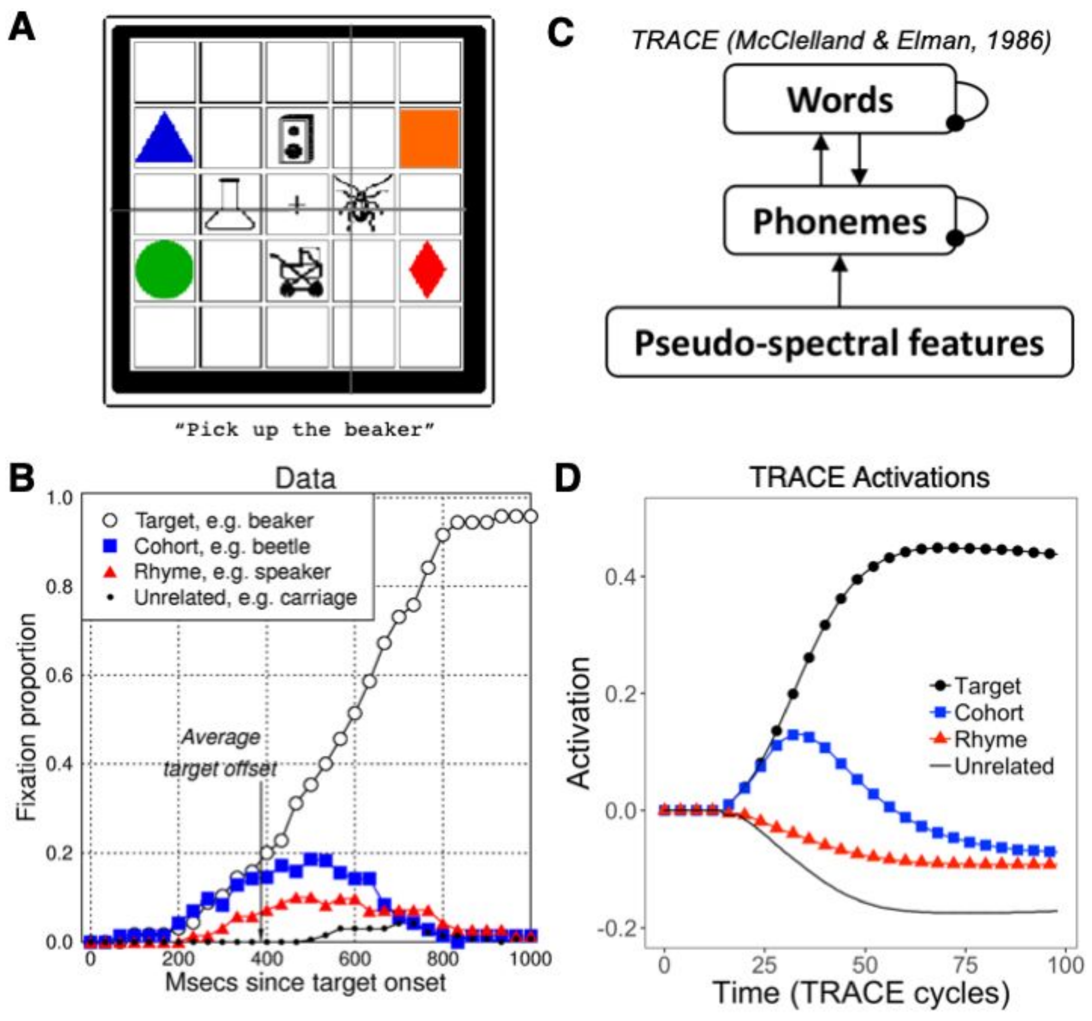
EARSHOT: *Emulating Auditory Recognition of Speech by Humans Over Time*

James Magnuson,[1] Heejo You,[1] Jay Rueckl,[1,2] Paul Allopenna,[1] Monica Li,[1] Sahil Luthra,[1] Rachael Steiner,[1] Monty Escabi,[3] Kevin Brown,[4] Rachel Theodore,[5] & Nicholas Monto[6]

[1]UConn Psychological Sciences, [2]Haskins Labs, [3]UConn Biomedical Engineering, [4]Oregon State, [5]UConn Speech, Language & Hearing Sciences

Crucial behavioral target

Human timecourse of lexical access and phonological competition

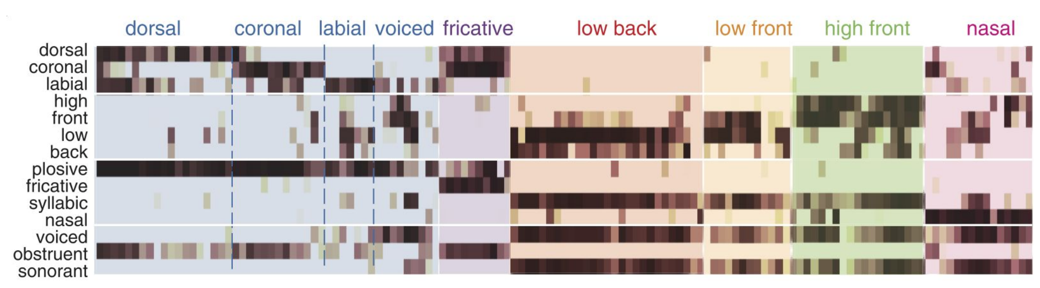


Dynamics of phonological competition in humans and models. Allopenna, Magnuson, & Tanenhaus (1998) asked listeners to follow simple spoken instructions to interact with simple displays (e.g., panel A), and tracked their eye movements as they did so (panel B). Fixation proportions over time were hypothesized to relate to internal lexical activation and competition. They used the gold-standard TRACE model (C; McClelland & Elman, 1986) to simulate their paradigm. Panel D: model activations clearly resemble human performance (simulations of all words in the TRACE lexicon, with mean activations for targets, cohorts [onset competitors, e.g., CAT-CAB-CANDLE], rhymes [CAT-BAT; CANDLE-HANDLE], and unrelated items).

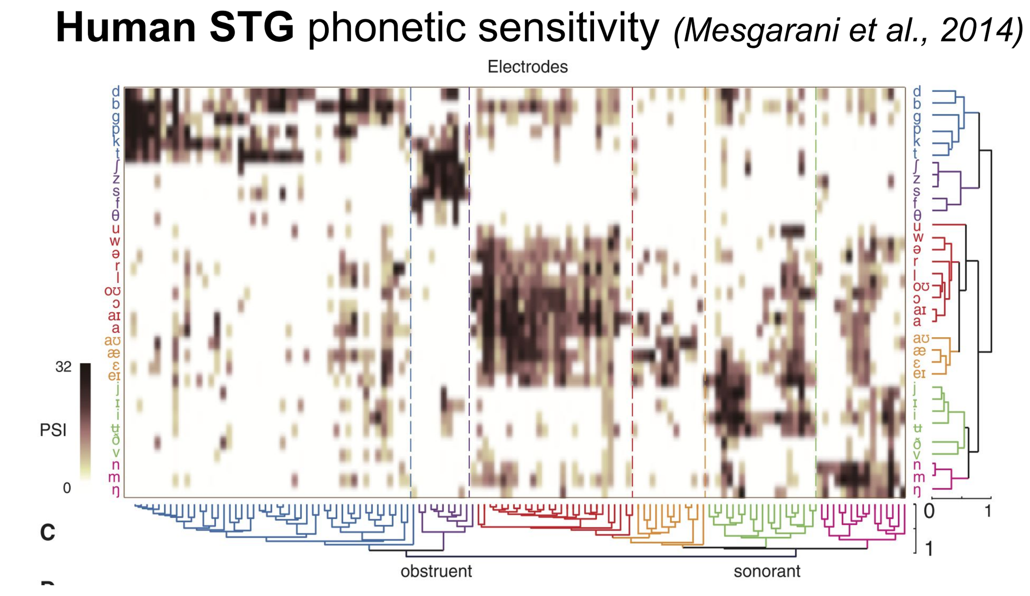
Potential neurobiological target

Phonetically-structured responses in human superior temporal gyrus (STG)

Human STG feature sensitivity *(Mesgarani et al., 2014)*



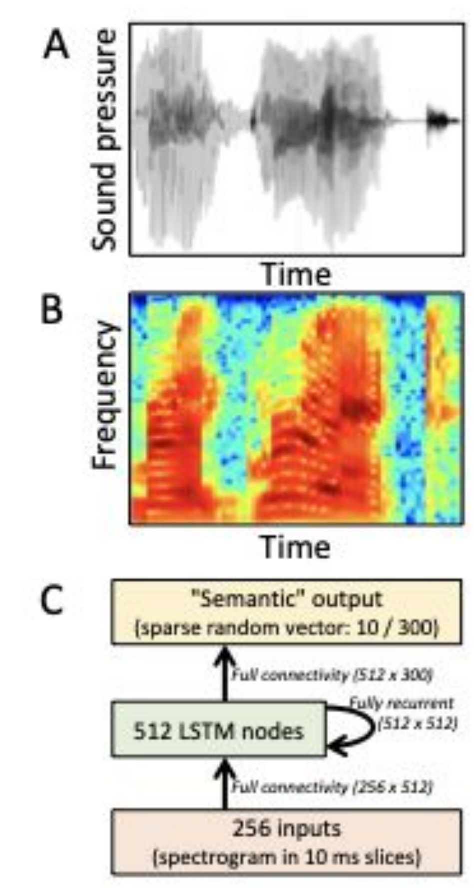
Human STG phonetic sensitivity *(Mesgarani et al., 2014)*



Electrocorticography results from:

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*, 1006-1010.

EARSHOT input & architecture

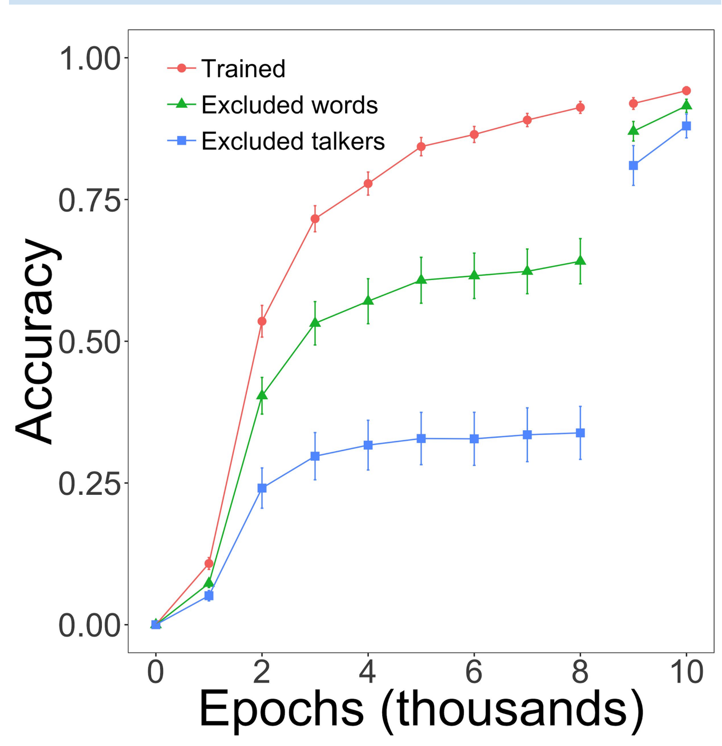


Model input and structure. (A) Audio files converted to 256-channel spectrograms **(B)**, w/10 ms steps. The model (C) is a standard recurrent network, except "long short-term memory" nodes are used in the hidden layer, allowing it to become sensitive to multiple temporal grains. Output targets are pseudo-semantics (sparse random vectors, a common simplification).

Training

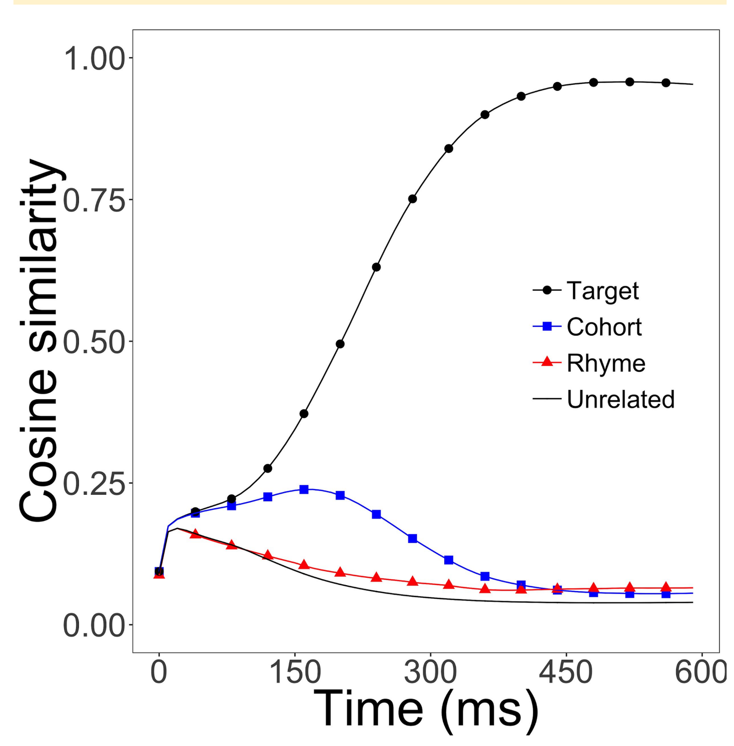
- Given *T* talkers and *W* words:
  - Train *T* models with *T*-1 talkers (leave one out for generalization)
  - For remaining talkers, exclude different subsets of [*W* / (*T* - 1)] words for generalization
- First simulations: 10 talkers, 1000 words, 1-3 syllables (1-8 phonemes)
- Talkers: 10 voices from Apple's *say* text-to-speech app (5 male, 5 female)

High accuracy, moderate generalization



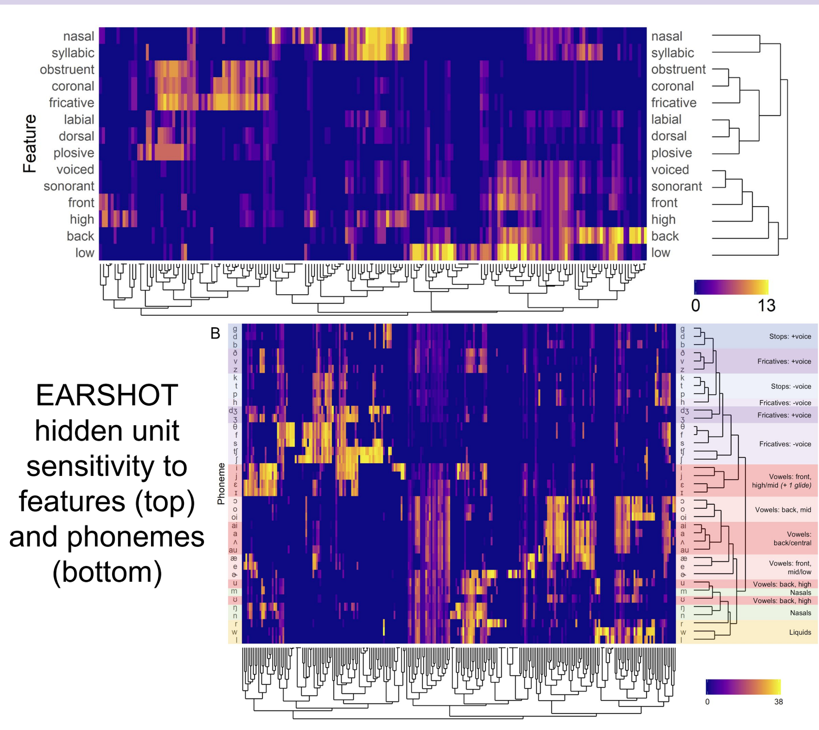
Accuracy by epoch averaged over ten models. When training resumed with all items (epochs 8001-10,000), high accuracy was achieved quickly for all talkers and excluded words.

Timecourse resembles human & TRACE pattern



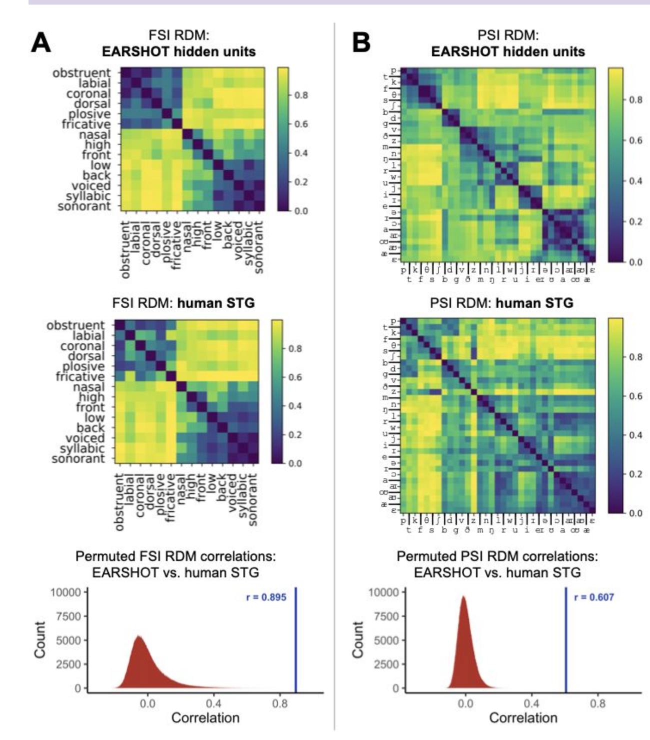
Timecourse of competition for accurate trials at epoch 8000, for cohorts and rhymes, closely resembling human performance (Allopenna et al., 1998) and TRACE simulations shown above.

Hidden unit responses reveal phonetic organization similar to that in human STG - *w/o phonetic training!*



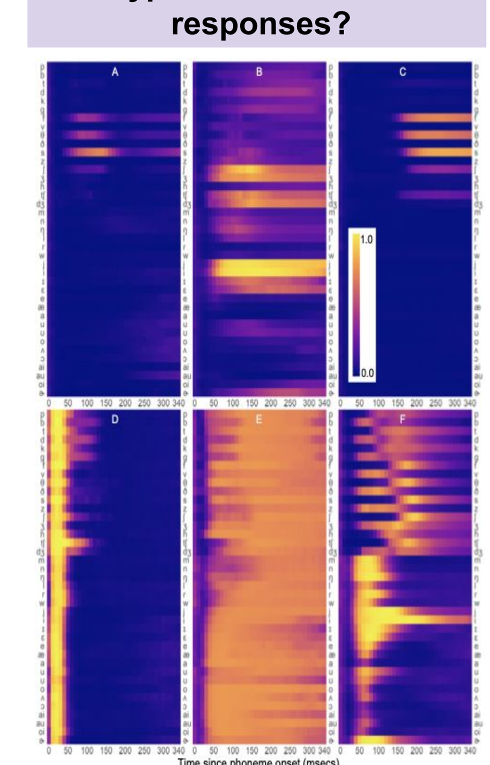
Phonetic sensitivity revealed by hierarchical clustering. TOP: *Featural Sensitivity Index (FSI, top)* based on hidden unit (x-axis) responses to phonetic features; for every hidden unit-feature pair, FSI was incremented for every feature to which the hidden unit responded substantially more weakly (yellow indicates high selectivity, with maximum FSI of 13, given 14 features). **BOTTOM**: *Phonetic Sensitivity Index (PSI)*. High PSI indicates selective responses to specific phonemes. Max = 38 (given 39 phonemes).

EARSHOT and STG responses are highly similar, indicating similar sensitivity to signal information, not necessarily homology



Representational similarity analyses quantifying similarity of EARSHOT and human neural responses. Panel A (left column) shows *Representational Dissimilarity Matrices* (RDMs) for featural sensitivity indices (FSIs) for EARSHOT and human STG ECoG data from Mesgarani et al. (2014). RDMs: calculate *dissimilarity* between vectors of FSIs for each hidden unit or electrode for each feature (low values=high similarity). Bottom left: correlation between FSI RDMs ($p < 1 \times 10^{-6}$), and correlation distribution when 1 was shuffled randomly (1M permutations); actual correlation (blue line) far from chance range. Panel B: RDMs for phonetic sensitivity indices (PSIs); again, RDM correlation is high ($p < 1 \times 10^{-6}$).

Variable hidden unit response profiles: new hypotheses for STG responses?

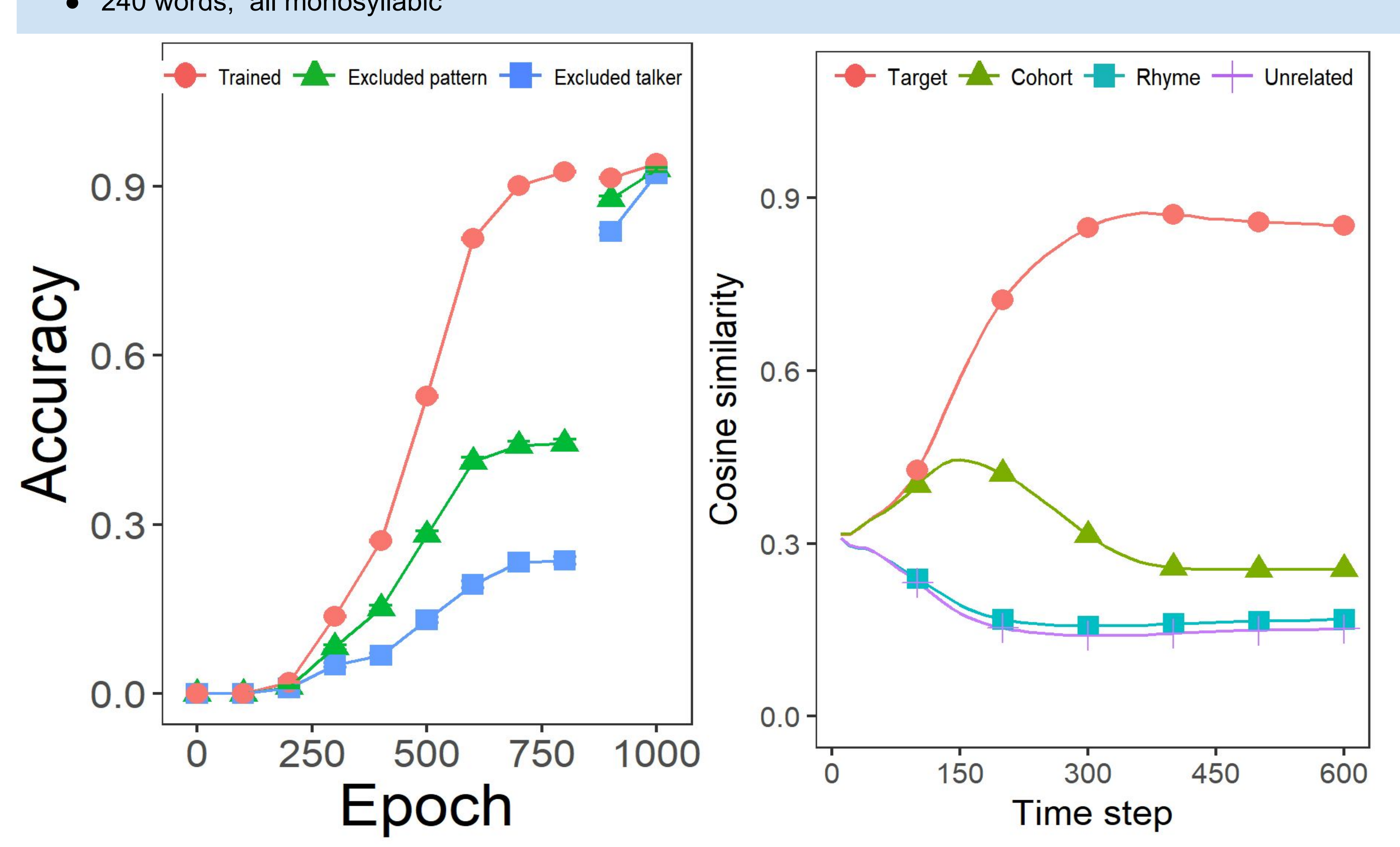


Hidden unit response profiles. Over-time response profiles of example hidden units for each phoneme (y-axis). **(A)** Time locked, discrete responses (~5% of units). **(B)** Time locked, sustained responses (~20%). **(C)** Delayed responses (~35%). **(D)** Early onset responses (~4%). **(E)** Post-onset inactivation (~3%). **(F)** Complex responses (~29% of HUs). *Additional ~4% are largely non-responsive.*

Expanding to 17 talkers, 240 words: faster learning

- 15 synthetic [10 by rule, 5 concatenative]
- 2 human
- 240 words, all monosyllabic

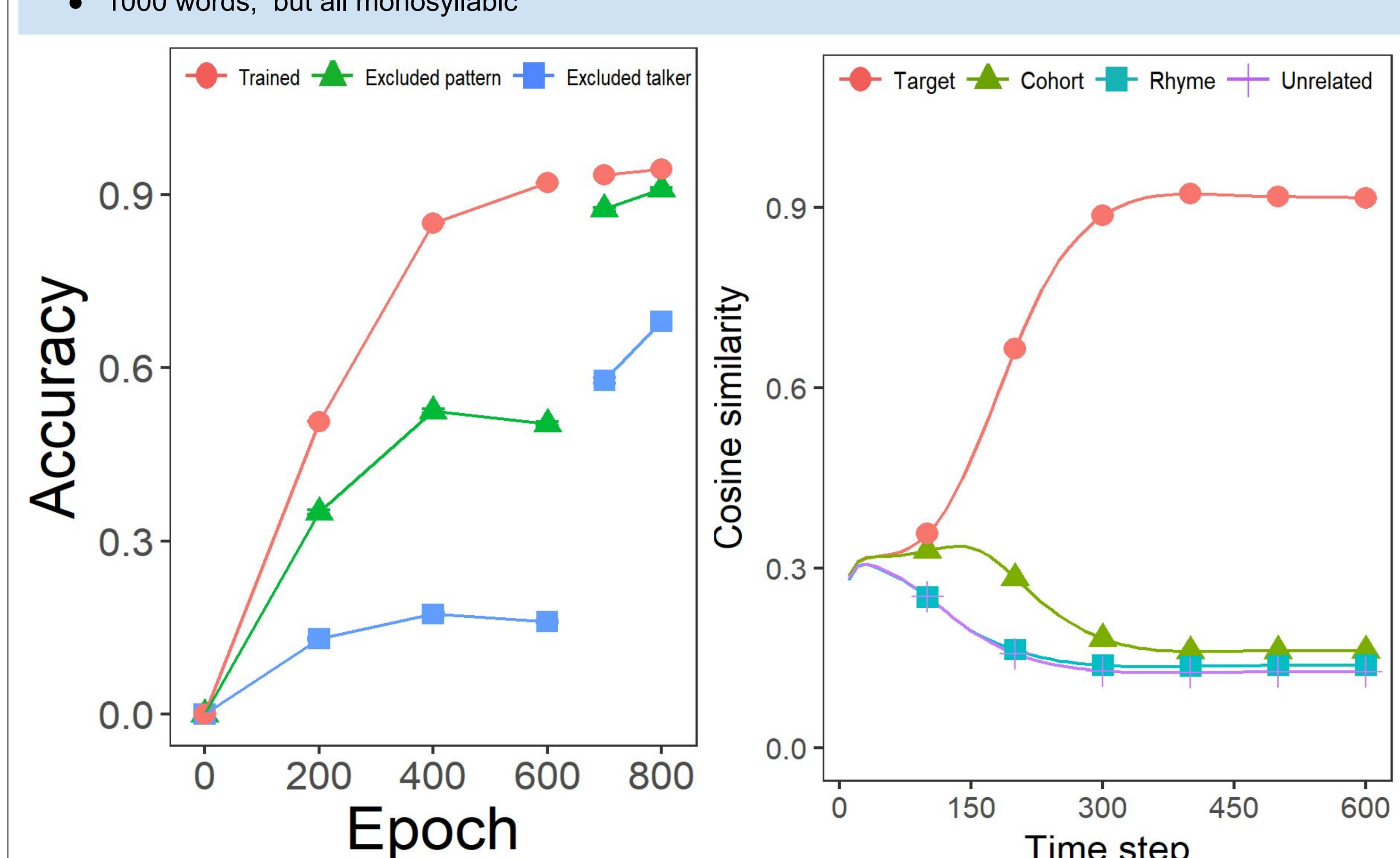
Is learning faster because of more talkers, smaller lexicon, or monosyllables?



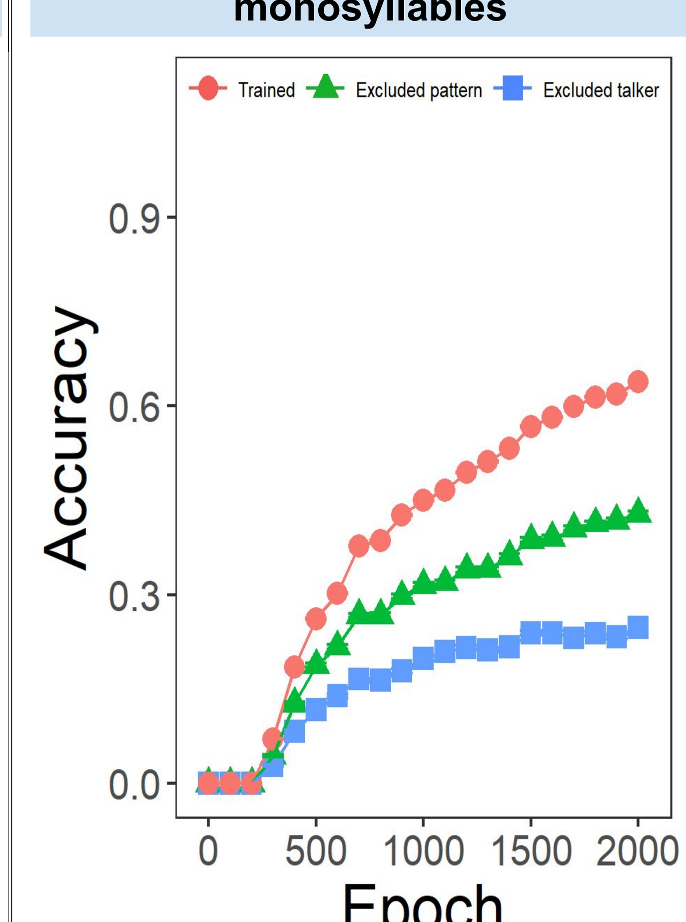
Expanding to 17 talkers, 1000 words: faster learning

- 15 synthetic [10 by rule, 5 concatenative]
- 2 human
- 1000 words, but all monosyllabic

Does not resolve whether learning is faster because of more talkers or monosyllables



15 synthetic talkers, original 1000-word lexicon: faster learning was due to monosyllables



EARSHOT is the first model of **human** speech recognition **that can recognize real speech** from multiple talkers with high accuracy (1000 words x 10 talkers)

EARSHOT is **simple enough** to provide a new platform for testing theories of human speech recognition

EARSHOT is not trained on phonemes, but develops **internal representations** that **resemble cortical responses to phonemes**

This project is supported by U.S. National Science Foundation grants 1754284 *Computational approaches to human spoken word recognition* (JSM, PI), IGERT 1144399 *Language plasticity* (JSM, PI), and NRT 1747486 *Science of learning & art of communication* (JSM, PI), and by the Connecticut Institute for the Brain and Cognitive Sciences, and the UConn Office of the Vice President for Research.