Computational Modeling of Spoken Language Processing: A hands-on tutorial

The 30th Annual Conference of the Cognitive Science Society

THE NEW SCHOOL

Haskins Laboratories

Computational Modeling of Spoken Language Processing: A hands-on tutorial

Ted Strauss

New School University

Dan Mirman

Jim Magnuson

University of Connecticut Department of Psychology

and

Haskins Laboratories

Goals

- General
  - Illustrate principles of modeling using spoken word recognition as an example domain
- Specific
  - Intensive instruction in using jTRACE to prepare participants to do their own modeling

Plan

- Module 1: Introduction to SWR and TRACE
- Module 2: Tour of jTRACE
- Module 3: Classic simulations
- Module 4: Scripting
- Module 5: Linking hypotheses
- Module 6: Lab time, Q&A, one-on-one

Module 1

- Motivations for modeling
- Review of speech perception and spoken word recognition (SWR) models
- Introduction to TRACE

Principles of spoken word recognition

- Current theories share three core principles (cf. Marslen-Wilson, 1993)
  - As a word is heard:
    1. Multiple words are activated
    2. Activation depends on
       a. Similarity to the input
       b. Word frequency (prior probability)
    3. Activated words compete for recognition

Fundamental problems for SWR

- Precisely which items are activated (similarity metric)?
- Segmentation / alignment problems
- How is competition resolved?
- Fluent speech vs. isolated words
- Learning
- Connections to production, semantics (word/sentence/beyond)

Why model?

- Minor differences in similarity metric, competition mechanism, etc., lead to intuitive differences
- What are the precise differences?
- With just a few assumptions operating simultaneously, analytic prediction becomes difficult if not intractable
- Prediction via simulation
  - Forces precise specification of assumptions
  - When faced with demands of real processing, simpler solutions may emerge
  - OR seemingly logical predictions may be falsified

Psychological models vs. ASR

- Keep in mind: our goal is to develop psychological models
- These will not perform as well as ASR systems
- No current psychological models of word recognition work directly with speech
- But ASR systems seem to operate very differently than human speech recognition **and** are not psychologically tractable

Different kinds of models

Verbal / box and arrow	Cohort (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987)
Mathematical	Neighborhood Activation Model (Luce, 1986; Luce & Pisoni, 1998)
Simulating	TRACE (McClelland & Elman, 1986) Shortlist/Merge (Norris, 1994; Norris et al., 2000) PARSYN (Auer, Luce et al., 2000) SRNs (Elman, 1990; Gaskell & Marslen-Wilson, 1997; Magnuson et al., 2003) Plaut & Kello (1999) ART (e.g., Grossberg et al., 1997, 2000)

Comparing types of models

- Nature of the competitor set
- Cohort and NAM make conflicting predictions
- Can simulated time course help resolve the conflict?

Verbal/box & arrow

- Cohort I, II: <u>precise </u>verbal models
- Make optimal use of speech: activate based on matches, inhibit based on mismatch
- Exploits temporal nature of speech for segmentation
- Predictions: ordinal/relative to information density
- Evidence: *cat* primes *sugar* (via *candy*) but not *chair* (via *sat*)

k	ae	t
toad		
ghost		
coat		
coast		
keen		
cave		
catch	catch	
cast	cast	
candy	candy	
castle	castle	
cat	cat	cat
cattle	cattle	cattle
catapult	catapult	catapult

Mathematical

- Neighborhood Activation Model (NAM)

  Luce (1986), Luce & Pisoni (1998)

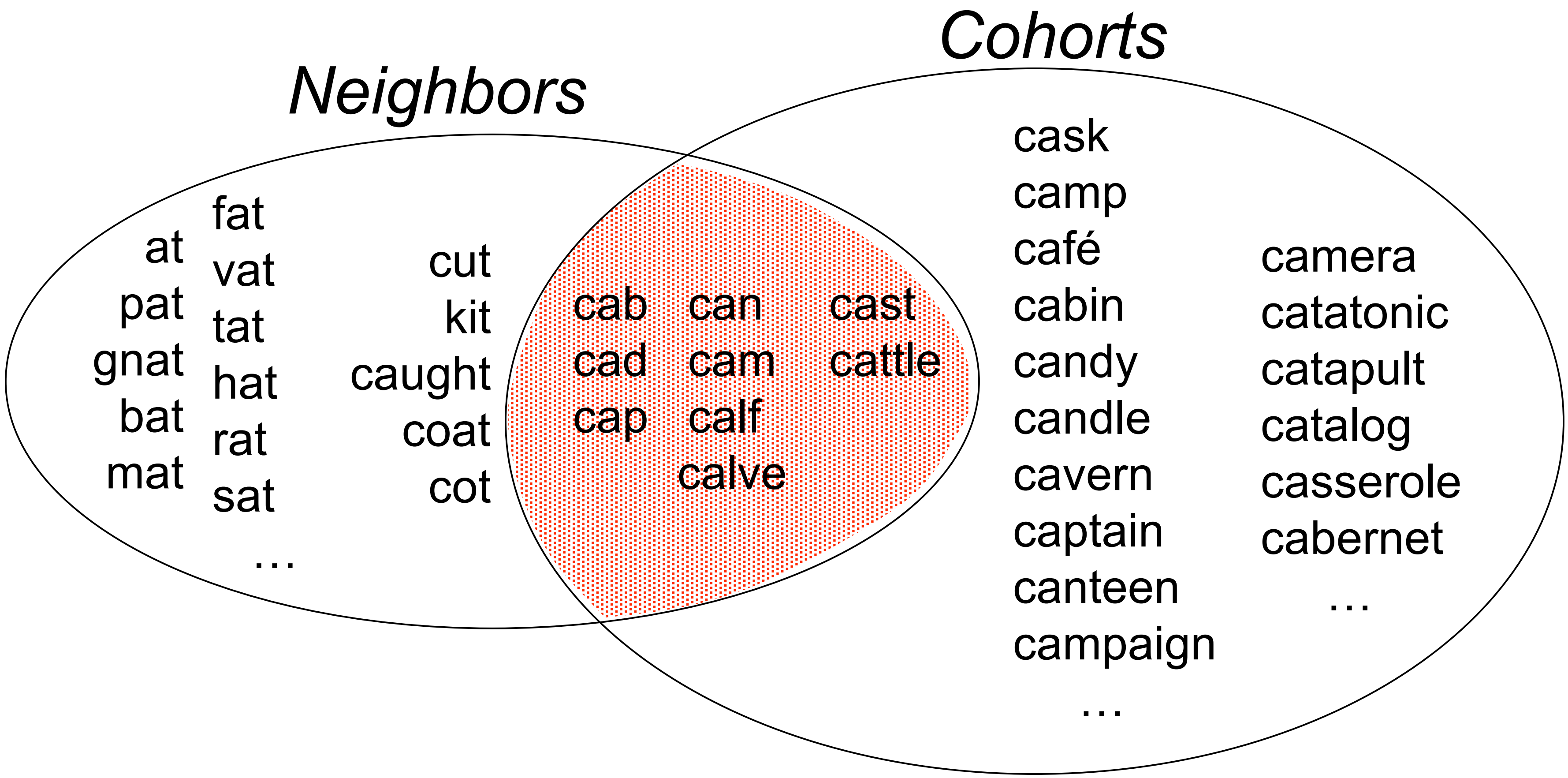
- Mathematical model
  - Described as a processing model, but most significant contribution: simple, concise encapsulation of theoretical assumptions
  - Does not address segmentation/alignment

1. Operationalize *neighbor* (1-phoneme shortcut, segment-by-segment similarity)
2. Recognition facility (frequency weighted neighborhood probability) $\approx \dfrac{f_t}{\sum f_n}$

- Evidence: FWNPR accounts for more variance than any other factor!

Competitor sets

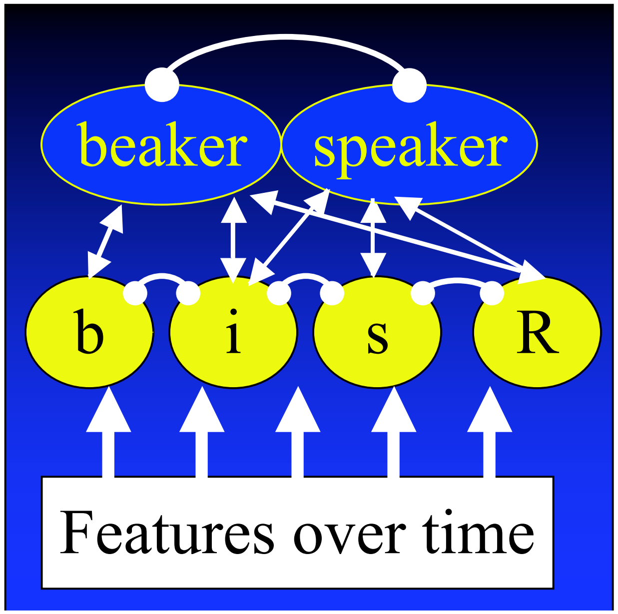
Example: cat



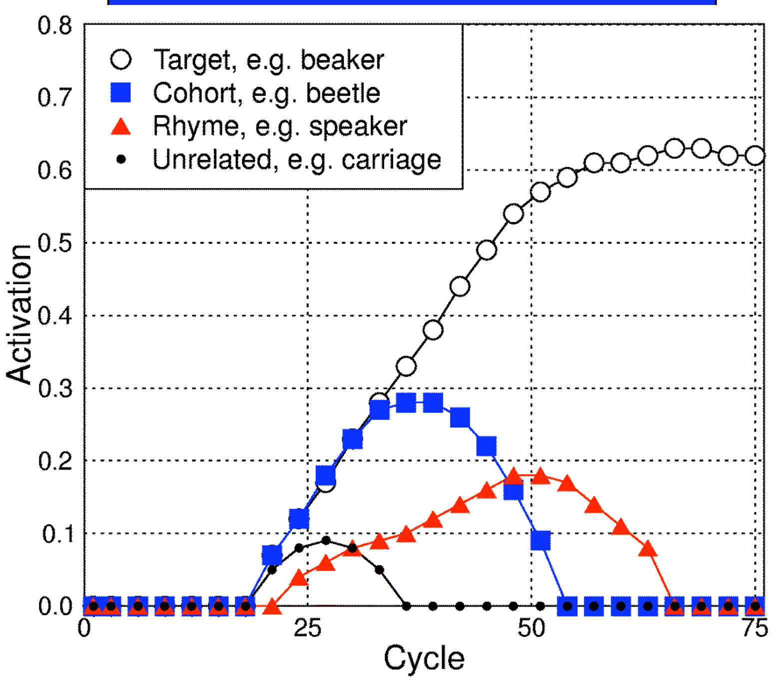
Simulating

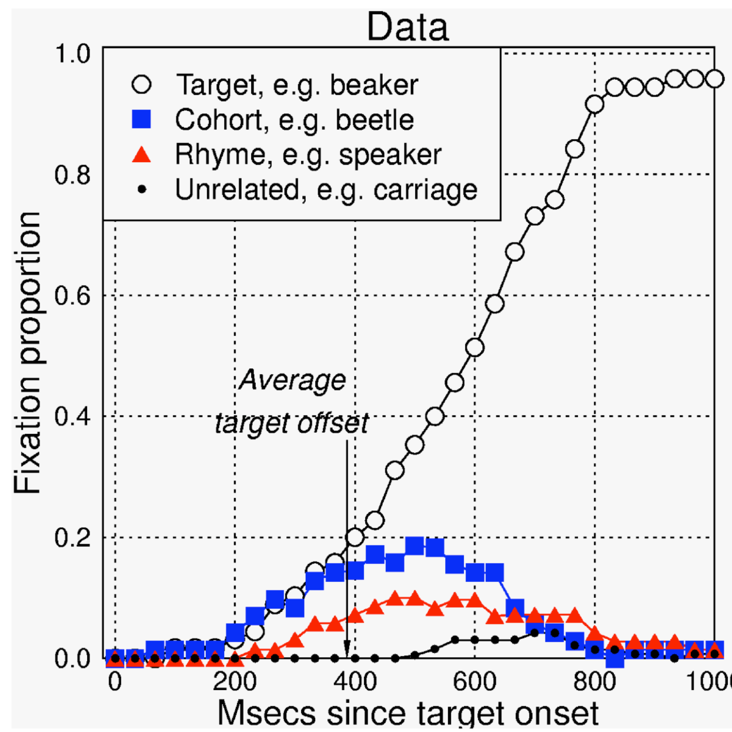
- To simulate, you must grapple with the practical implications of theoretical assumptions
- Also, many other details
  - How to make input analogous to speech
  - How to map model time to real time
  - How to link model performance to human task
  - How to gauge model success and failure

Simulating: TRACE (McClelland & Elman, 1986)



Allopenna, Magnuson, & Tanenhaus (1998): Human eye tracking data highly similar to TRACE predictions





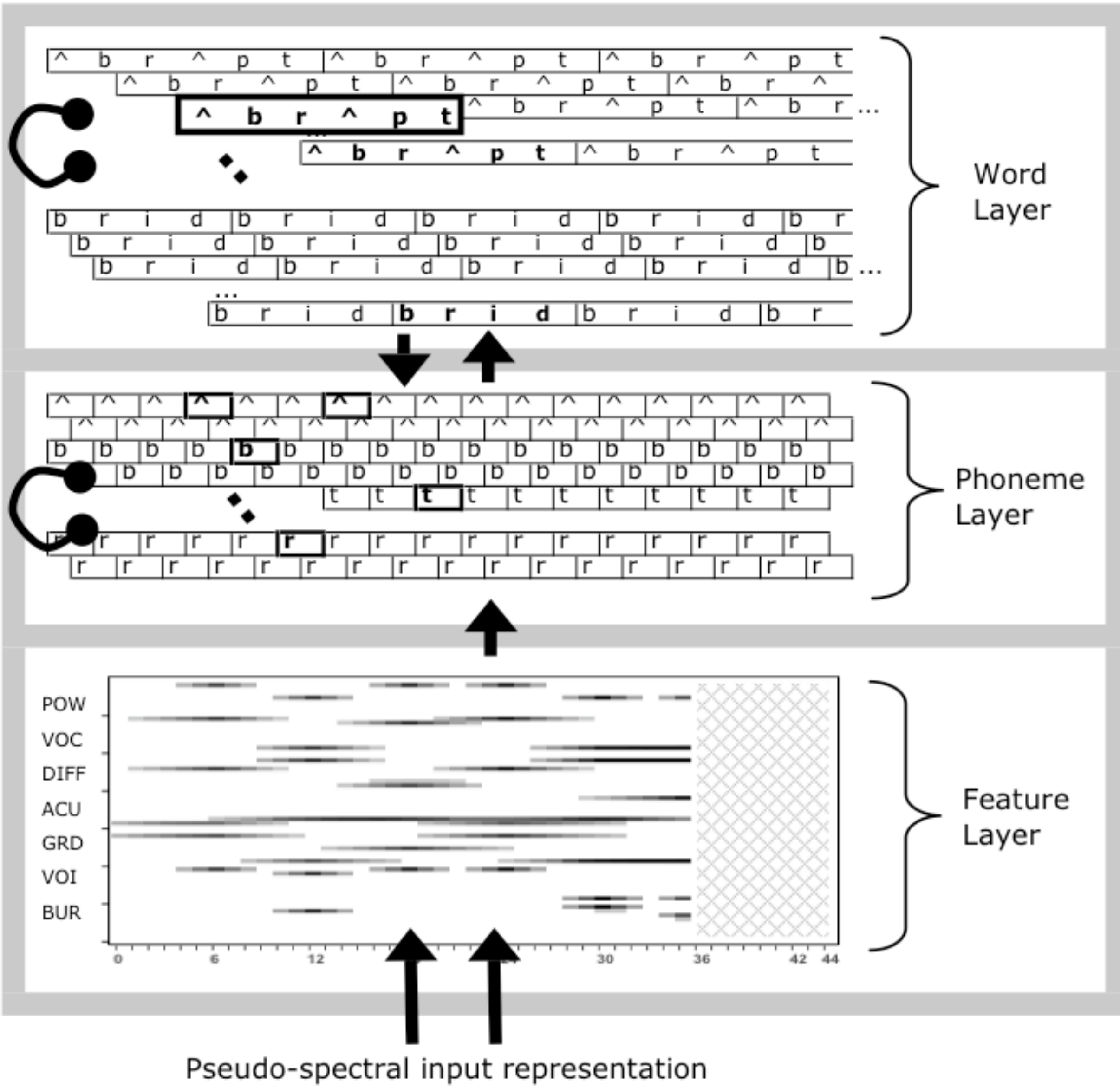
Why use TRACE?

- Excellent predictions for broad range of phenomena
- Representative of current models
  - Dynamics/time course
  - Embodies 3 key principles (multiple activation; activation proportional to similarity & prior probability; competition)
- Relatively transparent parameters
- Shortcomings
  - Brute force approach to solving alignment
  - Does not learn

Introduction to TRACE

- Architecture
- Connectivity
- Flow of activation
- Input representation
- Processing time vs. slice time

TRACE architecture: Connectivity



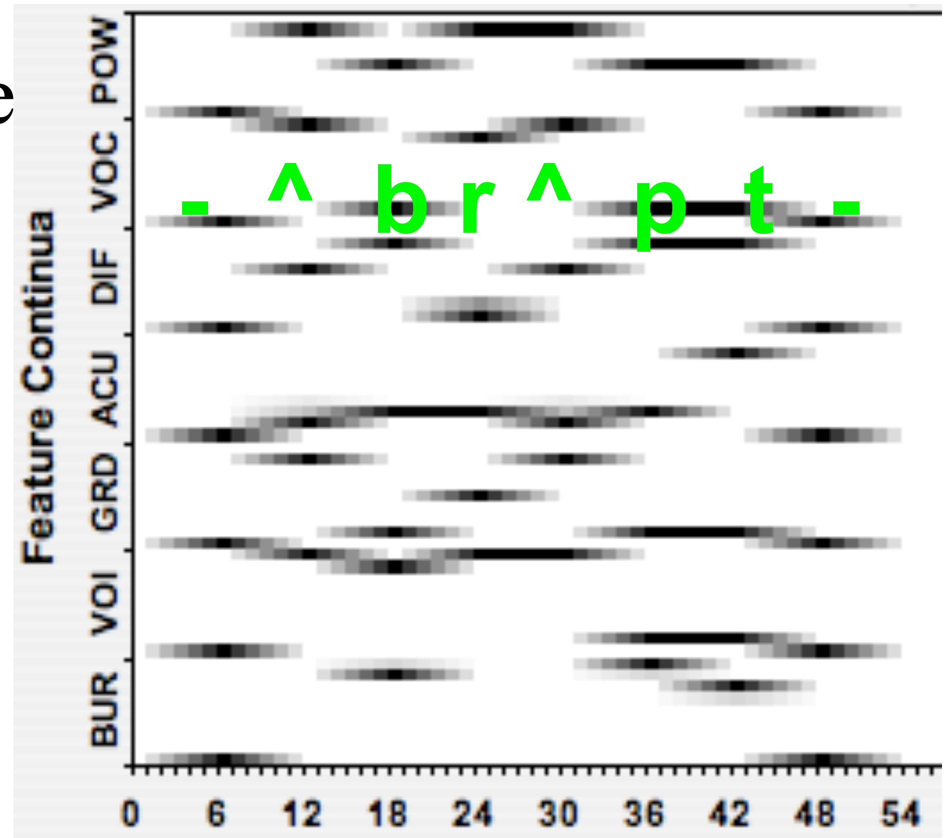
- Bottom-up (feed-forward) excitatory connections
  - Input→feature
  - Feature→phoneme
  - Phoneme→word
- Top-down (feed-back) excitatory connections
  - Word→phoneme
  - Phoneme→feature*
- Lateral inhibition within layers

TRACE input representation

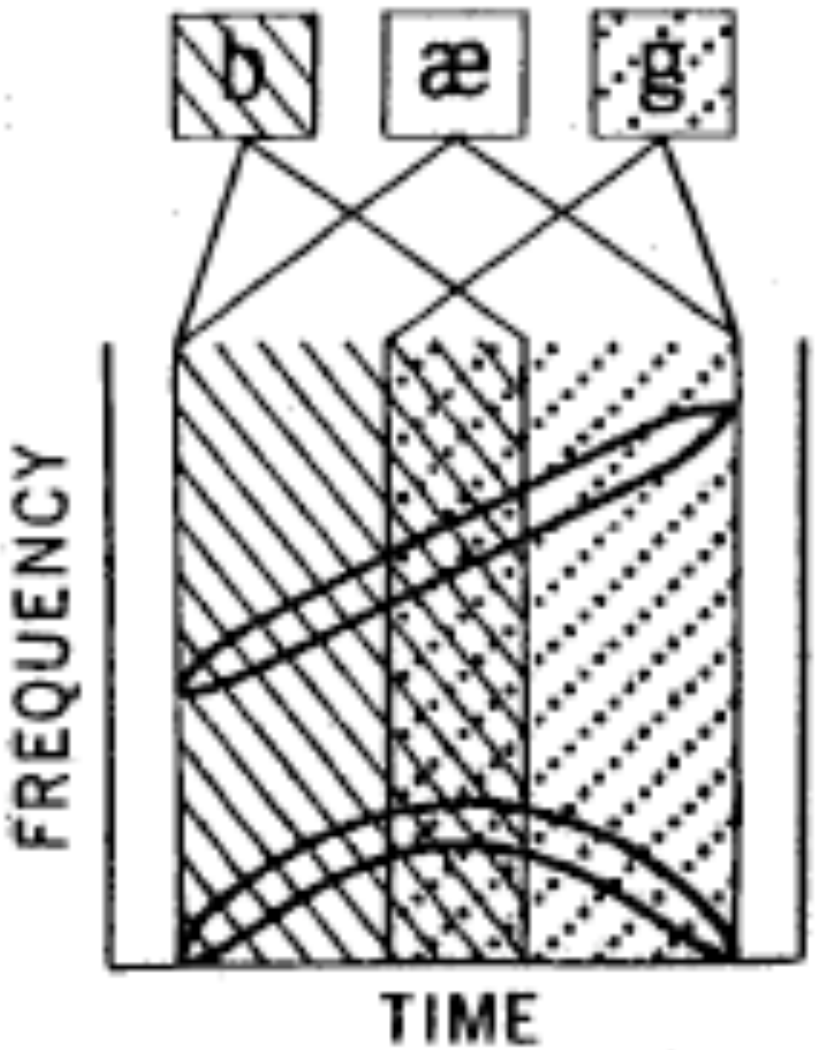
- Designed to approximate several important facts about the speech signal and speech perception
  - Perceptual similarity rooted in acoustic similarity rooted in event (production) similarity
  - Speech signal is extended over time
  - Speech sounds (phonemes) overlap in time from one to the next

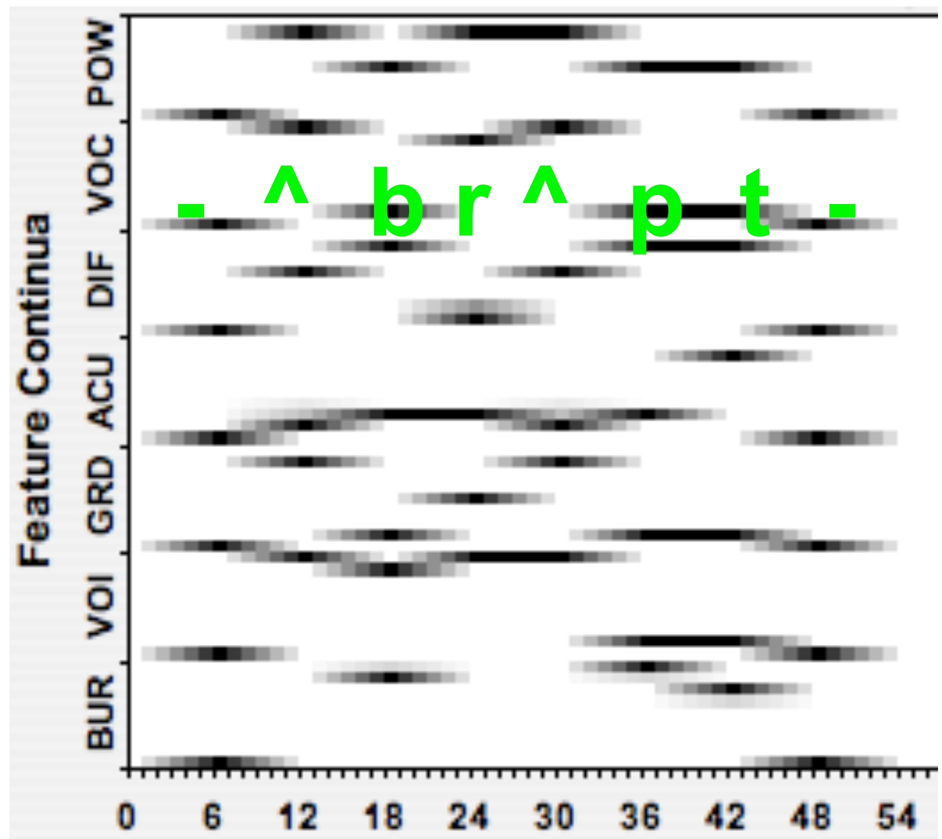
TRACE: Input representation

- The input to TRACE is a matrix of 7 feature vectors with 9 levels each
- Features are based on acoustic-phonetic features
  - consonantal, vocalic, diffuseness, acuteness, voicing, burst, power



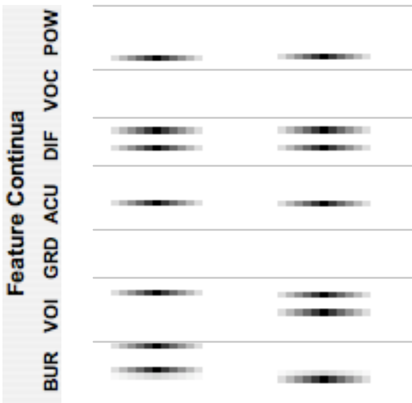
Coarticulation

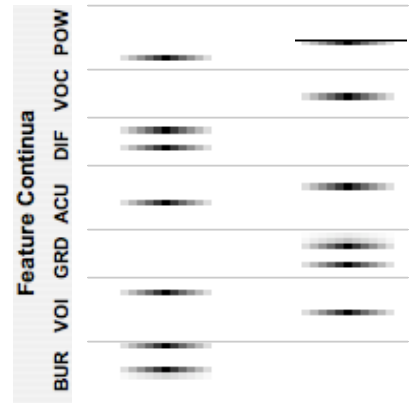


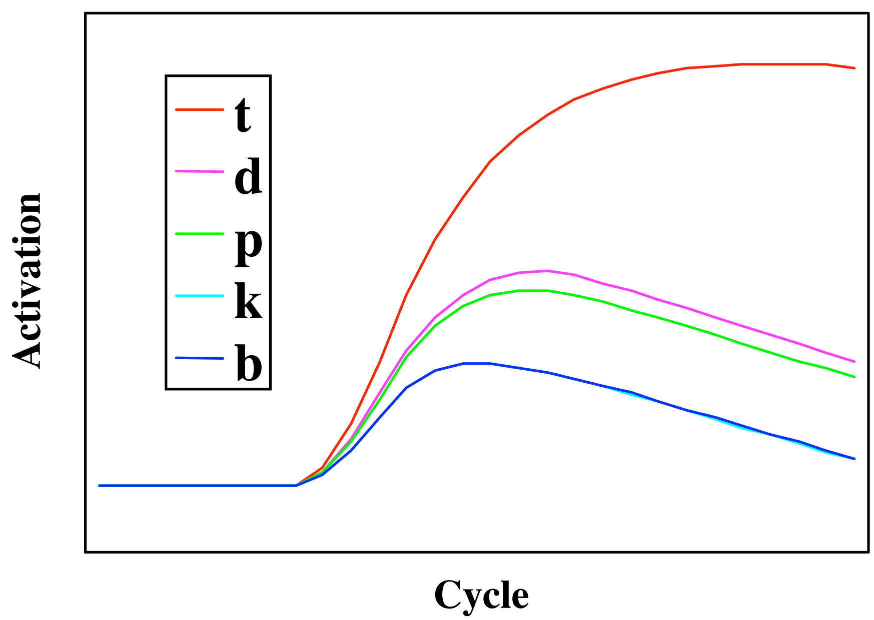


Similarity

/t/ ↔ /d/ vs. /t/ ↔ /a/

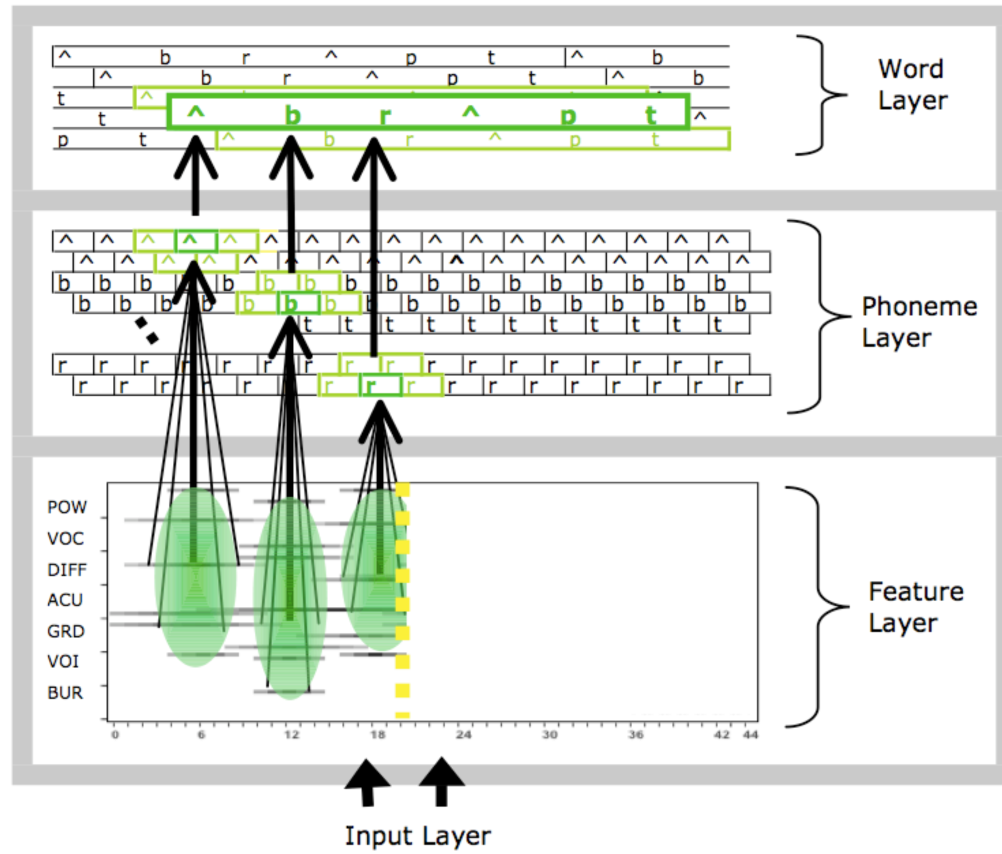






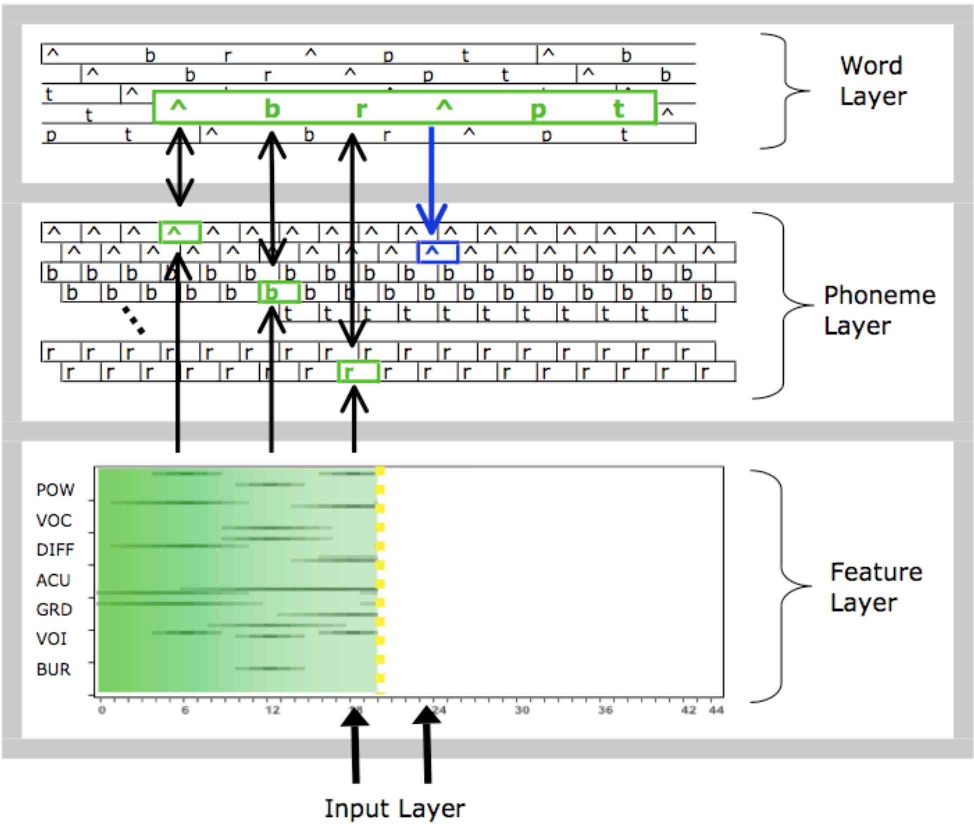
Activation of phonemes & words

- Bottom-up (**feed-forward**) excitation causes initial activation
- Excitation increases with temporal overlap and similarity

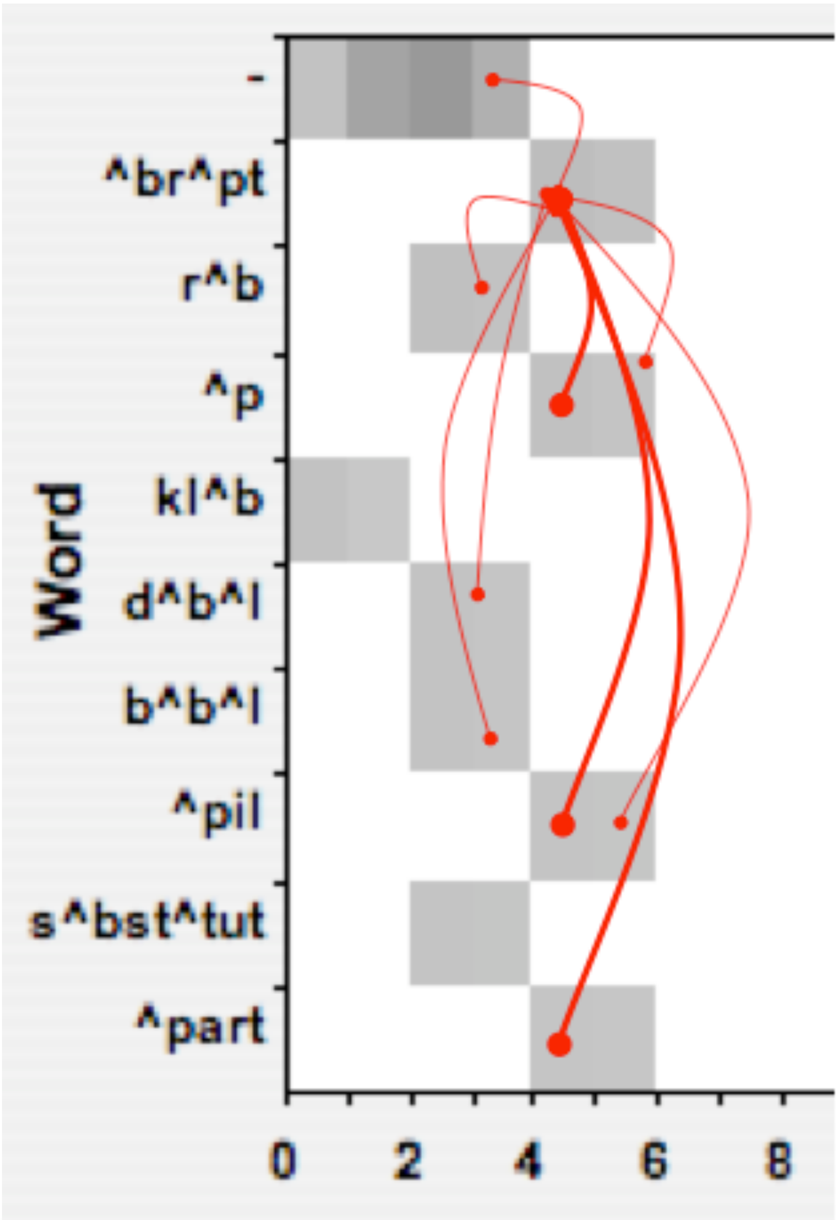


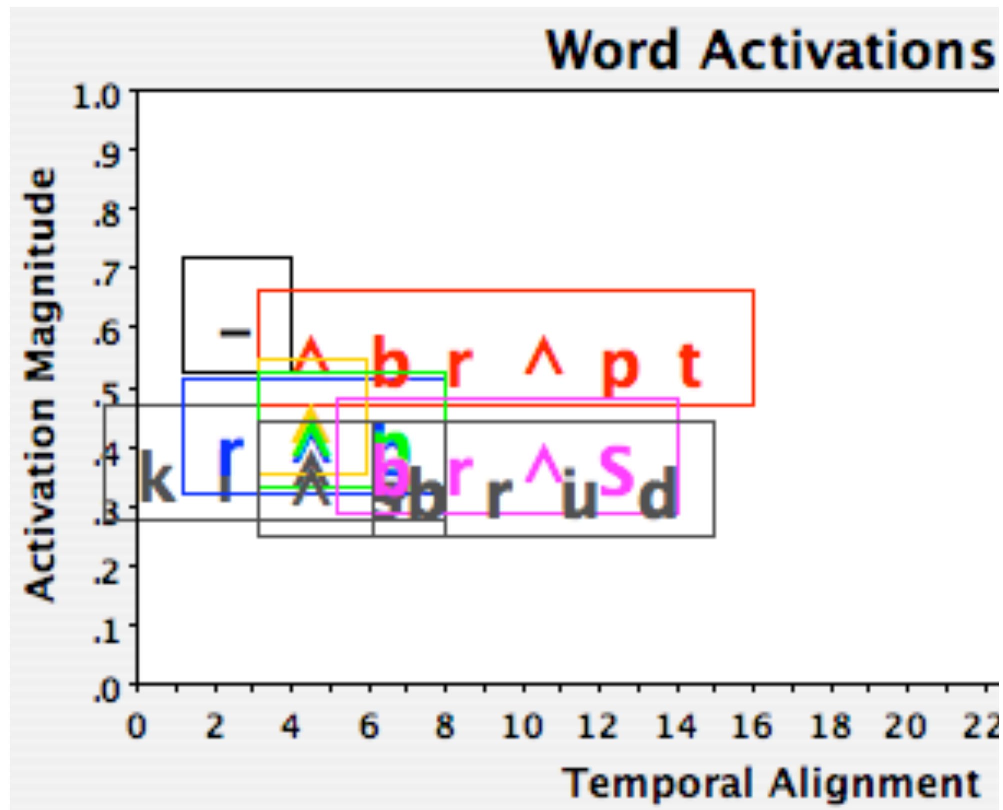
Lexical-to-phoneme feedback

- Word units feed back to constituent phonemes
- Constituent phonemes that were not “heard” (earlier or later) via lexical **feedback**



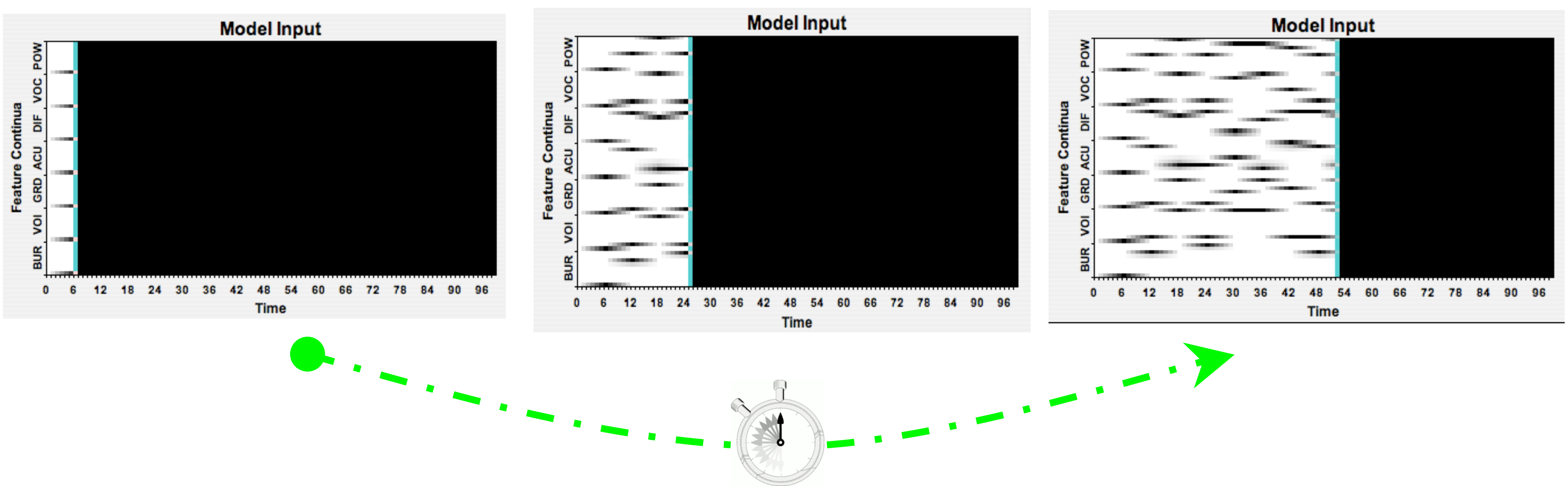
Within-layer competition (lateral inhibition)





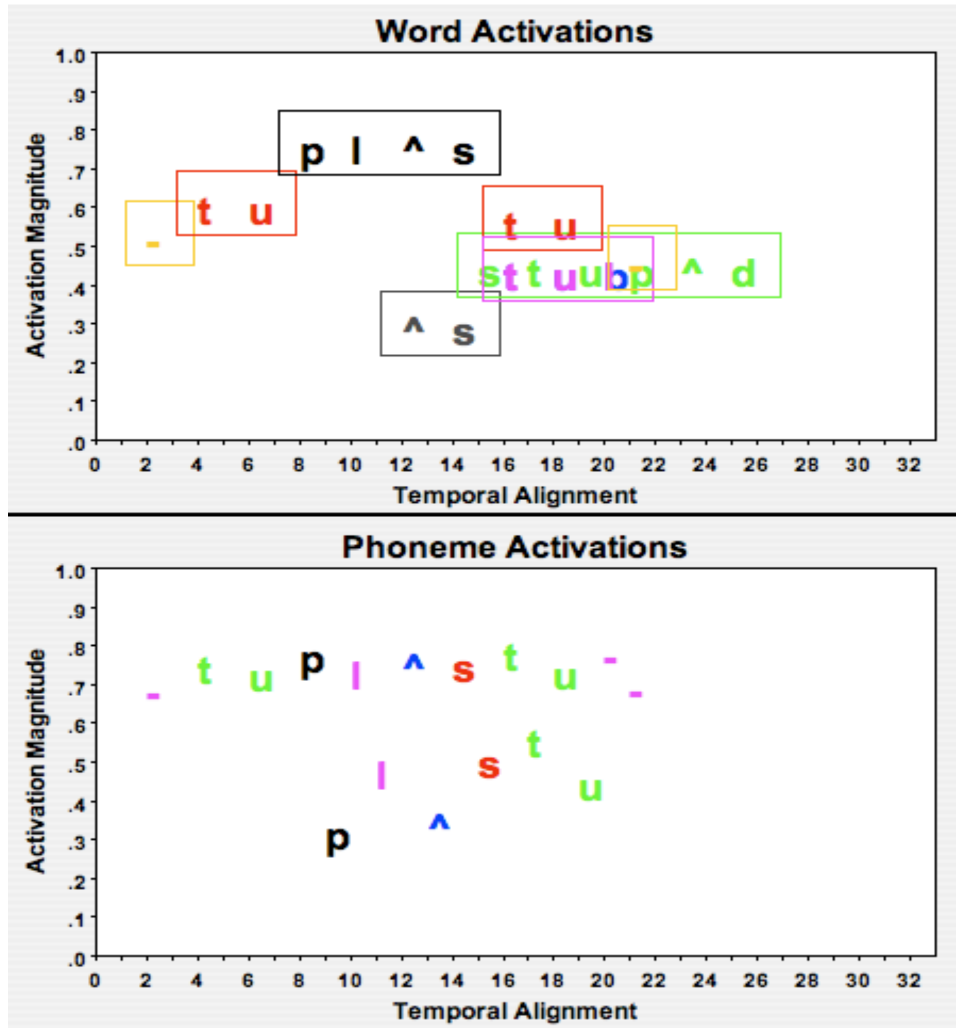
Two kinds of *time*

1. "Real" time (cycles): **passing** of time, during which speech input is being presented to the model continuously



Two kinds of *time*

2. Temporal **alignment** of units; “slice” number



THE NEW SCHOOL

Haskins Laboratories

Next: Module 2, tour of jTRACE