LING5702: Lecture Notes 7
Speech perception and phoneme recognition

Contents

7.1 Speech perception

Speech perception starts when speech sounds enter ears...

1. The **pinna** modifies sound behind you so you can localize front/back.
2. The **eardrum** is connected via bones in the middle ear to the membrane of the **cochlea**.

   (It protects the cochlea from water, bacteria, Q-tips.)
3. Phase-locked delay line carries low frequencies from cochlea to **nucleus laminaris**.

   The difference in phase helps triangulate the sound source in binaural audition.

   (It only works for waves larger than 1/2 size of your head.)
4. Vibrations in the cochlea resonate in locations proportional to wavelength of sound

   (High sounds w. short wavelength near entrance, low sounds w. long wavelength further in.
5. Vibrations in the cochlea vibrate small/med/large **cilia** in the **basilar membrane**.

   (These are killed when you go to a concert; the ringing tone afterward is tinnitus.)
6. Cilia in the basilar membrane are connected to neurons.

   (These can be stimulated artificially with cochlear implants.)
7. Spatially-encoded neural stimuli form features in a concept space.

   (These are like color, but resolve with 40 or so dimensions.)

7.2 Spectra

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

- We can multiply the signal by sin,cos function at different frequencies to get +/- resonances.

  (These are similar to hairs physically resonating with different signals.)

  (Sin,cos differ by 1/4 phase, triangulate magnitude of signal at any phase.)

william-iyihehaeuwuhoaaa.pdf

[i ɪ ɛ æ u ə ɔ a l r]

- Main **fundamental frequency** comes from the larynx: the note speaker is singing/saying.

williamehbehdehgeh-mono.jpg karinehbehdehgeh-mono.jpg claireehbehdehgeh-mono.jpg

male [ɛ b ɛ d ɛ g ɛ] female [ɛ b ɛ d ɛ g ɛ] child [ɛ b ɛ d ɛ g ɛ]

- Also, **harmonics** at each integer multiple above fundamental frequency.

  Like pushing a swing (actually, pushing hair); pushing every n^{th} cycle still works.
- The first big peak across harmonics is **first formant**: resonance in pharynx.
- The second big peak is **second formant**: resonance in oral cavity.
- Above that, a few more formants (timbre).
- Then **frication** noise, from back obstruction to front: ch/k, sh, s/t.

william-fvszshzhthdh.jpg

[f v s z ʃ ʒ θ ð]

(Telephones cut off at 8kHz, can't tell /s/ from /f/.)

7.3 Phoneme recognition

Phonemes are partially defined by formant/frication frequencies:

- Sine wave speech is 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:

  We can detect difference between phonemes better than within phonemes.

But speech does not consist of simple sequence of phonemes:

- There is no sound for stop phonemes; isolating them on a tape just gives chirps.

  williamehbehdehgeh-mono.pdf

  [ɛ b ɛ d ɛ g ɛ]

- Voiced/unvoiced stops are both technically unvoiced, differ only in VOT.
- Phonemes are coarticulated (distributed across signal):
  - progressive assimilation: frication in seat higher frequency than in suit
  - regressive assimilation: frication in key higher frequency than in koo
  - regressive assimilation: vowel in con more nasal than cop
  - regressive assimilation: cannonball → /kænəmbɔl/

Phonemes are not fixed classes either:

- There is variability across speakers:
  - age, sex produce different fundamental frequencies, formants
  - accent changes phoneme characteristics
  - voice characteristics (voices are identifiable)
- There is 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 takes information from several sources:

- from visual cues:
  - McGurk effect: play audio of /ba/, video of /ga/ → subjects hear /da/

    (Why /da/? Closer in sound to /ba/ than /ga/, but has open mouth)
- from vocal stress
- from orthography: apsurd
- 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

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

Speech perception interleaves with production

- split utterances: A: we didn't finish... B: our sentences? I noticed.