reading: Borden, et al. – Ch. 6 (today); Keating (1990): "The window model of coarticulation" (Tues)

Theories of Speech Perception

1. Theories of speech perception must be able to account for certain facts about the acoustic speech signal, e.g.:
   - There is inter-speaker and intra-speaker variability among signals that convey information about equivalent phonetic events.
   - The acoustic speech signal is continuous even though it is perceived as and represents a series of discrete units.
   - Speech signals contain cues that are transmitted very quickly (20 to 25 sounds per second) and simultaneously.

   They must also be able to account for various perceptual phenomena, e.g.:
   - categorical perception
   - phonemic restoration
   - episodic memory

   plus, various word recognition effects (e.g., frequency effects, priming, etc.)

2. Theories of speech perception differ with respect to their views of what is perceived and how.

• Auditory – listeners identify acoustic patterns or features by matching them to stored acoustic representations	• Motor – listeners extract information about articulations from the acoustic signal
• Bottom-up – perception is built from information in the physical signal	• Top-down – listeners use higher level sources of information to supplement the acoustic signal
• Active – cognitive/intellectual work is involved in perception	• Passive – perception relies on passive responses (e.g., thresholds)

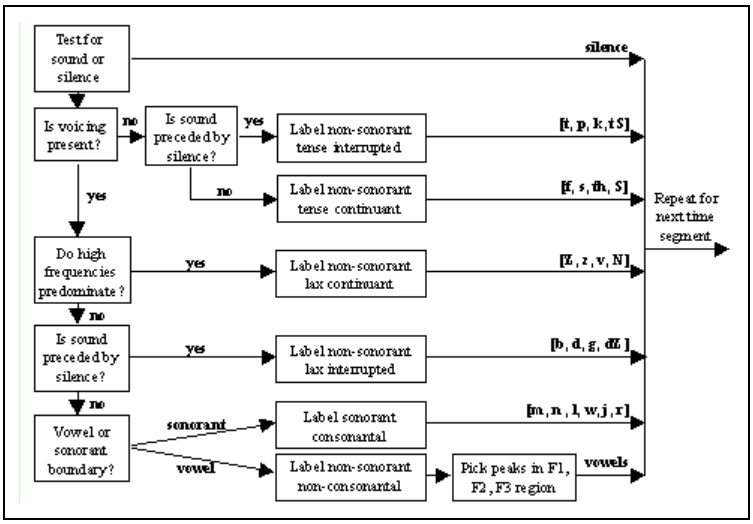
<u>Auditory theories</u>

3. Auditory Model (Fant, 1960; also Stevens & Blumstein, 1978)
   - The assumption of this model is that invariance can always be found in the speech signal by means of extraction into distinctive features.

   Listeners, through experience with language, are sensitive to the distinctive patterns of the speech wave. → We have feature detectors (that may be more or less specialized).

- template matching: When we listen to speech, we match the incoming auditory patterns to stored templates (phonemes or syllables) to identify the sounds.

  Templates may be more abstract than the patterns or features found in spectrograms (especially to represent place of articulation).



  - After being decoded, the perceptual units have to be recombined to access lexical items.

- Auditory Enhancement Theory (Diehl & Kluender, 1989)

  Various acoustic properties may 'work together' to increase the auditory salience of phonological contrasts. Contrasts between sounds are robust because phonological systems have evolved to enhance the perceptual distinctiveness of the contrasts.

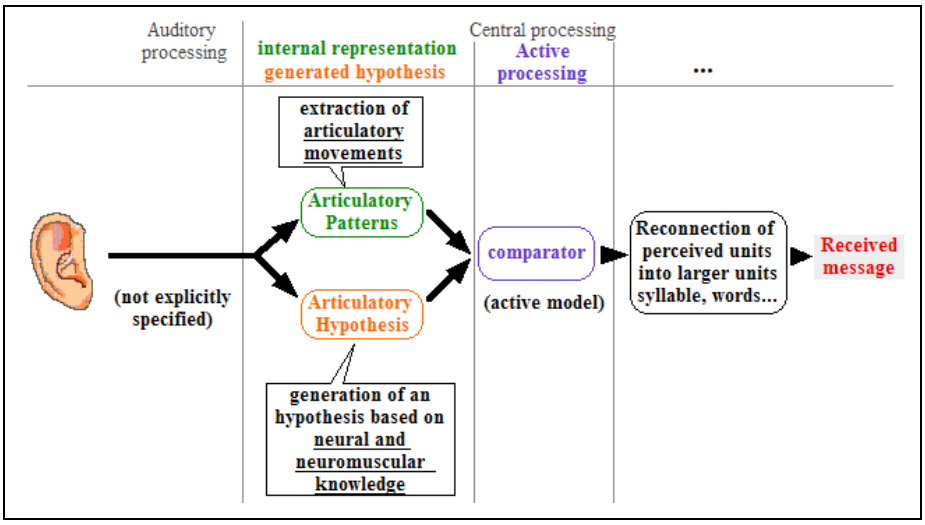
<u>Motor theories</u>

4. Motor Theory (Liberman, et al., 1967; Liberman & Mattingly, 1985)
   - Given the lack of acoustic invariance, we can look for invariance in the articulatory domain (i.e., maybe the representational units are defined in articulatory terms).

   Motor theory postulates that speech is perceived by reference to how it is produced; that is, when perceiving speech, listeners access their own knowledge of how phonemes are articulated. Articulatory gestures such as rounding or pressing the lips together are units of perception that directly provide the listener with phonetic information.

   Biological specialization for phonetic gestures prevents listeners from hearing the signal as ordinary sound, but enables them to use the systematic, special relation between signal and sound to perceive the gestures.

   - Originally, the <u>motor commands</u> that control articulation were considered to be the invariant phonetic features.
   - The revised theory says that it is <u>intended gestures</u> that are the invariant object of perception.



(from Fougeron web tutorial)

- We perceive sounds discretely (categorically) because sounds are produced with discrete articulators/gestures.

- The McGurk effect suggests that we represent at least some features as articulatory.

5. Analysis by Synthesis (Stevens & Halle, 1960)
   - In this model, speech perception is based on auditory matching mediated through speech production.

   When a listener hears a speech signal, he or she analyzes it by mentally modeling the articulation (in other words, the listener tries to synthesize the speech his or herself). If the 'auditory' result of the mental synthesis matches the incoming acoustic signal, the hypothesized perception is interpreted as correct.

6. Direct Realist Theory (Fowler, 1986)
   - Direct realism postulates that speech perception is direct (i.e., happens through the perception of articulatory gestures), but it is not special. All perception involves direct recovery of the distal source of the event being perceived (Gibson).
     In vision, you perceive *objects* (e.g., trees, cars, etc.). Likewise with smell you perceive e.g., cookies, roses, etc. Why not in the auditory perception of speech?
     - So, listeners perceive tongues and lips.

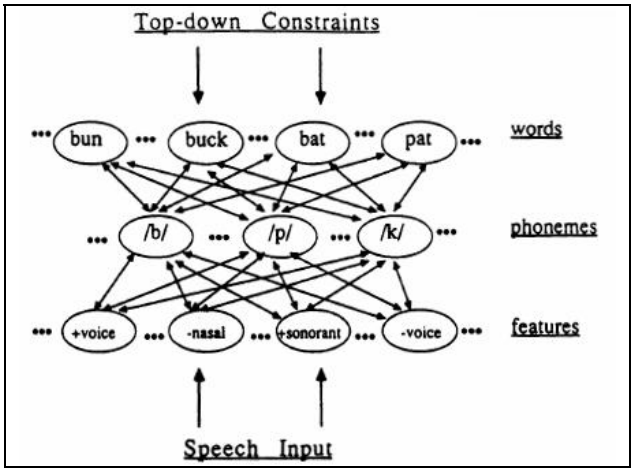
   The articulatory gestures that are the objects of speech perception are not *intended* gestures (as in Motor Theory). Rather, they are the *actual* gestures.

<u>Word recognition</u>

7. TRACE (McClelland & Elman, 1986)
   - TRACE is a connectionist network model of speech perception / lexical perception.

   Different levels of speech units (e.g., features, phonemes, words) are represented on different levels of the network.

- Influences across levels share excitatory activation; i.e., activated features lead to the activation of the related phoneme; activated phonemes activate units on the word level.
- Influences within a level (those that are inconsistent with eachother) are inhibitory; i.e., the activation of one phoneme level unit inhibits the activation of other competing phonemes.



8. Cohort Theory (Marslen-Wilson, 1980)
   - Cohort theory models spoken word recognition.

   Based on the beginning of an input word, all words in memory with the same word-initial acoustic information, the cohort, are activated. As the signal unfolds in time, members of the cohort which are no longer consistent with the input drop out of the cohort.

   input: *cap-* (e.g., of *captivate*) *capt-* (of *captivate*)

   cap, captain, capsize, captive, caption, capital, captivate, etc. ⇨ ~~*cap*~~*, captain, ~~capsize~~, captive, ~~caption~~, ~~capital~~, captivate,* etc.

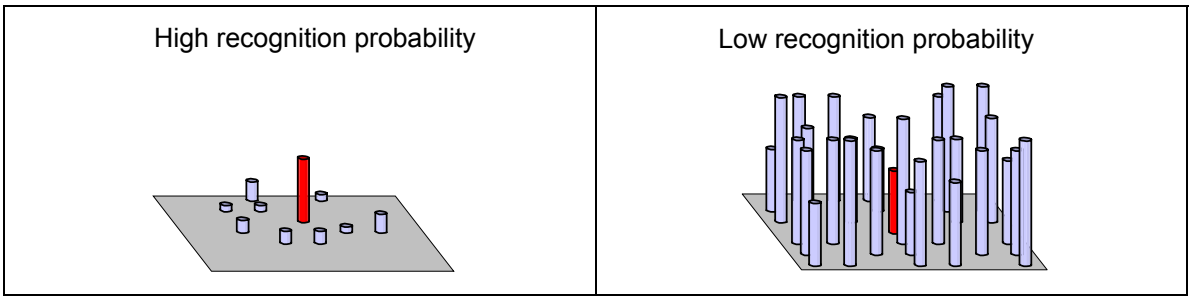
   Cohort elimination continues until a single word remains (i.e., is identified).

   The point (left to right) at which a word diverges from all other members of the cohort is called the uniqueness point.

9. Neighborhood Activation Model (Luce, 1986; Luce & Pisoni, 1998)
   - The Neighborhood Activation Model (NAM) models spoken word recognition as the identification of a target from among a set of activated candidates (competitors).

   All words phonologically similar to a given word are in the word's neighborhood. Recognition of a word is based on the probability that the stimulus word was presented compared to the probability that other words in the neighborhood were in fact presented.

   Probability is also influenced by lexical frequency.



10. Exemplar Models – Non-analytic approaches (e.g., Johnson, 1997; Goldinger, 1997; Pierrehumbert, 2002)

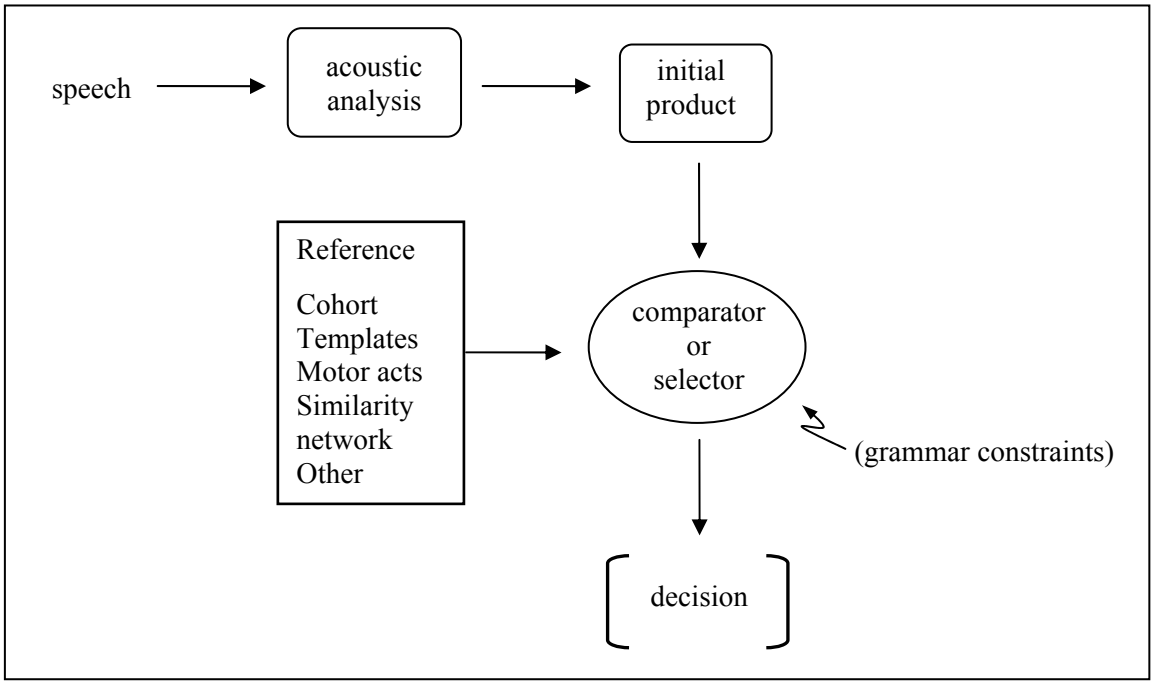
- In most models of speech perception, the objects of perception (or the representational units) are highly abstract. In fact, information about specific instances of a particular word are abstracted away from and 'discarded' in the process of speech perception.
  So information about a particular speaker or speech style or environmental context can play no role in the representation of words in memory.
- Exemplar models postulate that information about particular instances (episodic information) is stored.
  - Mental representations do not have to be highly abstract.
  - They do not necessarily lack redundancy.

Categorization of an input is accomplished by comparison with all remembered instances of each category (rather than by comparison with an abstract, prototypical rep'n).

- Often, exemplars are modeled as categorizations of words, but they might also be categorizations of segments or syllables or whatever.

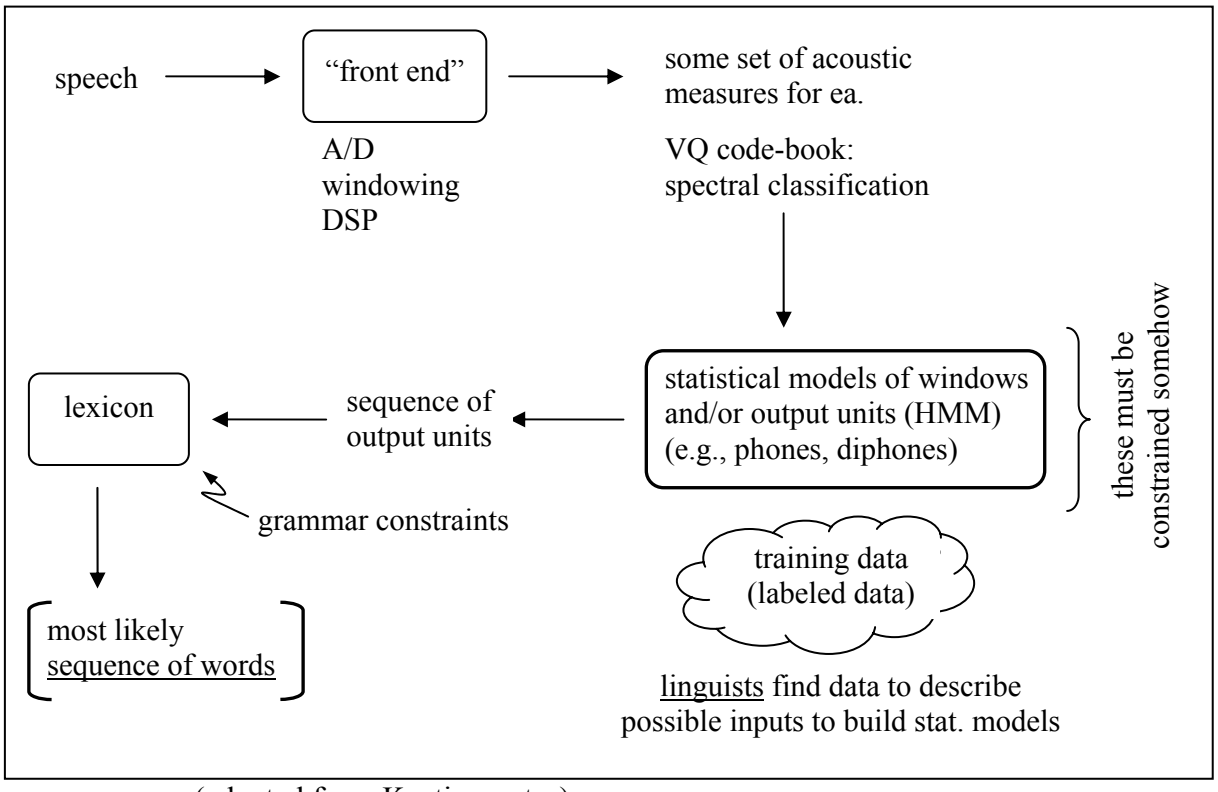
Stored exemplars are activated to a greater or lesser extent according to their degree of similarity to an incoming stimulus; activation levels determine categorization.

11. Generalized model of speech perception



(adapted from Kent, 1997)

12. Machine speech recognition



(adapted from Keating notes)