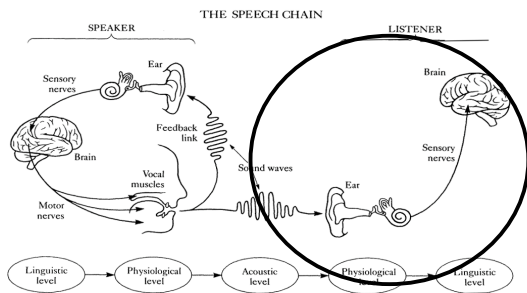
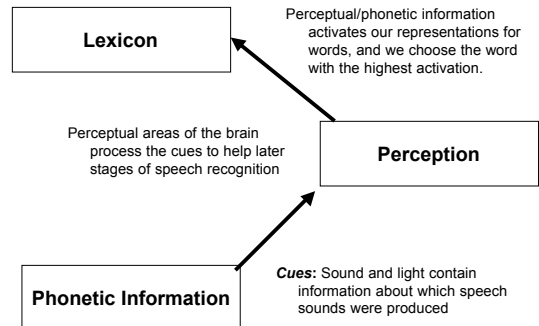


## Week 7: The challenge of speech perception



## A simple model of speech recognition



## The rest of the course

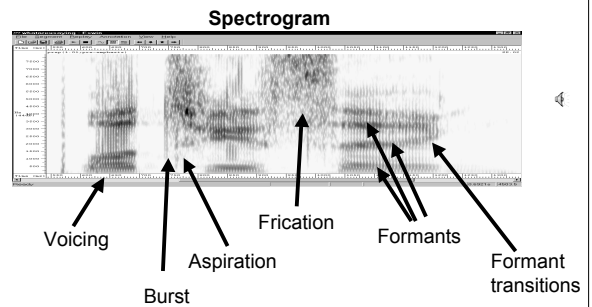
**Today:** Why speech perception is a challenge, and more about acoustics

**Week 8:** Low-level solutions

**Week 9:** Lipreading, plus high-level solutions (i.e., linguistic processing)

**Week 10:** How we learn how to recognize speech

## Review: Acoustic cues



## A simplistic view of perception from what you know about acoustics

- Speech recognition just involves checking off a list of acoustic cues
- e.g., /b/ = stop gap, short VOT, and low locus frequencies
- Cues can be easily heard, just like they can be read from spectrograms

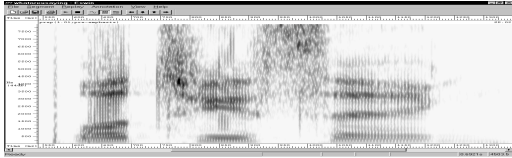
**Wrong!!!**

## Ways in which speech perception is challenging

(1) "Phonemes are not like beads on a string" -- Hocket

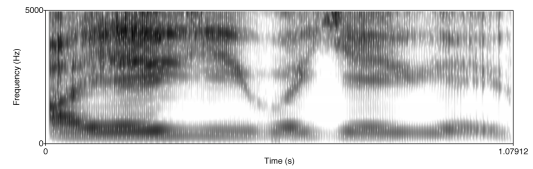
Movie example

## The segmentation issue



No clear markers for the beginning or end of words or phonemes

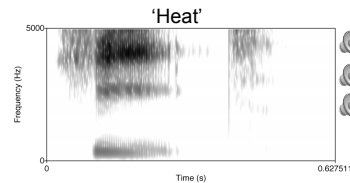
## The segmentation issue: Another example



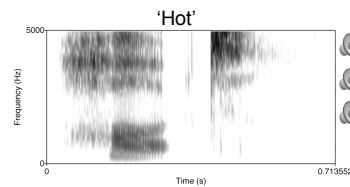
## Coarticulation

- The acoustic realization of phonemes is affected by neighboring sounds...

## Coarticulation Example 1

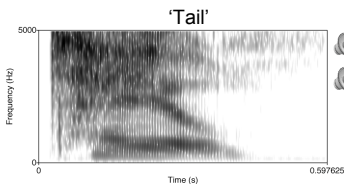


- Acoustic form of /h/ takes on the resonances of the following vowel
- The /h/ sounds the same in each vowel despite acoustic differences



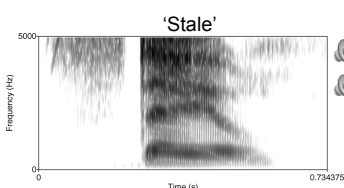
- We can hear the differences in isolation and in 'cross splices'

## Coarticulation Example 2



- Initial /t/ is aspirated with a long VOT. The /t/ in /st/ is unaspirated with a short VOT.

- They sound the same.
- The /t/ in 'stale' sounds like /d/ when the /s/ is removed
- The /t/ in 'tail' sounds unnatural when an /s/ is spliced on



## Coarticulation

- The production of phonemes interacts with the production of neighboring phonemes
- The acoustic cues are broadly distributed in time
- Makes segmentation hard, because there are no clear dividing lines where the cues for one phoneme starts and the next begins
- The listeners must somehow learn that two acoustically different phonemes are meant to sound the same.

## Ways in which speech perception is challenging

(2) "We speak to be understood, and only to be understood" -- Passy

## Where did I leave my keys?

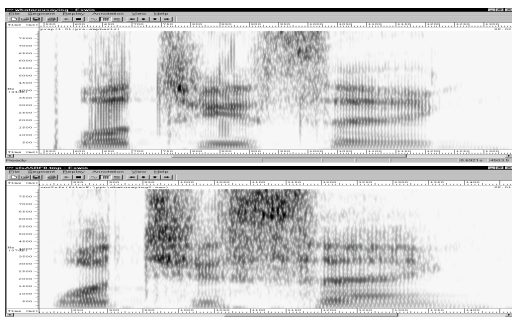
I do not know

I don't know

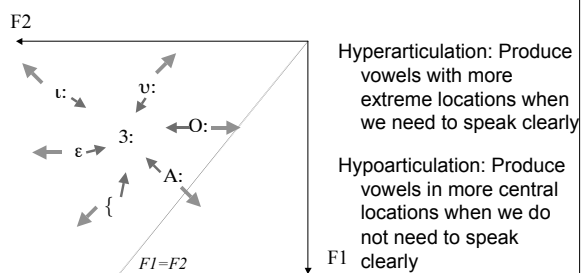
I dunno

(pitch)

## Within-talker variability



## Hyper and Hypo articulations of vowels



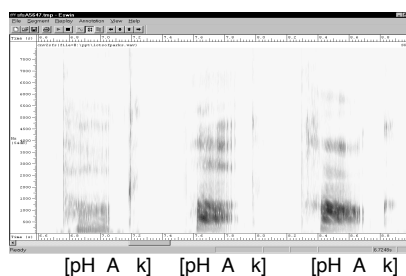
## Ways in which speech perception is challenging

(3) Between-talker variability

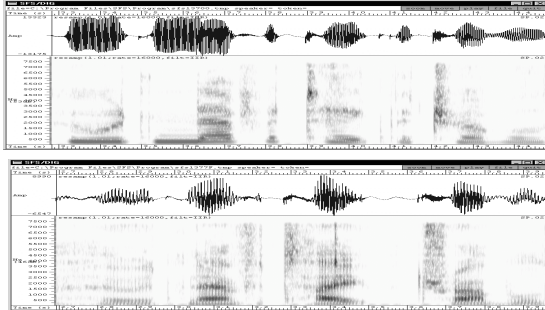


## Between-talker variability

What do these patterns have in common?



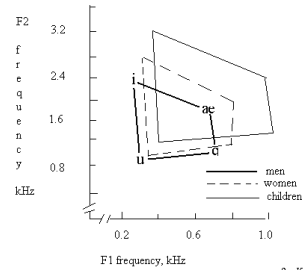
## Female and male talker variability



who would never take the trouble...!

## Vowel spaces and vocal tract size

- Formant frequencies are higher and the vowel space is larger for shorter vocal tracts
- The same formant frequencies will be different vowels for different talkers

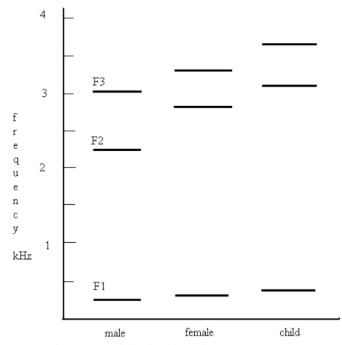


after Kent and Read, 2002

From <http://users.ipfw.edu/dalbyj/>

## Vowel spaces and vocal tract size: Another view

These are *formant frequency* differences, not differences in pitch or fundamental frequency!



Average frequency of the first three formants for the vowel [i].

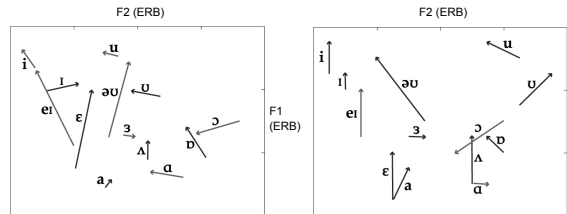
Data from Peterson and Barney, 1952

From <http://users.ipfw.edu/dalbyj/>

## Vowel spaces and accent

Southern British English

Native German Speaker's English Vowels



Iverson & Evans, 2003

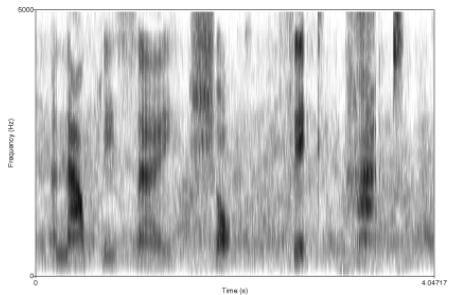
## Some sources of talker variability

- The acoustic form of speech varies a great deal because of:
  - Anatomy (e.g., vocal tract size)
  - Accents
  - Chosen speaker gestures

## Ways in which speech perception is challenging

(4) Noise and channel effects


## Speech in noise




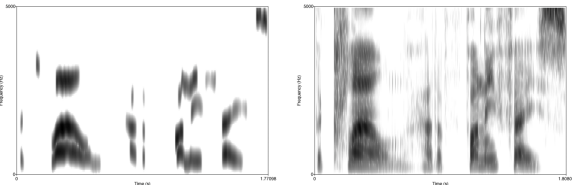
## Ways in which speech perception is challenging

(5) No necessary or sufficient features

## Sinwave Speech


sinwave 


regular 

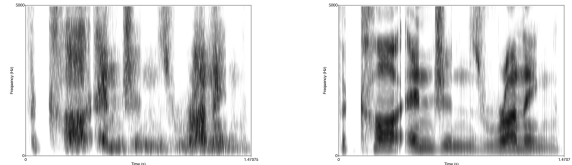


Formants can be replaced with sinwaves:  
Completely unnatural source  
No 'normal' acoustic cues

## Cochlear implant simulations

simulation 

regular 



Speech can be replaced by bands of noise:  
Completely unnatural source  
Poor spectral information

## Summary of the challenges

- 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

## Despite all this variability...

- High speed of processing
  - 25-30 phones a minute
- Amazing ability to 'normalise' across speakers.
  - Example of sentence with every sound spoken by a different speaker

## **How do we do it?**

- No one knows the full picture
- Rest of the term will be about the parts that we *do* know
  - Phonetic perception (lecture 8)
  - Linguistic effects (lecture 9)
  - Development and 2nd language acquisition (lecture 10)