Language and Perception

Theories of Speech Perception

Theories of Speech Perception

- Theories specify the “objects” of perception and the mapping from sound to object.
- Theories must provide for robustness and graceful degradation. A key element to graceful degradation is the principle of least commitment.
- Theories must be sufficiently specific to be falsified (perhaps by being implemented as a model of perception).

Speech Oddities

- Perceptual constancy, but lack of invariants
- Categorical perception
- Segmentation
- Audio-visual integration
- Duplex perception
- Rate of speech sounds

Where is the Invariant?

Three types of theories:

1. In the signal, but we haven't been looking in the right place (e.g., Stevens & Blumstein)
2. In the production of the signal: Motor Theory (Liberman, Mattingly, et al.)
3. In the mind of the perceiver: TRACE (McClelland & Elman)

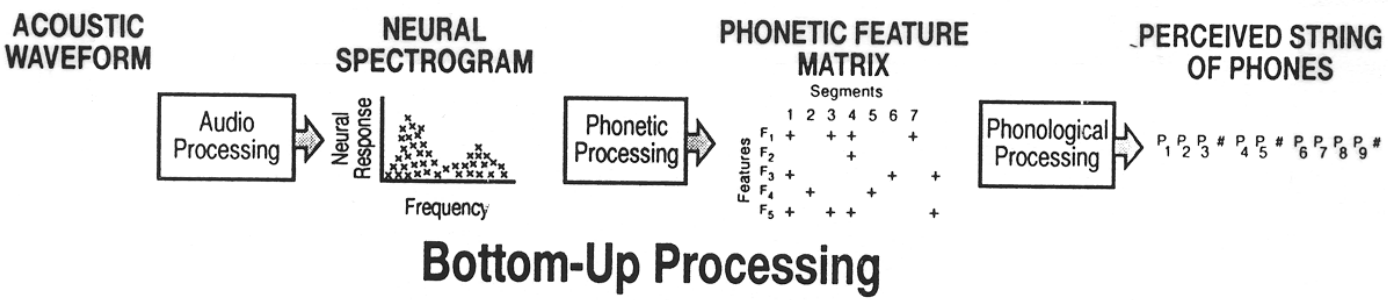
Categories of Theories

- Active vs. Passive
- Bottom-up vs. Top-Down
- Autonomous vs. Interactive

Active vs. Passive Theories

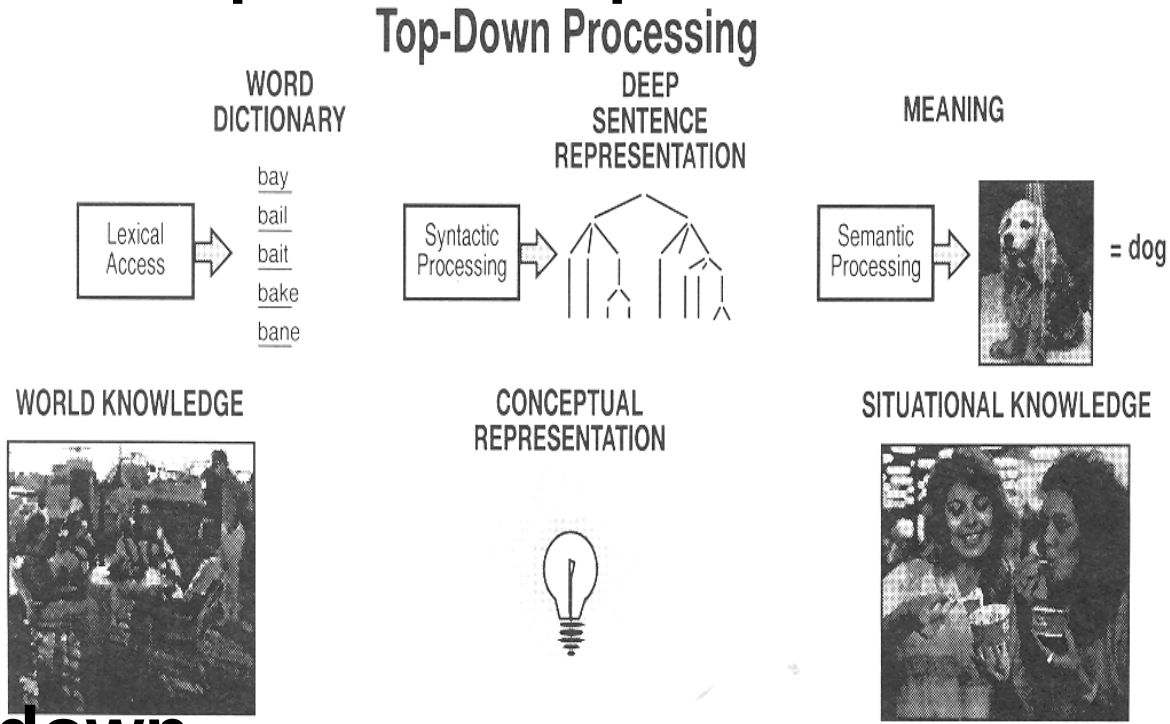
- **Active theories**
  - the process of speech perception involves some aspect of speech production, with the listener viewed as having an active part in the process.
  - Speech sounds are sensed, analyzed for their phonetic properties by reference to how such sounds are produced, and thereby recognized.
- **Passive theories**
  - the process of speech perception is primarily sensory and the listener is relatively passive in this process.
  - The listener has a filtering mechanism with knowledge of speech production and vocal tract characteristics playing a minor role and only in difficult listening situations.

Bottom-up vs. Top-Down Theories



- **Bottom-up**
  - All the information necessary for the recognition of sounds is contained within the acoustic signal.
  - The first stages involve the conversion of the incoming auditory information into a neural signal.
  - Some sort of neural spectrogram reveals the time-varying formant frequencies into speech.
  - From this neural code the perceptual system has to derive the critical phonetic features.
  - The listener doesn’t need to involve linguistic and cognitive processes in decoding sounds.

Bottom-up vs. Top-Down Theories



- **Top-down**
  - higher-level linguistic and cognitive operation plays a crucial role in the identification and analysis of sounds.
  - The listener makes use of stored knowledge that serves to constrain the number of plausible alternative messages.

Autonomous vs. Interactive Theories

- **Autonomous**
  - the signal is processed in a serial manner, from the phonetic to lexical stages, to syntactic stages and so on.
  - The listener's perceptual decision making can be made in a closed, autonomous system that contains all the necessary perceptual operations for such decisions, with no need for other sources of information (e.g., info provided by context).
  - The output of one stage of processing provides the input to the next stage
- **Interactive**
  - information and knowledge from many sources are available to the listener and are involved at any or all stages of processing the signal on it's way through the speech perception system.

Motor Theory - Classification

- Active
- Bottom-up
- Autonomous

Stevens & Blumstein Acoustic Landmarks

1) Landmark detection. Points of maximal and minimal change.
2) Measure acoustic correlates in vicinity of landmarks.
3) Estimate distinctive features and syllable structure.
4) Match to lexicon, use lexical info to synthesize a set of landmarks and cues, compare to results of step 2.

Landmarks

- The landmarks and cues are derived from considerations of the articulators. That is, the representation is distinctive features that are useful in speech production.
- The analysis of the signal is based on a process of segmentation and landmark identification. Again, the landmarks are motivated by articulatory considerations.
- Only one underlying representation is present for each lexical item.

Landmarks

There are three sets of landmarks: vocalic, glide, and consonantal.

- Vocalic - Find the maximum in the F1 region (frequency and amplitude) in a region of no spectral discontinuities.
- For glides (/w/, /j/, /h/) find the F1 profile and the reduction in amplitude in a region of no spectral discontinuities.
- For consonants, find the point of abrupt spectral discontinuity (change in source, closure). These occur in pairs (into and out of constriction).

Articulator-free Features

- The spectral information at the landmarks specify the articulator free features such as [vowel] and [consonant].
- The [consonant] can be further classified as [continuant], [sonorant], and [strident] based on closure ([-continuant]) and the distribution of energy at high frequencies ([+strident] for loud high frequencies).

Articulator-bound Features

- The spectral information around the landmarks is used to specify the features related to the position and movement of the articulators.
- For vowels, this includes high, low, back, round, etc. For consonants, this includes the location of the constriction (lips, tongue blade, tongue body), the state of the vocal folds, etc.

Articulator-bound Features

- The articulator bound features represent the merging of acoustic information, phonetic context, prosodic context, time intervals between landmarks (duration).
- The claim is that a module handles each articulator bound feature (e.g., one for place of articulation, voicing, nasality and liquid for consonants). The modules represent distinct brain structures and are (in a general form) part of the genetic endowment for language.

Landmark Theory - Critique

- The mapping of acoustic correlate to feature not yet sufficiently specified. This makes testing difficult.
- No psychological evidence for landmarks.
- If an iterative component is present, see earlier critique about analysis-by-synthesis.
- Does prosodic information influence early processing?

Landmark Theory - Classification

- Passive
- Bottom-up
- Autonomous

TRACE

- Elman and McClellan proposed TRACE as a multi-stage model that consists of an auditory (ear) front end, auditory feature extraction, a phonetic level, and a lexical level.
- TRACE is implemented in a connectionist architecture and has both ascending and descending (feedback) connections as well as connections within each level.
- TRACE is both a theory and a model of perception.

Connectionist Models

- a/k/a PDP or neural networks
- Class of neurally inspired information processing models that attempt to model information processing the way it actually takes place in the brain.
- A system of neural connections appeared to be distributed in a parallel array in addition to serial pathways.
- Different types of mental processing are considered to be distributed throughout a highly complex neural network.
- Information processing takes place through interactions of large numbers of simple processing elements called units, each sending excitatory and inhibitory signals to other units.

TRACE

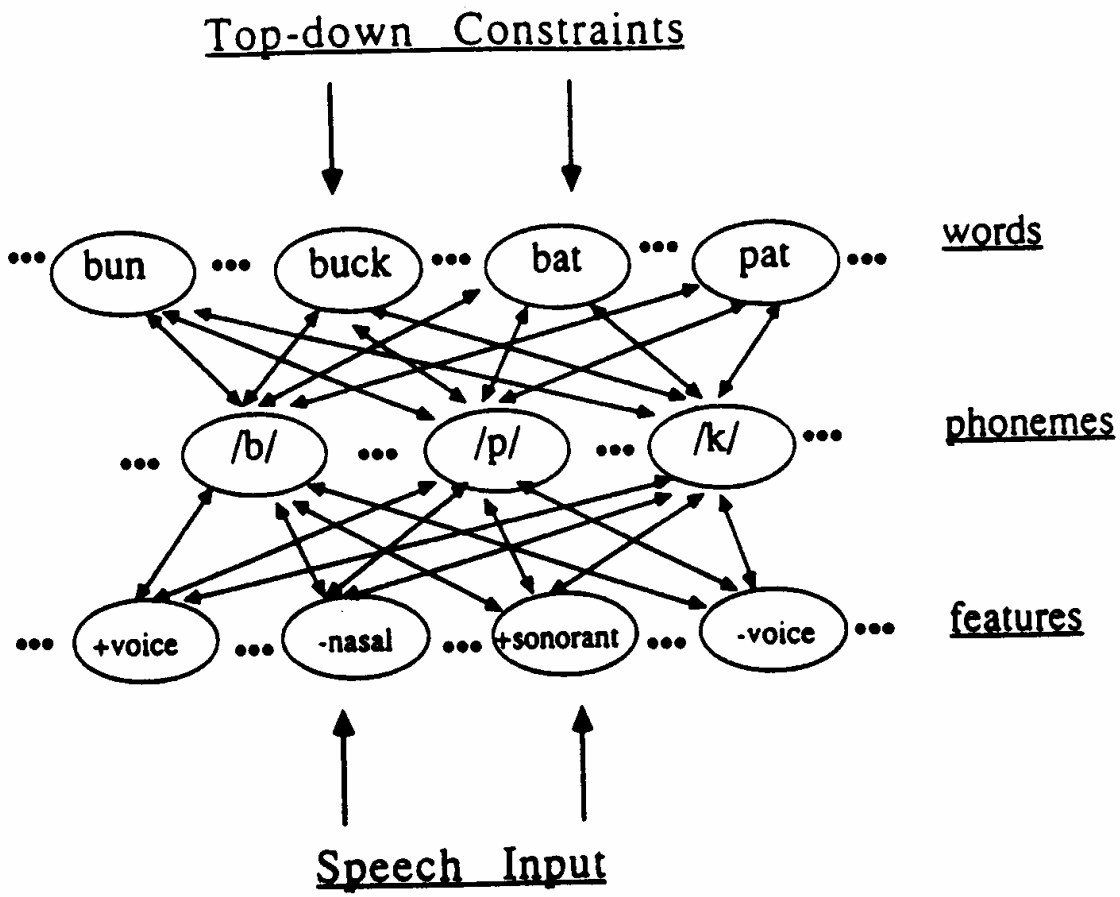


FIGURE 6-9 The TRACE Model of Speech Perception. *(Reprinted with permission from Goldinger, S.D., Pisoni, D.B., & Luce, P.A. [1996]. Speech perception and spoken word recognition: Research and Theory. In Lass, N.J. (Ed.),* **Principles of Experimental Phonetics.** *St. Louis: Mosby, Inc. 277–327.)*

TRACE

- Multiple levels of representation as well as feed-forward and feedback connections between processing units (nodes).
- Nodes are arranged on three levels that together, form a network
  - Phonetic feature
  - Phoneme
  - Word
- Activation on one level increases the activity of all connected nodes on adjacent levels (bottom-up or top-down).
- Within all levels, nodes are connected by inhibitory links, forcing rapid resolution of any ambiguity in the signal (i.e., suppressing competing nodes).

Trace – Key elements

- Invariant cues are not required. Perception is a result of a cascade of stages involving a one-to-many and many-to-one mapping (behaves like a prototype system).
- Feedback and competition among nodes at the same level are used to stabilize perception.

Trace - Critique

- Some aspects of connectionist architecture are very implausible.
- Only implements limited set of features, phonemes, and words. Unclear if this can be scaled to the full range of voices, speaking rates, phonemes and words of spoken language (is this robust?).
- No separate justification for mapping of cues to phonemes other than it can be learned by model (using back-propagation learning).

Trace - Classification

- Active
- Top-Down
- Interactive