nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00714-5

Decoding speech perception from non-invasive brain recordings

In the format provided by the authors and unedited

Table A.1: **Top-1 Accuracy.** Segment-level accuracy related to Table 2 in the main paper.

Model	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
Random model	0.5±0.0	0.1±0.0	0.1±0.0	0.1±0.0
Base Model	0.7±0.2	0.1±0.0	3.0±0.3	5.8±0.6
+ Contrastive	0.9±0.6	2.1±0.4	26.2±0.6	25.1±0.6
+ Deep Mel	**5.2**±1.1	4.2±0.7	34.8±1.2	31.6±1.0
+ wav2vec 2.0	**5.2**±0.8	**5.0**±0.4	**41.3**±0.1	**36.8**±0.4

Table A.2: **Ablations of the brain module.** Top-10 segment-level accuracy (%) for our model and its ablated versions. Stars indicate significant gain ($p < 0.001$) across participants.

Model	Broderick (EEG)	Brennan (EEG)	Schoffelen (MEG)	Gwilliams (MEG)	delta	p-val
Our model	**17.7** ± 0.6	25.7 ± 2.9	**67.5** ± 0.4	**70.7** ± 0.1		
- Spatial attention dropout	16.0 ± 1.7 *	**26.8** ± 0.7	67.5 ± 0.2	69.0 ± 0.2 *	0.4	0.009
- GELU + ReLU	16.4 ± 0.1	24.6 ± 2.1	65.8 ± 0.6 *	68.8 ± 1.3 *	1.6	$< 10^{-18}$
- Final convs	14.2 ± 1.1 *	19.0 ± 4.4 *	67.5 ± 0.3	68.9 ± 0.9 *	1.1	$< 10^{-10}$
- Non-residual GLU conv	8.4 ± 6.8 *	6.0 ± 0.2 *	67.0 ± 0.2	70.2 ± 0.2	1.6	$< 10^{-10}$
- Skip connections	13.9 ± 2.0 *	24.2 ± 2.7	65.4 ± 0.4 *	66.2 ± 0.3 *	2.4	$< 10^{-21}$
- Initial 1x1 conv	15.4 ± 0.6	22.1 ± 1.9 *	62.9 ± 0.9 *	67.7 ± 0.7 *	3.4	$< 10^{-26}$
- Spatial attention	15.4 ± 0.6 *	20.6 ± 2.2 *	65.9 ± 0.3 *	65.5 ± 0.4 *	2.5	$< 10^{-22}$
- Subj layer	8.1 ± 1.9 *	20.2 ± 1.3 *	42.4 ± 0.1 *	47.0 ± 1.3 *	14.4	$< 10^{-28}$
- Clamping	0.5 ± 0.0 *	14.1 ± 1.0 *	1.5 ± 0.3 *	23.6 ± 24.6 *	26.2	$< 10^{-29}$

A. Supplementary information

A.1. Datasets

The data from Schoffelen et al. [1] was provided (in part) by the Donders Institute for Brain, Cognition and Behaviour with a "RU-DI-HD-1.0" licence.The data for Gwilliams et al. [2] is available under CC0 1.0 Universal. The data for Broderick et al. [3] is available under the same licence. Finally, the data from Brennan and Hale [4] is available under the CC BY 4.0 licence. All audio files were provided by the authors of each dataset.

A.2. 'Brain module' evaluation.

To evaluate the elements of the brain module, we performed a series of ablation experiments, and trained the corresponding models on the same data. Overall, several elements impact performance. Performance systematically decreases when removing skip connections, the spatial attention module, the initial or final convolutional layers (Table A.2). These results also show how essential clamping is to train the model. Additional experiments confirm that the present end-to-end architecture is robust to M/EEG artefacts, and thus requires little

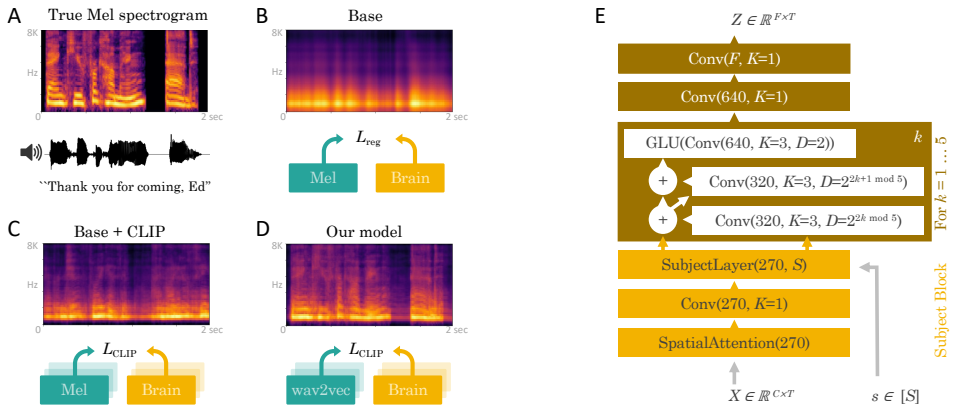

Figure A.1: **Design choices. A.** Illustration of a 3 s speech sound segment (bottom) and its corresponding Mel spectrogram (top). **B.** Mel-spectrogram predicted with a direct regression loss of a brain decoder (orange). **C.** Replacing the regression loss with a contrastive loss improves reconstruction in the same subject, still using the mel-spectrogram as the speech representation. **D.** Now replacing the mel-spectrogram with wav2vec 2.0. The probabilities given by Eq. (2) are used to rebuild a mel-spectrogram. **E. Architecture of the brain module.** Architecture used to process the brain recordings. For each layer, we note first the number of output channels, while the number of time steps is constant throughout the layers. The model is composed of a spatial attention layer, then a 1x1 convolution without activation. A "Subject Layer" is selected based on the subject index s, which consists in a 1x1 convolution learnt only for that subject with no activation. Then, we apply five convolutional blocks made of three convolutions. The first two use residual skip connection and increasing dilation, followed by a BatchNorm layer and a GELU activation. The third convolution is not residual, and uses a GLU activation (which halves the number of channels) and no normalization. Finally, we apply two 1x1 convolutions with a GELU in between.

preprocessing (Supplementary, Sections A.4 and A.3).

A.3. Impact of clamping

Clamping is here motivated by the fact that electro-magnetic recordings can be subject to perturbations orders of magnitudes larger than brain activity. As explained in the Method section, clamping is performed as follows: for each recording independently, we apply a robust scaler such that the data range [-1, 1] maps to the [0.25, 0.75] quantile range. The resulting M/EEG signals are thus expected to have a scale of the order of 1. In Table A.3, we provide the top-10 accuracy for our model similar to Table 2 in the main paper. Extending the clamping range to 100 standard deviations does not appear to help extracting more information. This result suggests that clamping effectively takes care of outlier without removing useful information. On the contrary, the removal of clamping leads to a collapse of decoding performance. This is expected, as extreme outliers will impact, for instance, the BatchNorm mean and standard deviation estimates, and one outlier can thus impact the entire batch. Outliers can also cause extreme gradients and throw off the optimization process. Overall, these analyses emphasize the importance of using clamping for M/EEG analyses. We report hereafter the impact of clamping in terms of top-10 accuracy.

Clamping	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
20	**25.7** ± 2.9	**17.7** ± 0.6	**70.7** ± 0.1	**67.5** ± 0.4
100	**27.1** ± 2.6	**17.6** ± 0.0	**70.6** ± 0.3	**67.2** ± 0.9
None	14.1 ± 1.0	0.5 ± 0.0	23.6 ± 24.6	1.5 ± 0.3

A.4. Comparison with `Autoreject`

To evaluate the potential usefulness of advanced M/EEG preprocessing techniques, we compare our model to one trained on M/EEG data preprocess with the `Autoreject` package [5]. This package aims to detect and correct corrupted channels based on their spatial neighborhood. Note that this package can also reject samples that are too corrupted. However, as this procedure would change the definition of the test set, we do not consider it here. Due to the added complexity of applying Autoreject, which requires in particular the loading of the full dataset in memory, we only evaluated it on the first 16 recordings of Gwilliams et al. [2] Overall, similar performances were obtained with and without Autoreject. This result suggests that our model does not trivially benefit from advanced preprocessing techniques.

Method	*Gwilliams2022* (16 rec.)
clamping at 20	54.7 ± 1.72
autoreject [5]	53.4 ± 1.51

A.5. Impact of EEG/MEG time offset

Our model uses a fixed delay of 150 ms to align speech and EEG/MEG representations. This decision was originally motivated by the non-compressible delay that exists between the cochlea's response and the cortical activations. To further investigate this decision choice, we here study the impact of this delay on the Gwilliams2022 dataset. We observe a small impact of this parameter, although setting it to 0 only reduces the top-10 accuracy by 0.5%.

Delay (ms)	*Gwilliams2022*
0	69.9 ± 0.27
50	70.1 ± 0.13
100	**70.4** ± 0.05
150	**70.4** ± 0.11
200	69.6 ± 0.59
250	68.5 ± 0.14
300	67.6 ± 0.17

A.6. Impact of the number of Mels

Are the models trained to decode Mel features (as opposed to latent representations) impeded by the number of Mel? To study this issue, we evaluate the impact of the number of frequency bands used for the Mel spectrogram for different versions of the model, while keeping the minimum and maximum frequencies fixed. For clarity, we only provide the average top-10 accuracy overall datasets. We observe a small increase of the accuracy when using more Mel bands for all the models.

	# Mel bands			
Model	20	40	80	120
Base model	9.3	9.8	9.8	10.0
+ CLIP	30.5	30.8	31.2	31.8
+ Deep Mel	40.0	40.5	38.2	41.3

A.7. Impact of the batch size and learning rate

We evaluate the impact of batch size and learning rate on the model performance. For simplicity, these evaluations are reported for the Gwilliams et al. [2] dataset only. For the main study, we used a batch size of 256 and a learning rate of 3e−4. We run an evaluation of a grid of 4 learning rates and 4 batch sizes. Overall, we observe better results with larger batch sizes, and we observe instabilities with larger learning rates.

	Batch size			
Learning rate	32	64	128	256
1e−4	65.1 ± 0.62	68.1 ± 0.29	68.2 ± 0.19	68.5 ± 0.28
3e−4	62.5 ± 0.59	65.8 ± 0.45	68.5 ± 0.59	70.4 ± 0.11
6e−4	20.5 ± 34.2	0.7 ± 0.02	67.0 ± 0.25	68.7 ± 0.22
1e−3	0.7 ± 0.0	0.7 ± 0.02	22.3 ± 37.3	36.9 ± 32.0

A.8. Zero-shot decoding of words

The contrastive objective allows our model to select the most likely speech sounds of a test set given brain activity. To what extent does this approach allow the decoding of words absent from the training set? To evaluate this issue, we report the top-10 accuracy at the word level, depending on whether the words are (1) present in or (2) absent from the training set, respectively. Overall, the results show that for EEG, such zero-shot decoding of words falls dramatically. For MEG, however, such zero-shot decoding of words is remarkably close to the accuracy obtained for words present in the training set. These elements strengthen the superiority of MEG over EEG, and show that our approach can be efficiently used to decode words never present during training.

	Brennan (EEG)	Broderick (EEG)	Gwilliams (MEG)	Schoffelen (MEG)
Words in test	9.7	8.0	64.0	61.6
Words in train	35.9	32.9	70.0	73.4

A.9. Decoding of isolated words

To what extent can our approach be used to decode words presented in isolation? To explore this issue, we evaluated our model using a subset from Schoffelen et al. [1], where subjects are presented with random word lists. We use a segment ranging from -300 ms to +500 ms relative to word onset. The results, displayed in Supplementary Figure A.4, show that our model reaches a zero-shot top-1 accuracy of 22.7% when we limit the vocabulary size to 50 at test set, with subjects peaking at 42.9%. While this performance is low, it is interesting to compare it to Moses et al. [6] who report a top-1 accuracy of 39.5% with a model trained to decode the production of individual words, without the use of a language model, *i.e.* independently of context.

A.10. Spatial attention

We provide a visualisation of the spatial attention described in the Method section in Figure A.3.

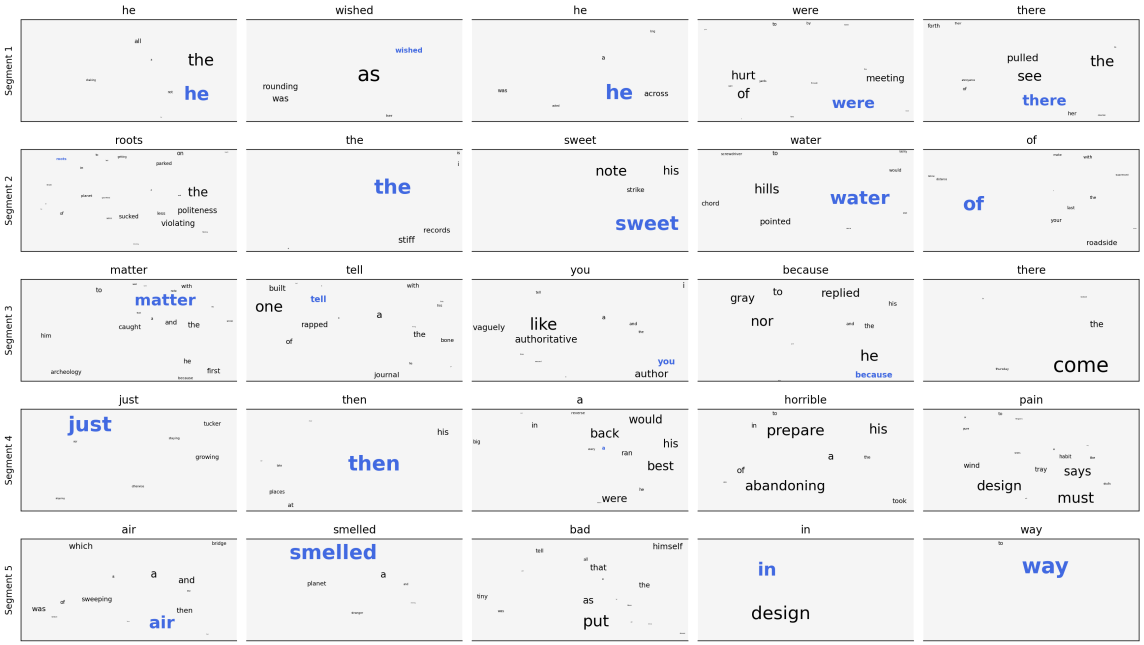

Figure A.2: Similar illustration to Figure 3 in the main paper but for five representative speech segments. The top and bottom segments are the easiest and hardest to decode, respectively. For each segment, we plot the predictions obtained for the subject with the median decoding scores across the cohort.

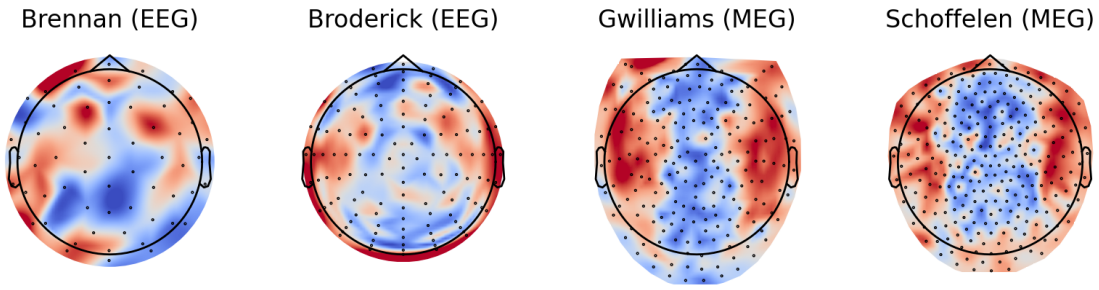

Figure A.3: **Attention weights.** Red color indicate that the M/EEG sensors is, on average, associated with a higher spatial attention weight. At the exception of the Brennan dataset, the topographies highlight channels typically activated during auditory stimulation.

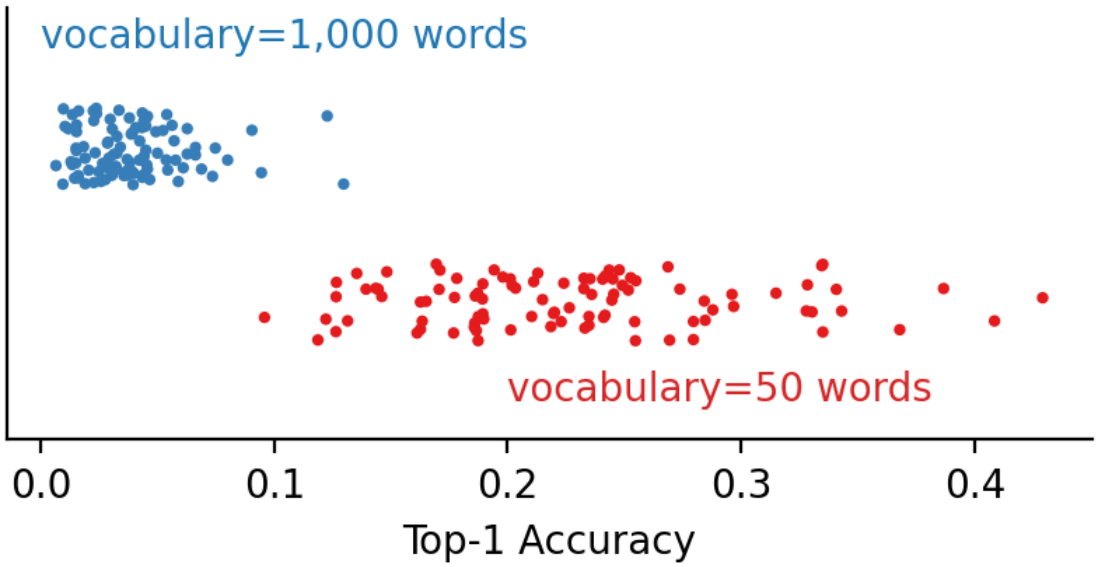

Figure A.4: Decoding performance for single-word, using a vocabulary of 1,000 (red) or 50 (blue) words.

References

[1] Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche H. L. Lam, Julia Uddén, Annika Hultén, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6(1):17, April 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0020-y. URL `https://data.donders.ru.nl/collections/di/dccn/DSC_3011020.09_236?0`.

[2] Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkanen, David Poeppel, and Jean-Remi King. Meg-masc: a high-quality magneto-encephalography dataset for evaluating natural speech processing. *arXiv preprint arXiv:2208.11488*, 2022.

[3] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

[4] Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

[5] Mainak Jas, Denis A Engemann, Yousra Bekhti, Federico Raimondo, and Alexandre Gramfort. Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, 159:417–429, 2017.

[6] David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.