Speech and speech processing

9.59 / 24.905
April 7, 2005

Ted Gibson

The structure of language

Sound structure: phonetics and phonology

"cat" = /k/ + /æ/ + /t/

"eat" = /i/ + /t/

"rough" = /r/ + /^/ + /f/

Language sounds

- wi**n** wi**n**g
- writer vs. rider
- Sounds, not the spelling: “rough” = /r^f/

Summary

- Articulatory properties of speech
  - Distinctive / articulatory features
  - English consonants and vowels
  - Information is smeared between segments: co-articulation
- Speech perception
  - Problems: Lack of invariance, smearing
  - Solutions: Acoustic features; Categorical perception; Motor theory of perception; Use of context
- What aspects of speech are learned / innate?

Phones vs. Phonemes vs. Allophones

- Phones: acoustically different speech sounds
- Phonemes: sounds that make a difference in meaning
  - pot vs. dot
- Allophones: different phones corresponding to the same phoneme
  - Spin vs. pin
  - S[p]in vs. [p^h]in

Source-Filter Model

- larynx: buzzy sound source
- Changeable resonators:
  - pharynx (throat);
  - mouth
  - lips
  - nose

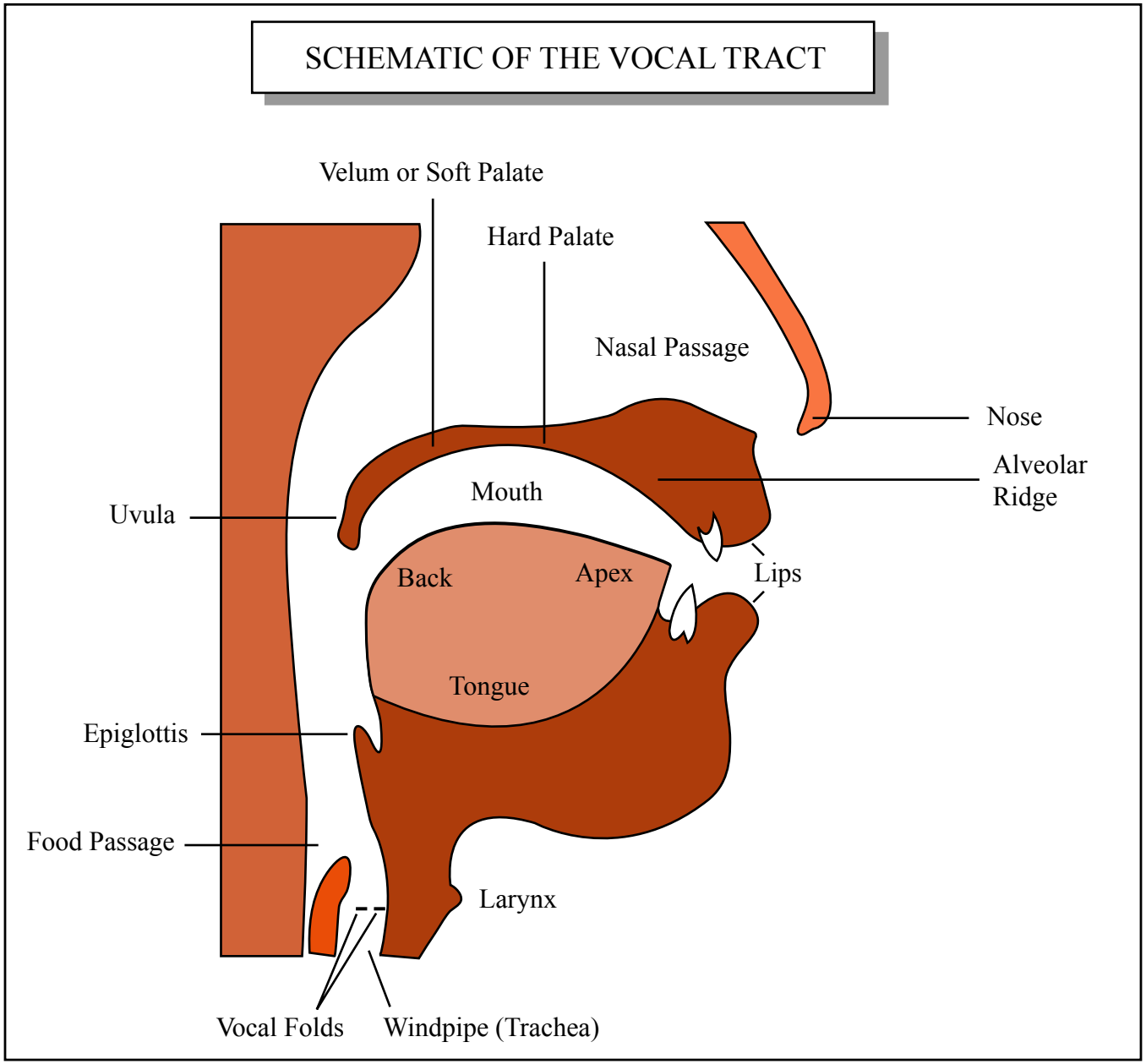


Figure by MIT OCW.

Key Properties of Speech

- Formants of voiced sounds (F_1, F_2, etc.) – Harmonics: Strongest frequencies
  (Result from the size and shape of the resonating cavities)

- Range of human hearing 20Hz-20,000Hz

- Sound is modulated by manipulating the articulators.
  - Changes resonance properties (frequencies of formants)
  - Changes airflow.

Table removed for copyright reasons.
The International Phoentic Alphabet (Phonemes of English).

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/):

Easy to produce, easy to distinguish

Speech sounds: Distinctive/Articulatory features

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)

English Stop Consonants

- /b/: voiced, labial, stop
- /p/: unvoiced, labial, stop

- /d/: voiced, dental, stop
- /t/: unvoiced, dental, stop

- /g/: voiced, velar, stop
- /k/: unvoiced, velar, stop

English Fricatives

- /f/: unvoiced, labio-dental, fricative
- /v/: voiced, labio-dental, fricative

- /s/: unvoiced, dental, fricative
- /z/: voiced, dental, fricative

- /sh/: unvoiced, alveolar, fricative
- /zh/: voiced, alveolar, fricative

English Nasals

- /m/: voiced, labial, nasal
- /n/: voiced, dental, nasal
- /ng/: voiced, velar, nasal

Speech sounds: Distinctive features

Vowels: Unrestricted vocal tract

1. part of tongue (front vs. back)
    - beet vs. boot; bet vs. butt

2. position of tongue (high, middle, low)
    - beet vs. bat; boot vs. bought

Table removed for copyright reasons.
The International Phoentic Alphabet (Phonemes of English).

“The dog snapped”

- The different types of segments and what they look like.
  - Stops vs. Vowels
  - Fricatives
    - White noise
- Generally it is not clear where one segment begins and another stops.
  - Information is smeared

Graphs of frequency vs. time removed for copyright reasons.

Voicing in a Spectrogram: The /ka/ - /ga/ continuum

- Voicing: differences in Voice Onset Time (VOT)
- Small VOT: voiced; Large VOT: unvoiced
- Plosion spike (stop) followed by formants (vowel)

Graphs of frequency vs. time removed for copyright reasons.

Phonemes are not produced serially

- Sounds are not produced serially
  "cat" is not just "/k/ + /æ/ + /t/"
  "eat" is not just "/i/ + /t/"
  "rough" is not just "/r/ + /^/ + /f/"

- Synthesized speech often sounds unnatural
- Parallel transmission
  - Context conditioned variation

Continuous speech

- **Coarticulate**: adjust pronunciation of current sound to take into account preceding and following sounds
  - *kill* vs. *cool*
  - *bog*
- Information for segments overlap so we can get out more in a shorter amount of time
- Fast (~15 sounds/sec): Articulators are not always in the ideal position so we need to cheat

/da/

/dee/

Graphs of frequency vs. time
removed for copyright reasons.

/doo/

Not independent segments, but Features

- Speech is a trajectory through a sequence of articulatory targets
- Rules are conditioned on distinctive features
  - Plural -s

bib	/z/	dog	/z/	dad	/z/	
tip	/s/	tick	/s/	cat	/s/	
				kiss	/iz/	wish /iz/
						pinch /iz/
				hen /z/	till /z/	bay /z/

- Example of assimilation – a feature spreads from one segment to an adjacent segment
  - Makes things easier to pronounce

Speech Perception

Problems for Speech Perception

- Fast, 15 sounds/sec up to 30 sounds/sec in fast speech
- Parallel transmission: Sounds blend into each other
  - Each chunk of signal contains evidence of multiple phonemes
  - Coarticulation

Problems for Speech Perception

- Prosody (suprasegmentals)
  - Stress – prominence within words
    - *perMIT* as a verb
    - *PERmit* as a noun
  - Rate – Changes formant transitions
    - Same sound can be produced for two different phonemes
      - /ba/ vs. /wa/
  - Intonation – Variations in pitch across a phrase
    - *Dad wants me to mow the lawn*.
    - *Dad wants me to mow the lawn*?

Problems for Speech Perception

- Emotional State
  - Smiling
  - Frowning
  - Stressed
- Different speakers

Problems for Speech Perception

- Context-conditioned variation
  - One-to-many variation: Same phoneme may be superficially realized in different ways
  - Many-to-one variation: Different phonemes can have the same sound in different contexts

Summary: Problems in Speech Perception

- Problems
  - Lack of invariance, smearing
- Solutions
  - Acoustic features
  - Categorical perception
  - Motor theory of perception
  - Context
    - Same level
      - Phonemic context, prosodic context
    - High level
      - Syntactic, semantic, lexical knowledge

Solutions to speech perception

There are *some* acoustic invariants:

- Stops
  - Bursts: aperiodic burst of energy in some frequencies
- 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
  - voicing
  - /m/ and /n/ differ in formant transitions

Solutions: Categorical Perception

- For consonants, much of the difficulty of telling sounds apart is at the boundaries among sounds

- We impose categories on physically continuous stimuli

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

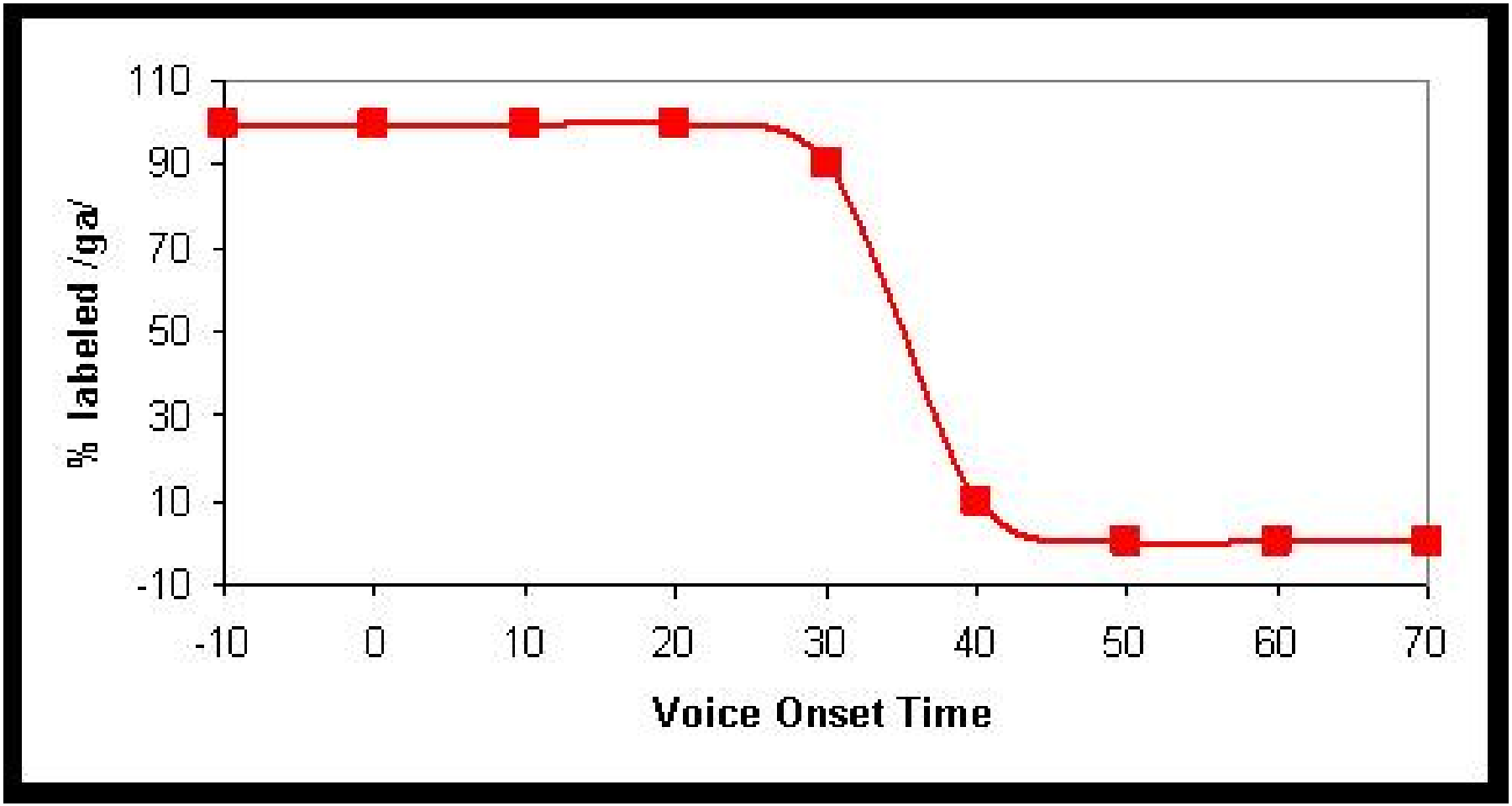
- Voicing: differences in Voice Onset Time (VOT)
- Small VOT: voiced; Large VOT: unvoiced

Graphs of frequency vs. time removed for copyright reasons.

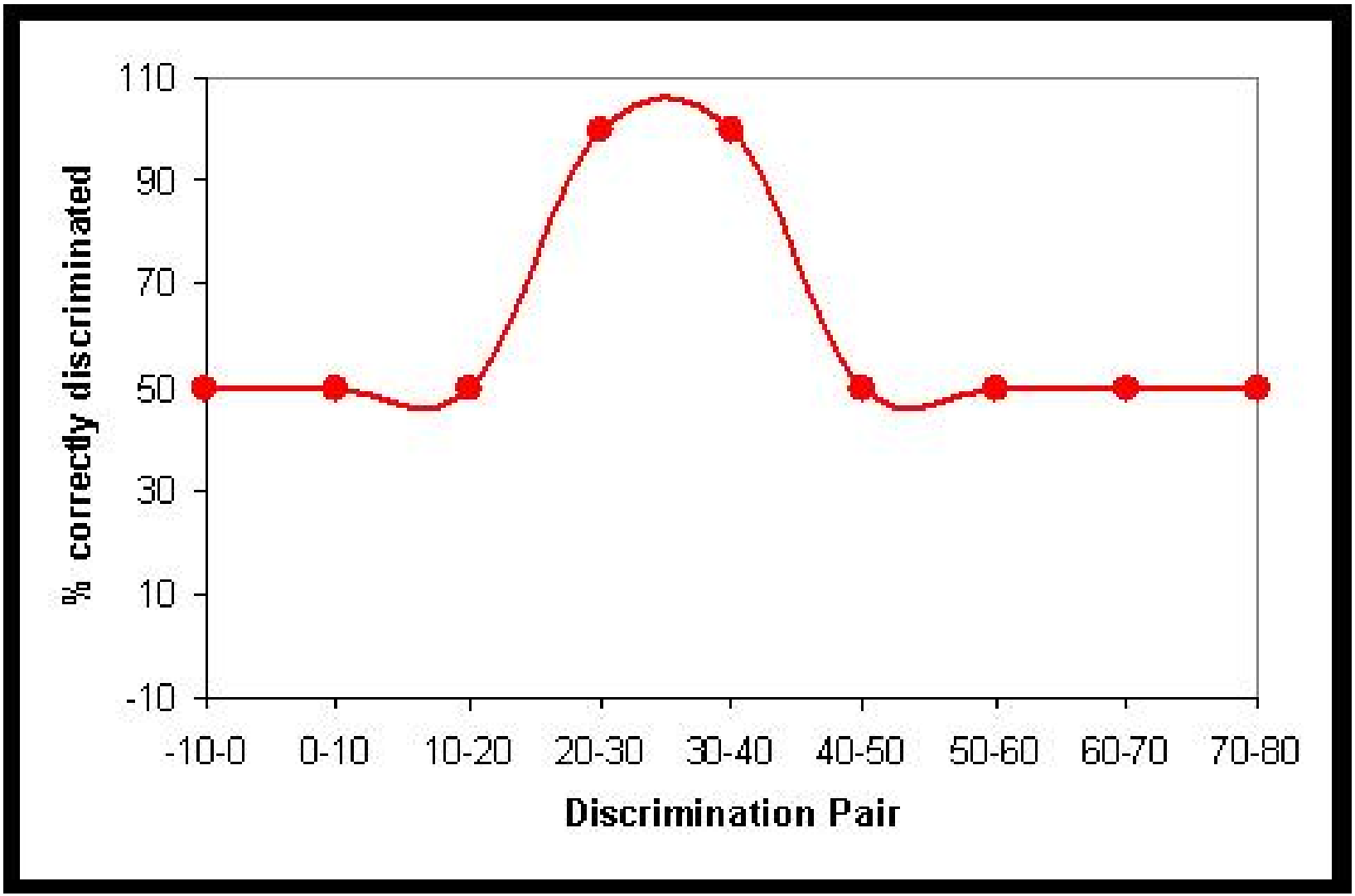
/ga/ - /ka/ in-class demonstration

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



Results of discrimination task: 10 msec intervals of VOT



- **Categorical Perception:** Can't discriminate stimuli any better than you can identify them.
  - Discriminate – tell two things apart
  - Identify – classify a sound
  - Perceptual phenomenon; Not a response strategy

What Good is Categorical Perception?

It helps to

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

Motor Theory of Perception

- McGurk Effect – Visual information automatically integrated into speech percept
- Place of articulation cued by visual input
- Manner cued by ear

Solutions: Phonemic Context

- Use knowledge of how surrounding segments are articulated to interpret ambiguous segments
  - /s/ is higher frequency than /sh/
  - White noise is higher preceding /a/ than /u/
  - A sound halfway between /s/ and /sh/ is interpreted differently depending on whether it is pronounced before a /u/ or an /a/

Graph removed for copyright reasons.

Solutions: Prosodic Context

Rate Normalization

- We correct for speaking rate
  - VOT discrimination
    - Categorical boundary shifts for /ga/-/ka/ if previous syllable is pronounced faster (e.g., short /da/ versus long /da/)
  - Formants
    - /ba/ vs. /wa/
    - If **succeeding** syllable is faster, then percept can change.

Solutions: Higher-Level Context

- Noisy perception (Miller, Heise, Lichten, 1951)

  Grammatical: *Accidents kill motorists on the highways.*

  Anomalous: *Accidents carry honey between the house.*

  Scrambled: *Around accidents country honey the shoot.*

- Shadowing – Echo speech you hear (Marslen-Wilson & Welsh, 1978)
  - Intentional mispronunciations
  - When corrected, they go completely unnoticed and do not delay shadowing
- Use syntax and semantics to perceive the input

Context can Affect Perception

- /pi/ vs. /bi/ demo: lexical knowledge affects categorical boundary

- Not just high-level percept, but perceptual discrimination is affected.

Summary: Problems in Speech Perception

- Problems
  - Lack of invariance, smearing
- Solutions
  - Acoustic features
  - Categorical perception
  - Motor theory of perception
  - Context
    - Same level
      - Phonemic context, prosodic context
    - High level
      - Syntactic, semantic, lexical knowledge