ADVANCED MEG ANALYSIS OF AUDITORY AND LINGUISTIC ENCODING IN SPOKEN LANGUAGE PROCESSING

Matteo Ciferri
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ciferri@students.uniroma2.eu

Matteo Ferrante
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
matteo.ferrante@uniroma2.it

Nicola Toschi
University of Rome, Tor Vergata
Department of Biomedicine and Prevention
A.A. Martinos Center for Biomedical Imaging
Harvard Medical School/MGH, Boston (US)

ABSTRACT

In this work, we explore brain responses related to language processing using neural activity elicited from auditory stimuli and measured through Magnetoencephalography (MEG). We develop audio (i.e. stimulus)-MEG encoders using both time-frequency decompositions and latent representations based on wav2vec2 embeddings, and text-MEG encoders based on CLIP and GPT-2 embeddings, to predict brain responses from audio stimuli only. The analysis of MEG signals reveals a clear encoding pattern of the audio stimulus within the MEG data, highlighted by a strong correspondence between real and predicted brain activity. Brain regions where this correspondence was highest were lateral (vocal features) and frontal (textual features from CLIP and GPT-2 embeddings).

1 INTRODUCTION

In recent years, the field of computational neuroscience has seen significant advancements in understanding how the brain processes language. While much of the existing research in brain encoding and decoding (Goldstein et al., 2022; Tang et al., 2023) relies on functional Magnetic Resonance Imaging (fMRI) data, this modality is somewhat limited, amongst other factors, by its low temporal resolution. In contrast, the temporal resolution offered by Magnetoencephalography (MEG), despite other limitations (e.g. lower sensitivity in deep brain structures), can provide a more detailed and dynamic insight into neural mechanisms underlying language comprehension and production. In this work, we aimed to further develop so called encoding models to advance our understanding of language processing through the lens of MEG data. An encoding model is a computational framework designed to map input stimuli to corresponding (elicited) neural activity patterns. We developed audio-to-MEG encoders using two types of representations for audio data, i.e. time-frequency decompositions derived from Short-time Fourier Transform (STFT) (Griffin & Lim, 1984), and latent spaces generated by the wav2vec2 library (Baevski et al., 2020). Additionally, we built text-to-MEG encoders that incorporate embeddings from the Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021) or GPT-2 (Radford et al., 2019) and compared the encoding performance between pipelines (Figure 1). This comparison was performed with the goal of gaining insight into the neural processes involved in auditory and linguistic perception and advancing the computational strategies used for interpreting complex neural signals. Code available at: https://github.com/mattciff5/spect-to-meg.

2 Related Work

So far, research in brain encoding for speech and language processing has primarily used functional Magnetic Resonance Imaging (fMRI) (Huth et al., 2012; Antonello et al., 2023; Caucheteux et al., 2023). These studies have contributed to the development of both linear and nonlinear models that map stimuli to brain activity from fMRI signals. Previous work focuses on e.g. enhancements in network scaling and uncovering correlations in auditory and semantic processing areas (Caucheteux & King, 2022). However, limitations in the temporal resolution of fMRI have led researchers to explore MEG data collected during exposure to auditory stimuli. On the encoding side, Oota et al. (2023) developed a model using Bidirectional Encoder Representations from Transformers (BERT) contextual embeddings (Devlin et al., 2018) to predict MEG signals. In terms of decoding, one paper (Défossez et al., 2023) successfully reconstructed audio from MEG signals through contrastive learning which was based on aligning signals with the latent space generated by the wav2vec2 library (Baevski et al., 2020). These efforts demonstrate the potential of MEG-related neural data to reconstruct the stimulus that has generated it. Our study builds upon these developments, aiming to augment encoding models by building both semantic and speech representations of the brain.

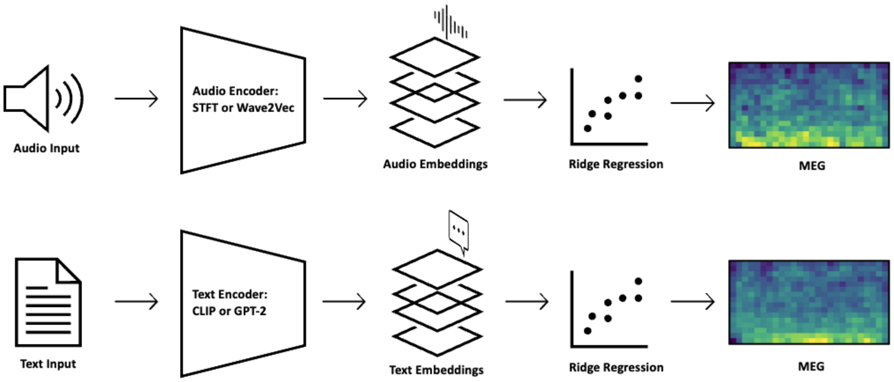

Figure 1: Schematic representation of the encoding pipeline. Left: initial input stimuli (audio). Center: two different encoders individually process the stimuli to generate embeddings. Right: regression process to predict MEG time-frequency decompositions using the embeddings. Bottom: repetition of the pipeline for text stimuli, mirroring the process for audio stimuli at the top.

3 Material and Methods

We used data from the MEG-MASC dataset (Gwilliams et al., 2023), specifically selecting 8 subjects as in the study by Oota et al. (2023). The dataset includes recordings from 208 MEG sensors as the subjects listened to a series of naturalistic spoken stories, selected from the Open American National Corpus, namely "*Cable Spool Boy*", "*LW1*", "*Black willow*", and "*Easy money*". We applied separate sampling rates to the two data types, in particular 16,000 Hz for the audio input and 1,000 Hz for the MEG. For pre-processing the raw MEG data, we employed the *MNE-Python* library (Appelhoff et al., 2019). Our methodology involved several key steps: a) bandpass filtering (0.5-30.0 Hz) (Marzetti et al., 2013) b) segmentation into time windows (length = 3 s) which begin in correspondence with every word (stimulus) onset, typically encompassing an average of 5 words; c) window-wise baseline correction of 200 ms before the stimulus d) channel-wise clipping between the fifth and ninety-fifth percentile. The whole pipeline resulted in 3200 time points for each of the 208 sensors and for each subject.
We employed time-frequency decompositions (i.e. spectrograms) as a unified representational for both the input (vocal signal) and the output (MEG signal). Spectrograms of the audio and MEG signals were generated using Short-Time Fourier Transform (STFT) applied to 3-second speech epochs. These segments were defined based on words marked by temporal onset and spanning the specified duration. This approach ensures temporal alignment between audio segments and the corresponding MEG data. The other approach employed for audio encoding is the pre-trained wav2vec2 model.

For this model, we processed inputs comprising the 3-second audio epochs, each sampled at a frequency of 16,000 Hz, obtaining 48000 time points. Epochs are then encoded into a $149D \times 768D$ embedding matrix from the last hidden layer.
We designed an approach for text encoders to link each MEG epoch to a corresponding linguistic phrase, thus forming sequences that represent the linguistic context for each set of MEG data as follows. Each epoch incorporated a contextual span of 20 preceding words (past context) and 5 words present in the 3 s MEG window under investigation. For the encoding of these sentences, we used the tokenizer from the pre-trained CLIP model. The tokenizer transformed sentences into a format suitable for the machine learning model, subsequently processed into final CLIP embeddings. We accommodated sentence structure and additional linguistic elements by including padding, beginning-of-stream (BOS), and end-of-stream (EOS) tokens. Each sentence is represented by a $33D \times 512D$ matrix, derived from the final hidden layer of the model. The first dimension corresponds to the encoded sentence, comprising both word tokens and padding tokens, and the second is the embedding dimension. We also incorporated the GPT-2 model for text analysis, applying a processing methodology similar to what is described above. A key distinction between the textual models, however, lies in the dimensional structure of their embeddings. Specifically, the GPT-2 model has a more expansive size in its last hidden layer, with a feature vector of length 768. This larger scale in the embedding space allows for a potentially richer and more nuanced representation of textual data. The most recent versions of GPTs were not used as they do not provide access to the embeddings.
Audio and text features were then used in encoding models to predict brain responses. As in e.g. Oota et al. (2023), we opted for ridge regression as our encoding model. The objective function of ridge regression is expressed as $f(X_s) = \min \|Y_b - X_s W_s\|_F^2 + \lambda \|W_s\|_F^2$. Here, X_s represents the input stimuli representation, $W_s \in \mathbb{R}^{F_s \times L}$ are the learnable weights, with F_s denoting the stimulus representation features (depending on audio or text input) and L the number of MEG sensors. The sample stimulus $s \in \mathbb{R}^{F_s}$, $\|.\|_F$ indicates the Frobenius norm, and $\lambda > 0$ is the regularization weight, a tunable hyper-parameter.
The dataset underwent subject-wise splitting into training and test sets. Specifically, the collection of 3-second MEG windows starting from the word stimulus, was divided for each participant with 70% allocated for training and 30% for testing. Training procedures involved leave-one-out cross-validation on the training set. Optimization of the parameter λ, was conducted by exploring different values (1, 10, 500, 5000). Following cross-validation, the model was retrained on the entire training split using the best-performing hyperparameter (5000).

4 Results

We evaluated the reconstructed time-frequency decompositions across the full spectrum (0.5-30 Hz) as well as for individual frequency bands below 30 Hz, which include delta, theta, alpha, and beta bands (Abhang et al., 2016). Delta frequencies are typically between 0.5 and 4 Hz, often associated with deep sleep or states of unconsciousness. Theta frequencies are generally between 4 and 8 Hz, associated with states between wakefulness and sleep. Alpha frequencies are typically from 8 to 12 Hz referring to relaxed, calm states while awake, and finally, beta frequencies between 12 and 30 Hz are linked to active, busy, or anxious thinking and concentration. Our evaluation focused on the computation of several key statistical metrics to assess the precision of our predictions of MEG spectrograms from audio data. These metrics encompassed the Pearson Correlation (Siems et al., 2016) and the coefficient of determination R², computed among every real and predicted pair of time-frequency decompositions after flattening both time and frequency dimensions. Evaluations were conducted for each sensor location and frequency band. The following figures show the anatomical distribution of the Pearson Correlation. The highest performances were observed in predominantly lateral areas for the audio encoders (Figure 2) and frontal regions for the textual models.

Table 1 shows an overview of all our results. We averaged the R² scores and the Pearson Correlation (PC) across all sensors and subjects for each specific frequency band. We observed that utilizing textual embeddings for MEG encoding resulted in improved accuracy metrics, compared to using vocal feature representations. However, it is noteworthy that the latter exhibited significant activation in distinct brain regions despite its comparatively lower performance.

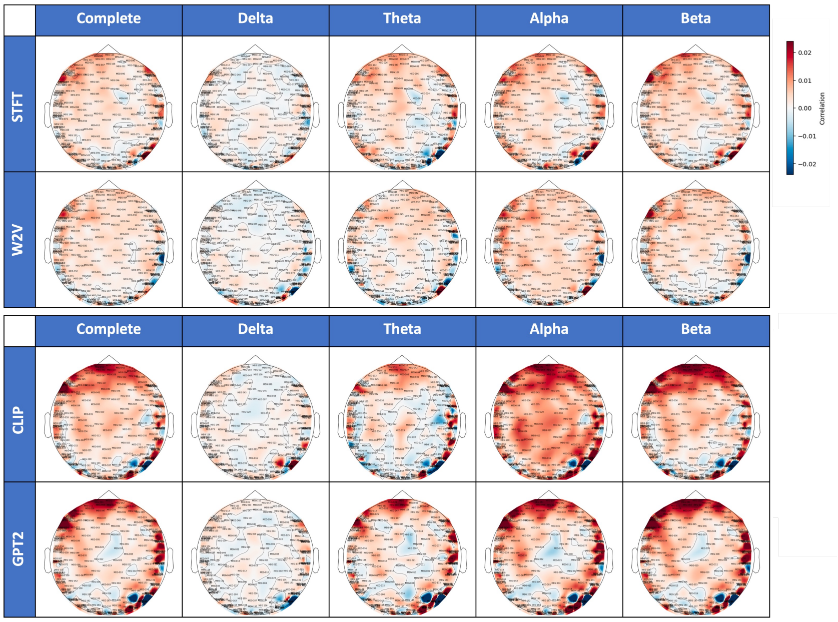

Figure 2: Pearson Correlation topography maps, visualizing of the neural encoding model's performance across different sensors and frequency bands. In the case of audio models, high values of correlation occur in lateral brain areas, while textual models exhibit significant performance also in frontal regions. The performance decreases notably across frequency bands, particularly in lower frequencies, which are typically associated with states of rest or sleep rather than concentration and cognitive processing.

Table 1: Comparative Pearson correlation and R² results from audio and text encoders.

Band	**Input**	**Model**	**PC** (10^{-3})		**R2** (10^{-4})	
			mean	**std.**	**mean**	**std.**
Complete	Audio	STFT	2.70	2.14	0.30	0.35
		wav2vec2	2.56	2.08	0.26	0.30
	Text	CLIP	5.60	4.80	1.87	3.19
		GPT-2	6.11	5.38	1.75	2.92
Delta	Audio	STFT	0.63	1.38	0.04	0.09
		wav2vec2	0.35	1.35	0.02	0.09
	Text	CLIP	1.31	1.77	0.10	0.15
		GPT-2	0.10	1.67	0.07	0.14
Theta	Audio	STFT	2.28	2.44	0.28	0.35
		wav2vec2	1.63	2.27	0.18	0.27
	Text	CLIP	1.52	4.62	0.64	1.56
		GPT-2	4.38	4.86	1.05	1.84
Alpha	Audio	STFT	3.12	2.52	0.42	0.45
		wav2vec2	3.85	2.44	0.53	0.49
	Text	CLIP	9.12	5.11	3.00	3.12
		GPT-2	6.65	5.95	2.20	3.12
Beta	Audio	STFT	3.10	2.56	0.43	0.53
		wav2vec2	2.84	2.57	0.35	0.46
	Text	CLIP	6.29	6.28	2.75	5.51
		GPT-2	7.45	6.61	2.70	3.84

In order to consolidate our results, we constructed a null distribution simulating the null hypothesis that the mean R² values obtained from the predicted MEG are not significantly different from zero, by randomly permuting the time-frequency decompositions derived from the encoding models.

Subsequently, we computed the R² values between the real MEG spectrograms and the permuted estimates. This process was iterated 30 times to derive the distribution of R², following the approach outlined in Tang et al. (2023).
This test was conducted for each model across the entire frequency band of 0-30 Hz. The null distribution was computed for each subject and each channel. Subsequently, z-scores and p-values were evaluated for each channel (Figure 3, averaging across subjects) and for each subject (Figure 4, averaging across channels), leading to the rejection of the null hypothesis and confirming the non-randomness of the results described above.

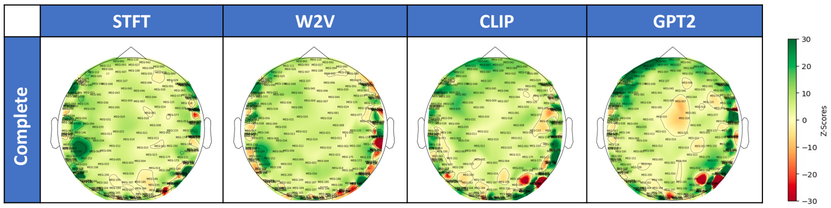

Figure 3: Z-score topography maps, visualizing the neural encoding model's z-score values across different sensors. Higher values mean a greater distance from the mean of the null distribution.

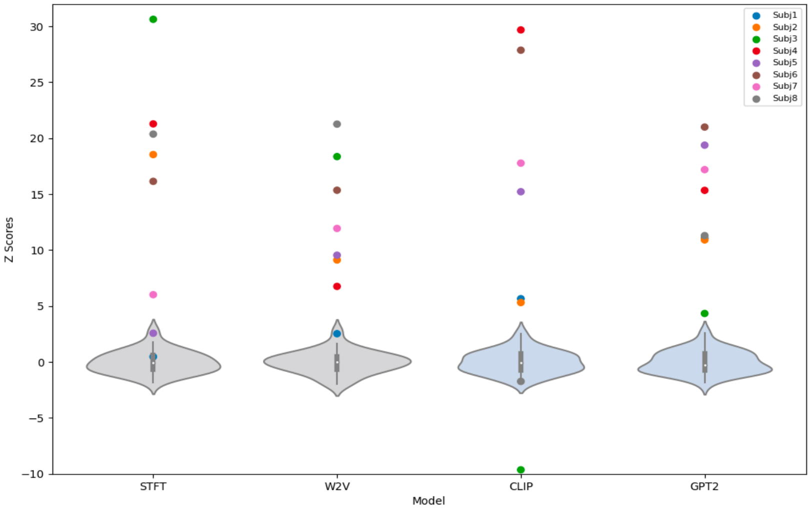

Figure 4: Violin plots illustrate the null distribution for each encoding model, with individual dots representing the z-score of each subject. The distance of each dot (i.e. single subject) from the mean reflects the degree of non-randomness in the predicted brain activity, emphasizing the robustness of the findings.

Table 2: P-values resulting from the analysis conducted with audio and text encoders, averaged across subjects and channels.

Band	**Input**	**Model**	**P-Value**	
			mean	**signif.**
Complete	Audio	STFT	0.7%	<5%
		wav2vec2	1.2%	<5%
	Text	CLIP	0.5%	<5%
		GPT-2	0.7%	<5%

5 Discussions and Conclusion

The examination across frequency bands illustrated by topography maps, reveals significant performance of audio and textual encoders in lateral and frontal areas, respectively, highlighting the nuanced interplay between neural encoding models and sensory data. The neural processing of both natural language and acoustic features exhibits a strong correlation with specific regions within the cerebral cortex responsible for the comprehension and production of spoken and written language. The use of advanced machine learning models, while considering their limitations and biases, will remain a key focus in our efforts to unravel the complexities of neural language processing and its applications. The application of larger, audio or text pre-trained models may influence the outcomes of neural representations, starting from more brain-like features (Antonello et al., 2023), suggesting a potential area for refinement in future studies. The inclusion of a broader range of subjects and the integration of multimodal data represent exciting avenues for future research. Such expansions would not only enhance the robustness of our findings but also pave the way for a more nuanced understanding of neural processes. Moreover, the potential application of our findings in predicting time series data within neural studies opens up new possibilities for advancing the field. Discussing the potential clinical applications of the research findings, such as in diagnosing language disorders or designing neurofeedback interventions, would highlight the translational significance of the study. As bidirectional brain-model mappings grow increasingly powerful, ethical concerns, especially around privacy and misuse, become crucial in neural data studies. It is essential to handle encoding and decoding carefully to prevent biases and protect personal thoughts, underscoring the need for strict ethical guidelines to ensure responsible and privacy-conscious neural research advancements.

6 Acknowledgements

This work is supported by NEXTGENERATIONEU (NGEU); the Ministry of University and Research (MUR); the National Recovery and Resilience Plan (NRRP); project MNESYS (PE0000006, to NT) - A Multiscale integrated approach to the study of the nervous system in health and disease (DN. 1553 11.10.2022); the MUR-PNRR M4C2I1.3 PE6 project PE00000019 Heal Italia (to NT); the NATIONAL CENTRE FOR HPC, BIG DATA AND QUANTUM COMPUTING, within the spoke “Multiscale Modeling and Engineering Applications” (to NT); the European Innovation Council (Project CROSSBRAIN - Grant Agreement 101070908, Project BRAINSTORM - Grant Agreement 101099355); the Horizon 2020 research and innovation Programme (Project EXPERIENCE - Grant Agreement 101017727).

References

Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra. Chapter 2 - technological basics of eeg recording and operation of apparatus. In Priyanka A. Abhang, Bharti W. Gawali, and Suresh C. Mehrotra (eds.), *Introduction to EEG- and Speech-Based Emotion Recognition*, pp. 19–50. Academic Press, 2016. ISBN 978-0-12-804490-2. doi: https://doi.org/10.1016/B978-0-12-804490-2.00002-6. URL https://www.sciencedirect.com/science/article/pii/B9780128044902000026.

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. Scaling laws for language encoding models in fMRI, December 2023. URL http://arxiv.org/abs/2305.11863. arXiv:2305.11863 [cs].

Stefan Appelhoff, Matthew Sanderson, Teon L Brooks, Marijn van Vliet, Romain Quentin, Chris Holdgraf, Maximilien Chaumon, Ezequiel Mikulan, Kambiz Tavabi, Richard Höchenberger, et al. Mne-bids: Organizing electrophysiological data into the bids format and facilitating their analysis. *Journal of Open Source Software*, 4(44), 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, December 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441, March 2023. ISSN 2397-3374. doi: 10.1038/s41562-022-01516-2. URL https://www.nature.com/articles/s41562-022-01516-2. Number: 3 Publisher: Nature Publishing Group.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, October 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00714-5. URL http://arxiv.org/abs/2208.12266. arXiv:2208.12266 [cs, eess, q-bio].

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-022-01026-4. URL https://www.nature.com/articles/s41593-022-01026-4.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

Laura Gwilliams et al. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific Data*, 10(1):862, 2023.

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, December 2012.

Laura Marzetti, Stefania Della Penna, Abraham Z Snyder, Vittorio Pizzella, Guido Nolte, Francesco de Pasquale, Gian Luca Romani, and Maurizio Corbetta. Frequency specific interactions of meg resting state activity within and across brain networks as revealed by the multivariate interaction measure. *Neuroimage*, 79:172–183, 2013.

Subba Reddy Oota, Nathan Trouvain, Frederic Alexandre, and Xavier Hinaut. Meg encoding using word context semantics in listening stories. In *INTERSPEECH 2023-24th INTERSPEECH Conference*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Marcus Siems, Anna-Antonia Pape, Joerg F Hipp, and Markus Siegel. Measuring the cortical correlation structure of spontaneous oscillatory activity with eeg and meg. *NeuroImage*, 129:345–355, 2016.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, May 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL https://www.nature.com/articles/s41593-023-01304-9. Number: 5 Publisher: Nature Publishing Group.