EARSHOT:
A minimal network model of human speech recognition that operates on real speech

James S. Magnuson (james.magnuson@uconn.edu)

Heejo You (hee_jo.you@uconn.edu)

Jay Rueckl (jay.rueckl@uconn.edu)

Paul Allopenna (paul.allopenna@uconn.edu)

Monica Li (monica.li@uconn.edu)

Sahil Luthra (sahil.luthra@uconn.edu)

Rachael Steiner (rachael.steiner@uconn.edu)
Psychological Sciences & CT Institute for the Brain and Cognitive Sciences, U. Connecticut, Storrs, CT 06269-1020

Hosung Nam (nam@haskins.yale.edu)
Korea University, Seoul, Korea, and Haskins Laboratories, New Haven, CT 06511

Monty Escabi (monty.escabi@uconn.edu)
Biomedical Engineering & Psychological Sciences, University of Connecticut, Storrs, CT 06269-3247

Kevin Brown (kevin.brown@oregonstate.edu)
Depts. of Pharmaceutical Sciences and Chemical, Biological, and Environmental Engineering, Oregon State University

Rachel Theodore (rachel.theodore@uconn.edu)

Nicholas Monto (Nicholas.monto@uconn.edu)
Speech, Language & Hearing Sciences, University of Connecticut, Storrs, CT 06269

Abstract

Despite the *lack of invariance problem* (the many-to-many mapping between acoustics and percepts), we experience *phonetic constancy* and typically perceive what a speaker intends. Models of human speech recognition have side-stepped this problem, working with abstract, idealized inputs and deferring the challenge of working with real speech. In contrast, automatic speech recognition powered by *deep learning* networks have allowed robust, real-world speech recognition. 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 human speech recognition. We developed a simple network that borrows one element from automatic speech recognition (*long short-term memory* nodes, which provide dynamic memory for short and long spans). This allows the network to learn to map real speech from multiple talkers to semantic targets with high accuracy. 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 human speech recognition.

Keywords: spoken word recognition; computational models; neural networks; deep learning

Introduction

Human speech recognition (HSR) poses some of the greatest unsolved scientific challenges in the cognitive and neural sciences. Despite a many-to-many mapping between acoustic patterns and percepts (for now, let us assume percepts are phonemes, i.e., consonants and vowels), listeners experience *phonetic constancy*: we hear what the speaker intends even though the same acoustic pattern can cue different phonemes depending on context, and different patterns can cue the same phoneme. This challenge is the *lack of invariance problem*.

Many factors complicate the acoustic-perceptual mapping: (a) coarticulation (temporal and articulatory overlap of phonemes in series; Liberman et al., 1967), (b) lack of robust boundaries between phonemes or words (Cole & Jakimik, 1980), and (c) shifts in the mapping due to variation in speaking rate (Miller & Baer, 1983), talker characteristics (Joos, 1948; Peterson & Barney, 1952), phonetic context (Liberman et al., 1967), coarticulation (Liberman et al., 1952), and novelty of message content (Fowler & Hosum, 1987). Similar problems are found in other perceptual domains (e.g., visual objects must be recognized despite variation in size, rotation, and illumination; DiCarlo & Cox, 2007). However, the temporal and transient nature of speech compounds the challenge.

Deep vs. minimal networks for speech recognition One might suppose that the lack of invariance problem has been solved in contemporary automatic speech recognition (ASR) systems, such as those used daily by billions of smartphone users. The deep-learning neural network models underlying the best ASR (Hinton et al., 2012) provide robust real-world application but little guidance for theories of HSR. Deep nets for ASR require many complex and richly connected layers, as well as complex, carefully engineered training regimens.

That said, researchers interested in HSR have developed less complex deep networks with the aim of illuminating possible mechanisms supporting audition and HSR. Nagamine et al. (2015), for example, examined hidden units of a 5-layer network trained explicitly on phoneme recognition and observed responses strikingly similar to phonetically-structured responses in human superior temporal gyrus (Mesgarani et al., 2014). Kell et al. (2018) used a deep network to achieve human-like accuracy on two unusual tasks: (1) recognizing the word at the *center* of a two second sample of speech and (2) musical genre identification. Their network had many layers and required complex training. The first 7 layers were shared for speech and music, but then it branched into specialized speech and music pathways (with 5 additional layers). The model surpassed standard spectrotemporal filter models of auditory cortex in predicting human cortical responses to natural sounds (measured with fMRI. Kell et al. suggested that deep networks might provide the only computational approach able to achieve human-like performance for natural stimuli.

We optimistically disagree. Our aim is to develop maximally simple (minimal) models of HSR. Theoretical progress will be difficult if our models approach the complexity of their biological target (the neural basis for HSR). At the same time, we aim to grapple with details that have been left out of deep learning models of auditory perception. First, several models have achieved high accuracy by side-stepping the temporal nature of speech (e.g., by treating an utterance or sound as a static image, with time as one axis) rather than as a time series. Furthermore, such models have not addressed the kinds of human data of greatest interest to psycholinguists who study human spoken word recognition, such as the time course of lexical activation and competition (Allopenna et al., 1998).

Simpler shallow computational models have been applied to grappled with over-time inputs and time course of lexical competition, but with two different limitations: (1) they do not use real speech as input (instead using, for example, |abstract distributed phonetic features over time (TRACE: McClelland & Elman, 1986) or human diphone confusion probabilities (Shortlist B: Norris & McQueen, 2008); (2) they tend not to address learning. Models developed since the mid 1980s have either adopted these simplifications in order to address the time course of spoken word recognition with large vocabularies, or have strived for greater realism but in small-inventory models (e.g., Grossberg et al., 1997), or have attempted to incorporate ASR approaches into cognitive models of spoken word recognition (e.g., Scharenborg, 2010; Scharenborg et al., 2005). Such approaches have led to genuine insights, but the models tend to have low accuracy, limited empirical coverage, or both.

Minimal models from long short-term memory nodes
Our aim is to develop a *minimal* cognitive model of HSR that could *learn to map over-time speech to semantics, without explicit phonetic training,* that remains simple enough to generate hypotheses for mechanisms that could support HSR. However, current network-based cognitive models of HSR do not appear adequate for processing real speech.

Thus, we examined a variety of network architectures and elements used in network models used for ASR. We found that a two-layer recurrent network provides the needed power for our goal domain if its hidden units are *long short-term memory* (LSTM) nodes (Hochreiter & Schmidhuber, 1997). LSTM nodes add 3 internal gates and a memory cell that allow nodes to develop sensitivity to information over long time scales, mitigating the *vanishing gradient problem* (Hochreiter et al., 2001). In the following sections, we describe a new neural network model of HSR, EARSHOT (*Emulation of Auditory Recognition of Speech by Humans Over Time*), that we believe approaches the minimal complexity required to map real speech to semantics.

Methods

Network structure and parameters

The EARSHOT network is schematized in Fig. 1. Its 256 input units are fully connected to 512 LSTM hidden units. The hidden layer is fully recurrent (i.e., every unit has a connection to every other unit). A *tanh* activation function is applied to hidden outputs. The hidden units are fully connected to 300 output units. High accuracy on our task (described below) required ~500 hidden units (performance is not improved by increasing to 750 or 1000 hidden nodes).

Materials

We pseudo-randomly selected 1000 words from a list of uninflected English words, with the constraints that (a) word length varied from 1-8 phonemes (mean = 5.5) and (b) every phoneme had to occur in at least 10 words. We created speech files for each of the 1000 words pronounced by 10 talkers in the Apple text-to-speech application, *say* (5 females [Agnes, Kathy, Princess, Vicki, Victoria] and 5 males [Alex, Bruce, Fred, Junior, Ralph]). Mean duration was 659 ms (range: 289-1121 ms). We also created 360 consonant-vowel (CV) and VC syllables for testing purposes (using 15 vowels and 24 consonants). Sound files were converted to spectrographic representations with 256 channels in 10 ms steps with sampling rate of 8000 hz.

We created random sparse vectors for each word as a proxy for semantic representations. Vectors had 300 elements, with 10 "on" (set to 1, others set to 0). This common simplification is considered acceptable given the largely arbitrary mapping from form to menaing (e.g., Lazlo & Plaut, 2012).

Training method

We trained 10 instantiations of EARSHOT. For each model, a different one of the 10 talkers was excluded from training (reserved to test generalization to a novel talker). We excluded 100 different randomly selected words from each trained-on talker (reserved to test generalization to unseen items from trained-on talkers). So for each model, the training set was 8100 input-output patterns, with all 10,000 pairs included for testing.

Each training epoch included one presentation of each of the 8100 training items in random order with no pause or other indication of word boundaries. The target pattern was the semantic vector for the current word, and it was compared to the output at each time step. To enhance learning, we used *minibatch gradient descent*, *Noam decay*, and *Adam optimizing* (Vaswani et al., 2017). Full details are available in a longer preprint (Magnuson et al., 2018). Connections were trained using backpropagation through time (Werbos, 1988). Training accuracy largely plateaued by 8000 epochs. We then resumed training with formerly excluded talkers included. The logic was that when humans encounter new talkers, we presumably learn to adapt to them by learning any idiosyncratic aspects of their acoustics-to-percepts mapping (e.g., by using lexical hypotheses to guide learning). In simple tests of generalization, the model cannot learn. We continued training for another 2000 epochs (8001-10,000).

Testing method Every 1000 epochs, models were tested with all 10,000 words (including excluded words and talkers). Successful recognition was operationalized as the output vector's cosine similarity to the target exceed any other item's cosine similarity to the output by at least 0.05 for at least 100 ms, and subsequently, no item could exceed the target's cosine similarity to the output before word offset.

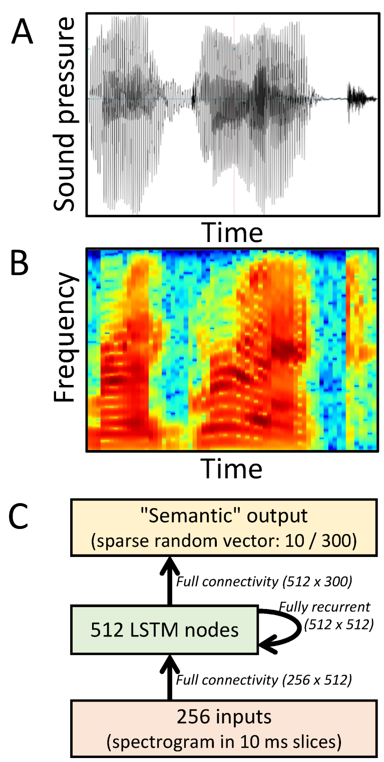

Figure 1. Model input and structure. (A) Audio files are converted to spectrograms **(B),** with 256 channels (rows) in 10 ms steps (columns). Color indicates amplitude (blue-red indicates low-high). **(C)**. The model is a standard recurrent network, except "long short-term memory" nodes are used in the hidden layer, allowing it to become sensitive to multiple temporal grains.

Replicability We trained all 10 models 3 times; only minor variations were observed between iterations. We present results from the first run of each model in this report.

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 required approximately 10 hours for training.

Alternative architectures In developing EARSHOT, we explored dozens of combinations of candidate architectures and model elements. We limited networks to 2 layers of forward connections (inputs→hidden→outputs). We varied 3 aspects of models: number of hidden units (typically from 100 to 1000 nodes before rejecting a model if accuracy plateaued 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 explored spectrograms at various resolutions, Mel Frequency Cepstral Coefficients (MFCCs), and cochleagrams. Most combinations failed to achieve high accuracy. Aside from the model reported here, the only combinations that achieved greater than 90% accuracy was an MFCC-based model that failed to show human-like time course despite high accuracy. Note that this does not mean that only a single set of parameters worked; the model described above begins achieving high accuracy with more than 256 LSTM hidden units, and *maximal* accuracy with ~500 or more LSTM nodes.

Results

Accuracy and time course

We present key model behavior results in Fig. 2. Mean accuracy on training items was quite high (88%) after 8000 epochs. Accuracy was 67% for excluded words from trained-on talkers but only 33% for excluded talkers, with a very wide range (4% to 78%). When training resumed with all talkers and items included, performance improved rapidly (to 89% and 86% for excluded words and talkers, respectively, 93% for previously trained-on items).

Next, we consider the challenge of simulating the time course of HSR (Allopenna et al., 1998). This is a central behavioral target in psycholinguistics but has not been addressed in deep learning models of speech (Kell et al., 2018; Nagamine et al., 2015). Our minimal model exhibits the correct qualitative pattern for phonological competition (Fig. 2B) and makes predictions similar to the gold-standard of HSR, TRACE (Fig. 2C; McClelland & Elman, 1986). This similarity might suggest that any model that can map speech inputs to word-form outputs (as in TRACE) or semantic outputs (EARSHOT) would exhibit this human-like time course. However, this is not the case. As we noted above, an MFCC-based model was able to achieve high accuracy, but could not simulate the patterns seen in Figs. 2B and 2C.

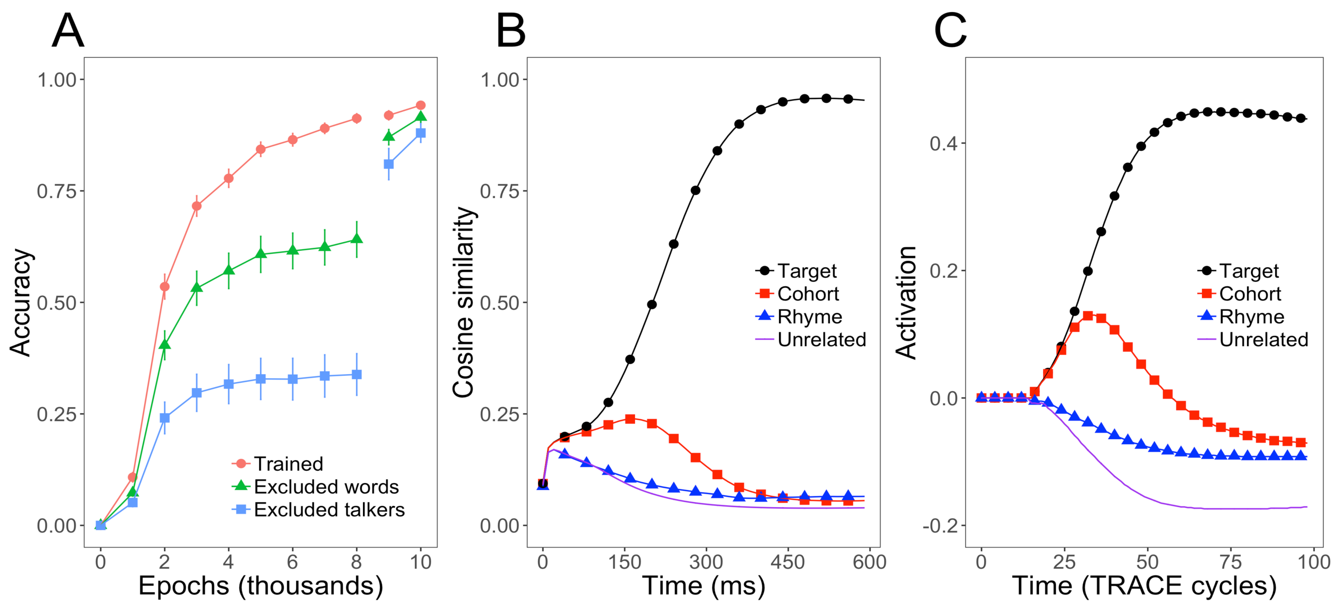

Figure 2. Model performance. (A) Accuracy by epoch averaged over 10 models. When training resumed with all items included (epochs 8001-10,000), high performance was achieved quickly for all talkers. **(B)** Competition time course (correct trials), for 2 criterial competitor types. For a target (e.g., CAT), "Cohort" represents mean cosine similarity for words overlapping in the first 2 phonemes (CAN, CASTLE). "Rhyme" words rhyme with the target (BAT, SAT). "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. Crucially, EARSHOT displays the same rank ordering and similar timing for competitor types as the gold-standard TRACE model.

Unpacking the model

How can we determine how the model works, and how can its mechanisms guide theories of HSR (both cognitive and neural)? To address this, we borrowed an approach that Mesgarani et al. (2014) developed for decoding human electrocorticography data. We presented the model with all possible CV and VC vowels, and examined the responses of every hidden unit over time. For every hidden unit paired with every phoneme, we calculated a *Phonetic Sensitivity Index* (PSI). For example, for unit 239, we would note its mean activation in response to /b/ from the onset of /b/ to 100 ms later. We then subtract unit 239's response to each other phoneme in turn from its response to /b/. When the difference is > 0.3, the PSI for {239, /b/} would be incremented. We repeat this for all 39 phonemes. The maximum PSI for a unit-phoneme pair would be 38 (indicating a unit that responded more strongly to that phoneme than to any other).

We calculated the PSI for all unit-phoneme pairs. Then, we subjected the resulting unit-by-phoneme matrix to hierarchical clustering (Fig. 3). This allows us to ask whether phonetic structure emerges as the model learns to map speech to semantics, even though no explicit information about phonetic features or phonemes is given in training.

About 50% of hidden units exhibited structured responses in the SI time window (20% of electrodes examined by Mesgarani et al. [2014] met their inclusion criteria). The hierarchically clustered PSI solution bears remarkable resemblance to that derived from electrodes in human superior temporal gyrus, with selective responses for phonetically similar phonemes.

The PSI analysis reveals an internal phonetic code that emerges over training. However, hidden units have more complex dynamics than are revealed by the PSIs. Profiles include strong responses at phoneme onset, but also delayed and sustained responses (see Magnuson et al., 2018). In future work, we will explore how the full combination of response profiles support EARSHOT's robust performance. It is also possible that the variety of response profiles observed in the model could be the basis for hypotheses regarding candidate response profiles that might occur in human cortical recordings.

Discussion

Decades after the *lack of invariance problem* – the absence of invariant cues to speech sounds (e.g., Joos, 1948; Liberman et al., 1952; Peterson & Barney, 1952) – was first described, speech science offers limited explanations for human phonetic constancy. A significant obstacle is that computational models of HSR have side-stepped the problem of working directly on the speech signal. Instead, models have focused on the challenges inherent in spoken word recognition beyond initial encoding, using simplified inputs such as gradient phonetic features (McClelland & Elman, 1986), phonemes (Hannagan et al., 2013; You & Magnuson, 2018), or human phoneme confusion probabilities (Norris & McQueen, 2008) instead of real speech. Ironically, simplifying assumptions can *complicate* theoretical challenges (Magnuson, 2008) by masking constraints (in this case, e.g., prosodic cues to phoneme identity or word length).

Simplifying assumptions about input were motivated by complexity concerns. As McClelland and Elman (1986) argued, models aimed at guiding psychological theory must prioritize psychological over computational adequacy, favoring simplicity and understandability over full, end-to-end modeling. A comprehensive and robust model that is itself too complex to understand offers little guidance to HSR theories.

In developing EARSHOT, our aim was to maximally conserve psychological adequacy (i.e., simplicity) in a model that takes real speech as input. Borrowing one tool from ASR – long short-term memory (LSTM) nodes (Hochreiter & Schmidhuber, 1997) – allowed a *shallow* recurrent network to *learn* to map from *speech* to pseudo-semantics while exhibiting human-like dynamics of lexical activation and competition (similar to TRACE; Fig. 2). Generalization (on items from trained-on talkers that were not included in training, as well as talkers wholly excluded from training) was fairly low and quite variable. On the one hand, this represents a major advance, since there simply are no other *cognitive* models of HSR that operate on real speech. This is the first time such a simple model has been applied to problems entailed by doing so (talker variability, etc.). On the other hand, relatively low and variable generalization may reflect the degree to which the model *memorizes* training patterns. In ongoing work, we are exploring the use of more variable inputs, but ultimately, we must move to using open-ended training items produced by natural talkers.

Another contrast with other models of HSR is that EARSHOT is a learning model. Although we have thus far used an unnatural training regimen, EARSHOT allows the exploration of more naturalistic learning.

Admittedly, *how* the model succeeds in learning to map speech to semantics is not yet completely clear. By importing techniques from human electrocorticography (Mesgarani et al., 2014), we were able to track responses of hidden units to specific phonemes (Fig. 3) and observe the model's emergent sensitivity to phonetic structure. It develops this sensitivity without any explicit training or information about phonetic features or phonemes. Deeper understanding will require more complex analyses of not just hidden units, but also output units and weight layers.

However, the preliminary similarity of EARSHOT's hidden unit responses to responses in human superior temporal cortex (Mesgarani et al., 2014) suggests that our approach has potential for new means of developing cognitive models that are potentially linkable to the neural substrates supporting HSR. Speculatively, we would propose that response profiles observed in hidden units in a model like

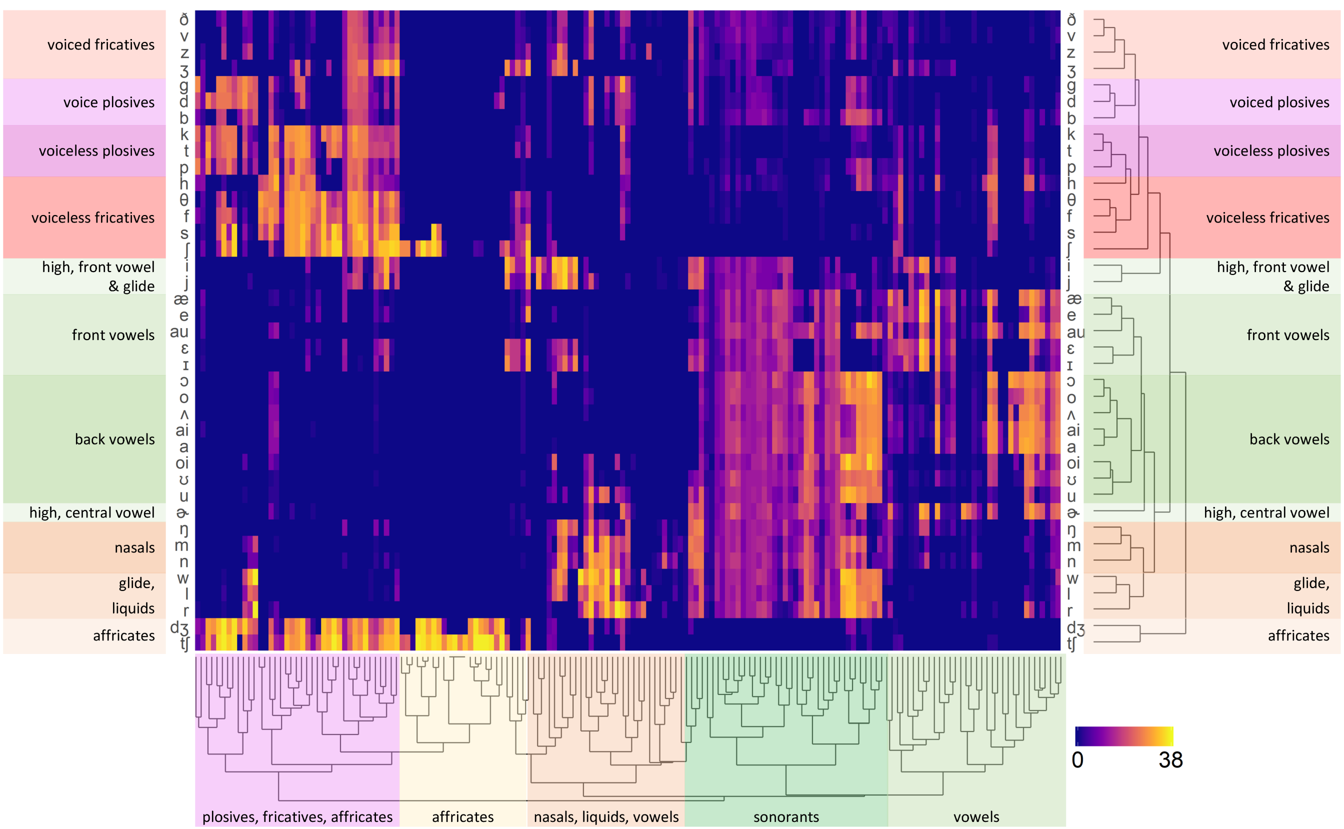

Fig. 3. Phonetic sensitivity revealed by hierarchical clustering. *Phonetic Sensitivity Index (PSI)* based on hidden unit (x-axis) responses in the presence of specific phonemes. For every hidden unit-phoneme pair, PSI was incremented for every phoneme to which the hidden unit responded substantially *more weakly* (yellow indicates high selectivity, with maximum PSI of 38, given 39 phonemes). 246 HUs showing selective responses are included. We used hierarchical clustering to sort both axes, revealing substantial structure in hidden unit responses.

EARSHOT could provide hypotheses for human cortical responses.

In conclusion, EARSHOT may provide a first step towards a comprehensive solution to the overarching challenge for theories and models of HSR – the *lack-of-invariance problem.* Simulations on previously out-of-reach topics (talker and rate variability, etc.) can be conducted with the *same materials* presented to human listeners. Our aim in this brief report is to provide a snapshot of the basic properties of EARSHOT. In a longer subsequent report, we will describe our ongoing work to more fully assess the capabilities of the model.

Acknowledgments

Supported by NSF 1754284, NSF IGERT 1144399, & NSF NRT 1747486 (PI: J.S.M.); NICHD P01 HD0001994 (PI: J.R.); and NSF 1827591 (PI: R.M.T.).

References

Allopenna, P.D., Magnuson, J.S., Tanenhaus, MK (1998). Tracking the time course 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). Mahweh, 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 & 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. doi: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*, *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. Kramer & J.F. Kolen (Eds.) *A Field Guide to Dynamical Recurrent Neural Networks* (pp. 237-374). 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.

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). Mahweh, NJ: Erlbaum Associates.

Magnuson, J.S., You, H., Nam, H., Allopenna, P.D., Brown, K., Escabi, M., Theodore, R.M., Luthra, S., Li, M., & Rueckl, J. (2018, December 13). EARSHOT: A minimal neural network model of incremental human speech recognition. https://doi.org/10.31234/osf.io/h7a4n

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.

Nagamine, T., Seltzer, M.L. & Mesgarani N. (2015). Exploring how deep neural networks form phonemic categories. Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 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*. doi:10.3758/s13428-017-1012-5.