Deep neural networks reveal context-sensitive speech encoding in single neurons of human cortex

Shailee Jain, **Rujul Gandhi**, **Matthew K. Leonard**, **Edward F. Chang**
Department of Neurological Surgery, University of California San Francisco
{firstname}.{lastname}@ucsf.edu

Abstract

Speech perception relies on continuously tracking information at different temporal scales and integrating it with past context. While prior studies have established that the human superior temporal gyrus (STG) encodes many different speech features—from acoustic-phonetic content to pitch changes and word surpisal—we are yet to understand the neural mechanisms of contextual integration. Here we used deep neural networks to investigate context-sensitive speech representations in hundreds of single neurons in STG, recorded using Neuropixels probes. Through this, we established that STG neurons show a broad diversity of context-sensitivity, independent of the speech features they are tuned to. We then used population-level decoding to investigate the role of this property in tracking spectrotemporal information, and found that neurons sensitive to long contexts faithfully represented speech over timescales consistent with higher-order word and phrase-level information (∼1sec). Our results suggest that heterogeneity in both context-sensitivity and speech feature tuning enable the human STG to track multiple, hierarchical levels of spoken language representations.

Keywords: Speech perception; Single neurons

Speech is a dynamic acoustic signal that requires listeners to continuously extract and integrate information at multiple timescales. Prior studies have shown that local neural populations (Bhaya-Grossman & Chang, 2022; Yi, Leonard, & Chang, 2019) and single neurons in human superior temporal gyrus (STG) (Leonard et al., 2023) encode many different phonological and linguistic speech features, including acoustic-phonetic features, prosodic cues like pitch and intensity, and word-level surprisal. This work has uncovered the neural encoding of perceptually-relevant features, but we are yet to understand how these elements are represented in the context of each other and surrounding speech inputs. Recent advances in human single neuron recording techniques and powerful speech recognition models make it possible to address key questions about the neuronal encoding of naturalistic speech.

Here, we used high-density Neuropixels probes to record the activity of hundreds of neurons across all layers of STG (Leonard et al., 2023) while participants listened to ∼200 English sentences from the TIMIT corpus (2±0.04sec duration) (Garofolo et al., 1993) during awake brain mapping. We hypothesized that beyond detecting the spectrotemporal content of speech, STG neurons integrate and represent higher-order speech features by encoding contextual information. We tested this by extracting continuous, contextual features from transformer-based deep neural networks (DNNs), which can capture relationships across multiple timescales in the input. We specifically hypothesized that individual neurons would be sensitive to particular context lengths, giving neural populations in STG access to multiple, hierarchical representations of spoken language.

DNN encoding models of neuronal spiking

We built encoding models that learned to predict the spiking activity of each neuron using hidden states of a pretrained DNN for the given speech stimulus (Fig. 2). To investigate the degree of context-sensitivity across neurons, we varied both the amount of prior context available to the DNN (20-1000ms) and the DNN layer from which states were extracted (Layers 1, 5, 9, 12) (Jain & Huth, 2018). We report results using two different DNNs: HuBERT-base (Hsu et al., 2021) and WavLM-base (Chen et al., 2022) (7×4=28 feature spaces per DNN). Encoding models were fit using cross-validated ridge regression and prediction performance was evaluated by computing the linear correlation (r) between true and predicted spiking activity of each neuron for a held-out set.

We hypothesized that the ability of the DNNs to capture context-sensitive speech representations could explain neuronal spiking. To test this, we compared DNN encoding performance with models fit using context-independent, linguistically-motivated speech features (e.g. word onsets, instantaneous pitch, phonemic content etc. (Leonard et al., 2023)). DNNs better predicted neuronal firing in 73% of neurons (3.5±5% more variance explained) and these improvements were independent of the speech feature a given neuron was tuned to.

To test whether better encoding performance was simply a function of features derived from a nonlinear model, we compared DNNs to a different encoding model that learns nonlinear spectrotemporal receptive fields without explicitly accounting for context (Keshishian et al., 2020). DNNs largely outperformed these models, with as much as 20% increase in explained variance. This shows that the DNNs' ability to integrate speech information over prior context is useful for modeling STG neurons. Finally, to evaluate the importance of speech representations learned through DNN pretraining, we trained encoding models using randomly-initialized DNNs. The pretrained DNNs outperformed these models in over 83% of neurons, demonstrating that the model architecture or high dimensionality is not enough to explain the improvement.

These effects were observed across all Neuropixels recording sites, with some site-specific differences in the degree to which the DNNs explained more variance than the alternatives. This suggests that in general, neurons throughout the STG and across cortical layers are characterized by nonlinear, context-dependent encoding.

Variable context-sensitivity of single neurons

To investigate the degree of context-sensitivity in each neuron, we analyzed how its encoding performance varied with the amount of prior context and the DNN layer. We normalized each neuron's performance across the 28 DNN features spaces and applied principal components analysis (PCA; Fig. 1). PC 1 captured 25.5% variance across neurons, differentiating between neurons that improve performance with more context up to 500ms, and those that either decrease or show no change. Thus, sensitivity to the DNN context

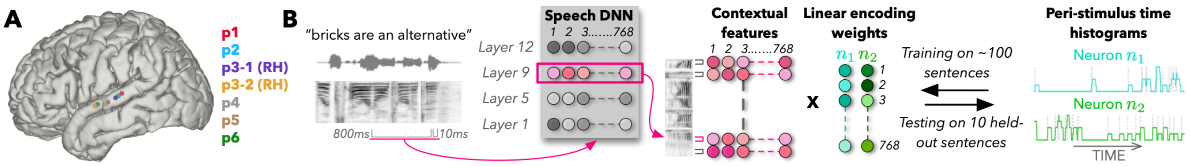

Figure 1: (A) 7 Neuropixels recording sites on an average cortical surface. (B) Single neuron encoding model where 768-D features are extracted from a speech DNN for every 10ms speech segment with 800ms of context.

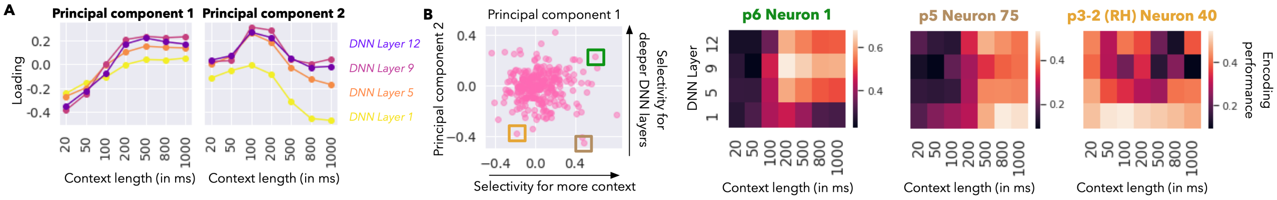

Figure 2: (A) First 2 principal components of encoding patterns. (B) Distribution of PC 1-2 across all Neuropixels sites with example neurons showing diverse context and DNN layer selectivity.

length was an important source of variation across neurons. PC 2 captured 11% variance, and differentiated between neurons that were better predicted by deeper DNN layers and those with either no or shallow layer selectivity. Prior work has shown that shallow layers linearly encode acoustic features like spectral modulations and envelope magnitude, while deeper layers capture higher-order features like acoustic-phonetic content (Vaidya, Jain, & Huth, 2022; Pasad, Chien, Settle, & Livescu, 2024). The distribution of PC2 thus suggests that STG neurons have substantial variability in feature tuning.

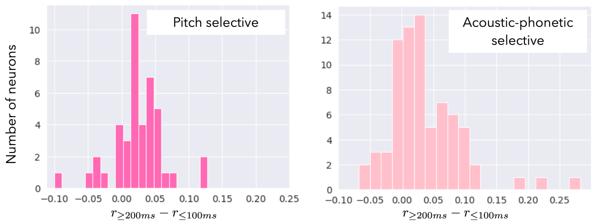

Figure 3: Neurons show variable improvements in encoding with context, even with similar linguistic tuning.

Next, we asked whether context-sensitivity is related to speech feature tuning. For example, it could be the case that longer context-sensitivity is beneficial for neurons that encode features which unfold over longer timescales (e.g., pitch) compared to shorter timescales (e.g., acoustic-phonetic features). However, we found substantial diversity in context-sensitivity even across neurons tuned to the same speech feature (Fig. 3). This suggests that the linguistic tuning and context-sensitivity of STG neurons play different roles in the neural code for speech. The context-sensitivity was also independent of DNN layer selectivity, cortical depth and putative cell type.

Decoding context from population activity

Lastly, given the heterogeneity in both tuning and context-sensitivity within and across sites, we hypothesized that population-level neural activity could capture an integrated representation of the speech input at the level of perceptually-meaningful units like words and phrases. We grouped neurons in either the top or bottom 15th percentile of the first PC of encoding patterns, ensuring matched distributions of second PC and encoding performance for spectrogram features). Then, we evaluated how well each group could linearly decode spectrogram content up to 900ms in the past. Models were trained using cross-validated ridge regression. Performance was measured as linear correlation between true and predicted spectrograms, averaged across 10 random test splits. We found that neurons and columns with long context-sensitivity faithfully represented speech over timescales consistent with higher-order word and phrase-level information (∼1sec) (Fig. 4).

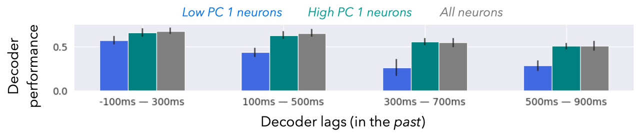

Figure 4: Unlike the high PC 1 group, the low PC1 group quickly degraded with increasing lags in the past. Errorbars represent standard deviation across random dataset splits.

Together, our results suggest that single neurons in the human STG encode speech in a context-dependent manner. The substantial heterogeneity of neurons in both feature tuning and context-sensitivity likely enables local populations in this brain region to track multiple levels of speech content rapidly and in parallel.

References

Bhaya-Grossman, I., & Chang, E. F. (2022, January). Speech Computations of the Human Superior Temporal Gyrus. *Annual review of psychology*, *73*, 79–102. Retrieved 2023-06-14, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9447996/ doi: 10.1146/annurev-psych-022321-035256

Chen, Z., Chen, S., Wu, Y., Qian, Y., Wang, C., Liu, S., . . . Zeng, M. (2022, January). Large-scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification. *arXiv:2110.05777 [cs, eess]*. Retrieved 2022-02-28, from http://arxiv.org/abs/2110.05777 (arXiv: 2110.05777)

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). *DARPA TIMIT:: acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1* (Tech. Rep. No. NIST IR 4930). Gaithersburg, MD: National Institute of Standards and Technology. Retrieved 2022-04-20, from https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir4930.pdf doi: 10.6028/NIST.IR.4930

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021, June). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units.* arXiv. Retrieved 2023-06-15, from http://arxiv.org/abs/2106.07447 (arXiv:2106.07447 [cs, eess]) doi: 10.48550/arXiv.2106.07447

Jain, S., & Huth, A. (2018). Incorporating Context into Language Encoding Models for fMRI. In *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc. Retrieved 2022-01-24, from https://proceedings.neurips.cc/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html

Keshishian, M., Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2020, June). Estimating and interpreting nonlinear receptive field of sensory neural responses with deep neural network models. *eLife*, *9*, e53445. Retrieved 2023-05-26, from https://doi.org/10.7554/eLife.53445 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.53445

Leonard, M. K., Gwilliams, L., Sellers, K. K., Chung, J. E., Xu, D., Mischler, G., . . . Chang, E. F. (2023, December). Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature*, 1–10. Retrieved 2024-01-28, from https://www.nature.com/articles/s41586-023-06839-2 (Publisher: Nature Publishing Group) doi: 10.1038/s41586-023-06839-2

Pasad, A., Chien, C.-M., Settle, S., & Livescu, K. (2024, January). *What Do Self-Supervised Speech Models Know About Words?* arXiv. Retrieved 2024-04-19, from http://arxiv.org/abs/2307.00162 (arXiv:2307.00162 [cs, eess])

Vaidya, A. R., Jain, S., & Huth, A. (2022, June). Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 21927–21944). PMLR. Retrieved 2022-10-16, from https://proceedings.mlr.press/v162/vaidya22a.html (ISSN: 2640-3498)

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019, June). The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron*, *102*(6), 1096–1110. Retrieved 2022-11-01, from https://linkinghub.elsevier.com/retrieve/pii/S0896627319303800 doi: 10.1016/j.neuron.2019.04.023