Speech perception starts when speech sounds enter ears...

1. **pinna** modifies sound behind you so you can localized front/back

2. **eardrum** connected via bones in middle ear to membrane of **cochlea** (protects cochlea from Q-tips, decorative 4th of July flags stuck in ears)

3. phase-locked delay line carry low freq from cochlea to **nucleus laminaris**
   difference in phase helps triangulate sound source in binaural audition
   (only works for waves larger than 1/2 size of head)

4. vibrations in cochlea vibrate small/med/large **cilia** in **basilar membrane**
   (these are killed when you go to concert, tone afterward is tinnitus)

5. cilia in basilar membrane connected to neurons
   (can stimulate artificially with cochlear implants)

6. spatially-encoded neural stimuli form features in concept space
   (like color, but resolve with 40 or so dimensions)
Spectra

We can ‘see’ speech / other sound using spectrogram (~how brain ‘sees’ it)

- multiply signal by sin, cos fn at different frequencies to get +/- resonances (similar to hairs physically resonating with different signals) (sin, cos differ by 1/4 phase, triangulate magnitude of signal at any phase)

  - main fundamental frequency is larynx: note speaker is singing/saying

  - also harmonics at each integer multiple above fundamental frequency like pushing swing (actually, pushing hair) every $n^{\text{th}}$ time still works

- first big peak across harmonics is first formant: resonance in pharynx

- second big peak is second formant: resonance in oral cavity

- above that, a few more formants (timbre)

- then frication noise, from back obstr. to front: ch/k, sh, s/t (telephone cuts off at 8kHz, can’t tell /s/ from /f/)
Phoneme Recognition

Phonemes are partially defined by formant/frication frequencies:
- synthetic speech with Lieberman & al’52 machine understood as speech
- sine wave speech also understood as speech
- phonemes correspond to contiguous regions of formant space (classifications of sound, as words are classifications of ideas)
- categorical perception / perceptual magnet effect: detect difference between phonemes better than within phonemes

But speech does not consist of simple sequence of phonemes:
- no stop sound, isolating tape just gives chirps
- voiced/unvoiced stops both technically unvoiced, differ only in VOT
- phonemes coarticulated (distrib. across signal):
  - progressive assimilation: frication in seat higher freq. than in suit
  - regressive assimilation: frication in key higher freq. than in koo
  - regressive assimilation: vowel in con more nasal than cop
  - regressive assimilation: cannonball $\rightarrow$ /kænəmbɔl/
Phoneme Recognition

Phonemes are not fixed classes either:

- variability across speakers
  - age, sex produce different fundamental frequencies, formants
  - accent changes phoneme characteristics
  - voice characteristics (voices are identifiable)

- variability across utterances
  - different speed (/b/ in slow speech equals /p/ in fast speech)
  - different emotional state raises/lowers frequencies
  - ambient noise
    - other voices, traffic, room walls, ... masks speech characteristics
    - Lombard effect: increase loudness, pitch, duration

- context
  - speakers increase distance from neighboring phones/words
  - whisper
  - sarcasm, humor, dopey voice
Phoneme Recognition

Phoneme recognition takes information from several sources

- from visual cues:
  - McGurk effect: play audio of /ga/, video of /ba/ → subjects hear /da/ (Why /da/? Closer in sound to /ga/ than /ba/, but has closed mouth)

- from vocal stress

- from orthography: absurd

- from language predictions:
  - lex knowledge in phone reconstruction: ‘s_lice’ → splice (not stlice, etc)
  - frequency: ‘a girl with kaleidoscope eyes’ → a girl with colitis goes by
  - semantic: ‘They hae slain the Earl o’ Moray and layd him on the green’ → They hae slain the Earl o’ Moray and Lady Mondegreen
  - desire for innuendo: master mate → /mæstəmbaɪt/ (Captain Pugwash)
  - ‘Mad Gab’ game (?)

Bottom-up / top-down processing (we’ll see later)

Speech perception interleaves with production

- split utterances: Alan: we didn’t finish... Bill: our sentences? I noticed.