Disentangling redundant and synergistic interactions in the alignment between auditory brains and machines

Christian Ferreyra (christian.ferreyra@etu.univ-amu.fr)
Institut des Neurosciences de la Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France.
Laboratoire d'Informatique et des Systèmes, UMR 7020, CNRS and Université Aix-Marseille, Marseille, France.

Marie Plegat (marie.plegat@univ-amu.fr)
Institut des Neurosciences de la Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France.

Giorgio Marinato (giorgio.marinato@univ-amu.fr)
Institut des Neurosciences de la Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France.

Maria Araújo Vitória (maria.araujo.vitoria@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University Maastricht, The Netherlands.

Michele Esposito (m.esposito@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University Maastricht, The Netherlands.

Elia Formisano (e.formisano@maastrichtuniversity.nl)
Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University Maastricht, The Netherlands.

Thierry Artières (thierry.artieres@lis-lab.fr)
Laboratoire d'Informatique et des Systèmes, UMR 7020, CNRS and Université Aix-Marseille, Marseille, France.

Bruno L. Giordano (bruno.giordano@univ-amu.fr)
Institut des Neurosciences de la Timone, UMR 7289, CNRS and Université Aix-Marseille, Marseille, France.

Abstract

Artificial neural networks (ANNs) have become increasingly useful for modeling how the brain builds representations from the natural world, yet the nature of their representational alignment with dynamic brain activity remains underexplored. Here, we introduce an information-theoretic framework to decompose representational geometries into redundant and synergistic components using partial information decomposition (PID). Combining magnetoencephalography (MEG) recordings from participants listening to natural sounds, and two sound-processing ANNs with categorical (CatDNN) and continuous (SemDNN) semantic outputs, we analyze time-varying brain-model alignment for two optimized stimulus sets. For low-agreement stimulus sets, where mutual information between models is minimized, SemDNN reveals higher mutual information with brain activity. PID further shows greater redundancy and synergy for SemDNN, suggesting sustained temporal integration of intermediate semantic features that can potentially afford a more accurate readout of the auditory environment. These results highlight the value of representational decomposition for detailing shared and complementary components of the alignment between brains and ANNs.

Keywords: Deep learning; Redundancy; Synergy; Auditory processing; Information theory

Introduction

Understanding how artificial neural networks align with dynamic brain representations is a central question in computational neuroscience (Sucholutsky et al., 2023). While previous research showed that ANNs better predict cerebral and behavioral responses to natural sounds (Giordano, Esposito, Valente, & Formisano, 2023), these efforts have largely focused on static representations, leaving the temporal dynamics of model-brain alignment underexplored. Furthermore, it has also been shown that brain-model representational similarity varies with the stimulus set (Araújo et al., 2024), motivating the use of optimized stimuli to better distinguish between models. Crucially, recent findings reveal that redundant and synergistic encoding serve distinct roles in brain function, with synergy supporting distributed representations (Greco, Moser, Preissl, & Siegel, 2024; Koçillari et al., 2023).

In this work, we introduce a novel information theoretic framework to analyze the dynamic alignment between brain activity and sound-processing ANNs by decomposing their representational geometries into redundant and synergistic components. Using sensor-level MEG responses to natural sounds and two similar acoustic-to-semantic ANNs (categorical and continuous semantic outputs), we investigate how well these models can align with brain activity over time, and, critically, how time-resolved cerebral representations interact throughout sound listening.

Methods

Experimental design. Magnetoencephalography (MEG) data were acquired from 21 participants as they listened to natural sounds while performing a one-back repetition detection task. The stimulus set comprised 600 natural sounds (duration = 2 s) where a common set of 150 sounds was shared across all participants (Araújo et al., 2024), and the remaining 450 were split across three distinct 150 sound sets, each assigned to a separate subgroup of 7 participants (unique sets). Each participant heard each sound 8 times throughout two subsequent MEG sessions.

Artificial neural networks. Previous studies have compared multiple sound-processing ANNs that vary in architecture, size and training objectives, and found that many are in hierarchical correspondence with cerebral representations (Tuckute, Feather, Boebinger, & McDermott, 2023). Here, in order to disentangle layer-specific representations in dynamic cerebral responses, we considered two nearly identical acoustic-to-semantic deep neural network models (Esposito et al., 2024). Both architectures were trained with the same dataset and have the same backbone (4 convolutional blocks and one global average pooling = layers 1-5), but one model outputs probabilities of sound-event categories (CatDNN), and the other a continuous Word2Vec semantic embedding learned from sound (SemDNN).

Decomposing representational geometries. To decompose how different models align with the brain's dynamical representations of natural sounds, we first computed representational dissimilarity matrices (RDMs) (Figure 1.A, top). For the preprocessed sensor-level MEG data (ICA correction; high-pass at 0.05Hz, low-pass at 70Hz and notch at 50Hz), we computed dissimilarities using cross-validated squared Mahalanobis distance across 306 sensors. A noise covariance matrix was estimated using Ledoit-Wolf method on pre-stimulus activity (Ledoit & Wolf, 2004). For the ANNs, we calculated RDMs as the normalized squared Euclidean distance.

We then applied an information-theoretic framework, the partial information decomposition (Williams & Beer, 2010), to analyze the relationship between model representations at different latencies of the cerebral response. PID allow us to decompose the unique, redundant and synergistic contributions to the total information that a set of source RDMs have about a target representation (Figure 1.A, bottom). To estimate redundancy (shared information) and synergy (combined information) between representations, we employed minimum mutual information PID (Barrett, 2015) alongside mutual information estimation using a Gaussian copula estimator (Ince et al., 2017).

We carried out two complementary analyses: **1) Mutual information** between the time-varying brain RDMs and the different layer RDMs of each ANN. **2) Time-to-time PID** using two brain RDMs from different time points as sources and a model layer RDM as target.

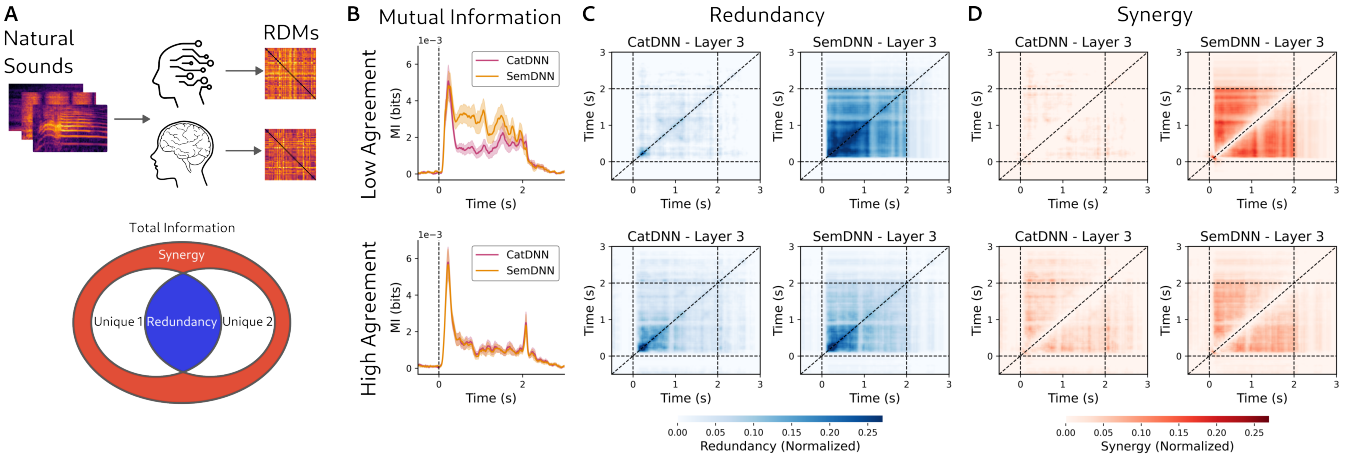

Figure 1: A) Analysis framework: for a given stimulus set, we compute brain (time-varying) and model RDMs (top). Partial information decomposition divides the total information into unique, redundant and synergistic components (bottom). B) Brain-model mutual information. C) Normalized redundant information for layer 3 of each model. D) Normalized synergistic information for the same systems as D. Top and bottom rows in B, C and D, show low- and high-agreement stimuli sets, respectively.

Optimized stimuli. Following a previous approach (Hosseini et al., 2024), we defined two types of optimized stimulus subsets. For each unique set, combined with the common set, we identified subsets of sounds that either minimized or maximized the mutual information between layer 3 RDMs of CatDNN and SemDNN, thereby defining the low- and high-agreement sets, respectively. A total of 6 optimized sets, 2 per unique set. We selected layer 3 because layer-wise brain-model mutual information differentiate at this processing stage (results not shown).

Results

Brain-model mutual information. We computed layer-wise brain-model mutual information for the low- and high-agreement stimulus selections. Figure 1.B shows the mean$\pm$SEM of the maximum mutual information across layers at each time point, averaged across participants, and for two stimulus subset optimizations. For the low-agreement set, SemDNN exhibited consistently higher mutual information than CatDNN throughout sound listening. In contrast, both models achieved comparable mutual information for the high-agreement set, providing evidence for shared representational axes.

Time-to-time partial information decomposition. Given a target model representation, we computed time-to-time partial information decomposition using two different brain RDMs as sources. For each model's layer 3, we averaged normalized redundant and synergistic maps across participants and optimized stimuli subsets (normalization by the total information). Results show expected redundancy between brain representations along the diagonal for the first second for both models and subsets (Figure 1.C). We observed an overall increase in redundancy for SemDNN in the low-agreement set possibly due to common encoded features that are highlighted between brain and layer 3 representations. In contrast, CatDNN RDMs do not exhibit long-lasting redundant interactions for this stimulus set. Figure 1.D shows an off-diagonal synergy pattern between brain representations, with highest values for SemDNN in the low-agreement set suggesting temporal integration of intermediate semantic features.

Conclusions

Overall, our analyses reveal a clear dominance of SemDNN throughout the cerebral response compared to CatDNN. PID further supports this distinction, showing that SemDNN exhibited greater redundancy and synergy with brain representations in the low-agreement regime. These synergistic interactions suggest that the integration of intermediate semantic features across latencies of the cerebral response could potentially afford a more accurate readout of the auditory environment. However, further work using source-localized MEG data with behavioral responses is needed to elucidate the nature of these synergistic interactions and their functional relevance. These results highlight the potential of decomposing representational geometries using PID as a powerful framework for disentangling shared and complementary contributions of neural and model representations.

Acknowledgments

This work was funded by the French National Research Agency (ANR-21-CE37-0027-01, BLG; ANR-16-CONV-0002 ILCB; ANR11-LABX-0036 BLRI), by the Dutch Research Council (NWO 406.20.GO.030 to EF), and by ERC-2024-SyG NASCE (Proj. 101167313) to EF and BLG. The authors thank the Institut du Cerveau (CENIR, Paris, France) for enabling the MEG acquisitions.

References

Araújo, M., Plegat, M., Marinato, G., Esposito, M., Herff, C., Giordano, B. L., & Formisano, E. (2024). Optimal stimulus selection for dissociating acoustic and semantic processing of natural sounds. *Conference on Cognitive Computational Neuroscience*.

Barrett, A. B. (2015). Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *Physical Review E*, *91*(5), 052802.

Esposito, M., Valente, G., Plasencia-Calaña, Y., Dumontier, M., Giordano, B. L., & Formisano, E. (2024). Bridging auditory perception and natural language processing with semantically informed deep neural networks. *Scientific Reports*, *14*(1), 20994.

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, *26*(4), 664–672.

Greco, A., Moser, J., Preissl, H., & Siegel, M. (2024). Predictive learning shapes the representational geometry of the human brain. *Nature Communications*, *15*(1), 9670.

Hosseini, E., Casto, C., Zaslavsky, N., Conwell, C., Richardson, M., & Fedorenko, E. (2024). Universality of representation in biological and artificial neural networks. *bioRxiv*, 2024–12.

Ince, R. A. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. *Hum Brain Mapp*, *38*, 1541–1573. doi: 10.1002/hbm.23471

Koçillari, L., Celotto, M., Francis, N. A., Mukherjee, S., Babadi, B., Kanold, P. O., & Panzeri, S. (2023). Behavioural relevance of redundant and synergistic stimulus information between functionally connected neurons in mouse auditory cortex. *Brain Informatics*, *10*(1), 34.

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, *88*(2), 365–411.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., ... others (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Tuckute, G., Feather, J., Boebinger, D., & McDermott, J. H. (2023). Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions. *Plos Biology*, *21*(12), e3002366.

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv*, 1004.2515. doi: 10.48550/arXiv.1004.2515