Invariance and Variability

- Listeners perception is relatively constant while the stimulus varies (perceptual constancy).
- Phonetic perception is an example of categorization.
- Are the categories defined by necessary and sufficient conditions (invariants) or probabilistically?

Variability

Sources of Variability in the Speech Signal:

1) Talker (vocal tract length, idiolect)
2) Speaking rate
3) Context (phonetic, lexical, semantic)
4) Other?

Perception: Sound to Phoneme

In order to understand phonetic perception:

1) Establish the acoustic correlates of phonetic categories.

2) Establish the "cue value" of acoustic correlates (both individually and in combination with other correlates).

3) Provide a description of an algorithm for mapping sound to category.

Allophonic Variation

When there are alternative versions of a phoneme, these are called allophones.

For example, the syllable initial /p/ in /pat/ (“pot” or [p^hat]) is aspirated while the /p/ in /spat/ (“spot” or [spat]) is not.

These two versions of /p/ - [p] and [p^h] are called allophones.

Does perception recover the allophone or the phoneme? Which is used in word recognition?

Stop Place

Acoustic Correlates:

1) Direction and extent of the second formant (F2) transition.

2) Direction and extent of the third formant (F3) transition.

3) Spectral structure, intensity and duration of release burst.

Stop Manner

Acoustic Correlates:

1) Closure (presence, duration)
2) Amplitude rise-time at release
3) Presence and amplitude of release burst
4) Duration of formant transitions
5) Extent of formant transitions

Stop Voicing

Acoustic Correlates (stressed syllable initial):

1) Voice Onset Time (VOT) - Duration (time) between the release and the onset of voicing.

2) Onset frequency and extent of F1 transition.

3) Duration of aspiration at onset.

4) Intensity and duration of release burst.

5) Duration of closure prior to release.

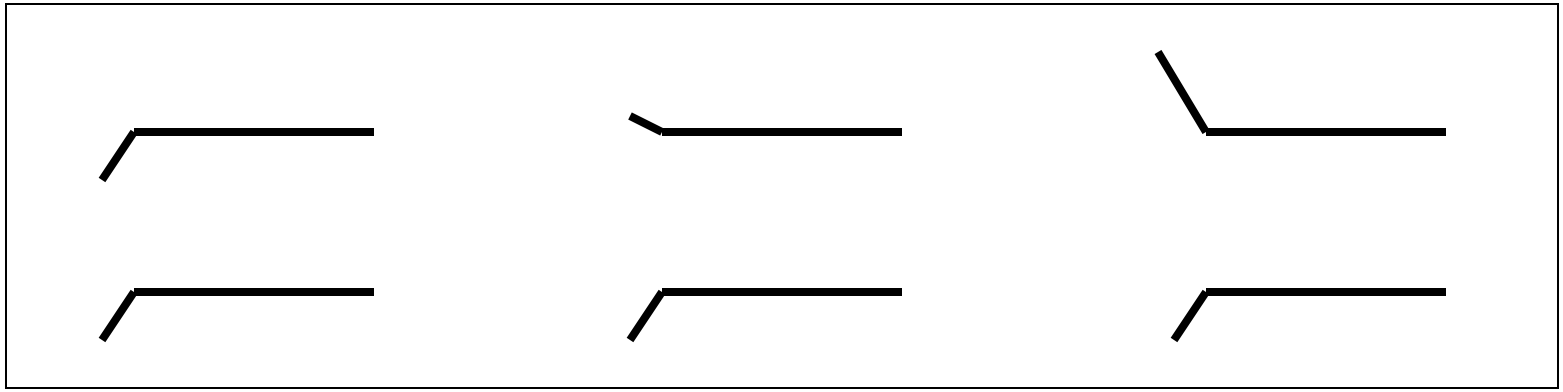
6) Fundamental frequency (F0).

Lisker (1986) notes other correlates for voiceless stops in intervocalic position.

Stop Place - F2 Transition

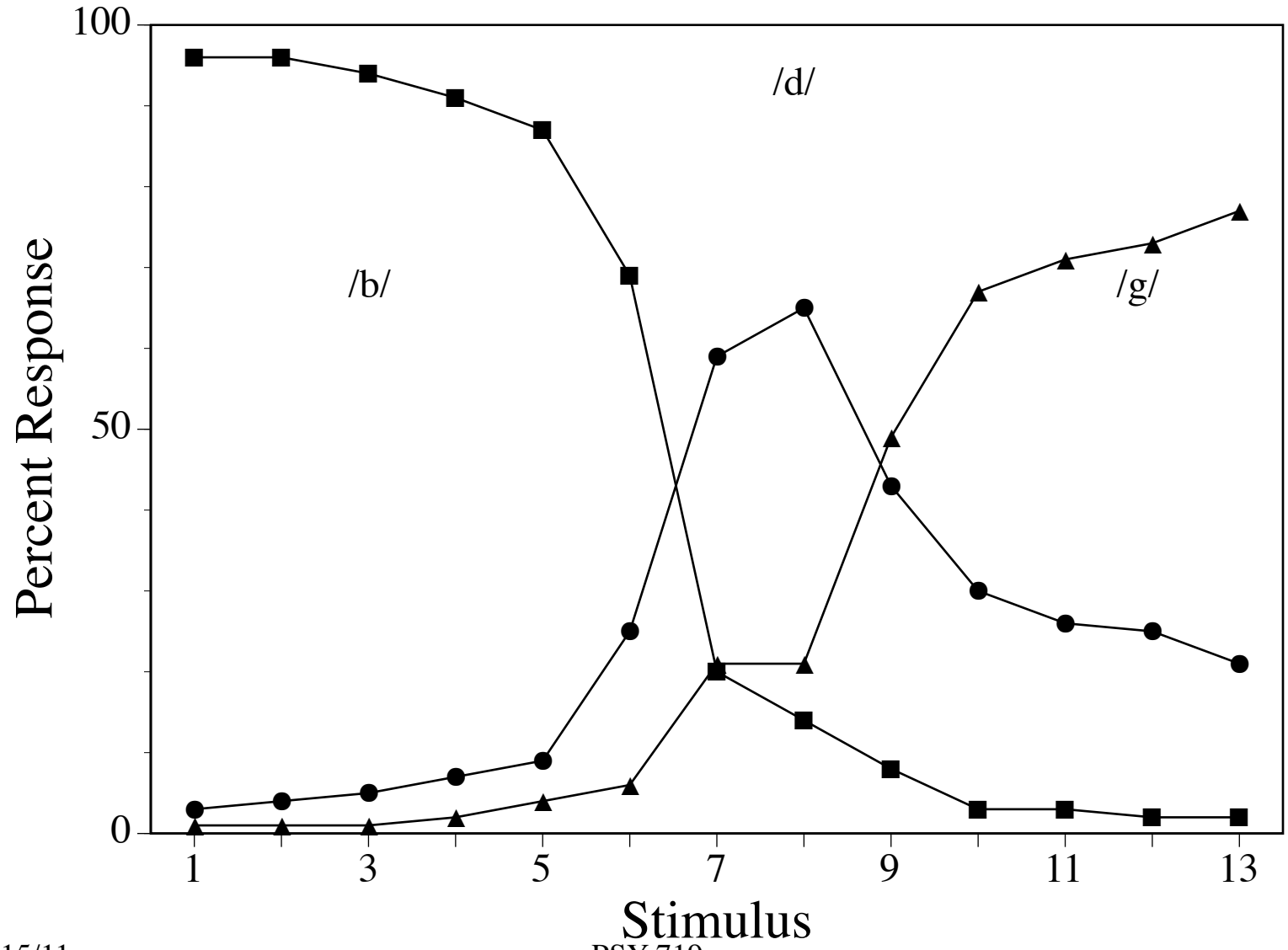
If the direction and extent of the F2 transition before a vowel is varied (from rising through flat to falling), we can examine the influence on listeners' identification of stop consonant place.

Two-Formant Syllables

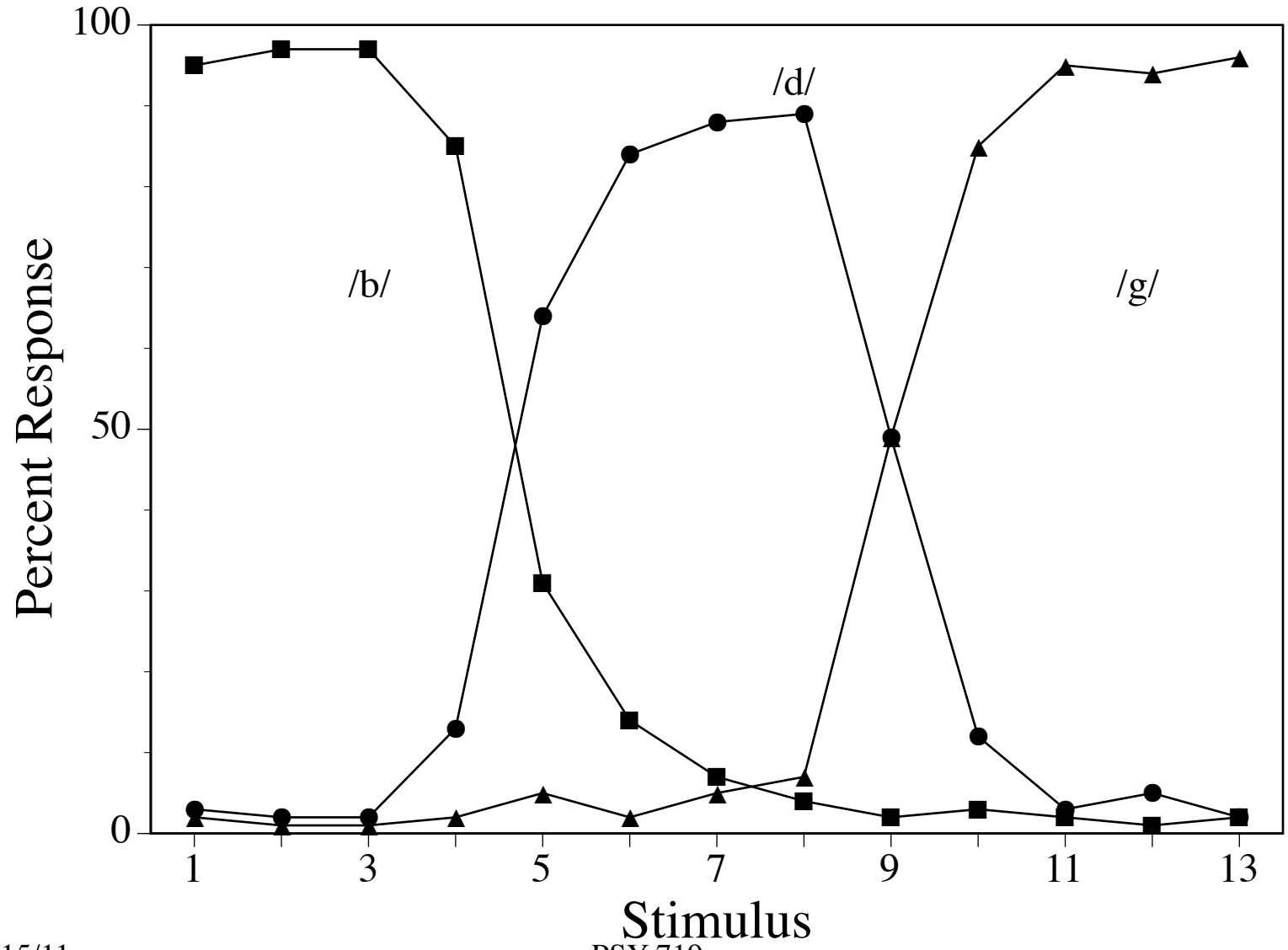


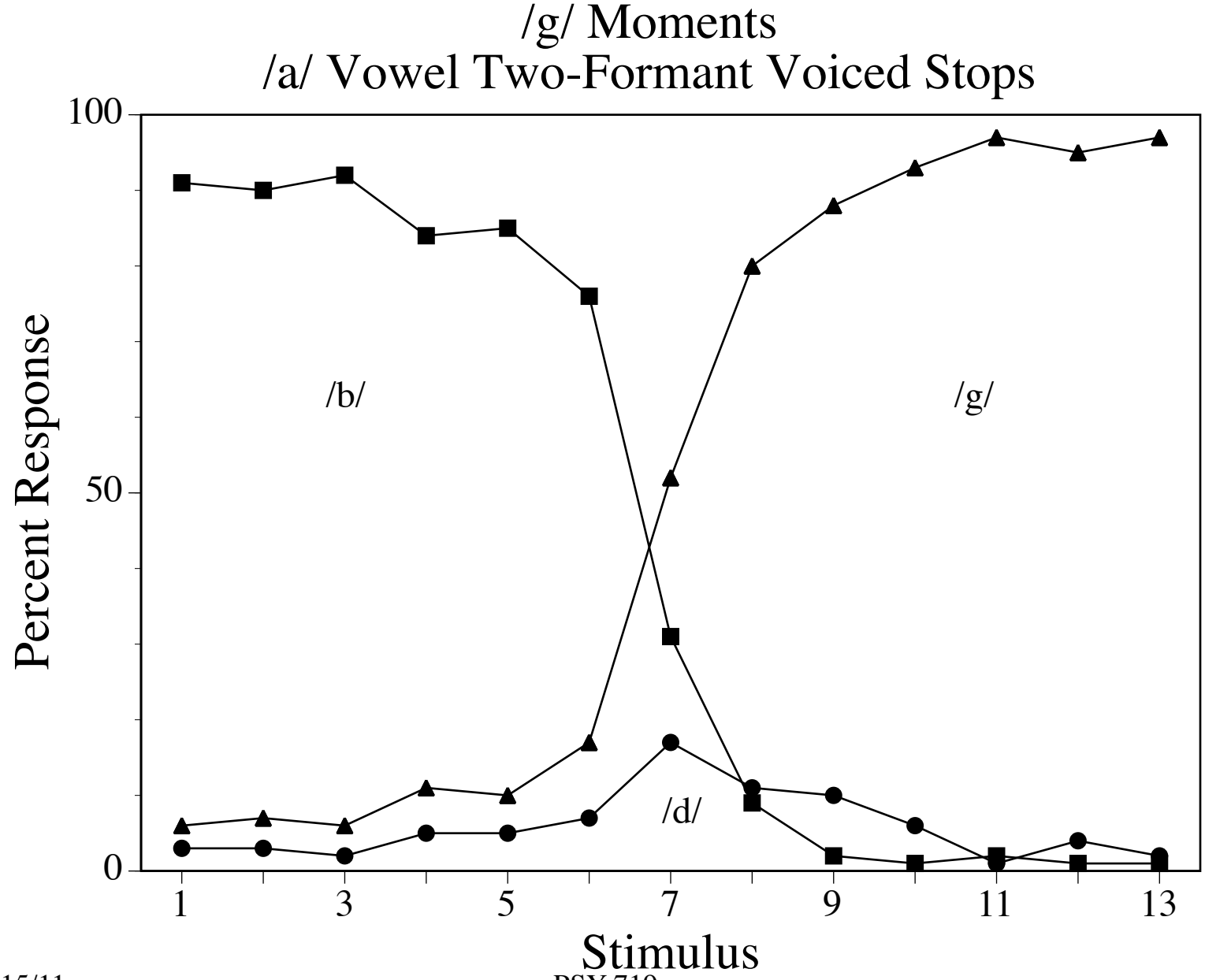
Stimulus 1 Stimulus 7 Stimulus 13

/a/ Vowel Two-Formant Voiced Stops



/a/ Vowel Two-Formant Voiced Stops

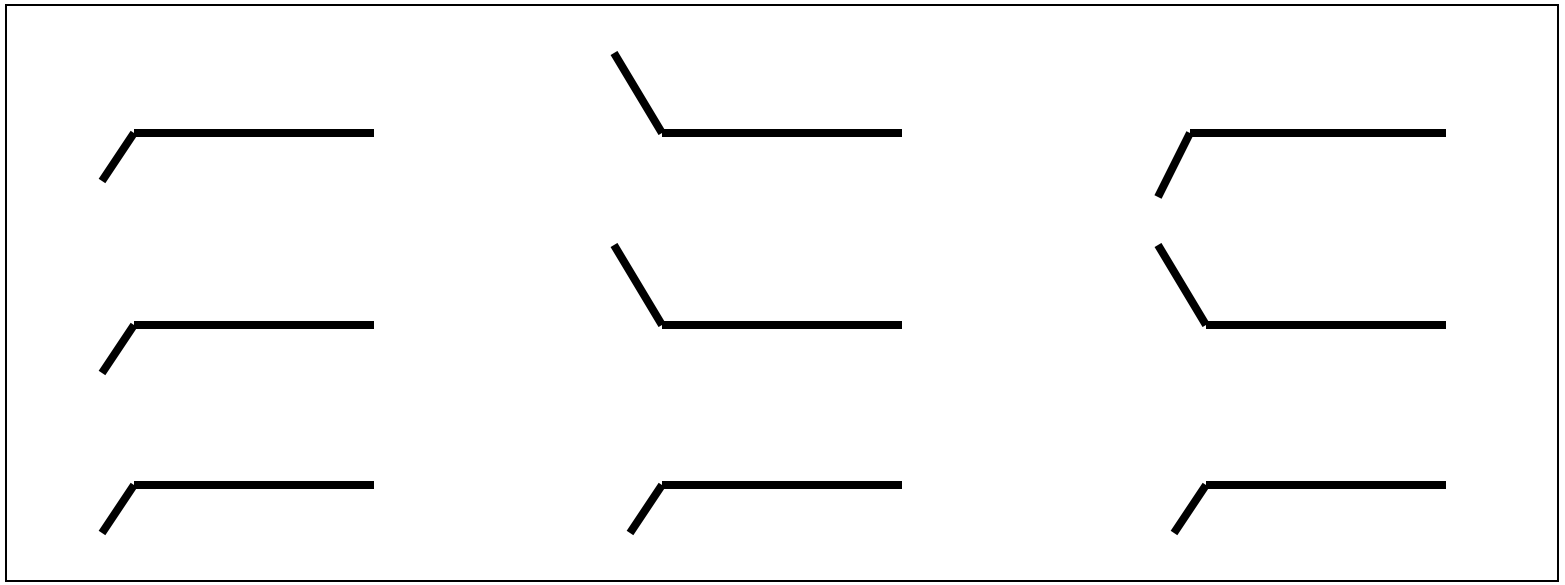



/g/ Moments
/a/ Vowel Two-Formant Voiced Stops
100
50
0
Percent Response
/b/
/g/
/d/
1
3
5
7
9
11
13
Stimulus

Stop Place - F3 Transition

The direction and extent of the F3 transition also influences listeners' identification of stop place.

Three Formant Stops



/bæ/ /dæ/ /gæ/

Stop Place - Release Burst

Varying the frequency of the release burst alters the listeners' identification of the tokens. In general, high frequency release bursts are identified as alveolar (/d/, /t/), middle frequency bursts as velar (/g/, /k/) and low frequency or diffuse, wide band bursts as bilabial (/b/, /p/).

Stop Place - Cue Integration

In syllable initial position, when the F2, F3 and burst cues are combined, the values that yield the best examples and the location(s) of the boundary between categories varies with:

1) Following vowel

2) Values of the other cues

3) VOT

Stop Place - Cue Integration 2

Listeners' identification of the tokens is relatively uninfluenced by formant amplitudes, shape or tilt of the spectrum (when voiced formant transitions are unambiguous), or F0 (fundamental frequency).

What is the Perceptual Cue to Stop Place?

1) A linear combination of the F2, F3, and burst cues.

2) A context dependent combination (nonlinear).

3) Is there an alternate algorithm (definition) for the acoustic correlates and what the perceptual system does (e.g. Stevens & Blumstein - Spectral Tilt or Forrest et al. – Spectral Shape)?

Linear Combinations - Locus Theory

Sussman has proposed the F2 locus as the dominant cue to place perception in voiced stops. Additional cues (F3, burst) are combined linearly to produce percept.

Locus is the trajectory for the F2 transition. The hypothetical start point for an initial stop and a value from the "end" of the stop are combined to determine the trajectory. This is compared against 4 "criteria" for classification (b, d, g-front, g-back).

Critique

1) Cues are not separate components that are linearly combined?
2) Fowler claims that loci measured from natural production don't match theory.
3) See BBS critiques

Pattern - Flow of Spectrum

Stevens (1975) proposed that the change in the distribution of energy in the spectrum (from release into following phoneme) characterizes the stops. Stevens did not specify what/how acoustic cues in the spectrum map onto the three patterns for initial stops.

The primary critique is that this is incomplete.

Pattern - 2

Stevens & Blumstein proposed that the shape of the spectrum, at stop release, cues stop place. This is also known as spectral tilt.

Diffuse - falling	corresponds to bilabial
Diffuse - rising	corresponds to alveolar
Compact	corresponds to velar

Spectral tilt critique

1) On further measurements, it isn't as "invariant" as originally thought.
2) Not computationally specified.
3) Can not handle two-formant stops.
4) When put in conflict with the formant transitions, perception dominated by formant transitions for voiced stops.

Spectral tilt revised

Revisions to the spectral tilt take one of two forms:

1) Lahiri et al. Change in spectral tilt over time. Specified for bilabial and alveolar.
2) Forrest et al. Spectral moments and the change in the spectral moments over time.

Further critique

Dorman showed that when the Lahiri proposal is put in competition with the formant transitions, perception follows the transitions.

Spectral moments, over time, can be de-coupled from the formant transitions. For voiced stops, perception follows the transitions (Richardson).

Next steps?

Base spectral moments on peaks in the spectrum? Alternative spectral/acoustic representations.

What does the auditory system do with complex signals?

What effect does using an exemplar based approach or changing the nature of the abstract representation have?

What about place of articulation in voiceless stops (see Forrest et al.) or nasals?