Speaking Rate

- Segment duration changes with speaking rate.
- Segment duration is a perceptual cue (to phoneme identity, stress and accent or prosodic structure).
- Other acoustic qualities (e.g. formant frequencies, rate of change in formant frequencies, degree of coarticulation) also may change when speaking rate changes.
- Listeners must somehow adjust perception to match the speaking rate and recover the intended utterance.

The Problem

In fluent speech, humans change their rate of speech dynamically (while talking) and different people speak at different intrinsic rates.

One consequence is that the physical duration of different parts of the speech signal - including the durations of gestures, phonemes and syllables - changes with a change in speaking rate.

This is one of the sources of variability in the signal that listeners must deal with in perception to recover the message intended by the talker.

Complications

The problem is complicated by the fact that one of the major acoustic correlates to many phonetic distinctions is segment duration. Examples include:

- The stop versus semi-vowel distinction
  - “big” versus “wig”
- The fricative-affricate distinction
  - “chop” versus “shop”
- The voiced-voiceless distinction
  - “bush” versus “push” or “bad” versus “bat”
- Some vowel distinctions
  - “had” versus “head”

As speaking rate changes, the durations of acoustic segments that cue these distinctions change.

Implications for Perception

This means that a physical duration that would cue a /w/, /sh/, or /t/ at a rapid rate of speech would be heard as /b/, /ch/, or /d/ at a slower rate of speech. However, listeners generally hear the message intended by the listener. Listeners compensate or normalize.

What information in the signal might listeners use to calibrate their perception? How do listeners normalize for speaking rate?

Complications - 2

English is a stress accent language. Accented syllables are longer and louder and tend to have a greater F0 change than unstressed syllables.

So, a syllable duration that is stressed (long) at a rapid rate of speech is equivalent to that of an unstressed syllable as a slow rate of speech.

Basically, absolute duration is not meaningful. Relative duration, however, cues speaking rate, accent (stress) and phonetic distinctions simultaneously.

Preliminaries - Speech Acoustics

A number of studies have had speakers produce sentences at different speaking rates. Acoustic measurements have then been made of segment durations and formant frequencies.

Some studies show evidence of vowel reduction for rapid speaking rates. Others do not. Formant frequency changes in stops and approximates are harder to quantify (see Miller, 1981).

Speech Acoustics - 2

When speakers change speaking rate, syllable and segment durations change. Within a talker, speaking rate typically varies over a 2:1 to 3:1 range (e.g. syllables of 120 to 360 msec average duration). This same range of variation is also found for stressed versus unstressed versions of the same syllable (see Crystal & House, 1990).

All phonemes (segments) change duration to some extent with changes in speaking rate. Some experiments suggest that the changes are not uniform (linear) and that intrinsically long segments (e.g. vowels) change proportionally more while voiced stops change very little.

Speech Acoustics - Some Limitations

Virtually all studies use read speech. This is done so that comparisons can be made across talkers for the same segments and within talkers for the same segments at different rates.

The problem here is that casual (spontaneous) and read speech may differ systematically. For example, there may be more reduction and assimilation in casual speech. It is unclear if the results for read speech will generalize to casual speech.

Speech Perception - Listener Compensation

Listeners compensate for variation in speaking rate. They appear to adjust perception so that speaking rate variation is "normalized".

The basic approach to investigating this is to create a phonetic contrast that is based on duration, such as /ba/ - /pa/. Then, vary the speaking rate information in the surrounding speech signal. Use listeners' responses to the phonetic contrast (the target or test series) to indicate whether they have used the speaking rate information to normalize perception.

Information (the what)

Numerous studies have show that the segment durations that precede the "target" influence perception. This influence of prior segments is composed of two components:

1) Segment duration immediately preceding the target. /i/ in /hiwɪl/ ("he will") influences perception of /w/.

2) The rhythm or prosodic structure of the phrase preceding the target.

See Summerfield, 1981 and Kidd, 1989.

A Process Overview

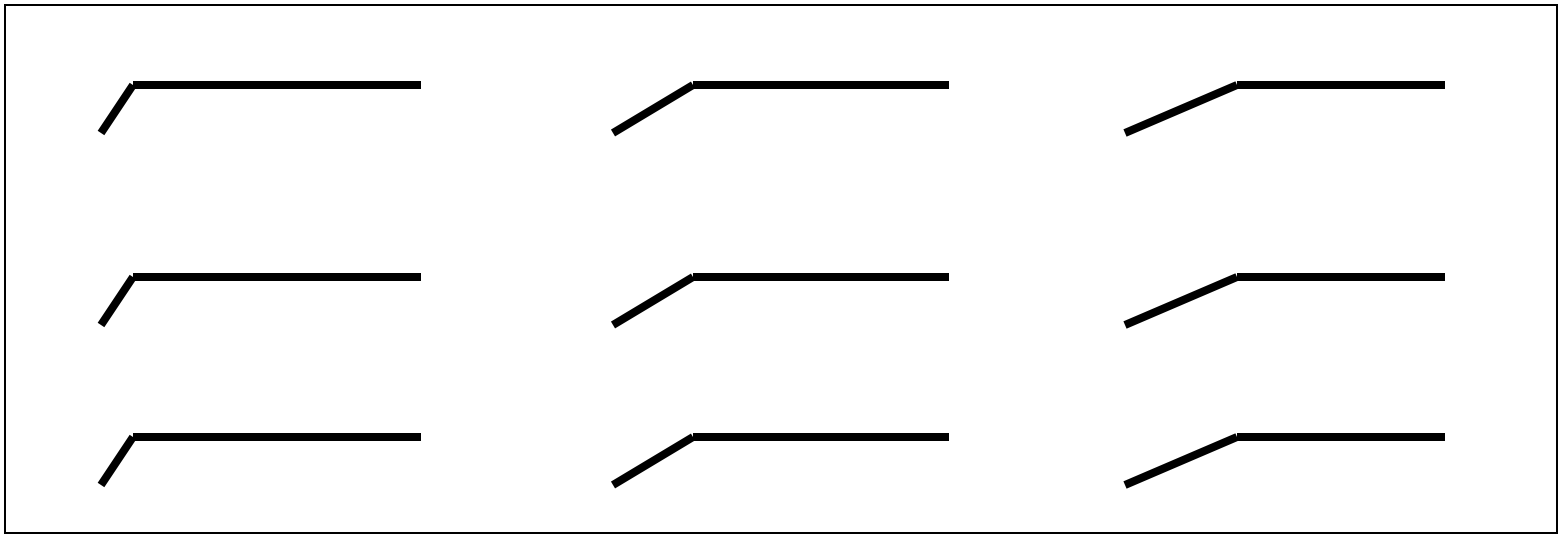
Kidd (see also Summerfield) proposed that there are two component processes in listeners' use of speaking rate information.

1) Long-term: Driven by the rate of stressed syllables.

2) Short term: Driven by segment durations contiguous to the target.

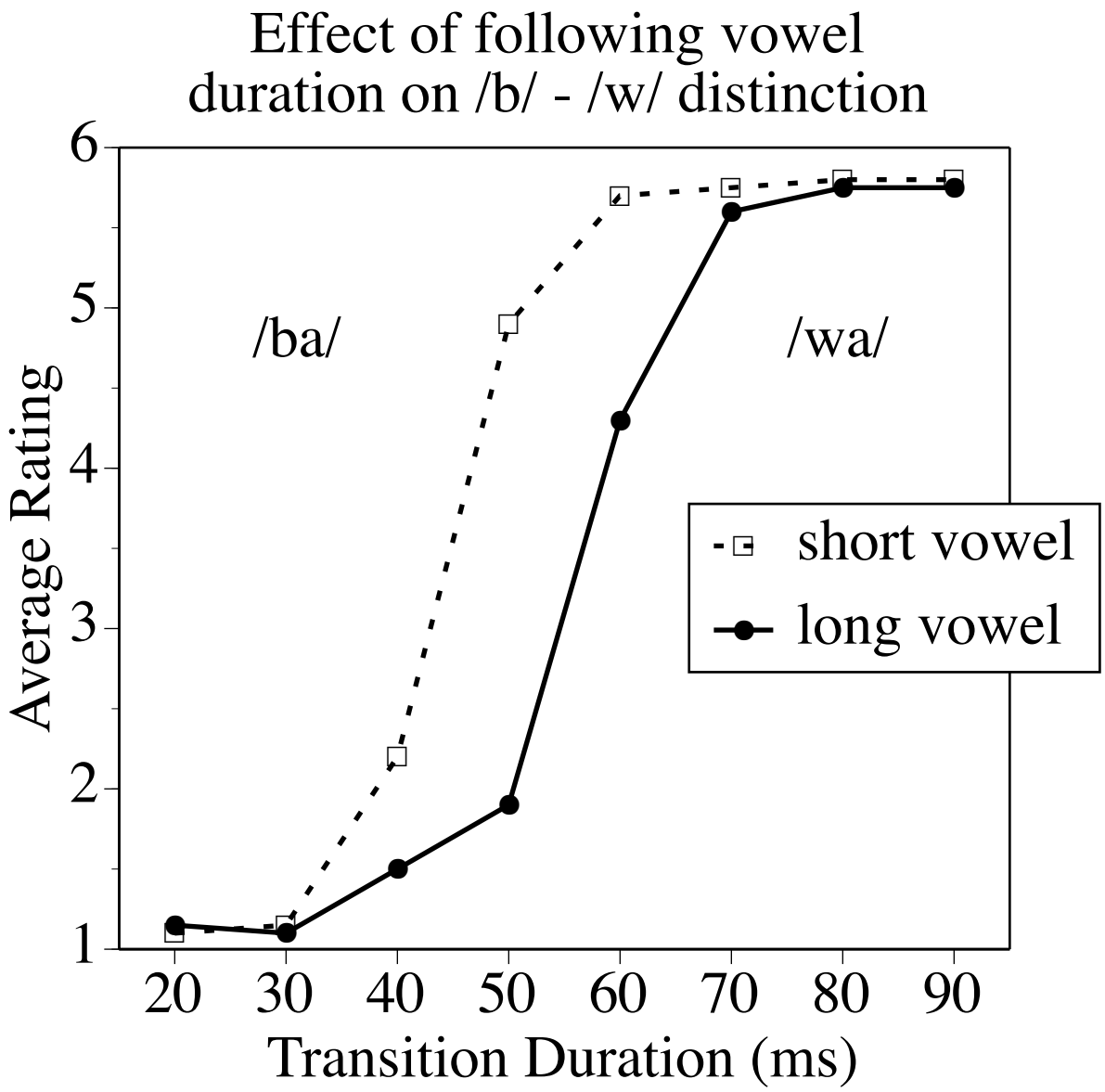
Information - Part 2

Miller and Liberman (1979) and others have also shown that the duration of the segment that *follows* the target will influence perception of the target. For example, synthetic speech series varying from /ba/ to /wa/ are constructed with long (below) and short vowel duration.



Data

Intermediate stimuli in the series are classified as /w/ with a short vowel but as /b/ with a long vowel.



Questions

What is the nature of the knowledge and/or process that influences rate normalization (short and long term)?

Is this knowledge or process domain specific?

Where in the process of perception does rate normalization take place?

Is this normalization? - or - Is this an exemplar based processing system where rate is part of what is stored?

Focus of Answers

In attempting to answer these questions, we will first focus on the "short-range" process and then take a look at the "long-range" process.

Some Answers

Event rate normalization can be found in the perception of other auditory events. Thus, all auditory perception seems to include a rate normalization process.

1. Fowler (1990) showed that nonspeech (auditory) perception seems to adjust for event rate.
2. Pisoni, Carrell & Gans (1983) and Diehl & Walsh (1989) have also shown that auditory perception of speech-like contrasts shows the same type of rate normalization as speech.
3. Welch, Sawusch, & Dent (2009) showed that parakeets exhibit speech normalization similar to humans.

Answers (cont.)

Rate normalization is not unique to speech, but that does not answer the question of whether speaking rate normalization is internal to speech processing.

The fact that non-humans normalize their perception of human speech is suggestive.

Answers - 2

Rate normalization seems to happen early in perception. Phonetic prototypes and phonetic category boundaries both move with changes in segment duration.

If listeners are forced to respond rapidly to a series with a long vowel, their data look like those of listeners hearing the same series with a short vowel (Miller and Dexter, 1988). Thus, processing seems to reflect the obligatory use of the duration information available at the time the phonetic decision is made.

Is this an auditory process prior to language specific processing?

Questions - Round 2

Rate normalization is a part of event perception. It is obligatory and early in processing.

Does it follow principles of speech?

1) Acoustic and/or Phonetic Similarity
2) Vocalicness
3) Phonotactics and Language Regularity

Questions - Alternatives

Does rate normalization follow principles of early, auditory perceptual processing:

1) Adjacency
2) Temporal proximity (contiguity)
3) Continuity of source

Data Summary

The basic pattern of results is easy to summarize. As long as the speech segment that follows the target occurs within a brief temporal window (after the target) then the duration of the segment will influence the perception of the target.

In the sets of stimuli used, no evidence of any influence of similarity, vocalicness or phonotactics could be found.

All speech segment duration information that occurs within a limited temporal window (250 msec) after the target will influence the perception of the target.

Data - 3

For /blos/ - /plos/ and /dlos/ - /tlos/, the is an effect of both adjacent /l/ duration and non-adjacent vowel.

For bush - push series, effect of both adjacent /ʊ/ and non-adjacent /ʃ/. (Similar results for parakeets.)

For /bʌlz/ - /pʌlz/ series, effect of both adjacent /ʌ/ and non-adjacent /l/.

When variation in non-adjacent segment is more than 200-250 msec removed from target, little or no influence (/s/ variation in /tʃæs/- /ʃæs/ or /æ/ in /tʃwæs/ - /ʃwæs/).

Further Exploration

What if there were acoustic cues that the signal changed talker? Would the speech from a second talker influence perception of the target in the first talker?

Alternative 1 - No. The auditory system would stream or segregate the two talkers into two different perceptual groups and rate normalization would be group specific. Lotto et al. proposed this based on results where the F0 of a vowel following a target changed part way through the vowel.

Alternative 2 - Yes. If the signal is phonetically coherent, all information within the temporal window is used.

Data - 4

Series 1: /bʌlz/ - /pʌlz/ with /bʌ/ spoken by female, /lz/ spoken by male. /l/ duration varied.

Series 2: /bi/ - /pi/ with initial part spoken by female but most of vowel spoken by male, male vowel duration varied.

In both cases, variation in duration of second talker altered perception of /b/ - /p/ spoken by first talker.

Data - 5

Series: /ba/ - /pa/ spoken by male. For two series, part way through vowel, stimulus changed to tone analog of male. Duration of vowel varied (control) or duration of tone analog varied.

Series 2: /bʌlz/ - /pʌlz/ spoken by female. /lz/ natural or tone analog. Duration of /l/ varied.

Listeners run with speech instructions or nonspeech instructions (regarding tone).

Data - 6

In speech mode, tone duration variation influences perception. Effect of tone similar to effect of natural speech.

In nonspeech mode, speech duration alters perception but tone duration as NO effect.

The influence of duration variation on a speech contrast occurs after entry into the speech mode.

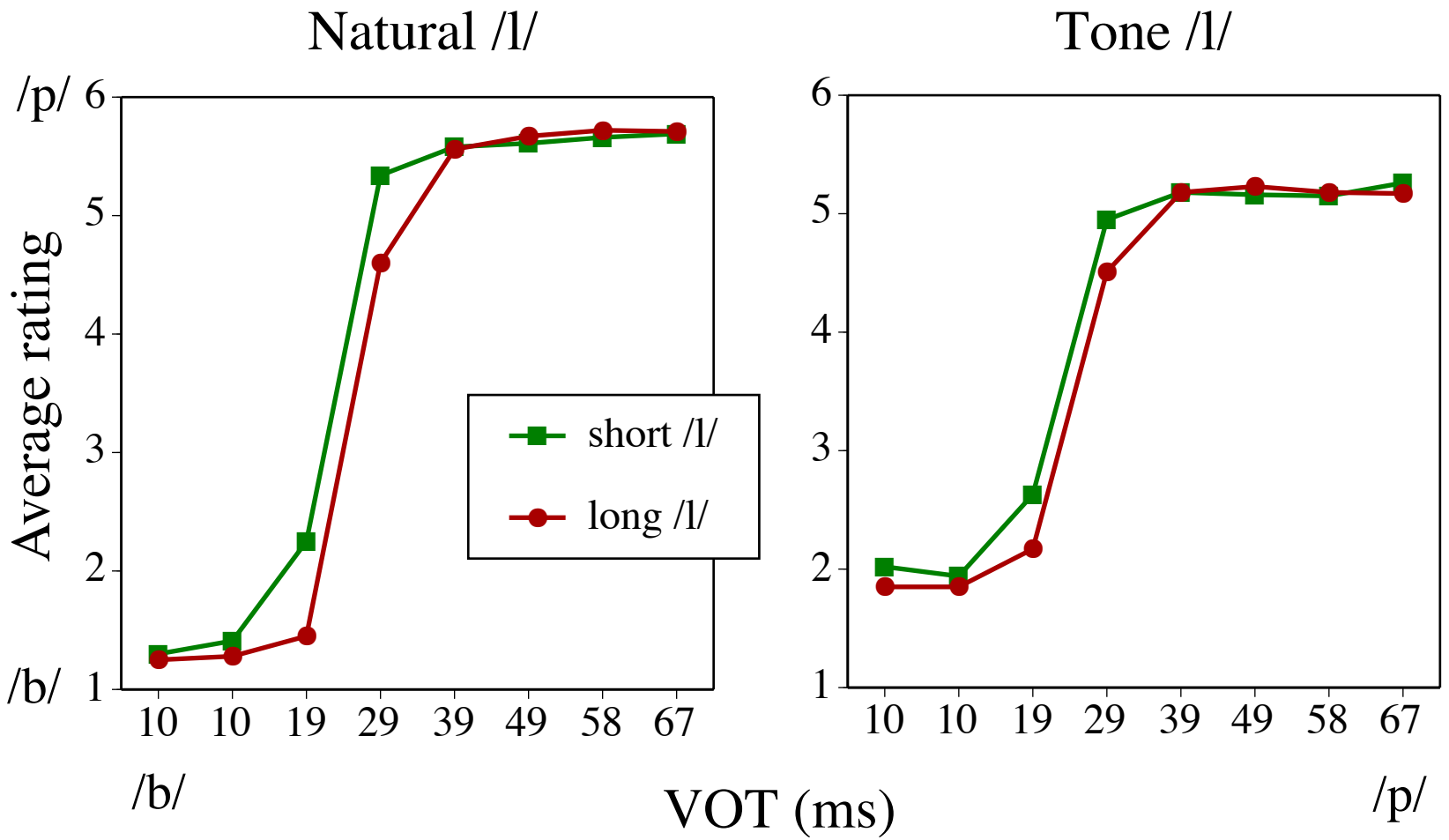
Yet Further Exploration

How does the speech mode use duration variation?

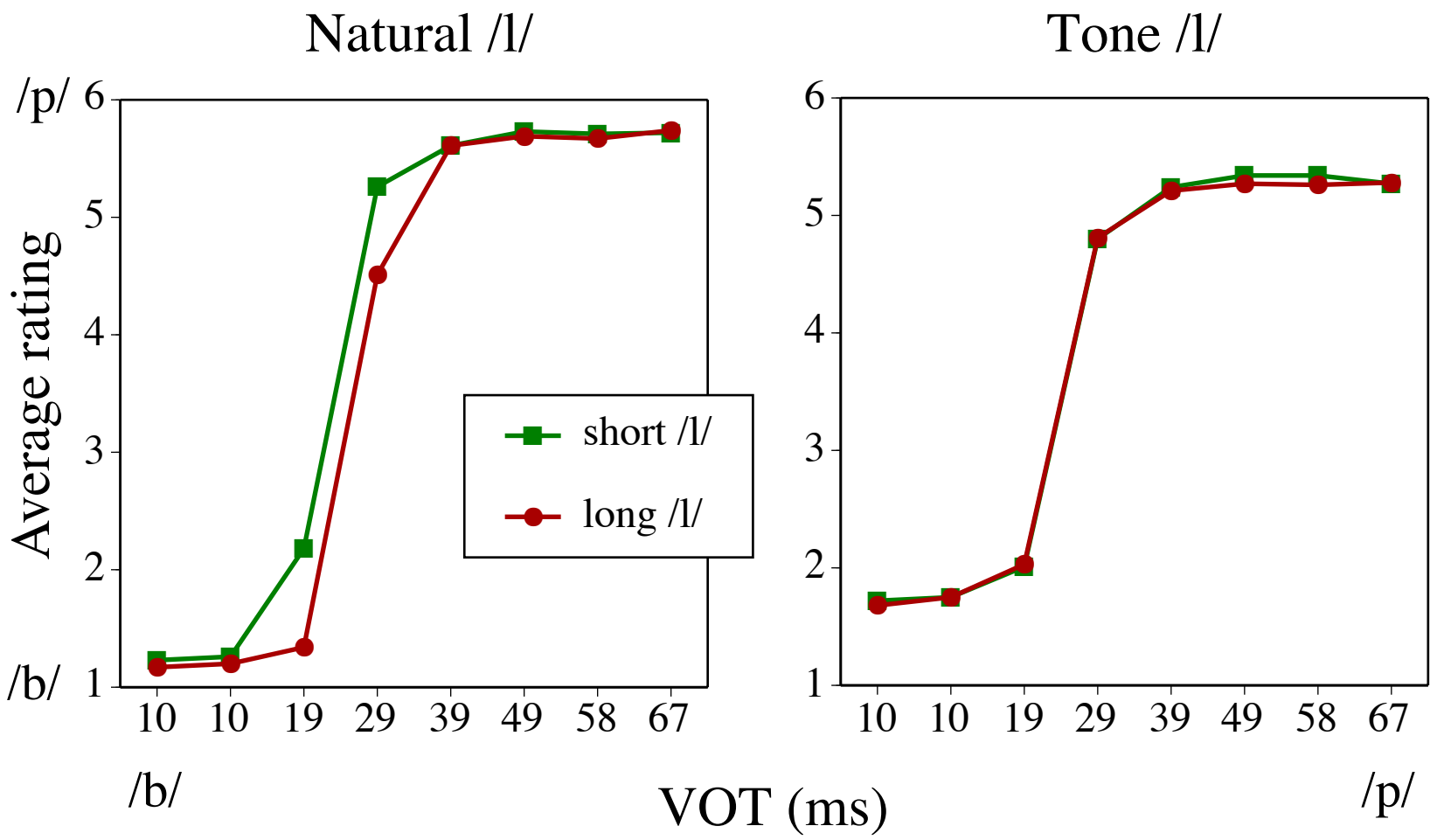
1) Phonetic parsing. The speech processor assigns variation to its source in articulation (Kidd).

2) Autonomous coding. Duration variation from all/any source within 250 msec of the target is used to adjust for speaking rate.

/bʌlz/ - /pʌlz/ Natural/Tone Hybrids
Speech Instructions



/bʌlz/ - /pʌlz/ Natural/Tone Hybrids
Nonspeech Instructions



Data - 7

Kidd proposed that the process of speaking rate normalization parses duration variation to its source in articulation. In order to test this, we need a series that:

1) Has a duration based initial contrast.
2) Has duration variation in the following segment.

But

3) The duration variation is not the result of variation in speaking rate.

Data - 8

Series: /bid/ - /wid/ and /bit/ - /wit/

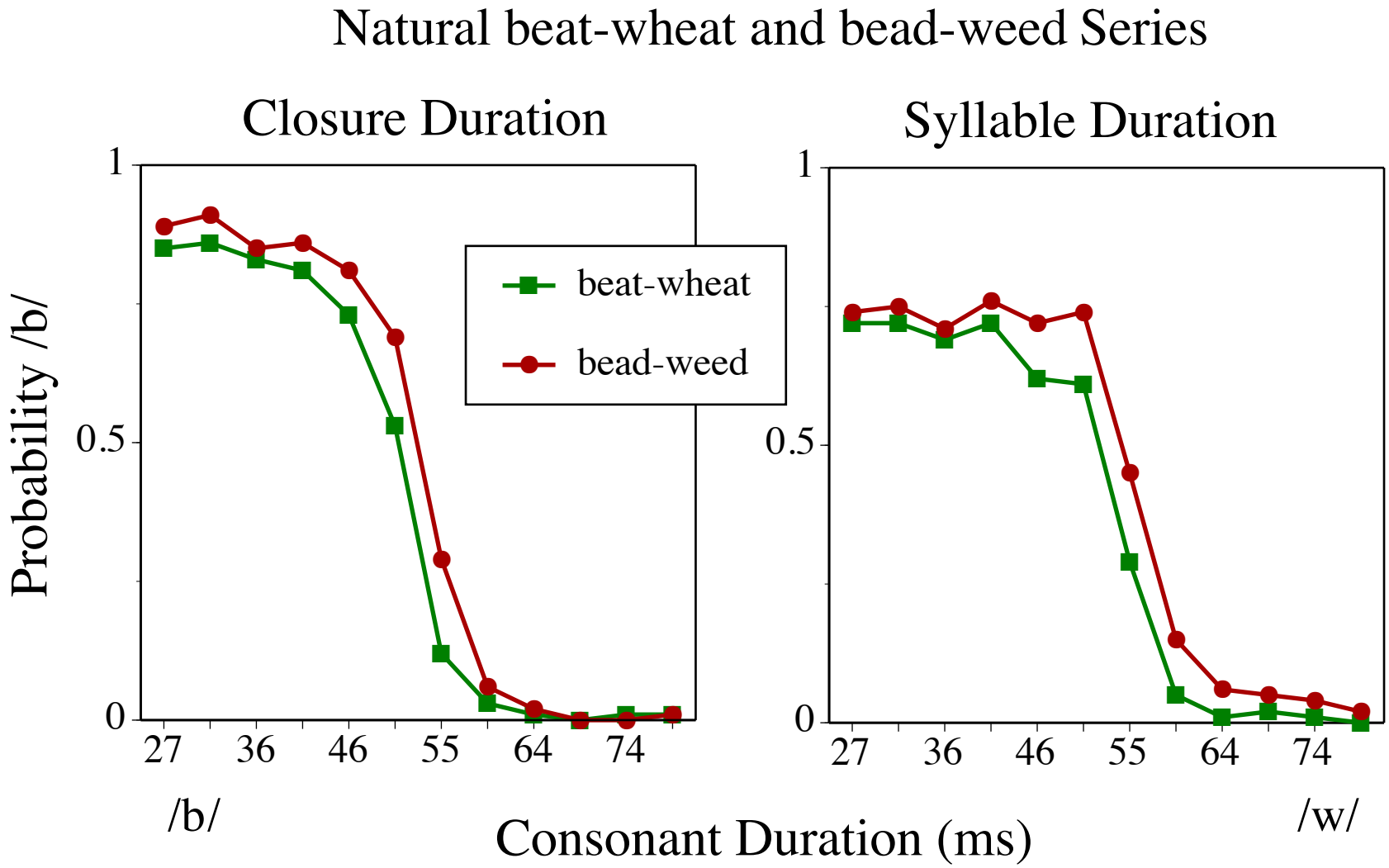
The series with final /d/ have a long vowel. Final /t/ has a short vowel. Vowel duration is the primary phonetic cue to the voicing of the final stop.

Natural tokens recorded in sentence context at a constant speaking rate. Ends of series tested with listeners for perceived speaking rate.

No differences in phonotactics or lexical neighborhood to bias listener.

Data - 9

Category boundary for /b/ - /w/ contrast varies with vowel duration. Long vowel (/bid/ - /wid/) results in more /b/ responses.


Natural beat-wheat and bead-weed Series
Closure Duration
Syllable Duration
Probability /b/
1
0.5
0
beat-wheat
bead-weed
27
36
46
55
64
74
/b/
/w/
Consonant Duration (ms)

Answers - 3

All information within temporal window influences perception as long as it is phonetically coherent and has not been segregated by early auditory grouping.

This process is autonomous (obligatory and independent of other processing) within the speech mode.

Effects of Preceding Phrase

Kidd's data indicate that the durations of the stressed syllables preceding the target was the basis for a long-range influence.

If this is correct, then ONLY speech precursors should be effective in altering perception of the target.

A number of studies have used non-speech precursors and found no influence on a speech target … but …

A Nonspeech Effect

Wade & Holt (2005) used a series of tones.

The durations of the individual tones in each sequence were ~50 msec or ~120 msec. The individual tones were steady-state, but the frequency varied randomly within the range of F2 (or F1). The long and short precursors had different numbers of tone components (so they had the same overall duration but differed in "event rate").

With a synthetic /ba/ - /wa/ test series, the short precursor led to more /w/ responses, the long precursor to more /b/ responses.

Data - 10

A nonspeech, phrase length precursor can influence a speech contrast.

Is this influence in the long-range system? Is it in the short-range (due to the few tones right before the /b/ - /w/ test item)?

Mantell has data that show the influence of the tones is in both the long-range and short-range components.

Now What?

Do precursor and post-cursor influences reflect the same mechanism? If the precursor effect of nonspeech is in the long-range system, how would this alter Kidd's proposal that the durations of stressed syllables governs the long-range effect.

If the nonspeech precursor effect is in the short-range system, what does this say about differences between precursor and post-cursor. Does attention and attentional capture modulate these effects (either directly or indirectly via perceptual grouping).