

Speech Perception

Ch. 8 p. 293+

Hearing

Brief overview

Hearing

- Best hearing video ever:
 - <https://youtu.be/PeTriGTENoc>
 - **Phone**: speech sound (term not in book)
 - Unspecified as to phonemic status
 - Phonetics = study of speech sounds (phones)
 - Phonology = study of organization of phones into language-specific categories of meaning and use
-

Speech Perception

- Components
- Big questions

Speech Perception

- Speech perception has two components:
 - **Discrimination:** tell the difference between (the phonetic features of) speech sounds
 - **Identification:** categorize/label speech sounds as phonemes
 - We mentally label big collections of phones as “the same” (our language’s phonemes)
 - e.g. negative to short VOT = English /b/ (but not French!)
 - It’s hard to discriminate between sounds in the same phoneme category

Big (unanswered) Questions

- **Segmentation Problem:** how do we (humans) segment a continuously changing stream of sound into separate units of meaning?
 - No breaks between phones, coarticulation
- **Lack of Invariance Problem:** how do we identify phonemes when they don't have the same acoustic features every time we hear them?
 - Coarticulation, undershoot, different speakers...

Redundancy

- Phonemes have multiple (redundant) **cues**
 - Acoustic features that indicate a phoneme or phonetic feature
- Coarticulation spreads features of phones onto their neighbors
 - Each “segment” in time has cues to the current, preceding, and following phones
 - While finishing one sound, also start making the next
 - Segmentation Problem

Redundancy

- Other linguistic levels also have redundancy
 - Language-specific knowledge: possible/common:
 - Phonotactics (phoneme orders)
 - Lexical (words), syntactic (word order), semantic...
- Very useful for perception, especially in difficult (noisy, distracting) listening conditions
 - If you miss one cue, the others are sufficient
 - Many speech perception experiments cut out or replace cues to figure out what cues are used/needed

Salient Features: Vowels

Salient Features of Vowels

* “salient” = “stands out,” perceptually important

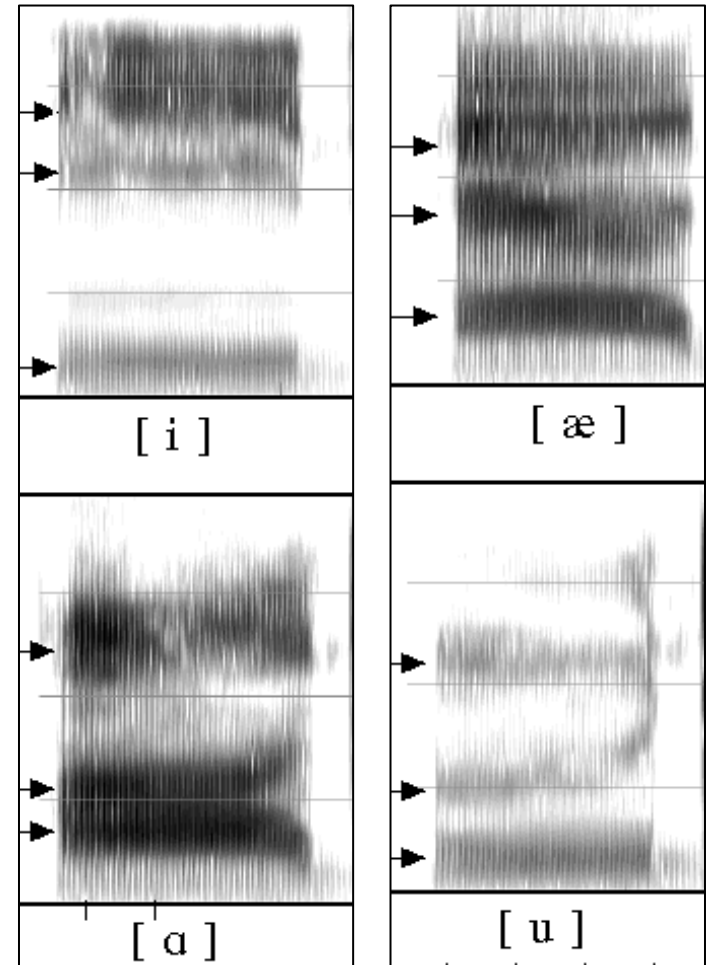
- Duration: longer than consonants
- Intensity: louder than most consonants
 - (Some sibilants may be louder)
- Periodic waveform
 - Feature of **sonorants** (vowels, glides, liquids, nasals) but not **obstruents** (stops, fricatives, affricates)
- Formants = cues to vowel quality/identity
 - Feature of sonorants (and sometimes /h/)

Contributions of Formants

- Raw formant values are not fixed; they differ:
 - Between speakers
 - Due to physiology, social factors...
 - Within speakers
 - Speaking rate, coarticulation, sentence position...
 - Social factors, context, mood, style, register...
- Lack of Invariance Problem
 - So we must use *relative* information about formants rather than *fixed/absolute*

Formant Relationships

- Formant patterns
 - F1, F2 far apart for /i/, close together for /a/
- Formant ratios (F2:F1)
 - Similar across age/sex
 - Large for /i/, small for /a/
 - But others overlap



Vowels in Connected Speech

- Further variability
 - Coarticulation: neighbors change vowel
 - Speaking rate: faster = reduction/undershoot
 - Vowels less distinct from each other
- Steady-state often short or non-existent
 - Formant transitions (direction, slope) and vowel duration more useful than target frequencies
- Context, experience help identify
 - Speaker identity, accent, topic, speech rate...

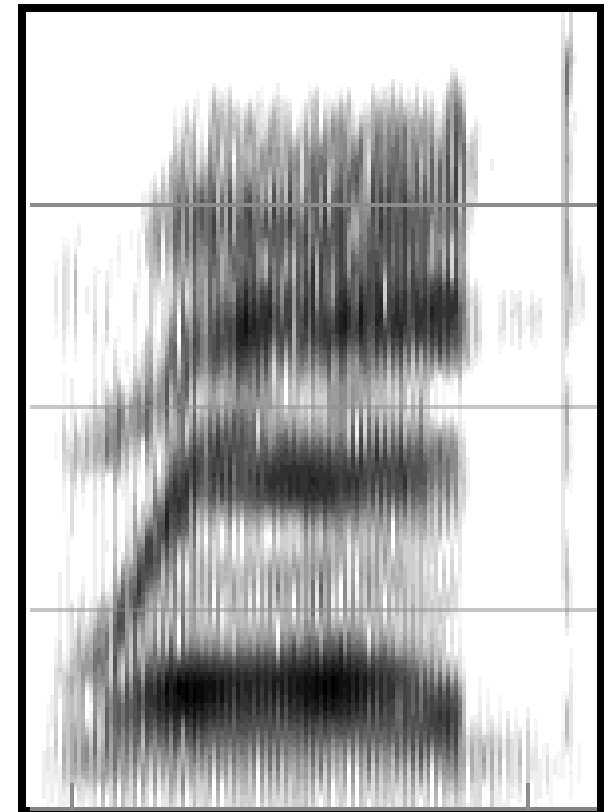
Salient Features: Sonorants

Perception of Diphthongs

- Most salient cue: How fast formants transition between nucleus and glide
 - And direction
 - But not formant frequencies of endpoints
 - Undershoot = “target” of glide is rarely reached

Salient Features of Glides

- Glides: /j, w/: similar to diphthongs but faster formant transitions
 - Duration < 40-60 ms = perceived as stop
 - Between 40-60 and 100-150 ms = glide
 - Duration > 100-150 ms = 2 vowels

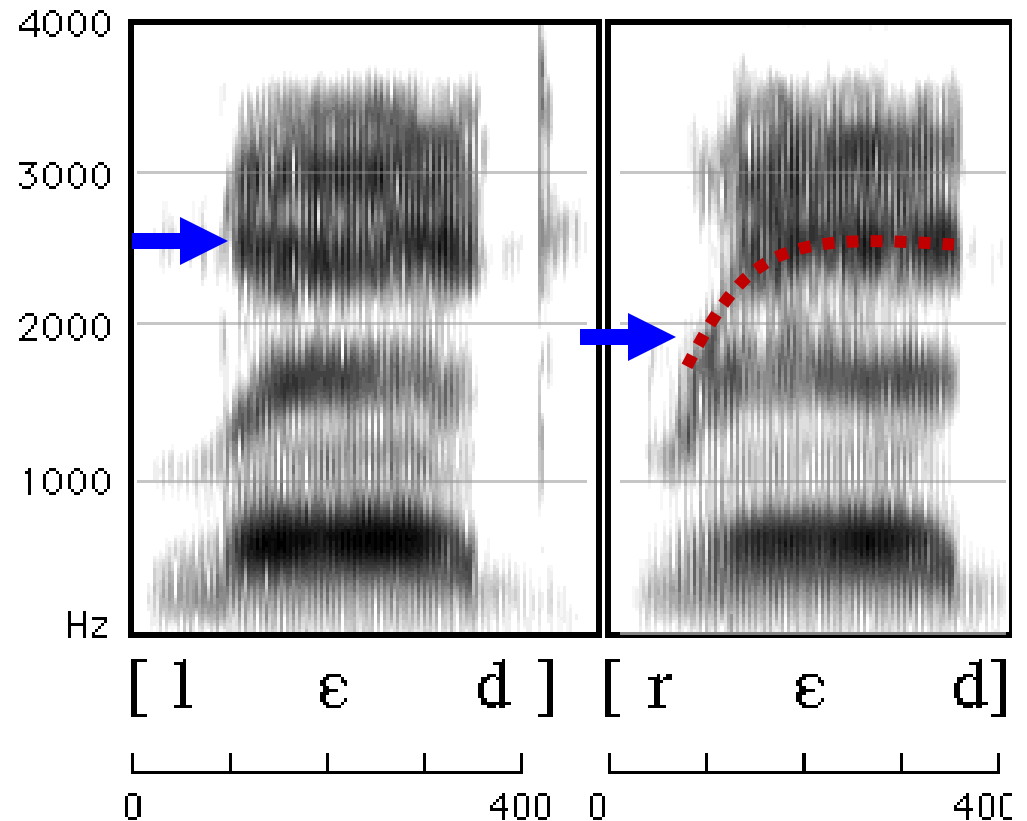


[w ɛ d]

0 400

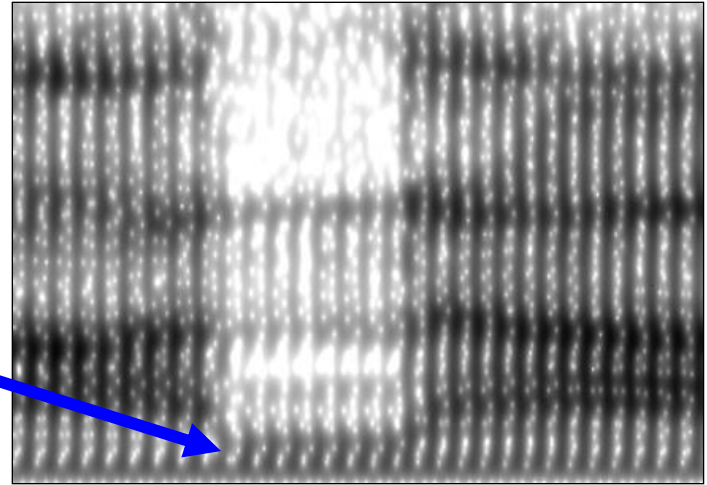
Salient Features of Liquids

- Liquids: /l, r/
 - Faster formant transitions than diphthongs
 - F3 distinguishes
 - Lower for /r/ than /l/



Salient Features of Nasals

- Formants less intense (quieter) than vowels
- Nasal formant

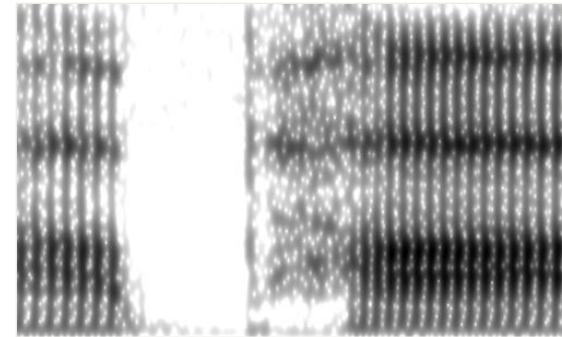


- Place of articulation cued by frequency and duration of formant transitions (esp. F2)
 - /m/: lowest frequency, shortest duration
 - /n/: intermediate
 - /ŋ/: highest frequency, longest duration

Salient Features: Obstruents

Salient Features of Stops

- Manner cued by:
 - Gap
 - Release burst
 - Rapid formant transitions
- Place of articulation cued by:
 - Frequency and duration of formant transitions
 - Esp. F2: Bilabial lowest, velar highest frequency
 - VOT
 - Bilabial shortest, velar longest
 - Frequency of release burst



gap ↑ VOT & aspiration
release burst

Salient Features of Stops

- Voicing cued by:
 - VOT
 - Short/negative = voiced, long/aspirated = voiceless
 - Aspiration
 - English voiceless (except after /s/)
 - **Cutback**: Beginning of F1 transition is cut off
 - > 30 ms cutback = perceived as voiceless
 - F0 of following vowel
 - Higher following a voiceless stop
 - Duration of preceding vowel
 - Longer vowel = perceived as voiced stop

Salient Features of Fricatives

- Frication noise

- Manner:

- > 130 ms = fricative
 - < 75 ms = perceived as stop

- Voicing:

- Longer, louder for voiceless

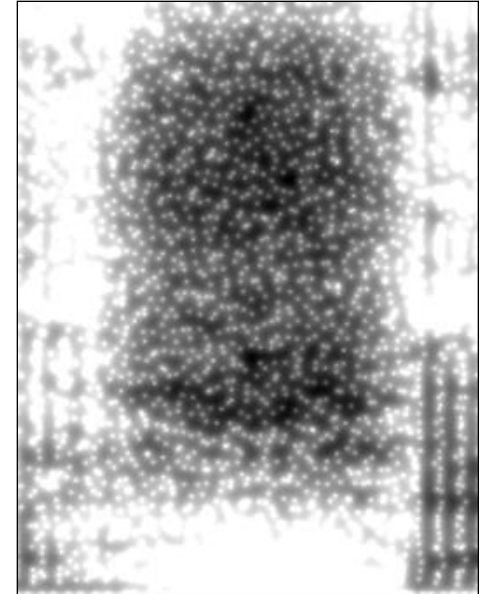
- Place:

- Intensity, spectral peak location

- Sibilants (alveolar/palatal) /s, z, ʃ, ʒ/: loud, high-frequency, peaked spectrum

- /s, z/: higher frequency concentration than /ʃ, ʒ/

- Labial/dental /f, v, θ, ð/: quiet, energy spread out



]

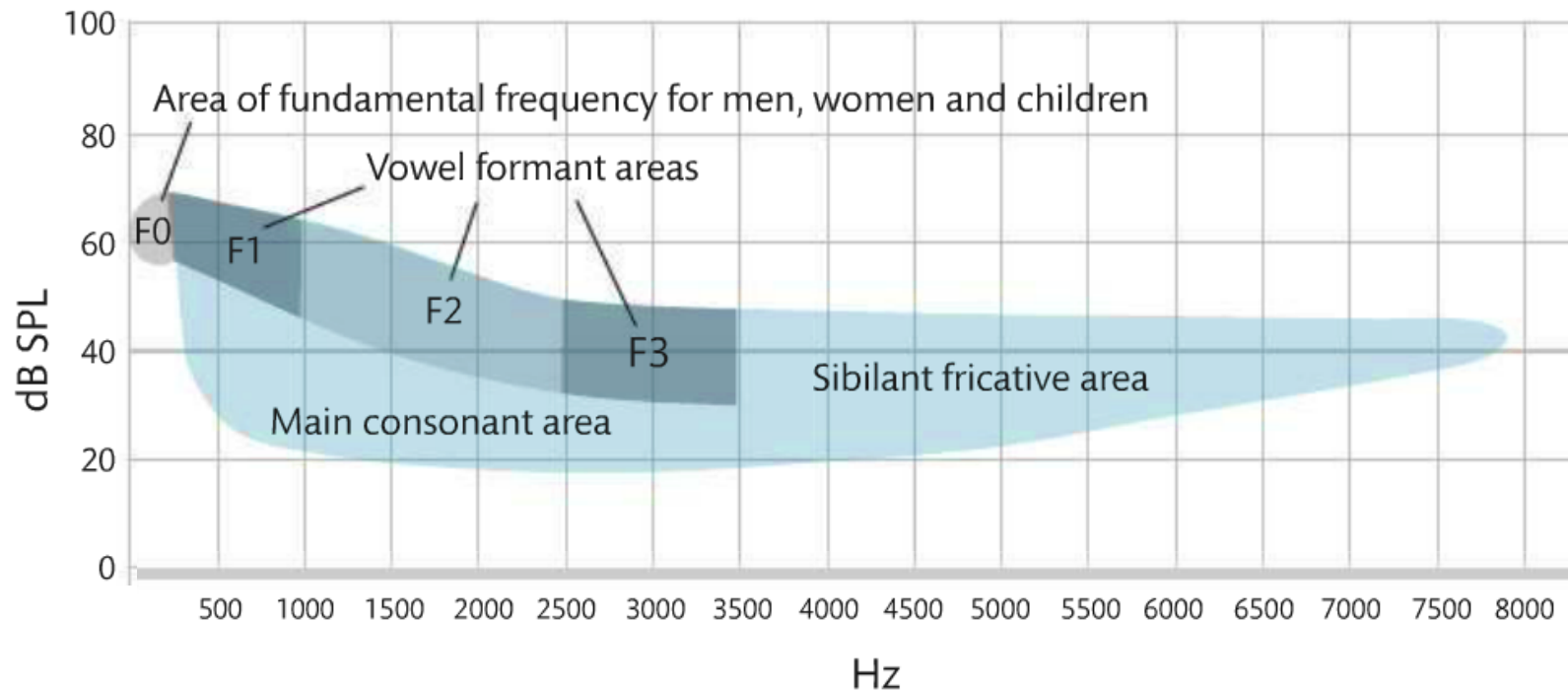
Salient Features of Affricates

- Manner (stop + fricative):
 - **Rise time:** duration until peak of amplitude envelope (~time to peak intensity)
 - ~33 ms = affricate, ~76 ms = fricative
 - Duration of frication noise shorter than for fricatives

Categorical Perception

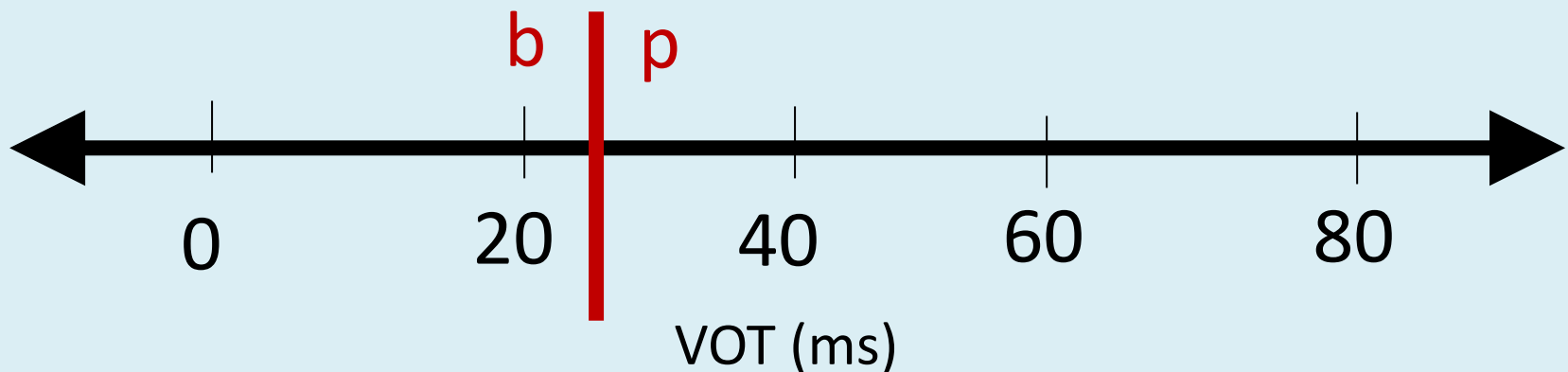
Perceptual Space

- Vowels contribute more information than consonants and are easier to perceive
 - Longer, louder
 - Also carry information about nearby consonants



Categorical Perception

- Many consonants are perceived **categorically**
 - Example: Voicing: English-speakers perceive all bilabials with $VOT < \sim 25$ ms as “the same” /b/ and all with $VOT > \sim 25$ ms as “the same” /p/
 - 25 ms = **crossover** (boundary)



Categorical Perception

– Example: Place of articulation

- Formant transitions (F1, F2, F3 direction, slope)
 - https://oup-arc.com/access/content/sensation-and-perception-5e-student-resources/sensation-and-perception-5e-activity-11-2?previousFilter=tag_activities



b



d



g

- We perceive these as each belonging to one category, /b, d, g/, not something in between (“b-d-ish”?)
- Even though the formant transitions change the same amount between each pair, we don’t notice differences within a category

Fun with Perception

Fun w/ Perception

- Vowel perception modified by context
 - Helps deal w/ lack of invariance
 - Lexical identification (bit, bet, but):
 - <https://engineering.purdue.edu/~malcolm/interval/1997-056/VowelQuality.html>
 - The vowels/formants of the preceding speech make you think “this must be “bit” or “bet” for this person”
 - Don’t believe it? Open “bit-bet1.wav” and “bit-bet2.wav” in Praat and play just the last word
 - Do quick measurements of F1 & F2 for each vowel

Fun w/ Perception

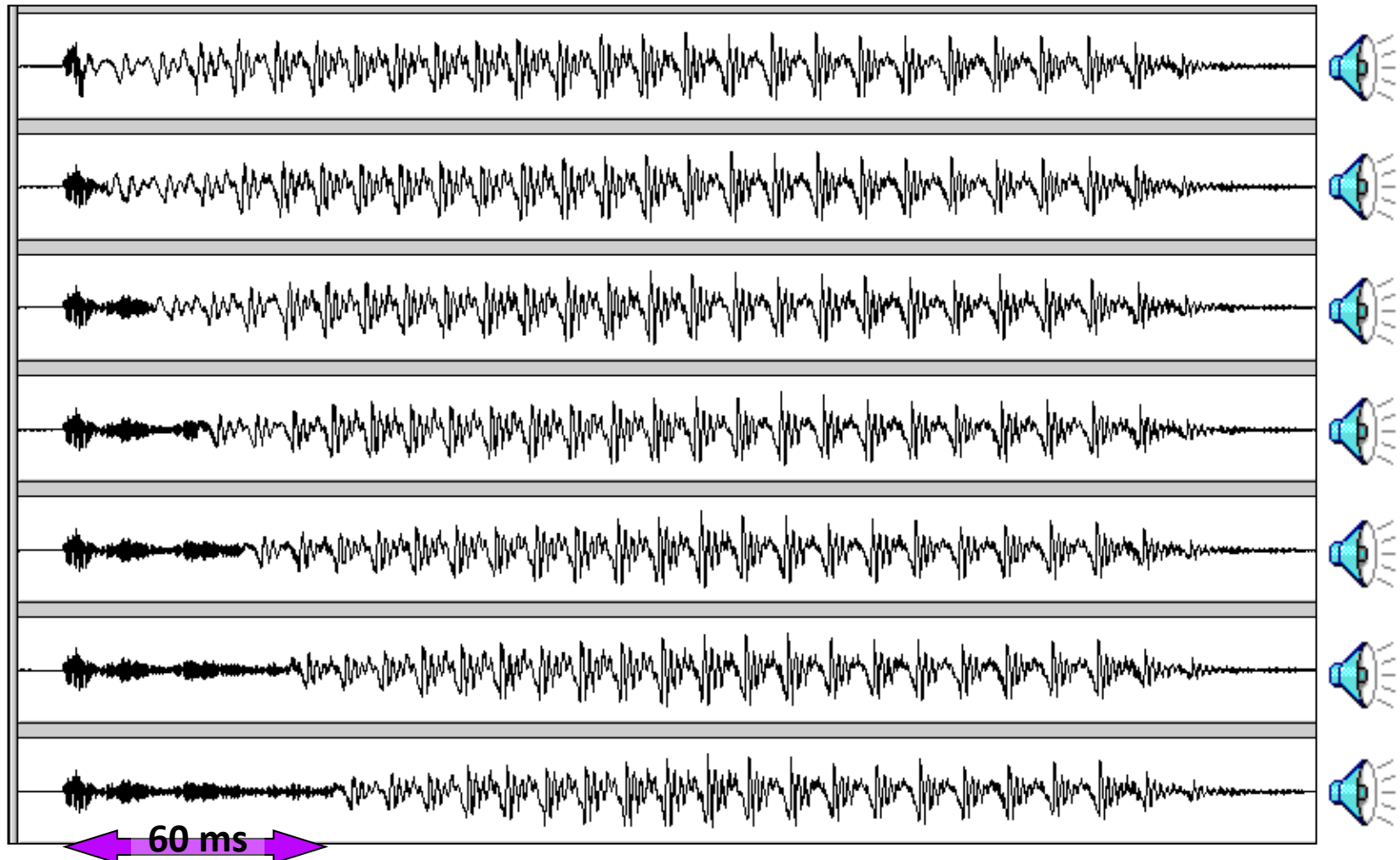
- Temporal ordering
 - Where is the superimposed sound?
 - http://www.vowelsandconsonants3e.com/chapter_10.html (10.7-10.8)
 - We perceive it in a prosodically logical place
 - Where we normally backchannel, interject, process...
- McGurk Effect: Audio-visual perception
 - https://oup-arc.com/access/content/sensation-and-perception-5e-student-resources/sensation-and-perception-5e-activity-11-3?previousFilter=tag_activities
 - We use visual information as another cue

Identification Experiments

Perception Experiments

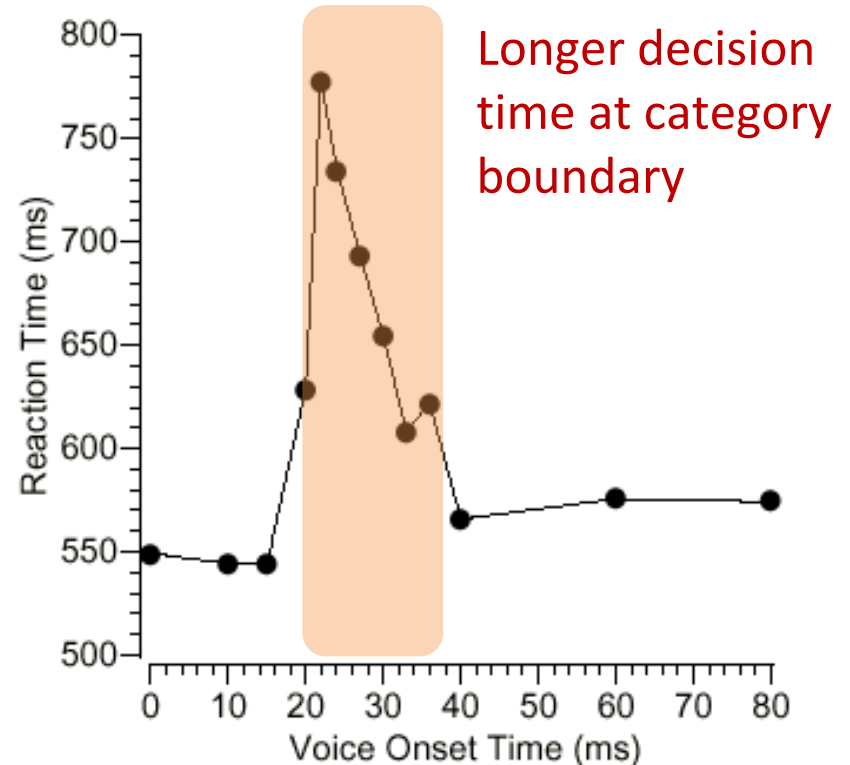
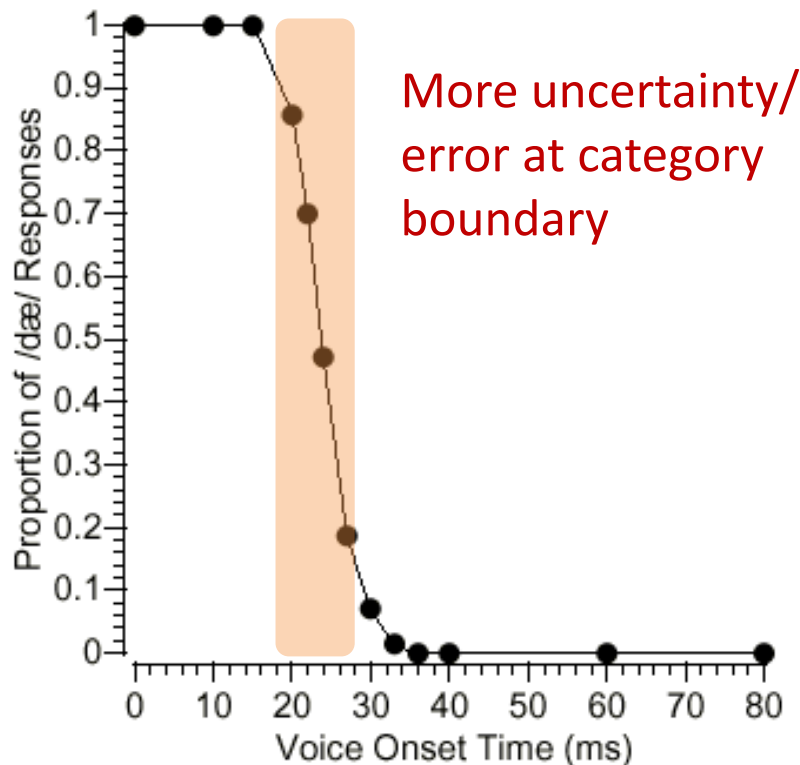
- Two main types of (classic) speech perception experiment tasks:
 - **Identification:** categorize/label phones/words
 - **Discrimination:** tell the difference between phones/words (“same or different?”)
 - Hard between sounds in the same category
 - “Hard” = more errors, choices at chance, long reaction times
- Example: VOT to distinguish voicing

VOT



Identification Task

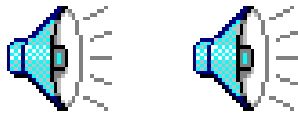
- Plot proportion of “d” responses
- Plot how long it took them to decide



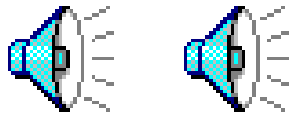
Discrimination Experiments

Discrimination Task

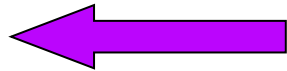
- “Are these two sounds the same or different?”



Same / Different
0ms 60ms



Same / Different
0ms 10ms



Why is this pair difficult?

- Acoustically similar?
- Same Category?



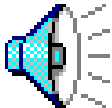
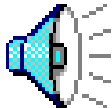
Same / Different
40ms 40ms

Discrimination Task

- If it were acoustic similarity, these would all be equally hard to tell apart:

D 0ms   20ms D

D 20ms   40ms T

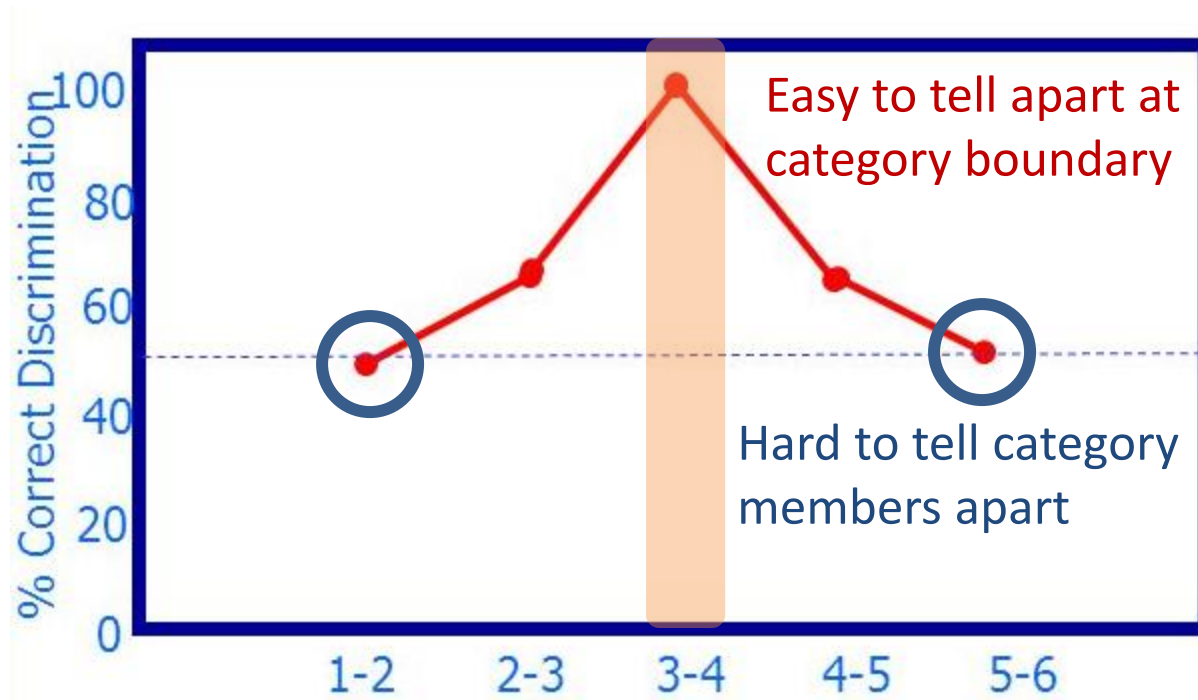
T 40ms   60ms T

Across-Category Discrimination is Easy

Within-Category Discrimination is Hard

Discrimination Task

- Plot proportion of correct discrimination
 - Said “different” when they were different sounds, “same” when the sounds were identical



Language-Specific Categories

Production to Facilitate Perception

- Speakers tend to avoid producing VOTs near the category boundary

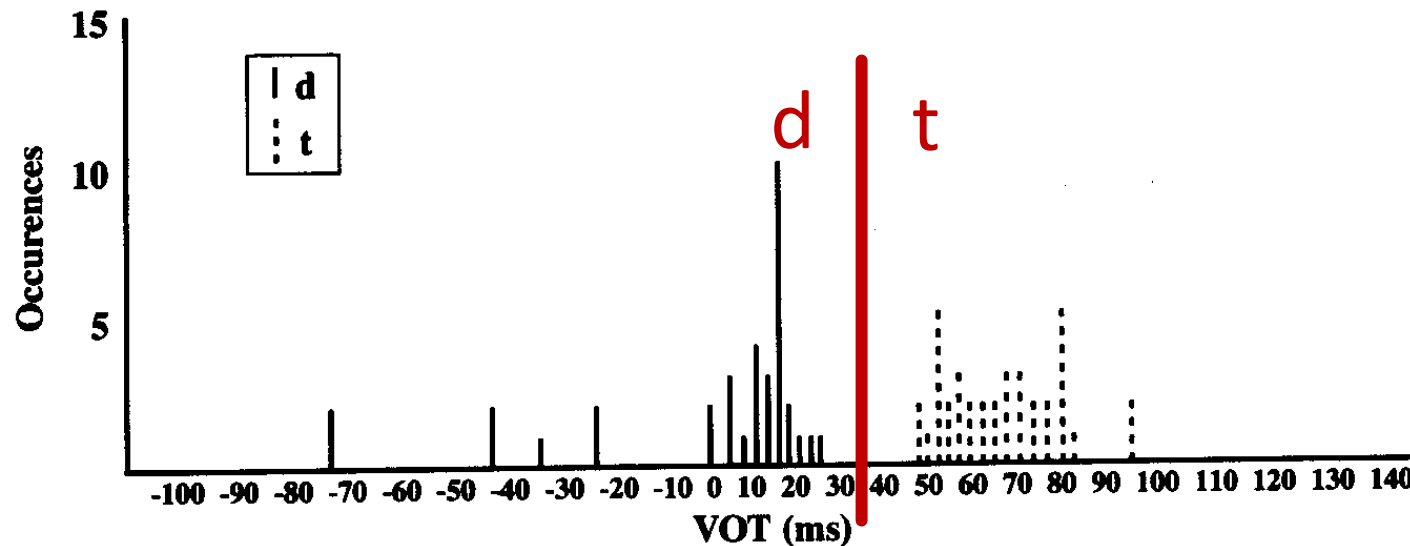
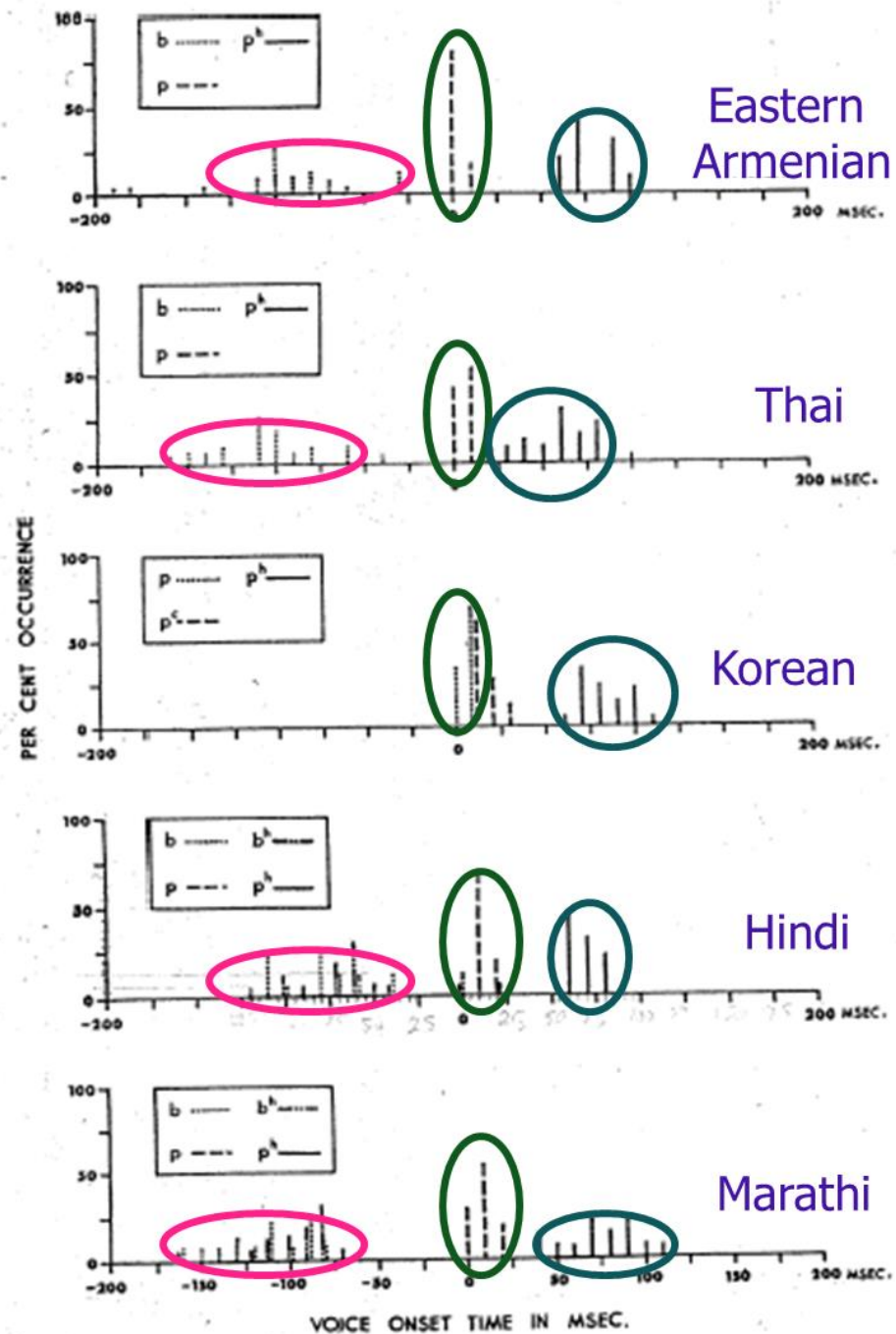
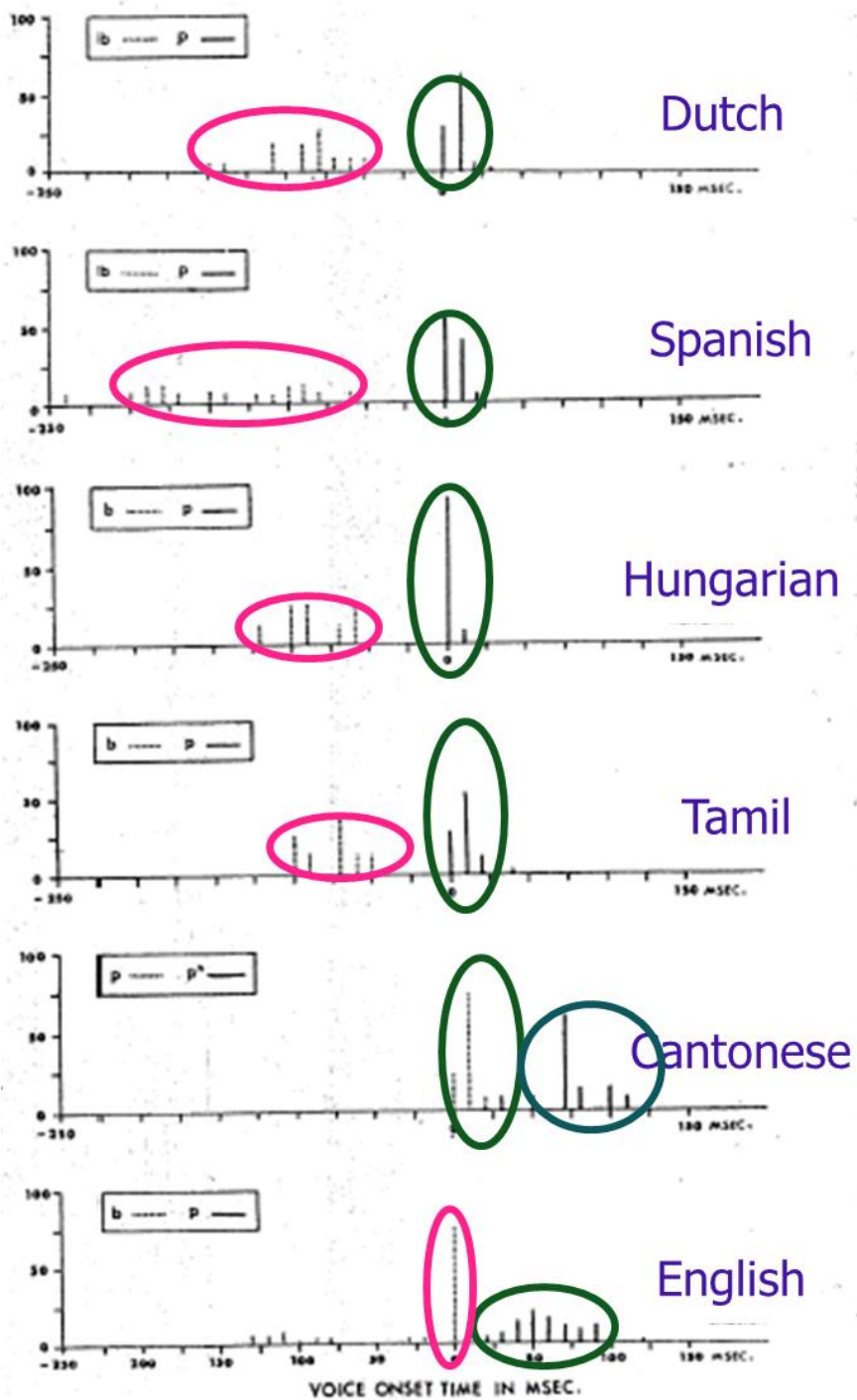
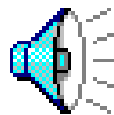
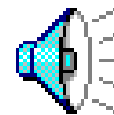
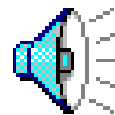
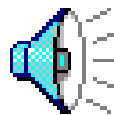
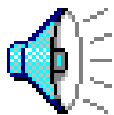
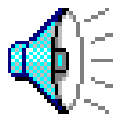
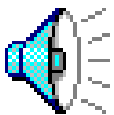
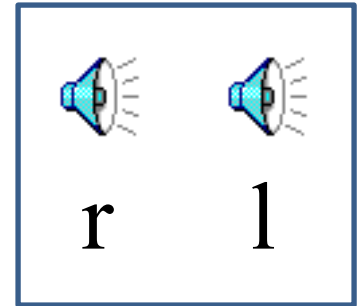


Figure 5–3. VOT productions of a single normal adult speaker of American English for words beginning with /d/ and /t/. (Figure adapted with permission from Blumstein, Cooper, Goodglass, Statlender, & Gottlieb, [1980]. Production Deficits in Aphasia: A Voice Onset-Time Analysis. *Brain and Language*, 9, 153–170. Copyright 1980 by Academic Press.)



Language-Specific Categories

- Phoneme category boundaries differ by language
 - English has two liquid phonemes, /r, l/
 - Japanese has one category for these
 - And can pronounce anything in between



r

l

Language-Specific Categories

- Identification task
 - English-speakers discriminate categorically
 - Japanese-speakers perform at chance
 - /r, l/ are in the same category

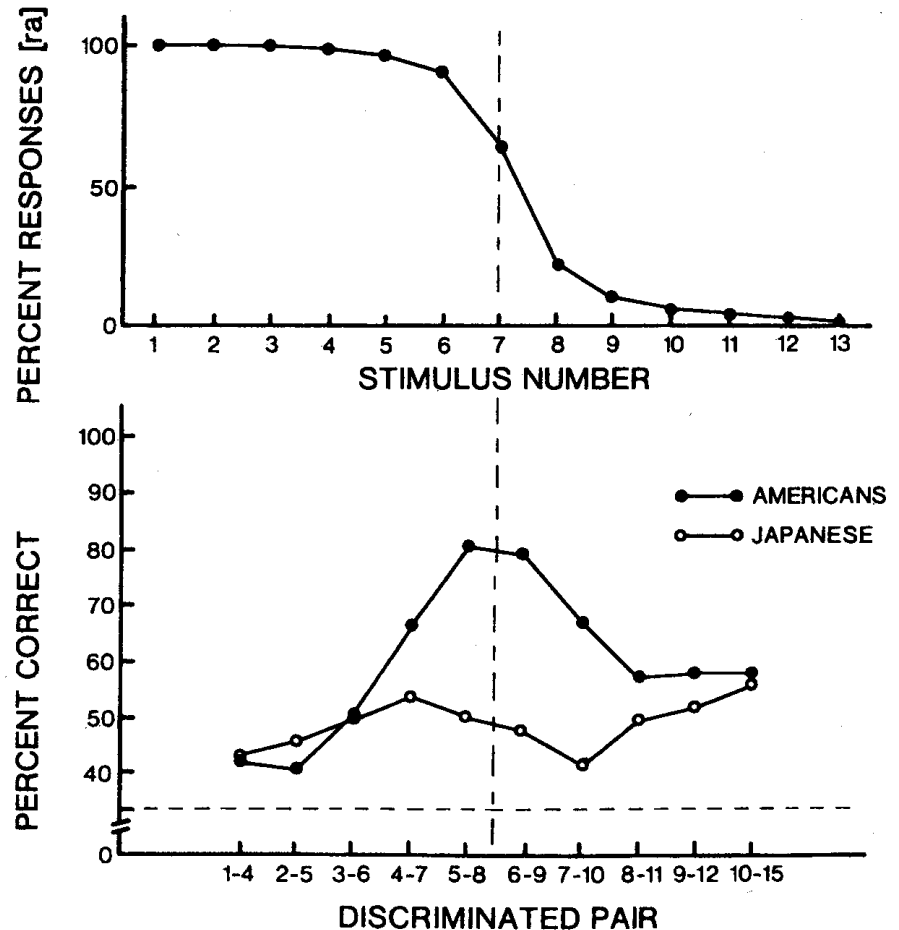


Figure 12.2. Test of the categorical perception of /ra/ and /la/ by American and Japanese adults. American listeners show the characteristic peak in discrimination at the phonetic boundary; Japanese listeners do not. (From Miyawaki et al., 1975.)