SUPPLEMENTARY MATERIALS

This document contains Supplementary Methods (§S1) and Supplementary Results (§S2) for:

Magnuson, J.S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P.D., Theodore, R.M., Monto, N., & Rueckl, J.G. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science, 44*, e12823. http://dx.doi.org/10.1111/cogs.12823

S1. SUPPLEMENTARY METHODS

S1.1 EQUATIONS FOR TRAINING AND TESTING

Training. We used three techniques to increase learning speed and performance (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017): minibatch gradient descent, Noam decay, and Adam optimizing. The 8100 words were divided into 5 mini-batches (4 x 2000, 1 x 100). A baseline learning rate of 0.002 was applied adaptively using Adam optimization and Noam decay as in Equation 1.

$$LR = 0.002 \times 4000^{0.5} \times min\,(E \times 5 \times 4000^{-1.5}, (E \times 5)^{-0.5}) \qquad [1]$$

LR and *E* denote learning rate and epoch. The hyper parameters ß$_1$, ß$_2$, and *ε* for Adam optimization were 0.9, 0.999, and 1e-08. These values are fairly standard, and the specific values are not crucial for any of the results.

Testing. To quantify the distance of the output vector at each time step to each word in the 1000-word lexicon, we computed the cosine similarity of the output vector to all 1000 semantic vectors:

$$Cosine\ similarity = \left(\sum_{i=1}^{n} T_i \times O_i\right) / \left(\sqrt{\sum_{i=1}^{n} T_i^2} \times \sqrt{\sum_{i=1}^{n} O_i^2}\right) \qquad [2]$$

T, *O*, and *n* indicate the target, output, and vector length, respectively. *O* indicates the output vector at one time step.

Simulation method (training and testing). The simulation process was as follows. First, a 10-ms spectrogram slice *I* was applied to the input layer at each time step. Hidden activation *H* of each time step was derived through input *I*. The following formulas (cf. Vaswani et al., 2017) were used for calculating *H*.

$$i_t = \sigma(I_t W_{Ii} + H_{t-1} W_{Hi} + c_{t-1} W_{ci} + b_i) \quad [3]$$

$$f_t = \sigma(I_t W_{If} + H_{t-1} W_{Hf} + c_{t-1} W_{cf} + b_f \quad [4]$$

$$c_t = f_t c_{t-1} + i_t \times \tanh(I_t W_{Ic} + H_{t-1} W_{Hc} + b_c) \quad [5]$$

$$o_t = \sigma(I_t W_{Io} + H_{t-1} W_{Ho} + c_t W_{co} + b_o) \quad [6]$$

$$H_t = \tanh(o_t) \times \tanh\,(c_t) \quad [7]$$

In the equations above, *i*, *f*, and *o* denote *input*, *forget*, and *output* LSTM gates, respectively. *c* is the LSTM cell memory, and *W* is the weight that connects two subscripted nodes. *b* is the bias of the subscripted node. *σ* and *tanh* are activation functions (sigmoid and tanh, respectively). The semantic output activation *O* was derived using the following equation.

$$O_t = \sigma(H_t W_{HO} + b_o) \quad [8]$$

O was derived for all time steps, and backpropagation was also performed for *O* at all time steps (and preceding steps, as this is backpropagation through time).

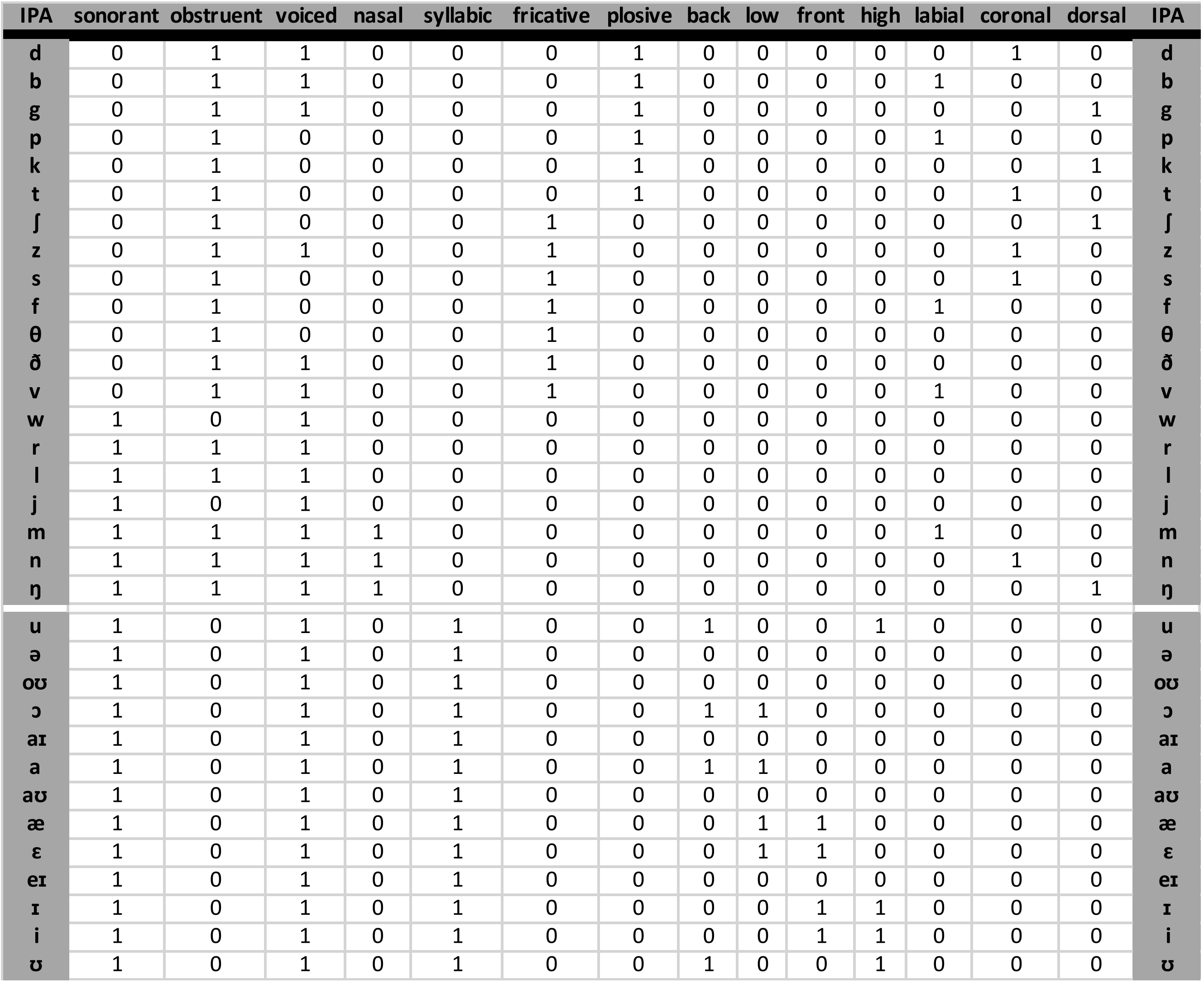
S1.2 SELECTIVITY INDICES

S1.2.1 The *phoneme selectivity index* (PSI) of hidden units was calculated as follows. First, we calculated the absolute value of each hidden unit's activation over time in response to all CV- and VC-diphones. Then, for each initial phoneme, we averaged each hidden unit's response to all diphones beginning with that segment, to derive the mean response of each hidden unit to each phoneme. We found that the modal maximal response period across all hidden units in response to all phonemes occurred from 0-60 ms after phoneme onset. We thus characterized the response of each hidden unit to each phoneme as the mean response to all diphones beginning with that phoneme over the 0-60 ms time period. Then for each phoneme-hidden unit pair, we calculated a PSI value as follows. Phoneme (*i*) - hidden unit (*j*) pair P_{ij} received 1 point for every phoneme to which hidden unit *j* responded more weakly than it did to phoneme *i* by a threshold (0.15). So, for example, if the activation of hidden unit 207 in response to /p/ exceeded its response to /b/ by 0.24, the PSI for $P_{/p/,207}$ was incremented. The maximum PSI was 38, which would indicate that the response of a hidden unit to a particular phoneme exceeded the threshold difference for all other phonemes.

In the study that motivated our use of the PSI (Mesgarani et al., 2014), the Wilcoxon rank sum test was used to compare electrode responses to phoneme pairs, with PSIs incremented when the difference was significant. However, we used a simple criterion in this study because very small differences easily reached significance. The threshold of 0.15 provided a level of sparsity similar to that reported in human selectivity indices[13].

To examine structured responses via the PSI, we used simple hierarchical clustering of each hidden unit's PSIs for all 39 phonemes (Fig. 4). Any hidden units that had PSIs of 0 for all phonemes were excluded.

S1.2.2 The *Feature Selectivity Index* (FSI) uses the same method as the PSI, but linked to features rather than phonemes. For example, for the FSI to "voiced", all diphones with a voiced segment in the first position were used. Our feature definitions are listed in Table S1.1.

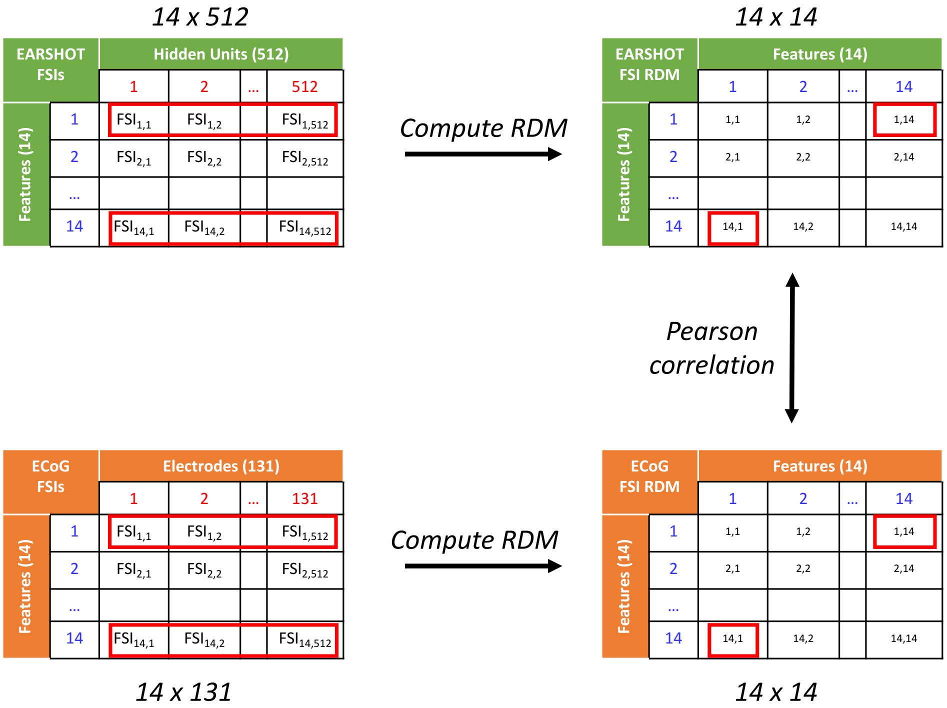


IPA	sonorant	obstruent	voiced	nasal	syllabic	fricative	plosive	back	low	front	high	labial	coronal	dorsal	IPA
d	0	1	1	0	0	0	1	0	0	0	0	0	1	0	d
b	0	1	1	0	0	0	1	0	0	0	0	1	0	0	b
g	0	1	1	0	0	0	1	0	0	0	0	0	0	1	g
p	0	1	0	0	0	0	1	0	0	0	0	1	0	0	p
k	0	1	0	0	0	0	1	0	0	0	0	0	0	1	k
t	0	1	0	0	0	0	1	0	0	0	0	0	1	0	t
ʃ	0	1	0	0	0	1	0	0	0	0	0	0	0	1	ʃ
z	0	1	1	0	0	1	0	0	0	0	0	0	1	0	z
s	0	1	0	0	0	1	0	0	0	0	0	0	1	0	s
f	0	1	0	0	0	1	0	0	0	0	0	1	0	0	f
θ	0	1	0	0	0	1	0	0	0	0	0	0	0	0	θ
ð	0	1	1	0	0	1	0	0	0	0	0	0	0	0	ð
v	0	1	1	0	0	1	0	0	0	0	0	1	0	0	v
w	1	0	1	0	0	0	0	0	0	0	0	0	0	0	w
r	1	1	1	0	0	0	0	0	0	0	0	0	0	0	r
l	1	1	1	0	0	0	0	0	0	0	0	0	0	0	l
j	1	0	1	0	0	0	0	0	0	0	0	0	0	0	j
m	1	1	1	1	0	0	0	0	0	0	0	1	0	0	m
n	1	1	1	1	0	0	0	0	0	0	0	0	1	0	n
ŋ	1	1	1	1	0	0	0	0	0	0	0	0	0	1	ŋ
u	1	0	1	0	1	0	0	1	0	0	1	0	0	0	u
ə	1	0	1	0	1	0	0	0	0	0	0	0	0	0	ə
oʊ	1	0	1	0	1	0	0	0	0	0	0	0	0	0	oʊ
ɔ	1	0	1	0	1	0	0	1	1	0	0	0	0	0	ɔ
aɪ	1	0	1	0	1	0	0	0	0	0	0	0	0	0	aɪ
a	1	0	1	0	1	0	0	1	1	0	0	0	0	0	a
aʊ	1	0	1	0	1	0	0	0	0	0	0	0	0	0	aʊ
æ	1	0	1	0	1	0	0	0	1	1	0	0	0	0	æ
ɛ	1	0	1	0	1	0	0	0	1	1	0	0	0	0	ɛ
eɪ	1	0	1	0	1	0	0	0	0	0	0	0	0	0	eɪ
ɪ	1	0	1	0	1	0	0	0	0	1	1	0	0	0	ɪ
i	1	0	1	0	1	0	0	0	0	1	1	0	0	0	i
ʊ	1	0	1	0	1	0	0	1	0	0	1	0	0	0	ʊ

Table S1.1: Phoneme-feature correspondences.

S1.3 REPRESENTATIONAL SIMILARITY ANALYSIS (RSA)

To quantify similarity between EARSHOT's hidden unit responses. and ECoG recordings from human STG (Mesgarani et al., 2014), we conducted a Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). This procedure is summarized in Fig. S1.1.



EARSHOT FSIs		Hidden Units (512)			
		1	2	…	512
Features (14)	1	$FSI_{1,1}$	$FSI_{1,2}$		$FSI_{1,512}$
	2	$FSI_{2,1}$	$FSI_{2,2}$		$FSI_{2,512}$
	…				
	14	$FSI_{14,1}$	$FSI_{14,2}$		$FSI_{14,512}$

EARSHOT FSI RDM		Features (14)			
		1	2	…	14
Features (14)	1	1,1	1,2		1,14
	2	2,1	2,2		2,14
	…				
	14	14,1	14,2		14,14

ECoG FSIs		Electrodes (131)			
		1	2	…	131
Features (14)	1	$FSI_{1,1}$	$FSI_{1,2}$		$FSI_{1,512}$
	2	$FSI_{2,1}$	$FSI_{2,2}$		$FSI_{2,512}$
	…				
	14	$FSI_{14,1}$	$FSI_{14,2}$		$FSI_{14,512}$

ECoG FSI RDM		Features (14)			
		1	2	…	14
Features (14)	1	1,1	1,2		1,14
	2	2,1	2,2		2,14
	…				
	14	14,1	14,2		14,14

Figure S1.1. General schematic of RSA process. The pattern of feature selectivity in EARSHOT's hidden layer can be expressed as a 14 x 512 (feature-by-hidden-unit) matrix of FSI values (top left). From this, we compute a 14 x 14 (feature by feature) representational dissimilarity matrix (RDM; top right), which compares the pattern of selectivity to one feature (across all hidden units) to the pattern of selectivity to another feature. A corresponding RDM is computed for the human electrocorticography (ECoG) data (bottom panels). Finally, the two RDMs are compared using a Pearson correlation. Permutation tests were used to estimate chance-level correlations. In particular, we shuffled the rows of the EARSHOT FSI matrix prior to computing the EARSHOT RDM and then computed the correlation with the original ECoG RDM; this procedure was repeated 1 million times, yielding a distribution of permutation values that would be expected by chance.

Analyses were limited to the set of 14 articulatory features (e.g., voiced, fricative, labial) used by Mesgarani et al. (2014). Dissimilarity matrices were computed using cosine similarity. Results (Fig. 5 in the main text) indicated that the pattern of feature selectivity in EARSHOT strongly resembled the pattern of selectivity in human ECoG, and the correlation was significantly above what would be expected by chance, $r = 0.895$, $p < 1\times10^{-6}$ (as established by a permutation test, where we randomized one RDM before calculating the correlation 1 million permutations times). An analogous assessment of RDMs based on PSI matrices (Fig. 5 in the main text, panel A) yielded similar results, with a strong correlation between EARSHOT and human neural data, $r = 0.607$, $p < 1\times10^{-6}$.

To understand which sub-phonemic details might be represented in EARSHOT and by humans, we might consider how much degree of similarity between phonemes is predictable from how much they overlap in terms of features. In examining the PSI RDMs (Fig. 5, panel B), one observes, for instance, that the patterns for /p/ and /t/ are quite similar, and, on classic binary acoustic-phonetic features, these two sounds differ only in their place of articulation (/p/ is labial, /t/ is alveolar). By contrast, the patterns for /p/ and /n/ are relatively dissimilar. They have the same place dissimilarity as /p/ and /t/ (labial vs alveolar), but also differ in manner of articulation (plosive vs nasal), and voicing (voiceless vs voiced). Thus, we can establish a baseline similarity of the EARSHOT and STG RDMs to an RDM based on the features associated with each phoneme (Fig. 5, panel C). The robust correlations between the EARSHOT and STG PSIs to the phoneme-feature RDM further suggest that both are sensitive to phonetic properties in the speech signal (rather than other acoustic information at a finer or coarser grain). In Fig. S1.2, we present the distributions of correlations observed in the permutation tests along with the actual correlations for unscrambled RDMs.

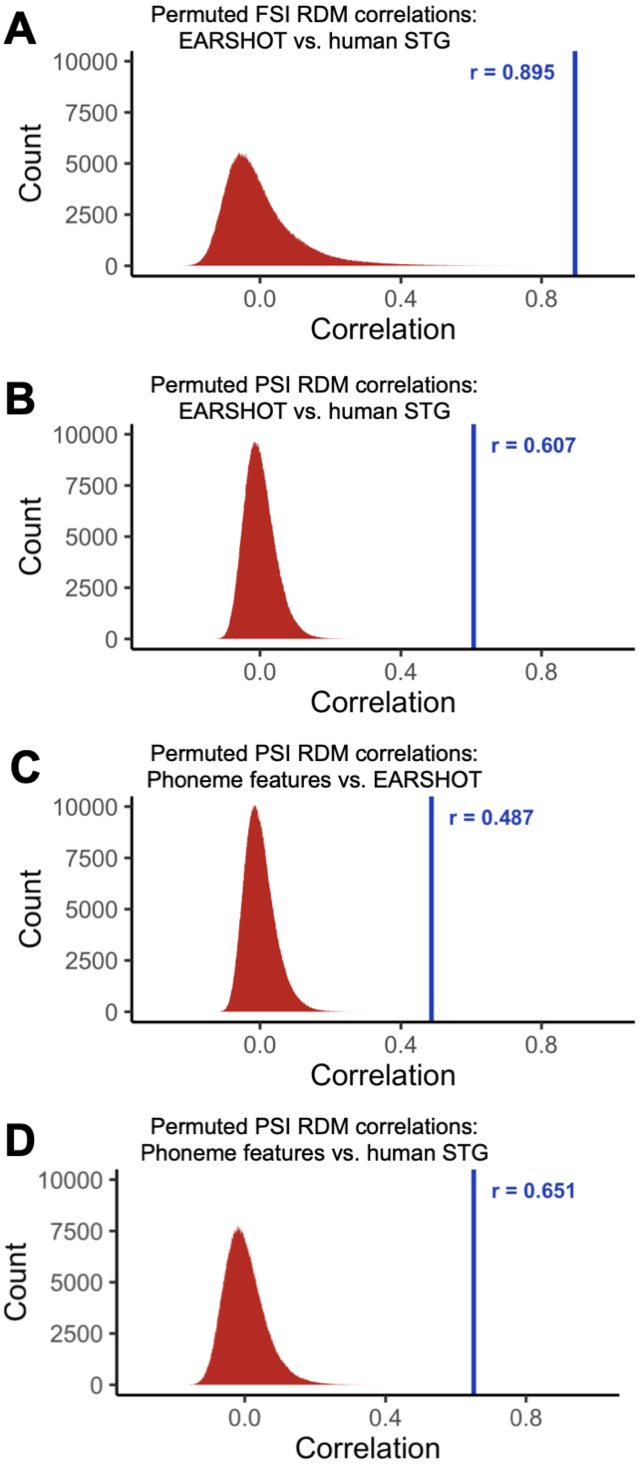


Figure S1.2. Permutation test results for RDM comparisons. In each case, one of the two RDMs being compared was shuffled randomly 1,000,000 times, and the correlation was computed each time. Red indicates the distribution of observed correlations. The line and value in each plot indicate the correlation for unshuffled RDMs.

S1.4 REPLICABILITY

Replicability was confirmed by repeating the complete training of 10 models three times; only minor variations were observed between iterations.

S1.5 HARDWARE AND SOFTWARE

Simulations were conducted on a Windows 10 workstation with an i7-6700k CPU, 64-gb of RAM, and a Titan-X (12-gb) graphics card. Simulations were implemented using Python 3.6 and TensorFlow 1.7. Each model requires approximately 10 hours to train on this workstation. The EARSHOT github repository (https://github.com/maglab-uconn/EARSHOT) provides an up-to-date Linux container with all necessary software and libraries for running our simulation code and analyses. However, conducting simulations will still require a high-performance workstation.

S1.6 ALTERNATIVE MODELS

In developing this model, we explored dozens of combinations of candidate architectures and inputs. All were limited to two layers (inputs-to-hidden and hidden-to-output). For architectures, we varied three aspects of models: number of hidden units (which we typically varied from 100 to 1000 nodes before rejecting an architecture for accuracy below 90%), hidden unit type (standard integrative nodes vs. LSTMs), and degree of recurrence (full recurrence, as in the model reported here, vs. single-step recurrence, as in simple recurrent networks; Elman, 1990). For inputs, we tried spectrograms at various resolutions, Mel Frequency Cepstral Coefficients (MFCCs), and cochleagrams.

Most combinations failed to achieve high accuracy. The only combinations that achieved greater than 90% accuracy were those reported here for EARSHOT (though similar results can be obtained with somewhat fewer hidden units) and a similar model using low-dimensional MFCCs rather than spectrographic inputs. However, the latter failed to show realistic timecourse (see Fig. S1.3).

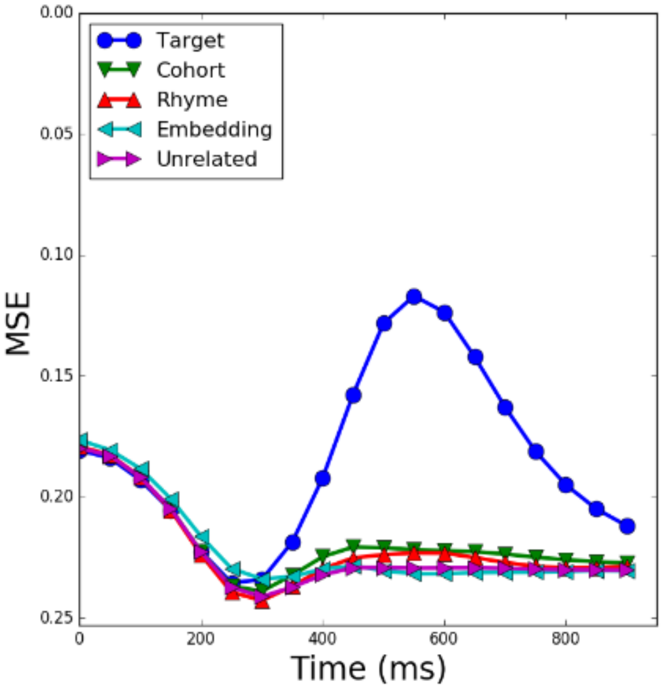


Fig. S1.3. Unrealistic timecourse in a high-accuracy model. This example illustrates that the correct timecourse does not necessarily emerge from any model with high accuracy. For this model, we used 13-element Mel Frequency Cepstral Coefficients (MFCCs, a common transformation used in ASR) as inputs to 500 LSTM nodes that mapped onto 300 semantic outputs. The model achieved 95% accuracy on a 200-word lexicon produced by 10 talkers. Here, we tracked mean squared error (MSE) to each pattern (note that the MSE scale is reversed to facilitate comparison to Fig. 2 in the main text), and a radically unrealistic timecourse (compared to human behavior; see Fig. 2 in the main text) emerged.

S2. SUPPLEMENTARY RESULTS

In this section, we present four figures that augment results presented in the primary article. In Fig. S2.1, we show the progression of training accuracy for all ten models that were trained. In Fig. S2.2, we show over-time activations of all 512 hidden units in response to phonemes, and their responses to features in Fig. S2.3. In Fig. S2.4, we show responses to phonemes and features by representative hidden units to illustrate how many units appear more sensitive to a phonemic rather than featural grain.

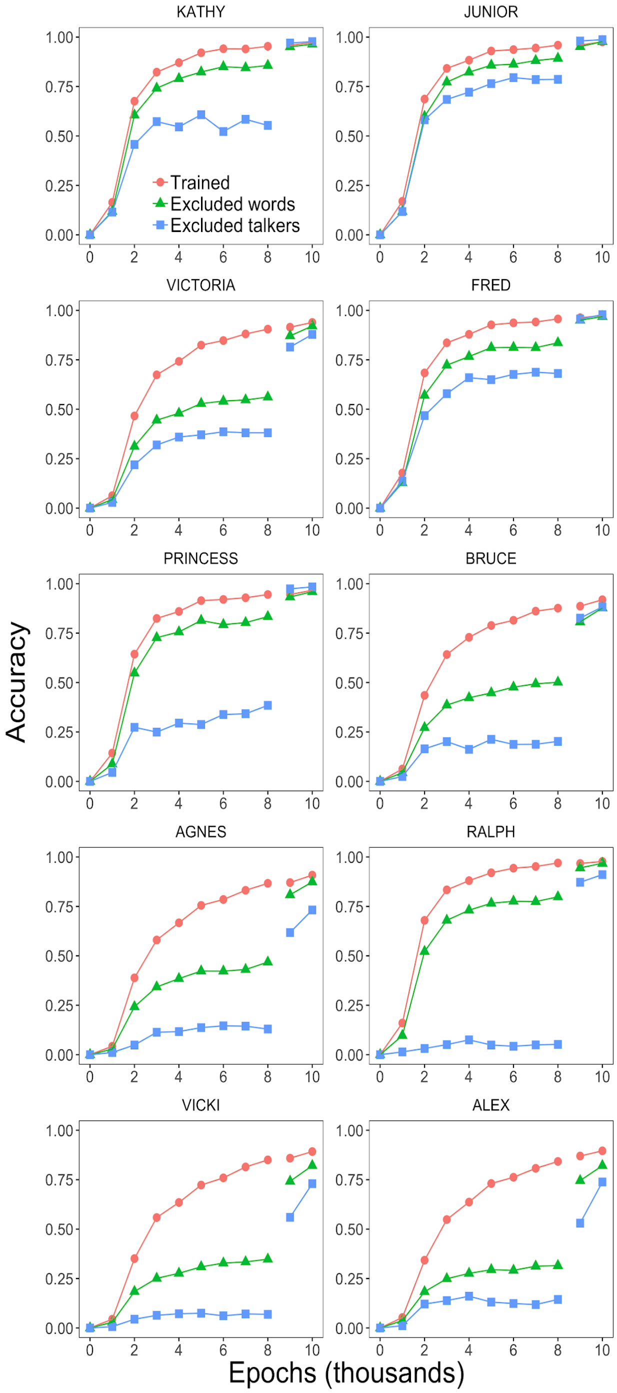


Fig. S2.1. Accuracy over epochs organized by talkers. In each panel, "Trained" indicates the performance on the listed talker (e.g., "KATHY") in the 9 simulations where it was included. "Excluded words" indicates performance on the listed talker's 100 excluded words in the 9 models that included that talker in training. The "Excluded talker" lines track performance on the listed talker when it was excluded. Training for each model was conducted for 8000 epochs with 100 words per training talker excluded and one talker excluded completely. For epochs 8001-10,000, excluded items were introduced to the training set. As can be seen from the figure, even for talkers for which generalization was initially poor, training allowed rapid improvement.

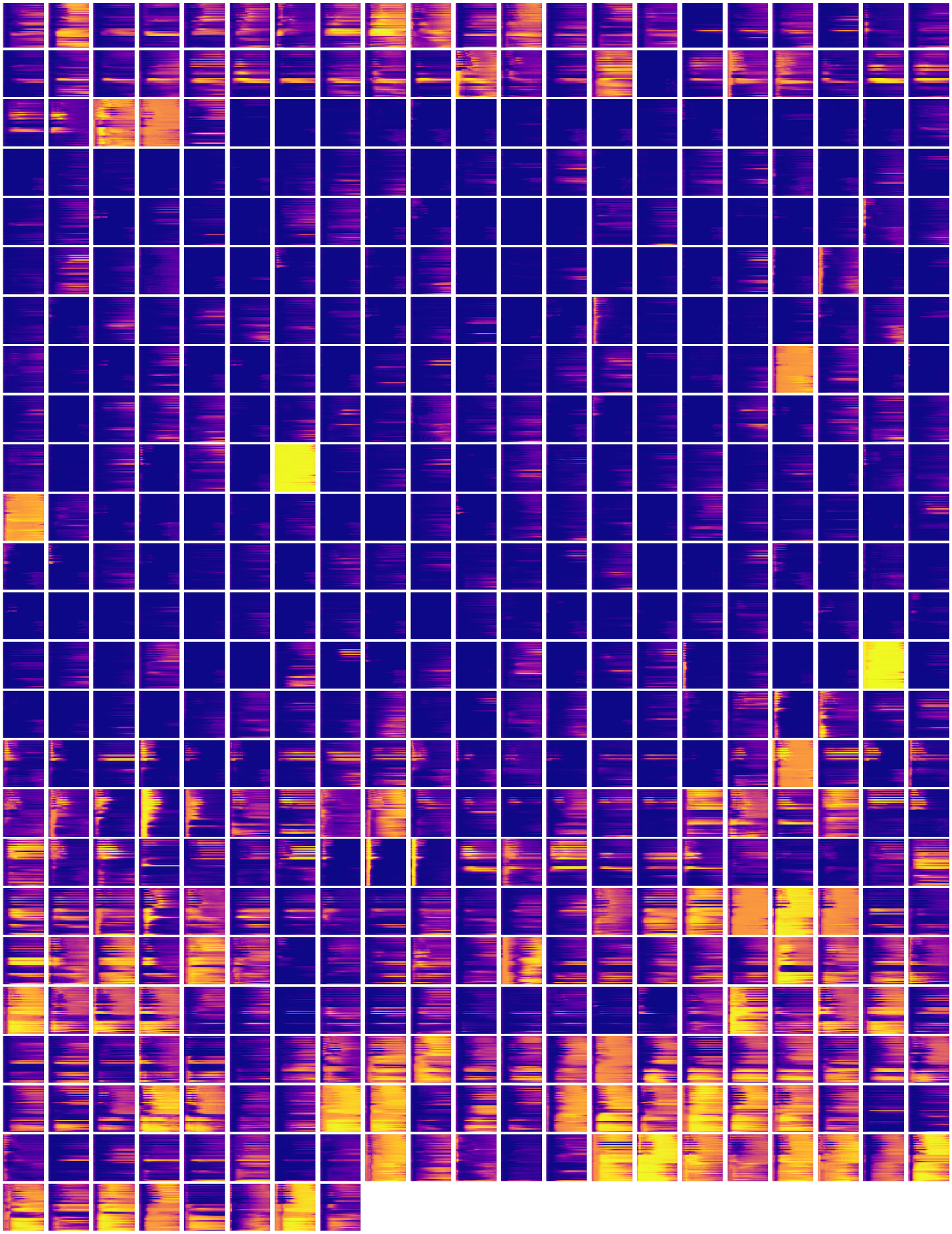


Fig. S2.2. Activations over time of all hidden units to each phoneme. Each panel displays one unit's response to each phoneme (y-axis, with the same ordering as Fig. 4 in the main article) over a period of 350ms. Units are ordered according to similarity in response profiles.

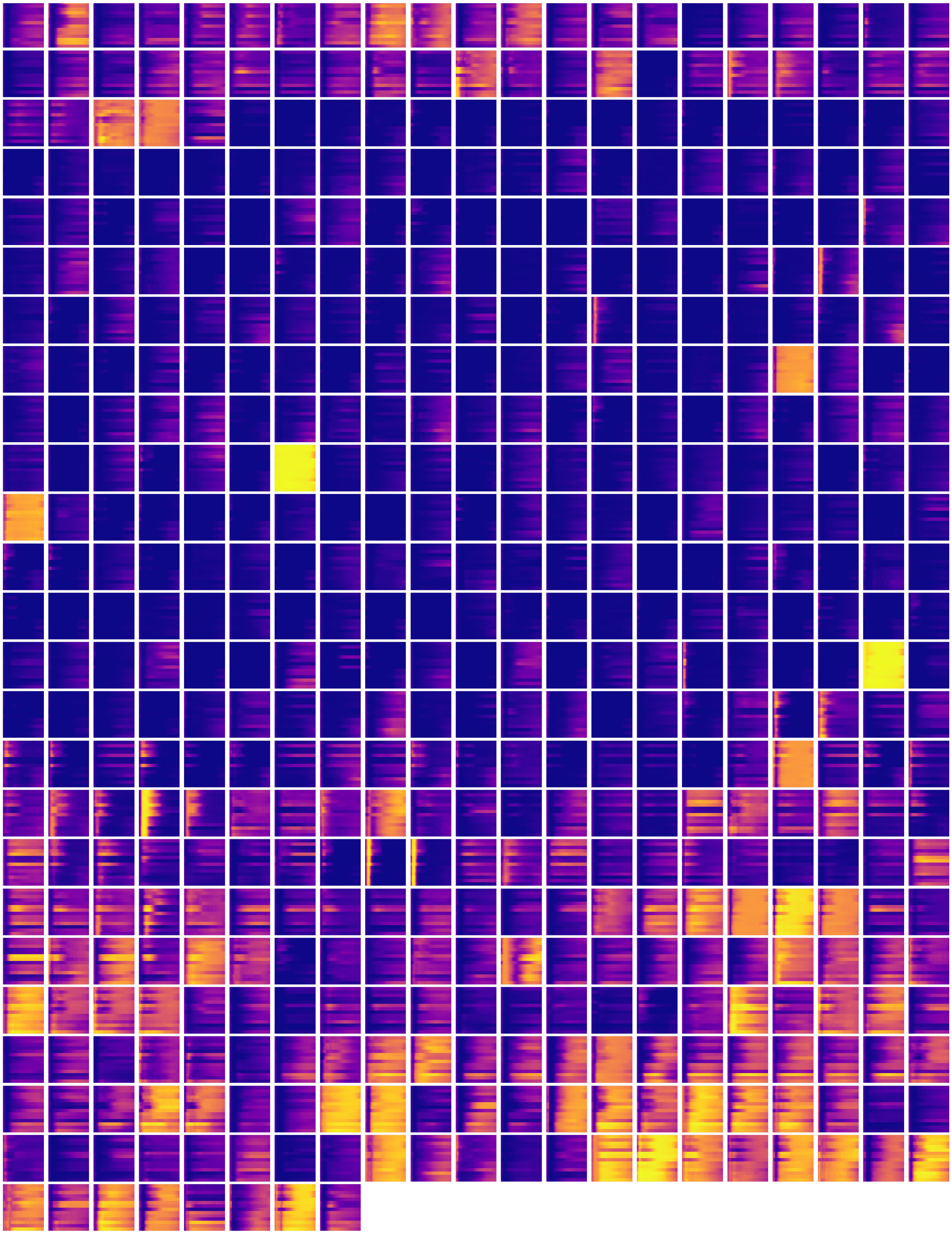


Fig. S2.3. Activations over time of all hidden units to features. Each panel displays one unit's response to each feature over a period of 350ms. Units are ordered according to similarity in phoneme response profiles (i.e., the same order as in Fig. S3). To see the order of features, see Fig. S5.

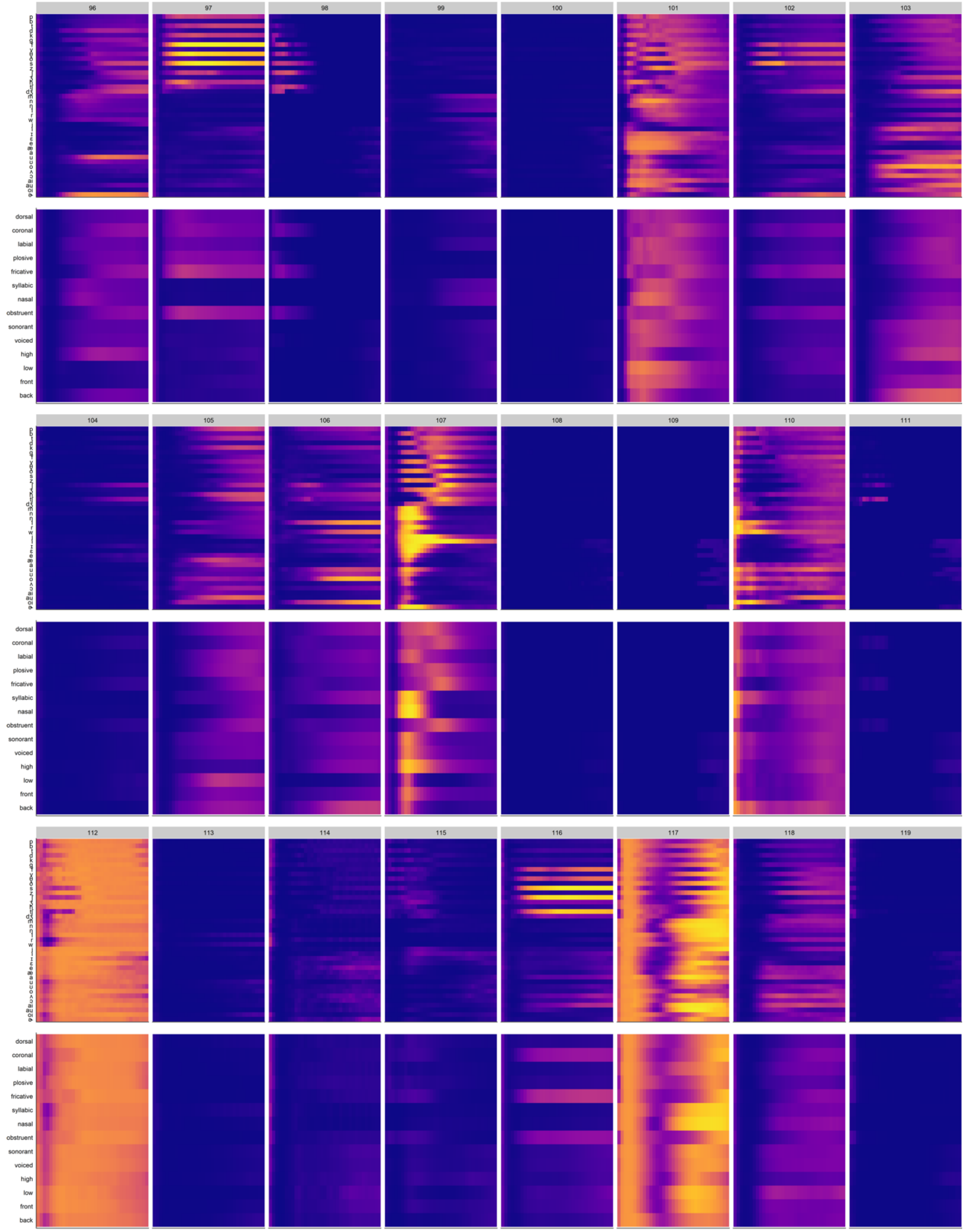


Fig. S2.4. Examples illustrating the tendency for greater selectivity to phonemes than features. Responses of units 96-119 to phonemes (rows 1, 3, and 5) and features (rows 2, 4, and 6). A moderate tendency for sharper, more selective responses to phonemes than features is apparent.