intelligibility – E.g., checkerboard speech
Human Speech Perception

- Speech contains much redundant information, much can be removed w/o impacting intelligibility – E.g., checkerboard speech
- Spectral transitions, derivatives might be key?
Human Speech Perception

- Speech contains much redundant information, much can be removed w/o impacting intelligibility – E.g., checkerboard speech
- Spectral transitions, derivatives might be key?
- Gaussian Scaled Speech (example)
Speech contains much redundant information, much can be removed w/o impacting intelligibility – E.g., checkerboard speech

Spectral transitions, derivatives might be key?

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)
Speech contains much redundant information, much can be removed w/o impacting intelligibility – E.g., checkerboard speech

Spectral transitions, derivatives might be key?

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 acoustic situations — e.g., background noise, cocktail party effect
Complicated tube analysis gives us LPC.
Summary

- Complicated tube analysis gives us LPC.
- Perhaps we can hence start with LPC analysis of speech and use that.