Varying sensitivity to vision and language across the STS during naturalistic social perception

Hannah Small (hsmall2@jhu.edu)
Johns Hopkins University, Department of Cognitive Science
Baltimore, MD 21218, US

Haemy Lee Masson (haemy.lee-masson@durham.ac.uk)
Durham University, Department of Psychology
Durham, DH13LE, UK

Ericka Wodka (wodka@kennedykrieger.org)
Kennedy Krieger Institute, Center for Autism Services, Science and Innovation
Baltimore, MD 21211, US
Johns Hopkins School of Medicine, Department of Psychiatry and Behavioral Sciences
Baltimore, MD 21205, US

Stewart H. Mostofsky (mostofsky@kennedykrieger.org)
Kennedy Krieger Institute, Center for Neurodevelopmental and Imaging Research
Baltimore, MD 21205, US
Johns Hopkins School of Medicine, Department of Neurology, Department of Psychiatry and Behavioral Sciences
Baltimore, MD 21205, US

Leyla Isik (lisik@jhu.edu)
Johns Hopkins University, Department of Cognitive Science
Baltimore, MD 21218, US

Abstract

Both vision and language signals contribute to real-world social processing, yet they have been mostly studied separately. To understand how these inputs are simultaneously processed, we model individual participant's brain responses (n=34) to a naturalistic movie using vision and language deep neural network embeddings. We find that these embeddings share very little similarity in a naturalistic movie, and they both predict brain responses in the superior temporal sulcus (STS). Within STS, we identified social interaction perception and language selective regions in individual participants to examine how they process vision and language signals in the movie. We found that 1) social perception regions are best explained by vision embeddings, but are also sensitive to sentence-, but not word-, level information, 2) language regions are well predicted by speech-, word-, and sentence-level embeddings and, surprisingly, equally well predicted by vision features as language features. However, 3) language regions are exclusively sensitive to high-level visual information, whereas social perception regions (and lower level visual regions) are sensitive to both low- and high-level visual information. This work suggests that social perception and language regions both integrate visual and language signals, but the specific nature of these integrated representations vary across the STS.

Keywords: naturalistic stimuli; fMRI encoding; multimodal social processing

Introduction

In our daily lives, we effortlessly integrate visual and linguistic signals, especially in social contexts. However, these two inputs are often studied separately. Previous work using controlled stimuli has mapped responses to diverse social signals, including visual (biological motion, faces) and linguistic (voices, theory of mind, and language) input in the superior temporal sulcus (STS), revealing regions highly selective for social stimuli as well as regions responding to multiple modalities (Deen et al., 2015). Recent work has shown that visual social perception regions of the STS are sensitive to communicative interactions (McMahon et al., 2023) and meaningful auditory interactions (Landsiedel & Koldewyn, 2023). This suggests a critical interface with language. However, this unimodal work using cannot address how social and language regions respond to simultaneous visual and linguistic inputs. We bridge this gap using both unimodal controlled and audiovisual naturalistic stimuli.

One powerful technique to analyze naturalistic data is building a linear mapping from stimulus to neural response. This encoding model approach has been used to analyze responses to listening to stories or podcasts (Huth et al., 2016; Schrimpf et al., 2021; Goldstein et al., 2022), and responses to watching silent movies (Huth et al., 2012), but not the neural processing of both visual and linguistic signals from the same naturalistic, socially-rich input. Here, we use deep neural networks to operationalize the visual and linguistic signals of the movie and link them to neural responses, examining responses within localized social perception and language regions. We conduct variance partitioning to pinpoint the specific contributions of these complex feature spaces. We find that both social perception and language regions integrate visual and linguistic signals, but to different extents.

Methods

fMRI. Participants (n=34, ages 19-35, 17 F) watched a 45 minute episode of the BBC series Sherlock. We followed the experimental procedures in Chen et al. (2017). All participants completed a social interaction perception localizer (Isik, Koldewyn, Beeler, & Kanwisher, 2017). A subset of participants (n=25) completed a language localizer (Scott, Gallée, & Fedorenko, 2017). We identified the top 5% motion, social interaction and language selective voxels within the MT, STS, and language parcels (which include STS as well as the broader temporal lobe), respectively. The social interaction and language voxels are non-overlapping in individual subjects (DICE coef. of 0.05 in left and 0.04 in right).

Encoding model For each participant, the fMRI BOLD series for each voxel within an intersubject correlation mask was predicted with a banded ridge regression model (Dupré la Tour et al., 2022) using neural network embeddings of vision and language signals. We used time delays of 1, 3, 4.5, and 6 seconds to account for variability in hemodynamic delays across cortex and fit the model using 5-fold cross validation. To account for temporal autocorrelation, we chunked the time series into continuous segments of 30s before splitting into train and test sets. For vision, we extracted embeddings from a motion energy model (pymoten (Nunez-Elizalde et al., 2021) and from the seven layers of AlexNet (Krizhevsky et al., 2012), which have previously been shown to predict visual responses in high-level visual cortex (Eickenberg et al., 2017). For language, we extracted activations from all layers of a speech transformer model (HuBert (Hsu et al., 2021)), a word-level semantic model (word2vec (Mikolov et al., 2013)), and a sentence-level transformer model (sBERT; all-mpnet-base-v2, huggingface.co) of the spoken content of the episode. We measured the similarity of these feature spaces with Canonical Correlation Analysis (Hotelling, 1936; Knapp, 1978).

From the encoding model, we examined the product measure, a measure of the predictive contribution of each feature space that considers the correlation between feature spaces (Dupré la Tour et al., 2022). We also performed structured variance partitioning (Lin et al., 2024) to examine layer-wise contributions in AlexNet.

Results

Vision and language embeddings share little similarity There was high similarity between the vision model embeddings (AlexNet and motion) and language model embeddings (speech, word, and sentence). Interestingly, there was little correlation between the vision and language feature spaces (Figure 1A).

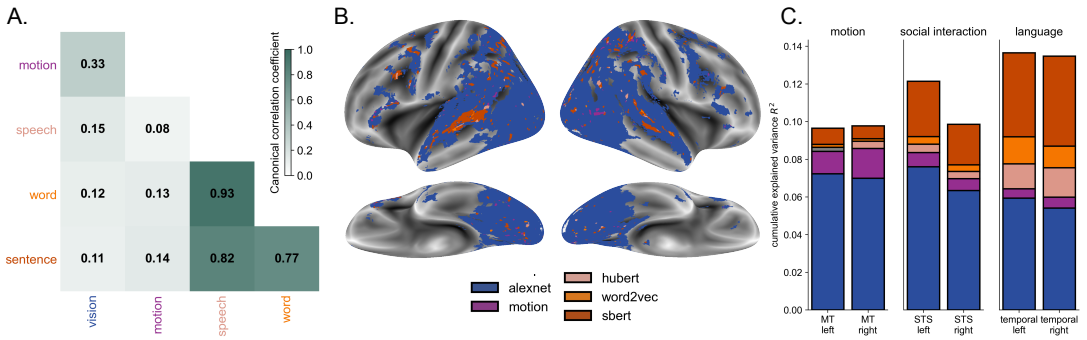

Figure 1: A) Feature similarity of vision and language model embeddings. B) Representative preference map showing which feature explains the most variance in each voxel (projected to surface) C) Product measure (averaged over participants) of each feature space within the joint model. Significant vision>language in motion and social interaction, no difference in language.

Vision and language embeddings both predict STS responses The joint encoding model explains significant group-level variance (corrected $p<0.001$) in all ISC mask voxels. Most voxels are best predicted by vision (AlexNet) features in individual brains, although there are portions best explained by sentence-level features, including in the STS (Figure 1B). However, in individually localized regions, we find strong visual feature predictivity in not only motion (a control visual region) and social perception regions, but also language regions. Follow-up work with more advanced vision and language models (SimCLR, GPT-2) showed a similar trend, with an even stronger advantage for vision models in language regions (not shown for space). We do find significant predictivity of speech, word, and sentence-level features in language regions, but only sentence features in the social perception regions (Figure 1C).

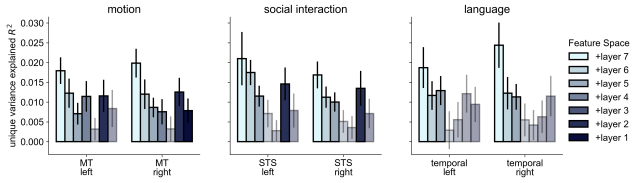

Figure 2: Structured variance partitioning of AlexNet layers. Opaque bars indicate a significant non-zero addition.

High-to-low-level visual feature contributions To delineate the contributions of high- to low-level visual features, we iteratively added AlexNet layers from the last to first layer to our encoding model, examining whether earlier layers explained additional variance on top of later layers (Lin et al., 2024). In motion and social perception regions, both late and early layers explained additional variance, indicating sensitivity to low-level visual features. However, in language-selective regions, no early layers explained additional variance, suggesting a sensitivity exclusively to mid-to-high-level visual features (Figure 2).

Discussion

In this work, we make several contributions towards understanding how humans process naturalistic, socially-rich, audiovisual stimuli.

We found that neural network embeddings of the visual and linguistic signals are not aligned over the course of a naturalistic movie. This illustrates an additional perspective (Small et al., 2024) to recent work emphasizing the alignment between language model representations of images and human high-level vision and deep vision models (Huh et al., 2024; Doerig et al., 2024).

Both vision and language neural network embeddings explained variance across the brain, but vision models dominated prediction in social visual regions and equaled language model predictivity in language regions. Structured variance partitioning revealed key differences between these nearby regions. Language regions were only explained by high-level visual features, building on previous work showing some sensitivity to meaningful visual semantic information (Sueoka et al., 2024). This contrasts with social perception regions, which displayed sensitivity to high- and low-level visual features. Although visual features dominated prediction, sentence features predicted social perception regions, while language regions were also well-predicted by speech and word features.

Our work points to an interaction between the visual and linguistic systems supporting social interaction perception. Specifically, social perception and language regions could be exchanging high-level social visual information and sentence-level language information, while each represent distinct information about lower-level vision/language information. Studying these multimodal regional interactions in natural contexts is an exciting direction for future work.

Acknowledgments

We are grateful to Elizabeth Im and Wenshuo Qin for help with data collection and members of the Isik lab for helpful discussion of this work. We thank Alyssa DeRonda, Natalie Alessi, Eva Freites, Cierra Smith, Alec Gonzaga, and Beatrice Ojuri for help with subject recruitment and testing. This

work was supported by NIMH R21MH129899 and NSF GRFP DGE2139757.

References

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, *20*(1), 115–125. (Number: 1 Publisher: Nature Publishing Group) doi: 10.1038/nn.4450

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(11), 4596–4609. doi: 10.1093/cercor/bhv111

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2024, July). *Visual representations in the human brain are aligned with large language models.* arXiv. (arXiv:2209.11737 [cs]) doi: 10.48550/arXiv.2209.11737

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*, 119728. doi: 10.1016/j.neuroimage.2022.119728

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194. doi: 10.1016/j.neuroimage.2016.10.001

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. doi: 10.1038/s41593-022-01026-4

Hotelling, H. (1936, December). Relations between two sets of variates. *Biometrika*, *28*(3-4), 321–377. doi: 10.1093/biomet/28.3-4.321

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. (arXiv:2106.07447 [cs]) doi: 10.48550/arXiv.2106.07447

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). *The platonic representation hypothesis.* Retrieved from https://arxiv.org/abs/2405.07987

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. (Publisher: Nature Publishing Group) doi: 10.1038/nature17637

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. doi: 10.1016/j.neuron.2012.10.014

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1714471114

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, *85*(2), 410–416. (Place: US Publisher: American Psychological Association) doi: 10.1037/0033-2909.85.2.410

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc.

Landsiedel, J., & Koldewyn, K. (2023, August). Auditory dyadic interactions through the "eye" of the social brain: How visual is the posterior STS interaction region? *Imaging Neuroscience (Cambridge, Mass.)*, *1*, 1–20. doi: 10.1162/imag$_{a0}$0003

Lin, R., Naselaris, T., Kay, K., & Wehbe, L. (2024, September). Stacked regressions and structured variance partitioning for interpretable brain maps. *NeuroImage*, *298*, 120772. doi: 10.1016/j.neuroimage.2024.120772

McMahon, E., Bonner, M. F., & Isik, L. (2023, December). Hierarchical organization of social action features along the lateral visual pathway. *Current Biology*, *33*(23), 5035–5047.e8. (Publisher: Elsevier) doi: 10.1016/j.cub.2023.10.015

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space.* arXiv. (arXiv:1301.3781 [cs])

Nunez-Elizalde, A. O., Deniz, F., Dupré la Tour, T., Visconti di Oleggio Castello, M., & Gallant, J. L. (2021). *pymoten: scientific python package for computing motion energy features from video.* doi: 10.5281/zenodo.6349625

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. doi: 10.1073/pnas.2105646118

Scott, T. L., Gallée, J., & Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive Neuroscience*, *8*(3), 167–176. doi: 10.1080/17588928.2016.1201466

Small, H., Masson, H. L., Mostofsky, S., & Isik, L. (2024, October). Vision and language representations in multimodal AI models and human social brain regions during natural movie viewing..

Sueoka, Y., Paunov, A., Tanner, A., Blank, I. A., Ivanova, A., & Fedorenko, E. (2024, June). The Language Network Reliably "Tracks" Naturalistic Meaningful Nonverbal Stimuli. *Neurobiology of Language*, *5*(2), 385–408. doi: 10.1162/nol$_{a0}$0135