

Modeling and Interpreting Correlations, Null Distributions and Significance Levels in Neural Tracking of Natural Stimuli

Simon Geirnaert^{a,b,*}, Alexander Bertrand^a, Tom Francart^b, Jonas Vanthornhout^b

^a*KU Leuven, Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing and Data Analytics,*

^b*KU Leuven, Department of Neurosciences, ExpORL,*

Abstract

Neural tracking - the time-locking of neural responses to continuous stimuli such as speech, music, and video - is widely used to study how the brain processes natural input. Tracking strength is typically quantified as the correlation between the recorded neural response and the stimulus, decoded and/or encoded through data-driven models, and this correlation is routinely used to compare stimulus features, models, or settings. However, its magnitude depends not only on how strongly the brain tracks the stimulus, but also on the statistical properties of the signals being correlated. For example, a smallband speech envelope carrying almost no information about speech content yields among the highest correlations, simply because it is easier to reconstruct. Meaningful interpretation therefore requires comparing each correlation to its null distribution: the correlations expected without any stimulus-response relationship. We show that the randomization procedures commonly used to construct this null distribution are not interchangeable: each implicitly encodes a different null hypothesis, and we motivate stimulus-response misalignment as the most practical and appropriate choice. Because reliable null distributions require many permutations, we introduce a semi-parametric model using the normal distribution after the Fisher transform that yields accurate significance levels from only 3-5 min of data and predicts them across analysis window lengths. Building on this, we propose the null-normalized tracking score (NNTS), an interpretable measure placing features and models on a common scale, which relates directly to the widely used match-mismatch accuracy. Applied to EEG from 121 participants listening to continuous speech, represented across eight acoustic and linguistic features, the framework reverses conclusions drawn from raw correlations, providing an efficient and principled methodology for interpreting neural tracking correlations and comparing features and models.

Keywords: neural tracking, electroencephalography, correlation, permutation testing, statistical significance, Fisher transformation

1. Introduction

When attending to a naturalistic, continuous stimulus, the brain's neural responses time-lock to that stimulus, a phenomenon known as *neural tracking* (Aiken and Picton, 2008; Lalor and Foxe, 2010; Ding and Simon, 2012; Pasley et al., 2012). It is typically observed using neurorecording modalities with high temporal resolution, such as electroencephalography (EEG), electrocorticography (ECoG), or magnetoencephalography (MEG), and has been demonstrated across a range of naturalistic stimuli, including natural speech, music, and natural video. Unlike traditional neuroscientific and brain-computer interface applications based on synthetic, repeated stimuli, neural tracking of naturalistic, continuous stimuli enables a broader range of

*Corresponding author

All authors are also affiliated with Leuven.AI - KU Leuven institute for AI.

Email address: `simon.geirnaert@kuleuven.be` (Simon Geirnaert)

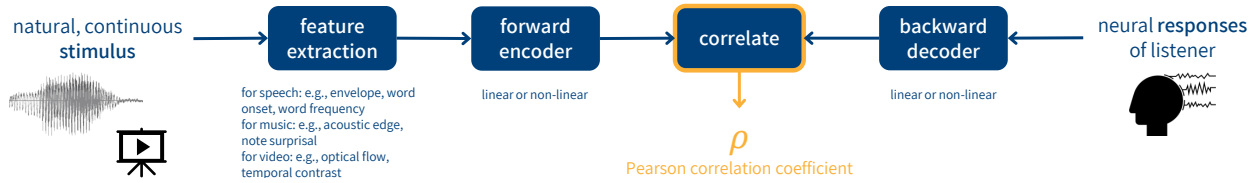


Figure 1: Data-driven models relate natural stimuli such as speech, music, and video to neural responses. Their outputs are typically evaluated using the Pearson correlation coefficient, whose interpretation requires comparison with the corresponding null distribution and significance level.

experiments, analyses, and applications in real-world scenarios. Examples include studying speech processing (Aiken and Picton, 2008; Ding and Simon, 2012; Pasley et al., 2012), music perception (Sturm et al., 2015; Di Liberto et al., 2020, 2021; Zuk et al., 2021), visual processing (Ki et al., 2020; Yao et al., 2024), attention in cocktail-party scenarios (O’Sullivan et al., 2014; Geirnaert et al., 2021b,a), and many more.

To study and utilize neural tracking across these applications, data-driven models are employed to relate stimulus features to neural responses (Figure 1). These models can decode stimuli from neural responses (backward modeling (Wong et al., 2018; Alickovic et al., 2019; Geirnaert et al., 2021b)), encode neural responses from stimulus features (forward modeling (Wong et al., 2018; Alickovic et al., 2019; Geirnaert et al., 2021b)), or jointly optimize both representations (Dmochowski et al., 2018; de Cheveigné et al., 2018; Geirnaert et al., 2021b; Reddy Katthi and Ganapathy, 2021), using either linear or non-linear models. Their outputs - or the original signals if no model is used - are typically compared using the Pearson correlation coefficient, the so-called ‘neural tracking correlation’. This similarity metric is widely used to compare models (Reddy Katthi and Ganapathy, 2021; Thornton et al., 2022), stimulus features (Di Liberto et al., 2015; Gillis et al., 2021), or frequency bands (Di Liberto et al., 2015; Etard and Reichenbach, 2019; Zuk et al., 2021; Thornton et al., 2022), and in several applications serves as a proxy for speech understanding (Vanthornhout et al., 2018; Verschueren et al., 2019) or auditory attention (O’Sullivan et al., 2014; Geirnaert et al., 2021b; Roebben et al., 2024). Excellent reviews and tutorials on neural response modeling to natural continuous stimuli can be found in Crosse et al. (2021), with a focus on linear models, and Puffay et al. (2023), with a focus on non-linear models.

The main premise of this paper is that, despite its widespread use, the Pearson correlation coefficient is not directly interpretable as a performance metric. Comparing raw correlations across different models, stimulus features, preprocessing strategies, or experimental conditions can therefore lead to wrong conclusions (Zuk et al., 2021; Puffay et al., 2023). As such, terms as ‘reconstruction accuracy’, sometimes used to denote this neural tracking correlation, while intuitive, suggest a well-calibrated performance scale for direct comparison that generally does not exist. Consider, for example, the correlations found in Figure 2, obtained by reconstructing different speech features from EEG using backward modeling. Based solely on the observed correlations, one might conclude that the smallband (1–1.1 Hz) envelope is one of the best features for relating speech to neural responses, influencing subsequent applications in, e.g., attention decoding. However, narrowing the frequency band even further - ultimately approaching a single sinusoidal component - would likely increase the correlation even more, as only the phase needs to be reconstructed accurately. Such a feature, however, carries little information about speech processing or attention. This raises the question of whether higher correlations always indicate better performance and whether the correlation coefficient in itself is a reliable performance metric. This interpretability issue becomes even more pressing in the age of deep learning, where black-box models and their outputs are far less interpretable and controllable. Furthermore, it connects to the finding that canonical correlation analysis (CCA)-based models, which explicitly maximize correlation, tend to favor very low over higher frequency components (de Cheveigné et al., 2018). This is expected, as low-frequency signals exhibit stronger temporal autocorrelation, making phase alignment easier without necessarily improving the underlying neural representation. Consequently, maximizing the correlation coefficient alone does not necessarily optimize the quantity of

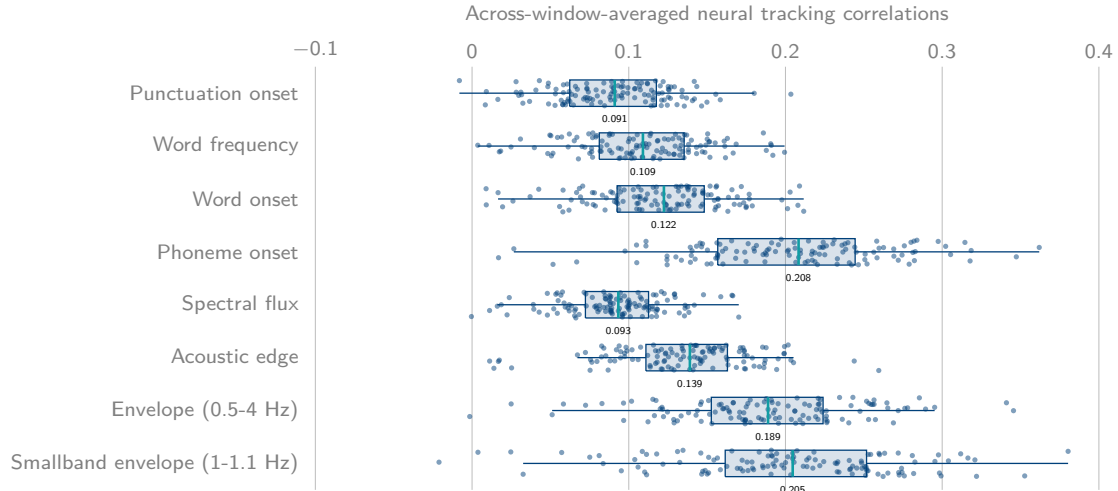


Figure 2: The per-participant average neural tracking correlations (Pearson correlation coefficient) across 5s-windows show clear differences between speech features, with the smallband envelope (1–1.1 Hz) yielding among the highest neural tracking correlations. The high correlation for a feature that carries hardly any information shows that correlation as such is not a proper metric to compare the performance of different features or models.

interest¹.

Key take-away #1

The Pearson correlation coefficient between neural responses and stimulus features is **not** a reliable performance metric **in itself**. Its interpretation depends on the corresponding null distribution, which can be influenced by various factors as the stimulus feature, model, or preprocessing.

To interpret neural tracking correlations meaningfully, they must be evaluated relative to the *null distribution*, representing the scenario in which no relationship exists between the stimulus and neural response. Under the null hypothesis, the expected correlation equals zero, but finite sample sizes inevitably produce non-zero correlations by chance (Combrisson and Jerbi, 2015). Statistical significance is therefore determined from the null distribution, either through a significance level defined as the $100(1 - \alpha)\%$ -percentile in the null distribution, with typical values for $\alpha = 0.01, 0.05, 0.10$, or even 0.005 as advocated in Benjamin et al. (2018) or through the associated p -value, the percentile of the null distribution corresponding to the observed correlation. Although significance is traditionally assessed by comparing the p -value with a fixed α -threshold, McShane et al. (2019) advocate instead interpreting p -values as continuous measures of evidence, among other factors, rather than binary decision criteria.

Constructing this null distribution is itself a non-trivial problem. Historically, standard parametric significance tests, such as the classical t-test, were used as it was impossible to empirically observe the underlying null distribution. However, these parametric tests rely on assumptions such as the independence of the individual time samples entering the correlation computation and a specific parametric distribution - assumptions that are typically violated. Nowadays, however, we have the means and computing power to generate correlations under the null hypothesis based on randomized stimulus-response pairings, for example, using circular shifting or phase scrambling (Crosse et al., 2021; Puffay et al., 2023). These non-parametric permutation approaches approximate the null distribution directly, removing the need to assume an underlying parametric distribution. The ideas of permutation testing to determine statistical significance were originally developed in the works of Fisher (Fisher, 1935) and Pitman (e.g., Pitman (1937)). Several works,

¹This also holds for unidirectional backward or forward models, shown to optimize correlation through minimizing the squared error (Biesmans et al., 2017).

also in biomedical research, have shown that permutation testing is superior to parametric testing, as it better operationalizes the null hypothesis by translating it directly into a computational procedure (Noguchi et al., 2021; Holt and Sullivan, 2023). However, although permutation tests are commonly described as assumption-free (e.g., Crosse et al. (2024)), every permutation procedure implicitly encodes assumptions about the null hypothesis, making the choice of the permutation procedure crucial and non-trivial for obtaining meaningful significance estimates (Holt and Sullivan, 2023). Excellent tutorials and overviews of permutation testing can be found in Maris and Oostenveld (2007); Holt and Sullivan (2023).

In this paper, we present a statistically principled framework for using and interpreting the neural tracking correlations, facilitating correct comparison of models, stimulus features, preprocessing, or experimental conditions. This framework addresses three challenges, discussed in the corresponding sections:

1. We analyze commonly used permutation procedures for neural tracking and argue that stimulus-response misalignment provides one of the most principled and practical estimates of the null distribution. Alternative approaches, including random shuffling, circular shifting, and phase scrambling, are shown to rely on different assumptions that can lead to biased significance estimates (Section 3).
2. We introduce a semi-parametric model of the permutation distribution based on the Fisher transform, combining the strengths of non-parametric and parametric modeling, and as such enabling accurate and efficient estimation of significance level and p -values. Permutation methods require thousands of permutations to generate reliable null distributions, imposing a heavy computational burden especially when comparing many models, features, or settings. We demonstrate that a parametric model on top of a non-parametric null distribution yields more reliable results than the raw empirical null distribution, especially when only a limited number of permutations is available (Section 4).
3. We combine both contributions into a unified framework for interpreting neural tracking correlations. Building on the modeled null distribution, we introduce the *null-normalized tracking score* (NNTS), a performance metric that enables meaningful comparison across models, stimulus features, and analysis settings, and demonstrate its relationship with the so-called match-mismatch accuracy metric (Section 5). For this purpose, we revisit the feature-comparison example from Figure 2.

Throughout the paper, “*Key take-away*” boxes summarize the main conclusions and can be read independently as a concise overview. A toolbox implementing the proposed methodology is available at Geirnaert et al. (2026b), and the code and data to reproduce analyses and results at Geirnaert et al. (2026a) and Geirnaert et al. (2026c).

2. Experimental setup

This section describes the experimental setup shared across all analyses in this paper: the stimulus features (Section 2.1), the dataset and preprocessing (Section 2.2), the backward decoding model (Section 2.3), and the validation procedure (Section 2.4). As the goal of this paper is to analyze and improve the underlying framework for using and interpreting the neural tracking correlations, and not to benchmark various models or stimulus features, a non-exhaustive list of example features and model is used.

2.1. Stimulus features

We use eight speech features to generate neural tracking correlations, visualized for a representative 5 s segment in Figure 3:

Envelope (see, e.g., Ding and Simon (2012)). Generated using a gammatone filter bank with 28 channels spaced by one equivalent rectangular bandwidth, and with centre frequencies from 50 until 5000 Hz. Subband envelopes, formed by raising absolute sample values to the power of 0.6, are averaged into a single envelope. This feature is prevalent in many studies.

Smallband envelope The smallband envelope is obtained by filtering the envelope feature between 1–1.1 Hz with a sixth-order Butterworth filter in the forward and backward direction. While this smallband is limited in application, it highlights potential pitfalls in null distribution estimation.

Acoustic edge (see, e.g., Brodbeck et al. (2018)). Acoustic edges are created by taking the first derivative of the broadband envelope above and setting negative values to zero. They are widely used alongside envelopes in studies focusing on acoustical features.

Spectral flux (see, e.g., MacIntyre et al. (2026)). Spectral flux quantifies the frame-to-frame change in the stimulus’ power spectrum, computed as the half-wave rectified sum of the differences between consecutive short-term spectra across the gammatone subbands. Unlike the envelope, it emphasizes spectral changes such as onsets, and is increasingly used as an acoustic feature in neural tracking studies (MacIntyre et al., 2026).

Phoneme onset (see, e.g., Di Liberto et al. (2015)). Phoneme onsets are created by annotating the stimulus with a value of one at every time step at which a phoneme starts, yielding a sparse impulse train. This feature, or a similar feature carrying more fine-grained information about the phonemes presented, is often used in studies investigating phonetic processing. Importantly, the (near-flat, broadband) spectrum of onset-based features can differ substantially from that of the neural response and of the other features. This spectral mismatch between feature and EEG is innocuous for reconstruction itself, but, as we will see, becomes important when constructing the null distribution.

Word onset (see, e.g., Broderick et al. (2018)). Word onsets are constructed analogously to phoneme onsets, placing an impulse of value one at the start of each word, and are likewise left unfiltered.

Word frequency (see, e.g., Gillis et al. (2021)). The word frequency feature uses the same word-onset impulses, but modulates each impulse by the log unigram frequency of the corresponding word, so that the feature additionally carries lexical information rather than only timing.

Punctuation onset Punctuation onsets, here defined to yield a highly sparse feature, place an impulse at the time of punctuation marks in the narrated text, marking prosodic and syntactic boundaries such as sentence and clause endings.

2.2. EEG dataset and preprocessing

We used a dataset involving 121 subjects who listened to the story “Milan”, written and narrated in Flemish by Stijn Vranken. The stimulus is 876 seconds long and was presented monaurally at 60 dBA. Stimuli were presented using the APEX 3 platform on a Windows laptop connected to an RME Multiface II sound card (RME, Haimhausen, Germany) and Etymotic ER-3A insert phones (Etymotic Research, Inc., IL, USA), which were electromagnetically shielded using CFL2 boxes (Perancea Ltd., London, UK). The setup was calibrated in a 2-cm³ coupler (Brüel & Kjær 4152) using stationary speech-weighted noise matched to the speech material. Experiments took place in an electromagnetically shielded and soundproofed room. The EEG was recorded using a BioSemi 64-channel ActiveTwo system. The dataset used here is part of the publicly available SparrKULee dataset (Accou et al., 2024), yet also contains data from additional participants that could not be made publicly available. All participants, including the additional ones, read and signed an informed consent approved by the Medical Ethics Committee UZ KU Leuven/Research (KU Leuven, Belgium), with reference S57102.

All EEG data were re-referenced to the common average and drift-filtered (first-order Butterworth, cutoff 0.5 Hz, applied forward and backward). All EEG and stimulus features were downsampled to $f_s = 8$ Hz using an anti-aliasing filter. For the smallband envelope condition, the EEG was additionally filtered to the matching 1–1.1 Hz band.

2.3. Backward modeling

A linear backward (decoding) model was used to reconstruct each stimulus feature from the neural responses (Crosse et al., 2021). Such a decoder is a spatio-temporal filter that linearly combines the EEG signals of all channels and their time-shifted versions to optimally reconstruct the stimulus feature. Let $r(t, c)$ denote the EEG sample at channel c at time t , and let $\mathbf{R} \in \mathbb{R}^{CL \times T}$ be the matrix collecting T EEG

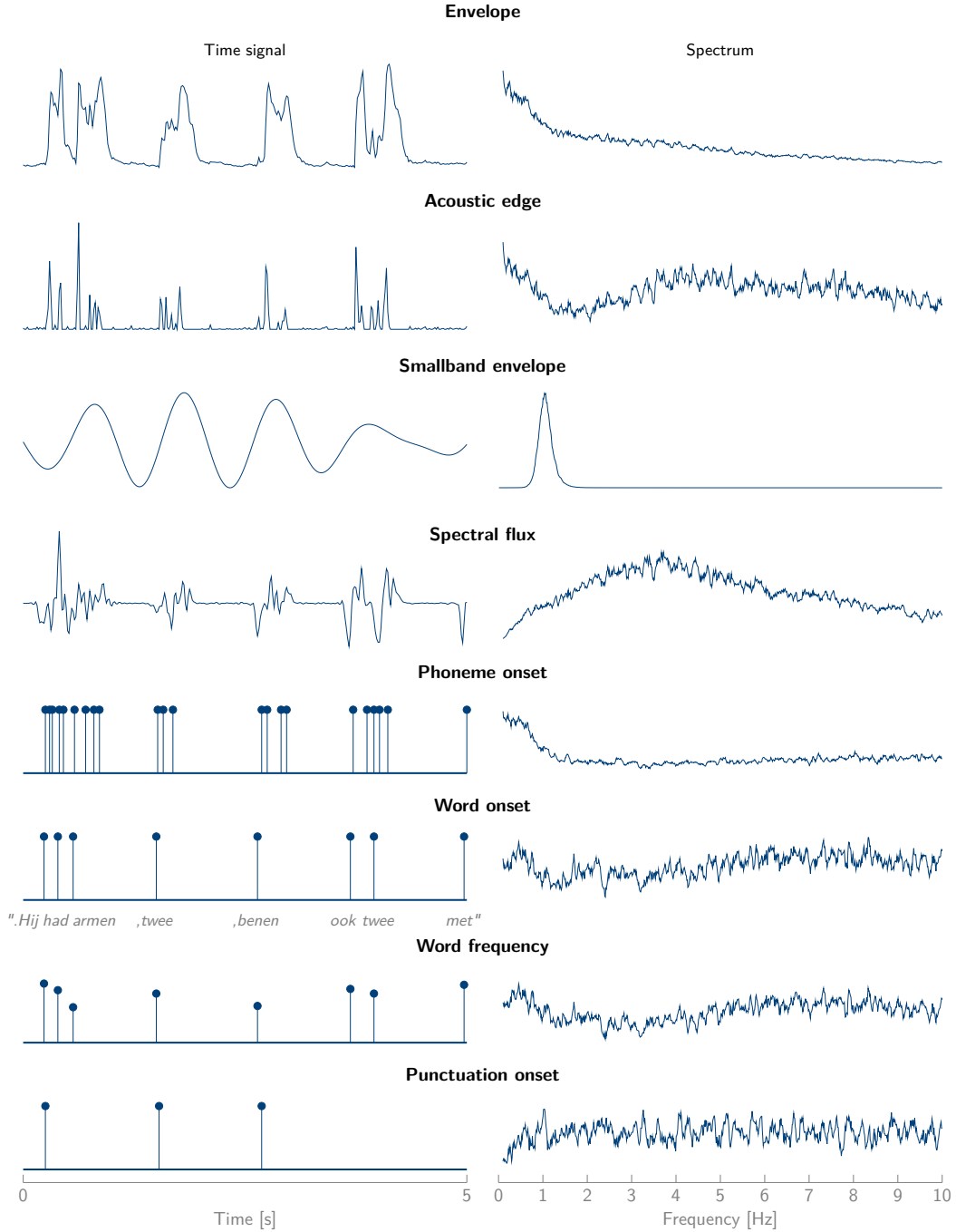


Figure 3: The eight speech stimulus features used throughout this paper, shown for a representative segment in the time and frequency domain (here at 64 Hz, shown up to 10 Hz to improve readability). The continuous acoustic features (envelope, smallband envelope, acoustic edge, spectral flux) are contrasted with the sparse, unfiltered onset-based features (phoneme onset, word onset, word frequency, punctuation onset), which differ markedly in their temporal and spectral statistics.

samples across all C channels (here: $C = 64$) expanded with L time-lagged copies with lags up to 500 ms. The reconstruction $\hat{s}(t)$ of the stimulus feature $s(t)$ is then obtained as

$$\hat{s}(t) = \sum_{c=1}^C \sum_{\tau=0}^{L-1} g_c(\tau) r(t + \tau, c),$$

with $g_c(\tau)$ the decoder weight for the τ^{th} time lag in channel c .

The decoder is estimated by minimizing the mean-squared error between the reconstruction and the actual feature:

$$\mathbf{g} = \arg \min_{\mathbf{g}} \mathbb{E} \left\{ (\hat{s}(t) - s(t))^2 \right\},$$

which yields the following closed-form, ridge-regularized least-squares solution:

$$\mathbf{g} = (\mathbf{R}\mathbf{R}^\top + \lambda \mathbf{I})^{-1} \mathbf{R}\mathbf{s},$$

where $\mathbf{g} \in \mathbb{R}^{CL}$ is the vector collecting the decoder weights across all channels and time lags, $\mathbf{s} \in \mathbb{R}^T$ is the vector containing T stimulus-feature samples, $\mathbf{I} \in \mathbb{R}^{CL \times CL}$ the identity matrix, and λ the ridge parameter counteracting overfitting (Crosse et al., 2021). This decoder is then used to reconstruct the stimulus feature, after which the Pearson correlation coefficient between the reconstructed and actual stimulus feature is computed across N samples, corresponding to the window length $w = \frac{N}{f_s}$.

Note that a linear decoder is suboptimal for the sparse onset features (phoneme and word onsets, word frequency, and punctuation onset). A backward decoder reconstructs the feature as a weighted sum of continuous, band-limited EEG channels and their time lags, so its output is an inherently smooth, continuously valued signal. An onset feature, in contrast, is mostly zero with occasional unit impulses - a sparse signal that no linear combination of smooth EEG signals can reproduce; at best the decoder yields a low-pass, smeared estimate with broad bumps around the onset times rather than the impulses themselves, which lowers the attainable Pearson correlation. Recovering the sparse train would require a non-linear (e.g., thresholding) output stage that the linear model lacks. We nonetheless deliberately use the same linear model across all features to keep the comparison controlled and isolate the effect of the null-distribution construction, which is the focus of this work.

2.4. Validation procedure

Using leave-one-window-out cross-validation, the decoder was retrained on all but one window and applied to the held-out window to generate its feature reconstruction; this is repeated for every window of a given length, and for every window length of interest. For each window, the Pearson correlation coefficient between the reconstructed (decoded) stimulus feature from the EEG and the original one is computed. For example, for 5 s windows and the 876 s-long story, this results in 175 true correlations, yielding a robust average correlation estimate per participant, as visualized in Figure 2. The resulting correlations are available at Geirnaert et al. (2026c).

3. Constructing null distributions

Considering the Pearson correlation between a (decoded) neural response and (encoded) stimulus feature, the null distribution represents the correlations expected when no relationship exists between the two signals. In the example used here, the signals consist of a reconstructed stimulus feature, decoded from the EEG using a linear decoder, and the actual feature.

The null hypothesis

There is no relationship between the neural response and the stimulus feature, i.e., the observed correlation is due to chance.

For the null distribution to be correct, it must reflect the testing context, and in particular the statistical properties of the signals entering the correlation. Stimulus features are typically correlated over time, and this temporal structure (equivalently, their autocorrelation or spectrum) must be preserved when generating null correlations. Furthermore, since the decoder shapes the spectrum of the reconstructed signal and, therefore, also influences the measured correlation, it should also be taken into account when constructing the null distribution. This makes the null distribution decoder-specific, with each model assessed against its own null distribution.

Mathematically, the null hypothesis states that the population correlation between neural response and stimulus feature is zero. Consequently, the null distribution, which reflects this hypothesis, must itself have a mean correlation of zero. This is a non-trivial requirement, and one worth checking explicitly once the null distribution has been constructed:

Key take-away #2

The null distribution of the correlation coefficients should have a **mean of zero**, reflecting the null hypothesis of no relationship between neural response and stimulus feature.

A conceptually different family of approaches randomizes the training process itself, rather than the test-time alignment between a fixed decoder’s output and the stimulus feature. Such alternative approaches often correspond, in subtle ways, to different null hypotheses. For example, MacIntyre et al. (2026) train random decoders on misaligned data but test them on aligned data, where a true relationship exists between neural response and stimulus feature. Besides being considerably more computationally intensive, this procedure relocates the null hypothesis to the level of the model itself, testing whether a trained model extracts meaningful patterns beyond its inductive bias. As this conflates neural tracking with the effectiveness of a specific model’s training, we instead test for a relationship directly between the reconstructed signal and the stimulus feature.

3.1. Comparing different permutation methods

All non-parametric permutation methods construct the null distribution by repeatedly generating *surrogate signals*: signals stripped of any true temporal correspondence with the neural response, but that ideally retain all other relevant statistical properties of the original feature. Each surrogate is then correlated with the reconstruction of the stimulus feature from the neural response, yielding one sample of the null distribution. Crucially, the properties a permutation method holds fixed when generating surrogates define an implicit *null hypothesis*, and these preserved properties must be matched to the statistic under study (Lancaster et al., 2018). Different surrogate methods therefore do not approximate a single “true” null distribution more or less accurately. They operationalize *different* null hypotheses, and we consider there to be no method-independent best choice. Here, we consider four methods (random shuffling, circular shifting, phase scrambling, and misalignment), examine the null each encodes, and motivate our adoption of misalignment for neural tracking of natural stimuli. The modeling framework of Section 4 and 5 is agnostic to this choice: it requires only a set of valid null correlations, so any method can be substituted without affecting the subsequent methodology.

3.1.1. Comparing the null distributions

Because each method implicitly operationalizes the null hypothesis in a different way, there is no external ground-truth null distribution against which they can all be validated. Instead, we compare the null distributions the methods produce (Figure 4) and reason about the implicit null hypothesis each encodes. The only fully method-agnostic requirement is the necessary condition of a zero mean correlation (Key take-away #2). The four methods clearly produce different null distributions and therefore different significance levels, confirming that the choice of method is not a mere implementation detail.

3.1.2. Random shuffling

In random shuffling, the surrogate is generated by randomly permuting the samples of the original (synchronized) feature, i.e., by reordering its values according to a random permutation of the time indices. This

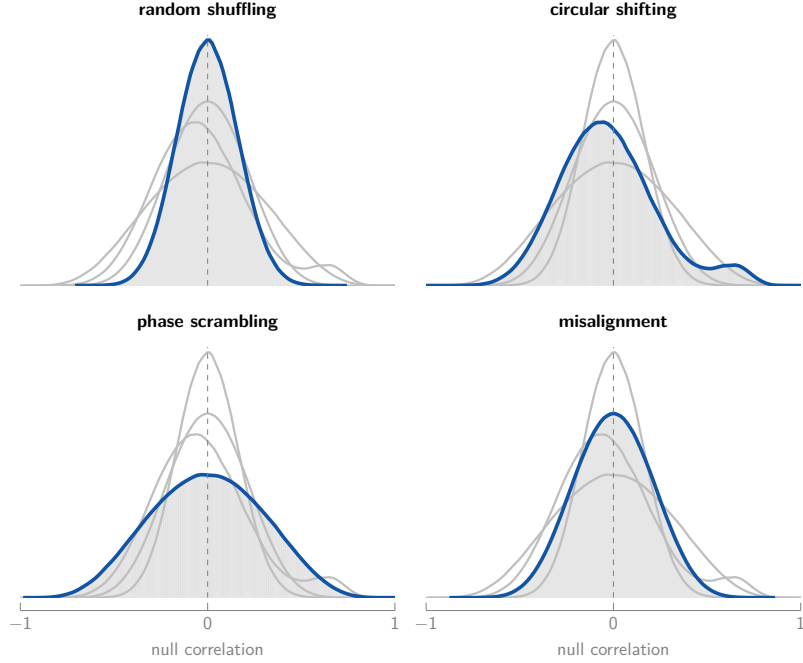


Figure 4: Null distributions of the 4 permutation methods for the envelope at 5 s windows, across all participants and windows. Each panel highlights one method (histogram and kernel-density estimate) against the other three (gray, identical across panels). Random shuffling whitens the feature, giving a narrow distribution around zero; circular shifting yields a non-zero mean with a secondary bump at high correlations due to autocorrelation; phase scrambling is broad, preserving the amplitude spectrum; and misalignment gives a unimodal distribution reflecting the natural variability of real, temporally unrelated segments.

implicitly encodes the null hypothesis that the neural response is unrelated to a temporally *unstructured* version of the feature: it destroys any temporal correspondence with the neural response, but also the temporal structure of the feature itself. Permuting the samples whitens the signal and flattens its spectrum: the speech envelope, for example, predominantly contains low-frequency energy, whereas its shuffled surrogate carries substantially more high-frequency energy. Because the correlation between two signals depends on their spectra, and as the whitened surrogate shares little structure with the smooth, low-frequency reconstruction, the resulting null correlations collapse tightly around zero, giving by far the narrowest null distribution (Figure 4). This artificially narrow null distribution clearly underestimates what the true spread of a null distribution would be, rendering the test liberal. As its implicit null (no tracking of a whitened feature) corresponds to no meaningful notion of neural tracking, random shuffling is inappropriate for this purpose.

3.1.3. Circular shifting

In circular shifting, the surrogate is obtained by shifting the original feature over a random number of samples in a circular fashion, such that samples shifted beyond the end of the signal reappear at its beginning. This encodes the null hypothesis that the neural response is unrelated to the feature *at this particular temporal alignment*, testing purely for temporal alignment. It has a clear theoretical appeal: unlike all other methods, it preserves the exact spectrum (and autocorrelation) of the feature. However, precisely because stimulus features can be strongly autocorrelated, small shifts leave the signal - and hence its correlation with the neural response - almost unchanged, producing surrogates that remain partially aligned. In Figure 4, this appears as a broad null distribution with a non-zero mean, thus violating the necessary condition of the null distribution, and a secondary bump at high correlation. Conceptually, the null in circular shifting holds the spectrum exactly fixed, testing only alignment rather than whether the response tracks this stimulus content over natural alternatives. Practically (and more decisively), the strong

autocorrelation means that, for short windows (typically below a few minutes²), one cannot draw enough shifts that avoid the autocorrelation to obtain sufficient independent surrogates, making circular shifting practically infeasible.

3.1.4. Phase scrambling

In phase scrambling, the surrogate is created in the frequency domain by randomizing the phases (while preserving the conjugate symmetry required for a real-valued signal) and recombining them with the original amplitude spectrum. This encodes the null hypothesis that the feature is a *stationary linear Gaussian process* with a given power spectrum (Lancaster et al., 2018). This technique removes the practical limitation of circular shifting as it allows generating unlimited independent surrogates from a single short window and avoids the autocorrelation problem. It introduces more randomness than circular shifting, since the phase spectrum changes drastically, but it still exactly preserves the amplitude spectrum, and randomizing the phases destroys the crucial phase-aligned structure (such as acoustic onsets) characteristic of natural auditory features. Various advanced signal processing techniques such as iterative amplitude-adjusted Fourier transform (IAAFT) surrogates have been proposed, that, for example, additionally preserve the amplitude distribution of the feature (Schreiber and Schmitz, 1996), but still constrain the surrogate to the original power spectrum and therefore share the same fundamental limitation. In Figure 4, this yields a broad, smooth null distribution. Like circular shifting, phase scrambling treats the stimulus and/or neural spectrum as fixed and known, which does not reflect the natural variability of stimulus content across a real experiment.

3.1.5. Misalignment

In misalignment, the null correlation is computed between the reconstruction $\hat{s}(t)$ and a segment of the *actual* feature taken from a non-corresponding part of the recording (or a different recording with similar characteristics), so that feature and response are misaligned in time. The trained decoder and its reconstruction are kept exactly as in the real analysis, and only the feature against which the reconstruction is correlated is replaced by a genuine but temporally unrelated excerpt. This encodes the null hypothesis appropriate for *content-specific* neural tracking: the neural response is no more related to the presented stimulus than to any other genuine stimulus of the same kind.

The property distinguishing misalignment from the spectrum-fixing methods is that the feature’s statistics such as the spectrum and fine structure are allowed to *vary* across surrogates, because each surrogate is a different real segment with its own statistics. Moreover, by the symmetry of the correlation, the pooled null distribution across windows can equally be interpreted as pairing each stimulus segment with reconstructions from many different epochs, and hence many different background-EEG realizations, such that this variability is present on both the stimulus and the neural-response side. This is important because a valid surrogate should reproduce the variability that would be observed across genuine repetitions of the measurement - a critical property of any surrogate method (Lancaster et al., 2018). Misalignment reproduces precisely the variability in stimulus-response null pairings that one would expect when the same stimulus is presented repeatedly, rather than treating the stimulus (or neural) spectrum as fixed and known.

Misalignment does rest on the assumption that the stimulus, and its statistical relationship to the neural response, is approximately stationary across the set of segments from which surrogates are drawn (e.g., the recording). This assumption constrains how the set of misaligned segments should be constructed. Overall, misalignment produces a clean, unimodal null distribution without the pathologies of the other methods (Figure 4). Its main practical limitation is that it requires sufficiently long recordings to draw enough misaligned segments, one of the motivations for the semi-parametric modeling introduced in Section 4.

Finally, even with misalignment, care must be taken when selecting misaligned segments, to avoid segments that retain non-zero autocorrelation with the original, aligned segment. While circular shifting is vulnerable to autocorrelation at any timescale, even over a few samples, misalignment is vulnerable only to

²For the envelope, whose autocorrelation extends to approximately ± 2.5 s, yielding 100 valid and independent shifts, spaced by the autocorrelation length, requires a window of at least 4.25 min at 8 Hz.

autocorrelation over much longer timescales (seconds or even minutes). One example is fNIRS-based neural tracking (Wilroth et al., 2026), where the very long autocorrelation of the hemodynamic response even causes misalignment to yield spurious correlations, requiring a tailored approach to ensure the null hypothesis is properly satisfied. Long autocorrelation must also be considered when drawing multiple misaligned segments for the same participant, to ensure the resulting null correlations are approximately independent.

3.1.6. Summary: misalignment as the adopted default

The four methods encode genuinely different null hypotheses and produce correspondingly different null distributions (Figure 4). Random shuffling tests against a whitened feature, an uninteresting null that collapses the null distribution and renders the test liberal. Circular shifting and phase scrambling both hold the stimulus spectrum fixed (exactly, or in expectation), testing whether the response tracks this specific realization of the spectrum rather than this stimulus content over natural alternatives. Circular shifting is additionally practically infeasible on short windows below a few minutes due to autocorrelation. Misalignment instead lets the natural statistics of the stimulus vary across surrogates, operationalizing the null hypothesis of content-specific tracking that we here consider most appropriate for neural tracking of natural stimuli. Based on the arguments in Section 3.1.5, we adopt misalignment throughout the remainder of this paper, with the subsequent framework accommodating any valid alternative.

Key take-away #3

Each surrogate method to create the null distribution encodes a different null hypothesis. We adopt **misalignment**, as it operationalizes the null hypothesis of *content-specific* neural tracking: it preserves the natural statistics of the stimulus, assuming stationarity, while letting them vary across surrogates, rather than holding the spectrum fixed (circular shifting, phase scrambling) or destroying it (random shuffling). The subsequent framework is agnostic to this choice.

4. Accurate and efficient semi-parametric modeling of the null distribution and estimation of significance levels

In the previous section, we adopted misalignment for generating null correlations. However, reliable null distributions require thousands of permutations, since significance levels are determined by the far tails of the distribution. When data is limited, generating that many misaligned segments may not be feasible. Even with ample data, computing thousands of permutations imposes a substantial computational burden, particularly when null distributions are needed across multiple features, models, or window lengths.

Therefore, we introduce semi-parametric modeling of the null distribution to estimate significance levels more efficiently and accurately. We first introduce the Fisher transformation as the basis for modeling null correlations (Section 4.1), then demonstrate how semi-parametric modeling improves significance level estimation, for example, when permutations are limited (Section 4.2), and finally show how significance levels can be predicted across window lengths without regenerating the null distribution each time (Section 4.3).

4.1. The normal distribution after Fisher transform as a semi-parametric null model

To model the null distribution of Pearson correlation coefficients between neural responses and stimulus features, we use the normal distribution after the Fisher transformation. Fisher proposed a normalizing and variance-stabilizing transformation on the Pearson correlation coefficients r using the inverse hyperbolic tangent function (Fisher, 1921), defined as:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = \operatorname{artanh}(r).$$

This Fisher transformation maps the correlation coefficients $r \in [-1, 1]$ to the real line $z \in]-\infty, +\infty[$, such that

$$z \sim \mathcal{N}(\mu_z, \sigma_z^2)$$

is approximately normally distributed with mean $\text{artanh}(\rho)$ and variance $\frac{1}{N-3}$, with ρ the true population correlation and N the number of samples in the correlation computation (assuming the N samples are independent). As this variance only depends on N and not on ρ , it is called a ‘variance-stabilizing’ transformation (Hotelling, 1953). Asuero et al. (2006); Fouladi and Steiger (2008) showed that this normal approximation holds even for N as low as 11, making it suitable for modeling null correlations even at short window lengths. In Appendix A, we show it outperforms alternatives such as the truncated normal or Student’s t -distribution.

For null correlations $r^{(\text{null})}$, where the true population correlation $\rho = 0$ by construction, the mean of $z^{(\text{null})}$ is zero and is modeled as such: $z^{(\text{null})} \sim \mathcal{N}\left(0, \sigma_z^{(\text{null})^2}\right)$. Crucially, the variance is not set to $\frac{1}{N-3}$, which assumes independent time samples - an assumption violated by the autocorrelated neural and stimulus signals typical of neural tracking (Section 3). Instead, it is estimated empirically from the misalignment permutations $\{z_i^{(\text{null})}\}_{i=1}^n$ using the zero-mean unbiased estimator³:

$$\sigma_z^{(\text{null})^2} \approx \hat{\sigma}_z^{(\text{null})^2} = \frac{1}{n} \sum_{i=1}^n z_i^{(\text{null})^2},$$

with n the number of permutations. This semi-parametric approach retains the key advantage of permutation testing, i.e., correctly capturing the autocorrelation-inflated null variance, while gaining the efficiency of a parametric model for characterizing and extrapolating the tail shape. The significance level at α -level can then be derived from this modeled null distribution as

$$\text{significance level}_{\alpha} = \tanh\left(P_{100(1-\alpha)\%}\left(\mathcal{N}\left(0, \hat{\sigma}_z^{(\text{null})^2}\right)\right)\right), \quad (1)$$

with $P_x(\cdot)$ the x^{th} percentile. The Fisher-transformed normal distribution has been used sporadically in the neural tracking literature, mostly to facilitate t -tests on correlation (Fiedler et al., 2019; Chen et al., 2023; Lopez-Gordo et al., 2025; Daeglau et al., 2025), but no explicit or systematic semi-parametric modeling of the null distribution using this approach has been proposed.

To evaluate the semi-parametric model, we compare it against empirical null distributions for the same eight features, using the same data as in Section 2: the envelope, the smallband envelope, the acoustic edge, spectral flux, phoneme onsets, word onsets, word frequency, and punctuation onsets. Null correlations are generated by misalignment with 100 000 permutations per participant and window length (20, 10, 5, 2, 1 s at 8 Hz)⁴, and are treated as the true underlying null distribution. These null correlations (Geirnaert et al., 2026c) and the code reproducing all subsequent results (Geirnaert et al., 2026a) are publicly available.

The semi-parametric model is assessed globally (Section 4.1.1), via fitted PDFs against histograms and QQ-plots, and by how accurately it estimates significance levels from the tail of the distribution (Section 4.1.2).

4.1.1. Global model assessment

Figure 5 shows PDFs and QQ-plots of the semi-parametric model against the empirical null distribution (using 100 000 permutations randomly sampled from the accumulated distribution across all participants) for four representative examples, including some edge cases.

For typical features and window lengths above 1 s (the large majority of scenarios) the original null distribution already closely follows the normal distribution (Figures 5a, 5c), in which case the influence of the Fisher transform is negligible. This is expected: at longer windows, N is large enough such that correlations remain well within $[-1, 1]$ in practice, and the Pearson correlation coefficient, being a normalized sum over many terms and therefore loosely based on the central limit theorem, converges toward normality. In edge

³Given the mean is assumed zero by construction, no degree of freedom needs to be sacrificed to estimate the mean, resulting in the $\frac{1}{n}$ -scaling instead of the generally used $\frac{1}{n-1}$ -scaling for unbiased estimation of the variance.

⁴For 20 and 10 s windows, only 9288 and 37 758 permutations, respectively, were used due to the limited number of non-overlapping windows.

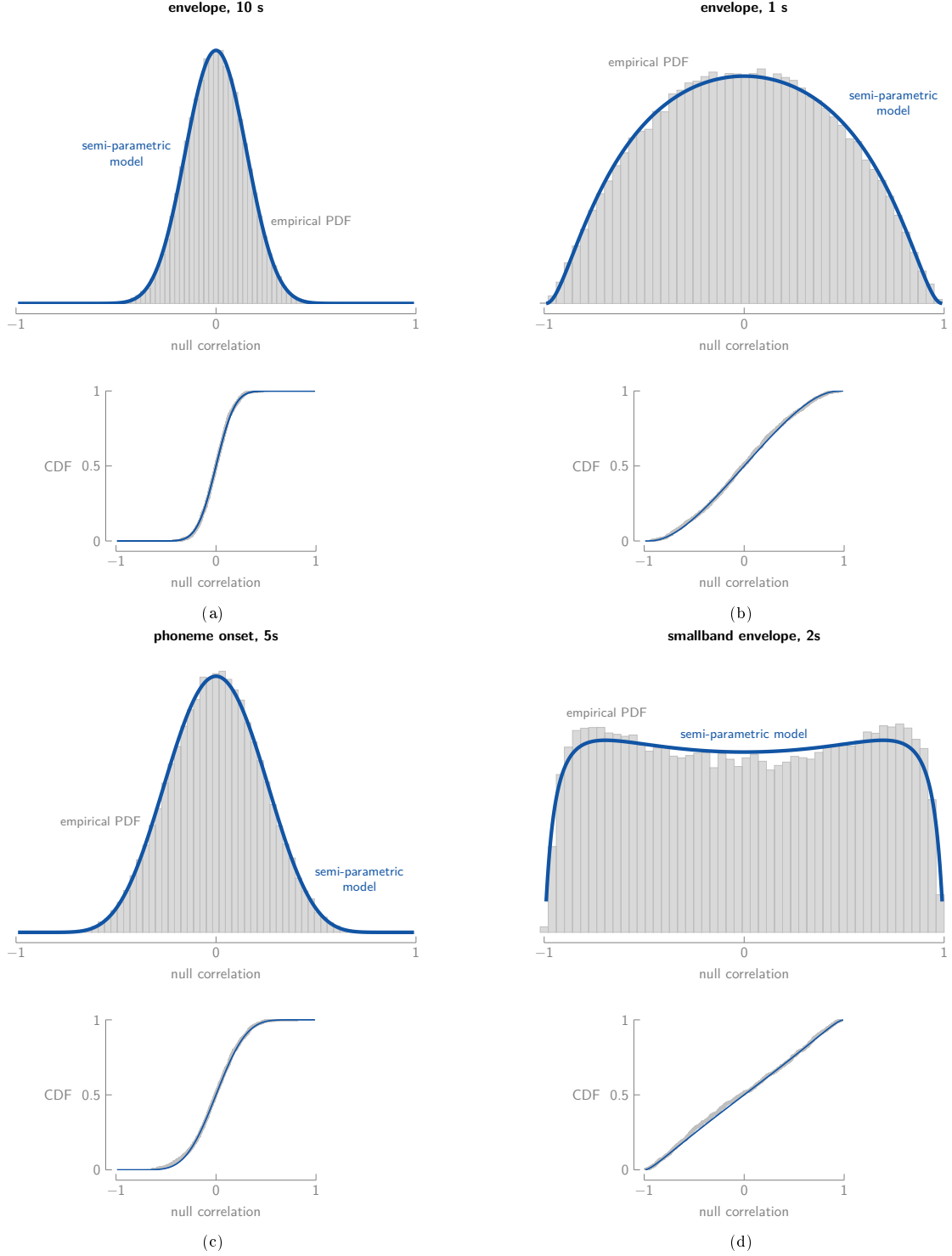


Figure 5: PDFs and QQ-plots of the semi-parametric model (normal distribution after the Fisher transform) against the empirical null distribution (100 000 samples accumulated across all participants) for **(a),(b)**: the envelope at 10s and 1s windows, **(c)**: phoneme onsets at 5s windows, and **(d)**: the smallband envelope at 2s windows. All panels are shown in the original correlation domain. **(a),(c)** represent the typical case: the null distribution of raw null correlations closely matches the normal distribution. **(b)** illustrates the effect of short windows, which widens the distribution and increases non-normality. **(d)** shows an edge case - short windows combined with a spectrally narrow feature - where heavier tails are present but still well captured by the semi-parametric model.

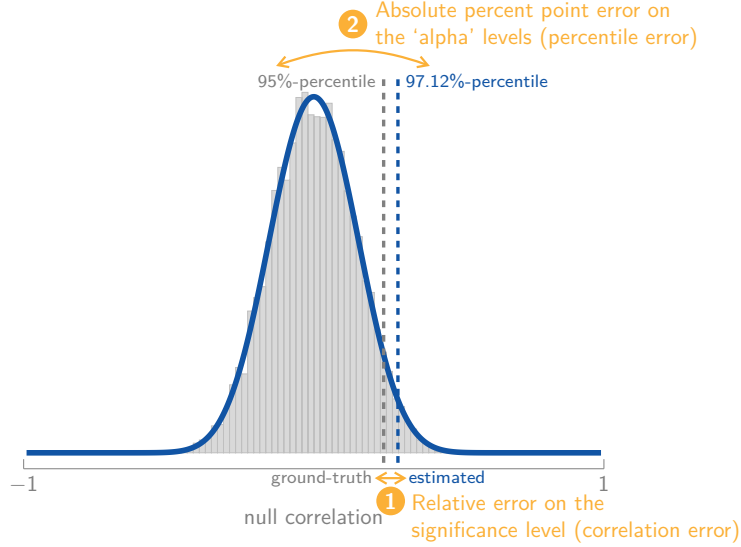


Figure 6: To evaluate the significance level estimation from the modeled distribution, i.e., given a certain α -level, what is the corresponding correlation-threshold based on the parametric model, we use two performance metrics. The *correlation error* is the relative error between the estimated and true significance-level thresholds (in correlation units). The *percentile error* is the resulting absolute deviation from the nominal α -level in percentage points.

cases such as short windows or spectrally narrow features (Figures 5b, 5d), the empirical null distribution has substantially heavier tails, yet the semi-parametric model tracks these faithfully. Since significance levels are determined precisely by the tails, we evaluate this accuracy directly in the next section.

4.1.2. Significance level assessment

We evaluate the quality of the semi-parametric model specifically in the context of significance levels, i.e., based on the tail of the null distribution. The estimated significance level from the parametric model for a given α -level is assessed using two complementary metrics illustrated in Figure 6: the *correlation error* (relative error on the estimated significance-level threshold in correlation units) and the *percentile error* (absolute deviation from the nominal α -level in percentage points). Both metrics are needed because a small relative error on the correlation threshold does not necessarily translate into a small error on the nominal α -level, depending on the probability mass in the underlying distribution.

Figure 7 shows both metrics per feature and window length, averaged across all 121 participants. Correlation errors are generally very low (Figure 7a), averaging 0.51% (standard deviation (SD): 0.26%) across all features and window lengths. Percentile errors are equally small (Figure 7b), averaging 0.13 percentage points (SD: 0.14), with the maximum at phoneme onsets and 1s windows and remaining below 0.5 percentage points in all other cases. The difference in relative correlation error and absolute percentile error for the phoneme onset feature at 1s shows the importance of having both metrics: while the error on the significance threshold itself is quite low, in terms of the underlying empirical distribution, the percentile error is larger than for other window lengths due to the wide phoneme onset null distribution. In practical terms, the semi-parametric model estimates the $\alpha = 0.05$ significance level typically with an effective α between [0.045, 0.055].

These results show that the normal distribution after Fisher transformation yields highly accurate modeling results of the null distribution of correlations and especially of the significance levels found in the tail of these null distributions. Moreover, characterizing the model as a normal distribution with a single estimated parameter, $\hat{\sigma}_z^{(\text{null})}$, brings along several interesting properties and is easy to work with, as shown in the following analyses.

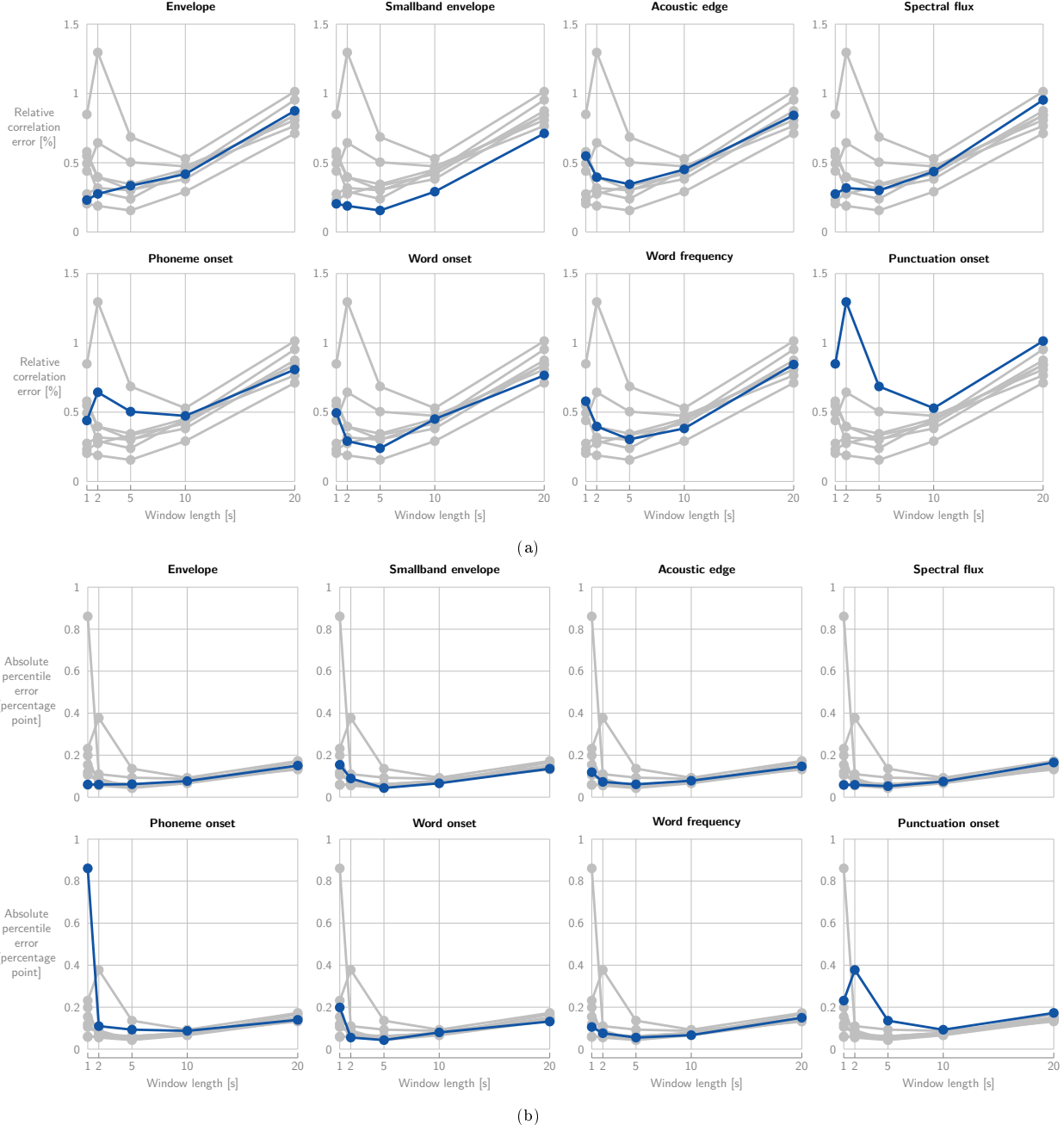


Figure 7: **(a)** The average (across participants) relative correlation error on the estimated 95%-significance level using the normal distribution after Fisher transformation, remaining below 1.25% across features and window lengths. **(b)** The average (across participants) absolute percentile error on the nominal $\alpha = 0.05$ level remains below 0.9 percentage point across features and window lengths.

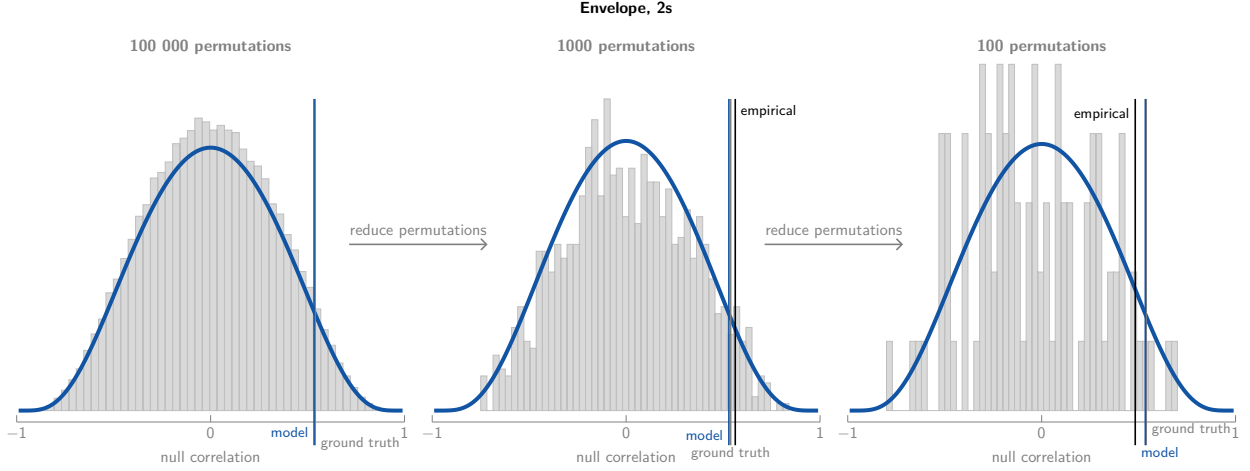


Figure 8: With few permutations, the tail of the null distribution is undersampled, making direct empirical estimation of the significance level, indicated by the vertical line ‘empirical’, unreliable w.r.t. the ground-truth based on 100 000 permutations (vertical line ‘ground truth’). The semi-parametric model mitigates this by modeling the full null distribution based on all available permutations and deriving the significance level, indicated by the vertical line ‘model’, from the analytical tail.

4.2. Efficient estimation of significance levels

Computing 100 000 misalignment permutations is computationally expensive, particularly when null distributions are needed across multiple features, models, or window lengths. Depending on the window length, generating sufficient permutations might even be infeasible if data is limited. Yet reliable significance level estimation requires dense sampling of the distribution’s tail, which is severely undersampled when few permutations are available, leading to worse significance level estimation based on the empirical null distribution, as illustrated in Figure 8.

The semi-parametric model resolves this tension: rather than estimating the significance level directly from an empirical histogram of null correlations, it models the full null distribution based on all available permutations and extracts the significance level from the modeled, more reliable parametric tail. This pooling of information stabilizes the estimate and reduces its sensitivity to the number of permutations used (Figure 8).

To evaluate this, we compare semi-parametric and empirical significance level estimates as the number of permutations is reduced. The null distribution and significance level based on 100 000 permutations serves as ground truth. For each combination of number of permutations K , feature, participant, and window length, we draw 100 random subsets (i.e., 100 sets, each with K samples of permuted correlations) and estimate the 95% significance level both empirically and via the semi-parametric model from each K -sample subset. Figure 9 shows the resulting average correlation and percentile errors. The semi-parametric model outperforms the empirical estimate for all numbers of permutations below $K = 10\,000$ permutations, with a negligible difference at very high numbers of permutations. Notably, the percentile error of the semi-parametric estimate remains below one percentage point on average when at least $K = 100$ permutations are used.

Figure 10 shows the mean $\pm$ SD estimated significance level across participants and resamplings for the punctuation onset feature at 5s for both the semi-parametric model and empirical estimate across number of permutations K , allowing to better explain why the semi-parametric model yields more accurate significance level estimates by analyzing the bias and variance. Firstly, Figure 10 clearly reflects that the empirical estimator is a biased yet consistent estimator, as the bias is clearly nonzero for small numbers of permutations, yet asymptotically becomes zero. The semi-parametric model is also biased due to model approximation errors, yet this bias is consistently lower except for very high numbers of permutations where it is negligible. The dominant advantage of the semi-parametric estimate, however, is the variance: across different resamplings, the empirical estimate shows much higher variance than the semi-parametric model,

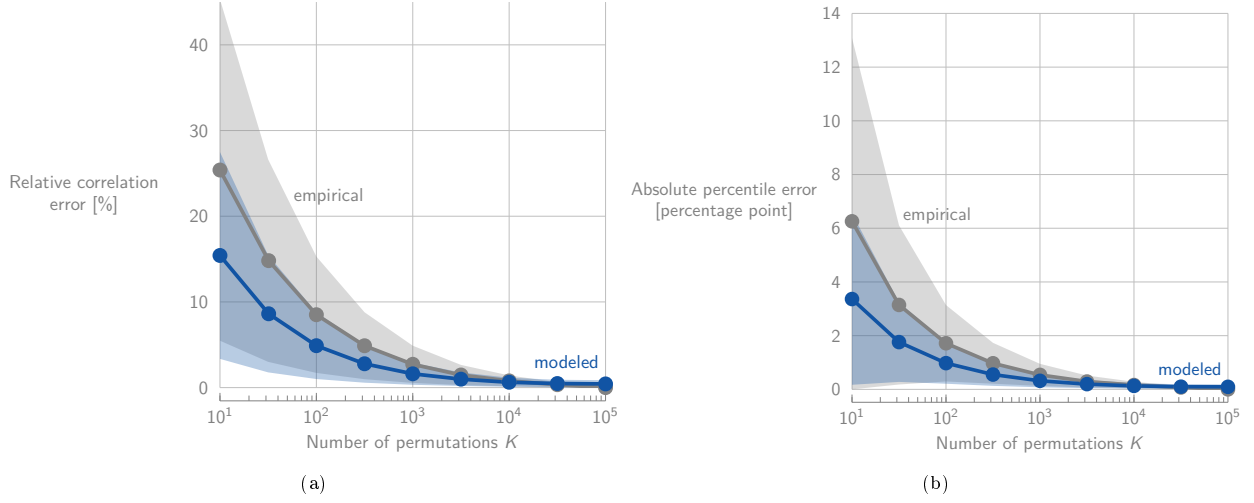


Figure 9: **(a)** The average (across participants, features, window lengths, and resamplings; shaded area: ± 1 SD) relative correlation error on the estimated 95%-significance level as a function of the number of permutations K . **(b)** The corresponding average absolute percentile error shows that the semi-parametric model outperforms the empirical estimate below $K = 10\,000$ permutations, with consistently below 1 percentage point errors when more than $K = 100$ permutations are used.

as it must characterize the tail from the few permutations that happen to fall there. In contrast, the semi-parametric model pools all available (limited number of) permutations to fit the full null distribution and extrapolates the tail analytically from a more robust distribution model, yielding much lower variance across a wide range of number of permutations.

In summary, the semi-parametric model yields substantially superior significance level estimates up to 10 000 permutations, with accurate estimates already from 1000 permutations (mean percentile error: 0.32 percentage points; SD: 0.24 across participants, features, window lengths, and resamplings).

4.3. Extrapolating significance levels across window lengths

While the previous section shows that the semi-parametric model drastically reduces the number of permutations required for accurate significance level estimation, the entire procedure still needs to be redone for every window length of interest. Here, we propose a method to extrapolate the significance level from one window length to any other, using the normal distribution after the Fisher transform.

Fouladi and Steiger derived expressions for the first eight cumulants of the Fisher transform under the null hypothesis of zero population correlation (Fouladi and Steiger, 2008). For the variance of $z^{(\text{null})} = \text{atanh}(r^{(\text{null})}) \sim \mathcal{N}(0, \sigma_z^{(\text{null})^2})$, this yields the following expression⁵:

$$\sigma_z^{(\text{null})^2} = V(N) = \begin{cases} \frac{\pi^2}{12} - \frac{1}{2} \sum_{k=1}^{\frac{N-4}{2}} k^{-2}, & N \text{ even} \\ \frac{\pi^2}{4} - 2 \sum_{k=1}^{\frac{N-3}{2}} (2k-1)^{-2}, & N \text{ odd}, \end{cases} \quad (2)$$

with $N = wf_s$ the number of samples used to compute the correlation, w the window length in seconds, and f_s the sampling frequency in Hz. Crucially, consistent with the explanation in Section 4, this expression $V(N)$ for the variance *only* depends on N , and hence the window length. Following the same approach

⁵Except at very low sample sizes ≤ 10 , this variance is well approximated by $\frac{1}{N-3}$, as proposed in Section 4.

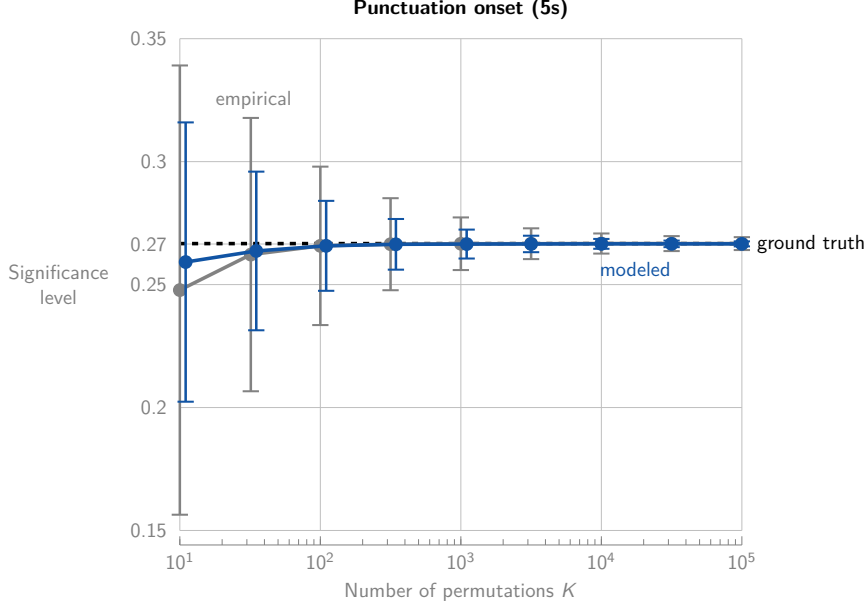


Figure 10: Mean $\pm$ SD (across participants and resamplings) estimated 95%-significance level for the punctuation onset feature at 5 s windows, as a function of the number of permutations. The semi-parametric model shows lower bias and substantially lower variance than the empirical estimate, explaining its superior performance.

used for performance curve modeling in selective auditory attention decoding (Geirnaert et al., 2025), we use Equation (2) not to compute the variance directly (given the independence assumption is violated), but to *rescale* an already-estimated variance from one window length to another. Given an estimated variance $\hat{\sigma}_{z,\text{base}}^{(\text{null})^2}$ at a base window length $w_{\text{base}} = N_{\text{base}}/f_s$, the variance at any target window length $w_{\text{target}} = N_{\text{target}}/f_s$ is:

$$\hat{\sigma}_{z,\text{target}}^{(\text{null})^2} = \hat{\sigma}_{z,\text{base}}^{(\text{null})^2} \frac{V(N_{\text{target}})}{V(N_{\text{base}})}. \quad (3)$$

Combined with Equation (1), the significance level at any target window length w_{target} can then be extrapolated from a single estimated variance $\hat{\sigma}_{z,\text{base}}^{(\text{null})^2}$ at w_{base} :

$$\text{significance level}_{\alpha}(w_{\text{target}}) = \tanh\left(P_{100(1-\alpha)\%}\left(\mathcal{N}\left(0, \hat{\sigma}_{z,\text{base}}^{(\text{null})^2} \frac{V(N_{\text{target}})}{V(N_{\text{base}})}\right)\right)\right). \quad (4)$$

To evaluate extrapolation performance, we extrapolate the significance level between every pair of window lengths, using 10 000 permutations per participant and feature (sufficient for the semi-parametric model as established in Section 4.2). Figure 11 shows the average significance level across participants for the word frequency feature, taking each window length in turn as the baseline. Overall, the extrapolation is highly accurate. As expected, the error is smallest at the baseline window length and grows with distance from it, most notably when extrapolating to the shortest window length (1 s), where the null distribution deviates most from normality and has the largest variance. A baseline window length near the middle of the considered range therefore performs best; here, 5 s. This is confirmed by the percentile error across all features (Figure 12): using 5 s as baseline, the average correlation error remains below approximately 1 percentage point.

This extrapolation methodology allows, for example, to predict how long neural responses need to be measured before a given target correlation ρ_{target} , determined a priori for example based on population data, becomes significant. Using Equation (4), one can sweep w_{target} to find the shortest window length for which $\rho_{\text{target}} > \text{significance level}_{\alpha}(w_{\text{target}})$, without empirically testing each window length in turn (Section 5.1

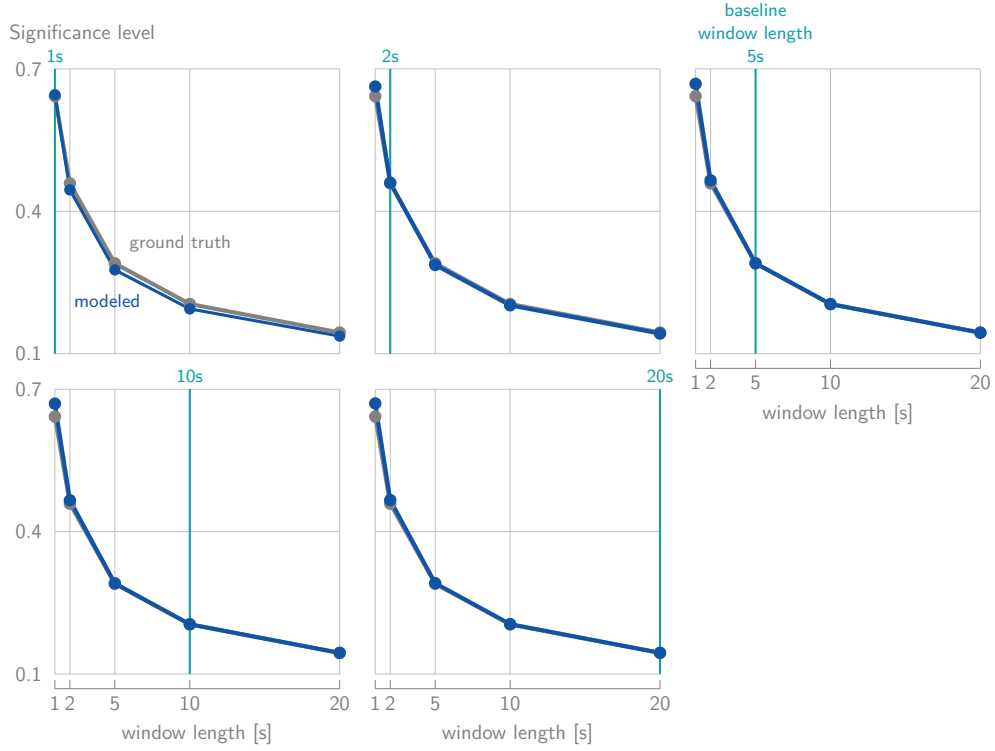


Figure 11: Extrapolating the significance level using the semi-parametric model gives accurate predictions across window lengths, shown here starting from each window length in turn as baseline (shown by the vertical line). The mean significance level (correlation domain) is shown across participants for the word frequency feature.

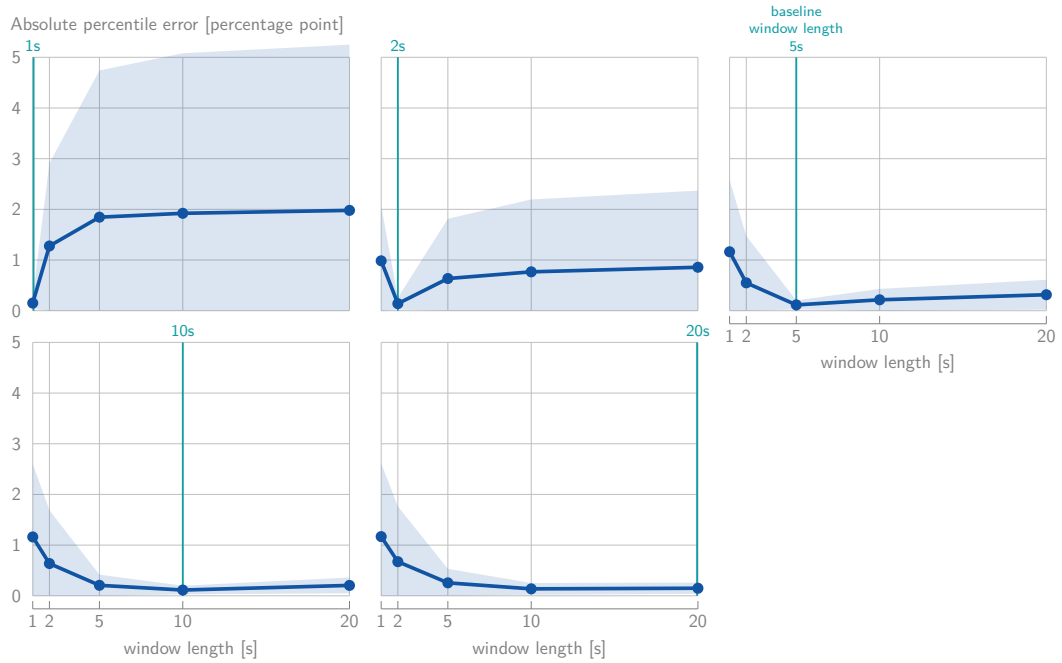


Figure 12: The average (across features and participants; shaded area: ± 1 SD) percentile error, in percentage points, on the extrapolated significance level, starting from various baseline window lengths. As expected, the error is smallest at the baseline window length itself; choosing a baseline near middle of the range, e.g., at 5 s, minimizes error overall.

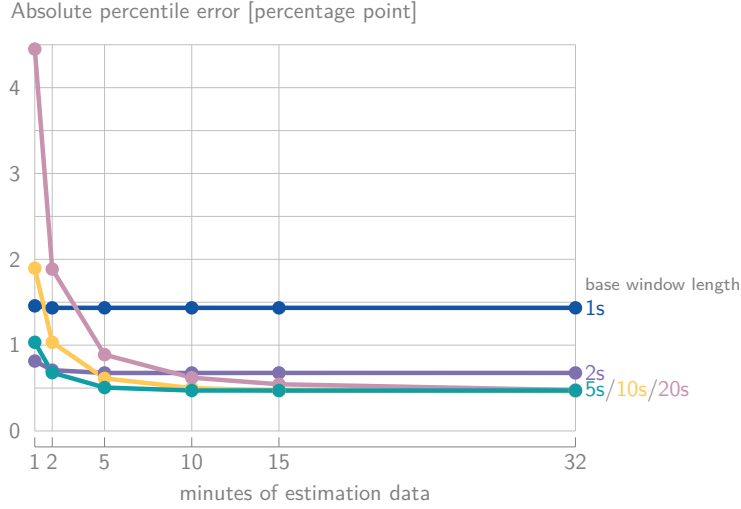


Figure 13: The average (across features, participants, 100 resamplings, and target window lengths) percentile error, in percentage points, on the extrapolated significance level, as a function of the amount of estimation data available, for various baseline window lengths. For each number of minutes of estimation data, the number of null correlations to estimate the significance level from using the semi-parametric model changes depending on the base window length. Overall, the error is smallest starting from 5s, except when only 1 min of estimation data is available; using this baseline, the average percentile error remains below 0.5 percentage points once more than 5 min of data is available.

gives an example). This is relevant, for example, in clinical applications of neural tracking, where it can serve as an objective measure of hearing or speech understanding. In such contexts, determining in advance (so not online and simply based on population data) the minimum recording length needed to reliably detect tracking is valuable for informing the design of efficient protocols.

In summary, the semi-parametric model allows the significance level to be predicted at any window length from a null distribution estimated at a single baseline window length (e.g., 5s), with an average percentile error below 1 percentage point, eliminating the need to populate the null distribution separately at each window length.

4.4. Summary: a practical algorithm

In practice, only a limited amount of estimation data is available to generate null correlations via misalignment (Section 3.1.5), constrained by experiment length and number of trials. This creates a tradeoff: the number of permutations obtainable from a fixed amount of estimation data depends strongly on the baseline window length. For example, 5 min of estimation data yields more than 10 000 permutations at 1 and 2s windows, 3540 permutations at 5s, but only 870 and 210 permutations at 10 and 20s windows, respectively. With very limited estimation data, it can therefore be preferable to extrapolate the significance level from a shorter baseline window length to the desired target window length, rather than estimating it directly at the target length. Figure 13 quantifies this tradeoff, showing the average percentile error on the extrapolated significance level (including at the baseline window length itself) as a function of the amount of estimation data available. Using 5s as baseline (established above as a good default), the average error stays below 0.5 percentage points once more than 5 min of estimation data is available. The resulting procedure using the proposed semi-parametric modeling approach of null distributions, summarized in Algorithm 1, provides an efficient and accurate approach for estimating significance levels across window lengths in neural tracking of natural stimuli. A toolbox, providing among others all code for Algorithm 1 is available in (Geirnaert et al., 2026b).

Key take-away #4

Using the normal distribution after the Fisher transform, the null distribution of neural tracking correlations can be modeled such that only 3 – 5 min (~ 1000 permutations) of data suffice to reliably and efficiently estimate the significance level across window lengths, features, ...

Summarizing algorithm

Algorithm 1 Efficient and accurate estimation of per-window null distributions and significance levels in neural tracking of natural stimuli (Geirnaert et al., 2026b)

Input: (decoded) neural responses and (encoded) stimulus features, sampling frequency f_s , target window length w_{target} , α -level

Output: null distribution and significance level at the α -level for target window length w_{target}

- 1: Use the misalignment method of Section 3.1.5 to generate misaligned pairs of (decoded) neural responses and (encoded) stimulus features. If a single target window length w_{target} is of interest, set $w_{\text{base}} = w_{\text{target}}$; otherwise, set $w_{\text{base}} = 5\text{s}$. Select the misaligned segments so that they are separated from the aligned segment by more than the autocorrelation length of the signals. Aim for at least $n = 1000$ permutations, yielding empirical null correlations $\{r_i^{(\text{null})}\}_{i=1}^n$. Verify the necessary zero-mean condition: $\bar{r}^{(\text{null})} = \frac{1}{n} \sum_{i=1}^n r_i^{(\text{null})} \approx 0$; a substantially non-zero mean indicates residual dependence between misaligned and aligned segments and requires revising the segment selection.
- 2: Apply the Fisher transformation to the empirical null correlations $\{r_i^{(\text{null})}\}_{i=1}^n$:

$$z_i^{(\text{null})} = \text{artanh}(r_i^{(\text{null})})$$

- 3: Model the null distribution as a normal distribution by fitting the variance:

$$\hat{\sigma}_z^{(\text{null})^2} = \frac{1}{n} \sum_{i=1}^n z_i^{(\text{null})^2},$$

such that

$$z^{(\text{null})} \sim \mathcal{N}(0, \hat{\sigma}_z^{(\text{null})^2}).$$

- 4: Compute the significance level at the α -level and at any target window length w_{target} as:

$$\text{significance level}_{\alpha}(w_{\text{target}}) = \tanh\left(P_{100(1-\alpha)\%}\left(\mathcal{N}\left(0, \hat{\sigma}_z^{(\text{null})^2} \frac{V(N_{\text{target}})}{V(N_{\text{base}})}\right)\right)\right),$$

with $N_{\text{base}} = w_{\text{base}} f_s$, $N_{\text{target}} = w_{\text{target}} f_s$, and

$$V(N) = \begin{cases} \frac{\pi^2}{12} - \frac{1}{2} \sum_{k=1}^{\frac{N-4}{2}} k^{-2}, & N_{\text{even}} \\ \frac{\pi^2}{4} - 2 \sum_{k=1}^{\frac{N-3}{2}} (2k-1)^{-2}, & N_{\text{odd}} \end{cases}$$

5. Interpreting and comparing neural tracking correlations

Building on the null distribution model and significance level provided by Algorithm 1, we return to the use case of comparing different features for neural tracking. We show how the significance level can be used to interpret and compare neural tracking correlations across features (Section 5.1), and how the proposed null-normalized tracking score (NNTS) and d-prime, together with their connection to match-mismatch accuracy, offer a more principled alternative (Section 5.2).

In principle, the null distribution and significance levels could be constructed at three different levels of granularity, always separately per feature (since inter-feature differences are expected to dominate):

1. *Across participants*: all windows across all participants are used to build a *single* null distribution. While this has the advantage of simplicity, resulting in one significance level across all participants that is easy to use, it ignores potential differences in EEG distributions and SNRs between participants.
2. *Per participant*: a null distribution across all windows is built per participant. This results in one significance level per participant across all windows, which is a bit more complex to work with, yet respects the differences in EEG distributions and SNRs noted above.
3. *Per individual window*: a null distribution is built for every participant and individual window, resulting in a different significance level per individual window. While this allows taking differences in stimulus properties and non-stationarities in the EEG within participants into account, it is highly cumbersome to work with, and typically insufficient permutations can be generated for reliable per-window null distributions. See also the discussion in Section 3.1.3: to fully exploit per-window characteristics, circular shifting would be the natural choice, yet infeasible in practice.

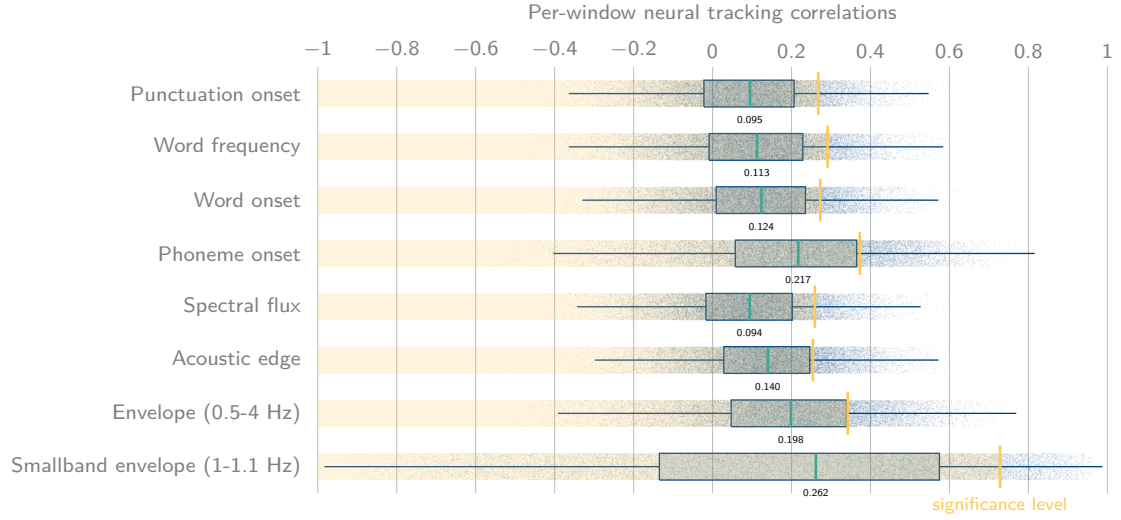
These options trade off convenience against sensitivity to known sources of variability (across features, participants, and individual windows). Here, we adopt the option 2, characterizing a single null distribution - and thus a single significance level - per participant, across all windows. This way, sufficient permutations can be generated per participant, and this granularity naturally aligns with the misalignment technique. Adopting a single per-participant significance level still leaves two ways of using it, both illustrated in the next section: (i) *at the window level*, by comparing each individual window’s correlation against this *same*, shared significance level - this is precisely what distinguishes the second option from the third, where every window would instead have its own, window-specific level; and (ii) *at the participant level*, by deriving the corresponding significance level for the across-window-averaged correlation (Algorithm 2) and testing the participant as a whole.

The window-level significance level used in (i) is obtained directly from Algorithm 1. For testing a participant as a whole as used in (ii), we average the per-window correlations as in Figure 2. Since averaging narrows the null distribution, this requires its own significance level. As shown in Appendix B, the null distribution of the average (non-Fisher-transformed) correlation across m windows per participant is well modeled by a normal distribution, following directly from the central limit theorem⁶. Its parameters follow directly from the same per-window null correlations $\{r_i^{(\text{null})}\}_{i=1}^n$ used in Algorithm 1 (assuming independence between windows, which is typically fulfilled for neural tracking correlations based on ≤ 1 s windows (Heintz et al., 2025)), as summarized in Algorithm 2. An implementation of this algorithm can be found in the accompanying toolbox (Geirnaert et al., 2026b).

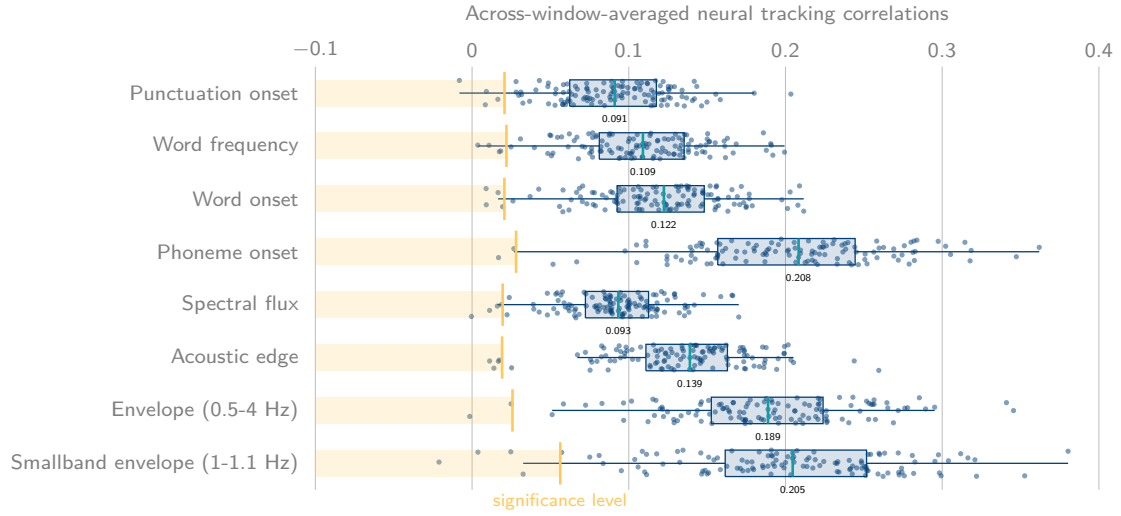
5.1. Comparing features via the significance level

Figure 2 shows, for every feature, the average neural tracking correlations per participant. Without a proper characterization of the null distribution and significance level, one could conclude that the smallband envelope yields the highest neural tracking when interpreting the raw correlation as a direct performance metric. However, computing and plotting the significance level of individual correlations with Algorithm 1, as shown in Figure 14a, provides the necessary context for interpreting and comparing neural tracking correlations across features.

⁶Unlike single correlation coefficients, for which the Fisher transform provides an exact normalizing transformation, no equivalent closed-form characterization exists for *averaged* correlation coefficients; here we rely on the central limit theorem instead.



(a)



(b)

Figure 14: **(a)** The individual neural tracking correlations (one dot = one window) and boxplots across all 5s-windows and participants, for every feature, including the significance levels characterized using Algorithm 1. All correlations below the significance level are non-significance. The per-window significance levels clearly differ across features and show that, for example, smallband envelopes yield, as expected, far fewer significant correlations than other features. **(b)** The per-participant across-window-averaged correlations (one dot = one participant) and boxplots show how the significance level, characterized using Algorithm 2, changes when averaging correlations.

Algorithm 2 Efficient and accurate estimation of per-participant across-window-averaged null distributions and significance levels (Geirnaert et al., 2026b)

Input: per-window empirical null correlations $\{r_i^{(\text{null})}\}_{i=1}^n$ on window length w , number of windows averaged per participant m , α -level

Output: across-window-averaged null distribution and significance level at the α -level

- 1: Estimate the variance of the per-window raw correlations with mean assumed 0:

$$\hat{\sigma}_r^{(\text{null})^2} = \frac{1}{n} \sum_{i=1}^n r_i^{(\text{null})^2},$$

- 2: Compute the significance level of the across-window-averaged null distribution across m (approximately independent) windows ($\bar{r}_m^{(\text{null})} = \frac{1}{m} \sum_{i=1}^m r_i^{(\text{null})}$, modeled as a normal distribution) at the α -level as:

$$\overline{\text{significance level}}_\alpha = P_{100(1-\alpha)\%} \left(\mathcal{N} \left(0, \frac{\hat{\sigma}_r^{(\text{null})^2}}{m} \right) \right).$$

As opposed to Figure 2, Figure 14a shows the individual per-5s-window correlations across all participants, together with the median single-window significance level (usage (i)). This fully reveals the spread of correlations at a given window length, and is consistent with using Algorithm 1, which characterizes the null distribution at the level of *individual* windows. Testing the across-window-averaged correlations of Figure 2 against this same per-window level would be far too conservative, because averaged correlations have a much narrower spread. Applying instead the across-window-averaged significance level (usage (ii); Algorithm 2) yields the participant-level test on 5s windows (averaging 175 windows per participant) in Figure 14b: because the null distribution narrows substantially under averaging, the significance level drops sharply. While most individual 5s-window correlations are non-significant (Figure 14a), neural tracking can be established at the participant level for all features in most participants (Figure 14b).

Figure 14a shows why the correlation scale is not a good scale for comparing across different features. Because the per-window significance level differs between features, the neural tracking correlations are also *incomparable*. While the smallband envelope might yield the highest correlations, its significance level is also much higher than that of other features, resulting in far fewer windows showing above-significance neural tracking. As explained in Section 1, the smallband envelope feature mostly requires phase alignment on a low-frequency signal, which is much easier to achieve, as now confirmed by its much higher significance level. When the significance level is not the same, the raw correlation is therefore not directly comparable across features. Note that using a forward temporal response function instead of a backward model, predicting neural responses from different features, would place every feature’s correlation in the same (EEG) space, which by itself would make significance levels more comparable across features. This does not, however, fully resolve the underlying interpretability problem raised in Section 1: because the significance level still depends on each predictor feature’s own autocorrelation structure, features dominated by low-frequency content would still tend toward inflated correlations and inflated significance levels, regardless of model direction.

As explained in Section 4.3, using Algorithm 1, we can predict how long neural responses to natural stimuli must be minimally measured before significant neural tracking can be established, or, in other words, what minimum window length should be chosen to establish significance for a given feature. Based on the mean correlation for every feature across windows and participants, we extrapolate the overall 5s-window significance level (based on the global null distribution, with 10 000 samples accumulated across windows from all participants) to increasing target window lengths, until the mean correlation exceeds the significance level. Implicitly, we assume that the mean correlation remains fairly constant across window

lengths, as shown in Lopez-Gordo et al. (2025). The resulting predicted measurement times in Table 1 show that, for the average participant, the envelope, acoustic edge, and phoneme onset require much less time to establish significant neural tracking (around 17s) than, for example, the smallband envelope (around 93s). Note that these measurement times are only illustrative of the methodology and specific to this dataset and recording setup, rather than general recommendations. Depending on the requirements, more stringent parameters, e.g., significance with a very strict significance level and for most participants in the population, can be inserted.

Feature	Predicted measurement time to significance [s]
Envelope	16.8
Smallband envelope	93.4
Acoustic edge	17.4
Spectral flux	38.8
Phoneme onset	17.3
Word onset	25.1
Word frequency	35.6
Punctuation onset	43.0

Table 1: Predicted measurement time for the average participant required to measure significant neural tracking for different features, when extrapolating the significance level based on 5 s windows until it is below the mean correlation across participants and windows.

Based on the correlations (per-window or averaged across windows) and their corresponding modeled significance levels, statistical testing could now be performed: features could be compared, for example, based on the number of correlations above the significance level. However, we advocate against statistical testing in the original correlation space using a fixed significance-level threshold. While the significance level remains a valuable tool for interpreting raw correlations, thresholding at a fixed level inherits all the disadvantages of classical statistical testing. When comparing many windows or participants against the significance level, the question of multiple-comparisons correction arises. For example, when simply using the number of windows or participants yielding above-significance correlations as a quantitative metric, there is a case for omitting corrections; however, once windows or participants are selected or reported based on significance, corrections become necessary. Second, thresholding requires an arbitrarily chosen significance level, discarding all continuous evidence. Better alternatives are therefore to use p -values as continuous variables, or NNTS and d-prime (Section 5.2).

As a first alternative, the percentile of an observed neural tracking correlation within the null distribution, i.e., its p -value, can be used as a continuous measure of neural tracking without thresholding (McShane et al., 2019), thereby also quantifying the strength of neural tracking. Multiple comparisons can be naturally controlled by combining p -values using the harmonic mean (Wilson, 2019). However, testing based on p -values is not straightforward, since they are bounded and compressed near the tails. We therefore do not explore p -value-based comparison further here; instead, the following section introduces NNTS as a better alternative.

5.2. Comparing features via NNTS

MacIntyre et al. (2026) proposed using the Z-score of the raw correlations w.r.t. the null distribution as a metric to compare features. Here, we use the same idea, but based on the Fisher-transformed correlations. Given that the Z-score assumes normality of the underlying variables, it is the natural metric to combine with the Fisher-transformed correlations, which follow the normal distribution as shown in Section 4. To emphasize that this neural tracking metric contextualizes the correlations with their null distribution, we dub it the “null-normalized tracking score” (NNTS). Given m Fisher-transformed neural tracking correlations $\{z_i = \text{artanh}(r_i)\}_{i=1}^m$ with sample variance $\hat{\sigma}_z^2$ and null distribution variance $\hat{\sigma}_z^{(\text{null})^2}$, the window-level

NNTSw and participant-level NNTSp are defined as⁷:

$$\begin{cases} \text{NNTSw}_i = \frac{z_i}{\hat{\sigma}_z^{(\text{null})}}, & (\text{window-level}) \\ \text{NNTSp} = \frac{\hat{\mu}_z}{\sqrt{\frac{1}{2}(\hat{\sigma}_z^{(\text{null})^2} + \hat{\sigma}_z^2)}}, & (\text{participant-level}) \end{cases} \quad (5)$$

with $\hat{\mu}_z = \frac{1}{m} \sum_{i=1}^m z_i$, $\hat{\sigma}_z^2 = \frac{1}{m-1} \sum_{i=1}^m (z_i - \hat{\mu}_z)^2$. Note that the denominator differs between the window- and participant-level definitions of NNTS, such that $\text{NNTSp} \neq \frac{1}{m} \sum_{i=1}^m \text{NNTSw}_i$, unless $\hat{\sigma}_z^{(\text{null})^2} \approx \hat{\sigma}_z^2$ (which is often the case). The window-level NNTSw only references the null variance, keeping it a local, per-window metric. The participant-level NNTSp, on the other hand, uses the pooled standard deviation between the null and real distributions, and therefore differs from the averaged NNTSw. In fact, NNTSp is defined this way precisely because it becomes equal to d-prime (if $\hat{\sigma}_z^{(\text{null})^2} = \hat{\sigma}_z^2$, in theory), a signal detectability/sensitivity metric already used in measuring neural tracking (e.g., Di Liberto et al. (2019); Crosse et al. (2021)). Similar to d-prime, this implies that NNTSp is a measure of distributional separability that also accounts for the variability in real correlations across windows. Both NNTSw and NNTSp, as well as the conversion to match-mismatch accuracy (See Section 5.2.1), are implemented in (Geirnaert et al., 2026b).

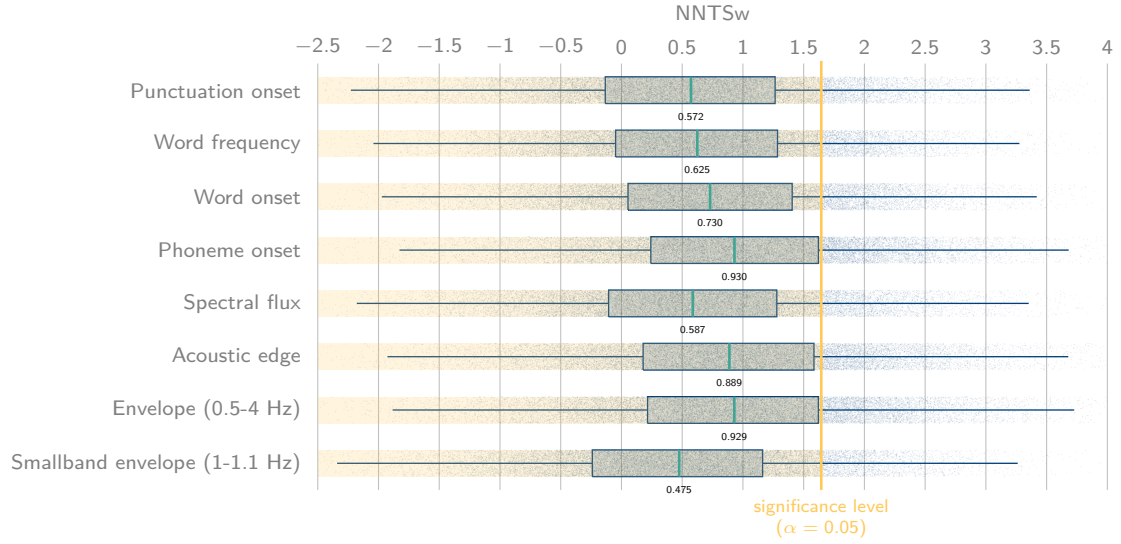
Figure 15a shows the NNTSw across features at the window level, and is therefore comparable with Figure 14a. Here, however, the NNTSw values are directly comparable across features, since normalizing with respect to the null distribution places them on the same scale. This also implies that the significance level, here for $\alpha = 0.05$, lies at the same value for every feature, i.e., $\text{NNTSw} = 1.645$, the 95%-percentile of the standard normal distribution $\mathcal{N}(0, 1)$. It is now clear that the smallband envelope is the worst feature. Furthermore, the envelope, acoustic edge, and phoneme onset seem to perform on par, despite the acoustic edge showing considerably lower raw correlations in Figure 14a. The reason is clear from Figure 14a: the significance level is also lower for the acoustic edge. The NNTS has the advantage that it combines all of these considerations into a single number, facilitating easy comparison between features (and models or conditions, if necessary). A similar conclusion can be drawn from Figure 15b, showing the NNTS at the participant level.

If desired, the NNTSp (shown in Figure 15b) can also be used for statistical testing. Here, just as an example, we run a linear mixed-effects model with feature as a fixed effect and a random intercept per participant, fitted by REML (lme4, R). A type-III ANOVA with Satterthwaite degrees of freedom was used as an omnibus test, showing a significant main effect of feature ($F = 354.87, p < 0.0001$). Given that we are not interested here in any particular comparison and use this merely as a showcasing example, we also run post-hoc pairwise comparisons across all features with Tukey multiple-comparisons correction (emmeans, R). The results in Table 2 show that phoneme onset, the envelope, and acoustic edge form a top-performing tier, each significantly outperforming all remaining features (all $p < 0.0001$, $d > 1.3$). Within this tier, the pairwise comparisons are not fully transitive: the envelope does not significantly differ from either phoneme onset ($p = 0.9950$, $d = 0.10$) or acoustic edge ($p = 0.1475$, $d = 0.34$), yet phoneme onset and acoustic edge do differ significantly from each other ($p = 0.0168$, $d = 0.44$). Word onset forms its own, statistically distinct tier, falling below the top three (all $p < 0.0001$, $d > 1.3$) but above the remaining features (all $p < 0.0001$, $d > 1.0$). Among spectral flux, word frequency, and punctuation onset, effect sizes are generally small ($d < 0.4$ for two of the three pairwise comparisons), although word frequency and punctuation onset still differ significantly ($p < 0.0001$, $d = 0.71$). Finally, the smallband envelope is now clearly the worst-performing feature, significantly underperforming all others (all $p < 0.0001$, $d > 0.9$).

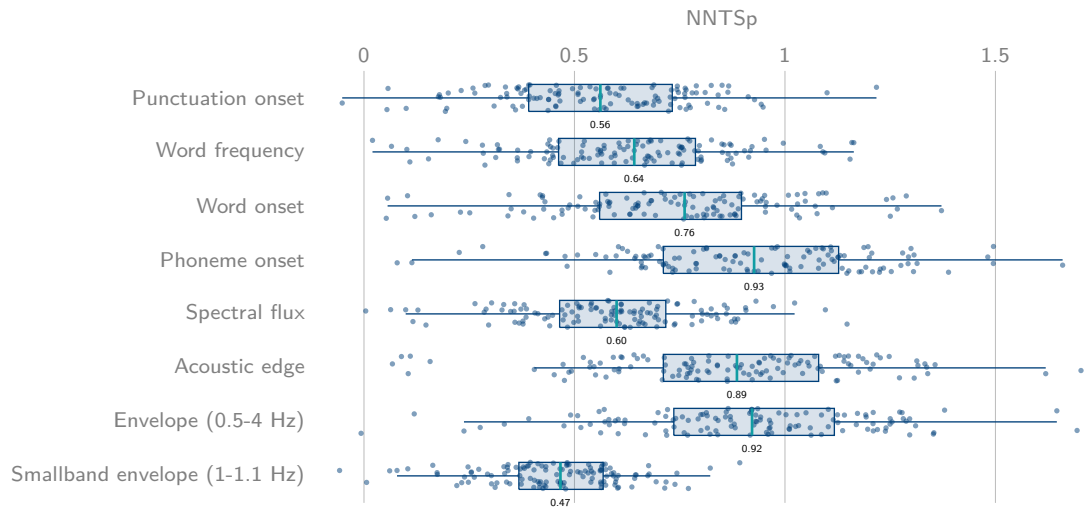
5.2.1. Connection with match-mismatch accuracy

de Cheveigné et al. (2018); de Cheveigné et al. (2021) put forward the match-mismatch task and its

⁷Note that we here implicitly assume that the real correlations too, with mean different from zero, can be modeled using the normal distribution after the Fisher transform. Given the theoretical explanations in Section 4.1 and the verifications on null correlations in Section 4.1.1, we consider this reasonable. Furthermore, this modeling technique has already been successfully employed in (Lopez-Gordo et al., 2025; Geirnaert et al., 2025).



(a)



(b)

Figure 15: **(a)** The null-normalized tracking score (NNTS) at the window level (NNTSw) with 5 s windows allows comparing features on the same scale and with a uniform significance level. **(b)** The NNTS at the participant level (NNTSp) summarizes neural tracking per participant and, again, allows comparing features on the same scale.

Comparison		Estimate	95% CI	$t(840)$	p (Tukey)	Cohen's d
Phoneme onset	Envelope	0.0101	[-0.030, 0.051]	0.76	.9950	0.10
Phoneme onset	Acoustic edge	0.0452	[0.005, 0.086]	3.39	.0168	0.44
Phoneme onset	Word onset	0.1846	[0.144, 0.225]	13.83	< .0001	1.78
Phoneme onset	Word frequency	0.2913	[0.251, 0.332]	21.82	< .0001	2.81
Phoneme onset	Spectral flux	0.3308	[0.290, 0.371]	24.79	< .0001	3.19
Phoneme onset	Punctuation onset	0.3645	[0.324, 0.405]	27.31	< .0001	3.51
Phoneme onset	Smallband envelope	0.4636	[0.423, 0.504]	34.74	< .0001	4.47
Envelope	Acoustic edge	0.0351	[-0.005, 0.076]	2.63	.1475	0.34
Envelope	Word onset	0.1744	[0.134, 0.215]	13.07	< .0001	1.68
Envelope	Word frequency	0.2811	[0.241, 0.322]	21.07	< .0001	2.71
Envelope	Spectral flux	0.3207	[0.280, 0.361]	24.03	< .0001	3.09
Envelope	Punctuation onset	0.3543	[0.314, 0.395]	26.55	< .0001	3.41
Envelope	Smallband envelope	0.4535	[0.413, 0.494]	33.98	< .0001	4.37
Acoustic edge	Word onset	0.1393	[0.099, 0.180]	10.44	< .0001	1.34
Acoustic edge	Word frequency	0.2461	[0.205, 0.287]	18.44	< .0001	2.37
Acoustic edge	Spectral flux	0.2856	[0.245, 0.326]	21.40	< .0001	2.75
Acoustic edge	Punctuation onset	0.3192	[0.279, 0.360]	23.92	< .0001	3.08
Acoustic edge	Smallband envelope	0.4184	[0.378, 0.459]	31.35	< .0001	4.03
Word onset	Word frequency	0.1067	[0.066, 0.147]	8.00	< .0001	1.03
Word onset	Spectral flux	0.1462	[0.106, 0.187]	10.96	< .0001	1.41
Word onset	Punctuation onset	0.1799	[0.139, 0.220]	13.48	< .0001	1.73
Word onset	Smallband envelope	0.2791	[0.238, 0.320]	20.91	< .0001	2.69
Word frequency	Spectral flux	0.0395	[-0.001, 0.080]	2.96	.0621	0.38
Word frequency	Punctuation onset	0.0732	[0.033, 0.114]	5.48	< .0001	0.71
Word frequency	Smallband envelope	0.1723	[0.132, 0.213]	12.91	< .0001	1.66
Spectral flux	Punctuation onset	0.0337	[-0.007, 0.074]	2.52	.1877	0.32
Spectral flux	Smallband envelope	0.1328	[0.092, 0.173]	9.95	< .0001	1.28
Punctuation onset	Smallband envelope	0.0992	[0.059, 0.140]	7.43	< .0001	0.96

Table 2: Pairwise comparisons of NNTSp across features using Tukey multiple comparisons correction, ordered by descending feature mean.

corresponding accuracy as a metric to compare models (or features) in neural tracking of speech. This match-mismatch task, which has been used in several other papers to compare features and models (e.g., Di Liberto et al. (2019); Accou et al. (2021); Puffay et al. (2023); Yao et al. (2024); Monesi et al. (2024); Di Liberto (2026)), consists of deciding whether a particular segment of the stimulus (feature) and neural response are temporally aligned or not. de Cheveigné et al. (2021) consider the match-mismatch accuracy a complementary metric to the correlation and the sensitivity index, akin to the NNTSp/d-prime metric defined in Equation (5).

Here, we close the circle by showing that the match-mismatch accuracy is simply one transformation away from the NNTSp metric. While neural tracking correlations have value, we have shown throughout this paper that careful interpretation with respect to the modeled null distribution is paramount. To facilitate comparison between features and models, we therefore proposed the NNTS metric. Because NNTS is built on the modeled null distribution via Algorithm 1 - which itself relies on the misalignment method - a direct connection with match-mismatch accuracy is to be expected, since match-mismatch accuracy is itself based on comparing aligned with misaligned segments of neural responses and stimulus features.

Mathematically, this can be shown using a similar methodology as in modeling and predicting accuracies in selective auditory attention decoding (Lopez-Gordo et al., 2025; Geirnaert et al., 2025). Considering the Fisher-transformed actual correlations $z = \text{artanh}(r) \sim \mathcal{N}(\hat{\mu}_z, \hat{\sigma}_z^2)$ and misaligned null correlations $z^{(\text{null})} = \text{artanh}(r^{(\text{null})}) \sim \mathcal{N}(0, \hat{\sigma}_z^{(\text{null})^2})$, we have therefore already modeled the matched and mismatched correlations. As artanh is a monotonically increasing function, the match-mismatch decision rule becomes $\text{argmax}(r, r^{(\text{null})}) = \text{argmax}(z, z^{(\text{null})})$. Assuming independence between z and $z^{(\text{null})}$ ⁸, the match-mismatch decision variable can therefore equivalently be defined as:

$$\Delta z = z - z^{(\text{null})} \sim \mathcal{N}(\hat{\mu}_z, \hat{\sigma}_z^2 + \hat{\sigma}_z^{(\text{null})^2}) \lesseqgtr 0.$$

As visualized in Figure 16a, the match-mismatch accuracy can be easily derived from the normal distribution of Δz , with known parameters derived from the modeled real (matched) and null (mismatched) distributions:

$$\begin{aligned} \text{match-mismatch accuracy} &= P(\Delta z > 0) \\ &= \frac{1}{2} \text{erfc} \left(\frac{-\hat{\mu}_z}{\sqrt{2(\hat{\sigma}_z^2 + \hat{\sigma}_z^{(\text{null})^2})}} \right) \\ &= \frac{1}{2} \text{erfc} \left(-\frac{\text{NNTSp}}{2} \right). \end{aligned}$$

This shows that there is a direct mathematical relationship between the NNTSp (d-prime) metric and match-mismatch accuracy. This correspondence is empirically verified in Figure 16b, comparing the predicted (NNTSp-based) and real match-mismatch accuracy at 5s windows. To empirically compute the real match-mismatch accuracy, for every feature and participant, the correlation between the neural response and the actual stimulus window is compared with the correlation against a mismatched stimulus window; the proportion of windows for which the matched correlation exceeds the mismatched one is defined as the match-mismatch accuracy. Across all participants and features, there is a close correspondence between the predicted and actual match-mismatch accuracy, with an overall 0.37 percentage point absolute error, confirming that NNTS and match-mismatch accuracy carry the same information about the underlying distributional separability. Given the well-founded popularity of match-mismatch accuracy, as it implicitly accounts for the nature of the model, stimulus feature(s), SNR, ... when evaluating performance, this mathematical relationship supports the validity of NNTS as a better-suited performance metric for quantifying neural tracking of natural stimuli than raw Pearson correlation coefficients, especially when comparing

⁸While this is not necessarily true, as the underlying neural responses and stimuli are shared, the results in Figure 16b show that, in practice, it is defensible.

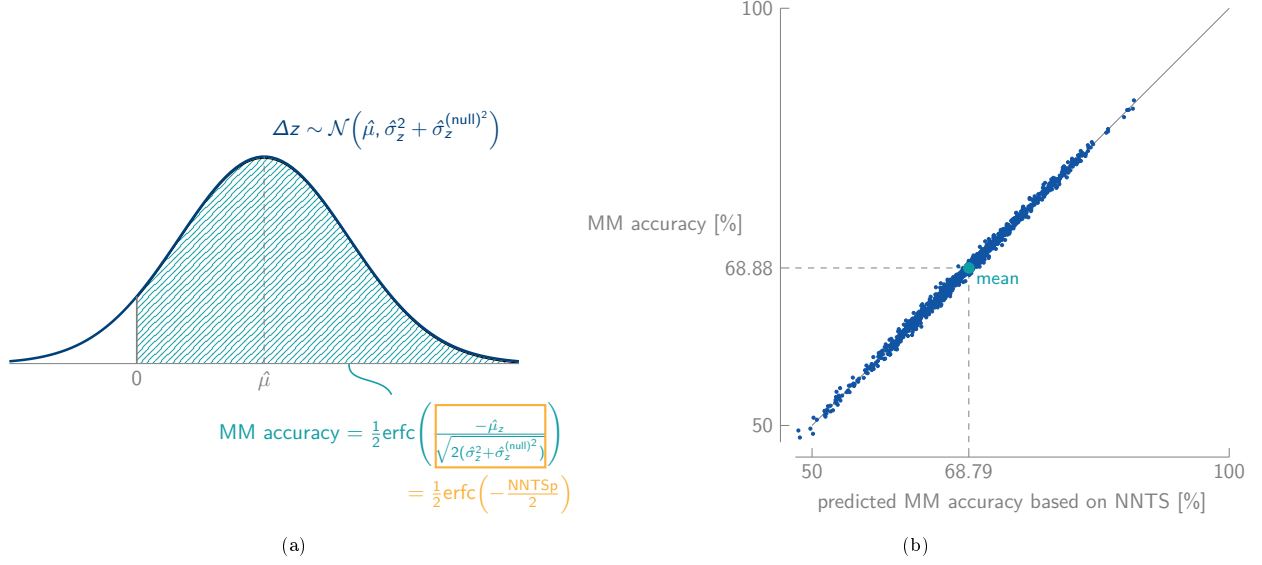


Figure 16: **(a)** The match-mismatch decision variable Δz follows a normal distribution with parameters derived from the modeled real and null distributions, enabling a direct transformation from NNTSp to the match-mismatch accuracy metric. **(b)** The NNTS-predicted match-mismatch (MM) accuracy at 5 s windows across all features and participants closely corresponds with the real, measured MM accuracy.

models and/or stimulus features. Different ways of choosing the mismatched segments, as proposed in Puffay et al. (2023), then simply correspond to different criteria for performing the misalignment used to generate null correlations. Lastly, the extrapolation of the null variance (Equation (3)) and of the real correlation distribution’s variance, following the procedure in Geirnaert et al. (2025), can be readily used to predict the match-mismatch accuracy across window lengths, by correspondingly extrapolating the variance in the computation of the NNTSp (Equation (5)).

Although the two metrics are equivalent, NNTS is preferable in practice. The match-mismatch accuracy is a bounded proportion and therefore saturates: once neural tracking is strong or at long window lengths, accuracies compress towards 100% and become insensitive to further differences between features or models, whereas NNTS is unbounded and continues to resolve them. The empirical accuracy moreover counts binary comparisons and thereby discards the magnitude of the underlying correlations, while NNTSp exploits the full continuous information. Finally, being unbounded and approximately normally distributed, it is also better suited to the parametric statistical testing illustrated above than a bounded proportion.

This more efficient use of the available information also makes the *modeled* match-mismatch accuracy based on NNTSp more reliable when few windows are available. Figure 17 compares the absolute error across features, participants, and 100 resamplings on the match-mismatch accuracy obtained through NNTSp with that of the empirical estimate, as a function of the number of windows m . Each of the m windows is compared with $m - 1$ mismatches, while correspondingly $m(m - 1)$ null correlations are used to compute the NNTSp. For small m , the modeled estimate is clearly more accurate (9.1 versus 11.9 percentage points at $m = 5$), as the empirical proportion is estimated from a limited number of binary comparisons and is correspondingly more variable and quantized. This advantage gradually shrinks as m grows, with both approaches performing equivalently from roughly $m = 100$ windows onward. The residual error of the modeled approach at the largest m reflects its (small) model bias, consistent with the 0.37 percentage point agreement reported above. Deriving the match-mismatch accuracy from NNTSp, if desired, is therefore particularly useful for short recordings, where the empirical accuracy becomes unreliable.

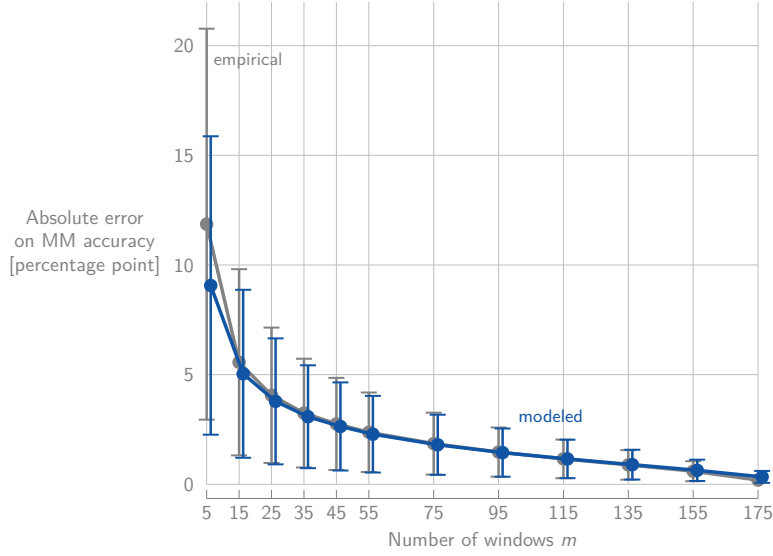


Figure 17: The mean $\pm$ SD (across participants, features, and 100 resamplings) absolute error on the match-mismatch accuracy, obtained either through the modeled NNTSp or empirically, as a function of the number of windows m . The modeled approach is more accurate for small m , with both converging for larger m .

Key take-away #5

Based on the modeled (null) distributions, raw neural tracking correlations can be more reliably interpreted. Features and models can be directly compared using the null-normalized tracking score (NNTS), which has a direct connection with the match-mismatch accuracy, yet is continuous and unbounded, and yields more reliable estimates when only limited data is available.

6. Conclusion

In this paper, we have provided a comprehensive framework to interpret and compare correlations between features and models in neural tracking of natural stimuli such as speech, music, and video. This framework consists of efficient and accurate modeling of null distributions and significance levels, and a new metric to more accurately quantify neural tracking, i.e., the null-normalized tracking score (NNTS). Interpreting neural tracking correlations in light of the null distribution and correct significance level is crucial, as the correlation can differ substantially in nature across, e.g., features and models. To facilitate adoption of the framework, all proposed methods are implemented in an accompanying toolbox (Geirnaert et al., 2026b).

First, we showed that different permutation techniques for generating null correlations are not assumption-free: each implicitly assumes different properties of the underlying signals. We adopted misalignment as the default method for generating null correlations, arguing that it operationalizes the null hypothesis appropriate for content-specific tracking. Second, to address the need for generating many null correlations, we showed that neural tracking correlations can be well modeled using the normal distribution after the Fisher transform. This semi-parametric model enables accurate estimation of the significance level. Moreover, because the null distribution is well modeled even from few null correlations, it enables *efficient* significance-level estimation. The semi-parametric model yields better significance-level estimates than the empirical null distribution for fewer than 10 000 permutations, with excellent approximations for as low as 1000 permutations. Third, the significance level can be reliably extrapolated from one window length to another using the model, without needing to recompute null correlations at different window lengths, greatly reducing computational cost. Combined, we established that, using 5 s windows, 3 – 5 min of data suffices to reliably estimate the null distribution and significance level.

Beyond this null-distribution and significance-level methodology, we also showed how features can be interpreted using the significance level at the window and participant level. To enable direct comparison of features and models on the same scale, we proposed the null-normalized tracking score (NNTS) and demonstrated its use in statistical testing of neural tracking. Moreover, we mathematically and empirically showed that NNTS has a direct relationship with the often-used match-mismatch accuracy and that deriving it from NNTS is more reliable when few windows are available, providing a coherent framework to interpret and compare features and models in neural tracking, beyond simply - and incorrectly - using raw correlations as a performance metric.

The validation and use case in this paper were performed on a single dataset, using backward modeling of speech features in normal-hearing adults. The proposed statistical framework itself is agnostic to stimulus modality and recording technique, and we see no specific reason it would not extend to other naturalistic stimuli (e.g., music, video) or modalities (e.g., MEG, ECoG). However, this generalization remains to be empirically demonstrated.

Data and Code Availability

The raw EEG data are part of the publicly available SparrKULee dataset (Accou et al., 2024); data from the additional participants included here could not be made publicly available. The neural tracking correlations and null correlations derived from these recordings, which underlie all analyses from Section 4 onward, are available at Geirnaert et al. (2026c). All MATLAB and R code required to reproduce the results, figures, and tables from Section 4 onward is available at Geirnaert et al. (2026a). A toolbox implementing the proposed methodology, including the misalignment procedure, the semi-parametric null distribution model, significance level estimation and extrapolation, and the NNTS metric, is available at <https://github.com/Gonaz/NTC> (Geirnaert et al., 2026b).

Author Contributions

Simon Geirnaert: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization, Project administration, Funding acquisition. **Alexander Bertrand:** Methodology, Writing - Review & Editing, Supervision, Funding acquisition. **Tom Francart:** Writing - Review & Editing, Supervision, Funding acquisition. **Jonas Vanthornhout:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Data Curation, Funding acquisition.

Funding

Financial support was provided by the Research Foundation Flanders (FWO) (junior postdoctoral fellowship fundamental research 1242524N for S. Geirnaert; project No G081722N and G026026N), Internal Funds KU Leuven (projects IDN/23/006, C14/25/108, C3/25/017), the European Research Council (ERC) under the European Union’s research and innovation programme (grant agreement No 802895 and grant agreement 101138304), and the Flemish Government (AI Research Program). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or any of the granting authorities. Neither the European Union nor the granting authorities can be held responsible for them.

Declaration of Competing Interests

The authors declare no competing interests.

Acknowledgments

Claude Sonnet (version 4.6; Anthropic) was used as a writing support for language editing and grammar checking, and as a coding assistant for software development and implementation support. All text and code was originally written by the authors, and all algorithms, analyses, and interpretations were designed, verified, and validated by the authors, who take full responsibility for the results. All authors have reviewed and approved the final content of the manuscript and take full responsibility for the paper’s scientific integrity.

References

- Accou, B., Bollens, L., Gillis, M., Verheijen, W., Van hamme, H., Francart, T., 2024. SparrKULee: A Speech-Evoked Auditory Response Repository from KU Leuven, Containing the EEG of 85 Participants. Data 9. doi:10.3390/data9080094.
- Accou, B., Jalilpour Monesi, M., Van hamme, H., Francart, T., 2021. Predicting speech intelligibility from EEG in a non-linear classification paradigm. *Journal of Neural Engineering* 18, 066008. doi:10.1088/1741-2552/ac33e9.
- Aiken, S.J., Picton, T.W., 2008. Human Cortical Responses to the Speech Envelope. *Ear and Hearing* 29, 139–157. doi:10.1097/AUD.0b013e31816453dc.
- Alickovic, E., Lunner, T., Gustafsson, F., Ljung, L., 2019. A Tutorial on Auditory Attention Identification Methods. *Frontiers in Neuroscience* 13, 153. doi:10.3389/fnins.2019.00153.
- Asuero, A.G., Sayago, A., González, A.G., 2006. The Correlation Coefficient: An Overview. *Critical Reviews in Analytical Chemistry* 36, 41–59. doi:10.1080/10408340500526766.
- Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., et al., 2018. Redefine statistical significance. *Nature Human Behaviour* 2, 6–10. doi:10.1038/s41562-017-0189-z.
- Biesmans, W., Das, N., Francart, T., Bertrand, A., 2017. Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 402–412. doi:10.1109/TNSRE.2016.2571900.
- Brodbeck, C., Hong, L.E., Simon, J.Z., 2018. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology* 28, 3976–3983. doi:10.1016/j.cub.2018.10.042.
- Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2018. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology* 28, 803–809. doi:10.1016/j.cub.2018.01.080.
- Chen, Y.P., Schmidt, F., Keitel, A., Rösch, S., Hauswald, A., Weisz, N., 2023. Speech intelligibility changes the temporal evolution of neural speech tracking. *NeuroImage* , 119894doi:10.1016/j.neuroimage.2023.119894.
- de Cheveigné, A., Slaney, M., Fuglsang, S.A., Hjortkjaer, J., 2021. Auditory stimulus-response modeling with a match-mismatch task. *Journal of Neural Engineering* 18, 046040. doi:10.1088/1741-2552/abf771.
- Combrisson, E., Jerbi, K., 2015. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods* 250, 126–136. doi:10.1016/j.jneumeth.2015.01.010.
- Crosse, M.J., Foxe, J.J., Molholm, S., 2024. PERMUTOOLS: A MATLAB Package for Multivariate Permutation Testing. arXiv preprint arXiv:2401.09401. doi:10.48550/arXiv.2401.09401.

- Crosse, M.J., Zuk, N.J., Di Liberto, G.M., Nidiffer, A.R., Molholm, S., Lalor, E.C., 2021. Linear Modeling of Neurophysiological Responses to Speech and Other Continuous Stimuli: Methodological Considerations for Applied Research. *Frontiers in Neuroscience* 15, 705621. doi:10.3389/fnins.2021.705621.
- Daeglau, M., Otten, J., Grimm, G., Mirkovic, B., Hohmann, V., Debener, S., 2025. Neural speech tracking in a virtual acoustic environment: audio-visual benefit for unscripted continuous speech. *Frontiers in Human Neuroscience* 19, 1560558. doi:10.3389/fnhum.2025.1560558.
- de Cheveigné, A., Wong, D.D., Di Liberto, G.M., Hjortkjær, J., Slaney, M., Lalor, E., 2018. Decoding the auditory brain with canonical component analysis. *NeuroImage* 172, 206–216. doi:10.1016/j.neuroimage.2018.01.033.
- Di Liberto, G.M., 2026. Robust Evaluation of Neural Encoding Models via ground-truth approximation. arXiv preprint arXiv:2604.14694. doi:10.48550/arXiv.2604.14694.
- Di Liberto, G.M., Marion, G., Shamma, S.A., 2021. Accurate Decoding of Imagined and Heard Melodies. *Frontiers in Neuroscience* 15. doi:10.3389/fnins.2021.673401.
- Di Liberto, G.M., O’Sullivan, J.A., Lalor, E.C., 2015. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology* 25, 2457–2465. doi:10.1016/j.cub.2015.08.030.
- Di Liberto, G.M., Pelofi, C., Bianco, R., Patel, P., Mehta, A.D., Herrero, J., de Cheveigné, A., Shamma, S.A., Mesgarani, N., 2020. Cortical encoding of melodic expectations in human temporal cortex. *eLife* 9, e51784. doi:10.7554/eLife.51784.
- Di Liberto, G.M., Wong, D., Melnik, G.A., de Cheveigné, A., 2019. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *NeuroImage* 196, 237–247. doi:10.1016/j.neuroimage.2019.04.037.
- Ding, N., Simon, J.Z., 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology* 107, 78–89. doi:10.1152/jn.00297.2011.
- Dmochowski, J.P., Ki, J.J., DeGuzman, P., Sajda, P., Parra, L.C., 2018. Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. *NeuroImage* 180, 134–146. doi:10.1016/j.neuroimage.2017.05.037.
- Etard, O., Reichenbach, T., 2019. Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise. *Journal of Neuroscience* 39, 5750–5759. doi:10.1523/JNEUROSCI.1828-18.2019.
- Fiedler, L., Wöstmann, M., Herbst, S.K., Obleser, J., 2019. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* , 33–42doi:10.1016/j.neuroimage.2018.10.057.
- Fisher, R., 1921. On the ‘Probable Error’ of a Coefficient of Correlation Deduced from a Small Sample. *Metron* 1, 1–32.
- Fisher, R.A., 1935. *The Design of Experiments*. Hafner Press, New York.
- Fouladi, R.T., Steiger, J.H., 2008. The Fisher transform of the Pearson Product Moment Correlation Coefficient and Its Square: Cumulants, Moments, and Applications. *Communications in Statistics - Simulation and Computation* 37, 928–944. doi:10.1080/03610910801943735.
- Geirnaert, S., Bertrand, A., Francart, T., Vanthornhout, J., 2026a. Code to reproduce results of paper “Modeling and Interpreting Correlations, Null Distributions and Significance Levels in Neural Tracking of Natural Stimuli” (Version v1.0.0) [Computer software]. URL: <https://doi.org/10.5281/zenodo.21821416>, doi:10.5281/zenodo.21821416.

- Geirnaert, S., Bertrand, A., Francart, T., Vanthornhout, J., 2026b. Neural Tracking Correlation Toolbox (Version v1) [Computer software]. URL: <https://github.com/Gonaz/NTC>, doi:10.5281/zenodo.21827915.
- Geirnaert, S., Bertrand, A., Francart, T., Vanthornhout, J., 2026c. Neural tracking correlations and null distributions of paper 'Modeling and Interpreting Correlations, Null Distributions and Significance Levels in Neural Tracking of Natural Stimuli' [Dataset]. URL: <https://doi.org/10.5281/zenodo.21819899>, doi:10.5281/zenodo.21819899.
- Geirnaert, S., Francart, T., Bertrand, A., 2021a. Unsupervised Self-Adaptive Auditory Attention Decoding. *IEEE Journal on Biomedical and Health Informatics* 25, 3955–3966. doi:10.1109/JBHI.2021.3075631.
- Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigné, A., Lalor, E., Meyer, B.T., Miran, S., Francart, T., Bertrand, A., 2021b. Electroencephalography-based auditory attention decoding: Toward neurosteered hearing devices. *IEEE Signal Processing Magazine* 38, 89–102. doi:10.1109/msp.2021.3075932.
- Geirnaert, S., Vanthornhout, J., Francart, T., Bertrand, A., 2025. Performance Modeling for Correlation-Based Neural Decoding of Auditory Attention to Speech, in: 2025 33rd European Signal Processing Conference (EUSIPCO), pp. 1467–1471. doi:10.23919/EUSIPCO63237.2025.11226424.
- Gillis, M., Vanthornhout, J., Simon, J.Z., Francart, T., Brodbeck, C., 2021. Neural Markers of Speech Comprehension: Measuring EEG Tracking of Linguistic Speech Representations, Controlling the Speech Acoustics. *Journal of Neuroscience* 41, 10316–10329. doi:10.1523/JNEUROSCI.0812-21.2021.
- Heintz, N., Francart, T., Bertrand, A., 2025. Post-processing of EEG-based Auditory Attention Decoding Decisions via Hidden Markov Models. *arXiv preprint arXiv:2506.24024*. doi:10.48550/arXiv.2506.24024.
- Holt, C.A., Sullivan, S.P., 2023. Permutation tests for experimental data. *Experimental Economics* 26, 775–812. doi:10.1007/s10683-023-09799-6.
- Horrace, W.C., 2015. Moments of the truncated normal distribution. *Journal of Productivity Analysis* 43, 133–138. doi:10.1007/s11123-013-0381-8.
- Hotelling, H., 1953. New light on the Correlation Coefficient and its Transforms. *Journal of the Royal Statistical Society. Series B (Methodological)* 15, 193–232. doi:10.1111/j.2517-6161.1953.tb00135.x.
- Kendall, M.G., Stuart, A., 1979. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. 4th ed., Charles Griffin & Company, London.
- Ki, J.J., Parra, L.C., Dmochowski, J.P., 2020. Visually evoked responses are enhanced when engaging in a video game. *European Journal of Neuroscience* 52, 4695–4708. doi:10.1111/ejn.14924.
- Lalor, E.C., Foxe, J.J., 2010. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience* 31, 189–193. doi:10.1111/j.1460-9568.2009.07055.x.
- Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., Stefanovska, A., 2018. Surrogate data for hypothesis testing of physical systems. *Physics Reports* 748, 1–60. doi:10.1016/j.physrep.2018.06.001.
- Lopez-Gordo, M.A., Geirnaert, S., Bertrand, A., 2025. Unsupervised Accuracy Estimation for Brain-Computer Interfaces Based on Selective Auditory Attention Decoding. *IEEE Transactions on Biomedical Engineering* 72, 2388–2399. doi:10.1109/TBME.2025.3542253.
- MacIntyre, A.D., Gaultier, C., Goehring, T., 2026. Decoding of speech acoustics from EEG: going beyond the amplitude envelope. *Journal of Neural Engineering* 23, 016014. doi:10.1088/1741-2552/ae3ae1.

- Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods* 164, 177–190. doi:10.1016/j.jneumeth.2007.03.024.
- McShane, B.B., Gal, D., Gelman, A., Robert, C., Tackett, J.L., 2019. Abandon Statistical Significance. *The American Statistician* 73, 235–245. doi:10.1080/00031305.2018.1527253.
- Monesi, M.J., Bollens, L., Accou, B., Vanthornhout, J., Van Hamme, H., Francart, T., 2024. Auditory EEG Decoding Challenge for ICASSP 2023. *IEEE Open Journal of Signal Processing* 5, 652–661. doi:10.1109/OJSP.2024.3376296.
- Noguchi, K., Konietzschke, F., Marmolejo-Ramos, F., Pauly, M., 2021. Permutation tests are robust and powerful at 0.5% and 5% significance levels. *Behavior Research Methods* 53, 2712–2724. doi:10.3758/s13428-021-01595-5.
- O’Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2014. Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex* 25, 1697–1706. doi:10.1093/cercor/bht355.
- Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing Speech from Human Auditory Cortex. *PLoS Biology* 10, e1001251. doi:10.1371/journal.pbio.1001251.
- Pitman, E.J.G., 1937. Significance Tests Which May Be Applied to Samples From Any Populations. Supplement to the *Journal of the Royal Statistical Society* 4, 119–130. doi:10.2307/2984124.
- Puffay, C., Accou, B., Bollens, L., Monesi, M.J., Vanthornhout, J., Van hamme, H., Francart, T., 2023. Relating EEG to continuous speech using deep neural networks: a review. *Journal of Neural Engineering* 20, 041003. doi:10.1088/1741-2552/ace73f.
- Reddy Katthi, J., Ganapathy, S., 2021. Deep Correlation Analysis for Audio-EEG Decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29, 2742–2753. doi:10.1109/TNSRE.2021.3129790.
- Roebben, A., Heintz, N., Geirnaert, S., Francart, T., Bertrand, A., 2024. ‘Are you even listening?’-EEG-based decoding of absolute auditory attention to natural speech. *Journal of Neural Engineering* 21, 036046. doi:10.1088/1741-2552/ad5403.
- Schreiber, T., Schmitz, A., 1996. Improved Surrogate Data for Nonlinearity Tests. *Physical Review Letters* 77, 635–638. doi:10.1103/PhysRevLett.77.635.
- Sturm, I., Dähne, S., Blankertz, B., Curio, G., 2015. Multi-Variate EEG Analysis as a Novel Tool to Examine Brain Responses to Naturalistic Music Stimuli. *PLOS ONE* 10, e0141281. doi:10.1371/journal.pone.0141281.
- Thornton, M., Mandic, D., Reichenbach, T., 2022. Robust decoding of the speech envelope from EEG recordings through deep neural networks. *Journal of Neural Engineering* 19, 046007. doi:10.1088/1741-2552/ac7976.
- Vanthornhout, J., Decruy, L., Wouters, J., Simon, J.Z., Francart, T., 2018. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope. *Journal of the Association for Research in Otolaryngology* 19, 181–191. doi:10.1007/s10162-018-0654-z.
- Verschueren, E., Somers, B., Francart, T., 2019. Neural envelope tracking as a measure of speech understanding in cochlear implant users. *Hearing Research* 373, 23–31. doi:10.1016/j.heares.2018.12.004.

- Wilroth, J., Silva, N.S., Tafakkor, A., de Avo Mesquita, B., Ip, E.Y.J., Lau, B., Hannah, J., Di Liberto, G.M., 2026. Investigating neural speech processing with functional near infrared spectroscopy: considerations for temporal response functions. *bioRxiv preprint bioRxiv:2026.03.20.713212*. doi:10.64898/2026.03.20.713212.
- Wilson, D.J., 2019. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* 116, 1195–1200. doi:10.1073/pnas.1814092116.
- Wong, D.D., Fuglsang, S.A., Hjortkjær, J., Ceolini, E., Slaney, M., de Cheveigné, A., 2018. A Comparison of Regularization Methods in Forward and Backward Models for Auditory Attention Decoding. *Frontiers in Neuroscience* 12. doi:10.3389/fnins.2018.00531.
- Yao, Y., Stebner, A., Tuytelaars, T., Geirnaert, S., Bertrand, A., 2024. Identifying Temporal Correlations Between Natural Single-shot Videos and EEG Signals. *Journal of Neural Engineering* 21, 016018. doi:10.1088/1741-2552/ad2333.
- Zuk, N.J., Murphy, J.W., Reilly, R.B., Lalor, E.C., 2021. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLOS Computational Biology* 17, 1–32. doi:10.1371/journal.pcbi.1009358.

Appendix A. Comparison of the Fisher transform with other modeling techniques

We consider the following three alternatives to the normal distribution after the Fisher transform for modeling the null distribution of Pearson correlation coefficients between neural responses and stimulus features:

Normal distribution The null correlations $r^{(\text{null})}$ can be modeled directly with a zero-mean normal distribution $r^{(\text{null})} \sim \mathcal{N}\left(0, \sigma_r^{(\text{null})^2}\right)$, with the zero mean following automatically from the null hypothesis and the variance estimated directly from the generated null correlations. This is loosely motivated by the central limit theorem, since the Pearson correlation coefficient involves a summation over many samples. Strictly, normality does not hold here: the time samples are not independent, and the correlation is bounded to $[-1, 1]$. Nonetheless, it can be a convenient choice in practice, especially for long windows, where many samples are summed and the variance of the correlation is small.

Truncated normal distribution To make the normal model theoretically more sound, the truncated normal distribution bounds it to $[-1, 1]$. The mean remains zero, while the standard deviation is obtained by the method of moments, i.e., by equating the sample variance with the variance of the distribution (Horrace, 2015).

Student’s t-distribution Inspired by significance testing for Pearson correlation coefficients, the null distribution can also be modeled with a Student’s t-distribution. As shown in Kendall and Stuart (1979), under the transformation

$$t = r \sqrt{\frac{\text{DOF}}{1 - r^2}} \sim \text{t-distribution}(\text{DOF}),$$

t follows a Student’s t-distribution when the underlying variables are normally distributed and uncorrelated, with DOF the degrees-of-freedom parameter. Here, we treat DOF as a parameter to be estimated, again by the method of moments after transformation.

We compare these methods against the normal distribution after the Fisher transform using the same significance-level estimation metrics as in Section 4.1.2: the relative correlation error (Table A.3a) and absolute percentile error (Table A.3b). These per-feature errors, averaged across participants and window lengths, show that the normal distribution after the Fisher transform consistently outperforms all alternatives, most clearly for edge cases such as the smallband envelope and phoneme and punctuation onsets, and for shorter windows, where the distributions deviate most from normality.

	normal	truncated normal	Student t	normal after Fisher
Envelope	0,64	1,08	0,77	0,43
Smallband envelope	4,96	2,47	2,24	0,31
Acoustic edge	0,51	0,63	0,49	0,52
Spectral flux	0,56	0,55	0,47	0,46
Phoneme onsets	0,86	1,12	1,13	0,57
Word onsets	0,43	0,67	0,42	0,45
Word frequency	0,49	0,76	0,46	0,50
Punctuation onsets	1,04	0,98	0,91	0,87
Mean	1,19	1,03	0,86	0,51

(a)

	normal	truncated normal	Student t	normal after Fisher
Envelope	0,14	0,24	0,17	0,08
Smallband envelope	1,88	1,05	1,04	0,10
Acoustic edge	0,09	0,12	0,09	0,10
Spectral flux	0,10	0,10	0,09	0,08
Phoneme onsets	0,36	0,46	0,15	0,26
Word onsets	0,09	0,15	0,09	0,10
Word frequency	0,09	0,14	0,09	0,09
Punctuation onsets	0,17	0,20	0,20	0,20
Mean	0,37	0,31	0,24	0,13

(b)

Table A.3: The average (across participants and window lengths) **(a)** relative correlation error on the estimated 95%-significance level and **(b)** absolute percentile error on the nominal $\alpha = 0.05$ level. Both are lowest overall for the normal distribution after the Fisher transform, with the largest differences at shorter window lengths for the smallband envelope, phoneme onset, and punctuation onset.

Appendix B. Modeling the null distribution of across-window-averaged correlations

As opposed to single Pearson correlation coefficients, there is no rigorous characterization for averaged (null) correlation coefficients $\bar{r}_m^{(\text{null})} = \frac{1}{m} \sum_{i=1}^m r_i^{(\text{null})}$ across m windows. The normal distribution after Fisher transform for single correlations does not offer a way out, as the hyperbolic tangent function does not commute with the sum. Therefore, the most relevant mechanism to characterize the distribution of $\bar{r}_m^{(\text{null})}$ seems to be central limit theorem, which states that its distribution should converge to a normal distribution. Given estimated variance of the per-window raw correlations (with mean assumed 0):

Unlike single Pearson correlation coefficients, averaged null correlations $\bar{r}_m^{(\text{null})} = \frac{1}{m} \sum_{i=1}^m r_i^{(\text{null})}$ across m windows have no rigorous distributional characterization. The normal distribution after the Fisher transform, used for single correlations, offers no way out here, since the hyperbolic tangent does not commute with averaging: the Fisher transform of the mean correlation is not the mean of the Fisher-transformed correlations. The most natural remaining route is the central limit theorem, which - assuming the per-window null correlations are approximately independent - implies that $\bar{r}_m^{(\text{null})}$ converges to a normal distribution as m grows. Given the estimated per-window null variance (with mean assumed zero),

$$\hat{\sigma}_r^{(\text{null})^2} = \frac{1}{n} \sum_{i=1}^n r_i^{(\text{null})^2},$$

the averaged null correlation is then modeled as

$$\bar{r}_m^{(\text{null})} \sim \mathcal{N}\left(0, \frac{\hat{\sigma}_r^{(\text{null})^2}}{m}\right).$$

This procedure is summarized in Algorithm 2.

We compare the estimated 95%-significance levels for the average neural tracking correlation on 5s windows, obtained via Algorithm 2, against the empirical estimate, for m ranging from 2 to 175 (the

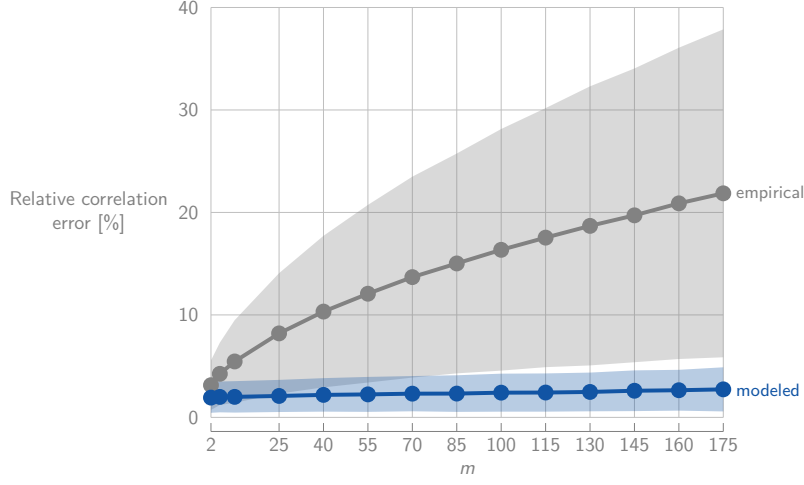


Figure B.18: The average (across participants and features; shaded area: ± 1 SD) relative correlation error on the estimated 95%-significance level of the average correlation as a function of the number of averaged windows m .

maximum number of 5 s windows available for averaging). The ‘ground-truth’ null distribution of average correlations is constructed by computing, 10 000 times, the average of m correlations subsampled from the original 100 000 null correlations. For each m , participant, and feature, both the empirical and the modeled significance-level estimate are based on 1000 randomly selected null correlations (based on the findings in Section 4.2). For the empirical estimate specifically, the null distribution is built from 10 000 averages, each subsampled from these 1000 selected correlations.

Figure B.18 shows the average relative correlation error on the 95%-significance level across participants and features. The modeled approach based on the normal distribution achieves low errors throughout, averaging 2.31% across m . In contrast, the empirical estimate degrades substantially as m grows, since fewer independent draws can be formed for larger m . While the model’s relative errors are slightly higher than for individual correlations (cf. Figure 9a at 1000 permutations, with 1.61% error), they remain low overall.