Optimizing fMRI Data Acquisition for Decoding Natural Speech with Limited Participants

Louis Jalouzot
CEA, ENS, Université Paris-Saclay, France
jalouzot.louis@gmail.com

Alexis Thual
karavela.ai, France

Yair Lakretz
ENS, EHESS, CNRS, Université PSL, France

Christophe Pallier
INSERM, CEA, CNRS, Université Paris-Saclay, France

Bertrand Thirion
INRIA, CEA, Université Paris-Saclay, France

Abstract

We investigate optimal strategies for decoding natural speech from fMRI data with limited participants[1]. Using data from LeBel et al. (2023) from 8 participants, we show that deep neural networks can effectively predict LLM-derived text representations with performance directly scaling with the amount of training data. Then, in this data regime, we observe that multi-subject training does not improve decoding accuracy compared to a single-subject approach. Furthermore, we find that our decoders better model syntactic than semantic features. Our results highlight deep phenotyping benefits and suggest multi-subject decoding needs more data per subject or a substantially larger cohort.

Keywords: fMRI; decoding; natural speech; deep learning; deep phenotyping

Introduction

Recent advances in neuroscience have demonstrated the feasibility of decoding complex percepts from brain activity. Using functional MRI (fMRI), several studies have achieved impressive decoding of auditory stimuli (Tang et al., 2023; Ye et al., 2025). These successes often rely on "deep phenotyping" datasets, where a large amount of data is acquired from each participant. A major challenge is to leverage data from multiple participants, and a key issue is to determine if inter-subject variability can be overcome.

In this work, we investigate optimal data acquisition strategies for decoding perceived natural speech from fMRI when only a limited number of participants are available. We train DNNs with a contrastive objective to predict LLM-derived text representations from fMRI. Our contributions are: (1) We show that this approach effectively decodes natural speech from fMRI. (2) We find that decoding performance scales with the quantity of data per participant. (3) We demonstrate that in a low-N regime (N=8), multi-subject training does not improve, and can even hinder, decoding performance compared to single-subject models.

Methods

Our goal is to learn a mapping from fMRI signals to high-level text representations.

Data and Task We use the fMRI dataset from LeBel et al. (2023), which contains recordings from eight participants listening to natural stories. The amount of data per participant is imbalanced, with three participants having significantly more data (∼16.5 hours each) than the others (∼6 hours each).

Preprocessing and Representations fMRI data were preprocessed using `fmriprep`, temporally smoothed, and standardized. We selected the top 4096 voxels for each subject based on their performance in a simple encoding model. For text, we used LLM2Vec (BehnamGhader et al., 2024) to generate 4096-dimensional embeddings of text chunks. To account for the hemodynamic delay, we introduced a lag ($\tau = 6s$) between the fMRI signal and the target text embedding. We also enriched text representations by including a context of preceding text ($c = 6s$).

[1] A preprint version of this work is available (c.f. Jalouzot et al. (2025)).

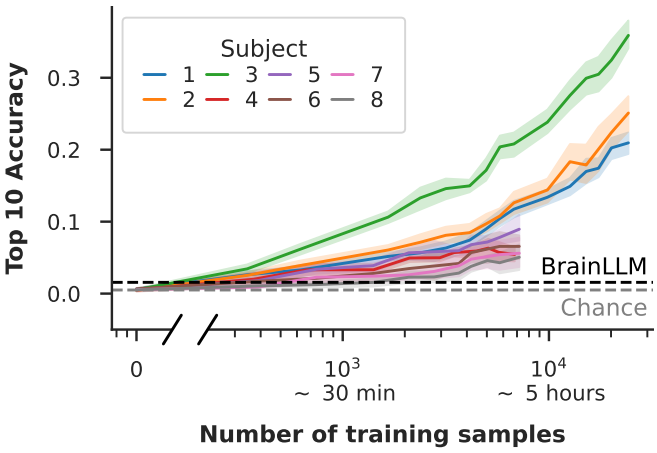

Figure 1: Impact of the amount of training data on single-subject decoding performance. Top-10 accuracy increases with more training data per subject, not yet plateauing at 13.5 hours.

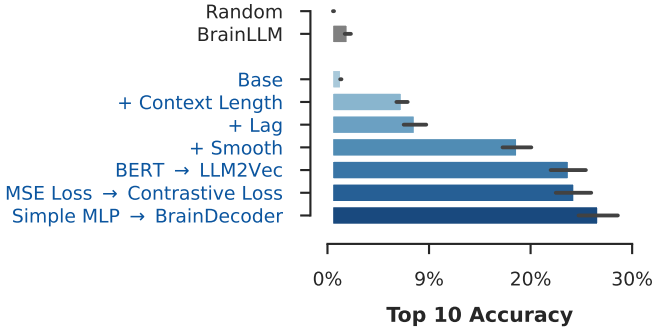

Figure 2: Impact of setup choices on decoding performance on the subjects with the most data (1, 2, and 3). Each modification (blue labels) incrementally improves top-10 accuracy.

Model and Training We trained a DNN composed of MLP layers, layer normalization, and skip connections, inspired by Scotti et al. (2024). The model was optimized using a contrastive loss (Radford et al., 2021). We compared two main setups:

- **Single-subject:** An independent decoder is trained for each participant.
- **Multi-subject:** A single decoder with a shared backbone is trained on data from all subjects. The first layer is subject-specific to project individual brain data into a common space.

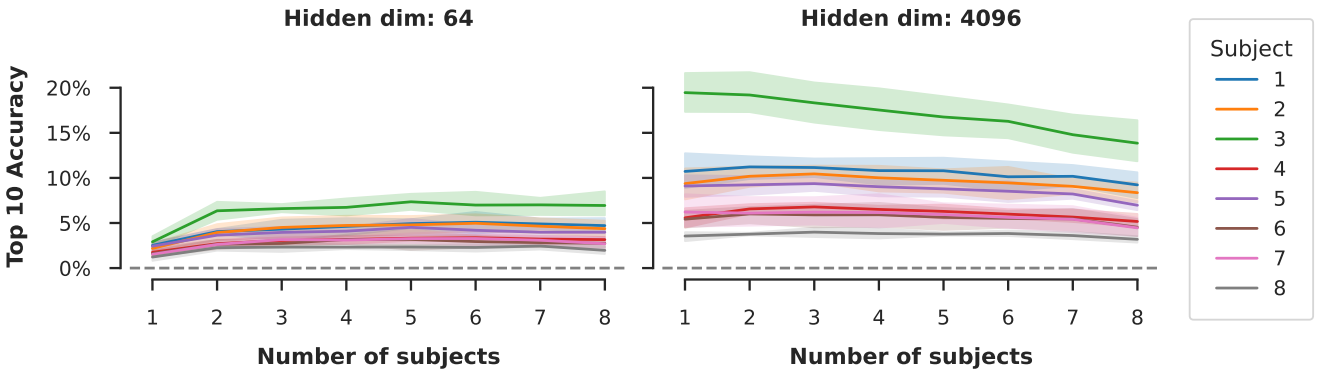

Figure 3: Impact of multi-subject training. For each subject (color), we plot the best decoding accuracy achieved when included in a multi-subject model of size N (x-axis). Performance does not improve over single-subject models (N=1), and can even degrade for larger models (right panel).

Evaluation Models were evaluated on a retrieval task. For a given test fMRI volume, the decoder predicts a text embedding. We then rank all embeddings in a large test set (approx. 2000 candidates) based on their cosine similarity to the prediction. Our primary metric is top-10 accuracy: the frequency at which the ground-truth embedding appears in the top 10 candidates.

Baseline As a point of comparison, we use a baseline derived from the work of Ye et al. (2025), who also used this dataset for a text generation task. We evaluated their predicted embeddings within our retrieval framework. This *BrainLLM* baseline achieves an average top-10 accuracy of 1.6%, serving as a strong reference point above chance level.

Results

Single-Subject Performance Scales with Data As shown in Figure 1, the performance of single-subject decoders strongly correlates with the amount of training data. For the three participants with the most training data ($\sim$13.5 hours), we achieved an average top-10 accuracy of 27%, with a peak of 36% for one subject. This performance is substantially above both chance level (0.05%) and the *BrainLLM* baseline of 1.6% (Figure 1). For participants with less training data ($\sim$4 hours), accuracy was lower (avg. 6%). Crucially, performance does not plateau, suggesting that acquiring even more data per individual would yield further improvements. This result provides strong evidence for the value of deep phenotyping.

Systematic Pipeline Optimization Our final decoding pipeline was the result of several incremental improvements. Figure 2 shows the contribution of each component. Starting from a basic MLP with MSE loss, we found that adding a hemodynamic lag, using contextualized text embeddings, switching from BERT representations to LLM2Vec, employing a contrastive loss, and using a more sophisticated DNN architecture all provided significant gains in performance.

Multi-Subject Training Does Not Improve Performance Contrary to what might be expected, training a multi-subject model did not improve decoding accuracy for individual subjects. As shown in Figure 3, performance for any given subject (colored lines) does not increase as more subjects are added to the training pool (x-axis). In fact, for larger models, performance often degrades compared to the single-subject baseline (N=1). In this data regime, the model struggles with inter-subject variability, and subject-specific patterns are better learned by dedicated models.

Decoder is More Sensitive to Syntax than Semantics To understand what linguistic features drive performance, we analyzed the best- and worst-decoded stories. A quantitative analysis revealed that our decoders are more sensitive to syntactic structure than to semantic content; the syntactic similarity between the ground-truth and retrieved text drops more sharply than semantic similarity for top candidates. This is corroborated by a qualitative analysis showing that stories with simpler, conversational syntax are decoded more accurately than those with complex sentences and abstract ideas. Current decoders, while effective, may be biased towards more superficial linguistic features, and future work should aim to better capture semantic nuances.

Discussion

The central result of our study is the clear superiority of a ”deep phenotyping” strategy over a multi-subject approach when working with a small cohort. Performance scales directly with the amount of data per participant, while pooling data across subjects fails to provide a benefit, likely due to high inter-subject variability in the neural representation of language. This has strong practical implications for experimental design in fMRI-based decoding: for studies with a limited number of participants, resources are better spent maximizing the data collected from each individual rather than increasing the number of subjects with less data each.

Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011016055).

References

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., & Reddy, S. (2024). LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders.. Retrieved 2025-06-12, from https://openreview.net/forum?id=IW1PR7vEBf#discussion

Jalouzot, L., Thual, A., Lakretz, Y., Pallier, C., & Thirion, B. (2025). *Optimizing fMRI Data Acquisition for Decoding Natural Speech with Limited Participants.* Retrieved 2025-06-12, from https://arxiv.org/abs/2505.21304v1

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., ... Huth, A. G. (2023, August). A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, *10*(1), 555. Retrieved 2023-09-17, from https://www.nature.com/articles/s41597-023-02437-z doi: 10.1038/s41597-023-02437-z

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021, February). *Learning Transferable Visual Models From Natural Language Supervision.* arXiv. Retrieved 2023-09-17, from http://arxiv.org/abs/2103.00020 doi: 10.48550/arXiv.2103.00020

Scotti, P. S., Tripathy, M., Villanueva, C. K. T., Kneeland, R., Chen, T., Narang, A., ... Abraham, T. M. (2024, March). *MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data.* arXiv. Retrieved from http://arxiv.org/abs/2403.11207

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023, May). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, *26*(5), 858–866. Retrieved 2023-09-17, from https://www.nature.com/articles/s41593-023-01304-9 (Number: 5 Publisher: Nature Publishing Group) doi: 10.1038/s41593-023-01304-9

Ye, Z., Ai, Q., Liu, Y., de Rijke, M., Zhang, M., Lioma, C., & Ruotsalo, T. (2025). Generative language reconstruction from brain recordings. *Communications Biology*, *8*(1), 346.