

reading: Wright (2004) chapter; excerpt from Borden, et al. (2003) chapter

### Speech Perception

1. The listener's task is to extract meaning from the acoustic signal.

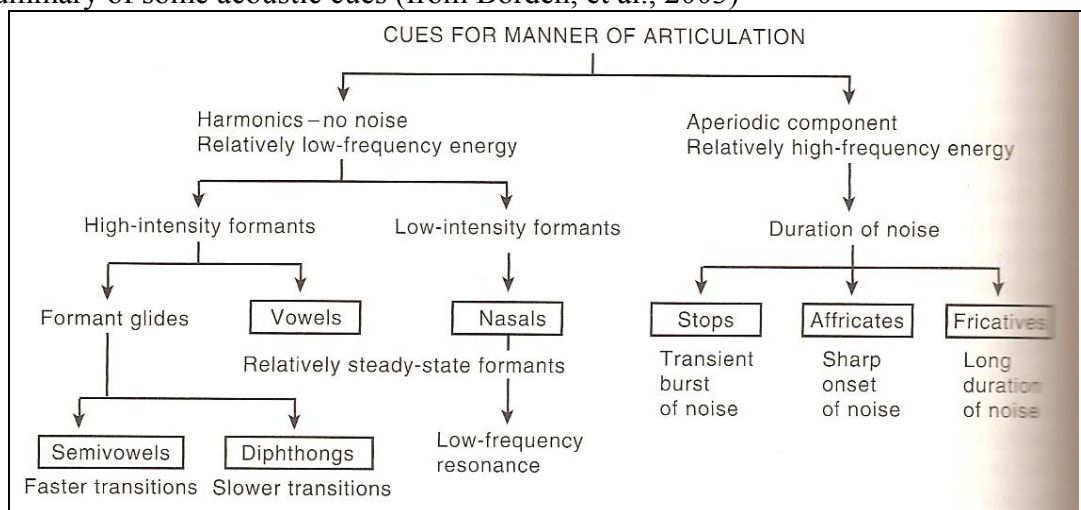
electrochemical transform'n of acoustic signal (the product of *audition*) → representational units (e.g., features, segments, syllables, words) → meanings (usually through the lexicon)

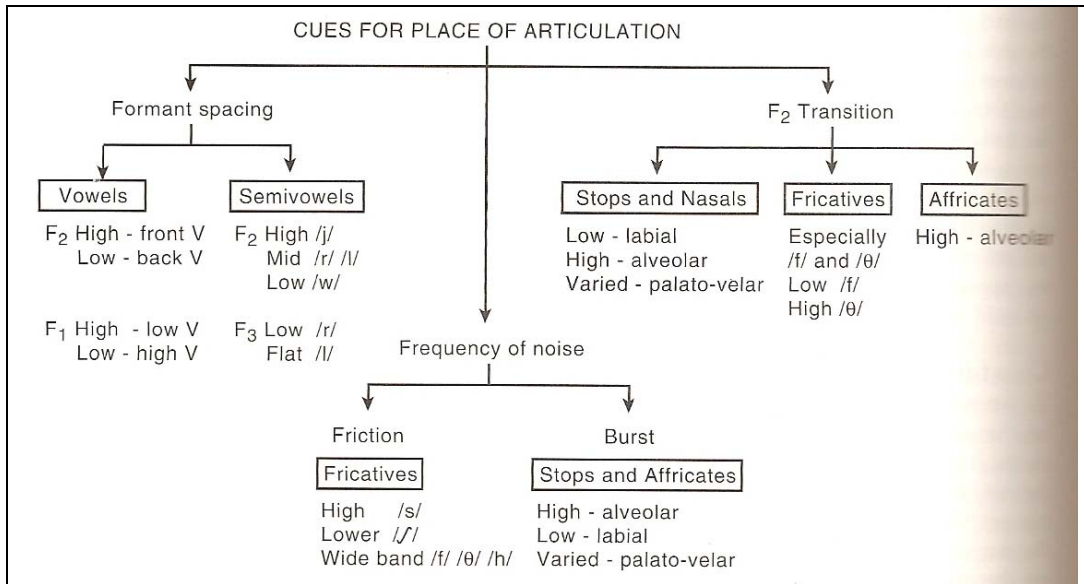
- Most theories posit representational units that are phonological – e.g., segmented and featural or gestural  
Words are stored as strings of feature bundles or gestures; speech perception involves recovering an abstracted segmental/featural/gestural representation from the acoustic signal, which provides the basis for lexical access.
- But representations could be directly auditory – i.e., direct, with no intermediate processing  
The lexicon contains auditory representations. Extra processing steps are unnecessary, and in fact, would lead to the loss of too much information.

2. So speech perception research is largely concerned with figuring out how listeners recover an abstracted (segmental/featural) representation from the acoustic signal.  
→ cues used to identify segments

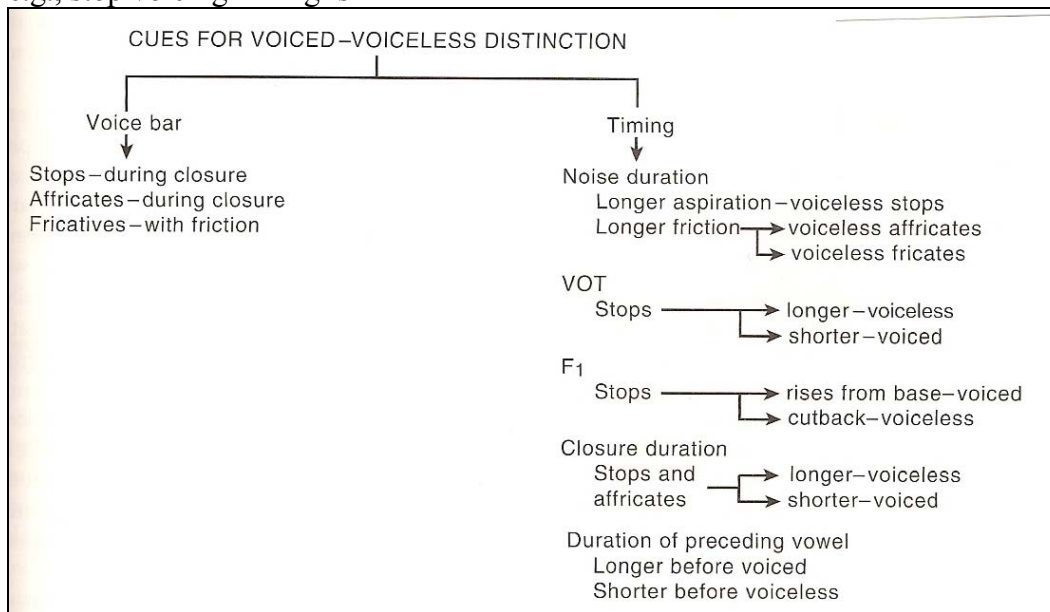
Even if we reject the assumption that there are segments to be found, it is still clear that languages use a limited number of ways to distinguish words (i.e., contrasts) – so in either case, it is important to clarify how these contrasts are distinguished in the acoustic signal.

3. Summary of some acoustic cues (from Borden, et al., 2003)





4. There are multiple cues to every contrast – the speech signal is highly redundant.  
e.g., stop voicing in English



- also, amplitude of aspiration, amplitude of release burst (Repp, 1979), F0 adjacent to closure (Haggard, et al, 1970), amplitude of F1 at release (Lisker, 1986)

5. Do listeners use all of these cues? Are they all equally important?

We can manipulate various parameters of these cues – by synthesizing speech or by editing natural speech – and test the effect on the perception of listeners.

6. Why are there so many cues?

- Multiple/redundant cues make the signal robust (e.g., Wright 1996, 2001, 2004)

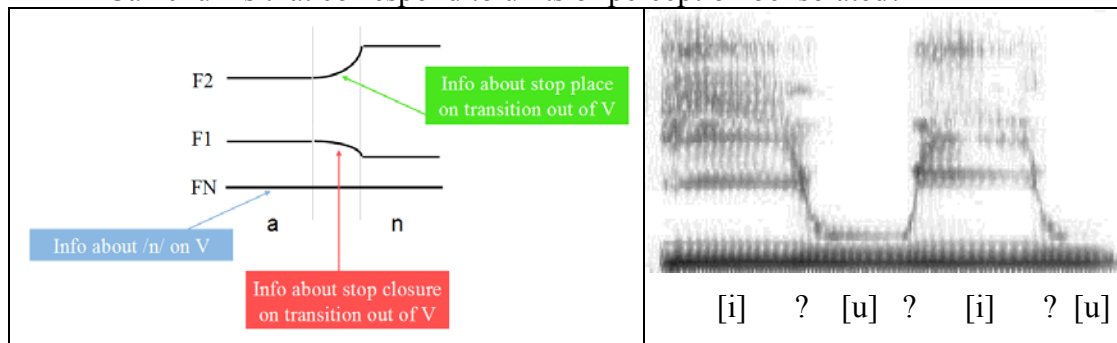
robust to environmental noise, momentary distraction by the listener, auditory weaknesses, etc.

7. Since cues to a contrast are temporally distributed, cues to more than one contrast may be present in the signal simultaneously (i.e., no strict segmentation).

- the segmentation problem

Where are the boundaries in the continuous signal?

Can chunks that correspond to units of perception be isolated?



(from Beddor slides)

- 2 possible types of segmentation problem:
  - overlapping (=coarticulatory) information could be “noise” between segments that might interfere with perception and processing
  - overlapping (=coarticulatory) information could contain complex cues that must be extracted by the perceiver

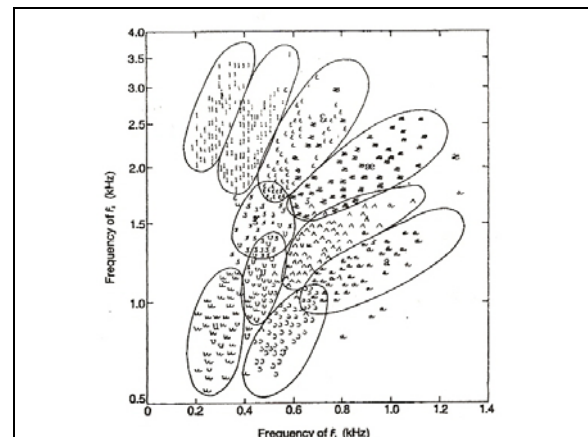
8. But the “segmentation problem” doesn’t seem to be a *problem* for listeners.

- Listeners make use of cues in overlap – systematic variation – rather than ignore it.
  - e.g., Nasality on a vowel can serve as a cue to an upcoming nasal consonant. (e.g., Beddor & Krakow, 1999)

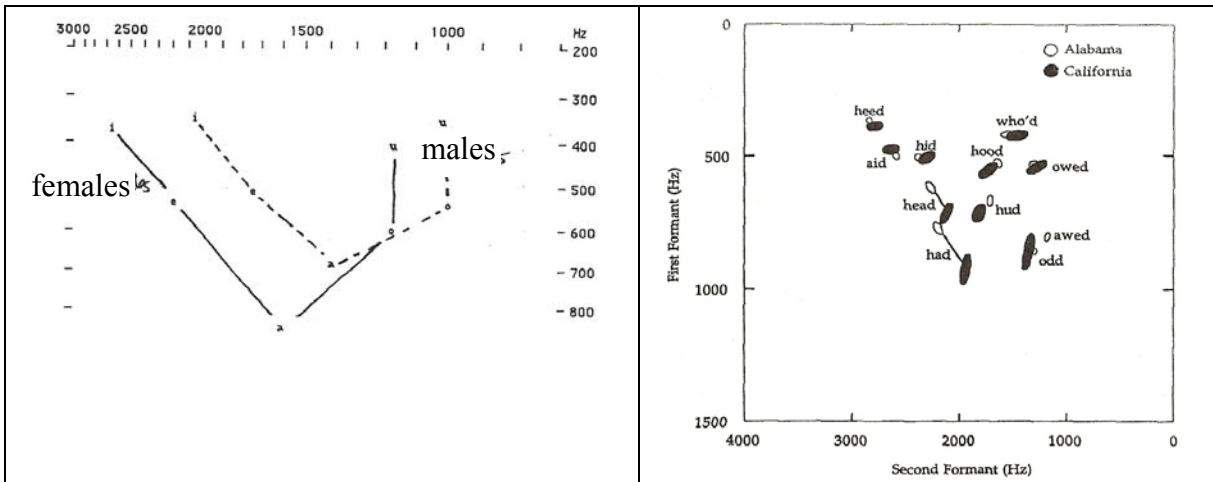
9. The acoustic realization of segments, and even of words, is highly variable. So the listener has to identify diverse acoustic signals as representing the same thing.

Sources of variability:

- linguistic context – segmental context (coarticulation), prosodic position (incl. stress, accent)
- speaker – vocal tract size, articulatory habits, language/dialect, etc.
  - physical & emotional state, etc.
- environment – ambient noise, room acoustics, etc.
- communicative context – register, speech rate, etc.



(from Lieberman & Blumstein, 1988)



In some cases, there don't appear to be acoustic properties that reliably correspond to the units/segments of linguistic analysis and/or perception.

- the problem of the lack of invariance
  - A given acoustic signal doesn't always evoke the same percept.
    - e.g., bursts in stops
      - A 1440 Hz (synthetic) burst is heard as [p] when followed by [i], but as [k] when followed by [a].
    - e.g., vowels
      - A given set of vowel formants can elicit different vowel percepts depending on the formant frequencies of the surrounding vowels.
      - systematic variability – it turns out listeners can use this information
  - Multiple acoustic signals can evoke the same percept.
    - e.g., place of articulation in stops
      - A burst or F2 transition reliably cues place information.
    - e.g., voicing in English
      - redundancy leads to robustness

10. Speaker normalization is the problem of perceiving speech across different speakers with different vocal tracts (so e.g., different formants, F0), with different dialects, etc.

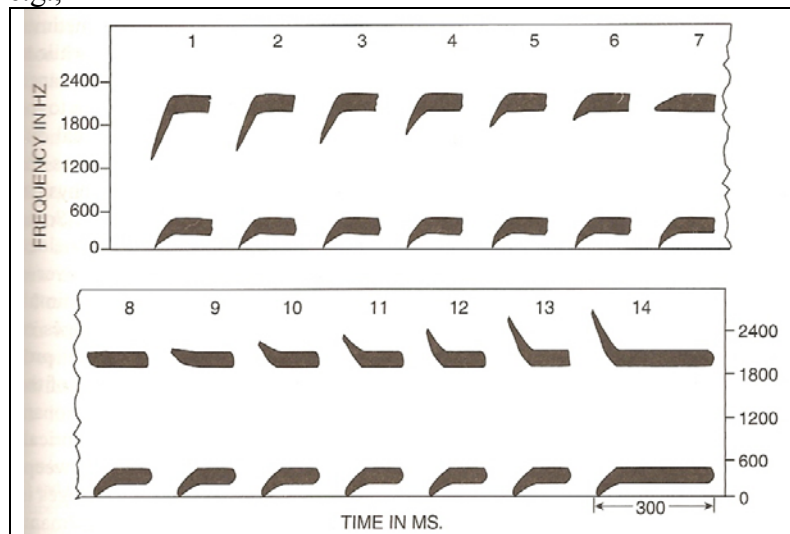
- This is problematic for machine speech recognition, but people handle it quite well.
- e.g., for vowel perception – Speakers have different sized vocal tracts (compare for instance men, women, and children), causing the acoustic resonant frequencies for a given vowel to vary widely across speakers. But listeners hear different /i/s from different speakers as /i/s, etc.
- How do listeners accommodate?
  - scaling – speakers' formants are scaled relative to one another; e.g., a child's formants might consistently be 4/3 an adult's
  - ratios – relative formant patterns are constant; e.g.,  $F2/F1 = 10$  for /i/
  - types of theories:

- *intrinsic* – normalization uses only information within the syllable (e.g.,  $f_0$ , higher formant frequencies, etc.) vs. *extrinsic* – syllable external properties are used (e.g., vowel formant range, average  $f_0$ , etc.)
  - *direct* – normalization information is used directly in constructing perceptual representations of segments vs. *indirect* – normalization information is used to create a frame of reference for the interpretation of segments
- Although listeners can normalize almost instantly, so it must be at least partly intrinsic, normalization is subject to extrinsic influences as well. Listeners do use information in the whole context of a speaker's utterances to influence the identification of particular parts of the signal.
- e.g., Ladefoged & Broadbent, 1957 – identification of an ambiguous 'bit/bet' stimulus is influenced by the formants of the preceding vowels in the sentence

## 11. Categorical perception

- A change in some variable along a continuum is perceived not as a gradual change from one type to another, but as sets of instances of discrete categories.

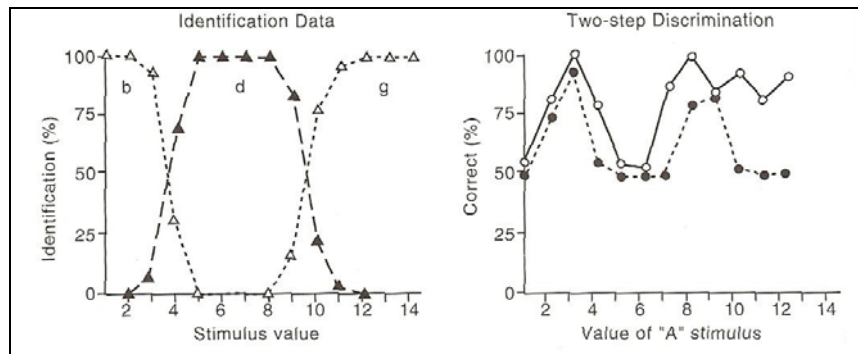
e.g.,



(from Borden, et al., 2003)

2 pieces of evidence:

- identification: when subjects are asked to label each of the stimuli, subjects divide them into groups with steep boundaries between categories  
e.g., into /b/, /d/, and /g/ groups, where the same intermediate stimuli are consistently identified in the same way
- discrimination: when subjects are presented with adjacent pairs of stimuli and asked to state whether they are the same or different, subjects are unable to perceive differences within a category, but discrimination peaks across category boundaries  
e.g., differences among stimuli categorized as /b/ in the identification task are imperceptible; the difference between the pair that cross the boundary is consistently perceived



(from Borden, et al., 2003)

- Non-speech sounds are generally not perceived categorically.
- Not all speech sounds are either: consonants generally are, but vowels are perceived essentially continuously (although listeners can define boundaries between categories).

→ Categorical perception shows that the perceptual system is organized in a learned way.

#### 12. Two models of categorization:

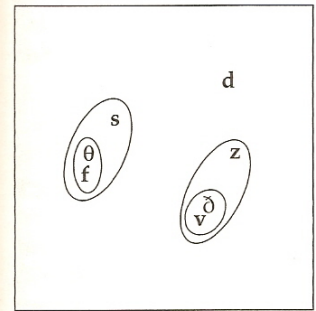
- Prototype models: Listeners construct prototypes for categories (segments or words), and categorization proceeds by matching incoming instances to prototypes in memory.
  - Prototypes are constructed from specific instances, but only the prototype, which abstracts away from many specifics of the instances, is stored.
- Exemplar models: Listeners store multiple categorized instances in memory. Categorization proceeds by matching incoming instances to the set of exemplars for each category.
  - We store instances (or at least a large number of types, along with frequency information) in considerable detail.
    - Often, exemplar models imply that an auditory representation can be used to directly probe the lexicon, but exemplars might also be categorizations of segments or syllables, rather than words.

#### 13. Not only is there an organization of instances or types into categories, but there is also an organization of categories relative to one another in a “perceptual space”.

Some sounds are more similar to one another than other sounds. I.e., some sounds are more confusable with one another.

- A confusion matrix tabulates the frequency with which one sound is confused with another (when heard in noise).

	"f"	"v"	"th"	"dh"	"s"	"z"	"d"
[f]	1.0						
[v]	.008	1.0					
[θ]	.434	.010	1.0				
[ð]	.003	.345	.000	1.0			
[s]	.025	.000	.170	.000	1.0		
[z]	.000	.026	.000	.169	.000	1.0	
[d]	.000	.000	.000	.012	.000	.081	1.0



(from Johnson, 2003)

We can calculate from these confusions how similar two sounds are.

$$S_{ij} = \frac{p_{ij} + p_{ji}}{p_{ii} + p_{jj}} \quad S_{ij}: \text{similarity of } i \text{ and } j, p_{ij}: \text{probability of } i \text{ given } j$$

Perceptual distance is the negative natural log of the similarity:  $d_{ij} = -\ln(S_{ij})$

Shepard's Law states that the relationship between perceptual distance and similarity is exponential.

So given perceptual distances between sounds, we can map sounds in perceptual space (as above) using multidimensional scaling (MDS). The map reflects graphically observations about perceptual similarity and likely confusions.

- A person's perceptual map (i.e., their categories and the relations among them) is language-specific.