

## Speech Perception

9.35

Josh McDermott

Question: Why are we studying speech in a perception class?

I mean, isn't language some high-level cognitive thing?

Answer: Speech is received by the brain as a sound signal.

Perceptual processes must transform the sound signal into a form that semantic and syntactic processes can handle.

This requires solving some very difficult problems.

## Phonemes

Phonemes are the smallest unit of sound that can make a difference in the meaning of speech.

e.g. pot vs. dot

\*\*\* phonemes are not letters\*\*\*

We can think of the problem of speech perception as that of extracting a string of phonemes from the speech signal.

This is hard to do.

## Speech Articulation

- **Specialized for speech**
  - Creates choking hazard.
  - Breathing affected

(Image removed due to copyright considerations.)

## Source-Filter Model

- **larynx: buzzy sound source**
  - Show movie
- **Changeable resonators filter the sound produced:**
  - pharynx (throat);
  - mouth
  - lips
  - nose

(Image removed due to copyright considerations.)

## Key Properties of Speech

- **Fundamental frequency ( $F_0$ )**
  - Men: 80-240Hz
  - Women: 140-500Hz
  - Children: 170-600Hz(determined by length and thickness of vocal chords)
- **Harmonics**
- **Resonators – cavities amplify certain frequencies and dampen others**
  - Bigger cavities = low sounds
  - Smaller cavities = high sounds

(Image removed due to copyright considerations.)

## Key Properties of Speech

- **Formants ( $F_1$ ,  $F_2$ , etc.) – Strongest frequencies**  
(Result from the size and shape of the resonating cavities)
- **Sound is modulated by manipulating the articulators.**
  - Changes resonance properties (frequencies of formants)
  - Changes airflow.

## Phonemes of the world

40 phonemes in English

Range: 11 in Polynesian – 141 in Khoisan (“Bushman”)

Total inventory across languages: thousands

However, some are very common across all languages (e.g., /m/, /n/, /t/, /d/, /k/, /g/, /s/, /z/).

## Phonemes

Consonants: Restricted vocal tract

1. place of articulation (dental vs. velar etc.)

2. manner of articulation (stop vs. nasal vs. fricative etc.)

3. voicing (voiced, unvoiced)

## Examples: stops

- /b/: voiced, labial, stop
- /p/: unvoiced, labial, stop
- /d/: voiced, dental, stop
- /t/: unvoiced, dental, stop
- /g/: voiced, velar, stop
- /k/: unvoiced, velar, stop

### Examples: fricatives and nasals

- /z/: voiced, dental, fricative
- /s/: unvoiced, dental, fricative
  
- /m/: voiced, labial, nasal
- /n/: voiced, dental, nasal

### Phonemes, continued

Vowels: Unrestricted vocal tract

Different vowels are distinguished by how the sound produced by the vocal cords is filtered.

Resonances are altered in two ways:

1. part of tongue (front vs. back)
  - bet vs. butt
2. position of tongue (high, middle, low)
  - beet vs. bat

Changing the resonances alters the formants of the vowel:

(Image removed due to copyright considerations.)

### Speech spectrogram.

Darkness indicates intensity of sound at a given moment at a given frequency.

(Image removed due to copyright considerations.)

### “I can see you”

- Note different types of segments and what they look like.
  - Stops vs. Vowels
  - Fricatives
    - White noise
- Generally it is not clear where one word begins and another ends.

Formant transitions characterize consonants;  
formant positions characterize vowels.

(Image removed due to copyright considerations.)

Question: what happens when you whisper?

Vowels can be perceived just fine, but the vocal chords do not vibrate...

So, all we have to do is build detectors for each phoneme and we're set, right?

Nope. Phonemes are produced differently depending on many factors. This results in a constancy problem much like those in vision (e.g. object recognition across viewpoints).

### Coarticulation

Phonemes are produced differently depending on what comes before and after them.

But they sound the same!

Somehow they are recognized as the same despite producing different patterns of sound energy.

(Image removed due to copyright considerations.)

### Other factors affecting sound

- **Prosody**
  - **Stress** – prominence within words
    - *perMIT* as a verb
    - *PERmit* as a noun
  - **Intonation** – Variations in pitch across a phrase
    - *Dad wants me to mow the lawn.*
    - *Dad wants me to mow the lawn?*
- **Emotional State**
  - Smiling
  - Frowning
  - Stressed

### Other factors affecting sound

- **Different speakers sound different**
  - Accents
  - Gender
  - Age

So the same phoneme may be realized in many, many different ways.

### Solutions to speech perception

There are *some* invariants:

- **Stops**
  - bursts
- **Fricatives**
  - Turbulence – broad spectrum energy
- **Vowels**
  - Steady state formants
  - relations between formants
- **Nasals**
  - Low frequency band of energy along with absence of high frequency noise
  - /m/ and /n/ differ in formant transitions

### Solutions: Categorical Perception

- We impose categories on physically continuous stimuli, which aids their detection.

### In-class demonstration: the /ka/ - /ga/ continuum

- Voicing: differences in Voice Onset Time (VOT)
  - Small VOT: voiced; Large VOT: unvoiced
- (Image removed due to copyright considerations.)

### /ga/ - /ka/ in-class demonstration

1. 0 msec (/ga/)
2. 70 msec (/ka/)
3. 60 msec (/ka/)
4. 30 msec (usually /ga/)
5. 10 msec (/ga/)
6. 20 msec (/ga/)
7. 40 msec (usually /ka/)
8. 50 msec (/ka/)

### % labeled /ga/ in /ga/-/ka/ continuum

(Image removed due to copyright considerations.)

Categorical perception. Discrimination is best at a category boundary.

(Image removed due to copyright considerations.)

### What Good is Categorical Perception?

Helps to

- Ignore irrelevant information
- Quickly classify transient events
  - consonants versus vowels

### Solutions: Knowledge of Words

- People are biased to hear phonemes that would result in a known word.

-beef/peace demo.

### Solutions: Visual Input

- McGurk effect  
Visual input (lipreading) is integrated with auditory input to determine the phoneme that is perceived.

Show demo.

### McGurk effect: vision & speech interact.

(Image removed due to copyright considerations.)

### Summary: Problems in Phoneme Recognition

- Problem
  - Lack of invariance
- Solutions
  - Acoustic features
  - Categorical perception
  - Visual input
  - Context

### Segmenting words is also hard:

Here the words are artificially separated.

(Image removed due to copyright considerations.)

In real speech everything runs together.

(Image removed due to copyright considerations.)

### No physical boundaries between words

- Not like written language
  - Mares eat oats and does eat oats and little lambs eat ivy. A kid'll eat ivy too. Wouldn't you?
- Oronyms
  - *It's a doggy-dog world*
  - *I don't really think it's a parent.*
  - *The girl with kaleidoscope eyes*
- Top down information
  - We need a decanter.*
  - We needed a cantor.*
  - Context influences lexical processor