APPENDIX A. Theta band results (4-10 Hz)

In the theta band, the cluster-based permutation test revealed no differences between the word list and sentence TRFs for the word frequency feature (see figure A1 below).

Like in the delta band, the full model was *accuracies ~ condition * (frequency + entropy + surprisal) + (1|subject)*. Removing the interaction between *frequency* and *condition*, or *surprisal* and *condition*, decreased model fit (marginally; frequency: $\chi2(1) = 3.80$, $p = 0.051$; surprisal: $\chi2(1) = 3.95$, $p < 0.05$), but removing the interaction between *entropy* and *condition* did not ($\chi2(1) = 0.47$, $p = 0.49$). We continued with the model *accuracies ~ condition * (frequency + surprisal) + entropy + (1|subject)*. The AIC comparison confirmed that this model was the best descriptor of the data.

In theta, too, there was main effect of condition ($\beta = 2.09*10\text{-}3$, SE = 6.90 * 10−4, $t(1530) = 3.02$, $p < 0.01$), with reconstruction being accuracies higher in the word list condition. In addition, there was a main effect of *frequency* ($\beta = 1.17*10\text{-}3$, SE = 5.64 * 10−4, $t(1530) = 2.07$, $p < 0.05$) indicating that generally, the addition of word frequency improved reconstruction accuracy. The interaction between *frequency* and *condition* approached, but did not reach significance ($\beta = 1.56*10\text{-}3$, SE = 7.97 * 10−4, $t(1530) = 1.95$, $p = 0.051$), indicating a potential trend for this effect to be larger in the sentence condition than in the word list condition (Figure A2).

With respect to the other predictors, there was a positive effect of *entropy* ($\beta = 2.43*10\text{-}3$, SE = 3.99 * 10−4, $t(1530) = 1.95$, $p < 0.01$), and an interaction between *condition* and *surprisal* ($\beta = 1.55*10\text{-}3$, SE = 7.92 * 10−4, $t(1530) = 1.99$, $p < 0.05$), indicating that *surprisal* enhanced reconstruction accuracies more in the sentence condition than in the word list condition.

Again, we performed post-hoc t-tests comparing the two largest models (Entropy/Surprisal and Full) to gain more insight in the effect of word frequency on the reconstruction accuracies. These showed that the word frequency predictor enhanced reconstruction accuracies in the sentence condition ($t(101) = 5.67$; $q < 0.01$), but not in the word list condition ($t(101) = 1.48$; $q = 0.57$). There were no effects of condition for these two models (all $q = 1$).

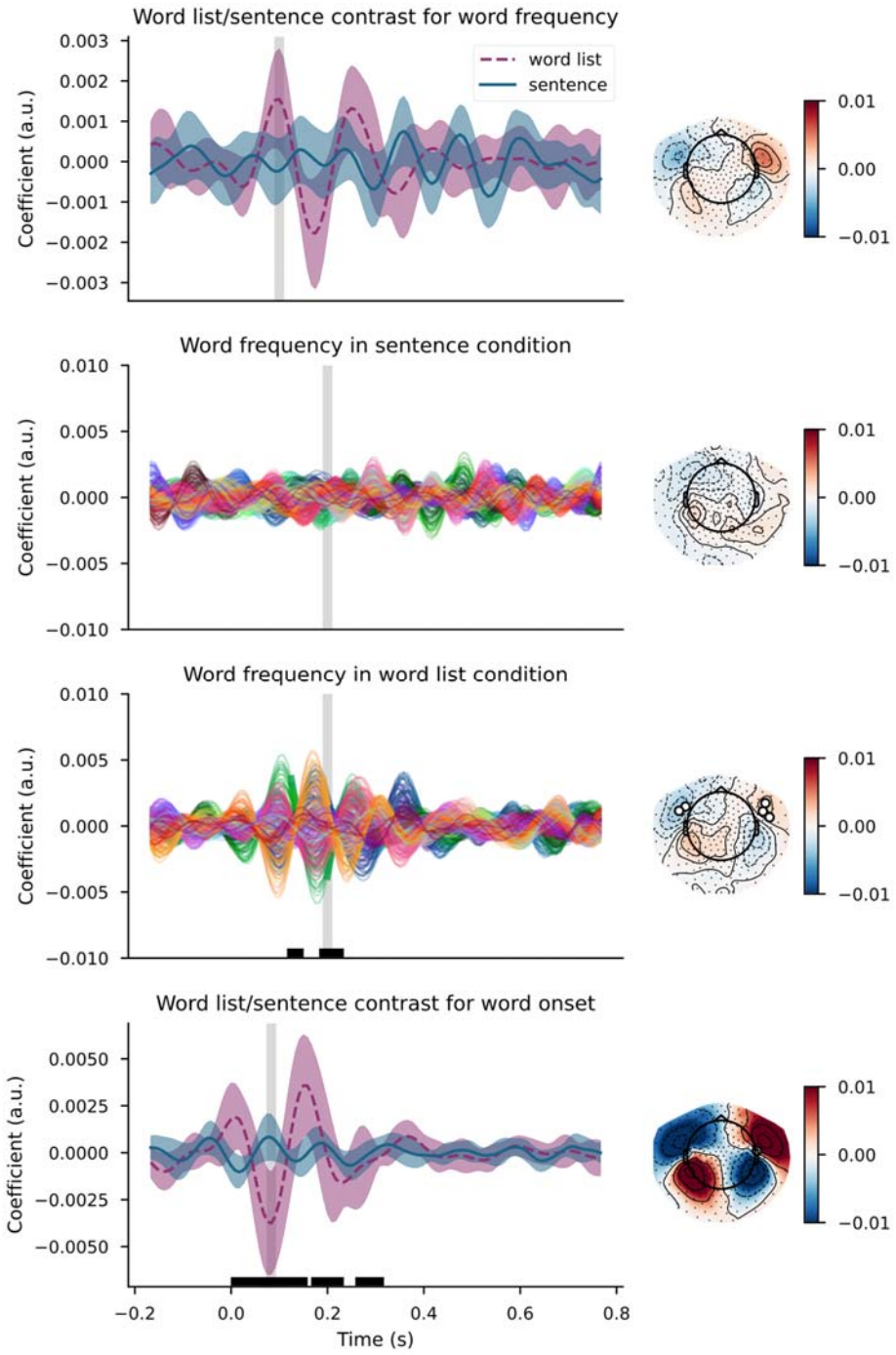


Figure A1. *(A) top: The word frequency TRF in both conditions in the theta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points of those significant clusters. Shaded area indicates standard deviation. (B) top middle: word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. (C) bottom middle: word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. (D) bottom: The word onset TRF in both conditions in the delta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points of those significant clusters. Shaded area indicates standard deviation.*

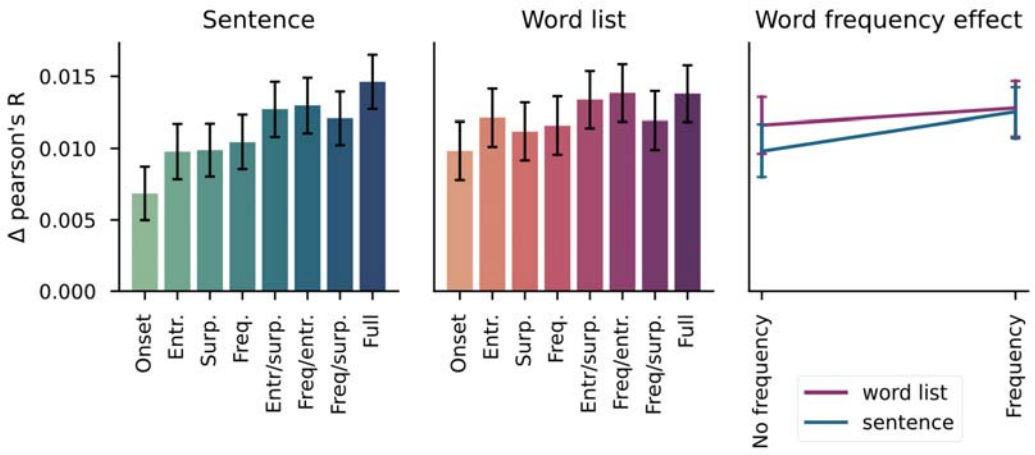


Figure A2. *(A) left: reconstruction accuracy difference with the envelope model for each model in the sentence condition. Error bars represent the standard error of the mean. (B) middle: reconstruction accuracy difference with the envelope model for each model in the word list condition. Error bars represent the standard error of the mean. (C) right: The interaction between* condition *and* frequency *on the reconstruction accuracies. Values on the y-axis are the difference with the envelope (as in A and B). Error bars represent the 95% confidence interval.*

Given that the permutation test in the sensor-based analysis did not reveal any effects in the theta band and we therefore could not select time-bins a priori, we performed a cluster-based permutation test on the full TRF. This revealed two clusters in the right hemisphere between 100 and 250ms. Both of these clusters likely reflect a larger amplitude across right frontal and temporal areas for the TRF in the word list condition than the sentence condition, as can be seen in the plots of the time courses of the clusters in figure A3 below. These effects, although visible in figure A2C, did not reach significance in the sensor-analysis; potentially due to the stringent threshold (recommended value multiplied by three) chosen there.

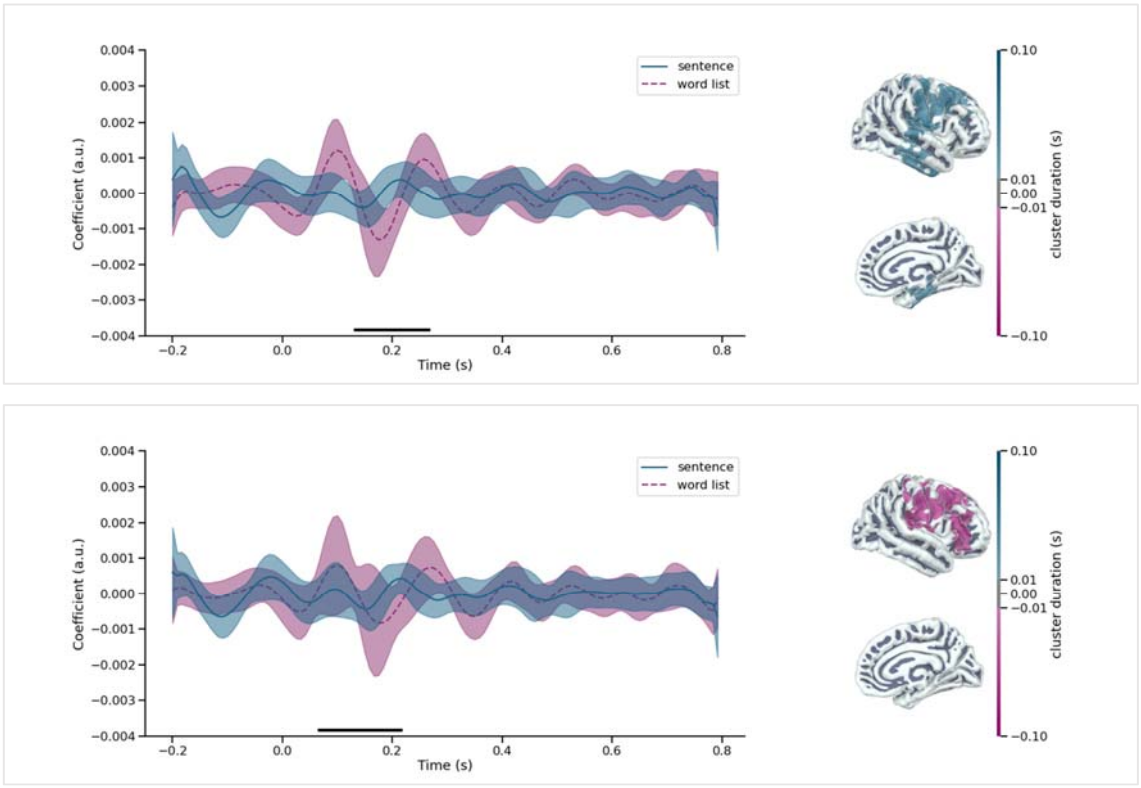


Figure A3. *Clusters from the theta-band TRFs in source space. Blue indicates that coefficients sentence > word list; pink indicates word list > sentence. Darkness of the color shows how long the conditions differ significantly in the given source; see the colorbars. (A) top: right-lateralized cluster where TRF sentence > word list. (B) bottom: right-lateralized cluster where TRF word list > sentence.*

In summary, in the delta band, we observed two bilateral clusters in frontal areas where the coefficients were higher for the sentence condition in the early time-bin (200-400ms). In the late time bin (500-700ms), we observed four significant clusters. In right frontal and inferior parietal cortex, premotor area and cingulate gyrus, coefficients were higher for the word list condition; in left inferior frontal cortex, coefficients were higher for the sentence condition. In the theta band, we observed two right-lateralized clusters across frontal and temporal areas in early time-lags (100-250ms).

APPENDIX B. Extra dataset results

1. Introduction

To examine potential effects of the pause in the word list condition, we analyzed a second dataset of 16 participants listening to word lists and sentences using the same methods. Importantly, the word lists in this condition were naturally spoken, as were the sentences. This means that there are no pauses between words in the word list condition, and coarticulation is present naturally. Replicating our findings in this analysis would be strong support for the interpretation that the pauses do not (fully) cause our findings.

2. Materials and methods

The data were supplied by Ten Oever et al. (submitted).

2.1. Participants

A total of 20 native speakers of Dutch (4 men, 16 women) participated in the experiment with a mean age of 39.5. Four participants were excluded from this analysis due to a variety of reasons (e.g., session was not finished). All participants were right-handed, reported normal hearing, had normal or corrected-to-normal vision, and had no history of neurological, developmental or linguistic deficits. All participants provided informed consent and the study was approved by the ethical Commission for human research Arnhem/Nijmegen (project number CMO2014/288). Participants were remunerated for their participation.

2.2. Materials

The stimuli were identical to the stimuli used in Kaufeld et al. (2020). The experiment consisted of three conditions in total: sentences, Jabberwocky, and word lists. Only the sentences and the word lists are analyzed here. The stimuli consisted of 10 words, which were all disyllabic except for “de” (the) and “en” (and). Sentences had a fixed syntactic structure of two coordinate clauses: [Adj N V N conj Det Adj N V N]. The word lists were scrambled versions of these sentences, and care was taken that there were no plausible internal combinations of words. The stimuli were recorded by a female native speaker of Dutch at a sampling rate of 44.1 kHz (mono). After recording, any pauses were normalized to ~150 ms in all stimuli and the intensity was scaled to 70 dB using the Praat voice analysis software (Boersma & Weenink, 2018).

Participants were asked to perform four different tasks on these stimuli: a passive task, a syllable task, a word task, and a word combination task. In this analysis, we did not distinguish between tasks. For a description of the tasks performed, see Ten Oever et al. (submitted).

2.3. Procedure

At the beginning of each trial, participants were instructed to look at a fixation cross presented at the middle of the screen on a grey background. The audio was presented binaurally through tubes after an interval randomly jittered between 1.5 and 3 seconds. One second after audio offset, the task was presented, which required participants to press a button on a button box. There were eight blocks of approx. 8 minutes. After each block, participants could take a break, during which head position was corrected.

MEG was recorded using a 275-channel axial gradiometer CTF MEG system at a sampling rate of 1200Hz. After the session, head shape was collected using the Polhemus digitizer (using as fiducials the nasion and the entrance of the ear canals as positioned with earmolds). For each participant, MRI was collected with a 3T Siemens Skyra system using the MPRAGE sequence (1mm isotropic). For MRI acquisition, participants wore earmolds with a vitamin E pill to optimize alignment.

2.4. MEG preprocessing

The MEG data were processed with custom-written Python scripts using MNE-Python (Gramfort et al., 2013). As in the main analysis, the raw MEG data was filtered using a windowed-sinc Finite Impulse Response (FIR) filter between 0.5 and 4 Hz for the delta band, and 4 and 10 Hz for the theta band, after which the data was epoched from audio onset to audio offset and resampled to 120 Hz for TRF estimation.

2.5. Stimulus representation

In this analysis, we used the *envelope, word onset,* and *word frequency* representations from the main analysis. For a full description, see section 2.7 in the main text.

2.6. Model fitting

We used the model-fitting approach described in section 2.8 of the main text. The main difference is that in this analysis, we only fit three models: Envelope (with only the *envelope* feature), Onset (*envelope* and *word onset* features), and Frequency (*envelope, word onset,* and *word frequency* features). The data was split pseudo-randomly into a training- and a testing set at a 80/20 ratio, ensuring that the sets contained 50% items from each condition.

The regularization parameter was optimized individually per participant and model, using an eight-fold cross-validation procedure with 20 log-spaced values around 60000 ($\lambda = 60000$) ranging from λe^{-2} to λe^{2}.

2.7. Model evaluation

For model evaluation, we used the procedure described in section 2.9 of the main text.

2.8. Statistical analysis

The TRFs for both conditions were compared using a cluster-based permutation test put in place using the *spatio_temporal_cluster_test* from the MNE-Python library (Gramfort et al., 2013) with the t-test as the test statistic and 1024 permutations. In the sensor space analysis, in addition, the responses were compared to zero with the same cluster-based permutation test with a one-sample t-test as the test statistic. We calculated the threshold on the basis of the two-tailed t-distribution with a significance level of 0.05 with 16 (number of subjects – 1) degrees of freedom. Only clusters with a p-value smaller than 0.01 were considered. Subsequent comparisons were done with a threshold calculated using a Bonferroni adjusted significance level (i.e., divided by two) to correct for multiple comparisons; all else was the same. For comparison to the main analysis, we also compared the word onset response between conditions with the methods described above.

To evaluate the effect of *word frequency* in each condition, we compared the reconstruction accuracies from the Onset and Frequency models in interaction with condition. The reconstruction accuracies were averaged over all sensors (conservative measure). After checking for normality and sphericity through (1) visual inspection of QQ-plots and histograms, (2) statistical testing using the Shapiro-Wilk test, Anderson-Darling test, and D'Agostino's K^2 test for kurtosis and skewness as implemented in SciPy, and (3) the Mauchly test for sphericity as implemented in Pingouin, the averaged reconstruction accuracy values were submitted to a repeated measures ANOVA using Statsmodels.

3. Results

The cluster-based permutation test revealed no significant differences between the word frequency response in the word lists and sentences (Figure B1A). To evaluate if this was due to there being no detectable response or due to no differences between conditions, we performed one-sample cluster-based permutation tests. Here we observed a response in the sentence condition over a large array of left-posterior sensors that was significant from word onset to about 400 milliseconds. The peak appears around 200 milliseconds. Although figure B1C suggests a potential response around 400 milli-

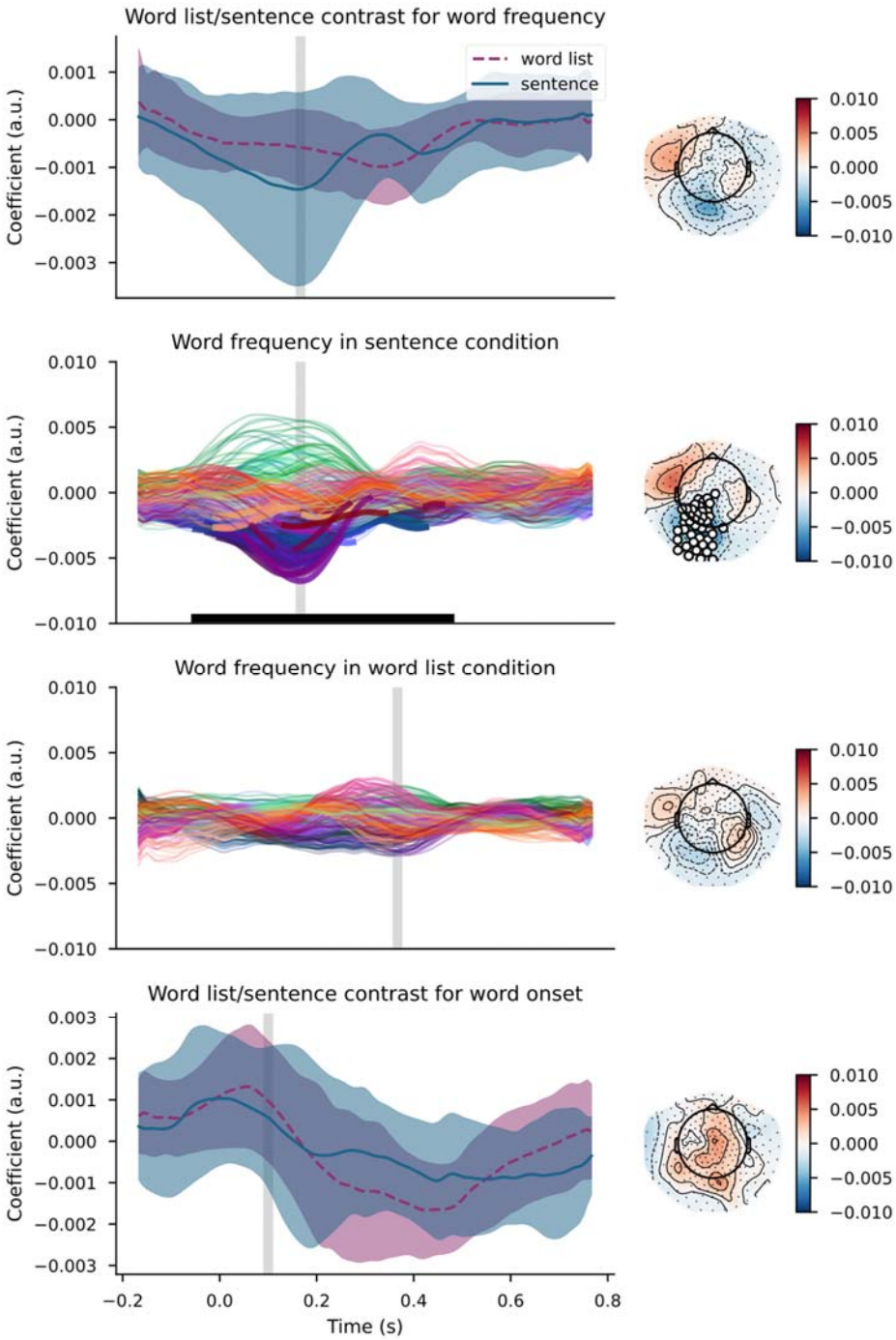


Figure B1. *(A) top: The word frequency TRF in both conditions. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points of those significant clusters (none). Shaded area indicates standard deviation. (B) top middle: word frequency TRF in the sentence condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters. (C) bottom middle: word frequency TRF in the list condition. Sensors in bold were significant in the one-sample cluster-based permutation test. Sensors in bold were significant in the one-sample cluster-based permutation test. Black bars indicate time points of the significant clusters (none). (D) bottom: The word onset TRF in both conditions in the delta band. Shown here is the mean of the sensors that were included in clusters that were different between the two conditions. Black bars indicate time points of those significant clusters (none). Shaded area indicates standard deviation.*

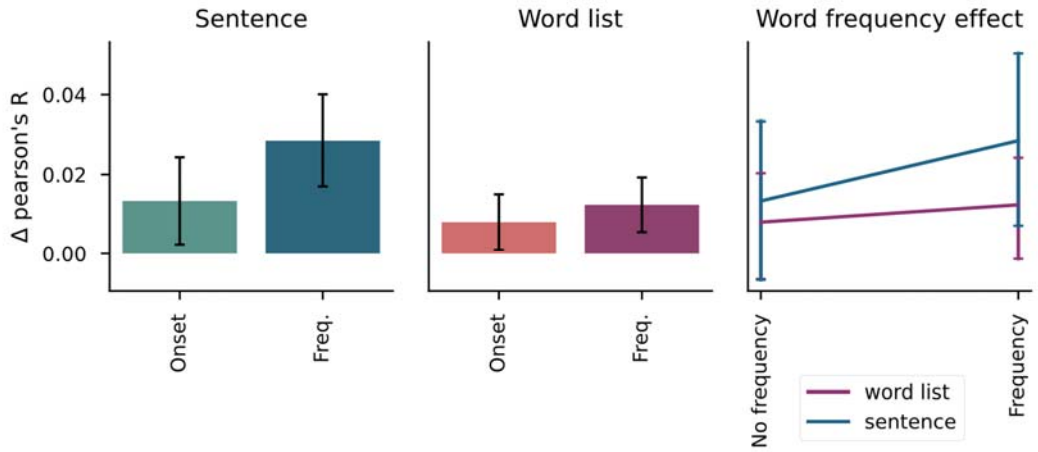


Figure B2. *(A) left: reconstruction accuracy difference with the envelope model for each model in the sentence condition. Error bars represent the standard error of the mean. (B) middle: reconstruction accuracy difference with the envelope model for each model in the word list condition. Error bars represent the standard error of the mean. (C) right: The interaction between* condition *and* frequency *on the reconstruction accuracies. Values on the y-axis are the difference with the envelope (as in A and B). Error bars represent the 95% confidence interval.*

seconds, there were no significant clusters in the word list condition. Like in the main analysis, there were no significant differences between conditions in the responses to word onset (Figure B1C).

The ANOVA on the reconstruction accuracies revealed a main effect of model ($F(1,15)=38.01$; $p < 0.01$), indicating that the word frequency predictor enhanced reconstruction accuracy (see Figure B2A and B), and an interaction between condition and model ($F(1,15)=6.79$; $p < 0.05$), suggesting that this effect is larger for the sentence condition than for the word list condition (Figure B2C). There was no main effect of condition ($p = 0.16$). See Table B1 for the full model results.

Table B1. *ANOVA results of the Condition by Model effect.*

	F value	Num DF	Den DF	p-value
Condition	2.14	1.0	15.0	0.16
Model	38.01	1.0	15.0	< 0.01
Condition:Model	6.79	1.0	15.0	0.02

4. Discussion

The analysis reported here suggests that under natural presentation conditions, the response to word frequency is more present in the sentence condition than in the word list condition. This is most apparent in the interaction on the reconstruction accuracy of the neural signal: adding the word frequency feature to the model improves reconstruction accuracy more in the sentence condition than in the word list condition. In addition, the response as captured by the TRF time course suggests

that the word frequency feature evokes a stronger response in the sentence condition than in the word list condition.

The finding on the accuracy of the reconstruction of the neural signal is in line with the findings in our main analysis. There, too, we observed a larger increase in the sentence condition than in the word list condition when the word frequency predictor is added to the model. Both of these analyses suggest that the word frequency feature explains less variance in the neural signal in the word list condition than in the sentence condition. This suggests that sentence context enhances the presence of lexical information in the neural signal, akin to how structure and meaning can enhance tracking of speech (Kaufeld et al., 2020).

The fact that we failed to observe a significant difference between the conditions in the TRF time courses likely has multiple causes. First, there are methodological differences between the studies, most notably the number of participants (N=16 versus N=102). Responses to linguistic features tend to have much lower power than, for example, responses to acoustic predictors like the speech envelope, but also acoustic edges and the spectrogram. Having a larger sample size increases overall statistical power, increasing the likelihood that effects are found in statistical tests. Furthermore, the content words chosen to create the stimuli in the present study were all of relatively high frequency. Low variability in the predictor makes it more difficult to capture effects – especially when they are already small and of relatively low power, as is the case for linguistic parameters. Needless to say, the fact that we do not observe differences in the tests used here, does not mean that they do not exist.

A second set of potential causes is psychological. As is seen in the one-sample tests, the response to word list is decreased in this study relative to the original analysis. This could be due to the lack of pauses between the words, which enhance the perceptual salience of the words and could lead to responses that are more detectable. A difference between conditions can then not be captured by statistical tests because (1) the response in the word list condition is not different from zero (i.e., the difference between the word list and the sentence response is smaller than the difference between the sentence response and zero), while not being strong enough to show up in one-sample tests; and (2) the response in the sentence condition is not strong because of the small sample size.

Taken together, the present analysis can tell us the following about the results in the main study. Firstly, given the absence of a response to word frequency in the word list condition in the present analysis, it is likely that the pauses between the words in the main analysis enhanced perceptual salience of the words in the word list condition. This allowed us to estimate lexical responses in the word list condition, and compare them between conditions in time and space: it has allowed us to show that much of the response to words in word lists and sentences overlaps in space

(superior temporal gyrus, inferior temporal gyrus, parahippocampal gyrus, and motor cortex). Secondly, the present analysis suggests that the word frequency feature explains less variance in the neural signal in the word list condition than in the sentence condition; the effects found on the reconstruction of the neural signal in the main analysis appear to be independent of the existence of a pause between the presented words.

Unfortunately, the fact that we cannot find a clear response in the word list condition means that we cannot draw conclusions about the timing of the responses in sentences and word lists. At this point, we know that the response to word frequency in the main analysis is delayed in the word list condition. The response to word onset was not delayed. Based upon the results from the paper, it remains possible that the brain employs a conservative processing strategy when it comes to language comprehension, in which word-internal features are processed no faster than absolutely required. Introducing a pause could then delay responses to lexical features. However, the possibility of a delay being introduced by the absence of sentence context is not excluded on the basis of the present analysis, either. Further research is required to examine these possibilities.

Appendix C

Hugo Weissbart & Sophie Slaats

Temporal Response Function: properties and caveats

We will overview some properties of the forward linear model used to estimate "temporal response functions" (TRF), i.e. a convolutive linear model. We aim at controlling for and estimating the bias between the conditions used in the article: one continuous stream of speech sound versus an isochronous presentation of words with some spacing (silence) between them. In summary we will focus on two aspects of the TRF modelling that relate with the experimental design: - how the inter word interval does not directly impact model evaluation, but rather the length of the evaluated signal together with its broad band signal-to-noise ratio do; - how bandpass filtering the data has no impact on further statistical anbalysis, that is, any model comparison within a frequency band of interest allows for valid inferences.

Model description

The system under scrutiny is simply a linear-time invariant system which sees its input convolved with a (or several) kernel to generate an output signal. In neuroimaging, we would refer to this linear model as an *encoding model*, or *forward model*, if y represents the brain response (and x the stimulus representation) and as a *decoding model*, or *backward model*, conversely.

However, brain signals recorded from any device will carry measurment noise on top of what the model is uncapable of representing (non-linear relationship to the stimulus features). Thus the full description of this linear model in the context of modelling brain activity encompasses an additive noise component, capturing all remaining part of the signal that is not described by the linear-time invariant system itself.

The equation driving the noisy output of this system reads:

$$y(t) = \sum_{\tau} \beta(\tau) x(t-\tau) d\tau + \eta(t)$$

For real data, sampled at a frequency F_s, we use the discrete time $t = ndt$, where $dt = \frac{1}{F_s}$. We can also add several predictors x_i and sum over each contribution, where each predictor gets its proper kernel β_i to be convolved with. One can thus rewrite the above as:

$$y[n] = \sum_{i} \sum_{k} \beta_i[k] x_i[n-k] + \eta[n]$$

Finally, a further reduction is possible by vectorizing the equation, and concatenating along the dimension of summation on kernels (dummy index i above):

$$\mathbf{y} = \mathbf{X_{lagged}} \beta + \eta$$

This equation stands for a given sensor (channel), but we could equally concatenate along a new dimension each sensor equation, then $\mathbf{y}$ becomes a matrix $\mathbf{Y}$ and β too. $\mathbf{X_{lagged}}$ remains the same. This ultimate concatenation is referred to as *multiple* regression in statistics literature (the solution of sensor specific equations are independent from one to another).

Terminology

- N: number of samples, reflecting duration of signals x and y
- m: dimension of y, i.e. number of sensors in the measured output signal
- p: dimension of x, i.e. number of predictors (including, when used, the intercept)
- k: number of discretised lags
- **Temporal Response Function** and **Impulse Response** or **Kernel** here represents the exact same elements, i.e. the filter kernel that convolves with the stimulus to output the response signal. Its vectorised version being $\beta \in \mathbb{R}^{k \cdot p}$ (or $\in \mathbb{R}^{kp \times m}$ for the multiple regression form, with all sensors concatenated)
- By **response**, we mean the **output** signal y_t, (or $\mathbf{y} \in \mathbb{R}^{N \times 1}$ for the sampled, time discrete, signal), also referred to as the **dependent** variables in a statistical framework (for the multiple reggression form, $\mathbf{y} \in \mathbb{R}^{N \times m}$)
- By **input**, we refer to signal x_t (or $\mathbf{x} \in \mathbb{R}^{N \times p}$ for the sampled, time discrete, signal), which implements the representation of the stimulus, also known as **independent variable** or **regressors** in the context of linear regression. With the encoding model in mind, those are also called **predictors**.
- $\mathbf{X_{lagged}}$ (or simply $\mathbf{X}$ form now on), from the vectorized formulation, is the matrix of time-lagged predictor time series. $\mathbf{X} \in \mathbb{R}^{N \times kp}$

Model estimation

In order to evaluate the quality of the model, one has to first estimate the kernel coefficients from the actual data $\{\mathbf{X}, \mathbf{y}\}$ and thus obtain the estimate $\hat{\beta}$. This is an ill-posed problem, as often in neuroimaging, there will more samples than lags and predictors: $N > kp$. The system of equation is thus *over-determined* (more equations than variables) and no solution uniquely exists. In other words, $\mathbf{X}$ is not squared and not invertible. To circumvent this, it is possible to turn the estimation problem into an optimization problem, and find, where possible, the global minimum of a specific cost function J. We opt here for the least square solution, that is, finding coefficients $\hat{\beta}$ that minimize the square error between predicted and actual values of y: $J(\beta) = \Sigma(\hat{y}[n] - y[n])^2 = \|\mathbf{X}\beta - \mathbf{y}\|_2^2$.

There is a closed-form solution to this minimization problem:

$$\hat{\beta} = (\mathbf{X^T X})^{-1} \mathbf{X^T y}$$

Often, when any predictor x presents some amount of autocorrelation, $\mathbf{X}$ will see several of its column dangerously correlated, meaning that $\mathbf{X^T X}$ might have eigenvalues close to zero. Numerically, the matrix will still be invertible, but those small eigenvalues will tend to blow up when inverting the moment matrix. A solution is to add positive elements to its diagonal ($\mathbf{X^T X} + \lambda \mathbf{I}$), this known as Tikhonov regularisation, or ridge regression.

Scoring

Then one would generate an estimate $\hat{\mathbf{y}}$ from $\hat{\beta}$ and $\mathbf{X}$. A different set of data $\{X, y\}$ must be used for model estimation (training) and evaluation (testing).

Finally $\mathbf{y}$ and $\hat{\mathbf{y}}$ are compared via a scoring metric such as Pearson's correlation coefficient or the coefficient of determination r^2.

When computed from the same sample on which the linear model coefficients have been estimated, the correlation coefficient r reads as:

$$r^2 = \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = \frac{var(\hat{y})}{var(y)}$$

Where the second equality can be shown when noticing that residuals are uncorrelated with reconstructed signal $\hat{y}$ (their vectors are othognal). Importantly, when there is no information contained in $\mathbf{X}$, the variance of $\hat{y}$ is zero while the variance of y will be dominated by noise η, hence $r \to 0$. This equation is not exact if computed from unseen data, however the contributions will follow a similar trend.

Notably, as we are representing stimulus with impulse trains, the reconstructed signal is virtually zero (or equal to the estimated mean, $\bar{\hat{y}} = \bar{y}$) where there is no information in the input signal (when $x[n] = 0$). In other words, no matter the inter word intervals (regardless whether the words are pronounced with natural prosody or spearated by silences), the scoring will be the same **as long as the signal length are the same**. In fact, what truly impacts the score is the broad band signal to noise ratio of the measured y signal, as shown above. This is observed in the simulation resuts seen in fig. 1 and fig. 2.

Spectral properties

One key property of the convolution is that its *dual* operator is the multiplication. Meaning that a convolution in time corresponds to a multiplication in Fourier domain, and vice-versa.

In other words, one can rewrite the time domain (continuous) equation of our model in spectral space as:

$$\hat{Y}(f) = X(f)B(f)$$

Where $X(f) = \mathcal{F}\{x\}$ is the Fourier spectrum of x and so forth for each variables. First of all we can observe directly from the above equation how each frequency contribute **independently** to the reconstructed signal, any effect observed in response works in a *per frequency* basis. The causal effect on the spectral content in the response is *bounded* to the spectral region observed in the stimulus. In other words:

> Such systems cannot generate response with frequency component that are not in the input.

Another way to obtain the kernel coefficients is therefore by dividing each spectra. This will eventually blow up as x is most likely a band limited signal (which shows from a spectral aspect why we might need regularisation):

$$\hat{B}(f) \sim \frac{X^*(f)Y(f)}{X^*(f)X(f)} = \frac{X^*(f)Y(f)}{|X(f)|^2}$$

We observe that the estimated kernel coefficients will only contains significant power for frequencies where y contains power (assuming a well behaved, i.e. broad-band, x).

Another important point is that filtering along the time axis is a convolution, therefore has the effect of multplying the spetrum Y by the filter spectrum in Fourier space. In other words, any filtering applied to y will be reflected directly on the estimated weights.

Let's take $w(\tau)$ as a FIR filter (with $\mathcal{F}\{w\} = W(f)$), filtering y then means our new filtered signal $\tilde{y}$ will have the spectrum $\tilde{Y}(f) = Y(f) \times W(f)$. Thus the estimated kernel coefficients (in Fourier space) become:

$$\hat{\tilde{B}}(f) \sim \frac{X^*(f)Y(f)W(f)}{|X(f)|^2} = \hat{B}(f) \times W(f)$$

> Filtering y is equivalent to filtering β

So we notice that we would recover the same coefficients by **not** filtering y but only filtering our estimated broadband $\hat{\beta}$.

Another implication is that $\hat{Y}(f) = X(f)\hat{B}(f) \sim Y(f)$, in the sense that they will share power at same frequencies.

> Estimated $\hat{\beta}$, and *a fortiori* $\hat{y}$, have a spectrum with energy concentrated respecitvely where y has spectral energy or where both x and y share energy.

In conclusion, when considering a model estimated on a narrow-band signal (e.g. δ-band, [0-4] Hz), we are exposing the relevance of the band-specific signal-to-noise ratio. The energy contained in a band specific region compared to the broad band signal will proportionally inflate the score estimated in the given frequency span. In other words the spectral power in a narrow specific band that is shared with the input signal will positively impact the score estimate for the filtered signal.

When considering multiple kernels (and multiple input predictors x_i), we might wonder what happens when different regressor contribute to different frequency bands. From the above reasoning, we can conclude that each contribution will scale independently from each other, as a linear-time invariant system acts independently on each frequencies. Therefore, we will observe higher score in a given frequency band *if and only if* y and x_i share power in the frequency span being analysed. Simulations shown in fig. 5, fig. 7 and fig. 6 highlight results relevant to those arguments.

Simulations

Code of all simulations available at this link (private for now).

ISI and signal length

We first tested the influence of inter-stimulus interval, and in our case inter-word onset timings, over the estimated reconstruction accuracy for a TRF model. An interactive example can be epxlored at this link (private for now).

The simulated response was equivalent to the forward model, namely a noisy output of a convolution between a predefined kernel (the ground truth for the TRF estimate) and an impulse train (for the input signal). We generated those data with variable amount of noise (i.e., explicitly manipulating the broadband signal-to-noise ratio) and with varying inter-stimulus interval (ISI) while keeping the signal length the same and the number of impulses, or events, constant (in which case shorter inter-stimulus interval results in the end portion of the output signal containing only noise, as seen in fig. 3). We then scored the forward model by computing both the R^2 score and the Pearson's correlation coefficient between the reconstruction $\hat{y}$ and the true signal using a test portion of the data, not used to estimate the coefficients β. Importantly, we then computed the scores:

- either from the fixed signal length data described above; since we also used a fixed number of impulses, or events, this resulted in some portion of the stimulated output signal to contain only noise;
- or from a shorten signal, where we truncated all signals so they extend only up to the last stimulus event. This resulted in shorter signals for shorter ISI.

Strikingly, ISI has no direct influence on reconstruction score, although the length of the segment on which we estimate the score does. In this case, difference in ISI, which eventually leads to difference in data lengths, shows how the bias in score observed between conditions is solely due to the difference in duration. The bias, however, is constant and should be controlled for when directly comparing models within conditions. Moreover, we actually observe the opposite effect in our MEG analysis: our score differences are over and beyond any bias generated from the stimulus differences.

Filtering effect and band specific SNR-ratios

These simulations demonstrate the spectral properties mentionned earlier, notably that we can equivalently filter the broadband-estimated kernels or the response signal to recover a narrow-band TRF; and also that the score obtained from narrow band TRF directly scales with the relative energy contained in that frequency band.

First we simulated an electrophysiological response, M/EEG-like, with stimulus-related signal (modelled as a forward convolutive model, such that the ground truth is indeed the kernel the TRF seek to estimate) in two neighbouring frequency bands. We add braodband (pink) noise, and stimulus-unspecific power in the alpha band. This can act as a control, as we design stimulus-related activity to be outside of the alpha band, spectra of relevant signals are shown in fig. 4. The coloured noise simulate the background noise in 1/f. This will imply a generally higher power in lower frequency bands, although this is not related to stimulus information. We will see that the absolute power does not matter, as the score scales with the relative energy contained in narrowband region where the stimulus aslo has power.

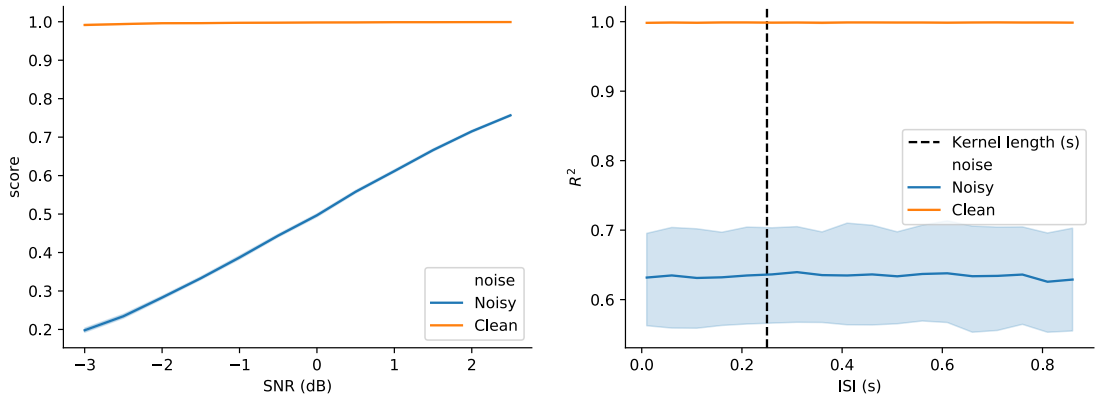


Figure 1: (Non) Influence of ISI on score. Signal length are matched (hence the shorter ISI signal will have a portion containing solely noise (and no stimulus response). The left panel shows the (proportional) influence of broadband SNR on score. The right panel, shows that for every ISI values, we actually measure the same score.

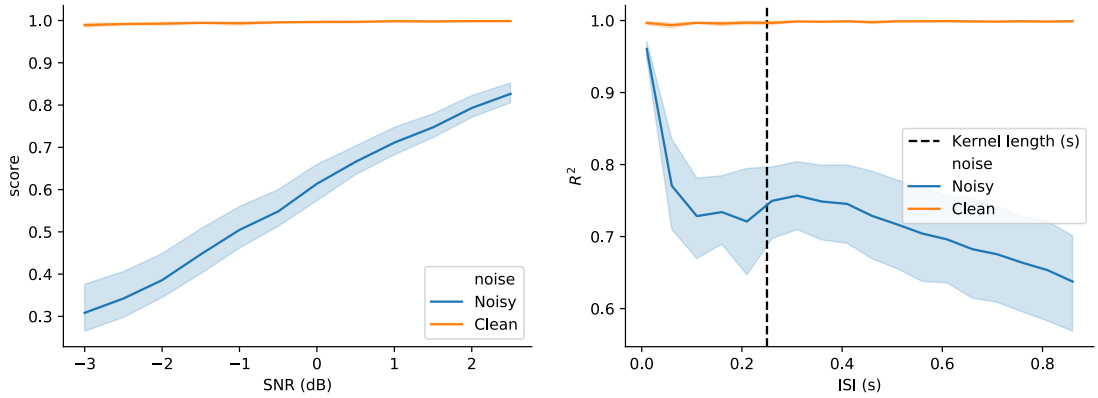
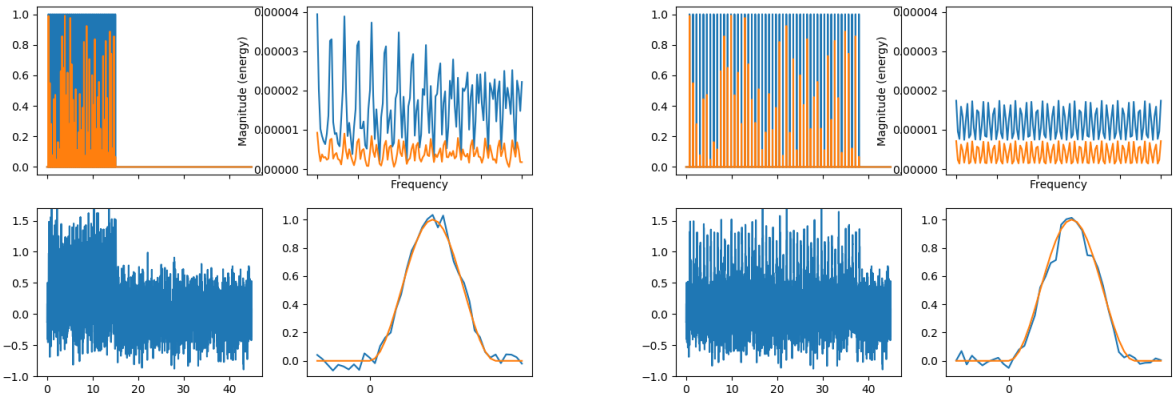


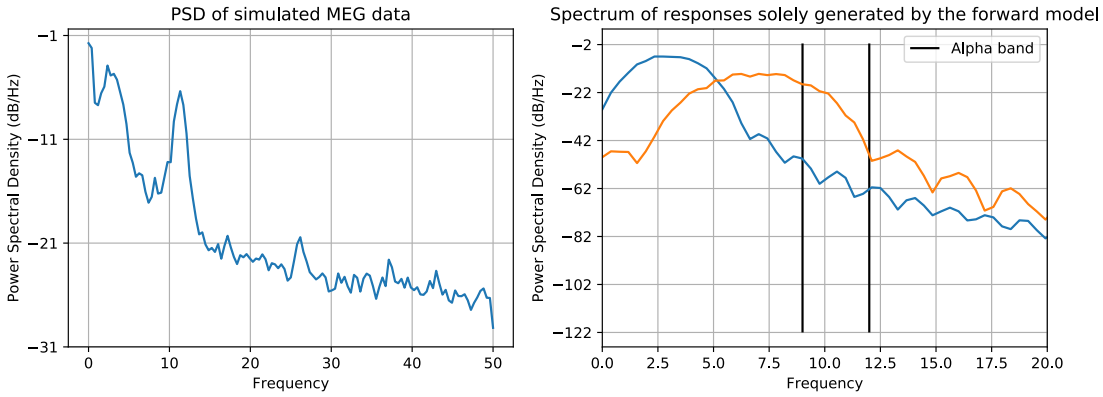
Figure 2: Influence of data length on score: longer ISI giving longer data segment for same number of events. The score is deflated as more noise is being evaluated in the scoring of the longer signal.



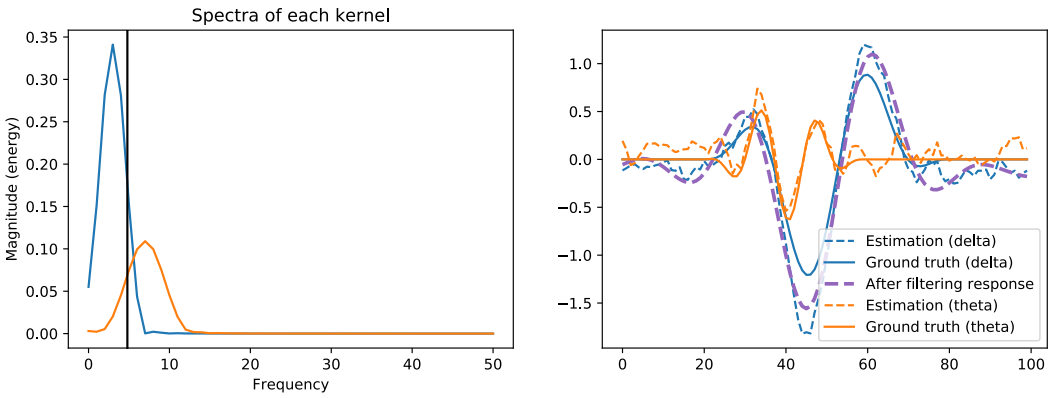
(a) Data used (example with short ISI)

(b) Data used (example with long ISI)

Figure 3: Example of data used in the simulations. On the left is a short ISI signal producing the output signal and on the right a long ISI signal. There is no impact on the estimated TRF.



(a) Spectra of simulated response



(b) Spectra of kernel used and estimated kernel coefficients.

Figure 4: Spectra of signal of interests. **Top**: *left*: noisy response spectrum; *right:* spectrum of response solely generated by either kernel (delta and theta). We observe also an overlap of the alpha band with the theta-band response. **Bottom**: Spectra of delta and theta kernel, we can observe a slight overlap; and estimated TRF for singled out response (response generated by either one of the kernels). The purple dashed line in the bottom right panel shows a TRF estimated on a nfiltered (delta band) signal.;

In fig. 5 we can observe the score of a model where the output signal is the combined convolution of two kernels in two separate frequency bands, with different stream of impulse train (one input signal for each kernel). We then computed score of a signal reconstructed using either the TRF from the filtered signal which matches the ground truth kernel's own frequency span or conversely using the TRF of the other frequency band (unmatched filtering condition). To summarize:

- "Matching TRF": Filtering y in the frequency band of kernel A (e.g., delta band) and computing scores with signal reconstructed from the TRF A only (both TRFs A and B were jointly estimated from the unfiltered y signal)
- "Unmatched TRF" Filtering y in frequency band of kernel A and computing scores with signal reconstructed from TRF B
- Additionally, computing the score in frequency band C, filtering y in C (alpha band) and using both TRF (delta and theta) to estimate $\hat{y}$

Results show that, as expected, besides the original signal having more power in the delta band, we only get a better score relative to the unmatched models in the frequency band where the original kernel actually has power.

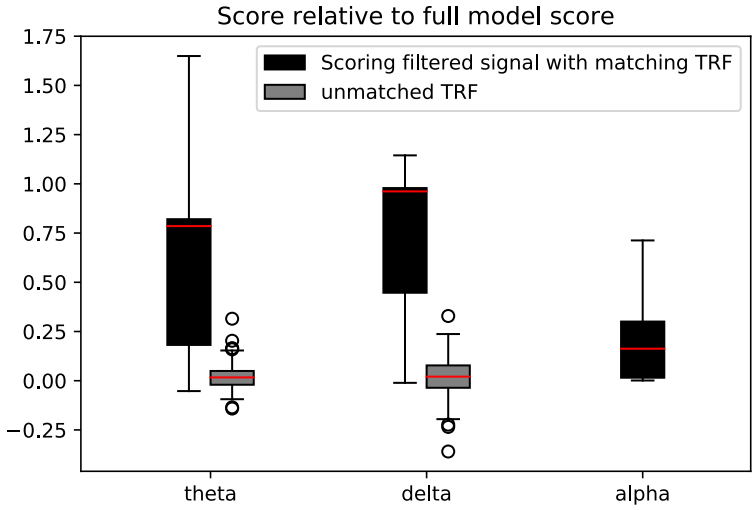


Figure 5: Score (ratio with full model) of reconstructed data compared to filtered data. Using either the "delta" TRF, or other.

We ran the simulation with varying amount of band-specific noise in both delta and theta. That is, we had different signal-to-noise ratio for each band, so we could differentially amplify the contribution of either kernel into the response signal. We observe that there is a trade-off between SNR of one frequency band and score in the other frequency band (see fig. 6). This is expected as the kernel spectra overlap in this case. However, importantly, for similar order of magnitude, the relationship is trivial, in the sense that no matter the band-specific SNR, we can reconstruct the signal faithfully in either frequency band.

The last simulations aim again at showing how the band-specific SNR affects score for band-pass filtered signal. Howver for those simulations, we simply modelled a noisy response from convolved kernel (no colour noise, or stimulus-unrelated oscillatory activity). The goal is to highlight that the score is directly proportional to the *signal energy*, as measured from the band-specific power. We normalised these values by the total energy (across all frequencies), and similarly for scores we computed the score in a given freuqnecy band, using the frequency specific TRF normalised by the score of the unfiltered signal using both kernel. Results are shown in fig. 7. A one to one mapping between the SNR in a given band and the score reconstructed in that same band are observed. The constant bias between each frequency band reflects a bias observed for different ISI, as the relative energy contained in a specific frequency band will vary according to how the signal is spread out. For shorter ISI, for the same level of spectral energy theta reads a higher score, while, conversely, at higher we will observe a higher score in delta for the same relative energy.

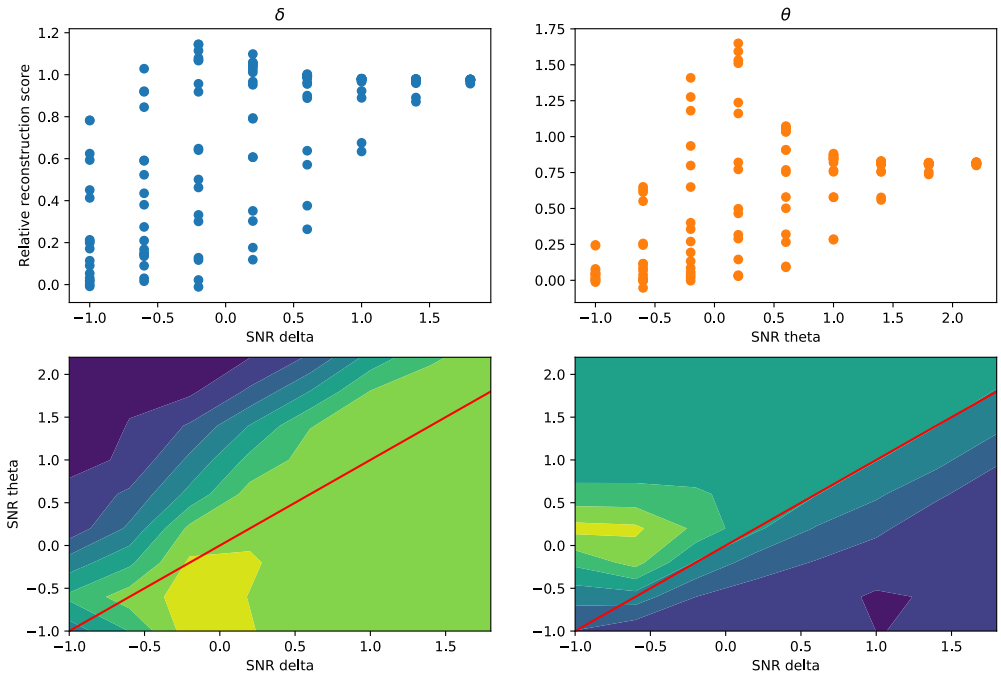


Figure 6: Using different SNR (wrt band-specific background noise) influence how well a band-specific kernel reconstruct the signal. Since our bands of interest are slightly overlapping, we also observe that higher SNR in the "other" band deflates one band's score.

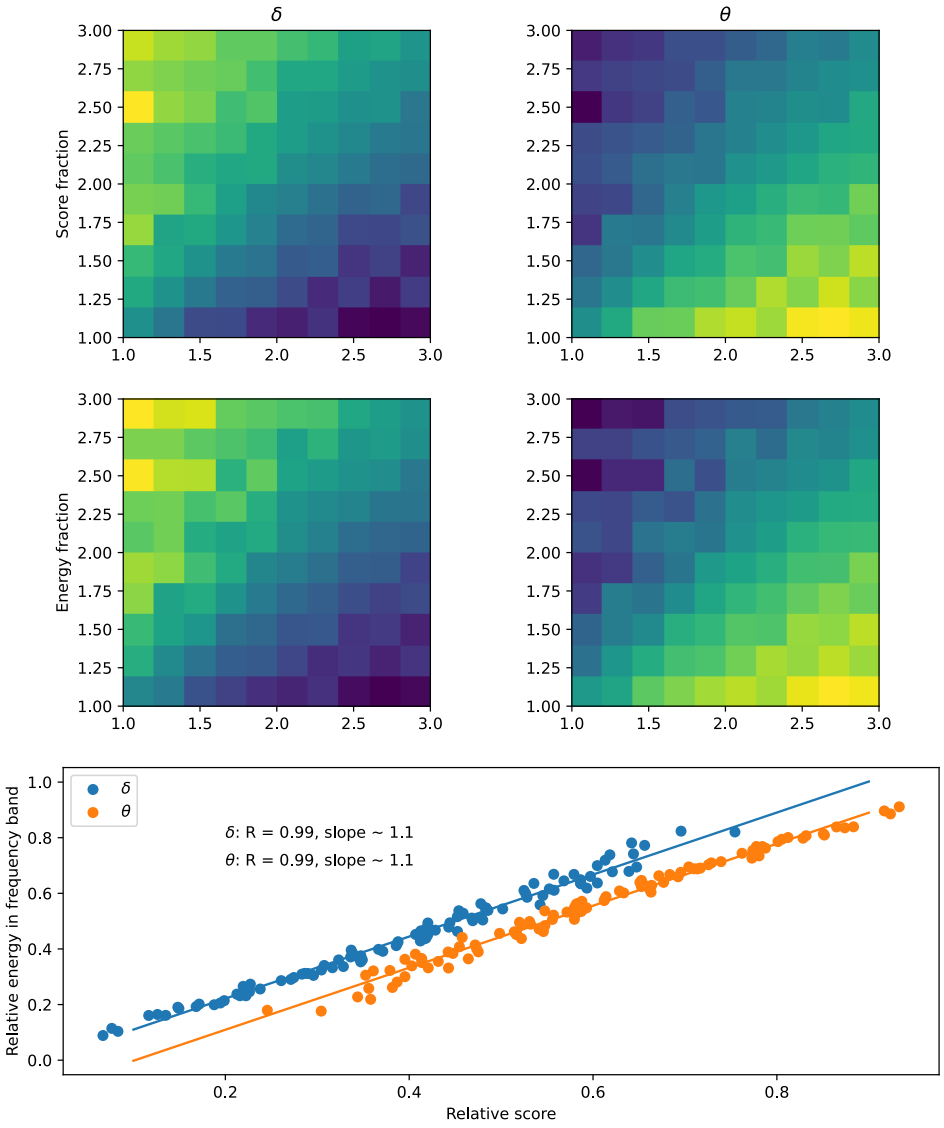


Figure 7: Relative score in a specific frequency band matches its relative energy (spectral SNR) in the given band