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).

Types of Theories

Theories of speech perception fall into one of three broad classes:

1) Motor theories - Perception involves processes related to the production of speech. Examples include Motor Theory and Analysis-by-Synthesis.

2) Direct perception - Perception recovers the sound producing objects directly (via invariants?). Examples include Fowler’ s Direct Realist approach.

3) Stage theories - Perception involves a sequence of transforms from sound to object. Examples include TRACE and LAFS.

Motor Theory

Motor Theory has, as its core, the premise that perception involves a reference to articulation. This view is often associated with the idea that speech is somehow “special” and involves specialized, species-specific mechanisms in perception.

Motor Theory - 2

Liberman and Mattingly describe a Motor Theory that involves analysis-by-synthesis. That is, the mapping of auditory patterns or features onto a language specific representation (features or phonemes or gestures) is accomplished by a mechanism that takes a proposed unit, generates the equivalent auditory pattern and matches it against the input. The degree of match/mismatch is used as feedback to correct the proposed unit and this cycle continues (iterates) until a sufficiently accurate match is achieved.

Motor Theory - Key Elements

The distinctive elements are:

1) Analysis-by-synthesis (perception is an iterative process, the iteration involves articulation).
2) Perception is specialized for speech (biologically).
3) Perception does not have to involve invariant attributes from the signal. The “invariance” is in the perceiver.

Empirical Support

Evidence cited as supporting Motor Theory - Speech is Special.

1) Categorical perception. Listeners can discriminate among tokens only to the extent that they give them different labels. Auditory and speech classification diverge.

2) Trading relations. Listeners appear to integrate acoustic correlates to phonetic categories so that they match articulatory constraints. Put another way, articulation provides a rationale for why cues trade and are integrated.

Empirical Support - 2

3) Auditory - Visual integration. The specialized module for speech integrates all sources of phonetic information.

4) Competition between speech and auditory modes. Duplex perception illustrates that aspects of a stimulus can be used for phonetic perception or “ordinary” perception, but not both.

5) Provides a “natural” account for the nature of the relation between production and perception.

6) Mirror neurons – link between perception and production

Critique of Motor Theory

1) Many of claims are not unique. For example, if perception is categorical, this does not imply Motor Theory is correct (etc.).

2) Some of arguments rest on intuition about nature of percept that has no empirical foundation. These arguments are tenuous.

3) Analysis-by-synthesis mechanism has not been specified. Furthermore, this type of iterative computation tends to be slow and may not result in a “real-time” system.

Critique of Motor Theory - 2

4) Many of predictions of differences between auditory perception and phonetic perception rest on null results.

5) Mirror neurons may or may not reflect mechanisms in perception.

Direct Realism

Based on the work of Fowler, this is a neo-Gibsonian perspective.

Perception is a direct mapping from acoustic qualities to the gestures that produced them. This is framed within a perspective where all of perception is the direct recovery of the distal source of the event being perceived. (Note - Emphasis on perceptual unit being events.)

Direct Realism - Key Elements

The distinctive elements are:

1) Perception is a single step from signal to percept.
2) The percept is the gesture or object that produced the event.
3) An invariant must be present (?) to mediate the mapping.

Direct Realism - Empirical Support

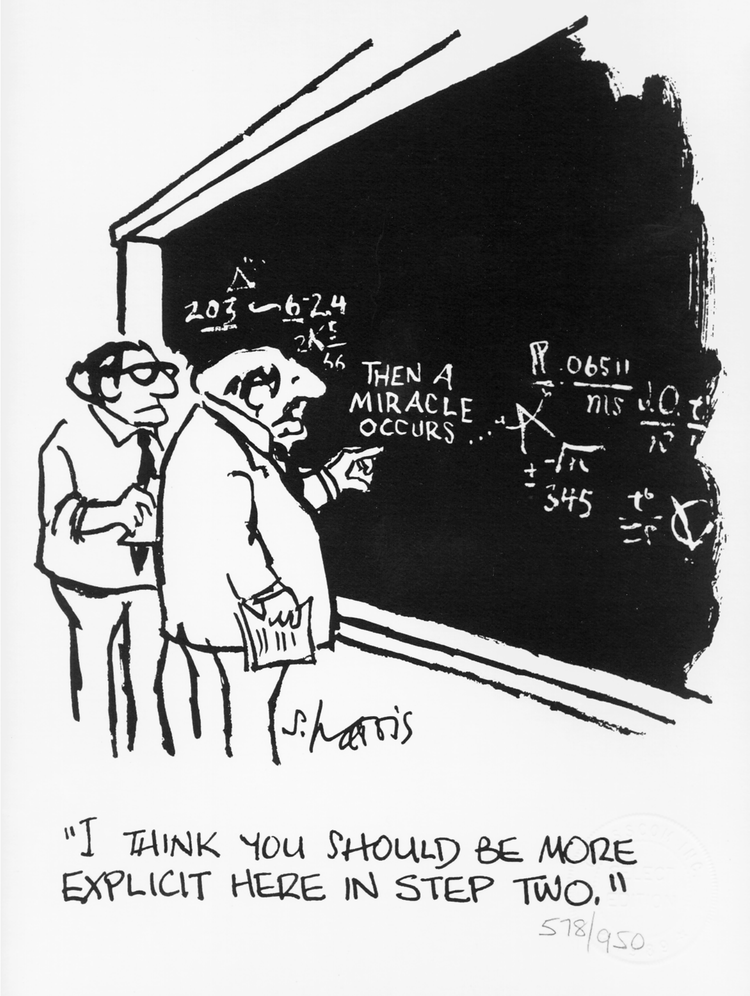
All of data cited for Motor Theory could be used here.

Claims to provide a more unified account by placing speech perception within the larger context of event perception.

Critique of Direct Realism

1) See Motor Theory critique.
2) Vastly underestimates computational complexity of perception.
3) Evidence for intervening representations in perception. Contradicts “direct” aspect?
4) Lack of evidence for invariant qualities in signal.

You say you have a theory?


THEN A MIRACLE OCCURS . . .
S. Harris
"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."
578/950

The result of underestimating the complexity of perceptual processing in a theory.

Stage Theories

Diverse set of theories that do not assume a link between production and perception. Some use invariant attributes (LAFS, Stevens & Blumstein, locus theory), others do not (TRACE). Some have intermediate representations (phonemes or features, TRACE), others map directly to lexicon (LAFS). Some implemented in a connectionist architecture (TRACE), others use an algorithmic format.

Role and nature of segmental (phonetic) representation is diverse.

Stage Theories - Key Elements

1) Coding is based on auditory processes.
2) All use intermediate representations though nature of representations is diverse.
3) All use an information processing framework (perception is the result of a sequence of transformations).

LAFS - Lexical Access From Spectra

In LAFS, Klatt proposed that the input is an auditory representation of the signal. This representation is a series of spectral sections.

A finite-state network parses the input. The path through the network that results from parsing an input is a word. That is, this system maps a sequence of spectral sections onto a word. Parts of the network that correspond to sequences of spectral sections are isomorphic to “diphones” (a type of context sensitive allophone).

LAFS - Key Elements

1) The invariant for perception is a characterization of the spectral shape, over time.
2) The "perceptual unit" is the context sensitive allophone, but listeners have no direct access to this representation (*phonetic perception is lexically mediated*).
3) Processing is controlled by a temporal parsing process (implemented as a finite state machine).
4) Note that Hillenbrand's model of vowel recognition is similar to LAFS.

Critique of LAFS

1) Empirical data do not support a strong role for spectral sections as the “cues” to phonemes for voiced speech. (Nor do they disconfirm.)

2) No good evidence for a diphone-like representation in speech. Conversely, no evidence against it.

3) The finite-state architecture is prone to catastrophic failure (does not show graceful degradation). This is due to violating least commitment, the lack of intervening stages of processing, and the serial nature of the finite state machine.

TRACE

Elman and McClellan proposed TRACE as a stages 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, in its two versions, a model of perception.

TRACE - Key Elements

1) 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).
2) There are two variants of TRACE. One uses a triphone (context-sensitive allophone) representation and the other an abstract phoneme.
3) Feedback and competition among nodes at the same level are used to stabilize perception.

Critique of Trace

1) Some aspects of (TRACE's) connectionist architecture are very implausible.
2) 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?).
3) No separate justification for mapping of cues to phonemes other than it can be learned by model (using back-propagation learning).

Stevens - Acoustic Landmarks and Distinctive Features

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.

See *JASA*, *111*, 1872-1891.

Stevens - Key Elements

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 frequency region (frequency and amplitude) in a temporal 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 - 2

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 (cf. Motor 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 critique above about analysis-by-synthesis.

Does prosodic information influence early processing?

Summary

While the theories appear to differ substantially, distinguishing among them is difficult for two reasons.

First, all are designed to account for the same basic phenomena. This ensures that they will also make similar predictions in many cases.

Second, they are not always explicit on many aspects of perception so that one can always simply claim that a revised model/theory can account for new results.