Disentangling of Spoken Words and Talker Identity in Human Auditory Cortex

Akhil Bandreddi* **(akhil_bandreddi@urmc.rochester.edu)**
Department of Biostatistics & Computational Biology, University of Rochester
601 Elmwood Ave., Rochester, NY 14642

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

David Skrill (david_skrill@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

*Co-first authors

Abstract

Complex natural sounds such as speech contain many different types of information, but recognizing these distinct information sources is computationally challenging because sounds with shared information differ widely in their acoustics. For example, variation across talkers makes it challenging to recognize the identity of a word, while variation in the acoustics of different words makes it challenging to recognize talker identity. How does the human auditory cortex disentangle word identity from talker identity such that each type of information can be coded invariant to acoustic variation all other information sources? To address this question, we measured neural responses to a diverse set of 338 words spoken by 32 different talkers using spatiotemporally precise intracranial recordings from the human auditory cortex. We developed a simple set of model-free experimental metrics for quantifying representational disentangling of word and talker identity, both within individual electrodes as well as across different dimensions of the neural population response. We observed individual electrodes that show a representation of words that is partially robust to acoustic variation in talker identity, but no electrodes or brain regions showed a robust representation of talker identity. However, at the population level, we observed distinct dimensions of the neural response that nearly exclusively reflected either words or talker identity, and were completely invariant to acoustic variation in the non-target dimension. These results suggest that while there is partial specialization for talker-robust word identity in localized brain regions, robust disentangling is accomplished at the population level with distinct representations of words and talker identity mapped to distinct dimensions of the neural code for speech.

Keywords: audition, invariance, speech

Sounds in the natural environment contain many different types of information, but extracting this information from the waveform that reaches the ear is challenging because sounds with shared information vary enormously in their acoustics. For example, the same word spoken by different talkers can vary widely in its acoustic properties, and utterances from the same talker can vary substantially depending on the word spoken. For successful communication, the brain must therefore "disentangle" these different types of information, such that each can be recognized "invariant" to acoustic differences in other dimensions (Figure 1). Representational disentangling has been extensively studied in the visual system (Cavanagh, 1978; DiCarlo & Cox, 2007), but less is known about how the human auditory cortex solves this challenge. Much of the relevant research in the auditory system has been conducted using animal models (Heller, Hamersky, & David, 2024; de la Mora & Toro, 2013; Petkov, Logothetis, & Obleser, 2009) or non-invasive human neuroimaging methods (Anderson, Davis, & Lalor, 2024). But non-human animals lack the speech-specific mechanisms present in the human auditory cortex (Landemard et al., 2021) and non-invasive methods lack the spatiotemporal precision to track rapidly varying speech structures (e.g., words). As a consequence, many open questions remain about how speech information is disentangled in the human auditory cortex.

One hypothesis is that anatomically localized brain regions specialize in coding word identity or talker identity. For example, prior studies have suggested that there are localized neural populations tuned for speech content (e.g,. phonemes) or voice identity (Yi, Leonard, & Chang, 2019). Alternatively, word identity and talker identity may represented by distinct dimensions of the neural code at the population level instead of in localized neural populations. To test these hypotheses, we used spatiotemporally precise intracranial recordings and novel experimental paradigms and methods to study representational disentangling in the human auditory cortex, and computational methods to measure invariance both in individual electrodes and at the population level.

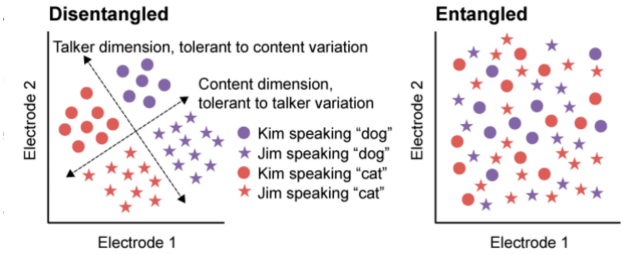

Figure 1: Schematic illustrating an entangled vs. disentangled representation of word and talker identity (see text for details).

Partially invariant representations of word identity in individual electrodes

We measured human cortical responses (broadband gamma power, 70-140 Hz) from 17 epilepsy patients, implanted with stereotactic depth electrodes at the University of Rochester and the University of Iowa. We measured responses to 338 words ("segments") spoken by 32 different talkers and selected 167 electrodes with a reliable response to sound.

If there is a representation of word identity that is invariant across different talkers, then there should be a time lag where the neural response to a word is the same across different talkers (i.e., Jim vs. Kim speaking "dog"). Similarly, a word-invariant talker representation should produce the same response if two different words are spoken by the same talker (i.e., Jim speaking "cat" vs. "dog"). We developed a simple, time-varying metric to quantify this idea. Specifically, we measured the response timecourse aligned to the onset of each word, yielding a **[word, time]** matrix (D). To measure talker-invariant word representations, we computed a second data matrix ($D_{\text{talker-swap}}$), in which we shuffled the **word** axis of the data matrix, such that each row contains the neural response to the same word as D, but spoken by a different talker. We then correlated the columns of D and $D_{\text{talker-swap}}$ separately for each time lag, which we refer to as the **invariant correlation (IC)**. If the response to a word is the same across different

talkers, the IC will be 1 in the absence of noise, while if the response to words is uncorrelated across talkers the IC will be 0, providing a graded metric of invariance. To account for noise, we performed the same calculation using two instances of the identical word measured in independent data, which provides a ceiling for the IC (**ceiling correlation, CC**. To measure word-invariant talker representations, we computed the IC by swapping the word identity, but preserving the talker identity ($D_{\text{word-swap}}$) (i.e., Jim speaking "cat" $\rightarrow$ Jim speaking "dog"). To ensure that our metrics only reflect the target word/talker and not what came before or after, we measured responses to words presented in a natural sentence, as well as a sequence of words in a random order. We computed D using responses to the sentences and $D_{\text{word-swap}}$, $D_{\text{talker-swap}}$ using responses to the random word sequence. Because what came before or after the target word/talker is independent between the two sequences, any reliable non-zero correlation must reflect the target word/talker. Our metrics, therefore, provide a clean and dynamic measure of the strength of invariance over time relative to the onset of a target word or talker.

We found many electrodes for which the IC was near zero for both types of invariance (data not shown due to space limitations), underscoring the computational challenge of representational disentangling. To test if we could reliably find any electrodes with invariant representations of talkers or words, we start by selecting electrodes with an average CC > 0.025 (between 100-500 ms; n ≈ 50) and then selected the top 10% of those electrodes with the highest IC/CC ratio for each metric (averaged between 100-500 ms); we then re-measured our metrics in independent data to avoid bias/circularity. This analysis revealed a small set of electrodes that showed a representation of words that was partially invariant to talkers, as evidenced by an IC that was approximately half that of the CC (Figure 2A). However, word-invariant representations of talkers were very weak, even in those electrodes that were selected to show the strongest invariance (Figure 2B).

Full disentangling in the neural population

We next investigated whether there existed distinct dimensions of the neural population response that show an invariant representations of words and talkers (Fig 1). Specifically, we attempted to learn a projection of the neural response (weighted sum of electrodes) that showed an invariant representation of words or talkers. We found that standard methods such as regularized linear discriminant analysis were ineffective at learning invariant dimensions (data not shown) and thus designed customized loss function explicitly designed to search for invariant response dimensions (equation 1). The first term in the loss is the average squared difference between the IC and CC, which will be 0 if the dimension's response is invariant. Because the IC and CC will also be 0 if the response is unreliable, we included a second term that reflects the overall reliability (r) of the response across two presentations of the same stimulus. The third term is the variance of the response timecourse across all stimuli (v), which

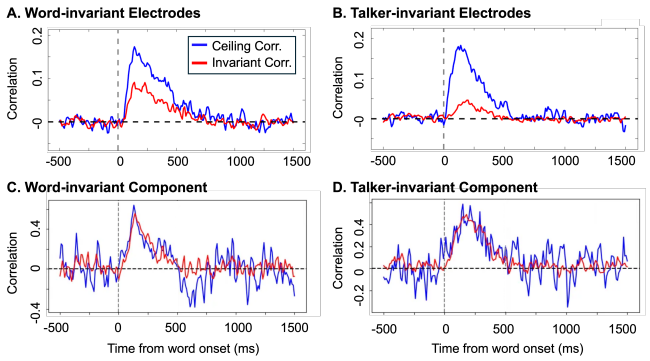

Figure 2: **A.** Invariant correlation (IC, red curve) and ceiling correlation (CC, blue curve) averaged across the most word-invariant electrodes (n = 5). Metrics re-measured in independent data for plotting. **B.** The most talker-invariant electrodes (n = 5) do not show strong invariance (IC $<$ CC). **C & D.** Component analyses pool across electrodes to uncover strongly word-invariant (**C**) and talker-invariant (**D**) responses.

acts a classic regularizer (analogous to an L2 penalty) that encourages the learned dimension to have high variance. The second and third term have hyper-parameters (α, β) that control their influence and we took the log of the first and third term to ensure they did not dominate the loss:

$$\log \sum_t (CC_t - IC_t)^2 - \alpha r - \beta \log v \qquad (1)$$

To prevent overfitting and ensure generalization, we used a train/test/validation split, splitting across the word axis ($\frac{1}{3}$ of words/fold). We used train data to learn the electrode weights, validation data to select the hyperparameters (α, β), and test data to measure the invariance of the learned projection (given the optimized weights and hyperparameters). The hyperparameters were selected using a grid search as those parameters which yielded the highest IC/CC ratio in validation data, excluding parameters where the ratio had high variance (measured using bootstrapping across words).

We found that this analysis was strikingly successful in that we were able to find dimensions that only reflected words or talkers and were completely invariant to the other dimension (Figure 2C and 2D), far surpassing the levels of invariance we observed in single electrodes (Figure 2A and 2B). This indicates that the auditory cortex is able to perform representational disentangling, but that this disentangling is performed at the population level and not in individual electrodes.

Acknowledgments

This study was supported by the National Institutes of Health (R01-DC020960 to S.N.H.).

References

Anderson, A. J., Davis, C., & Lalor, E. C. (2024, 11). Deep-learning models reveal how context and listener attention shape electrophysiological correlates of speech-to-language transformation. *PLOS Computational Biology*, *20*(11), 1-27. Retrieved from https://doi.org/10.1371/journal.pcbi.1012537 doi: 10.1371/journal.pcbi.1012537

Cavanagh, P. (1978). Size and position invariance in the visual system. *Perception*, *7*(2), 167-177. Retrieved from https://doi.org/10.1068/p070167 (PMID: 652474) doi: 10.1068/p070167

de la Mora, D. M., & Toro, J. M. (2013). Rule learning over consonants and vowels in a non-human animal. *Cognition*, *126*(2), 307-312. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027712002260 doi: https://doi.org/10.1016/j.cognition.2012.09.015

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333-341. Retrieved from https://www.sciencedirect.com/science/article/pii/S1364661307001593 doi: https://doi.org/10.1016/j.tics.2007.06.010

Heller, C. R., Hamersky, G. R., & David, S. V. (2024, June). Task-specific invariant representation in auditory cortex. *eLife*. Retrieved from http://dx.doi.org/10.7554/eLife.89936.2 doi: 10.7554/elife.89936.2

Landemard, A., Bimbard, C., Demené, C., Shamma, S., Norman-Haignere, S., & Boubenec, Y. (2021, nov). Distinct higher-order representations of natural sounds in human and ferret auditory cortex. *eLife*, *10*, e65566. Retrieved from https://doi.org/10.7554/eLife.65566 doi: 10.7554/eLife.65566

Petkov, C. I., Logothetis, N. K., & Obleser, J. (2009). Where are the human speech and voice regions, and do other animals have anything like them? *The Neuroscientist*, *15*(5), 419-429. Retrieved from https://doi.org/10.1177/1073858408326430 (PMID: 19516047) doi: 10.1177/1073858408326430

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096-1110. Retrieved from https://www.sciencedirect.com/science/article/pii/S0896627319303800 doi: https://doi.org/10.1016/j.neuron.2019.04.023