COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12823

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12823

EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition

James S. Magnuson,[a,b] Heejo You,[a,b] Sahil Luthra,[a,b] Monica Li,[a,b,c] Hosung Nam,[c,d] Monty Escabí,[a,b,e,f] Kevin Brown,[g] Paul D. Allopenna,[a,b] Rachel M. Theodore,[a,h] Nicholas Monto,[a,h] Jay G. Rueckl[a,b,c]

[a]*Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut*
[b]*Psychological Sciences, University of Connecticut*
[c]*Haskins Laboratories*
[d]*Department of English Language and Literature, Korea University*
[e]*Electrical and Computer Engineering, University of Connecticut*
[f]*Biomedical Engineering, University of Connecticut*
[g]*Departments of Pharmaceutical Sciences and Chemical, Biological, and Environmental Engineering, Oregon State University*
[h]*Speech, Language, and Hearing Sciences, University of Connecticut*

Received 27 August 2019; received in revised form 11 December 2019; accepted 5 February 2020

Abstract

Despite the *lack of invariance problem* (the many-to-many mapping between acoustics and percepts), human listeners experience phonetic constancy and typically perceive what a speaker intends. Most models of human speech recognition (HSR) have side-stepped this problem, working with abstract, idealized inputs and deferring the challenge of working with real speech. In contrast, carefully engineered deep learning networks allow robust, real-world automatic speech recognition (ASR). However, the complexities of deep learning architectures and training regimens make it difficult to use them to provide direct insights into mechanisms that may support HSR. In this brief article, we report preliminary results from a two-layer network that borrows one element from ASR, *long short-term memory* nodes, which provide dynamic memory for a range of temporal spans. This allows the model to learn to map real speech from multiple talkers to semantic targets with high accuracy, with human-like time-course of lexical access and phonological competition. Internal representations emerge that resemble phonetically organized responses in human superior temporal gyrus, suggesting that the model develops a distributed phonological code despite no explicit training on phonetic or phonemic targets. The ability to work with real speech is a major advance for cognitive models of HSR.

Keywords: Human speech recognition; Computational modeling; Neurobiology of language

Correspondence should be sent to James S. Magnuson, Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT. E-mail: james.magnuson@uconn.edu

1. Introduction

Phonetic constancy in human speech recognition (HSR) poses a significant theoretical challenge for the cognitive and neural sciences. Despite a *lack of invariance* (a many-to-many mapping between speech acoustics and linguistic percepts such as consonants, vowels, syllables, and words), listeners achieve *phonetic constancy*, (usually) perceiving a speaker's intended message with apparent ease. The acoustic patterns specifying different phonemes overlap in time (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), with few boundaries between phonemes or words (Cole & Jakimik, 1980), and shift with factors such as speaking rate (Miller & Baer, 1983), talker characteristics (Joos, 1948; Peterson & Barney, 1952), phonetic context (Liberman et al., 1967), coarticulation (Liberman, Delattre, & Cooper, 1952), and novelty of message content (Fowler & Housum, 1987). Although similar problems exist in other domains (e.g., robust visual object recognition over variation in size, rotation, and illumination; DiCarlo & Cox, 2007), the temporal nature of speech exacerbates the challenge; the elements of a spoken word are a series of overlapping events that do not persist in the environment (unlike a visual object or written word, which can be resampled, with all elements simultaneously and persistently present).

Deep-learning neural network models underlying automatic speech recognition (ASR) provide robust real-world computer speech recognition for billions of users (Hinton et al., 2012). As Kietzmann, McClure, and Kriegeskorte (2019) have argued, deep networks can guide theoretical understanding to the degree that they can predict real-world behavior and/or neural activity. On our view, current deep networks for speech recognition offer little guidance to theories of HSR. They have many layers of richly connected nodes and require carefully engineered training regimens that are not (typically) constrained by biological considerations. Bridging the gap between deep networks that provide robust ASR and cognitive theories of HSR will require the development of models that progressively span the current divide.

Scientists have used less complex deep networks to investigate mechanisms that might support audition and speech. For example, hidden units of a five-layer network trained explicitly on phoneme recognition (Nagamine, Seltzer, & Mesgarani, 2015) exhibited phonetically organized responses similar to those observed in human superior temporal gyrus (Mesgarani, Cheung, Johnson, & Chang, 2014). Another (Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018) achieved human-like accuracy on one speech task (identifying the word at the center of a 2-s speech sample) and one music task (genre identification), with many layers and a complex training regimen. This network consisted of seven initial layers shared across the two tasks, which then branched into two separate five-layer pathways, one for each task. The model was better at predicting human fMRI responses to natural sounds than a standard spectrotemporal filter model of auditory cortex. Kell et al. suggested that deep learning may be the only computational approach capable of human-like performance in perceptual domains. However, this approach has three important limitations: (a) unpacking emergent mechanisms in many-layered networks and linking them to theories of human capacities is a formidable challenge; (b) many deep

learning approaches to auditory processing do not take over-time input (speech is often input like an image, as though an entire utterance occurred instantaneously rather than over time); and crucially, (c) these models have not been applied to the complex timecourse of human lexical activation and competition (a primary focus of cognitive theories; see Fig. 1).

Simpler models (e.g., McClelland & Elman, 1986) have guided theories of the timecourse of HSR for decades, but they have two different limitations. First, they do not use real speech as input; since the 1970s, most modeling of HSR has adopted the simplifying assumption that speech perception provides something like phonemic input to processes for word recognition, and abstractions from real speech are used as inputs (such as phonetic features spread over time [as in TRACE; McClelland & Elman, 1986], or human diphone confusions [as in Shortlist B; Norris & McQueen, 2008]). Neurally inspired modeling of human speech perception continued (Grossberg, Boardman, & Cohen, 1997), but in small-inventory models rather than large vocabulary, signal-to-word models. Recent attempts at linking automatic speech recognition approaches to cognitive models (Scharenborg, 2010; Scharenborg, Norris, ten Bosch, & McQueen, 2005) generated interesting insights, but with low accuracy and limited empirical coverage. Second, they set aside the problem of learning, using fixed parameters in neural network (McClelland & Elman, 1986) or Bayesian approaches (Norris & McQueen, 2008). Nonetheless, these models simulate the fine-grained timecourse of lexical activation and phonological competition (Allopenna, Magnuson, & Tanenhaus, 1998; see Fig. 1) and have significantly advanced theories and understanding of HSR dynamics.

The persistence of these two "temporary" simplifying assumptions decades later reflects the tension between *computational adequacy* (maximizing task realism and performance) and *psychological adequacy* (capturing key details of human behavior, while providing an understandable *account* of how the model works; McClelland & Elman, 1986). A model with high computational adequacy but opaque mechanisms offers little guidance to theories that seek to provide mechanistic explanations of HSR at a fine grain. Thus, it is imperative that we address the gaps of *signal* and *learning* via models sufficiently simple that they can guide cognitive theory.

Our goal is to develop a *minimal* (and thus more readily analyzable and understandable) cognitive model of HSR that can learn to map over-time speech to semantics, without explicit phonetic training as a first step in bridging the gap between HSR and ASR (ultimately through progressively more complex and realistic models). A minimal model could reveal representations that emerge in a simple learning system, providing hypotheses for cognitive and neural mechanisms supporting HSR. After exploring approximately three dozen models (varying in input, hidden unit types, and architecture; see Appendix S1), we achieved human-like performance with a shallow (two-layer) network equipped with *long short-term memory* (LSTM) hidden nodes (Hochreiter & Schmidhuber, 1997). LSTM nodes add three internal gates and a memory cell that allow nodes to develop sensitivity to information over varied ranges of short and long time scales, mitigating the *vanishing gradient problem* (Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001): Recurrent networks (with connections between nodes within a layer or from a superior to an inferior layer) can theoretically

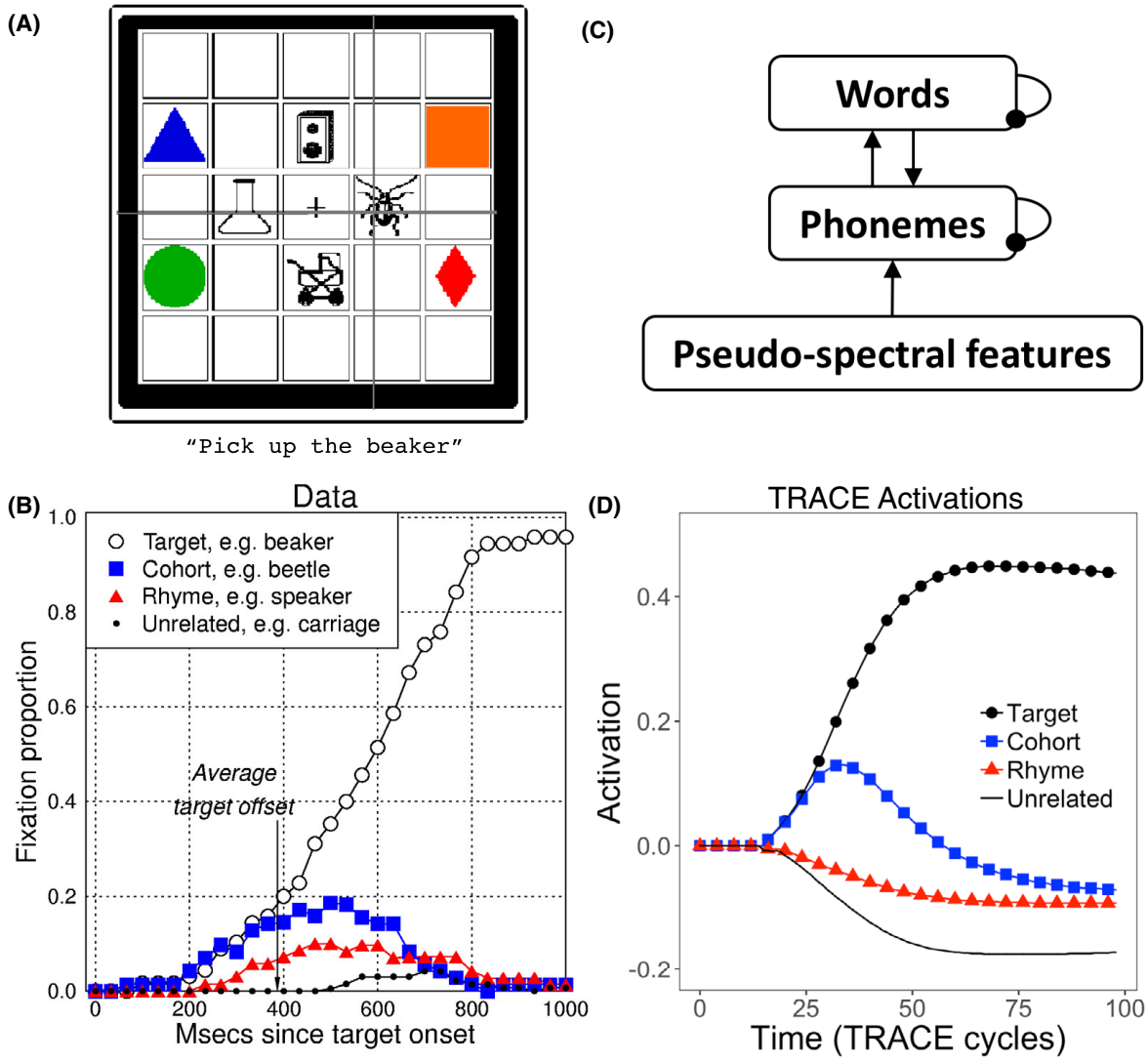

Fig. 1. Dynamics of phonological competition in humans and models. Allopenna et al. (1998) asked listeners to follow simple spoken instructions to interact with simple displays (e.g., A), and tracked their eye movements as they did so (B). Fixation proportions over time were hypothesized to relate to internal lexical activation and competition. They used the TRACE model (C; McClelland & Elman, 1986) to simulate their paradigm. Raw model activations (D) clearly resemble human performance (C); see Allopenna et al. (1998) for a more complex linking hypothesis and quantitative comparison. The TRACE results in (D) are new results created by considering all possible cohort and rhyme pairs in the TRACE lexicon; in Allopenna et al.'s (1998) simulations (not shown here), activations were clipped at 0.0, while we included items with negative activations. Panels A and B are adapted from Allopenna et al. (1998) as allowed under Elsevier's policy for personal use of published materials by authors. Panel C is Magnuson (2019).

become sensitive to dependencies over long time spans, but in practice, context degrades with each time step, severely limiting the span of learnable dependencies. LSTM nodes allow networks to learn variable-span dependencies, making them an excellent candidate for linking simple network models to real speech.

How does the choice to borrow LSTMs compare with other aspects of ASR we might consider with respect to utility for modeling HSR? On the one hand, LSTMs add complexity in the form of additional connections and recurrence (the additional input and recurrent connections required of the LSTM gates), which allows them to circumvent the vanishing gradient problem, and allows the network to operate on real speech. Other alternatives, such as adding more hidden layers to create a deeper network, are more likely to exacerbate rather than mitigate this issue. While it will be useful for future work to explore other possibilities, our aim here is to report on the demonstrable efficacy of LSTMs for processing the over-time speech signal with minimal increase in complexity.

2. Methods

2.1. Network architecture

We constructed a network (Fig. 2) dubbed EARSHOT, for *Emulation of Auditory Recognition of Speech by Humans Over Time*, emphasizing the key aims of working with real speech and emulating the timecourse of lexical access and competition. The model has 256 spectrographic inputs, 512 LSTM hidden nodes, and 300 pseudo-semantic outputs (random sparse vectors; a common simplification [Laszlo & Plaut, 2012] given the largely arbitrary mapping from sound to meaning). There are feedforward connections between layers, and the hidden layer is fully recurrent (each node has a connection to every other). A *tanh* activation function is applied to hidden unit outputs. Connections were trained using backpropagation through time (Werbos, 1988). The training target at each time step was the semantic vector corresponding to the current word. We trained several models on subsets of 1,000 words produced by 10 talkers (see Section 2.2). The maximum mean performance was approximately 90%, which required ~500 hidden units (we used multiples of 32 units at each level to facilitate easy expression of unit ratios).

2.2. Materials

A total of 1,000 words with length varying from 1 to 8 phonemes (mean: 5.5) were selected randomly from a list of uninflected English words, with the constraint that each of 39 English phonemes occur at least 10 times. Pronunciations of all 1,000 words were generated from 10 talkers included in the Apple text-to-speech application, *say* (five females and five males). Mean duration was 659 ms (range: 289–1,121 ms). For analysis purposes, we created 719 diphone syllables for all legal consonant-vowel and vowel-consonant combinations for each talker (24 consonants, 15 vowels). Each sound file was converted to a spectrogram with 256 channels in 10 ms steps, and a sampling rate of 8,000 Hz. For each word, a unique semantic target pattern was created: a sparse vector, with 10 of 300 randomly selected elements set to 1, and all others set to 0. Word-pattern mappings were randomized for each model.

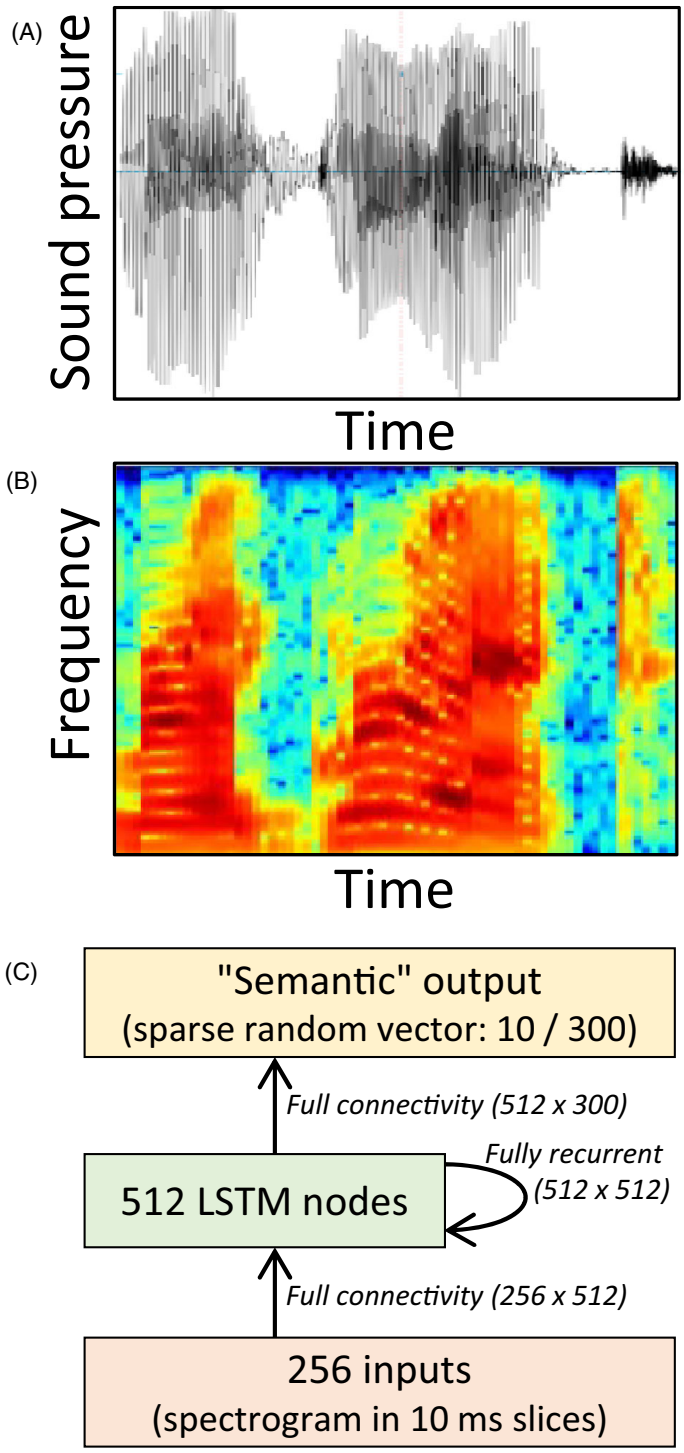

Fig. 2. Model input and structure. (A) Audio files are converted to spectrograms (B), with 256 channels (rows) in 10 ms steps (columns). Color indicates amplitude (blue-red indicates low-high). (C). The model is a standard recurrent network, except "long short-term memory" nodes are used in the hidden layer, allowing it to become sensitive to multiple temporal grains.

2.3. Training method

We created 10 different models. For each, a different talker was entirely excluded from training. For each of the nine trained-on talkers, 100 different, randomly selected words were excluded from training. Thus, each model's training set consisted of 8,100 input-output patterns (900 words × 9 talkers). Tests were conducted on all 10,000 items (10 talkers × 1,000 words, including items excluded from training). Training was organized as epochs. Each epoch included one presentation of each of the 8,100 training items in random order with no pause or other indication of word boundaries except that the training target pattern changed.

We used three techniques to increase learning speed and performance (Vaswani et al., 2017): minibatch gradient descent, Noam decay, and Adam optimizing. The 8,100 words were divided into five mini-batches (4 × 2,000, 1 × 100). A baseline learning rate of 0.002 was applied adaptively using Adam optimization and Noam decay (see Appendix S1).

2.4. Testing

To quantify the distance between the output vector at each time step to each word in the 1,000-word lexicon, we computed the cosine similarity of the output vector to all 1,000 semantic vectors (see Appendix S1 for equations). Because higher cosine indicates greater similarity, we defined a simple but conservative metric for word recognition accuracy. Accuracy was operationalized based on a two-parameter threshold: the output vector's cosine similarity to the target had to exceed any other item's cosine similarity to the output by a minimum of 0.05 for at least 100 ms; subsequently, no item could exceed the target's cosine similarity to the output before word offset (this is more conservative than a simple "maximum similarity" threshold). At the end of every 1,000 epochs, each model was tested with all 10,000 words (including excluded words and talkers). For additional details, see Appendix S1.

2.5. Additional details

See Appendix S1 and the EARSHOT Github repository (https://github.com/maglab-uconn/EARSHOT), which includes all simulation and analysis code.

3. Results

3.1. Accuracy

Models achieved high accuracy after 8,000 training epochs (Fig. 3): 88% for trained-on items, 67% for excluded words from trained-on talkers, and 33% for excluded talkers (range: 4%–78%). Generalization was poor for some talkers, but human listeners can *learn* to adapt to novel talkers. Thus, we explored how quickly models could learn when we resumed training with all items. Performance improved rapidly (to 89% for excluded

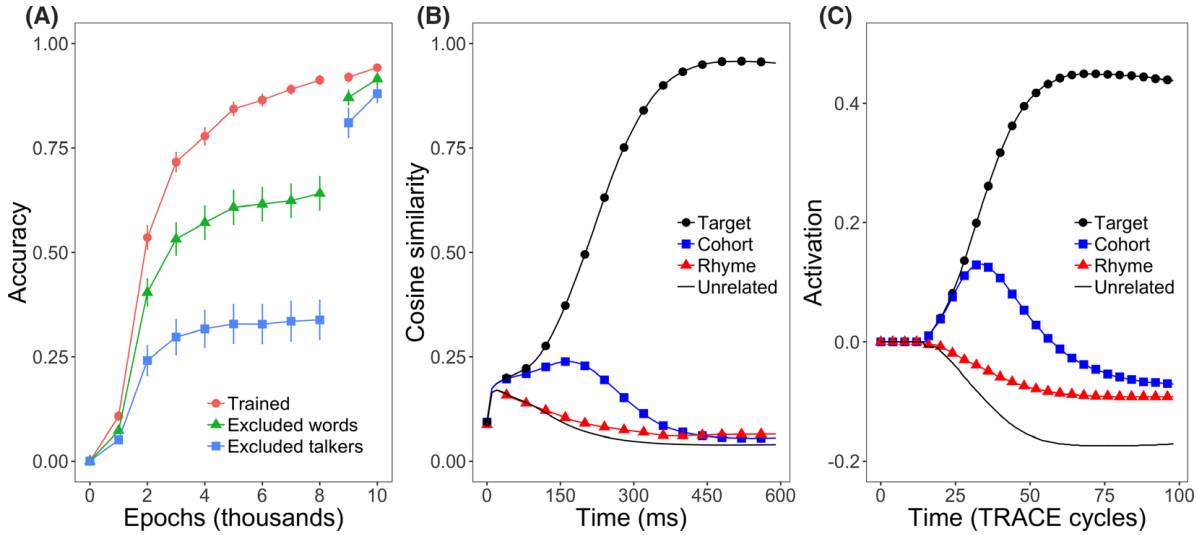

Fig. 3. Model performance. (A) Accuracy by epoch averaged over 10 models. When training resumed with all items (epochs 8,001–10,000), high accuracy was achieved quickly for all talkers. (B) The timecourse of competition for accurate trials, for two criterial competitor types. For a target (e.g., *cat*), "Cohort" represents mean cosine similarity for words overlapping in the first two phonemes (*can*, *castle,* . . .). "Rhymes" rhyme with the target (*bat*, *sat, at, scat* . . .). "Unrelated" is the average for all words phonologically dissimilar from the target. This pattern closely follows human performance (Allopenna et al., 1998). (C) For comparison, we conducted simulations with the TRACE model, with its standard 212-word lexicon, 14-phoneme inventory, and idealized "pseudo-spectral" inputs. Panel C (same data as panel D in Fig. 1) shows average activations by competitor types for all possible pairs in the TRACE lexicon. Crucially, EARSHOT exhibits the same rank ordering and similar timing for competitor types as the gold-standard TRACE model.

words and 86% for excluded talkers, and a boost to 93% for previously trained items). While the amount of training is too large to be analogous to rapid adaptation to new talkers by human listeners, it suggests a promising avenue for future investigation. Accuracy details by talker are presented in Appendix S1.

3.2. Timecourse

While high accuracy is a prerequisite for a valid model, a greater challenge is simulating the timecourse of HSR (Allopenna et al., 1998). The timecourse of HSR is a crucial explanatory target in speech science, but previous deep learning models of speech (e.g., Kell et al., 2018; Nagamine et al., 2015) have not been applied to timecourse. Our minimal model exhibits the correct qualitative pattern for phonological competition (Fig. 3). It is not necessarily the case that any model that can map speech inputs to semantic outputs would exhibit human-like timecourse; in Appendix S1, we describe a high-accuracy model with timecourse behavior that differs starkly from human performance. This suggests that there are important architectural choices that underlie EARSHOT's ability to simulate the correct qualitative timecourse pattern of lexical access and competition (Fig. 3B). First, it may be important that the input has a similar temporal resolution as the speech signal. Transformations that remove/compress too much spectral detail, such

as the Mel Frequency Cepstral Coefficient (MFCC) transformation used in the example model in Appendix S1, may transform the information processing task in ways that depart from the challenges faced by human listeners.

3.3. Hidden unit sensitivities

The next challenge is determining how the model works, with the aim of guiding cognitive theories of HSR. Since the model is learning to map speech input to pseudo-semantics, we also hypothesized that it might mediate that transformation by developing internal phonetic encoding. To begin, we adapted two "selectivity indices" (SIs) used with human electrocorticography data (Mesgarani et al., 2014). The *Phonemic Selectivity Index* (PSI) for a hidden unit-phoneme pair is the count of phonemes that evoke a substantially weaker response in that unit compared with the target phoneme. For example, given 39 phonemes, if a hidden unit responds more strongly to /p/ than any other phoneme, its PSI for /p/ would be 38 (maximum). The *Feature Selectivity Index* (FSI) does the same for *features* shared by classes of phonemes (e.g., nasal, labial, voiced). The SI approach allows us to ask whether phonetic structure emerges as the model learns to map speech to semantics, despite not being given explicit information about phonetic features or phonemes. To calculate PSI and FSI, we tracked the absolute amplitude of hidden node responses to each phoneme and feature over 100 ms. For example, for unit 239, we would note its mean activation in response to /b/ from the onset of /b/ to 100 ms later in all /b/-initial diphones. We would then subtract unit 239's response to every other phoneme in turn from its response to /b/. For each difference >0.3, the PSI for {239 /b/} would be incremented. We repeated this for all 39 phonemes.

We used hierarchical clustering to sort hidden units based on SIs (Fig. 4). Approximately 50% of nodes exhibit structured responses in the SI time window (in the human electrocorticography study our SI analyses are based on [Mesgarani et al., 2014], approximately 20% of electrodes met criteria for inclusion in SI analyses). The FSIs and PSIs *appear* remarkably similar to those derived from electrodes recording from human superior temporal gyrus (Mesgarani et al., 2014), with selective responses to features and phonetically similar phonemes, but it is important to quantify that similarity.

We used a *representational similarity analysis* (RSA; Kriegeskorte, Mur, & Bandettini, 2008) to quantify similarity of feature and phoneme selectivity in EARSHOT and human electrocorticography (ECoG) data (Mesgarani et al., 2014). On this approach, two systems with different kinds and numbers of elements can be compared in terms of their responses to classes of stimuli. We characterized EARSHOT's response to each phonetic feature as the mean vector of hidden unit responses when that feature is present (creating a 14 [feature] × 512 [hidden unit] matrix). This allowed us to compare the cosine similarities of EARSHOT's responses (or SIs) to each feature, and create a feature × feature dissimilarity (1 − similarity) matrix. We did the same with human ECoG data from Mesgarani et al. (2014, graciously provided by Chang and Mesgarani) in terms of the vector of electrode SIs in the presence of each feature, resulting in another feature × feature

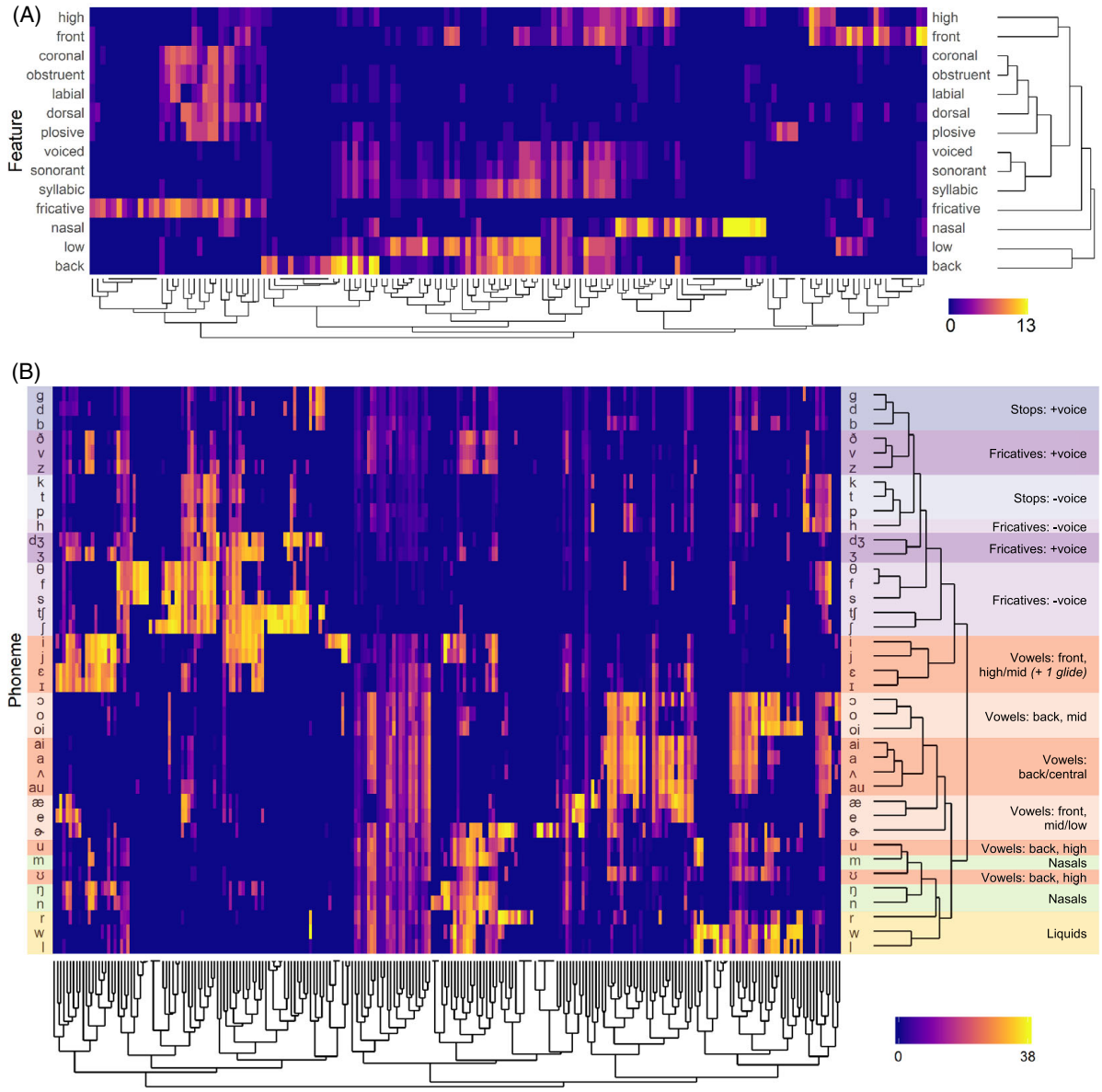

Fig. 4. Phonetic selectivity revealed by hierarchical clustering. (A) *Feature Selectivity Index* (*FSI*) based on hidden unit (*x*-axis) responses to phonetic features; for every hidden unit-feature pair, FSI was incremented for every feature to which the hidden unit responded substantially more weakly (yellow indicates high selectivity, with maximum FSI of 13, given 14 features). Vowel features pattern together (from high to back). One hundred and fifty-six hidden units with strongly selective responses are included. (B) *Phonemic Selectivity Index* (*PSI*). High PSI indicates selective responses to specific phonemes. Maximum score is 38, given 39 phonemes. Two hundred and thirty-nine hidden units showing selective responses are included.

matrix. We then compared EARSHOT and STG electrode dissimilarity matrices for both FSI and PSI. The results are summarized in Fig. 5 (see Appendix S1 for details), where we see strong correlations between model and human responses to phonetic features and phonemes.

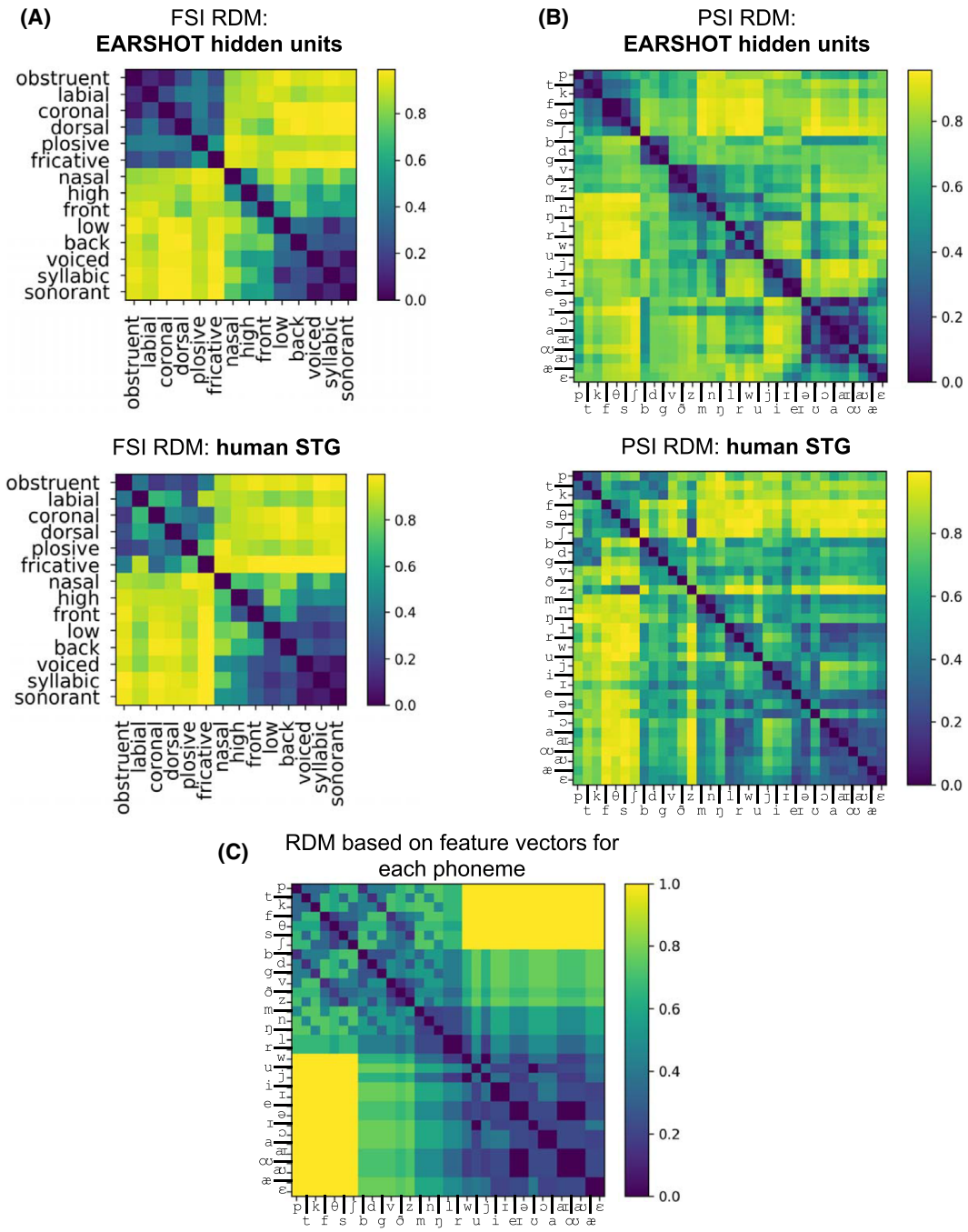

Fig. 5. Representational similarity analyses comparing EARSHOT and human neural responses. (A) *Representational Dissimilarity Matrices* (RDMs) for feature selectivity indices (FSIs) for EARSHOT and human STG ECoG data from Mesgarani et al. (2014). RDMs are created by calculating *dissimilarity* between vectors of FSIs for each hidden unit or electrode for each feature (low values [darker] indicate high similarity). The correlation between EARSHOT and STG RDMs was high: $r = .895$, $p < 1 \times 10^{-6}$ (based on a permutation test; see Appendix S1). (B) RDMs for EARSHOT and STG PSIs were also highly correlated ($r = .607$, $p < 1 \times 10^{-6}$). (C) A baseline RDM based on feature definitions for each phoneme. The correlation of this phoneme-feature RDM with the EARSHOT and STG PSIs were similar to the EARSHOT-STG PSI correlation (0.487 and 0.652, respectively; $p < 1 \times 10^{-6}$ for both).

We cannot conclude that EARSHOT-ECoG similarities indicate that EARSHOT directly implements mechanisms supporting HSR in the brain. However, the similarities demonstrate that both systems are sensitive to the phonetic structure available in the speech signal (again, even though EARSHOT is not trained on phonetic targets) and are sensitive to the structure in similar ways. Together with the fact that EARSHOT exhibits fine-grained timecourse of phonological competition similar to that seen in humans, similarity in internal selectivity to phonetic information suggests that EARSHOT may be a promising tool for discovering the mechanisms supporting HSR.

3.4. More complex hidden unit responses

Hidden units also have more complex dynamics than are revealed by the SIs (Fig. 6). Some develop strong, onset-locked responses, while others develop responses that include significant delays, and/or sustained responses. These response profiles suggest novel hypotheses for human cortical responses that could be explored in electrocorticographic recordings. The mapping from hidden states over time to semantic outputs likely depends both on intuitive profiles like the time-locked responses assumed by the SI analyses and on complex over-time patterns of combinations of those and other profiles. Additional details of hidden unit profiles and responses are presented in Appendix S1.

4. Discussion

Decades after the discovery of the *lack of invariance problem*—the absence of invariant cues to speech sounds (e.g., Joos, 1948; Liberman et al., 1952; Peterson & Barney, 1952)—speech science offers limited explanations of how humans achieve phonetic constancy despite the many-to-many mapping between acoustics and percepts. Computational models of HSR have provided little insight, since most current models sidestep the vagaries of the signal and use idealized, abstract elements such as phonetic features (McClelland & Elman, 1986), phonemes (Hannagan, Magnuson, & Grainger, 2013; You & Magnuson, 2018), or human phoneme confusion probabilities (Norris & McQueen, 2008) rather than real speech as input. Such assumptions can ultimately complicate rather than simplify problems (Magnuson, 2008), as the details they bypass may contain constraints essential to the mechanisms underlying human performance.

Complexity is a primary motivation for abstract inputs. As McClelland and Elman (1986) argued, a computational model aimed at guiding a psychological theory must prioritize psychological adequacy over computational adequacy. That is, such a model must favor simplicity and understandability over full, end-to-end modeling when the latter results in a model too complex to understand. However, we would expand beyond psychological and computational adequacy when considering the tension between simplicity and realism. We divide psychological adequacy into three parts. *Behavioral adequacy* is the ability of a model to simulate key details of human performance. *Developmental adequacy* is the ability of a model to learn as well as the degree to which learning by the

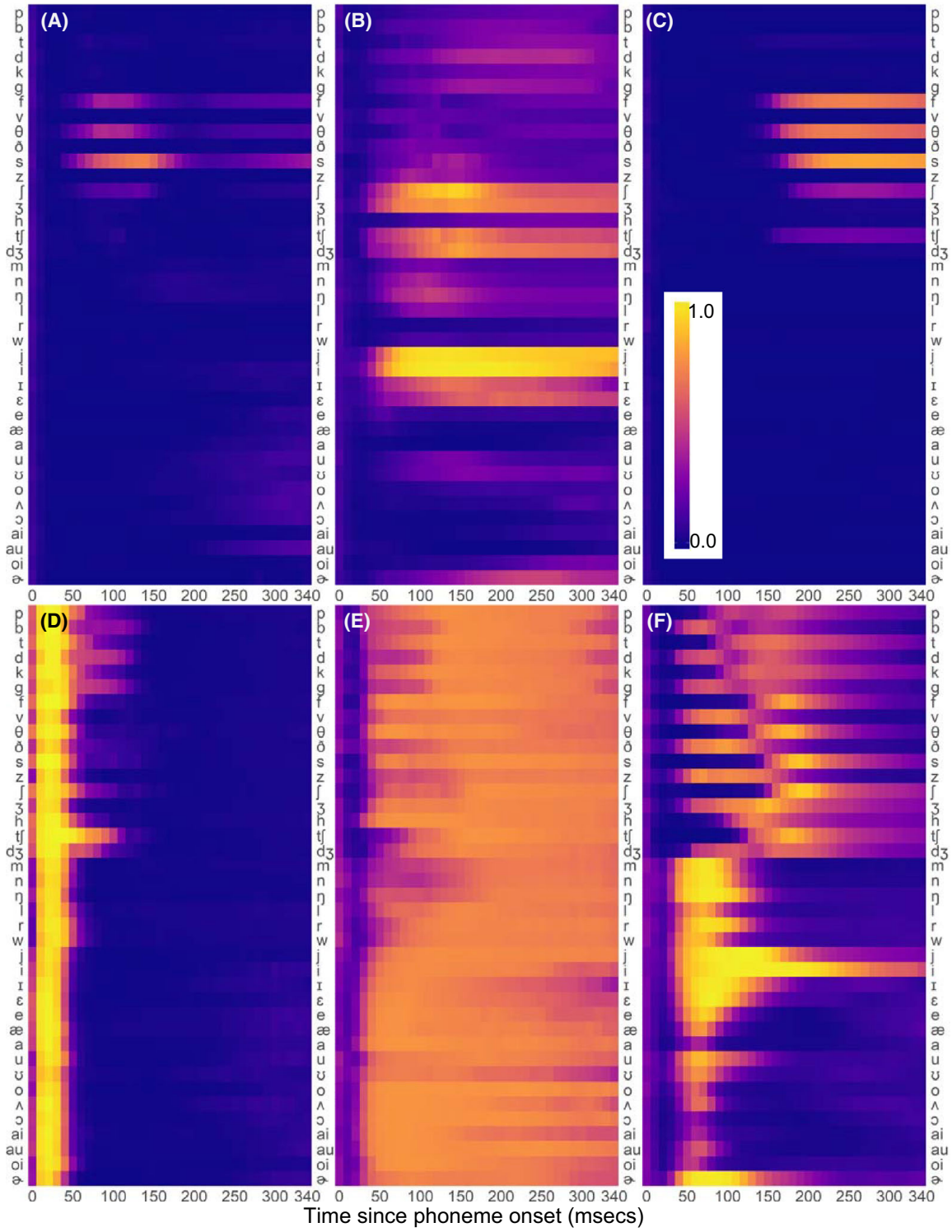

Fig. 6. Hidden unit response profiles based on absolute activation values. Over-time response profiles of example hidden units for each phoneme (*y*-axis). (A) Time locked, discrete responses (~5% of units). (B) Time locked, sustained responses (~20%). (C) Delayed responses (~35%). (D) Early-onset responses (~4%). (E) Post-onset inactivation (~3%). (F) Complex responses (~29% of HUs). An additional ~4% are largely non-responsive.

model is relatable to trajectories in human development. *Explanatory adequacy* is the degree to which the mechanisms of the model are analyzable and understandable; a model could have high adequacy in every other domain, but its utility in guiding theories of human capabilities would be limited if the mechanisms implemented in the model are inscrutable. We would also complement computational and psychological adequacy with *neural adequacy*: the ability of the model to relate to knowledge or theories of the neurobiological mechanisms underlying the modeled capacities.

We demonstrated that borrowing one element from ASR—long short-term memory (LSTM) nodes (Hochreiter & Schmidhuber, 1997)—is sufficient to allow a shallow recurrent network to learn to map from speech to arbitrary outputs (pseudo-semantic vectors), while also demonstrating a timecourse of lexical activation and competition (Fig. 3) that resembles that observed in human subjects and current gold-standard models of HSR (Allopenna et al., 1998). This represents a major advance in computational adequacy; EARSHOT is capable of performance with real speech that is unprecedented for a simple model aimed at guiding cognitive theory. On the other hand, EARSHOT's modest generalization to excluded talkers and excluded words from trained-on talkers should give us pause. Low and variable generalization may indicate that the model memorizes training patterns to some degree. In ongoing work, we are striving to use more variable inputs and ultimately will train the model on large numbers of human talkers. It will also be necessary to assess EARSHOT's ability to account for the full range of phenomena that models that work on abstractions of the speech signal (e.g., TRACE [McClelland & Elman, 1986]; TISK [Hannagan et al., 2013]; and Shortlist B [Norris & McQueen, 2008]) are able to simulate.

EARSHOT also has the potential to address developmental adequacy, since it is a learning model. In this preliminary work, we have not yet attempted to make the training of the model realistic or to link its development to human developmental trajectories. This will be a priority in future work.

Regarding EARSHOT's explanatory adequacy, we do not yet fully understand *how* the model succeeds in learning to map speech to semantics. We demonstrated that we can *begin* to unpack how EARSHOT learns to map speech to semantics by using techniques from human electrocorticography (Mesgarani et al., 2014) to track responses of hidden units to specific phonetic features and phonemes (Figs. 4–6). EARSHOT's emergent sensitivity to phonetic structure, *despite receiving no explicit phonetic training*, provides preliminary clues as to how such a simple learning system could achieve a speech-to-semantics mapping. Of course, fully understanding how the model works will require analyses beyond the phonetic structure apparent from the feature and phoneme SIs. First, it is apparent from the variation in hidden unit response profiles (Fig. 6) that the phonetic responses of the hidden units are substantially more complex than the selectivity analyses suggest. We expect that complex population responses are essential to how the model transforms spectral slices to semantics (rather than a system where individual units function as simple detectors for specific phonemes; see Morcos, Barrett, Rabinowitz, & Botvinick, 2018). Second, a full understanding of the model will also require unpacking the transformation from hidden unit states to semantic outputs. However, even the

preliminary similarity of EARSHOT's hidden unit responses to human electrocorticographic data suggests that such a model holds promise for addressing neural adequacy. Indeed, an intriguing possibility is that the variations in response profiles observed could generate hypotheses for potential response profiles in human cortical encoding of speech.

5. Conclusions

By borrowing one minimal element from ASR (long short-term memory nodes), EARSHOT opens new territory to computational exploration of HSR thanks to its ability to operate on real speech inputs. The fundamental challenges of the lack-of-invariance problem, which have been outside the scope of cognitive models of HSR for decades, are now addressable. These include variation in talker characteristics, speaking rate, and acoustic context, and integration of theories of development and processing. Simulations can be conducted with the same materials presented to human listeners, instead of idealized, abstract analogs of those materials. Finally, the fact that the distributed phonological code that emerges as the model learns to map speech to semantics resembles responses observed in human cortex (Mesgarani et al., 2014) demonstrates the promise of this approach as a testbed for theories of neurobiological mechanisms that may support HSR.

Acknowledgments

We thank Rachael Steiner for technical assistance and comments. We thank Inge-Marie Eigsti, Blair Armstrong, Thomas Hannagan, and Ram Frost for helpful comments. This work was supported by the following grants: NSF 1754284 (PI: JSM), NSF IGERT 1144399 (PI: JSM), NSF NRT 1747486 (PI: JSM), NICHD P01 HD0001994 (PI: JR), and NSF 1827591 (PI: RMT). We thank Eddie Chang and Nima Mesgarani for supplying us with data from Mesgarani et al. (2014) used to compare EARSHOT and human STG responses. We note again that our source code for simulations and analyses is available at the EARSHOT github repository (https://github.com/maglab-uconn/EARSHOT).

Author contributions

J.S.M., H.Y., H.N., and J.R. conceived of the initial project; all authors contributed to interpretation of results and development of the data analysis strategies, and to the writing of the manuscript; H.Y. implemented the model and analysis scripts; S.L. and H.Y. conducted RSA analyses; J.S.M. wrote the original draft of the manuscript. Overall, the first two authors made similar contributions to this project.

Note

1. Because the output vectors are arbitrary, they could stand for anything, including discrete word form patterns (analogous to lexical nodes in TRACE [McClelland & Elman, 1986]). Future work will use distributed semantic vectors based on corpus analyses.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the timecourse of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Cole, R. A., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 133–163). Hillsdale, NJ: Erlbaum.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*, 333–341.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*, 489–504.

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 418–503.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, *4*, 563. https://doi.org/10.3389/fpsyg.2013.00563

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*, 82–97.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer & J. F. Konlen (Eds.), *A field guide to dynamical recurrent neural networks* (pp. 237–374). New York: IEEE Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Joos, M. (1948). *Acoustic phonetics*. Baltimore, MD: Linguistic Society of America.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norma-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*, 630–644.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. In S. Murray Sherman (Ed.), *Oxford research encyclopedia of neuroscience*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 1–28.

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential reading data. *Brain and Language*, *120*, 271–281.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology*, *65*, 497–516.

Magnuson, J. S. (2008). Nondeterminism, pleiotropy, and single word reading: Theoretical and practical concerns. In E. Grigorenko & A. Naples (Eds.), *Single word reading* (pp. 377–404). Hillsdale, NJ: Erlabaum.

Magnuson, J. S. (2019). Very simple TRACE schematic (Version 1). *figshare*, https://doi.org/10.6084/m9.figshare.8273261.v1

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

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

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America*, *73*, 1751–1755.

Morcos, A. S., Barrett, D. G. T., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv*:1803.06959v4.

Nagamine, T., Seltzer, M. L., & Mesgarani, N. (2015). Exploring how deep neural networks form phonemic categories. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015-January, (pp. 1912–1916).

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. *Journal of the Acoustical Society of America*, *127*, 3758–3770.

Scharenborg, O., Norris, D., ten Bosch, L., & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, *29*, 867–918.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv*:1706.03762v5 [cs.CL].

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*, 339–356.

You, H., & Magnuson, J. S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*, *50*(3), 871–889. https://doi.org/10.3758/s13428-017-1012-5

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Supplementary methods, details, and results.