Measuring and Modeling Transformations of Information Between Brain Regions with fMRI

Stefano Anzellotti, Evelina Fedorenko, Alfonso Caramazza, Rebecca Saxe

Investigating how information is transformed from brain region to brain region is a crucial step to understand the neural foundations of cognitive processes. This investigation requires a characterization of the representations encoded in different regions, and models of how they are transformed that can match the complexity of neural processes. We introduce an approach in which representations are characterized as points in multidimensional spaces, and processes transforming representations from region to region are modeled as nonlinear functions using artificial neural networks. Across multiple experiments with different stimuli and tasks, we show that this approach reveals functionally relevant network structure and outperforms comparable linear models at predicting independent data.

Cognition consists of processes operating on representations. A key challenge for cognitive neuroscience is therefore to characterize the transformation of representations that occurs during neural processing. Representations in cortex can be characterized in terms of distinct patterns of neural population activity (using direct electrophysiological recording[1]) or of blood oxygenation (e.g. using functional magnetic resonance imaging, fMRI[2]). For example, different object categories evoke distinct spatial patterns of activity across ventral temporal cortex[3]. Over the past decade, several new methods have been developed to characterize neural representations by exploiting the rich variability in multivariate neural responses: to extract information about stimuli or tasks[2-6], and to directly model the encoding of this information within a cortical region[7]. The next open question is: How are representations transformed as they are processed, between brain regions? This article describes an approach to investigate this question. The approach is described and tested using fMRI data, but it is more generally applicable to other data acquisition techniques.

Existing methods for capturing inter-region interactions using fMRI data assume, implicitly or explicitly, that such interactions are linear. For example, widely used techniques, like functional connectivity[8], psychophysiological interaction analyses[9], granger causal modeling[10] and dynamic causal modeling[11], measure the relations between the average magnitude of response across regions (that is, the univariate response). A few relatively new techniques measure the relations between multivariate responses across regions[12-15], but are still limited to linear interactions. However, the transformation of representations between regions is likely multivariate and nonlinear. An example is the case of invariance in object recognition[16] . Early visual processing can be modeled as representing images in "pixel-space": frontal views of two different faces are more similar in pixel brightness than a frontal and a profile view of the same person. Processing along the ventral visual stream is then modeled as series of non-linear transformations[17], for example to create a "face-space" in which highly heterogeneous images of the same face are all seen as similar (i.e. occupy a tight convex region, supporting linear classification), allowing observers to recognize a specific individual, even across modalities (i.e. face and voice). An even more complex transformation occurs in the brain of a reader, to transform the brightness of pixels into representations of letters, words, sentences, and then a scientific argument.

In this article, we introduce a modeling framework that can capture multivariate and non-linear interactions between brain regions. In the first step, a data-driven multidimensional model of the representational space in each brain region is generated. In the second step, the dynamic mappings between representational spaces are modeled as multivariate and nonlinear transformations, and the predictive power of these nonlinear mappings is tested in independent data. We apply this framework to the analysis of two fMRI datasets: an experiment on the recognition of person identity from faces and voices, and an experiment on language understanding. Across both experiments, we find that the nonlinear models 1) explain more variance in independent data than linear processes, 2) reveal

structural properties of networks driven by hemispheric laterality, stimulus type, and interhemispheric homology, and 3) can be modulated by the task performed by the participants.

Results

Nonlinear processes explain more independent variance than linear processes during the recognition of faces and voices

Nonlinear processes during the recognition of faces and voices were investigated with a paradigm in which participants were asked to detect a target famous person identity from images of faces and recordings of voices.

Brain regions showing selective responses to faces and to voices were functionally defined in individual participants using two independent functional localizers employing a one-back task. Face selective regions were defined as regions showing stronger responses to images of faces than to images of buildings, and voice selective regions as regions showing stronger responses to recordings of voices than recordings of tool sounds. The contrast for face-selective regions (Supplementary Figure 1A) identified the fusiform face area (FFA) bilaterally, the superior temporal sulcus (STS) bilaterally, the ventral anterior temporal lobes (vATL) bilaterally, and posterior cingulate (PCvis). The contrast for voice-selective regions (Supplementary Figure 1B) identified the posterior, middle and anterior superior temporal gyri (STG) bilaterally, posterior cingulate (PCaud), and ventromedial prefrontal cortex (vmPFC). Some overlap between face-selective and voice-selective regions was found in the posterior cingulate and in superior temporal cortex.

The nonlinear interactions between these functionally-defined regions were examined, using data from five experimental runs of mixed faces and voices. Across all voxels in each region, principal component analysis (PCA) was applied to find the dimensions that best capture the variance in the region's responses over time. Each dimension corresponds to a spatial pattern of activity across voxels within a region. These spatial patterns can be interpreted as axes of the region's representational space. For the current analyses, we selected the top five dimensions in each region (based on prior experiments)[18].

Nonlinear interactions between each pair of brain regions were modeled with artificial neural networks having as inputs the responses in the dimensions of one region and as outputs the responses in the dimensions of the other region (and viceversa). The networks were trained using the entire timecourses in all but one run, and the variance explained in the excluded run was calculated, generating an independent measure of "nonlinear correlation" |r| (see Methods). This procedure was repeated for each choice of the excluded run, and the results from the different iterations were averaged.

The first question we set out to answer is whether the use of nonlinear models of the interactions between brain regions in fMRI data is justified. To address this question, we compared the performance of nonlinear models to the performance of linear models. For the linear models, the data were analyzed in the same way (with identical denoising and dimensionality reduction), but multivariate regression between the regions was calculated.

Nonlinear models explained more variance in independent data than linear models across all choices of the degrees of nonlinearity tested ($t(10) = 2.61$, $p < 0.05$ when using one hidden node, $t(10)>5$ and $p<0.0005$ when using more than one hidden node; Figure 1A-C). Because independent data are used for the training and testing of the models, this difference cannot be due to the greater complexity of nonlinear models. Indeed, the |r| values for nonlinear models do not continue to increase with the number of hidden nodes, but rather asymptote reaching a peak at 5 hidden nodes. Thus, the use of nonlinear models over linear models is justified: nonlinear models explain more variance than linear models in independent data. Furthermore, the number of hidden nodes at which the |r| value asymptotes is an estimate of the complexity of the nonlinear process that can be measured within the limits of the available fMRI data.

A more accurate characterization of a network should lead to improved discrimination between strong and weak inter-region interactions. Therefore, we hypothesized that moving from simpler to more complex models, an overall increase in the |r| values would be accompanied by an increase in the variability of |r| values across different region pairs, with greater differences between pairs of regions with strong interactions and pairs of regions with weak interactions. To test this hypothesis, we calculated the variance of the set of |r| values in the connectivity matrices generated with different models. The variance of |r| values across region pairs was indeed found to be greater for nonlinear than linear models, and to asymptote for the same number of nodes as the |r| values themselves (Figure 1D). In other words, the models that explained most variance in independent data also provided maximal discrimination between strong and weak interactions between regions.

As an additional control, we attempted to predict the multivariate timecourses in each region using the experimental conditions convolved with the standard haemodynamic response function (hrf) from SPM. Both linear and nonlinear models based on multivariate timecourses in other regions generated better predictions than the model based on the experimental conditions alone (Figure 1C, red line).

The network of face- and voice- selective regions is organized by modality and hemispheric laterality

In the experimental task, participants viewed faces and voices, and extracted a common invariant, the person's identity. Nevertheless, we hypothesized that pairs of brain regions involved in processing the same modality (faces, or voices) should have stronger interactions during this task. As hypothesized, stronger interactions were found between pairs of regions selective for the same modality (Figure 2A). (Note that this result is true even though the mean response of each region over time, the univariate timecourse, was removed from the data prior to analysis; these interactions reflect only relations between the spatial patterns of activity in different regions).

In the end, multi-dimensional scaling (MDS) was used to visualize the structure of the functional network (using closeness to represent the strength of each measured interaction, see Figure 2B, Supplementary Video 1). In the MDS visualization, regions were found to cluster by modality (preferring faces or voices in the independent localizer) and by hemispheric laterality (left versus right hemisphere, Figure 2B, Supplementary Video 1). This clustering was quantified with statistical tests on the strength of interactions. Interactions between pairs of regions both within the face-selective group or both within the voice-selective group are significantly stronger than between pairs of regions in which one is face-selective and the other voice-selective ($t(10) = 4.0900$, $p<0.005$). Interactions between pairs of regions within the same hemisphere are significantly stronger than interactions between pairs of regions in which one is in the right hemisphere and the other in the left hemisphere ($t(10) = 5.0781$, $p<0.005$).

Nonlinear processes explain more independent variance than linear processes during language understanding

To probe the versatility and potential of this framework for modeling nonlinear processes, in a second experiment we explored its application to the study of language understanding. Participants (N=16) watched a set of short stories (presented visually word-by-word), with each story contained in a separate functional run. In two of the stories participants watched passively, while in two different stories they engaged in an unrelated demanding task (a two-back task on the orientation of a line).

Language-selective regions were defined in individual participants with an independent language localizer[19] as regions showing stronger responses during the reading of sentences than the reading of sequences of nonwords. This contrast identified six regions in the left hemisphere (Supplementary Figure 2): three in the frontal lobe (LIFGorb, LIFG, and LMFG), three in the temporal and parietal cortices (LAntTemp, LPostTemp, and LAngG), as well as six right-hemisphere homologues.

The analysis approach used in Experiment 1 was repeated twice in Experiment 2, once for the passive reading task, and once for reading under a cognitive load. A comparison between the performance of nonlinear models of the interactions between regions to the performance of linear models (using identical preprocessing and dimensionality reduction) revealed that nonlinear models explain more variance than linear models in independent data for both tasks (passive listening: $t(15) > 23$, $p < 0.0005$ for all numbers of hidden nodes, Figure 3A-C; cognitive load: $t(15) > 11$, $p<0.0005$ for all numbers of hidden nodes, Figure 3 E-G). Also, the variance of $|r|$ values across region pairs was higher for nonlinear than linear models, and peaked at a similar number of nodes as the $|r|$ values themselves for both experimental tasks (Figure 3D for passive listening, Figure 3H for the cognitive load task). Thus, nonlinear models again explained more variance in independent data, and were more sensitive to the difference between strong and weak interactions, than linear models.

Network structure in language understanding is stable across different tasks

In an MDS visualization, homologous regions in the two hemispheres showed strong inter-hemispheric interactions (Figure 4, Supplementary Videos 2, 3). This structure is intriguing because the homologous regions are physically distant; nevertheless, the trajectory in representational space was strongly related between these pairs of regions. A similar network structure was found in the two tasks (passive listening: Figure 4A, cognitive load: Figure 4B).

Subtle differences in nonlinear processes between tasks make it possible to reliably classify the task based on network structure

The transformation of representations, and therefore the pattern of interactions between brain regions, should be different when participants are doing different tasks on the same input. In the current study, participants performed two tasks on similar verbal material: passive reading, or reading under cognitive load. To test whether this difference in mental activity was reflected in patterns of interregional interactions, linear discriminant analysis (LDA) was applied to the nonlinear interactions between regions in the two tasks. An LDA classifier was trained with data from all but one participant, and the accuracy at classifying the tasks in the left-out participant was assessed. This procedure yielded a mean classification accuracy of 63% ($p < 0.05$, permutation test, 1000 iterations). By contrast, the pattern of linear interactions between brain regions in the same data could not be used to classify the participants' task (accuracy = 53%, $p = 0.25$). Classification based on nonlinear interactions was significantly higher than classification based on linear interactions (one-tailed $t(15) = 1.86$, $p < 0.05$).

Discussion

This article introduces a method to investigate the transformation of representations between brain regions, using data from noninvasive neuroimaging in human participants. Each brain region's response over time is represented as a multivariate timecourse: the trajectory in the dimensions of its own representation space. Then, the relation between these trajectories across regions is estimated, using nonlinear function approximators (artificial neural networks). This is a new analysis technique, but also a new way to conceive the study of interactions between brain regions, shifting from measuring correlated fluctuations in the overall response of regions, to studying complex transformations of representations: the neural foundations of cognitive processes.

This approach stands at the confluence between existing techniques for multivariate pattern analysis (MVPA) and for measuring functional connectivity. As compared to other work that integrates MVPA and connectivity[11-14], the method described in this article leverages the potential of multivariate maps between representational spaces introduced by Multivariate Pattern Connectivity (MVPC)[13], combining it with nonlinear functions in order to predict the multivariate responses in one regions as a function of complex interactions between multiple dimensions represented in another region. For

example, one region might encode information about face parts (e.g. each dimension reflects the relative size or distance of one feature); another region might encode information about facial emotional expression; the proposed method has the potential to measure the nonlinear transformation of dimensions in feature-space required to predict responses in the space of facial emotional expressions.

The results show that this approach is not only theoretically principled, but also practically viable. Nonlinear interactions between brain regions explain more variance in independent data than linear interactions, demonstrating that even the limited data obtained with a standard fMRI protocol justify the use of nonlinear models. In other words, the data are sufficiently rich that nonlinearity can be introduced without leading to overfitting or compromising generalization to independent data. Furthermore, nonlinear models reveal meaningful and reliable structure in the networks of brain regions: stronger interactions between regions based on their preferred stimulus modality (faces or voices), and based on their position in the cortical hierarchy (homologous regions across hemispheres).

Several lines of research can be pursued in the future to increase the potential of this modeling approach. A critical step will be to relate the principal components of responses in each individual region to properties of the stimuli (i.e. an encoding model of each region's representational space[20]). Next, data driven models of nonlinear interactions between brain regions could be used to constrain algorithmic models of neural computation, for instance models based on deep neural networks. In the end, graph analysis methods[21] could be used to study the structure of networks composed of nonlinear interactions. Modeling neural processes complements the investigation of neural representations, providing the instruments to study how these representations are transformed from region to region to give rise to cognition.

Methods

Experimental design

In the first experiment 11 participants completed a face localizer and a voice localizer. In the face localizer, participants watched 16s blocks of images of faces and houses while performing a 1-back task. In the voice localizer, participants performed a 1-back task on 16s blocks of voice and tool sounds. After the localizers, participants were administered five experimental runs in which they were asked to detect a target person identity, defined before the beginning of the experiment. Images of faces and recordings of voices were presented in 4s long trials arranged in a pseudorandomized order generated with Optseq 2 (http://surfer.nmr.mgh.harvard.edu/optseq/). Participants were asked to detect a famous target identity. Each run consisted of 120 trials and lasted approximately 8 minutes.

In the second experiment, 16 participants completed a language localizer in which they passively read 18 seconds long blocks of sentences or nonwords. Each block consisted of 3 sequences of nonwords or of words forming a sentence. At the end of each sequence a cue was presented and participants had to press a button. After the localizer, participants read stories presented visually word-by-word. Each run contained a full story, lasting approximately 6 minutes. In two of the runs, participants watched the stories passively; in two other runs they performed a two-back task on the orientation of a line.

Data acquisition

Data for the first experiment were collected on a Bruker BioSpin MedSpec 4T at the Center for Mind/Brain Sciences (CIMeC) of the University of Trento using a USA Instruments eight-channel phased-array head coil. Before collecting functional data, a high-resolution ($1\times1\times1$ mm^3) T1-weighted MPRAGE sequence was performed (sagittal slice orientation, centric phase encoding, image matrix = 256 × 224 [Read × Phase], field of view =256 × 224 mm [Read × Phase], 176 partitions with 1-mm thickness, GRAPPA acquisition with acceleration factor = 2, duration = 5.36 minutes, repetition time = 2700, echo time = 4.18, TI = 1020 msec, 7° flip angle).

Functional data were collected using an echo-planar 2D imaging sequence with phase oversampling (image matrix = 70 × 64, repetition time = 2000 msec, echo time = 21 msec, flip angle = 76°, slice thickness = 2 mm, gap = 0.30 mm, with 3 × 3 mm in plane resolution). Over four runs, 1260 volumes of 43 slices were acquired in the axial plane aligned along the long axis of the temporal lobe.

Data for the second experiment were collected on a Siemens Trio 3T scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at McGovern Institute for Brain Research at MIT. Before collecting functional data, a high-resolution ($1\times1\times1$ mm^3) T1-weighted MPRAGE sequence was performed (sagittal slice orientation, centric phase encoding, image matrix = 256 × 224 [Read × Phase], field of view =256 × 224 mm [Read × Phase], 128 axial slices, GRAPPA acquisition with acceleration factor = 2, duration = 5.36 minutes, repetition time = 2530, echo time = 3.48, TI = 1020 msec, 7° flip angle).

Functional data were collected using an echo-planar 2D imaging sequence (image matrix = 96 × 96, repetition time = 2000 msec, echo time = 30 msec, flip angle = 90°, 31 slices, slice thickness = 4 mm, 10% distance factor, with 2.1 × 2.1 mm in plane resolution). Prospective acquisition correction[22] was used to adjust the positions of the gradients based on the participant's motion one TR back. The first 10 s of each run were excluded to allow for steady-state magnetization.

Data analysis

Data were preprocessed with SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) running on MATLAB 2015b. After slice-timing correction and realignment, the functional volumes were coregistered to the anatomical volume and normalized. No smoothing was applied.

Functional regions of interest (ROIs) were defined in individual subjects with t-contrasts in the functional localizers for faces>houses and voices>tool sounds in the first experiment, for sentences>nonwords in the second experiment. In the first experiment, ROIs were defined as 9mm radius spheres centered in the t-contrast peaks in the anatomical areas where activations were expected based on prior studies. In the second experiment, ROIs were defined by taking the 100 voxels showing highest t-contrasts within each of six search spaces generated based on a probabilistic activation overlap map for the localizer contrast in 220 participants[23]; these were similar to the original search spaces reported in Fedorenko et al.[19], but the two anterior temporal and two posterior temporal search spaces ended up being morphed together.

Patterns of response in each ROI were extracted and denoised with CompCor[24], regressing out the first five principal components extracted from a control ROI in the ventricles. The data were then demeaned, and the runs entering in each analysis were concatenated. In the first experiment, all runs were analyzed together, in the second experiment, the runs with passive listening and the runs with the active task were analyzed separately. Dimensionality reduction was performed with PCA, the first five principal components were preserved on the basis of previous results[17]. The data from different runs were subsequently split for the training/testing of the interaction models. This procedure was identical for the linear and for the nonlinear interaction models.

For the linear models, the interactions between each pair of brain regions were modelled with a multiple regression taking as inputs the values along the dimensions in one brain region in the pair and taking as outputs the values along the dimensions in the other region.

For the nonlinear models, interactions were modelled with one-hidden-layer artificial neural networks with the same inputs and outputs. The networks used hyperbolic tangent transfer functions and were trained with backpropagation using the Levenberg-Marquardt algorithm[25]. Variance explained was calculated with a leave-one-run-out cross-validation procedure in which the networks were trained with data from all runs except from one, and prediction accuracy was tested on the left-out run that was not used for training. In order to facilitate comparison with functional connectivity measures, which use r values rather than R^2, the squared root of variance explained was computed obtaining a "generalized correlation" index $|r|$. Thanks to the independence of the training/testing procedure, the estimates of $|r|$

are not biased by the complexity of the models, and in fact they asymptote for a small number of hidden nodes (Figure 1A, 2A, 3A-B). Nonmetric multidimensional scaling was performed with the MATLAB function “mdscale” using default parameters.

Acknowledgments

This work was funded by NIH Grant 1R01 MH096914-01A1 to Prof. Rebecca Saxe, by a grant from the NICHD to EF (HD057522), a grant from the Simons Foundation to the Simons Center for the Social Brain at MIT to EF and RS, and by the Center for Mind/Brain Sciences at the University of Trento. Stefano Anzellotti was supported by a fellowship from the Simons Center for the Social Brain. We would like to thank Tyler Bonnen, Dae Houlihan, Zach Mineroff, Alex Paunov, and Briana Pritchett for their assistance during the acquisition of the data, and the implementation of experiment and analysis scripts. We would like to acknowledge the Athinoula A. Martinos Imaging Center at McGovern Institute for Brain Research at MIT, and the support team (Steve Shannon, Atsushi Takahashi, and Sheeba Arnold).

Bibliography

1. Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, 330(6005), 845-851.
2. Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7), 523-534.
3.Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425-2430.
4. Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). " Who" Is Saying" What"? Brain-Based Decoding of Human Voice and Speech. *Science*, 322(5903), 970-973.
5. Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage*, 38(4), 649-662.
6. Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126-1141.
7. Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19), 1641-1646.
8. Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnetic resonance in medicine*, 34(4), 537-541.
9. Friston, K. J., Buechel, C., Fink, G. R., Morris, J., Rolls, E., & Dolan, R. J. (1997). Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage*, 6(3), 218-229.
10. Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25(1), 230-242.
11. Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, 19(4), 1273-1302.
12. Coutanche, M. N., & Thompson-Schill, S. L. (2013). Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Frontiers in human neuroscience*, 7, 15.
13. Coutanche, M. N., & Thompson-Schill, S. L. (2015). Creating concepts from converging features in human cortex. *Cerebral Cortex*, 25(9), 2584-2593.
14. Anzellotti, S., Caramazza, A. & Saxe, R. (2016). Multivariate Pattern Connectivity. *bioRxiv*, 046151.
15. Geerligs, L., & Henson, R. N. (2016). Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation. *Neuroimage*, 135, 16-31.
16. Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619-8624.
17. DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8), 333-341.
18. Diez, I., Erramuzpe, A., Escudero, I., Mateos, B., Cabrera, A., Marinazzo, D., ... & Cortes Diaz, J. M. (2015). Information flow between resting-state networks. *Brain connectivity*, 5(9), 554-564.
19. Fedorenko, E., Hsieh, P. J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*,104(2), 1177-1194.
20. Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. Neuroimage, 56(2), 400-410.

21. Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3), 186-198.
22. Thesen, S., Heid, O., Mueller, E., & Schad, L. R. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magnetic Resonance in Medicine*, 44(3), 457-465.
23. Jouravlev et al., submitted
24. Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, 37(1), 90-101.
25. Moré, J. J. (1978). The Levenberg-Marquardt algorithm: implementation and theory. In *Numerical analysis* (pp. 105-116). Springer Berlin Heidelberg.

Figures

Figure 1

r

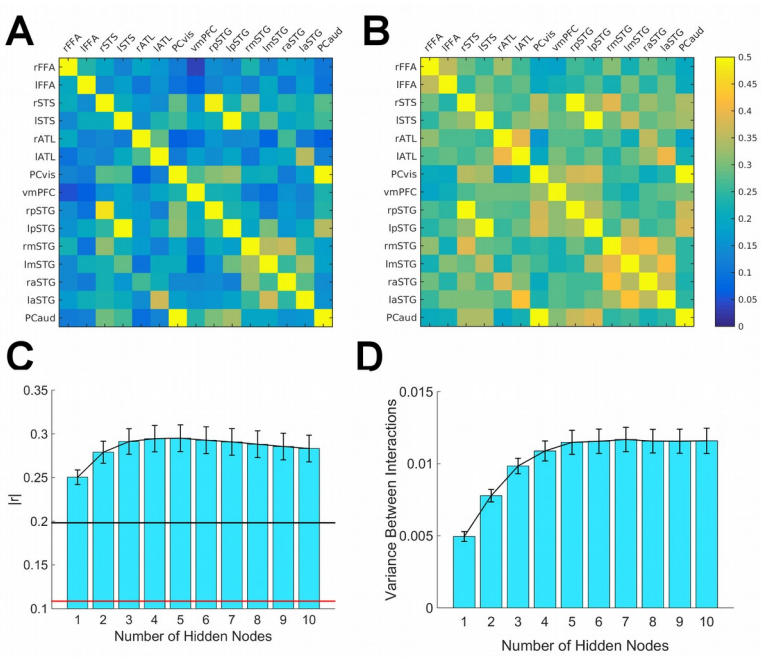


Fig. 1. A) Linear multivariate connectivity matrix generated with multiple regression. Each cell of the matrix depicts the |r| index (square root of independent variance explained) for a pair of regions. B) Nonlinear multivariate connectivity matrix generated using the same data with one-hidden-layer artificial neural networks, using a network with 5 hidden nodes. Each cell depicts the |r| index for a pair of regions. C) Bars depict the average |r| across all connections for nonlinear multivariate connectivity as a function of the number of nodes in the hidden layer. The |r| index increases and asymptotes at about 5 hidden nodes. The black horizontal line denotes the average |r| for linear multivariate connectivity, and the red line the |r| obtained using a linear model based on the set of experimental conditions. D) Bars depict the variance between |r| values within the connectivity matrix as a function of the number of hidden nodes. As the number of hidden nodes increases, the differences between low |r| and high |r| interactions increase, reaching an asymptote at 5 hidden nodes, the same number of nodes at which the average |r| asymptotes.

Figure 2

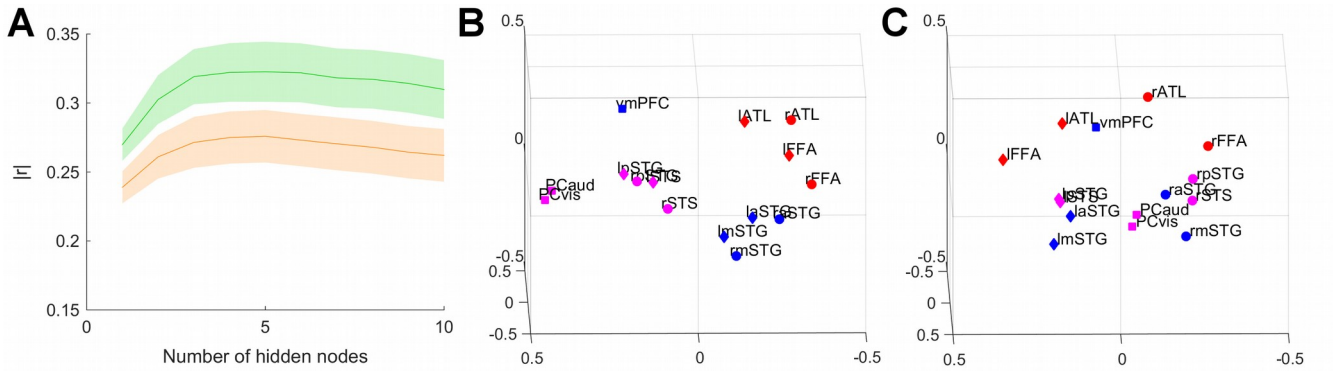


Fig. 2. A) |r| index as a function of number of hidden nodes for pairs of regions selective for the same modality (green) and for pairs of regions one selective for faces and the other for voices (orange). B) 3D multidimensional scaling based on the nonlinear multivariate connectivity matrix shows clustering of face-selective regions (red), voice-selective regions (blue) and regions with overlap between face- and voice- selectivity (purple). C) the same multidimensional scaling, seen from a different vantage point, shows separation between right hemisphere regions (circles, on the right) and left hemisphere regions (diamonds, on the left).

Figure 3

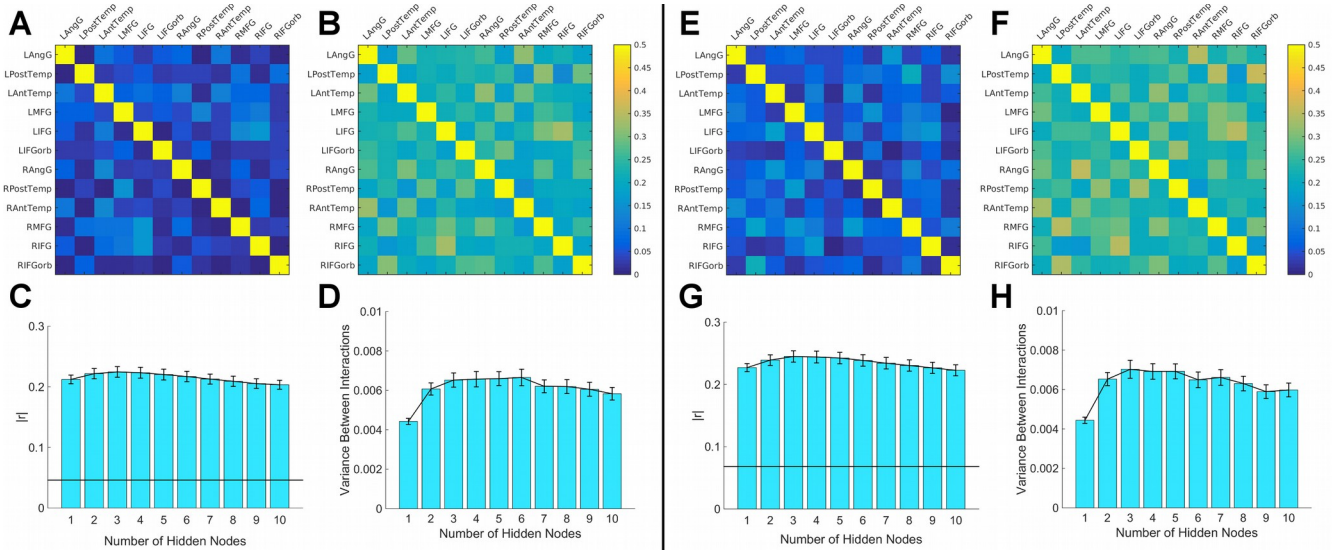


Fig. 3. A-D: passive listening task. A) Linear multivariate connectivity matrix. Each cell of the matrix depicts the |r| index for a pair of brain regions. B) Nonlinear multivariate connectivity matrix for 3 hidden nodes. C) Average |r| index as a function of the number of hidden nodes. D) Variance of the |r| index across region pairs as a function of the number of hidden nodes. E-H: two-back task. A) Linear multivariate connectivity matrix. B) Nonlinear multivariate connectivity matrix for 3 hidden nodes. C) Average |r| index as a function of the number of hidden nodes. D) Variance of the |r| index across region pairs as a function of the number of hidden nodes.

Figure 4

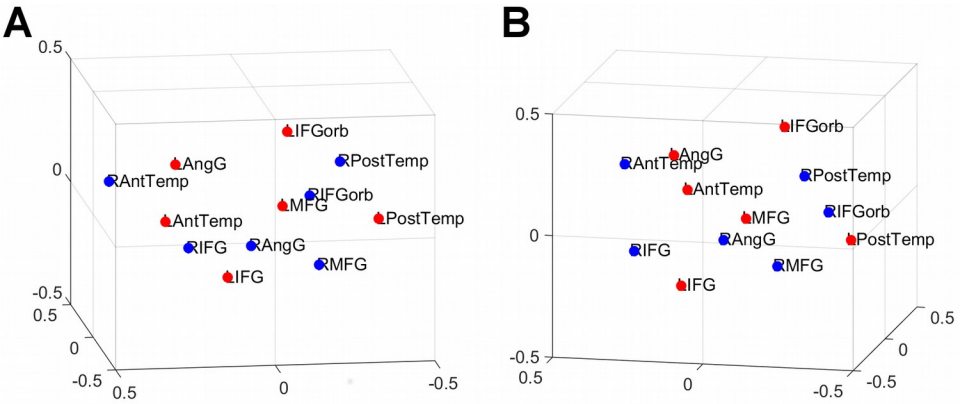
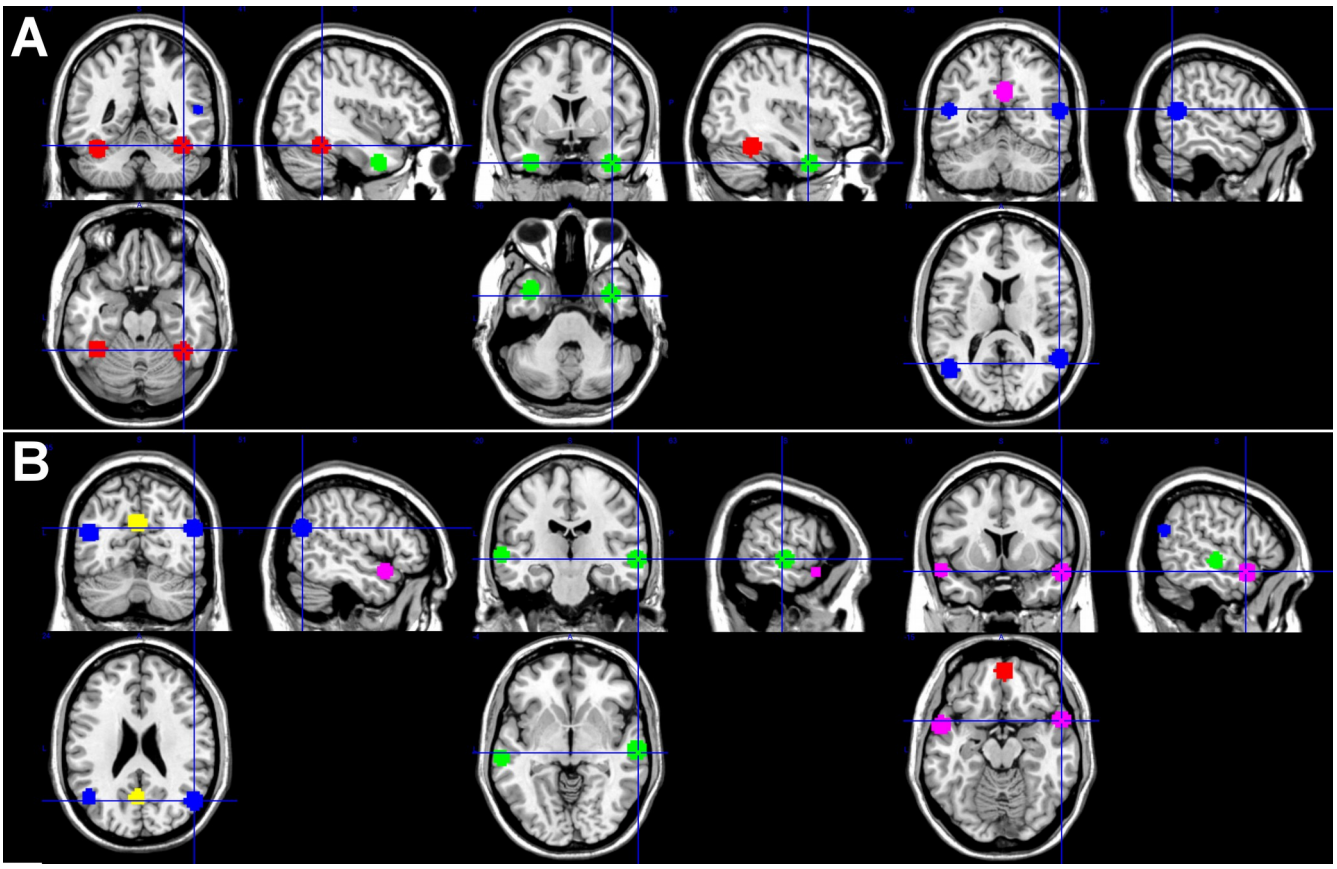


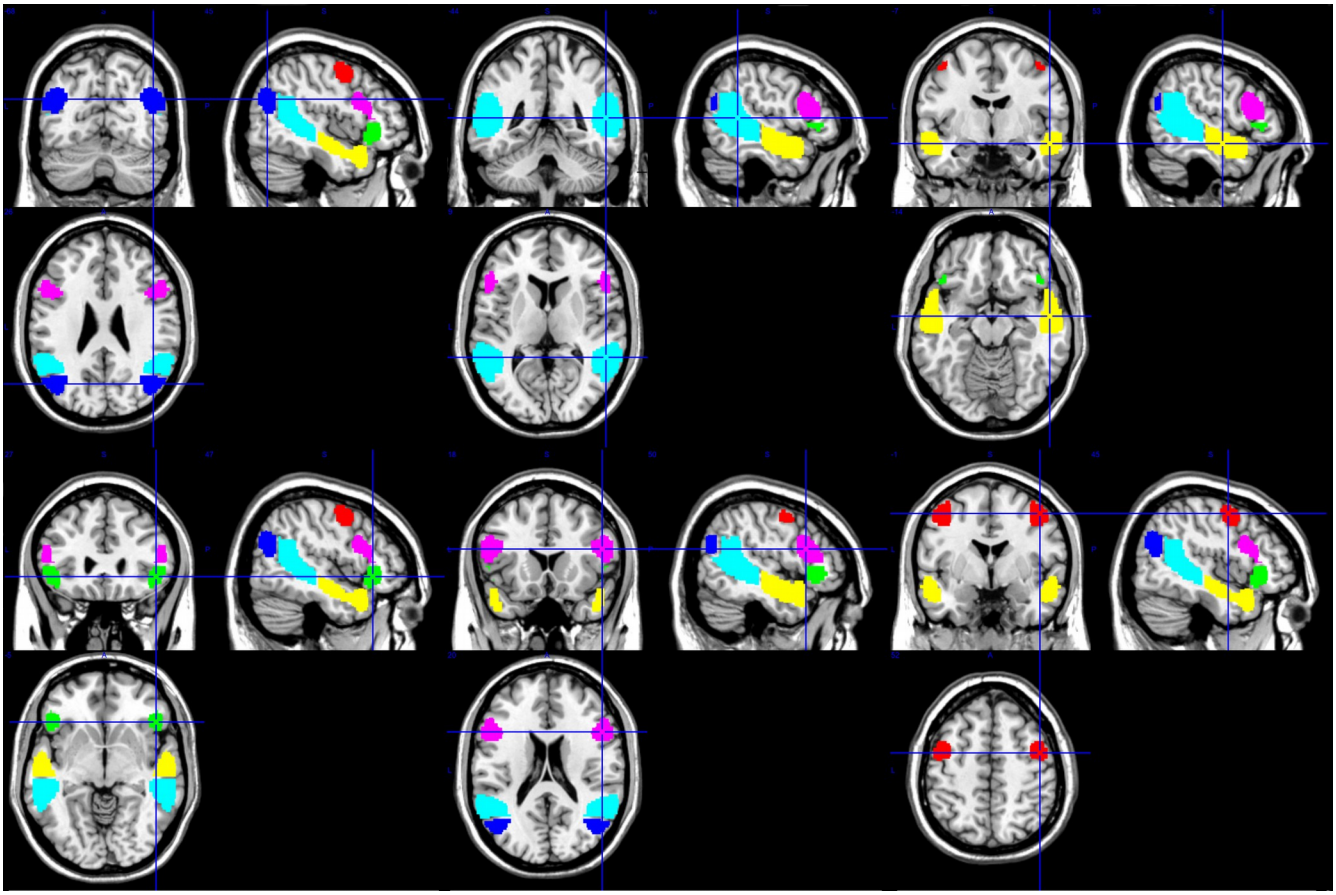
Fig. 4. A) 3D multidimensional scaling based on the connectivity matrix of the passive listening task. B) 3D multidimensional scaling based on the connectivity matrix of the two-back task. The brain regions are arranged similarly in space across the two tasks.

Supplementary Figure 1



Supp. Fig. 1. A) Face-selective ROIs. FFA (red), ATL (green), STS (blue), PC (magenta). B) Voice-selective ROIs. pSTG (blue), mSTG (green), aSTG (magenta), vmPFC (red), PC (yellow).

Supplementary Figure 2



Supp. Fig. 2. Search spaces for the language ROIs. AngG (blue), PostTemp (cyan), AntTemp (yellow), MFG (red), IFG (magenta), IFGOrb (green).