The role of context in neural representational alignment to audio- and text-based language systems

Marianne de Heer Kloots (m.l.s.deheerkloots@uva.nl)
Institute for Logic, Language, & Computation, University of Amsterdam, Science Park 107
1098 XG Amsterdam, The Netherlands

Josef Parvizi (jparvizi@stanford.edu)
Department of Neurology and Neurological Sciences, Stanford University, 300 Pasteur Dr
Palo Alto, CA 94304 United States

Laura Gwilliams (laura.gwilliams@stanford.edu)
Department of Psychology; Stanford Data Science; Wu Tsai Neurosciences Institute,
Stanford University, 450 Jane Stanford Way
Stanford, CA 94305 United States

Abstract

Speech understanding requires integrating the current input with surrounding context. Prior research has found that increasing context size in artificial text-based language systems leads to improved predictivity of human brain activity. Here, we investigate (i) how the *type* of context (unidirectional; bidirectional) influences brain alignment; (ii) how the information contained in speech and text model embeddings changes as a function of context size and context type; (iii) what changes in model representations could explain brain alignment. We recorded intracranial EEG of participants listening to audiobooks, and extracted corresponding layerwise embeddings from a speech model (Wav2Vec2) and a language model (RoBERTa) under different context sizes and types. We find that context *type* rather than size has the biggest influence on the linear decodability of linguistic structure, the intrinsic dimensionality of the underlying representations, and ultimately, brain alignment. This work represents an important step towards understanding the representational basis of model-brain alignment, and identifies context type as an important driver of models extracting brain-relevant information.

Keywords: representational alignment; speech comprehension; audio- and text-based models; intracranial EEG

Introduction

Language understanding – in both biological and artificial systems – involves continuously integrating incoming input with available contextual information (Gwilliams, King, Marantz, & Poeppel, 2022; Heilbron, Armeni, Schoffelen, Hagoort, & de Lange, 2022). It has been proposed that this common processing strategy leads to the the success of artificial language systems in predicting human brain activity during speech and text processing (Vaidya, Jain, & Huth, 2022; Schrimpf et al., 2021). In line with this, model-brain alignment improves as more context is provided to the model (Abnar, Beinborn, Choenni, & Zuidema, 2019; Toneva & Wehbe, 2019; Anderson, Davis, & Lalor, 2024).

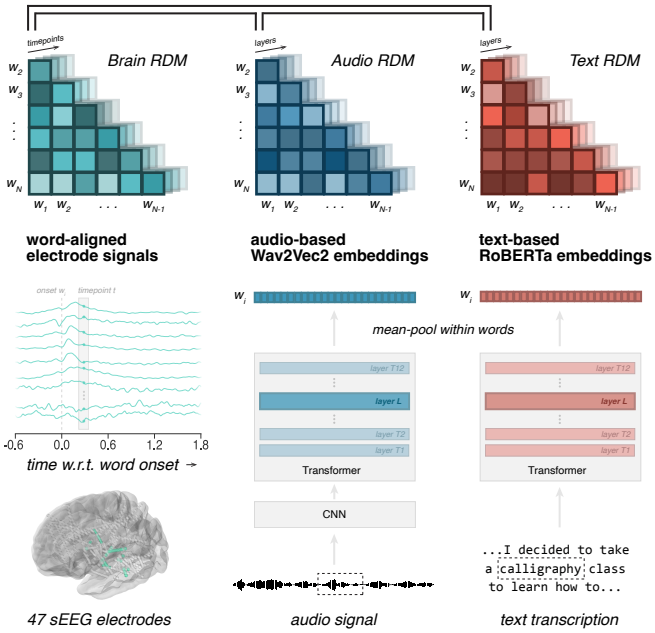

Figure 1: We compute Representational Dissimilarity Matrices (RDMs) for word-aligned sEEG data (across timepoints), and for audio- and text-based word embeddings extracted from Transformer models (across layers). We then compare RDMs within and across language processing systems.

The observation that adjusting context affects brain alignment implies that model representations change to become more 'brain-like' with increased context. In addition to context size, context can also differ in *kind*. For example, a model may only consider the past (unidirectional context), or also the future (bidirectional context). The human brain has been shown to use both past and future inputs to derive ultimate understanding (Gwilliams, Linzen, Poeppel, & Marantz, 2018).

Here, we aim to improve our understanding of what explains representational alignment across language processing systems, by studying different *types* and different *sizes* of context. What contextually driven differences in speech- and text-

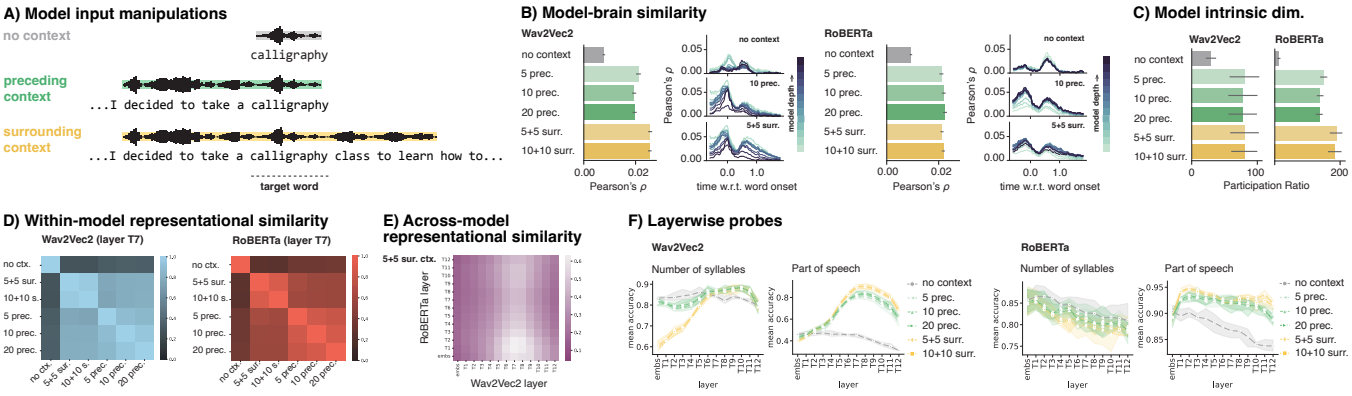

Figure 2: **A)** Different context windows provided alongside the target word as model input (shown are the no context, 5-preceding and 5+5 surrounding conditions). **B)** Context effects on model-brain similarity, across layers and sEEG timepoints (right), and averaged over layers and timepoints for all context windows (left). **C)** Context effects on model intrinsic dimensionality do not fully mirror those on model-brain similarity. **D)** Within-model representational similarities are larger between similar context types than between similar context sizes. **E)** Across models, similarity peaks between late Wav2Vec2 and early RoBERTa layers. **F)** Model encoding of part-of-speech categories benefits more from contextualization than model encoding of word length (in syllables).

model representations of linguistic input have consequences for their alignment to human neural activity?

Methods

Neural data Stereo-EEG (sEEG) data was recorded while participants listened to audiobook snippets. We extracted epochs between 600 ms before and 1800 ms after word onset from aggregated data across 1082 electrodes and 7 participants. From these, we selected 47 electrodes for analysis, which showed highest activity within 400 ms after word onset.

Transformer model embeddings Based on the audiobook recordings and aligned text transcriptions, we extracted layerwise word-level embeddings from one audio-based (Wav2Vec2-base; Baevski, Zhou, Mohamed, & Auli, 2020) and one text-based (RoBERTa-base; Liu et al., 2019) Transformer model. Both models are pre-trained with a bidirectional masked objective; Wav2Vec2 operating on audio frame representations, and RoBERTa on sub-word text token embeddings. For both models, we extract the input embeddings and hidden state representations at each Transformer layer, mean-pooling across audio frame and text token representations within each word (Fig. 1). We provide each model with three different types and several different sizes of linguistic context to investigate contextualization effects across the generated embeddings (Fig. 2A).

Representational analysis techniques We compute Pearson's correlations between cosine-distance RDMs to quantify representational alignment between models, as well as model-brain alignment (Fig. 1). Alignment to the sEEG activity is computed separately for 241 timepoints across the epoch. To quantify the intrinsic dimensionality of model representations we computed Participation Ratios (Gao et al., 2017), and to quantify the decodability of linguistic features across model layers we trained multinomial logistic regression probes.

Results & Discussion

Across our analyses, we observe large differences between model representations of isolated words on the one hand, and representations that integrate linguistic context on the other (Fig. 2BCD). Furthermore, differences are generally largest between context type (i.e. preceding, surrounding or no context) rather than context size, suggesting that representational alignment is not simply driven by the amount of shared context. While studies of text-based models have found that layerwise brain-alignment scores closely follow geometric measures of model intrinsic dimensionality (ID; Antonello & Cheng, 2024), we find a different pattern when comparing across context conditions. Both brain-similarity (Fig 2B) and ID (Fig 2C) increase when models integrate linguistic context; however, surrounding context increases model-brain alignment but not ID for Wav2Vec2, and vice versa for RoBERTa. Hence, context effects on model-brain alignment do not seem to be purely driven by intrinsic dimensionality either.

To further explore what linguistic properties might benefit from contextualization and potentially drive differences in brain alignment, we investigate the encoding of linguistic features across the layers of each model's contextualization module (i.e. the Transformer). We find considerable similarity between late layers of the audio-based Wav2Vec2 model and early layers of the text-based RoBERTa model, showing that model internal representations can align when processing similar linguistic content, despite differing in modality (Fig. 2E). When probing the decodability of linguistic features across model layers, we find that the same layers peak for more abstract features, such as part-of-speech category (Fig. 2F). Moreover, these features benefit most from the integration of linguistic

context: while more surface-level features such as word length are similarly decodable from isolated word vs. contextualized representations, higher-level features show larger differences between these conditions, which increase over model layers.

Our current results show that linguistic context shapes the internal representations of audio- and text-based Transformer models, as well as their alignment to human brain activity in naturalistic speech comprehension. Contextualization benefits model-brain alignment as well as the decodability of more abstract (but not surface-level) linguistic features from model internals. This raises the question of what contextual linguistic features actually drive more brain-like model representations. Future work could investigate causal effects by ablating specific contextual features from model input (Kauf, Tuckute, Levy, Andreas, & Fedorenko, 2023) or removing their decodability from model-internal representations (Oota, Gupta, & Toneva, 2023). Furthermore, we plan to investigate the temporal dynamics of context effects on model-brain alignment (Fig. 2B), as well as context effects on features below the word level.

References

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019). Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 191–203). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/W19-4820

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

Antonello, R., & Cheng, E. (2024). Evidence from fMRI Supports a Two-Phase Abstraction Process in Language Models. In *Proceedings of UniReps 2024: 2nd edition of the Workshop on Unifying Representations in Neural Models.*

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 12449–12460). Curran Associates, Inc.

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., & Ganguli, S. (2017). *A theory of multineuronal dimensionality, dynamics and measurement.* bioRxiv. doi: 10.1101/214262

Gwilliams, L., King, J.-R., Marantz, A., & Poeppel, D. (2022). Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature Communications*, *13*(1), 6606. doi: 10.1038/s41467-022-34326-1

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In Spoken Word Recognition, the Future Predicts the Past. *Journal of Neuroscience*, *38*(35), 7585–7599. doi: 10.1523/JNEUROSCI.0065-18.2018

Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., & de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2201968119

Kauf, C., Tuckute, G., Levy, R., Andreas, J., & Fedorenko, E. (2023, September). Lexical-Semantic Content, Not Syntactic Structure, Is the Main Contributor to ANN-Brain Similarity of fMRI Responses in the Language Network. *Neurobiology of Language*, 1–36. doi: 10.1162/nol_a_00116

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv. doi: 10.48550/arXiv.1907.11692

Oota, S., Gupta, M., & Toneva, M. (2023). Joint processing of linguistic properties in brains and language models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 18001–18014).

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). (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.2105646118

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc.

Vaidya, A. R., Jain, S., & Huth, A. (2022). Self-Supervised Models of Audio Effectively Explain Human Cortical Responses to Speech. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 21927–21944). PMLR.