Invariant representations of words are rapidly constructed across the auditory cortical hierarchy

Dana Boebinger (dana_boebinger@urmc.rochester.edu)
Department of Biostatistics & Computational Biology, University of Rochester
601 Elmwood Ave., Rochester, NY 14642

Guoyang Liao (guoyang_liao@urmc.rochester.edu)
Department of Biostatistics & Computational Biology, University of Rochester
601 Elmwood Ave., Rochester, NY 14642

Kirill Nourski (kirill-nourski@uiowa.edu)
Department of Neurosurgery & Iowa Neuroscience Institute, The University of Iowa
200 Hawkins Dr. 1815 JCP, Iowa City IA 52242

Matthew Howard (matthew-howard@uiowa.edu)
Department of Neurosurgery & Iowa Neuroscience Institute, The University of Iowa
200 Hawkins Dr. 1823 JPP, Iowa City IA 52242

Christopher Garcia (Christopher-garcia-1@uiowa.edu)
Department of Neurosurgery, The University of Iowa
200 Hawkins Dr. 1815 JCP, Iowa City IA 52242

Thomas Wychowski (thomas_wychowski@urmc.rochester.edu)
Department of Neurology, University of Rochester
601 Elmwood Ave., Rochester, NY 14642

Webster Pilcher (webster_pilcher@urmc.rochester.edu)
Department of Neurosurgery, University of Rochester
2180 South Clinton Ave., Rochester, NY 14618

Sam Norman-Haignere (samuel_norman-haignere@urmc.rochester.edu)
Departments of Biostatistics & Computational Biology, Neuroscience,
Brain & Cognitive Sciences, Biomedical Engineering, University of Rochester
601 Elmwood Ave., Rochester, NY 14642

Abstract:

The fundamental computational challenge of auditory word recognition is that instances of the same word vary enormously in their acoustics. The auditory system is thought to construct representations of sound that are invariant to such acoustic diversity by adaptively selecting and removing spectrotemporal acoustic variation. Yet despite the importance of word recognition, little is known about how invariant representations of words are organized in the human auditory cortex, in part due to the coarse spatiotemporal precision of human neuroimaging methods. Here, we developed a novel paradigm that leverages the spatiotemporal precision of human intracranial recordings to measure the strength and timing of invariant and non-invariant representations across many different words and types of acoustic variation. We show that invariant representations of words emerge rapidly after word onset (within 200 ms), increase substantially in strength across the cortical hierarchy for many different types of acoustic variation, and are delayed by ~30 ms compared with non-invariant representations. We show that these effects cannot be explained by standard spectrotemporal filtering models nor do they require an extended adaptation period. These results indicate that invariant representations of words are computed by fast, hierarchically organized, nonlinear computations that do not depend critically on adaptive spectrotemporal filtering.

Keywords: speech; invariant coding; auditory cortex; intracranial EEG

In real-world environments, the acoustics of a word differ enormously due to factors such as interactions with the environment (e.g., reverb), imperfect speech transmission (i.e., spectral filtering), background noise, and properties of the speaker's voice (e.g., voicing). Successful communication requires a listener's recognition of words to be robust to such variation, which is computationally challenging (Sharpee et al., 2011). Yet, compared with for example object recognition in vision (DiCarlo et al., 2012), much less is known about the neural mechanisms that support invariant word recognition in speech. Psycholinguistic models often ignore the invariance challenge by assuming an invariant input representation (e.g., phonemes) (Gaskell & Marslen-Wilson, 1997; Luce & Pisoni, 1998; McClelland & Elman, 1986; Norris, 1994). Much of the relevant neuroscience research has been conducted in nonhuman animal models (Carruthers et al., 2015; Mesgarani et al., 2014; Moore et al., 2013; Rabinowitz et al., 2013; Schneider & Woolley, 2013), but there is considerable evidence for speech-specific responses in the human brain (Landemard et al., 2021; Overath et al., 2015). Human neuroimaging methods, while useful (Ding & Simon, 2013; Kell & McDermott, 2019), lack the spatiotemporal precision needed to measure rapidly varying speech structures. One prior study measured intracranial responses to speech in the presence of three different background sounds, finding that auditory cortex adaptively suppresses responses to the background sounds over ~500 ms (Khalighinejad et al., 2019). However, it is unclear whether this prolonged adaptation effect is needed to construct invariant representations of words, because no prior studies have systematically measured the strength of invariant word representations across the human auditory cortex.

To address this gap, we used a novel experimental paradigm coupled with spatiotemporally precise human intracranial recordings to measure the strength and timing of invariant word representations across many words and types of acoustic variation. We measured responses from 136 sound-responsive electrodes from 16 neurosurgical patients implanted with stereotactic depth electrodes at the University of Rochester Medical Center and University of Iowa Hospitals and Clinics. We presented spoken sentences with and without acoustic variation (**Fig 1A**, left two columns). Variation included spectral filtering (lowpass, bandpass, or highpass), reverberation (convolution with 12 naturally recorded impulse responses; Traer et al., 2021), background sounds (12 natural backgrounds; 10 dB SNR), and voicing (replacing periodic excitation with noise excitation, simulating whispering). The type of acoustic variation changed every ~4 s. To isolate word representations, we also presented the constituent words in a random order without variation (**Fig 1A,** right column). We then correlated the response timecourses (broadband gamma; 70-140 Hz) across words with and without variation, separately for each time lag relative to word onset, which we term the **invariant correlation** (**IC**) (orange arch, **Fig 1A**). The IC measures the strength of responses, time-locked to word onset, that are consistent across words with different acoustics.

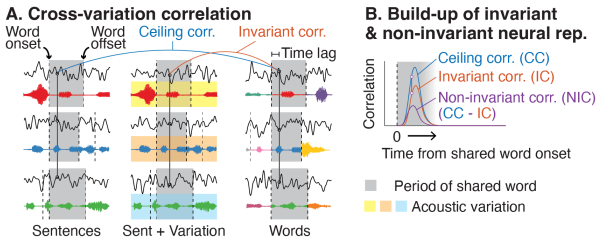

Figure 1. **(A)** Neural response timecourses (black lines) to sentences with (middle) and without (left) acoustic variation, and word sequences (right) without variation. Response timecourses were aligned to word onset, and correlated across words with and without variation as a function of the time (orange arch). The same calculation without variation (blue arch) provided a ceiling correlation. **(B)** Ceiling (blue), invariant (orange), and non-invariant (purple) correlations over time.

We computed a **ceiling correlation** (**CC**) for the IC by performing the same calculation without variation (blue arch, **Fig 1A**). If the response is fully invariant, the IC and CC will be identical, and we therefore used the difference between the IC and CC as a measure of the strength of non-invariant word representations (i.e., responses to words that differ when the acoustics vary), which we term the **non-invariant correlation (NIC)** (**Fig 1B**, purple). To compare the strength of invariance across regions, we compute the ratio between the IC and CC (**invariance ratio**).

Invariance increases in non-primary regions and at later latencies

We plot the IC, NIC and CC for example electrodes in primary and non-primary auditory cortex (**Fig 2A**). The primary electrode showed a rapid increase in both IC and NIC, peaking around ~100 ms after word onset. By comparison, the non-primary electrode showed an initial increase in both the IC and NIC during the first 180 ms, after which the NIC drops to zero and the response becomes nearly fully invariant 200 ms after word onset. This suggests that the strength of invariant word representations increases both across the cortical hierarchy and across time within a single cortical site. To test this hypothesis, we averaged our measures across electrodes based on their distance to primary auditory cortex (Norman-Haignere et al., 2022; Norman-Haignere & McDermott, 2018) (**Fig 2B**). This analysis revealed a substantial increase in invariance in non-primary regions (**Fig 2C**) for all types of acoustic variation tested, though the overall strength of invariance varies across types (**Fig 2D**, left). In all regions we found that the IC peaked within 200 ms, suggesting invariant representations are rapidly computed after word onset, but also observed a consistent delay of ~30 ms in the IC peak relative to the NIC peak (**Fig 2E**, left), suggesting that invariant representations require longer computation time than non-invariant representations.

Invariance changes cannot be explained by spectrotemporal modulation tuning

To determine whether the observed effects could be explained by spectrotemporal tuning, we fit spectrotemporal modulation encoding models to each electrode (Norman-Haignere et al., 2022) and applied our same analyses to the model predictions. We found that the model predictions showed similar levels of invariance across the auditory hierarchy (**Fig 2D**, right), and thus cannot explain the increased invariance strength in non-primary regions. The model predictions show a delayed IC peak relative to the NIC peak (~17 ms), but the effect is approximately half that observed neurally (**Fig 2E**, right).

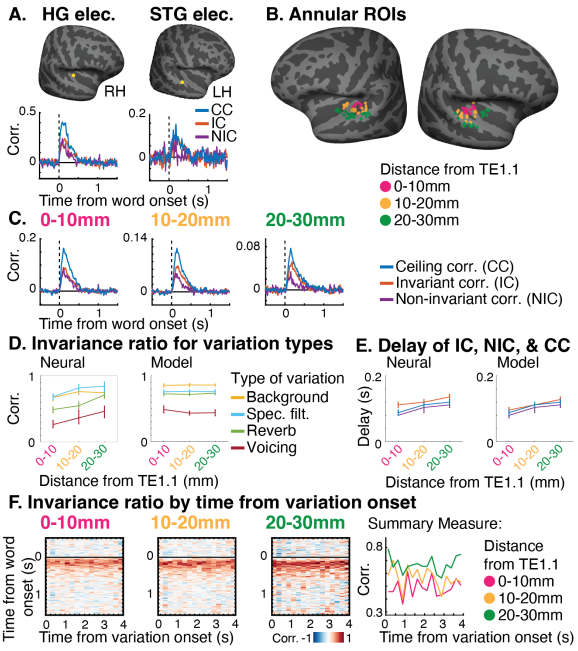

Figure 1. **(A)** Ceiling (CC), invariant (IC), and non-invariant (NIC) correlations for a primary (Heschl's gyrus, HG) and non-primary (superior temporal gyrus, STG) electrode. **(B)** Annular ROIs based on distance from primary auditory cortex (TE1.1). **(C)** CC, IC, and NIC for each ROI. **(D)** Invariance strength for different types of acoustic variation in each ROI. **(E)** Delay (peak time) of IC, NIC, and CC for each ROI. Left and right panels in **(D)** & **(E)** show data and spectrotemporal model predictions, respectively. **(F)** Invariance strength as a function of time from word onset and from the onset of a particular type of acoustic variation. Right: summary of IC peak as a function of time from variation onset.

Invariant word representations do not require an extended adaptation period

To test whether adaptation like that reported in Khalighinejad et al. (2019) might explain our results, we tested whether invariance increases as a function of time from the onset of a new type of acoustic variation, computing our same measures but binning words based on their relative time from variation onset. We did not observe any increase in invariance as a function of time from variation onset in either primary or nonprimary regions (**Fig 2F**). Thus, while longer-term adaptive mechanisms likely exist in auditory cortex, they cannot explain the rapid construction of invariant word representations observed here. Instead, our results suggest that invariant word representations are constructed by fast computations that are hierarchically organized both across auditory cortex and in time.

Acknowledgments

This study was supported by the National Institutes of Health (NIDCD-R00-DC018051 to S.V.N.-H.).

References

Carruthers, I. M., Laplagne, D. A., Jaegle, A., Briguglio, J. J., Mwilambwe-Tshilobo, L., Natan, R. G., & Geffen, M. N. (2015). Emergence of invariant representation of vocalizations in the auditory cortex. *Journal of Neurophysiology*, *114*(5), 2726–2740.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Ding, N., & Simon, J. Z. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. *Journal of Neuroscience*, *33*(13), 5728–5735.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5–6), 613–656.

Kell, A. J. E., & McDermott, J. H. (2019). Invariance to background noise as a signature of non-primary auditory cortex. *Nature Communications*, *10*(1), 1–11.

Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Adaptation of the human auditory cortex to changing background noise. *Nature Communications*, *10*(1), Article 1.

Landemard, A., Bimbard, C., Demené, C., Shamma, S., Norman-Haignere, S., & Boubenec, Y. (2021). Distinct higher-order representations of natural sounds in human and ferret auditory cortex. *eLife*, *10*, e65566.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, *19*(1), 1.

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

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences*, *111*(18).

Moore, R. C., Lee, T., & Theunissen, F. E. (2013). Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise. *PLOS Computational Biology*, *9*(3), e1002942.

Norman-Haignere, S. V., Long, L. K., Devinsky, O., Doyle, W., Irobunda, I., Merricks, E. M., Feldstein, N. A., McKhann, G. M., Schevon, C. A., Flinker, A., & Mesgarani, N. (2022). Multiscale temporal integration organizes hierarchical computation in human auditory cortex. *Nature Human Behaviour*, *6*(3), 455–469.

Norman-Haignere, S. V., & McDermott, J. H. (2018). Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. *PLoS Biology*, *16*(12).

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Overath, T., Mcdermott, J. H., Zarate, J. M., & Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*(6), 903–911.

Rabinowitz, N. C., Willmore, B. D. B., King, A. J., & Schnupp, J. W. H. (2013). Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. *PLoS Biology*, *11*(11).

Schneider, D. M., & Woolley, S. M. N. (2013). Sparse and Background-Invariant Coding of Vocalizations in Auditory Scenes. *Neuron*, *79*(1), 141–152.

Sharpee, T. O., Atencio, C. A., & Schreiner, C. E. (2011). Hierarchical representations in the auditory cortex. *Current Opinion in Neurobiology*, *21*(5), 761–767.

Traer, J., Norman-Haignere, S. V., & McDermott, J. H. (2021). Causal inference in environmental sound recognition. *Cognition*, *214*(May), 104627.