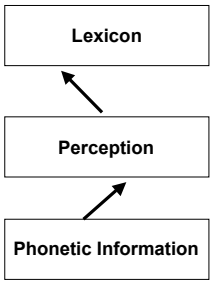
Week 8: The perception of multimodal phonetic cues

Summary of how speech perception is challenging

- The cues for phonemes are mixed, not like "beads on a string"
- Large within-talker variability; listeners can say the same thing different ways
- Large between-talker variability
- Noise can obscure acoustic cues
- No acoustic cues are necessary or sufficient

More of a challenge for science, than for the listener

Next two weeks: How we meet the challenges



- Perception (*low level)*
  - Auditory perception (today & next week)
  - Visual perception (today)
- Language structure (*high level*)
  - Linguistic processes (next week)

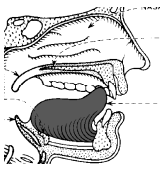
How we meet the challenge of speech perception

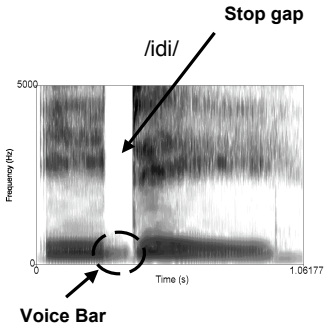
Part 1: We use multiple acoustic cues

Example of multiple acoustic cues: production of voiced plosives

Step 1: Complete closure of the vocal tract

- Blocks flow of air through the oral cavity (impeding vocal fold vibration) and absorbs acoustic energy
- Produces stop gap and voice bar

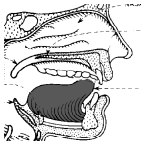




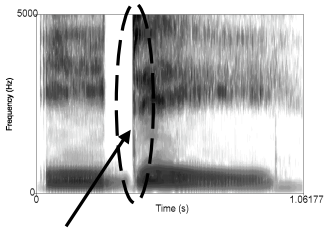
Example of multiple acoustic cues: production of voiced plosives

Step 2: Release of the closure

- Air rushes out through opening
- When only partially open, frication energy (I.e., noise due to turbulence) is produced.
  - Seen as a ***burst*** on a spectrogram



/idi/

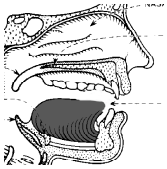


Burst

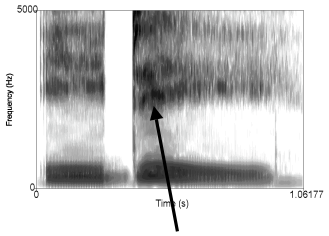
Example of multiple acoustic cues: production of voiced plosives

Step 3: Onset of voicing

- Air pressure is released, so vocal folds can vibrate again
- Articulators move into position for next phoneme
  - Movement of transitions change resonant frequencies, which are seen as ***formant transitions***



/idi/



Formant transitions

A single articulation thus has many acoustic consequences…

- Stop gap: Plosive manner (wide frequency range)
- Voice bar: Voicing (low frequencies)
- Burst: Place of articulation (variable frequency)
- F2 and F3 formant transitions: Place of articulation
- Voice onset time: Voicing (wide frequency range)
- F1 formant transition: Voicing and manner

Listeners pay attention to multiple cues (pieces of evidence about the articulation), rather than relying on only a single cue.

Multiple cues means that some information is *redundant* (i.e., several cues mean the same thing)

This is a big advantage for speech perception because if we cannot hear all of the cues, we may be able to still recognize speech based on what we *can* hear.

- Cues can be knocked out due to noise or hearing impairment
- Talkers do not always produce all of the cues clearly (i.e., they speak to be understood)
- Talkers with different accents may produce a different set of cues than you expect to hear

Just because a cue is *available* to listeners, it does not mean that they use it for speech recognition…

Listeners *weight* cues differently (i.e., depend on them to greater or lesser extents)

Cue weightings develop during infancy and childhood, as the individual works out what combination of cues gives the 'right' answer

Acoustic analyses tell us what cues are in the signal, but not how someone uses them to recognize speech

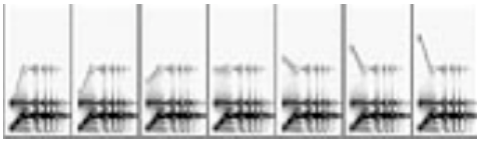
For that, we need to run *perceptual tests*

How do you determine what acoustic pattern info is important for perception?

- Why not just use natural speech?
  - Can be good at measuring real-world performance, but does not give us much control over the acoustic variation
- Use of synthetic speech
  - Controls acoustic cues
- Use of 'controlled' tests
  - Evaluate the perceptual effect of one or more speech patterns
  - Use a 'speech continuum'

How do you determine how people use an acoustic cue?

- Construct a 'speech continuum' that varies a particular acoustic cue



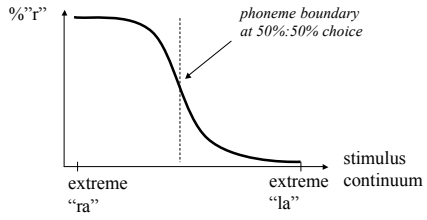
Step 1 Step 2 Step 3 Step 4 Step 5 Step 6 Step 7

How do you determine what constitutes an acoustic cue?

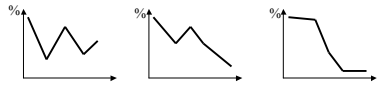
- Then, test whether the cue affects identification.

Is this /ra/ or /la/?

Labeling Graph



Labeling graphs and acoustic cues



•Cues that are most important to listeners (*primary* cues) will have *categorical* labeling functions.

•Cues that are less important to listeners (*secondary cues*) will have more progressive labeling functions

•Cues that are unimportant to listeners will have random or flat labeling functions

Why is it important to know how people perceive acoustic cues?

- Scientific reasons
  - Understand how speech perception works
  - Understand how language experience shapes perception
- Clinical applications
  - Enhancement of acoustic cues for hearing-impaired children
  - Evaluation of cause of perceptual difficulties (e.g. children with SLI)

How we meet the challenge of speech perception

Part 2: We use lipreading cues

Information provided by visual cues

- Global (non-phonetic) information
  - WHO is speaking
  - WHERE person is speaking
  - WHEN person is speaking
  - HOW (facial expression)
- Segmental information
  - Phonetic cues

Segmental information

- We are good at perceiving ***place*** information through lipreading *(demo)*
  - Harder to perceive place information for articulations inside the oral cavity *(demo)*
- ***Voicing*** information is invisible
  - Cannot see the vocal folds vibrating
- ***Manner*** information is partially visible
  - Can see some differences in timing and liprounding
  - Cannot see uvular flap (nasals)

Segmental information

- Because place and manner are hard to see, there are groups of phonemes that essentially look the same
  - /b/, /p/, /m/ (*demo*)
  - /d/, /t/, /s/, /z/, /n/, /l/ (*demo*)

Visemes: Groups of phonemes that look the same during lipreading

Audio/visual integration

- C*omplementarity* between auditory and visual cues
  - visually Place ☺ Manner ☻ Voicing ☹
  - auditory Place ☹ Manner ☻ Voicing ☺

Speech perception is *multimodal*

- e.g., Audition+vision, Audition+touch
- Notion of '*superadditivity*': speech perception with 2 sources of information is often greater than predicted on the basis of intelligibility for each source alone

Visual/Auditory cue integration

- Issue: Is visual information used as a backup, or are both Audio and Visual evaluated before a label is given to the sound?
- Test? McGurk experiments (*demo*)

McGurk effect

- Visual /g/ + Auditory /b/ = Perceived /d/
  - Visual cues
    - Tells us consonant ***could not*** be /b/ (because /b/ is so visible),
    - but it ***could*** be either /d/ or /g/ (hard to distinguish articulations in the mouth)
  - Auditory cues
    - Sounds like /b/,
    - but /d/ is the next closest consonant
  - We perceive the combination as /d/
    - It is the only phoneme that is possible given this combination of auditory and visual cues
    - We perceive this combination ***even though the auditory information is not ambiguous***
- Everyone who can see uses lipreading information, even people with normal hearing
  - Helps with unfamiliar accents, noise, or semantically difficult content

Can we use visual cues to help people with hearing impairments?

- When speaking to someone with a hearing impairment...
  - Be visible (face them in good light)
  - Open your mouth
  - Shave your moustache!
- Technological solutions for using telephones (***demo***)

Summary of cue perception

- Lipreading: Very good place, some manner, no voicing
  - Visemes
- Speech is multimodal
  - We combine cues across modalities as easily as we combine different acoustic cues
  - Lipreading gives us even more cues
  - We use lipreading cues even when we can clearly hear all of the auditory cues

Today's Lab: VCV Test

Test yourself on audio, visual, and audiovisual speech perception in order to examine how you perceive acoustic cues.